PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# Evaluation of methods for taxonomic relation extraction from text

ROGER LEITZKE GRANADA

Doctoral Dissertation presented as partial requirement for obtaining the Doctor's degree in Computer Science from Pontifical Catholic University of Rio Grande do Sul.

Advisor: Renata Vieira
Advisor: Nathalie Aussenac-Gilles
Co-Advisor: Cássia Trojahn dos Santos

**Porto Alegre**
**2015**

**Dados Internacionais de Catalogação na Publicação (CIP)**

G748e Granada, Roger Leitzke

Evaluation of methods for taxonomic relation from text / Roger Leitzke Granada. – 2015.
151 f.

Tese (Doutorado) – Faculdade de Informática, PUCRS.
Orientadores: Prof<sup>a</sup>. Dr<sup>a</sup>. Renata Vieira e Prof<sup>a</sup>. Dr<sup>a</sup>. Nathalie Aussenac-Gilles.

1. Ontologia. 2. Processamento da Linguagem Natural. 3. Sistemas de Informação. 4. Informática. I. Vieira, Renata. II. Aussenac-Gilles, Nathalie. III. Título.

CDD 23 ed. 006.35

**Ramon Ely CRB 10/2165**
**Setor de Tratamento da Informação da BC-PUCRS**

FACIN PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "*Evaluation of methods for taxonomic relation extraction from text*" apresentada por Roger Leitzke Granada como parte dos requisitos para obtenção do grau de Doutor em Ciência da Computação, aprovada em 28 de setembro de 2015 pela Comissão Examinadora:

Renata Vieira

Profa. Dra; Renata Vieira– Orientadora — PPGCC/PUCRS

Profa. Dra. Nathalie Aussenac-Gilles – Orientadora Cotutela — IRIT

Vera Lúcia Strube de Lima

Profa. Dra. Vera Lúcia Strube de Lima – — PPGCC/PUCRS

Prof. Dr. Renato Rocha Souza - — FGV

Finatto

Profa. Dra. Maria José Bocorny Finatto - — UFRGS

Profa. Dra. Cássia Trojahn dos Santos - — IRIT

Prof. Dr. Boughanem Mohand - — Université Toulouse III

Profa. Dra. Adeline Nazarenko - — Université Paris 13 Nord

Homologada em 16 / 06 / 2016, conforme Ata No. 012 pela Comissão Coordenadora.

Fernandes

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

*"Reality is merely an illusion, albeit a very persistent one."*

*Albert Einstein*

# ACKNOWLEDGEMENTS

First of all, probably I would never have been able to finish this thesis without the guidance of my advisors, help from friends, and support from my family.

I would like to express my deepest gratitude to my advisor, Renata Vieira, for her guidance, patience, and providing me with an excellent atmosphere for doing research. I would also like to express my very sincere gratitude to Nathalie Aussenac-Gilles from the Institut de Recherche en Informatique de Toulouse for the support to make this thesis possible. My very sincere thanks to Cássia Trojahn dos Santos for her support, guidance, generous care and the home feeling whenever I was in need during my stay in Toulouse.

Thanks to my colleagues from the NLP Research Group at PUCRS, Lucelene Lopes, Larissa Freitas, Sandra Collovini, Daniela do Amaral, and Evandro Fonseca who were always willing to help and give the best suggestions. It would have been a lonely lab without you. A special thanks to Carlos Ramisch for his support and encouragements while doing research.

I would like to also thank to all people from Fondazione Bruno Kessler (FBK) and from Institut de Recherche en Informatique de Toulouse (IRIT) for providing excellent working conditions and for the effort in trying to make me feel at home while abroad. Also, thanks to PUCRS Faculty of Informatics for its excellence in teaching and research, to whom I owe a great part of my education.

I am also greatly indebted to some teachers in the past: Stanley Loh for introducing me to the research activities while in college. Vera Lúcia Strube de Lima for getting me even more interested in Natural Language Processing while advising me as a master student.

I gratefully acknowledge the financial support received towards my PhD from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), that I otherwise would not have been able to develop this thesis.

Thank you to all people who have helped me in any way in this journey and who I did not mention here. Hopefully, you all know me well enough to know I am not good at remembering this kind of thing. Really sorry!

Last, but not least, I thank my family: my parents, Nilton and Gleci, for giving me life in the first place, for unconditional love and continuous support. Also for educating me with aspects from both sciences and arts. My brother Rodrigo and my sister Rossana, for supporting me spiritually throughout writing this thesis and my life in general.

Thank you all!

## Evaluation of methods for taxonomic relation extraction from text

# ABSTRACT

Modern information systems are changing the idea of "data processing" to the idea of "concept processing", meaning that instead of processing words, such systems process semantic concepts which carry meaning and share contexts with other concepts. Ontology is commonly used as a structure that captures the knowledge about a certain area via providing concepts and relations between them.

Traditionally, concept hierarchies have been built manually by knowledge engineers or domain experts. However, the manual construction of a concept hierarchy suffers from several limitations such as its coverage and the enormous costs of extension and maintenance. Furthermore, keeping up with a hand-crafted concept hierarchy along with the evolution of domain knowledge is an overwhelming task, being necessary to build concept hierarchies automatically.

The (semi-)automatic support in ontology development is usually referred to as ontology learning. The ontology learning from texts is usually divided in steps, going from concepts identification, passing through hierarchy and non-hierarchy relations detection and, seldom, axiom extraction. It is reasonable to say that among these steps the current frontier is in the establishment of concept hierarchies, since this is the backbone of ontologies and, therefore, a good concept hierarchy is already a valuable resource for many ontology applications.

A concept hierarchy is represented with a tree-structured form with specialization/generalization relations between concepts, in which lower-level concepts are more specific while higher-level are more general. The automatic construction of concept hierarchies from texts is a complex task and since the 1980 decade a large number of works have been proposing approaches to better extract relations between concepts. These different proposals have never been contrasted against each other on the same set of data and across different languages. Such comparison is important to see whether they are complementary or incremental, also we can see whether they present different tendencies towards recall and precision, *i.e.*, some can be very precise but with very low recall and others can achieve better recall but low precision.

Another aspect concerns to the variation of results for different languages. This thesis evaluates these different methods on the basis of hierarchy metrics such as density and depth, and evaluation metrics such as Recall and Precision. The evaluation is performed over the same corpora, which consist of English and Portuguese parallel and comparable texts. Both automatic and manual evaluations are presented. The output of seven methods are evaluated automatically and the output of four methods are evaluated manually. Results shed light over the comprehensive set of methods that are the state of the art according to the literature in the area.

**Keywords**: Ontology; Hierarchical Relations; Automatic Relation Extraction

## Évaluation de méthodes pour l'extraction de relations taxonomique à partir de textes

# RÉSUMÉ

Les systèmes d'information modernes sont en train de changer l'idée de «traitement des données» à l'idée de «traitement de concepts», ce qui signifie qu'au lieu de traiter des mots, ces systèmes traitent des concepts sémantiques, lesquels portent la signification des mots et partagent des contextes avec d'autres concepts. Les ontologies sont couramment utilisées comme une structure qui capture la connaissance sur un domaine donné à travers des concepts et des relations entre eux.

Traditionnellement, les hiérarchies de concepts sont construites de façon manuelle par des ingénieurs des connaissance ou des experts du domaine. Les hiérarchies qu'en résultent de ce processus manuel souffrent cependant de plusieurs limitations telles que leurs faible couverture du domaine et les coûts liés à leur évolution et maintenance. La construction automatique d'hiérarchies de concepts est donc nécessaire.

Le support (semi-)automatique dans la construction d'ontologies est généralement appelé 'apprentissage d'ontologie' et comporte généralement trois étapes : identification de concepts, détection de relations hiérarchiques et non hiérarchiques, et extraction d'axiomes. Il est raisonnable de dire que dans cette chaîne de traitement, la frontière est la construction de hiérarchies de concepts, car cela est l'épine dorsale des ontologies et, par conséquent, une bonne hiérarchie de concepts est déjà une ressource précieuse pour de nombreuses applications.

Une hiérarchie de concepts est représentée par une structure d'arbre qui traduit une relation de spécialisation/généralisation, où les concepts de niveau inférieur sont plus spécifiques et les concepts de niveau supérieur sont plus générales. La construction automatique d'hiérarchies est un processus complexe et, depuis 1980, de nombreux travaux proposent des approches pour mieux extraire les relations entre des concepts. Ces différentes propositions ont jamais été comparées les unes aux autres sur un même ensemble de données et ce à travers différentes langues. Cette étude vise à identifier leurs complémentarités, leurs différentes tendances en termes de précision e de rappel, *i.e.*, certaines approches peuvent être très précises en détriment du rappel, et d'autres peuvent obtenir un meilleur rappel en détriment de la précision.

Un autre aspect concerne la variation des résultats pour les différentes langues. Cette thèse évalue ces différentes méthodes en fonction d'un ensemble de métriques, telles que la densité et la profondeur des hiérarchies, et de mesures classiques d'évaluation, telles que le rappel et la précision. L'évaluation est menée sur un corpus composé par des textes parallèles et comparables en anglais et portugais. Sept méthodes sont évaluées automatiquement et quatre méthodes sont évaluées de façon manuelle. L'analyse des résultats apporte des lumières sur l'ensemble des méthodes de l'état de l'art.

**Mots-clés**: Ontologie; Relations hiérarchiques; Extraction automatique de relations

# Avaliação de métodos para extração automática de relações a partir de textos

## RESUMO

Sistemas de informação modernos têm mudado a ideia "processamento de dados" para a ideia de "processamento de conceitos", assim, ao invés de processarem palavras, tais sistemas fazem o processamento de conceitos que contêm ignificado e que compartilham contextos com outros contextos. Ontologias são normalmente utilizadas como uma estrutura que captura o conhecimento a cerca de uma certa área, provendo conceitos e relações entre tais conceitos.

Tradicionalmente, hierarquias de conceitos são construídas manualmente por engenheiros do conhecimento ou especialistas do domínio. Entretanto, este tipo de construção sofre com diversas limitações, tais como, cobertura e o alto custo de extensão e manutenção. Assim, se faz necessária a construção de tais estruturas automaticamente.

O suporte (semi-)automatico no desenvolvimento de ontologias é comumente referenciado como aprendizagem de ontologias e é normalmente dividido em etapas, como identificação de conceitos, detecção de relações hierarquicas e não hierarquicas, e extração de axiomas. É razoável dizer que entre tais passos a fronteira está no estabelecimento de hierarquias de conceitos, pois é a espinha dorsal das ontologias e, por consequência, uma boa hierarquia de conceitos é um recurso válido para várias aplicações de ontologias.

Hierarquias de conceitos são representadas por estruturas em árvore com relacionamentos de especialização/generalização, onde conceitos nos níveis mais baixos são mais específicos e conceitos nos níveis mais altos são mais gerais. A construção automática de tais hierarquias é uma tarefa complexa e desde a década de 80 muitos trabalhos têm proposto melhores formas para fazer a extração de relações entre conceitos. Estas propostas nunca foram contrastadas usando um mesmo conjunto de dados. Tal comparação é importante para ver se os métodos são complementares ou incrementais, bem como se apresentam diferentes tendências em relação à precisão e abrangência, *i.e.*, alguns podem ser bastante precisos e ter uma baixa abrangência enquanto outros têm uma abrangência melhor porém com uma baixa precisão.

Outro aspecto refere-se à variação dos resultados em diferentes línguas. Esta tese avalia os métodos utilizando métricas de hierarquias como densidade e profundidade, e métricas de evaliação como precisão e abrangência. A avaliação é realizada utilizando o mesmo corpora, consistindo de textos paralelos e comparáveis em inglês e português. São realizadas avaliações automática e manual, sendo a saída de sete métodos avaliados automaticamente e quatro manualmente. Os resultados dão uma luz sobre a abrangência dos métodos que são utilizados no estado da arte de acordo com a literatura.

**Palavras-chave**: Ontologia; Relaçoes hierárquicas; Extração Automática de Relações

# LIST OF FIGURES

Figure 2.1 Example of compounding terms containing one single word and two or more elements. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

Figure 2.2 Example of head-modifier relations containing mixed coumpoundings. . . . . 32

Figure 2.3 Dendrogram produced by hierarchical clustering methods. . . . . . . . . . . 39

Figure 2.4 Representation of coordinate terms by a binary tree. . . . . . . . . . . . . . 40

Figure 4.1 Overview of the methodology developed in the thesis . . . . . . . . . . . . 59

Figure 4.2 Excerpt of the extracted structure from Onto.PT. . . . . . . . . . . . . . . 65

Figure 4.3 Syntactic tree for the phrase “Parliament rejected the request from the president.”. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66

Figure 4.4 Tree structure generated by PALAVRAS parser to the phrase “O parlamento rejeita o pedido do presidente.” . . . . . . . . . . . . . . . . . . . . . . . 67

Figure 4.5 Tree structure generated by PALAVRAS parser to the phrase “Primeiramente o homem gostaria de realçar a ausência de lógica do senhor deputado Poettering.” . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70

Figure 4.6 Taxonomies for common semantic cotopy example. . . . . . . . . . . . . . . 77

Figure 4.7 Tree generated for the root “formação de solos” . . . . . . . . . . . . . . . 79

Figure 4.8 Examples of taxonomies represented as direct graphs . . . . . . . . . . . . 80

Figure 5.1 Precision, recall and F-measure of `DocSub` for data sets in English using 1,000 terms (1k) and 10,000 terms (10k) . . . . . . . . . . . . . . . . . . . . . 95

Figure 5.2 Precision, recall and F-measure of `DocSub` for data sets in Portuguese using 1,000 terms (1k) and 10,000 terms (10k) . . . . . . . . . . . . . . . . . . . 96

Figure 5.3 Precision, local and global recall, local and global F-measure of `HClust` for data sets in English using 1,000 terms and 10,000 terms. . . . . . . . . . . . 98

Figure 5.4 Precision, local and global recall, local and golbal F-measure of `HClust` for data sets in Portuguese using 1,000 terms and 10,000 terms. . . . . . . . . . 99

Figure 5.5 Transitive reduction from digraph D1 to D2. . . . . . . . . . . . . . . . . . 106

Figure 5.6 Ratios for relations shared by models in the English corpora. . . . . . . . . 110

Figure 5.7 Ratio of relations shared by models in the Portuguese corpora. . . . . . . . 111

Figure 5.8 Relative precision using the intersection of the models. . . . . . . . . . . . 112

Figure D.1 Google form used in the manual evaluation . . . . . . . . . . . . . . . . . 143

Figure E.1 Examples of taxonomies represented as direct graphs . . . . . . . . . . . . 145

# LIST OF TABLES

Table 2.1 Lexico-syntactic patterns for English, French and Portuguese . . . . . . . . . 29

Table 2.2 Lexico-syntactic patterns and dependency paths found by Snow *et al.* . . . . 43

Table 3.1 Works that manually evaluate the extracted relations. . . . . . . . . . . . . 55

Table 3.2 Works that automatically evaluate the extracted relations. . . . . . . . . . . 56

Table 4.1 Statistics about the corpora . . . . . . . . . . . . . . . . . . . . . . . . . . 63

Table 5.1 Precision, recall and F-measure for methods using the top 1,000 words with the highest number of contexts. . . . . . . . . . . . . . . . . . . . . . . . . 86

Table 5.2 Precision, recall and F-measure for methods using the top 10,000 words with the highest number of contexts. . . . . . . . . . . . . . . . . . . . . . . . . 89

Table 5.3 Precision, recall and F-measure for methods using the top 1,000 words with the highest number of contexts and selecting the best parent. . . . . . . . . 90

Table 5.4 Results for each pattern using Europarl and TED Talks corpora in English. . 91

Table 5.5 Results for each pattern using Europarl and TED Talks corpora in Portuguese. 94

Table 5.6 Lists of terms for manual evaluation . . . . . . . . . . . . . . . . . . . . . . 101

Table 5.7 Fleiss' Kappa inter-annotator agreement and percentage of equal votes. . . . 102

Table 5.8 Precision obtained by manual evaluation for each model in each corpus. . . . 103

Table 5.9 Metrics for taxonomies generated by models using the top 1,000 terms of each corpus in English. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107

Table 5.10 Metrics for taxonomies generated by models using the top 1,000 terms of each corpus in Portuguese. . . . . . . . . . . . . . . . . . . . . . . . . . . . 109

Table F.1 Ratio of relations shared by models in the English corpora. . . . . . . . . . . 149

Table F.2 Ratio of inverse relations shared by models in the English corpora. . . . . . . 150

Table F.3 Ratio of relations shared by models in the Portuguese corpora. . . . . . . . . 150

Table F.4 Ratio of inverse relations shared by models in the Portuguese corpora. . . . . 150

Table F.5 Relative precision values for models using the English corpora. . . . . . . . . 151

Table F.6 Relative precision values for models using the Portuguese corpora. . . . . . . 151

# LIST OF CONTENT

1. Introduction 23
   - 1.1 Motivations . . . 24
   - 1.2 Structure of this thesis . . . 24

2. Literature Review on Taxonomic Relation Extraction 27
   - 2.1 Little or no supervised algorithms . . . 27
     - 2.1.1 Methods based on Lexico-Syntactic Patterns . . . 27
     - 2.1.2 Head-Modifier Detection . . . 31
     - 2.1.3 Distributional Analysis . . . 33
     - 2.1.4 Distributional Inclusion . . . 36
     - 2.1.5 Hierarchical Clustering . . . 39
     - 2.1.6 Document subsumption . . . 41
   - 2.2 Supervised algorithms . . . 43
   - 2.3 Summary . . . 45

3. Approaches for Taxonomy Evaluation 47
   - 3.1 Manual evaluation . . . 47
   - 3.2 Automatic evaluation . . . 49
   - 3.3 Summary . . . 54

4. Materials and methods 59
   - 4.1 Resources . . . 60
     - 4.1.1 Corpora . . . 60
     - 4.1.2 Gold standard . . . 63
   - 4.2 Preprocessing . . . 65
     - 4.2.1 Parsing . . . 65
     - 4.2.2 Context extraction . . . 67
     - 4.2.3 Concept extraction . . . 69
   - 4.3 Relation extraction methods . . . 73
     - 4.3.1 Models for automatic evaluation . . . 73
     - 4.3.2 Models for manual evaluation . . . 75
   - 4.4 Evaluations . . . 76
     - 4.4.1 Automatic evaluation . . . 76
     - 4.4.2 Manual evaluation . . . 77

4.4.3 Metrics for characterizing taxonomies . . . . . . . . . . . . . . . . . . . . . . 78
4.5 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81

5. Evaluation of Taxonomic Relation Extraction Methods 85
5.1 Results for the automatic evaluation . . . . . . . . . . . . . . . . . . . . . . . . 85
5.1.1 All methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85
5.1.2 Results from Pattern-based model (`Patt`) . . . . . . . . . . . . . . . . . 91
5.1.3 Document subsumption-based analysis (`DocSub`) . . . . . . . . . . . . . 94
5.1.4 Hierarchical clustering-based analysis (`HClust`) . . . . . . . . . . . . . . 97
5.2 Results for the manual evaluation . . . . . . . . . . . . . . . . . . . . . . . . . 100
5.3 Metrics Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
5.4 Complementarity Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 109

6. Conclusions and further work 113
6.1 Thesis overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
6.2 Main results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114
6.3 Directions for further work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115

Bibliography 117

Appendix A. Tagset of Stanford Parser 133

Appendix B. Tagset of PALAVRAS parser 135

Appendix C. Heuristics application 137

Appendix D. Guidelines for the manual evaluation of taxonomic relations 143

Appendix E. Metrics for characterizing taxonomies 145

Appendix F. Relation tables 149

# 1. Introduction

The automatic construction of ontologies is still an open problem, despite being the subject of a large number of research initiatives since the last 30 years. This problem has became even more important with the omnipresent evolution of the Web and the overwhelming abundance of data. Considering that most of Web data is in natural language form, *i.e.* texts, the ontology learning from texts is of paramount importance.

According to Buitelaar and Magnini [13], ontology learning is the (semi-)automatic support in ontology development. The ontology learning from texts is usually divided in steps [13], going from concepts identification, passing through hierarchy and non-hierarchy relations detection and, seldom, axiom extraction. It is reasonable to say that among these steps the current frontier is in the establishment of concept hierarchies, since according to many authors [13, 21, 58], this is the backbone of ontologies and, therefore, a good concept hierarchy is already a valuable resource for many ontology applications.

The construction of concept hierarchies from texts is a complex process and the 1980 decade has seen the proposal of a large number of works in this field. Works as Amsler in 1981 [2], Lesk in 1986 [78] and Alshawi in 1987 [1] started the modern search of concept hierarchy from structured sources (dictionaries). Short after, the focus to build concept hierarchies changed to start from textual material, more specifically, domain corpus. Besides early initiatives of Firth in 1957 [35] and Harris in 1968 [51], the modern concept hierarchy construction from texts started with the works by Miller and Charles in 1991 [101], Hearst in 1992 [52], and Ganter *et al.* in 1997 [41]. Initially, these works measured the ability of a system to reproduce the relations of an already existing hierarchy, since they used pairs of terms as seeds to find relations. Nowadays, works such as presented by Navigli *et al.* [105] and Velardi *et al.* [149] attempt to build entire hierarchies from scratch.

A common point of all these works and the subsequent efforts is that the establishment of concept hierarchies is always based on the determination of hypernym and hyponym relations among concepts. As denoted by Jurafsky and Martin [60], a concept $c_1$ is considered hypernym of concept $c_2$ if $c_1$ is a generalization of $c_2$, *e.g.*, the concept "person" is the hypernym of "adult". Analogously, if $c_1$ is the hypernym of $c_2$, $c_2$ is the hyponym of $c_1$, *e.g.*, "adult" is the hyponym of "person". Once all hypernym/hyponym relations are know the concept hierarchy is constructed. Seeing through an ontology learning perspective, the methods presented in this thesis do not differentiate the relation between two different concepts, indicating a class inclusion (`car is-a vehicle`) and between a concept instance and its superordinate concept, indicating class membership relation (`Brazil is-instance-of country`)

## 1.1 Motivations

Although ontology learning from texts itself is a good reason for this thesis, in the past years we observe an increasing number of works addressing taxonomic relation extraction between terms. These works usually propose new methods or adapt old methods (*e.g.*, using patterns for extracting relations from the Web) and apply them in specific tasks to observe their performance. As these works usually use different data sets, their results are not comparable. For instance, the precision achieved by a method using Wikipedia can not be compared with the precision achieved by another method using the New York Times texts. For the best of our knowledge, these different methods have never been contrasted against each other on the same set of data and across different languages. As they have never been contrasted, a number of questions rises:

1. Is there a method that outperforms all other methods?

2. If changing the language, do the methods perform equally?

3. Do all methods generate similar taxonomies?

4. Are results generated by different methods complementary or dissimilar?

Thus, this thesis addresses the answers for these questions. In order to answer these questions, we developed a set of methods that are the state of the art according to the literature in the area. For the first question we performed automatic and manual evaluations, analysing precision, recall and f-measure scores of each method. For the second question we performed experiments using English and Portuguese corpora. for the third question we analyze the learned taxonomies on the basis of hierarchy metrics such as width and depth. For the last question we analyze the complementarity of the results generated by each method.

## 1.2 Structure of this thesis

This thesis is divided into five chapters besides this introduction. Chapters are divided as follows:
*Chapter 2: Literature Review on Taxonomic Relation Extraction* This chapter describes methods for extracting hierarchical or taxonomic relations from text corpora, *i.e.*, relations between a more specific and a more general term. Methods for automatically learning taxonomic relations from texts are divided into two major branchs: Little or no supervised algorithms, and supervised algorithms. The former includes lexico-syntactic patterns, head-modifier detection, distributional analysis, document subsumption and hierarchical clustering. Descriptions of the latter algorithms are a little short since they are out of the scope of this thesis.
*Chapter 3: Approaches for Taxonomy Evaluation* This chapter presents the main criteria used by works presented in Chapter 2 to evaluate automatically constructed hierarchies. The strategies for evaluating hierarchical relations are separated into two main groups: manual evaluation and

automatic evaluation. This chapter also discusses the trade-offs of using one or other approach to evaluate hierarchical relations.

*Chapter 4: Materials and methods* This chapter describes the methodology used in this thesis for developing and evaluating models that extract taxonomic relations from text corpora in Portuguese and English, as well as the resources used and the process of evaluation. The chapter is divided into 4 main parts: resources, preprocessing, models and evaluation. Resources presents all content used in this thesis to perform experiments, including corpora and gold standards. Preprocessing describes how texts are treated before being used in methods for generating taxonomic relations. Models describes how methods for the automatic evaluation and methods for the manual evaluation were developed. Evaluation presents the design of the automatic and manual evaluations.

*Chapter 5: Evaluation of Taxonomic Relation Extraction Methods* This chapter presents a series of experiments aiming to evaluate the methods presented in Chapter 4. In order to verify the quality of the extracted relations and indirectly the quality of the method that generated such relations, automatic and manual evaluations are performed. Besides the analysis of the results of all methods, a deeper analysis is performed on results generated by methods that generate a range of values of precision, recall and f-measure. This chapter also presents an analysis on the characteristics of the taxonomies generated by each method on the basis of hierarchy metrics such as depth and width, and an analysis on the complementarity of pairs of methods. The complementarity indicates whether relations generated by one method are complementar or similar to the relations generated by another method.

*Chapter 6: Conclusions and further work* In this chapter, we present our conclusions and a number of directions for further work.

# 2. Literature Review on Taxonomic Relation Extraction

This chapter describes methods for extracting hierarchical or taxonomic relations from text corpora, *i.e.*, relations between a more specific and a more general term. Seeing through an ontology learning perspective, the methods presented in this chapter do not differentiate the specific type of relation between two concepts, indicating a class inclusion (`car is-a vehicle`) and between a concept instance and its superordinate concept, indicating class membership relation (`Brazil is-instance-of country`). Throughout the text we will use `is-a` for both types of relations.

We divided these methods into two major forms of relation extraction: little or no supervised algorithms (Section 2.1) and supervised algorithms (Section 2.2). Little or no supervised algorithms include methods based on rules and methods that use the distribution of the words as an indicative of taxonomic relationship. On the other hand, supervised algorithms reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances [68]. These methods request a set of training examples which consists of an input object (typically a vector) and a desired output value. The algorithm analyzes the training data and produces an inferred function, which should predict the correct output value for any valid input object. As we do not have manually annotated data to use for the training step, we present only a brief description of works that use supervised algorithms, since this kind of algorithm is out of the scope of this thesis.

## 2.1 Little or no supervised algorithms

Works presented in this section use little or no supervised algorithms and rely on the thorough analysis of the text contents or the application of ground rules in order to estimate possible taxonomic relations. Works that use some version of distributional models, such as distributional analysis, distributional inclusion and hierarchical clustering, in order to identify hierarchical relations between terms are also included.

### 2.1.1 Methods based on Lexico-Syntactic Patterns

The idea of learning taxonomic relations from texts by using lexico-syntactic patterns in the form of regular expressions has been introduced by Hearst [52, 53]. The main idea underlying using patterns is that even if one has never encountered a term, he can infer its semantic relation. For example, consider the following phrases, patterns and extracted relations, where the pattern is expressed as a regular expression:

```
     Phrases:
1.   The bow lute, such as the Bambarandang, is plucked and has an individual curved neck for each string.
2.   Office equipments such as printers, fax machines, and copiers are commonly verified.
3.   Most European countries, especially France, England, and Spain.

     Patterns:
A.   NP such as {NP ,} * {or|and} NP
B.   NP {,} especially {NP ,}* {or|and} NP

     Extracted relations:
1.   is-a(Bambarandang, bow lute)
2.   is-a(printer, office equipment)
     is-a(fax machine,office equipment)
     is-a(copier, office equipment)
3.   is-a(France, European country)
     is-a(England,European country)
     is-a(Spain, European country)
```

The pattern "NP such as {NP ,}* {or|and} NP" means that a noun phrase (NP) must be followed by the words "such" and "as", which must be followed by an NP or by a list of NPs separated by comma, having before the last NP "or" or "and".

In the first phrase even if one have never encountered the term "Bambarandang", it can be inferred that "Bambarandang" is a kind of "bow lute". Pattern-based approaches in general are heuristic methods that apply regular expressions to match a sequence of words in the text. These patterns can extract one or more relations between words in the same phrase. For example, the same pattern was applied on the first and second phrase to extract one relation and three relations respectively. Another pattern is applied on the third phrase to extract a list of relations. The extracted relations is-a(term$_1$, term$_2$) are read as term$_1$ is a kind of term$_2$, or term$_1$ is a hyponym of term$_2$ (or its equivalent term$_2$ is a hypernym of term$_1$). As presented in the examples, the underlying idea of using lexico-syntactic patterns is very simple: to define regular expressions that capture expressions and to map the results of the matching expression to a taxonomic structure between terms.

Patterns proposed by Hearst [52, 53] were initially developed for English, but they have been widely spread to other languages such as Japanese [3], Dutch [143], Turkish [166], French [103] and Portuguese [7]. Table 2.1 summarizes and relates the patterns for English, French and Portuguese. This table is adapted from the work of Basegio [7] and presents the lexico-syntactic patterns where NP is a noun-phrase and LNP is a list of noun-phrases. LNP represents the regular expression "{NP ,}* {or|and} NP" which captures noun-phrases in a sequence.

Taxonomic relations can be extracted fairly accurately with the syntactic patterns [104]. In contrast, these patterns are usually brittle and may not occur very often in a corpus. Although approaches relying on lexico-syntactic patterns have a reasonable precision, their recall is very low [13, 21]. For example, Lopes and Vieira [89] found only 7 occurrences using these patterns in a corpus of the geology domain, containing 139 texts (39,974 sentences and 245,089 terms). Along

| # | **English** [52] | **French** [103] | **Portuguese** [7] |
|---|---|---|---|
| 1 | NP such as LNP | NP tel que LNP | NP {tal\|tais} como LNP |
| 2 | | NP comme LNP | NP como LNP |
| 3 | such NP as LNP | NP tel LNP | tal(is) NP como LNP |
| 4 | LNP or other NP | LNP ou d'autre NP | LNP ou {outro(s)\|outra(s)} NP |
| 5 | LNP and other NP | LNP et d'autre NP | LNP e {outro(s)\|outra(s)} NP |
| 6 | | LNP et notament d'autre NP | |
| 7 | NP including LNP | | NP incluindo LNP |
| 8 | NP especially LNP | NP, particulièrement LNP | NP especialmente LNP |
| 9 | | | NP principalmente LNP |
| 10 | | | NP particularmente LNP |
| 11 | | | NP em especial LNP |
| 12 | | | NP em particular LNP |
| 13 | | | NP de maneira especial LNP |
| 14 | | | NP sobretudo LNP |
| 15 | | {deux\|trois\|...} NP (LNP) | {dois\|duas\|três\|...} NP (LNP) |
| 16 | | {deux\|trois\|...} NP : (LNP) | {dois\|duas\|três\|...} NP: (LNP) |
| 17 | | {certain\|quelque\|d'autre} NP (LNP) | {certos\|certas\|qualquer} NP (LNP) |
| 18 | | | {outros\|outras\|alguns\|algumas} NP (LNP) |
| 19 | | {certain\|quelque\|d'autre} NP : (LNP) | {certos\|certas\|qualquer} NP: (LNP) |
| 20 | | | {outros\|outras\|alguns\|algumas} NP: (LNP) |
| 21 | | chez NP, LNP | |

Table 2.1: Lexico-syntactic patterns for English, French and Portuguese

with the patterns, Hearst [52] provides an algorithm for identifying new patterns, and unlike much statistical work, a single occurrence may be sufficient. This algorithm is composed by five steps:

1. Identify a lexical relation of interest;
2. Gather a list of terms for which this relation holds;
3. Find places in the corpus where these expressions occur syntactically near one another and record the environment;
4. Identify new patterns based on the new environments;
5. Use the new patterns to gather more instances of the target relation and go to Step 2.

Following the algorithm proposed by Hearst, it is possible to identify pairs of related terms and re-feed the system to find other relations. Using these new relations the system can extract new pairs of related terms and so on. Extracting relations only for terms that appear close in a corpus (usually at sentence level) is an intrinsic characteristic of lexico-syntactic patterns, what could lead to consider treating them out of the boundaries of a sentence. On the other hand, the size of the corpus may have an impact on the coverage on identifying the relations. Hence, big corpus should be exploited in order to overcome this limitation.

In order to minimize the drawbacks of low coverage Pantel and Pennacchiotti [112] exploit the web as a big corpus. They developed *Espresso*, a weakly-supervised, general-purpose and accurate algorithm for harvesting semantic relations. This algorithm begins with seed instances of a particular binary relation (*e.g.*, `is-a`). These seeds are used to infer a set of patterns $P$ that connect both words in the corpus. As noise in data can infiltrate the algorithm and steer it in a wrong direction (semantic drift), the retrieved patterns are ranked on the basis of their *reliability*. Reliability of a pattern ($r_\pi(p)$) and reliability of a relation ($r_\iota(i)$) are defined recursively as:

$$r_\pi(p) = \frac{\sum_{i \in I}\left(\frac{PMI(i,p)}{max_{PMI}} r_\iota(i)\right)}{|I|} \tag{2.1}$$

$$r_\iota(i) = \frac{\sum_{p \in P}\left(\frac{PMI(i,p)}{max_{PMI}} r_\pi(p)\right)}{|P|} \tag{2.2}$$

where $r_\pi(p)$ is the reliability of the pattern $p$, $r_\iota(i)$ is the reliability of the relation instance $i$, and $max_{PMI}$ is the maximum Pointwise Mutual Information (PMI) between all patterns and all instances. All instances supplied initially as seeds have $r_\iota(i) = 1$. As explained by Turney [146], PMI provides a way to measure the degree of co-occurrence of two words by comparing the number of co-occurrences to the number of individual occurrences. This value is maximal when all occurrences are co-occurrences. Equation 2.3 presents the PMI measure.

$$PMI(x,y) = log\frac{P(x,y)}{P(x)P(y)} \tag{2.3}$$

where $P(x,y)$ is the probability of $x$ and $y$ co-occur, $P(x)$ is the probability of $x$ and $P(y)$ is the probabilty of $y$. In their work, Pantel and Pennacchiotti adapted the PMI to estimate de degree of co-occurrence between an instance $i = \{x,y\}$ and a pattern $p$, and it is estimated as:

$$PMI(i,p) = log\frac{|x,p,y|}{|x,*,y||*,p,*|} \tag{2.4}$$

where $|x,p,y|$ is the frequency of pattern $p$ occurring with terms $x$ and $y$ and where the asterisk $(*)$ represents a wildcard. This ranking intends to keep patterns that are highly associated with the input instances. After ranking patterns, a threshold is applied for discarding the low ranked ones. Then, *Espresso* retrieves from the corpus a set of instances $I$ that match any of the patterns in $P$. Google search engine is used as a web expansion, gathering new instances to $I$ using the patterns contained in $P$. For example, given the instance {Italy,country} in $I$ and the pattern "`NP such as LNP`", a query "`country such as *`" is submitted to the search engine. In this step, *Espresso* also applies a filtering algorithm to automatically separate correct instances extracted by generic patterns, *i.e.*, patterns with high recall and low precision such as "`is-a`", from incorrect ones. The full *Espresso* algorithm was tested with many relations such as `is-a`, `part-of`, `succession`, `reaction` and `production`.

In order to improve recall and precision of terms extracted after applying the patterns proposed by Hearst, Cederberg and Widdows [16] use Latent Semantic Analysis (LSA) to filter out terms that are not semantically related. Applying LSA, they reduce the rate of error of the initial pattern-based hyponymy extraction by 30%, achieving a maximum precision of 64%. Ponzetto and Strube [119] attached the patterns presented by Hearst [52, 53] in a method for building a taxonomy based on the content of Wikipedia structure. In this approach, the semantic relations between categories are labeled either `is-a` or `not-is-a` using methods based on connectivity in the network and lexico-syntactic matching. In order to determine the quality of the taxonomy, a manual evaluation is performed. To determine its coverage, an automatic evaluation comparing the coverage with ResearchCyc and WordNet is carried out. The resulting taxonomy compares favorably in quality and coverage with broad-coverage manually created resources.

Even with drawbacks, many works keep using the patterns proposed by Hearst to extract hyponym relations between terms, sometimes mixing them with other techniques such as clustering [15, 29] or LSA [16], sometimes using them on the Web to extract contexts [114] or class instances [31, 70].

### 2.1.2 Head-Modifier Detection

According to Radford [121], the head of a phrase is the word which is grammatically most important in the phrase, since it determines the nature of the overall phrase. The changing in the sense of the head can be classified according to the particle attached to the head. In a term composed by affixation, an affix is added to a word forming a new word, such as "preprocessor" where "processor" is the head and "pre" is an affix. In a compounding term, two words together create the new meaning. These terms can be joined into one single word or be composed by two or more elements. Figure 2.1 shows examples of compounding words and their relations with their hypernyms.

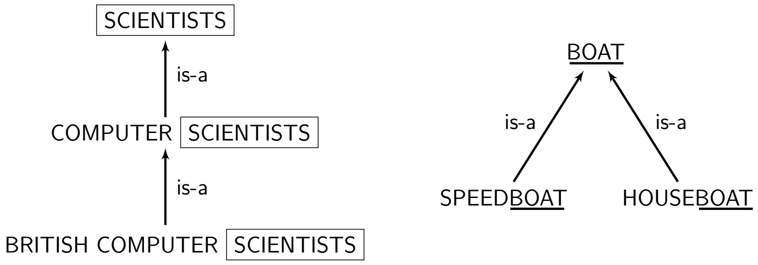


Figure 2.1: Example of compounding terms containing one single word and two or more elements.

When compounding terms are consisted by two or more elements, it is claimed that the arrangement of these elements reflects the kind of information being conveyed. The main element, known as the head, identifies the semantic category to which the whole term belongs. The other elements distinguish these members from other members of the same category. Thus, the whole compound term is related to the head as a hyponym. Following the example in Figure 2.1, the term "Computer Scientists" has as head "Scientists" which can be transformed into its hypernym, *i.e.*, "Computer Scientists" is a kind of "Scientist", and "British Computer Scientists" has the head "Computer Scientists" which can be its hypernym, *i.e.*, "British Computer Scientists" is a kind of "Computer Scientists". In contrast, compounding terms containing a single word have the head as part of the word. Thus, in constructions such as "Houseboat" and "Speedboat" the head element is "Boat", and may therefore be viewed as hypernym. The affixed terms "Houseboat" and "Speedboat" are hyponyms of "Boat", *i.e.*, a kind of "Boat". The modifiers "house" and "speed" act to distinguish the members of the set of hyponyms [56].

Following this idea, it is also possible to have a mix of compounding terms. Figure 2.2 presents the hierarchy using head-modifier to the term "Boat". As the reader can see, the terms "Competition Speedboat" and "Leisure Speedboat" are a kind of "Speedboat", as well as "Speedboat" is a kind of "Boat". Finally, the terms "Competition Speedboat" and "Leisure Speedboat" are connected to "Boat", through the term "Speedboat".

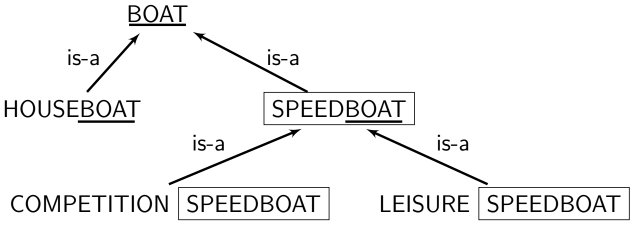


Figure 2.2: Example of head-modifier relations containing mixed coumpoundings.

Terms containing three or more words (multiwords, *e.g.*, "film society committee") could be linked to its hypernym (*e.g.*, "society committee") that is also linked to its own hypernym (*e.g.*, "committee"), thus creating natural sub-levels. However, the chunking of multiword terms can be ambiguous. Using the example above, the term "film society committee" can be decomposed in two ways: [film [society committee]] or [[film society] committee]. Note that the bracketing is important and it indicates the subconstituency of the term, and therefore its derivational history. The former example can be interpreted as "There exist committees, some of which are society committees. There are range of these, including society committees whose interest is film." and the latter be interpreted as "There exist committees, some of which are film society committees." [56].

In order to eliminate the ambiguity many techniques have been applied, such as using the term frequencies. For example, the frequency of [society committee] can be compared with that of [film society] to give the likelihood of the candidate bracketings. Velardi *et al.* [148] use Mutual Information [33] and Dice factor [134] to the terminology extraction and then calculate the Domain Relevance (DR) for term disambiguation. Vossen [150] developed a heuristic to extract the most likely chunking. In this heuristic, the system will first look for the most salient head and then try to decompose the remaining multiwords into modifiers. If there are any lexicalized multiwords embedded in the multiword phrase and the lexicalized multiword segments overlap, they select the most "salient" candidate. If there are no lexicalized multiwords embedded in the multiword phrase, they apply the same criteria to all multiword segments.

In many languages the morphological system is very rich and enables the construction of semantically complex compound words. Like English, German also offers the possibility of combining words, especially nouns. Usually noun chains in English are featured by spaces or hyphens between words, while in German normally the chain appears as one word (*e.g.*, the German word "Kreuzbandverletzung" corresponds to three English words: "cruciate", "ligament" and "injury"). In order to deal with these complex compounds, Buitelaar *et al.* [14] and Sintek *et al.* [133] present OntoLT, a plug-in to Protégé[1], which provides an environment for the integration of linguistic analysis in ontology engineering through the definition of mapping rules. One of this mapping rules splits nouns

[1] http://protege.stanford.edu/

into its chain of elements and uses the head of the chain to build the hypernym/hyponym relations between terms.

Lopes [84] presents a process for the automatic extraction of concept hierarchies from domain corpora in Portuguese. The process starts by the extraction of noun phrases, containing simple and compound terms, using the E$\chi$ATO$_{lp}$ tool [86]. From the list of noun phrases, the terms that best represent the domain are selected. Finally, from the selected terms, the head of each noun phrase is identified, assuming the hypernym role and the hierarchical structure is then generated.

Usually head-modifier approaches are used as part of the process of taxonomy generation. For instance, in the work of Ponzetto and Strube [119], head-modifer is applied to identify is-a relations in a conceptual network generated using the Wikipedia categories. Relations between terms are named is-a when both terms have the same lexical head (*e.g.*, the term "Scientists" in Figure 2.1). In case that the stem of the lexical head of one category occurs in a non-head position in other term, the relation is named as not-is-a (*e.g.*, the head "Islam" plays the role of modifier in the term "Islamic Mysticism", thus "Islamic Mysticism" "not-is-a Islam"). Ponzetto and Strube reported a good coverage by identifying 141,728 is-a relations by head matching on Wikipedia categories, describing such method containing high precision.

Ponzetto and Strube also claim that due to the polysemic nature of the words, approaches that use head-modifier tend to mix all the meanings of a word in only one instance of this word, and thus leading to errors. They point out that head matching will erroneously succeed, if the modifiers select different senses for the respective heads (*e.g.*, "Caucus chair" and "chair"), or the relation expressed is not an is-a relation (*e.g.*, meronymy as in "West Java" and "Java"). Also, the head matching will not succeed in those cases in which the head of a noun phrase modifies the head of another to select a compatible sense (*e.g.*, "Electronic music" and "Music genres").

### 2.1.3 Distributional Analysis

Many methods that process data involve semantic models, also known as corpus-based semantic models, semantic spaces, word spaces, or distributional similarity models (DSMs). Such models usually rely on some version of the distributional hypothesis [50] which states that words that occur in the same contexts tend to have similar meanings. In other words, the degree of semantic similarity between two words is related to the degree of overlaping among their contexts. A better understanding about semantic similarity is presented by Lemaire and Denhiére [75] who point out that it could be viewed as an association of two terms, that is, the mental activation of one term when another term is presented.

Usually DSMs vary according to the aspects of meaning they are designed to model. According to Medin *et al.* [96] there are two types of similarity, the attributional similarity, *i.e.*, a correspondence between attributes, and the relational similarity, *i.e.*, a correspondence between relations. As noted by Grefenstette [46], attributional similarity is typically addressed by word collocates, that is, words that co-occur more often than would be expected by chance. Thus, words that share many collocates denote concepts that share many attributes. For example, the contextual similarity between "automobile" and "car" is very high because they co-occur with "accelerate", "break", "color" and other words. On the other hand, when two word pairs have a high degree of relational similarity, we say they are analogous (*e.g.*, "traffic" is to "street" as "water" is to "riverbed").

Term co-occurrence has been studied since 1960 [92] and nowadays it is a common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words [20]. These co-occurrences can be within a certain limited distance in the context (using a window to generate these co-occurrences or even the size of each sentence) or within syntactic relations (*e.g.*, verb-object, verb-subject relations). Methods to achieve the similarity between terms can be separated in three groups according to the term co-occurrence order, namely first-order, second-order and third (or higher) order co-ocurrence methods.

In first-order co-occurrence, two terms appear in identical contexts [147]. This type of co-occurrence is based on the J.R. Firth saying "You shall know a word by the company it keeps." [35]. For instance, on the one hand, "bank" co-occurs with words and expressions such as "money", "notes", "loan", "account", "investment", "clerk", "official", "manager", and so forth. On the other hand, we find "bank" co-occurring with "river", "swim", "boat", and "east" depending on which meaning the word presents [49]. Approaches that explore first-order co-occurrence are usually statistics-based and do not need any (or need a minimum) linguistic pre-processing of the corpus, *e.g.*, using a context-window [20, 61], clustering words [26], or even web-based [146].

Usually methods to identify terms similarity using first-order co-occurrence rely on the extraction of contexts to all words in the corpus using a window, *i.e.*, every pair of terms occurring together within a window is extracted as the window is moving through a text. The size of this window should accommodate a few sentences and not having a too large computational load. Then all the words in the context of every occurrence of a word $w$ inside a bag are collected. That bag of words will represent the meaning of $w$ [125]. Measuring the correlation between terms is the last step to associate terms by their similarity. Kaji *et al.* [61] use Pointwise Mutual Information (Equation 2.3) [20] as the measure of correlation between terms. After applying PMI, the result consists of a list composed of terms with their related terms ranked by PMI values.

According to Lemaire and Denhiére [75] two words are associated by means of second-order co-occurrence if they share at least one word context. This view is based on the Harris' distributional hypothesis [50] which states that words that occur in the same contexts tend to be similar. Second-order co-occurrences usually are linguistics-based, which means that a pre-processing of the corpus is needed to find contexts (*e.g.*, syntactic contexts). In the works by Grefenstette [46] and Lin [80], for example, the basic idea is that two terms sharing more syntactic relations with respect to other terms are more similar in meaning. Syntactic relations between term pairs were captured by the notion of dependency triples (*e.g.*, <$\mathtt{w_1}$, r, $\mathtt{w_2}$>, where $\mathrm{w_1}$ and $\mathrm{w_2}$ are two words and r is the syntactic relation between them).

According to Buitelaar *et al.* [14] methods that rely on raw data or frequency counting, such as first and second-order co-occurrence methods, may lead to data sparseness. In order to overcome this problem, approaches based on dimension-reduction techniques like Latent Semantic Analysis

[72] should be applied on term by context matrices. Besides reducing the space dimension these techniques also identify third (or higher) order co-occurrences. Gamallo and Bordag [40] explain third (or higher) order co-occurrences as co-occurrences between words that do not co-occur in the corpus with the same words (or lexical-syntactic contexts) but between words that can be related through further indirect co-occurrences. These co-occurrences can be obtained by means of applying Singular Value Decomposition (SVD) methods. Thus, SVD methods try to represent a more abstract and generic word space which tries to capture higher-order associations by inducing a latent (hidden) structure that does not rely on word co-occurrences attested in the corpus.

SVD methods are based on linear algebra and apply mathematical operations on a term-document or term-context matrix. Initially this technique was called Latent Semantic Indexing (LSI) and it was used by Deerwester *et al.* [28] as the main application for Information Retrieval. Later, Landauer and Dumais [72] proved that truncated SVD can also be efficiently applied to word similarity, calling it as Latent Semantic Analysis (LSA).

Yang and Powers [164] apply SVD on a matrix of words by syntactic contexts, as presented by Grefenstette [46]. The contexts considered to nouns were head modifiers (*e.g.*, "adjective-noun"), nouns when they have "subject-verb" or "verb-object" relations. They reduced the original matrix containing thousand dimensions to 250 dimensions and applied the cosine similarity measure on the word vectors to compute the similarity between words. The cosine of the angle between vectors $x$ and $y$ in the n-dimensional space is defined as presented in Equation 2.5, where $\|x\|\,\|y\|$ are the length of $x$ and $y$:

$$cos\Theta = \frac{x \cdot y}{\|x\|\,\|y\|} = \frac{\sum_{i=1}^{n} x_i, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{2.5}$$

An evaluation of these methods that use different co-occurrence orders to get similarity between terms is performed by Granada *et al.* [45]. In that work, a manual evaluation is performed on each method and the results show that the second-order co-occurrence method has higher scores when compared with methods using other orders. A similar approach is performed by Gamallo and Bordag [40], evaluating the usefulness of SVD in a similarity extraction task. In that work, the authors argue that methods based on SVD are much less precise than other word space models for the task of extracting translation equivalents from comparable corpora. They pointed out that when applying SVD, the second-order similarity decreases to the increasing of the third (or higher) order similarity.

Recently, there has been much work on DSMs that identify related terms based on their semantic similarity in vector spaces. Consequently, a term can be represented by a vector of contexts in which it frequently appears. Any vector space model could then use the term' vectors to cluster semantically related terms. According to Budanitsky and Hirst [12], two terms are semantically related if they have any kind of semantic relation. Thus, methods that use distributional analysis to generate similar terms should have a post-processing in order to identify only hierarchical terms. This process can be performed using distributional inclusion (Section 2.1.4), hierarchical clustering (Section 2.1.5), or document subsumption (Section 2.1.6). For instance, Rios-Alvarado *et al.* [123] first cluster terms using a second-order matrix and then use patterns discovered by Hearst [52,53] and Snow *et al.* [135] to build Web queries. Each query is sent to a web search engine and the sentences containing the patterns in the retrieved pages are identified. The nouns linked by each pattern are extracted and the hierarchical relations between them is identified.

### 2.1.4 Distributional Inclusion

Hyponymy relation is transitive, *i.e.*, whenever an element `A` is related to an element `B`, and `B` is in turn related to an element `C`, then `A` is also related to `C`; and asymmetrical, *i.e.*, `A` is a kind of `B`, but `B` is not a kind of `A`. For example, "potato" is a kind of "vegetable" and "vegetable" is a kind of "plant". Thus, by transitivity "potato" is a kind of "plant" and by asymmetry "plant" is not a kind of "potato".

These relations generate a hierarchical semantic structure where the hyponym is below its superordinate, and where the hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate [100]. For example, "mapple" is an hyponym of "tree" because it inherits the properties of the latter but is distinguished from the other trees by the hardness of its wood, shape of its leaves *etc.*. On the other hand, Weeds *et al.* [156] verified that distributional generality is correlated with semantic generality, *i.e.*, hypernyms tend to occur in a larger variety of contexts than hyponyms. In this sense, the asymmetry is captured by means of co-occurrence retrieval [155], where a word has higher recall (hereafter `WeedsRec`) and lower precision (hereafter `WeedsPrec`) when compared with its hyponyms' co-occurrences, and higher precision and lower recall when compared with its hypernyms' co-occurrences. In other words, if $u$ and $v$ are related terms and `WeedsPrec`($u$,$v$) $>$ `WeedsRec`($u$,$v$), it is expected that $u$ is a hyponym of $v$. Equation 2.6 defines precision and recall, where $I(w, f)$ is the Pointwise Mutual Information (PMI) between the word $w$ and the feature $f$ and $F_w$ is the set of all features $f$ with $I(w, f) > 0$.

$$WeedsPrec(u,v) = \frac{\sum_{f \in F_u \cap F_v} I(u,f)}{\sum_{f \in F_u} I(u,f)} \quad WeedsRec(u,v) = \frac{\sum_{f \in F_u \cap F_v} I(v,f)}{\sum_{f \in F_v} I(v,f)} \tag{2.6}$$

Observing the distributional generality of words in the work of Weeds *et al.*, Geffet and Dagan [43] point out that sources of noise in the Weeds' work may be influenced in the results, not showing a great improvement when using feature vectors. The authors claim that the quality of similarity scores is often biased by inaccurate feature weights. Thus, they propose a recalculation on weighted vectors taking into account the set of most similar words generated by the Lin's measure [80].

Using the new weighted vectors, Geffet and Dagan propose the Distributional Inclusion Hypothesis. The hypothesis says that if the meaning of a word $u$ entails another word $v$, then it is expected that all the typical contexts (features) of $u$ will occur also with $v$. In other words, if a term $u$ is semantically narrower than term $v$, then a significant number of salient distributional syntactic features of $u$ is also included in the feature vector of $v$. Letting $u_i => v_j$ denote the directional entailment between the word senses of word $u$ and $v$, two hypotheses were created:

- Hypothesis 1: If $u_i => v_j$ then all the syntactic features of $u_i$ are expected to appear with $v_j$
- Hypothesis 2: If all the syntactic features of $u_i$ appear with $v_j$ then is expected that $u_i => v_j$.

The validity of the hypotheses was tested and results show that most of the pairs tested fulfill the condition proposed in both hypotheses.

Szpektor and Dagan [141] proposed the `balPrec` measure which combines the precision of `WeedsPrec` with the Lin's measure by taking their geometric average. Lin's measure [80] (Equation 2.7) works on `WeedsPrec` penalising vectors containing few features, where $I(w,f)$ is the positive PMI between word $w$ and feature $f$.

$$LIN(u,v) = \frac{\sum_{f \in F_u \cap F_v} [I(u,f) + I(v,f)]}{\sum_{f \in F_u} I(u,f) + \sum_{f \in F_v} I(v,f)} \tag{2.7}$$

Thus, `balPrec` is defined as a balanced version of `WeedsPrec`:

$$balPrec(u,v) = \sqrt{WeedsPrec(u,v) \cdot LIN(u,v)} \tag{2.8}$$

Clarke [24] formalised the idea of distributional generality and computes the entailment between two words using a variation of Weeds *et al.* [156] measures. It differs from the one proposed by Weeds *et al.* because it reduces the weight of included features if they have lower weight within the vector of the broader term. Precision (ClarkeDEPrec) and recall (ClarkeDERec) are defined as:

$$\begin{aligned} ClarkeDEPrec(u,v) &= \frac{\sum_{f \in F_u \cap F_v} min(I(u,f), I(v,f))}{\sum_{f \in F_u} I(u,f)} \\ ClarkeDERec(u,v) &= \frac{\sum_{f \in F_u \cap F_v} min(I(u,f), I(v,f))}{\sum_{f \in F_v} I(v,f)} \end{aligned} \tag{2.9}$$

Lenci and Benotto [77] expand the idea of Geffet and Dagan [43], and explore the possibility of identifying hypernyms in Distributional Similarity Models (DSMs) using directional (or asymmetric) similarity measures. They propose a new measure that takes into account not only the inclusion of the features of $u$ in $v$, but also the non-inclusion of the features $v$ in $u$. This measure, presented in Equation 2.10, is a variation of the measure proposed by Clarke [24] – Equation (2.9) which measures the inclusion of the features of a term $u$ in a term $v$. The idea behind `invCL` is that, if $v$ is a semantically broader term of $u$, then the features of $u$ are included in the features of $v$, but crucially the features of $v$ are not included in the features of $u$.

$$invCL(u,v) = \sqrt{ClarkeDEPrec(u,v) \cdot (1 - ClarkeDERec(u,v))} \tag{2.10}$$

Kotlerman *et al.* [67] crafted the `balAPinc` measure (Equation 2.11) which is optimized to capture a relation of feature inclusion between terms while using the relative relevance of features. This measure applies the IR evaluation method of *Average Precision* (`APinc`) in order to identify the feature inclusion, while uses the symetric similarity measure of Lin [80] to penalise low frequency words.

$$APinc(u,v) = \frac{\sum_{r=1}^{|F_u|} P(r) \cdot rel'(v,r,u)}{|F_u|} \tag{2.11}$$
$$balAPinc(u,v) = \sqrt{APinc(u,v) \cdot LIN(u,v)}$$

The idea of `APinc` is that the score increases with a larger number of features shared by $u$ and $v$ (given by $P(r)$ which calculates the precision at every rank $r$ among the shared $u$'s) dimensions), while giving higher weight to highly ranked features of the narrower term (given by $rel'(v,r,u)$).

In a recent work, Santus *et al.* [130] used an entropy-based measure named `SLQS` for the unsupervised identification of taxonomic relations in DSMs. `SLQS` is grounded on the idea that contexts of hypernyms are less informative than the most typical linguistic contexts of its hyponyms. For example, contexts like "has fur" and "bark" are likely to co-occur with a smaller number of terms than "move" and "eat". Thus, `SLQS` uses entropy [131] as an estimate of context informativeness. First, each context is weighted using Local Mutual Information (LMI) [32]. A semantic generality index ($E_{wi}$ – Equation 2.12) for a word $w_i$ is generated by calculating the median entropy ($M_e$) for a set of the top $N$ most associated contexts. Median entropy takes into account normalized contexts vectors with entropy $H_n(c_j)$. Normalization is performed in each vector of contexts using a range 0–1 with the Min-Max-Scaling [120]. Thus, the semantic generality index of each word is calculated as:

$$E_{wi} = Me_{j=1}^{N}(H_n(c_j)) \tag{2.12}$$

where entropy is defined as:

$$H(c) = -\sum_{i=1}^{n} p(f_i|c) \cdot log_2(p(f_i|c)) \tag{2.13}$$

where $p(f_i|c)$ is the probability of the feature $f_i$ given the context $c$, obtained through the ratio between the frequency of $\langle c, f_i \rangle$ and the total frequency of $c$. Finally, `SLQS` measures the semantic generality of word $u$ and word $v$ as the reciprocal difference between the semantic generality of the two words $u$ and $v$ as:

$$SLQS(u,v) = 1 - \frac{E_u}{E_v} \tag{2.14}$$

According to the formula, $u$ subsumes $v$ if `SLQS` $>$ 0, $v$ subsumes $u$ if `SLQS` $<$ 0, and we can not infer a taxonomic relationship between $u$ and $v$ if `SLQS` $\simeq$ 0.

### 2.1.5 Hierarchical Clustering

Besides the distributional similarity or the identification of semantically related terms [26, 80], clustering approaches have been applied to the identification of hierarchical relations. In order to identify hypernyms/hyponyms, hierarchical clustering algorithms group words according to their meanings in text. These groups are labeled using its members' lexical or syntactic dependencies and then an `is-a` relation is extracted between each cluster member and the cluster label. The process of clustering can either begin with individual terms or concepts, grouping the most similar ones (*i.e.*, bottom-up clustering, also known as agglomerative clustering), or with all terms or concepts and dividing them into smaller groups to maximize the similarity within the group (*i.e.*, top-down clustering, also known as divisive clustering).

A bottom-up approach begins with every term in a distinct (singleton) cluster, and successively merges clusters together until a stopping criterion is satisfied [57] or until all terms belong to one cluster. In details, hierarchical clustering performs the following steps:

1. create a node to each term;
2. compute the similarity between each pair of nodes using some similarity measure;
3. find the two most similar nouns and combine them by giving a common parent;
4. compute the similarity between the new node and all other nodes;

Steps 3 and 4 are repeated until all nouns have been placed under a common ancestor. This approach generates a binary tree (dendrogram) that can be visually presented as the structures in Figure 2.3.

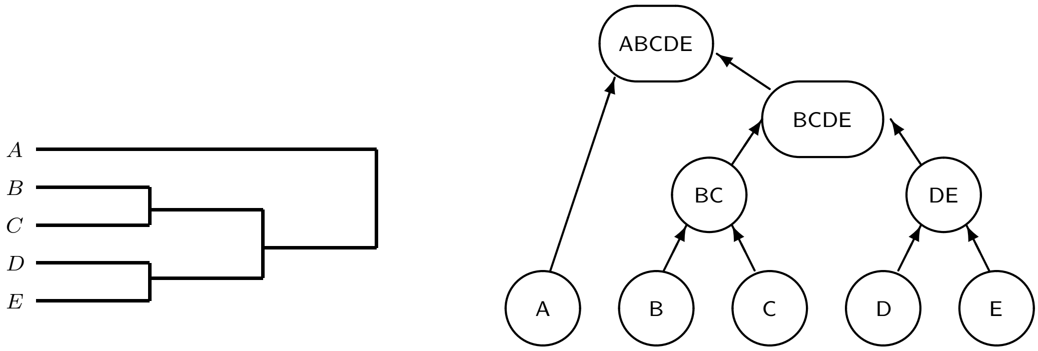


Figure 2.3: Dendrogram produced by hierarchical clustering methods.

The approach used by Caraballo [15] was the first attempt to create hypernym-labeled noun using hierarchical clustering. Initially, an unlabelled hierarchy of noun clusters using an agglomerative bottom-up clustering of vectors having conjunction and appositives as features is generated. Traditional hierarchical clustering algorithms construct binary trees. However, binary branches may not correctly describe the data structure. Figure 2.4 presents the structure generated by a traditional hierarchical clustering algorithm $(a)$, and how the tree should be structured $(b)$, since the words

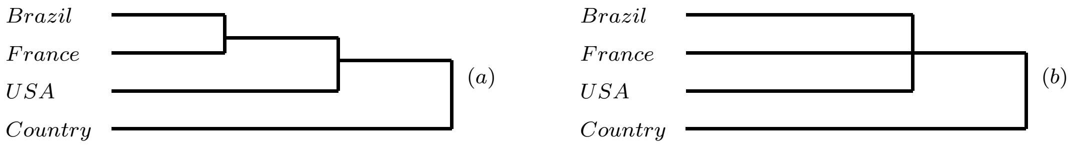


Figure 2.4: Representation of coordinate terms by a binary tree.

"Brazil", "France" and "USA" are coordinate terms, *i.e.*, terms that share the same hypernym (*e.g.*, "country").

In order to solve this problem, Caraballo [15] applies a step to compress the generated tree. The tree is compressed by looking when an internal node is unlabeled, meaning that a hypernym could not be found to describe its descendant nouns. Since there is a labeled node with a hypernym above the current node, delete the current node and connect the children of this node to the hypernym.

In order to label the parent node of each cluster, Caraballo used the patterns containing the word "other" as suggested by Hearst [52, 53], patterns (4) and (5) in Table 2.1. Using these patterns, vectors are constructed indicating whether a word has a hyponymy relation with other words. These vectors are associated to the leaves of the cluster and for each association between nodes, a vector of hypernyms is created, by adding together the vectors of its children. The label of the node is the hypernym with the largest value in this vector, *i.e.*, the hypernym which appeared with the largest number descendant nouns.

Cimiano *et al.* [22] applied a divisive clustering technique called Bi-Section-KMeans to generate a taxonomy of terms. Bi-Section-KMeans is initiated with a unique cluster containing all terms. The division loops follow the algorithm: it selects the cluster with the largest variance and it calls KMeans in order to split this cluster into exactly two subclusters. The loop is repeated $K-1$ times such that $K$ non-overlapping subclusters are generated. Using the Figure 2.3 as example again, in the top-down approach the arrows are in the opposite side, thus the cluster starts on the top, having all terms in the same cluster and recursively splits each (sub-)cluster until each term be in a single node.

Cimiano *et al.* also compare effectiveness (*i.e.*, quality of result, using a manually built gold standard), the efficiency (*i.e.*, run-time behaviour) and the traceability (*i.e.*, a qualitative discussion of how easy it is for the ontology engineer to comprehend why the taxonomy was constructed in a particular way by the corresponding method) of a bottom-up against a top-down approach. The results show that both approaches have a comparable performance regarding the task and that each approach has its own benefits. In general, all these approaches fall short of human achievement, however they seem to be good enough for supporting the task of ontology engineering.

Liu *et al.* [82] apply a multi-branched tree algorithm during the hierarchical clustering process, avoiding the problem of generating a binary tree. In order to build the multi-branched tree, they adopted a deterministic, agglomerative approach proposed by Blundell *et al.* [10] called Bayesian Rose Tree (BRT). This approach explores a Bayesian hierarchical clustering algorithm that can produce trees with arbitrary branching structure at each node.

Dietz *et al.* [30] present TaxoLearn, an approach to the automatic construction of domain taxonomies. This approach uses hierarchical clustering in order to generate the taxonomy of terms. In order to cluster terms, the average linkage clustering was used, since single linkage method has the drawback that clusters that are not very similar can be put together if just two single entities in each of the clusters are very close to each other, and complete linkage method has the drawback that outliers have a high influence on the clustering process. The average linkage clustering has a good balance between these two extremes. The label to each new cluster is chosen with a combination of two approaches. The first one considers the hypernym information retrieved from WordNet, finding the common ancestor to all terms in the cluster. The other approach considers as hypernym the term that is closest to the centroid of the cluster. Hence, for clusters where it is not possible to find a centroid, *i.e.*, clusters composed by two terms, it is checked whether these terms have any hyperym in common. If they do not have any hypernym in common, a concatenation of the two concepts is used as label. Clusters that contain more than two terms are labeled using the centroid method.

Similarly to Caraballo [15], De Knijff *et al.* [27] also build a taxonomy using hierarchical clustering methods. In that work De Knijff *et al.* claim that this type of method may generate hierarchies with terms containing multiple potential parents. Hence, they choose one potential parent to maintain the hierarchical tree structure. The decision is based on a score calculated for each potential parent, taking into account the distance between the target term and the list of ancestors parents. The score for each potential parent is defined as presented in Equation 2.15:

$$score(p,x) = P(p|x) + \sum_{a \in A_p} w(a,x) \cdot P(a|x) \tag{2.15}$$

where $p$ is the potential parent of term $x$ and $Ap$ is the list of ancestors of $p$. The co-occurrence probability $P(a|x)$ is multiplied by the weight $w(a,x)$. This weight is defined as:

$$w(a,x) = \frac{1}{d(a,x)} \tag{2.16}$$

where $d(a,x)$ is the path length between term $x$ and its ancestor $a$. After computing the $score(p,x)$ for all possible parents, the parent containing the highest score is chosen as parent of the term $x$.

### 2.1.6 Document subsumption

Similar to Distributional Inclusion, the document subsumption method (sometimes also called co-occurrence analysis) identifies hierarchical relations between terms through conditional probabilities of the occurrence of terms in documents [19, 37, 106, 128, 153]. This idea was initially presented by Forsyth and Rada [37], where they measured the degree of association between terms using cohesion statistic. The generality and specificity relations were determined by their document frequency (`df`). Thus, the more documents a term occurred in, the more general it is assumed to be.

Sanderson and Croft [128] present a measure based on the probabilities of term co-occurrences. This measure dictates that a term $x$ subsumes another term $y$, *i.e.*, term $x$ is an hypernym of $y$, if the relations in Equation 2.17 hold:

$$P(x|y) > \lambda \quad \text{and} \quad P(x|y) > P(y|x) \tag{2.17}$$

where $P(x|y)$ is the conditional probability of $x$ given $y$ and $\lambda$ is a threshold. In other words, a term $x$ is said to subsume $y$ if the documents in which $y$ occurs are a subset of the documents in which $x$ occurs. In their work the threshold was set to 1, but after some experiments the authors noticed that many terms were just failing to be included because a few occurrences of the subsumed term, $y$, did not co-occur with $x$. Thus, after an informal analysis of subsumption term pairs, the threshold was set to 0.8.

Later on, Njike-Fotzo and Gallinari [106] extended this idea replacing the conditional probability $P(x|y)$ as basic co-occurrence evidence by another probability expression. Assuming $\mathcal{D}$ the set of all documents, and $\mathcal{D}^{(x)}$ the subset of $\mathcal{D}$ where the term $x$ is found, the new probability expression is given by:

$$P(x,y) = \frac{|\mathcal{D}^{(x)} \cap \mathcal{D}^{(y)}|}{|\mathcal{D}^{(y)}|} \tag{2.18}$$

Given this distinction, Njike-Fotzo and Gallinari keep a similar expression of subsumption as Sanderson and Croft – Equation 2.17 – *i.e.*, a term $x$ will be considered hypernym of term $y$ if and only if:

$$P(x,y) > \lambda \quad \text{and} \quad P(x,y) > P(y,x) \tag{2.19}$$

However, the work by Njike-Fotzo and Gallinari also proposes a distinct method to estimate subsumption pairs, not by counting the documents as in Equation 2.18, but using an EM (Expectation-Maximization) algorithm [48]. Such approach is based on log-likelihood indices and it is less accurate than the one stated in Equation 2.18. However, such approach is faster to compute, thus, being applicable to larger text sources.

Chuang and Chien [19] also employ a sort of co-occurrence analysis, since they look for similar terms that occur in snippets from web searches. Since the problem tackled by the authors is to extract hierarchical relations from short text segments, the search space is reduced to the relations appearing in the snippets. As a consequence, the computation of the similarity between terms tends not to be a burden, since few terms are extracted from each text segment. The experiments report a good compromise between computation time and F-measure, where a decrease of time still results in an increase of F-measure from 68% up to 81%.

A more recent approach using co-occurrence is proposed by Wang *et al.* [153], making a recursive construction of topical hierarchy of phrases in textual bases. That work has a different approach for subsumption detection, since it starts from two arbitrary end nodes and then computes a network of co-occurrences. It is remarkable that this work employs a maximum likelihood method to compute parameters and it also performs an EM algorithm to infer model parameters. After setting up the co-occurrence model, the topical relations are constructed by frequency estimation followed by ranking of candidate relations.

Similarily to Sanderson and Croft [128], De Knijff *et al.* [27] use the conditional probabilities of the occurrence of terms in documents in order to identify taxonomic relations. The idea is that the generality and specificity relation is determined by their document frequency (`df`). Thus, the more documents a term occurred in, the more general it is assumed to be.

Co-occurrence methods may also be mixed with other methods in order to identify hierarchical relations between terms. For instance, Caraballo [15] labeled the hypernyms of a hierarchical cluster by verifying the term subnode of the association that has the largest number of children, as presented in Section 2.1.5. Instead of using that approach, Caraballo could use the co-occurrence analysis to choose the broader term of the association.

## 2.2 Supervised algorithms

Unlike Hearst [52] that relied on hand-built lexico-syntactic patterns, Snow *et al.* [135] studied the possibility of learning them automatically. In their work, Snow *et al.* extract examples of hypernym-hyponym pairs from WordNet [99] and for each pair, they find the sentences in which both words occur. These sentences are then parsed and the patterns are automatically extracted. A hypernym classifier is then trained using aproximately 70,000 patterns extracted from a corpus of 6 million newswire sentences. Having the classifier trained, a pair of ordered words is given and the classifier makes a binary decision whether the nouns are related by hypernymy or not.

Analysing the patterns used in the classifier, Snow *et al.* found that most of them give very weak evidence of a hypernym relation. On the other hand, the authors "rediscovered" the hand-designed patterns originally proposed by Hearst [52, 53] among the patterns with the highest precision. In addition, they were able to capture new patterns containing high scores, as presented in Table 2.2. The dependency path is based on the broad-coverage dependency parser MINIPAR [81] and represents a syntactic relation between two words.

| # | Pattern | Dependency path |
|---|---|---|
| 1 | NP like NP | N:PCOMP-N:PREP,like,like,PREP:MOD:N |
| 2 | NP called NP | N:DESC:V,call,call,V:VREL:N |
| 3 | NP is a NP | N:S:VBE,be,be,-VBE:PRED:N |
| 4 | NP, a NP (appositive) | N:APPO:N |

Table 2.2: Lexico-syntactic patterns and dependency paths found by Snow *et al.*

As pointed out by Cederberg and Widdows [16], patterns are useful only to classify noun pairs which happen to occur in the same sentence. As the pattern information within a sentence in a corpus is quite sparse, many hypernymy/hyponymy relations are lost. In order to improve the hypernym/hyponym extraction, Snow *et al.* detect coordinate terms using the distributional similarity [110, 117] between terms as well as the pattern based technique "$\mathtt{NP}_1$, $\mathtt{NP}_2$ and $\mathtt{NP}_3$" where the first NP is connected to a hypernym. The new classifier improved the results when comparing with the classifier without coordinate terms.

McNamee *et al.* [95] apply the same techniques described by Snow *et al.* to detect hypernymic relations between named entities in order to improve the performance of a Question Answering system. The model used by McNamee *et al.* differs from the work by Snow *et al.* because it is tailored in several ways. McNamee *et al.* use a support vector machine (SVM-Light) instead of the logistic regression model, the size of training corpora is increased to about 16 million sentences, thus increasing the coverage. Also, they include additional features not based on dependency parses (*e.g.*, morphology and capitalization).

Recently, Xuan Do and Roth [162] propose a system called TAxonomic RElation Classification (TAREC). This system consists of a training algorithm that learns from a supervised training data set a local classifier, and a testing algorithm. Given two terms, the system determines the taxonomic relation between them using a machine learning-based approach based on the information from Wikipedia. Output relations may be classified as hyponym, hypernym, coordinate terms (terms that share the same hypernym) or not related terms. The training model is presented in Algorithm 1, where the input is the Wikipedia data $\mathcal{W}$ and a supervised training data $\mathcal{D}$, composed by triples($x$, $y$, $rel$), where $x$ and $y$ are two terms and $rel$ is their taxonomic relation. The function `WikiRepresentation`($term$, $\mathcal{W}$) generates a Wikipedia-based semantic representation for the input $term$. The new triple ($\mathfrak{R}_x$,$\mathfrak{R}_y$,$rel$) contains the semantic representation of both terms and taxonomic relation between them. The new data are then used to train the local multi-class classifier ($\mathcal{C}$) to predict relations.

**Fragmento 1** TAREC training algorithm [162]

```
INPUT:
    Supervised data $\mathcal{D} = (x, y, rel)$
    Wikipedia $\mathcal{W}$
ALGORITHM:
    $\mathcal{D}' = 0$
    for each $(x, y, rel) \in \mathcal{D}$
        $\mathfrak{R}_x \leftarrow$ WikiRepresentation($x$, $\mathcal{W}$)
        $\mathfrak{R}_y \leftarrow$ WikiRepresentation($y$, $\mathcal{W}$)
        $\mathcal{D}' = \mathcal{D}' \cup (\mathfrak{R}_x, \mathfrak{R}_y, rel)$
    $\mathcal{C} \leftarrow$ ExtractFeaturesAndTrainClasifier($\mathcal{D}'$)
RETURN: $\mathcal{C}$
```

The TAREC testing algorithm uses the multi-class classifier ($\mathcal{C}$) to predict the taxonomic relation between a pair of terms ($x$, $y$), as presented in Algorithm 2, where $\mathcal{W}$ represents the same Wikipedia data used in the training algorithm, the function `WikiRepresentation`($term$, $\mathcal{W}$) generates a Wikipedia-based semantic representation of $term$. These representations of $x$ and $y$ are classified by $\mathcal{C}$, getting the probability distribution $\mathcal{P}_{x,y}$ over relation classes (hyponym, hypernym, coordinate terms or not related). Thus, the probability distribution is used in a relational constraint-based inference model that takes advantage of additional related concepts from the YAGO ontology [138] ($\mathcal{Z}_{x,y}$) to predict the taxonomic relation between $x$ and $y$.

**Fragmento 2** TAREC testing algorithm [162]

```
INPUT:
    Pairs of terms ($x$, $y$)
    Wikipedia $\mathcal{W}$
    Taxonomic relation classifier $\mathcal{C}$
ALGORITHM:
1.  $\mathfrak{R}_x \leftarrow$ WikiRepresentation($x$, $\mathcal{W}$)
2.  $\mathfrak{R}_y \leftarrow$ WikiRepresentation($y$, $\mathcal{W}$)
3.  $\mathcal{P}_{x,y} \leftarrow$ Classify($\mathfrak{R}_x$ , $\mathfrak{R}_y$ , $\mathcal{C}$)
4.  $\mathcal{Z}_{x,y} \leftarrow$ ExtractRelatedTerms($x$, $y$)
5.  rel $\leftarrow$ ConstraintBasedInference($\mathcal{P}_{x,y}$ , $\mathcal{Z}_{x,y}$ , $\mathcal{C}$)
RETURN: rel
```

## 2.3 Summary

Methods for extracting hierarchical or taxonomic relations from text corpora can be divided into two major groups: little or no supervised algorithms and supervised algorithms. Supervised algorithms use a machine learning methodology in which a predictive model is trained using examples of data that are provided along with their label. A bottleneck of such methods can be the availability of annotated data. Although there are works in the literature that use supervised algorithms to extract taxonomic relations between terms, we do not have manually annotated data to use for the training step, and thus, this kind of algorithm is out of the scope of this thesis. On the other hand, methods that use little or no supervised algorithms rely on the throughout analysis of the text contents or the application of ground rules in order to estimate possible taxonomic relations.

Over the years, many works have developed methods that use little or no supervised algorithms for extracting taxonomic relations between terms. Hearst [52] in her seminal work uses lexico-syntactic patterns in the form of regular expressions to identify relations between words in definitions. Pattern-based approaches in general are heuristic methods that apply regular expressions to match a sequence of words in the text. These patterns can extract one or more relations between words in the same phrase. However, hierarchical relationships that are expressed in "such as" or its derived patterns, mainly in domain corpus that are not very large are rather limited. Xuan Do and Roth [162] also state the problem with the sparsity of the terms, where infrequent terms are less likely to be covered, and may not be effectively extracted since they do not usually appear in close proximity with other terms. Also, due to the close proximity of the patterns, they inevitably make errors. For example, the sentence "I saw wild animals in Africa such as lions." will match the pattern `NP such as NP`, but according to it “Africa” is the hypernym of “lions”, instead of “wild animals”. Even though producing some mistakes, these patterns usually have a high precision. On the other hand, high quality patterns (in opposite to generic patterns) typically have very low recall. Patterns also have to be manually adapted when applied to other languages, and thus, if a system requires a new language, the patterns must be adapted.

Some approaches rely on the fact that the internal structure of noun phrases can be used to discover taxonomic relations. These head-modifier based methods are based on the heuristic where additional modifiers added to the head of a noun phrase typically define its hyponyms. That means, for instance, that “Computer Scientist” can be interpreted as a hyponym of “Scientist”. Although this heuristic has a higher precision, it is very limited and only occur when a noun phrase contains modifiers. Thus, relations between terms that do not contain the same head can not be inferred by this method. Ponzetto and Strube [119] also point out that due to the polysemic nature of the words, approaches that use head-modifier tend to mix all the meanings of a word in only one instance of this word, and thus leading to errors. This errors occurs when the modifiers select different senses for the respective heads (*e.g.*, “Conference chair” and “chair”) or the relation expressed is not an `is-a` relation (*e.g.*, meronymy as in “West Java” and “Java”).

In order to discover relationships that are not explicitly expressed by patterns in the corpus, many approaches have been developed. Some approaches are based on distributional analysis and use the similarity between vectors of co-occurrences to group semantic similar words together. However, Cimiano *et al.* [22] affirm that similarity-based methods do not provide a high level of traceability due to the fact that it is the numerical value of the similarity between two high-dimensional vectors which drives the clustering process and which thus remains opaque to the engineer. In agglomerative approaches of hierarchical clustering, initial merges of small-size clusters correspond to high degrees of similarity and are thus more understandable, while in divisive approaches the splitting of clusters aims at minimizing the overall cluster variance thus being harder to trace.

Widdows [158] emphasizes that taxonomies obtained through these approaches are very hard to evaluate by a human judge, since the relations are learned on the basis of statistical measures and prone to noise. Also, as noted by Xuan and Roth [162], algorithms based on distributional analysis typically suffer from a trade-off between precision and recall, resulting either in a relatively accurate resource with low coverage or a noisy resource with broader coverage.

# 3. Approaches for Taxonomy Evaluation

Evaluating hierarchical structures is still a hard task, being difficult even for humans, due to the fact that there is no unique way to generate correct structures and sometimes different taxonomies may model a domain equally well. An evaluation may be manually or automatically performed. The former is based on manual evaluation where domain experts assess the structure and its relations. The latter can be performed comparing the automatically extracted taxonomy against a gold standard by comparing both structures and relations, or checking the adequacy within an application when using such structure. This chapter discusses the results obtained by the works described in Chapter 2. The strategies for evaluating hierarchical relations are separated into two main groups: manual evaluation and automatic evaluation.

## 3.1 Manual evaluation

Hearst [53] extracts hierarchical relations to nouns using patterns as presented in Section 2.1.1. The manual evaluation is performed using a set of 200 consecutive instances of the "`LNP or other NP`" pattern (pattern 4 in Table 2.1) extracted from six months worth of text from New York Times texts. The results are classified into eight categories. Terms classified as "very good" or "pretty good" are meant to approximate the judgment that would be made by a WordNet lexicographer about whether or not to place the relation into WordNet. Using this evaluation, 104 out of 166 (63%) were either already present or strong candidates for inclusion in WordNet.

Caraballo [15] presents a hierarchical clustering of terms using Wall Street Journal Penn Treebank corpus [91] and 1987 Wall Street Journal [18]. The evaluation of the generated hierarchy was performed manually by three judges, using 10 internal nodes of the cluster containing at least 20 nouns under the selected node. Using a strict criteria, *i.e.*, considering only the hyponym judged as "the best", the algorithm achieved a precision of 33% considering that at least two judges evaluated a hyponymy relation as valid, and 39% considering that at least one judge evaluated it as correct. Using a looser criteria, *i.e.*, considering the second and third best hyponyms, the algorithm achieved a precision of 47.5% considering the majority of judges voting the relation as valid, and 60.5% considering that at least one voting it as valid.

Pantel and Pennacchiotti [112] evaluate *Espresso* using a sample of Aquaint (TREC-9) newswire text collection containing 5,951,432 words, and a small dataset of 313,590 words from a college level textbook of introductory chemistry, called CHEM [11]. *Espresso* was tested using two configurations: `ESP-` and `ESP+`, where `ESP+` exploits the same patterns of `ESP-` plus generic patterns, *i.e.*patterns with high recall and lower precision such as "`NP is a NP`". The evaluation is performed using two human judges who assessed a random sample containing 50 instances from the TREC corpus and 20 instances from the CHEM corpus. The precision for a given set of instances is the ratio between the sum of the judges' scores and the total of evaluated instances. Relative recall was calculated comparing to another system's recall. Thus, the relative recall of a system $A$ given system $B$, $R_{A|B}$ is calculated as:

$$R_{A|B} = \frac{R_A}{R_B} = \frac{\frac{C_A}{C}}{\frac{C_B}{C}} = \frac{C_A}{C_B} = \frac{P_A \times |A|}{P_B \times |B|} \tag{3.1}$$

where $R_A$ is the recall of $A$, $C_A$ is the number of correct instances extracted by $A$, $C$ is the total number of correct instances in the corpus, $P_A$ is $A$'s precision of the evaluated relations, and $|A|$ is the total number of instances discovered by $A$. The relative recall was calculated using ESP− as reference. Thus, ESP+ obtained the relative recall of 8.26 in TREC-9 corpus, which means that ESP+ outputs 8.26 times more correct relations than ESP− (at a cost of half of the precision). The *Espresso* algorithm without exploiting generic patterns (ESP−) achieved a precision of 73% using TREC corpus. Exploiting generic patterns (ESP+), the precision decreased to 36.2% but the recall increased 8 times more when comparing with the algorithm without exploiting them. Using CHEM corpus, the algorithm without exploiting generic patterns achieved a precision of 85%. It increased the recall about 7 times but the precision decreased to 76%. As claimed by Pantel and Pennacchiotti, the addition of generic patterns substantially improves recall without much deterioration in precision.

Cederberg and Widdows [16] extract hypernyms using Hearst patterns on the British National Corpus (BNC). A manual evaluation was performed using a random sample of 100 out of 513 extracted relations, from which 40% were correct relations. This precision is less than the one reported by Hearst [53]. They argue that this discrepancy could be due to the use of BNC instead of Grolier's encyclopedia. As noted by Hearst, the encyclopedia is designed to be especially rich in conceptual relationships. In order to improve the precision LSA is applied to filter out non-related terms. It is very efficient, since LSA groups semantically related terms and let non-related terms far apart. After applying LSA, Cedeberg and Widdows evaluated another random sample containing 100 relations, achieving a precision of 58%. Recall is then improved using coordination patterns, *i.e.*, patterns containing conjunction and appositives, identifying terms that belong to the same hypernym. Recall was increased almost fivefold and the precision increased from 40% when only pattern-matching was used, to 46%. Finally, Cederberg and Widdows combine all those techniques to extract hierarchical relations. Thus, they first extract hypernymy/hyponymy relations using Hearst patterns, then increase the quantity of hyponyms using coordination patterns, and finally, apply LSA to filter out non related terms. The final result is a precision of 64%, which is better than any of these techniques alone. They also noted that increasing the threshold of the Cosine similarity between terms would increase the precision, but on the other hand the recall would decrease.

Sanderson and Croft [128] present the generation of a taxonomy from TREC collection based on the probabilities of the terms co-occurrence, *i.e.*, the frequency in which terms co-occur through documents. The evaluation was performed by 8 human judges that evaluated 50 concept hierarchies, classifying them as "interesting" or "not interesting". It resulted in 67% of the relations judged as interesting, against 51% of baseline containing random associations. When evaluators were asked to classify these relations in one of the following categories, synonymy (*i.e.*, the same as the parent), antonymy (*i.e.*, the opposite of the parent), hyponymy (*i.e.*, a type of the parent), holonymy relations (*i.e.*, an aspect of the parent, *e.g.*, an actor is an aspect of a movie), or "don't know", most of the relations (49%) were evaluated as holonymy and 23% of them as hyponymy.

## 3.2 Automatic evaluation

Hearst [52] extracts hierarchical relations between nouns using patterns as presented in Section 2.1.1. The evaluation is based on the extraction of unmodified nouns (*i.e.*, single nouns without modifiers) for the pattern "`NP such as LNP`" (pattern 1 in Table 2.1) from 8.6M words of Grolier's American Academic Encyclopedia [47]. Extracted relations where compared with WordNet version 1.1 verifying the existence of both hypernym and hyponym, and the relation in the gold standard. Grolier's American Academic Encyclopedia contained 7067 sentences with the selected pattern and 61 out of 106 (58%) feasible relations (*i.e.*, relations in which both terms were already registered in WordNet).

Ponzetto and Strube [119] use the patterns presented by Hearst [52] and head-modifier identification algorithms (Section 2.1.2) for building a taxonomy based on the structure of Wikipedia categories. The automatic evaluation was performed in order to verify how well the generated taxonomy compares with other existing resources. Thus, ResearchCyc[1], the research version of the Cyc knowledge base [76] and WordNet [34] version 3.0 were used as gold standards. Matchings between the generated taxonomy and the gold standards are measured in terms of Coverage, Novelty and ExtraCoverage. Let $G_{Wiki} = \langle V_{Wiki}, E_{Wiki} \rangle$ be the taxonomy, where the vertices represent the categories and the edges the generated `is-a` relations, and $G_S = \langle V_S, E_S \rangle$ a gold standard taxonomy. A subgraph containing the mapping of $G_{Wiki}$ in $G_S$ is represented by $G'_{Wiki} = \langle V'_{Wiki}, E'_{Wiki} \rangle$, being the Coverage calculated as presented in Equation 3.2:

$$Coverage(G'_{Wiki}, G_S) = \frac{|E'_{Wiki}|}{|E_S|} \tag{3.2}$$

Coverage quantifies how many pairs in the generated taxonomy can be mapped to concepts in the gold standard to the total number `is-a` relations in the latter. Coverage thus measures the size of the intersection between the generated taxonomy and the gold standard. In order to verify the number of pairs of the generated taxonomy that are deemed to be an `is-a` relation but have no suitable mapping in the gold standard, the Novelty is calculated according Equation 3.3:

$$Novelty(G_{Wiki}, G'_{Wiki}) = \frac{|E_{Wiki} \setminus E'_{Wiki}|}{|E_{Wiki}|} \tag{3.3}$$

Finally, ExtraCoverage (EC) can compute the proportional gain of unmapped relations from the generated taxonomy against the total number of semantic relations in the gold standard, as presented in Equation 3.4:

[1] http://research.cyc.com/

$$EC(G_{Wiki}, G'_{Wiki}, G_S) = \frac{|E_{Wiki} \setminus E'_{Wiki}|}{|E_S|} \tag{3.4}$$

The generated taxonomy was evaluated and regardless of the gold standard employed the overall coverage was very low, having up to 1.6% for Cyc and 8.7% for WordNet. The authors claim that this low coverage may be due to the nature of Wikipedia, which provides a thematic meta-classification scheme for the resource's encyclopedic entries. On the other hand, the low coverage is counterbalanced by an extremely high novelty rate, having above 99% for all methods and gold standards, and substantial extra-coverage, having up to 28.2% for Cyc and up to 211.6% for WordNet. The best results were achieved by the combination of all methods. A manual gold standard was created using a random sample of 3500 annotated category pairs. These pairs were annotated as either `is-a` or `not-is-a` relationships. Precision ($\mathcal{P}$) is calculated by the ratio of correct `is-a` relations to total `is-a` labels generated by the system. Recall ($\mathcal{R}$) is calculated by the ratio of correct `is-a` relations to total `is-a` labels in the gold standard. F-measure ($\mathcal{F}$) is calculated as:

$$\mathcal{F} = \frac{2 \cdot \mathcal{P} \cdot \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{3.5}$$

Syntax, connectivity and pattern-based methods were evaluated separately, but the best results were obtained by the combination of all methods. When comparing with the best method alone, the combination of all methods achieved a 22.3% improvement in recall (from 69.9% to 84%) and a 2.1% decrease in precision (from 92.40% to 90.30%), resulting in an overall improvement of 14.1% F-measure (from 68.1 to 80.3). It was an improvement of 27.9% F-measure with respect to a baseline using a random classification scheme, *i.e.*, a category pair is randomly categorized as `is-a` or `not-is-a`.

Weeds *et al.* [156] expect that distributional generality is correlated with semantic generality, *i.e.*, a word has high recall/low precision retrieval of its hyponyms' co-occurrences and high precision/low recall retrieval of its hypernyms' co-occurrences. In order to verify the hypothesis, they used a list containing 2,000 nouns extracted from BNC, and all possible pairs between these nouns that exist in WordNet 1.6 (20,415 pairs). They found that the taxonomic relation is correlated with the direction predicted by precision and recall in 71% of cases.

Geffet and Dagan [43] propose the Distributional Inclusion Hypothesis 2.1.4 which says that if the meaning of a word $u$ entails another word $v$, then it is expected that all the typical contexts (features) of $u$ will occur also with $v$. In order to test the proposed hypotheses they perform a recalculation on weighted vectors taking into account the set of most similar words generated by the Lin's measure [80]. The validity of the hypotheses was tested on a sample of 400 manually annotated pairs extracted from 18 million tokens subset of the Reuters RCV1 corpus[2] and weigthed using Relative Feature Focus (RFF) [42]. The result shows that 86% of the entailments tested preserve the feature inclusion and 70% of the features inclusion preserve the entailment.

Kotlerman *et al.* [67] crafted the `balAPinc` measure (Equation 2.11) which is optimized to

[2] Known as Reuters Corpus, Volume 1, English Language, 1996-08-20 to 1997-08-19.

capture a relation of feature inclusion between terms while using the relative relevance of features. Directionality between a hypernym relation using balAPinc was tested with terms and dependency relation as their features extracted from Reuters RCV1 corpus. The directionality of pairs of terms was compared with a sample of 1,886 manually judged term pairs (each pair was assessed in both directions for lexical entailment, resulting in 1,067 valid and 2,705 invalid directional pairs [167]). Each pair obtained a score according to the directional measure, quantifying its belief that the pair is valid. All pairs are sorted by their scores and assessed the quality of the list by calculating its Average Precision score based on the manually annotated gold-standard. Using this configuration, balAPinc obtained the same average precision as ClarkeDE, AP=47%, with a precision P=32% and recall R=92%.

Lenci and Benotto [77] explore the possibility of identifying hypernyms using a directional similarity measure (InvCL 2.10) that takes into account not only the inclusion of the features of $u$ in $v$, but also the non-inclusion of the features $v$ in $u$. Distributional similarity measures were applied on lexical items using distributional feature vectors extracted from the TypeDM tensor [5]. TypeDM is a particular instantiation of the Distributional Memory (DM) framework containing a set of 30,693 lemmas (20,410 nouns, 5,026 verbs and 5,257 adjectives). The directional similarity was evaluated using a subset of *BLESS* [6] data set. *BLESS* consists of tuples expressing a relation between a target and a relatum concept. Target concepts include 200 distinct English concrete nouns, equally divided between living and non-living entities, and are grouped into 17 broader classes (*e.g.*, BIRD, FRUIT, VEHICLE, *etc.*). The subset of *BLESS* used is formed by 14,547 tuples with both concepts attested in the TypeDM word set. Evaluation was performed in terms of Average Precision (AP), a method that combines precision, relevance ranking and overall recall. Thus, the best possible score (AP=1.0) corresponds to the case in which all tuples containing the hypernym relation (HYPER) have higher similarity scores than the tuples that contain other relations (COORD, MERO, RANDOM-N). Reported results show that InvCL achieved the highest average precision AP=40% followed by AP=38% achieved by ClarkeDE.

Santus *et al.* [130] use an entropy-based measure named SLQS for the unsupervised identification of taxonomic relations in distributional similarity models. SLQS is grounded on the idea that contexts of hypernyms are less informative than the most typical linguistic contexts of its hyponyms. SLQS was tested in a directionality task using 1,277 pairs of hypernym–related pairs of *BLESS* [6] data set. Using content words (nouns, proper nouns, adjectives and verbs) as context from a combination of ukWaC (1.915 billion words) and WaCkypedia (820 million words) corpora they were able to successfully identify the correct direction in 87% of the pairs.

Snow *et al.* [135] use machine learning in order to automatically learn pattern spaces based on syntactic dependency paths. These paths represent the relationship between hypernym/hyponym word pairs from WordNet and are used as features in a logistic regression classifier. The syntactic dependency paths were generated using MINIPAR [80] in a corpus containing over 6 million newswire sentences. They trained a variety of classifiers, including multinomial Naive Bayes, complement Naive Bayes, and logistic regression, as well as two classifiers containing the patterns suggested by

Hearst [52] and the coordination patterns presented by Caraballo [15].

A model selection was performed using 10-fold cross validation on this training set, evaluating each model based on its maximum hypernym F-Score averaged across all folds. The binary logistic regression model shows a 132% relative improvement of average maximum F-score over the classifier based on Hearst's patterns, achieving an average maximum F-score F=34.8%. In a hybrid hypernym-coordinate classification task, a classifier based on the WordNet taxonomy (version 2.0) using only the first sense of a hyponym with a maximum distance of 4 between a hyponym and hypernym, and allowing any member of a hypernym synset to be a hypernym had the best F-measure. It achieves the F-measure of 33.57%, while a model containing only the patterns proposed by Hearst obtained a F-measure of 14.17% and the coordination patterns presented by Caraballo obtained a F-measure of 13.86%. The best classifier had a 43% relative maximum F-measure improvement over the best WordNet classifier. The combination of the linear interpolation Hypernym with the coordinate model had a 40% relative maximum F-measure improvement, while the logistic regression hypernym-only model had a 16% improvement.

McNamee *et al.* [95] apply a tailored version of the same techniques described by Snow *et al.* [135], using a corpus containing about 16 million sentences (a joint of TREC Disks 4 and 5, Aquaint and Wikipedia). This tailored version contains a support vector machine (SVM-Light) instead of a logistic regression model and additional features not based on dependency parses (*e.g.*, morphology and capitalization) are included. In order to evaluate this approach, the authors created a baseline containing a comparable weakly-supervised hypernym classifier of Snow *et al.* [135]. A model containing entity-enriched data extended the baseline training set by adding positive examples and a model that uses additional features besides dependency paths were created. A test set containing 75 categories was created and the results to each model were manually evaluated. The model using additional features gained 11% over the baseline condition, having a mean average precision of 53% and the maximum F-measure of 55% with 70% recall. The model containing additional entities gained 4% over the baseline, having a mean average precision of 50%.

Xuan Do and Roth [162] describe a machine learning approach that uses Wikipedia as a semantic resource to train a model that, given two terms, determines their taxonomic relation. They develop as well a global constraint-based inference process that leverages existing knowledge bases to enforce relational constraints among terms and thus improves the classifier predictions. To evaluate their approach, the authors use part of a manually constructed dataset containing 11,000 instances which were used to evaluate information extraction tasks [115], and a data set generated from 44 semantic classes of more than 10,000 instances [152]. Using the same datasets, the authors performed an evaluation comparing TAREC system with the systems presented by Ponzetto and Strube [118], Snow *et al.* [136] and Suchanek *et al.* [138]. The results showed that TAREC significantly outperformed other systems, achieving 85.34% of accuracy using the first dataset and 86.98% using the second dataset. The authors claim that other systems did not perform well because they rely heavily on string matching techniques to map input terms to their respective ontologies, being thus very inflexible and brittle. On the other hand, machine learning-based classifiers are very flexible in extracting features of the two input terms and are thus much better at predicting their taxonomic relation.

Cimiano *et al.* [23] use Formal Concept Analysis (FCA) to learn concept hierarchies from text corpora. Due to the fact that FCA does not produce appropriate names for the abstract concepts generated, it seems dificult to evaluate the learned taxonomy by computing the number of correct concepts generated against a gold standard hierarchy. Instead, they compute how similar the automatically learned hierarchy is with respect to the gold standard. Based on the work presented by Maedche and Staab [90] they use *Semantic Cotopy* (SC) to evaluate the automatically learned hierarchy, comparing the taxonomic overlap between two ontologies. Before introducing this measure, Cimiano *et al.* define a core ontology as:

> "A core ontology $O$ is a tuple $(C, root, \leq_C)$, where $C$ is a set of concept identifiers, $root$ is the top element of the partial order $\leq_C$ on $C \cup \{root\}$, which is called concept hierarchy or taxonomy."

In order to compare the taxonomy of two ontologies, the authors have adapted the Semantic Cotopy ($SC$), considering only the concepts in both learned and gold standard ontologies, and excluding the concept itself from its Common Semantic Cotopy ($SC''$), *i.e.*:

$$SC''(c_i, O_1, O_2) = \{c_j \in C_1 \cap C_2 \mid c_j <_{C_1} c_i \vee c_i <_{C_1} c_j\}$$

Thus, having the concept $c \in C_1$ and the concept $c' \in C_2$ maximizing the overlap between their respective semantic cotopies, the global Taxonomy Overlap $\overline{TO'}(O_1, O_2)$ between the two ontologies is defined as follows:

$$\overline{TO'}(O_1, O_2) = \frac{1}{|C_1 \backslash C_2|} \sum_{c \in C_1 \backslash C_2} max_{c' \in C_2 \cup \{root\}} \frac{S_1}{S_2}$$

where:

$$S_1 = |SC''(c, O_1, O_2) \cap SC''(c', O_2, O_1)|$$

$$S_2 = |SC''(c, O_1, O_2) \cup SC''(c', O_2, O_1)|$$

Finally, from $\overline{TO'}$, Cimiano *et al.* introduce precision $P(O_1, O_2)$, recall $R(O_1, O_2)$ and F-measure $F(O_1, O_2)$, calculating the harmonic mean of the taxonomic overlap in both directions:

$$P(O_1, O_2) = \overline{TO'}(O_1, O_2)$$

$$R(O_1, O_2) = \overline{TO'}(O_2, O_1)$$

$$F(O_1, O_2) = \frac{2 \cdot P(O_1, O_2) \cdot R(O_1, O_2)}{P(O_1, O_2) + R(O_1, O_2)}$$

Using Taxonomy Overlap, Cimiano *et al.* performed an evaluation on the tourism and finance domains. A manually built ontology for the tourism domain [90] and a finance ontology developed within the GETESS project [137] were used as gold standards. The text collections for the tourism domain were acquired from *LonelyPlanet*[3] and *All in all*[4], and for the finance domain were acquired from Reuters news from 1987. Furthermore, they used British National Corpus[5] as a general corpus. Using Taxonomy Overlap in the tourism domain, they achieved the best F-measure of 40.52% with a precision of 29.33% and recall of 65.49%. Taxonomies extracted from the financial domain had a lower performance, achieving a F-measure of 33.11% with almost the same precision 29.93% and a lower recall of 37.05%.

Rios-Alvarado *et al.* [123] build a concept hierarchy from corpus by using an adaptation of the clustering algorithm proposed by Pantel [110], a set of linguistic patterns identified by Hearst [52] and Snow *et al.* [135], and additional contextual information from the Web that improves the discovery of the most representative hypernym/hyponym relationships. The evaluation was performed over four different gold standards: LonelyPlanet, SmartWeb Football, Biology News Net website[6] and Java [74]. Precision, recall and F-measure were obtained for each dataset, having the best F-measure 80%, with precision 77% and recall of 83% using the SmartWeb Football dataset. LonelyPlanet data set achieved the best recall 89% with a precision of 53% and F-measure of 67%. The authors compared their results of precision for LonelyPlanet dataset with the work by Cimiano *et al.* [23] and had an improvement of 3% in precision. They also achieved an improvement of 3% in precision when compared the results obtained using the SmartWeb Football dataset with the work by Jiang and Tang [59].

## 3.3 Summary

Manual evaluation is usually performed on novel or specific domains where a gold standard is not available. Furthermore, a manual evaluation allows to assess not only the taxonomic relations between terms, but also the quality of the whole structure in a detailed view. On the other hand, the manual evaluation is rarely performed on the whole taxonomy, but on a subset of all relations. For domain experts deciding whether or not a term belongs to the domain is more or less feasible. Furthermore, deciding the quality of a taxonomic relation is a more complex task. As mentioned by Velardi *et al.* [149], when annotators are asked to blindly produce a taxonomy from a given set of terms, they struggled with the domain terminology and produced a quite messy organization. Manual evaluation also has the drawbacks of being a time consuming task and depending on the availability of domain experts (for some domains it may be quite hard to find experts that can dedicate time for evaluation). Due to these factors, it is highly desirable to have a gold standard to support an automatic validation.

Works described in this thesis that manually evaluate the relations generated by hierarchical relation extraction methods are summarized in Table 3.1, where $\mathtt{P}$ denotes precision, $\mathtt{P}_{\mathtt{is\text{-}a}}$ means

[3] `http://www.lonelyplanet.com`
[4] `http://www.all-in-all.de`
[5] `http://www.natcorp.ox.ac.uk`
[6] `http://www.biologynews.net`

the precision to the is-a relations, $\mathtt{P_B}$ is the best precision achieved by the system, ESP+ represents the Espresso algorithm exploiting generic patterns, ESP− represents the Espresso algorithm without generic patterns, $\mathtt{R_{Rel}}$ is the relative recall, $\mathtt{MAP_{+NE}}$ means the Mean average precision with additional Named Entities, and $\mathtt{MAP_{+Feat}}$ is the Mean average precision with additional features. Usually a manual evaluation is performed on a subset of all relations, observing the precision of the method, *i.e.*, quantifying how good relations are when compared with human judgments.

| Reference | Resources used | How it is evaluated | Results |
|---|---|---|---|
| [53] | New York Times texts | 166 relations of the<br>"LNP or other NP" pattern | P=62.7% |
| [15] | Wall Street Journal<br>Penn Treebank corpus [91]<br>1987 Wall Street Journal [18] | ≈ 10 internal nodes<br>with 20 nouns each node<br>(3 evaluators) | from P=33% to P=60.5%<br>depending on the criteria |
| [128] | TREC collection | 50 concept hierarchies<br>(8 evaluators) | $P_{\text{is-a}}$=23% |
| [16] | British National Corpus (BNC) | Random sample with 100 relations<br>(2 evaluators) | P=64.0% |
| [135] | Articles of Associated Press<br>Wall Street Journal texts<br>Los Angeles Times texts<br>WordNet 2.0 [99] | 511 noun pairs (4 evaluators) | $F_B$=33.57% |
| [112] | Aquaint (TREC-9) newswire<br>text collection | 50 instances (2 evaluators) | (ESP+) P=73% $R_{rel}$=1.00<br>(ESP-) P=36% $R_{rel}$=8.26 |
| | CHEM dataset [11] | 20 instances (2 evaluators) | (ESP+) P=85% $R_{rel}$=1.00<br>(ESP-) P=76% $R_{rel}$=6.66 |
| [95] | TREC Disks 4 and 5 (newswire)<br>AQUAINT (newswire)<br>Wikipedia (April 2004)<br>WordNet 2.0 [99] | 75 categories (1 evaluator) | $MAP_{+NE}$=53%<br>$MAP_{+Feat}$=50% |

Table 3.1: Works that manually evaluate the extracted relations.

While the manual evaluation is always laborious, the automatic evaluation gives us the possibility of testing a set of relations, comparing results against a gold standard. A gold standard structure is a hierarchy manually constructed by domain experts and serves as reference of domain terms and taxonomy (*e.g.* WordNet [34]). In this kind of evaluation the quality of the obtained hierarchy is expressed by its similarity to the gold standard hierarchy, measuring the ability to reproduce the relations between pairs of words. The problem of using such structure is that imperfections in the gold standard will impact the results. This kind of imperfections vary from missing terms, *i.e.* the gold standard has low coverage and it does not contain all terms contained in the generated hierarchy, to missing relations between terms. For example, WordNet has an entry for *Robert De Niro* (United States film actor) as an instance of *actor*, a subclass of *performing artist*, *entertainer*, *person* and *being*, but not a subclass of *man*. Another problem of the automatic evaluation refers to the availability of gold standards for certain domains.

Table 3.2 summarizes works presented in this thesis that automatically evaluate the results generated by methods that extract hierarchical relations from texts, where $\mathtt{P}$ denotes precision, $\mathtt{R}$ recall and $\mathtt{F}$ F-measure, $\mathtt{F_t}$, $\mathtt{P_t}$ and $\mathtt{R_t}$ are the F-measure, precision and recall respectively, using the tourism gold standard, $\mathtt{F_f}$, $\mathtt{P_f}$ and $\mathtt{R_f}$ are the F-measure, precision and recall respectively, using the finance gold standard, $\mathtt{F_B}$ is the best F-measure achieved by the system, $\mathtt{MAP_{+NE}}$ means the Mean average precision with additional Named Entities, $\mathtt{MAP_{+Feat}}$ is the Mean average precision with additional features, $\mathtt{AP}$ denotes average precision, $\mathtt{Cv}$ represents the coverage, $\mathtt{Nv}$ the novelty, and $\mathtt{EC}$ the extra coverage, $\mathtt{ACC_1}$ the accuracy of the Test-I and $\mathtt{ACC_2}$ the accuracy of the Test-II. In this kind of evaluation the quality of the extracted relations is expressed by their similarity with a gold standard, assuming that the gold standard represents well and accurately the domain. However, the results are usually influenced by the imperfections of the gold standard. Lexical databases such as WordNet [34] or hand-crafted lists containing terms and their semantic relation are commonly used as gold standards. Usually, the results of automatic evaluation are in terms of precision, recall and F-measure, and denote how well a method "mimic" the gold standard.

| **Reference** | **Resources used** | **Evaluation** | **Results** |
|---|---|---|---|
| [52] | Grolier's American Academic Encyclopedia [47] | 106 relations of the "NP such as LNP" pattern using WordNet v.1.1 [100] | P=57.5% |
| [156] | British National Corpus (BNC) | 20,415 pairs of BNC using WordNet v.1.6 | P=71% |
| [43] | Reuters RCV1 corpus | 400 annotated pairs from Reuters corpus | P=70% |
| [23] | LonelyPlanet<br>All in all<br>BNC corpus<br>Reuters 1987[7] | Semantic Cotopy and Taxonomy overlap using Tourism ontology [90]<br>Finance ontology [137] | $F_t$=40.52%<br>$P_t$=29.33%<br>$R_t$=65.49%<br>$F_f$=33.11%<br>$P_f$=29.93%<br>$R_f$=37.05% |
| [135] | 6 million newswire sentences | 10-fold cross-validation based on WordNet v.2.0 | $F_B$=34.8% |
| [95] | corpus of 16 million sentences | Annotated pairs from 75 categories | $MAP_{+NE}$=53%<br>$MAP_{+Feat}$=50% |
| [67] | Reuters RCV1 corpus | 3,772 annotated pairs [167] | AP=47%<br>P=32%<br>R=92% |
| [119] | Wikipedia (March 2008) | ResearchCyc[8]<br>WordNet v.3.0 [34]<br>3,500 annotated category pairs | Cv=1.6%, Nv=99.2%, EC=28.2%<br>Cv=8.7%, Nv=99.3%, EC=211.6%<br>P=90.3%, R=78.5%, $F_B$=84% |
| [77] | TypeDM [5] | 14,547 tuples of BLESS [6] | AP=40% |
| [162] | Wikipedia (July 2008) | YAGO ontology [138] | $ACC_1$=85.34%<br>$ACC_2$=86.98% |
| [123] | LonelyPlanet<br>SmartWeb Football<br>Biology news<br>Java [74] | A gold standards associated to each corpus | P=53%, R=89%, F=67%<br>P=77%, R=83%, F=80%<br>P=49%, R=56%, F=52%<br>P=76%, R=72%, F=74% |
| [130] | ukWaC<br>WaCkypedia | BLESS [6] data set | P=87% |

Table 3.2: Works that automatically evaluate the extracted relations.

As pointed out by Velardi *et al.* [149] it is not clear how to evaluate the concepts and relations not found in the gold standard. As these terms can be either wrong or correct, the evaluation is in any case incomplete. Automatic evaluation also has the drawback of choosing an adequate evaluation metric. As presented in Section 3.2, taxonomy overlap computes the ratio between the intersection and union of two sets. Therefore this metric do not provide a structural comparison between the taxonomies, thus, errors in the hierarchy structure are not indicated by this metric.

Another common practice to evaluate the quality of the system automatically is to compare it against the state of the art systems, verifying their performance in some task. For instance, Rios-Alvarado *et al.* [123] reproduced the approaches presented by Cimiano *et al.* [23] and by Jiang and Tang [59] in order to compare with their proposal. The problem in emulate the approaches proposed by someone else is that not always the resources to reproduce the system are available (*e.g.*, the corpus to train a machine learning system). Also, usually results are compared between similar approaches (*e.g.*, Santus *et al.* [130] compare their results with other works that also use directional similarity measures).

Regarding the results obtained in each work, many are not comparable since the resources used in each approach were different. For instance, Hearst [52] achieved 57.5% of precision using encyclopedia texts. When Cederberg and Widdows [16] applied the same method on the British National Corpus they achieved a precision of 40%. Thus, it would be impossible to compare these values of precision and recall with the work presented by Cimiano *et al.* in [23] that built them using documents from finance and tourism domain. This difference shows how important it would be to perform an evaluation of these methods using the same corpora.

# 4. Materials and methods

This chapter describes the methodology used in this thesis for the development and evaluation of the methods for taxonomic relation extraction from text corpora in Portuguese and English, as well as the resources used and the process of evaluation. We divided the chapter into 4 main parts: resources (Section 4.1), preprocessing (Section 4.2), relation extraction methods (Section 4.3) and evaluation (Section 4.4). Resources section presents all content used in this thesis to perform experiments, including corpora and gold standards. Preprocessing section describes how texts are treated before being used in methods for generating taxonomic relations. Relation extraction methods section describes the methods developed in the context of the thesis. Evaluation section discuss the evaluation methodology and criteria. Figure 4.1 presents an overview of the whole processing, where `Corpora` and `Gold standard` are described in Section 4.1, `Parsing`, `Term extraction` and `Concept extraction` are described in Section 4.2, `Patt` (patterns-based method), `DSim` (method based on distributional inclusion hypothesis), `SLQS` (entropy-based method), `TF` (term frequency-based method), `DF` (document frequency-based metho), `DocSub` (document subsumption-based method), `Hclust` (hierarchical clustering-based method) and `Hmod` (head-modifier-based method) are described in Section 4.3, and how both manual and automatic `Evaluations` were performed is described in Section 4.4.

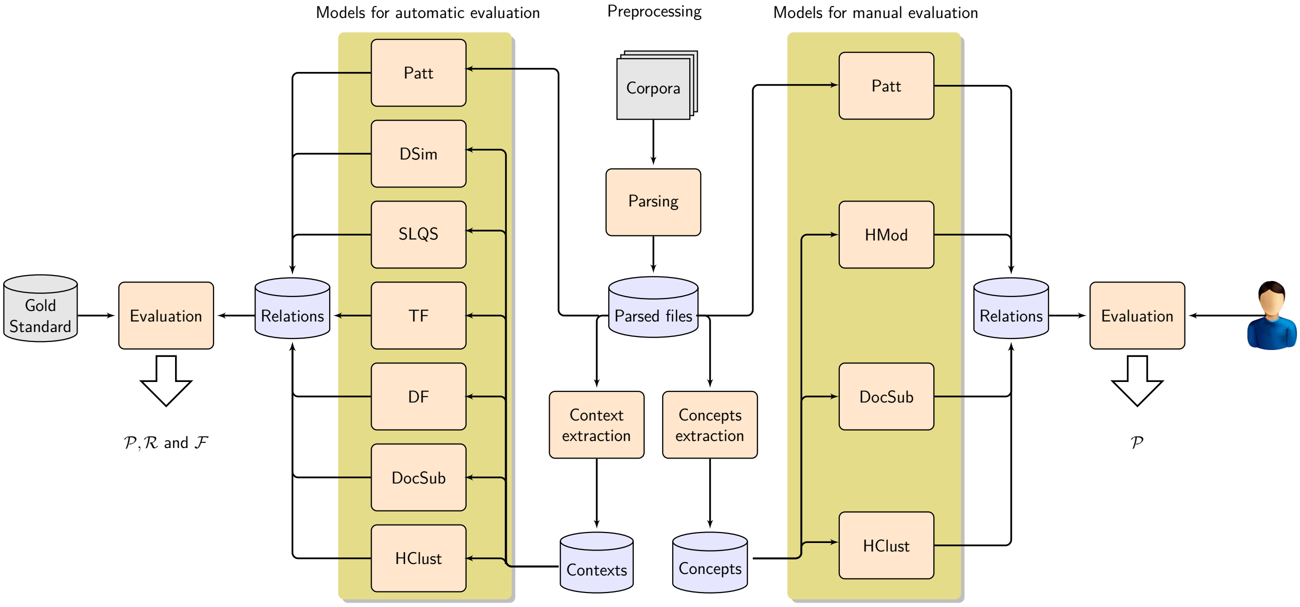


Figure 4.1: Overview of the methodology developed in the thesis

## 4.1 Resources

### 4.1.1 Corpora

Most of the works described in this thesis use text corpus as resource to generate taxonomies, *i.e.*, relations between words are extracted from plain text files. According to Liu *et al.* [82], although approaches that use this kind of resource achieve some success, they also have disadvantages. For example, for highly specific domains, it is difficult to find a text corpus that accurately characterizes them. It is easier to find a text corpus for general domains (*e.g.*, "computer science") than for specific domains or topics (*e.g.*, "big data for business intelligence") due to the fact that usually such topics are likely to be dispersed in many different places.

Since the Web has become a rich source of collective knowledge, many approaches have been using it as a resource to harvest new terms and identify relations between terms (*e.g.*, [69, 112]). Querying the Web in order to find related terms would be an alternative to text corpora. As noted by Granada *et al.* [45], the problem of using the Web as a corpus is the number of unrelated documents that the bootstrapping may retrieve, as well as the parsing of unstructured pages. Other works use Wikipedia as a source of information and its different aspects have been exploited, such as the hierarchical layouts [139, 140, 163], the categorical system [118, 119, 162], selected articles (*e.g.*, animals [54, 116]), infoboxes [160] or its whole content [95].

In this thesis we decided to use corpora of different natures. For instance, the automatic evaluation (Section 4.4.1) uses the Europarl and the Ted Talks corpora, while the manual evaluation (Section 4.4.2) is performed using the Europarl and Geology corpora. The corpora have differences in nature: for instance, a collection of speeches of the European Parliament, a corpora of transcribed talks which encompasses a vast area of knowledge, and encyclopedic-type documents related to the geology area. The nature of the documents may influence the number of extracted taxonomic relations, mainly using pattern methods, since such methods tend to capture relations in the definition of terms. The nature of the corpora also differs in terms of alignment between languages, having parallel and comparable corpora. According to McEnery and Xiao [94], a parallel corpus can be defined as a corpus that contains source texts and their translations, being bilingual or multilingual. In contrast, comparable corpus contains components that are collected using the same sampling frame and similar balance and representativeness, not being translations of each other. Although both terms contain many definitions, in this work, parallel corpus can be understood as a corpus containing bilingual translations of the phrases of the same document, and comparable corpus as containing non-sentence-alignment and non-translated documents but aligned under the same topic. An excerpt of aligned sentences from the Europarl parallel corpus is presented below:

| # | English: | Portuguese: |
|---|---|---|
| 1. | Madam President, I should like to draw your attention to a case in which this Parliament has consistently shown an interest. | Senhora Presidente, gostaria de chamar a sua atenção para um caso de que este Parlamento repetidamente se tem ocupado. |
| 2. | It is the case of Alexander Nikitin. | É o caso de Alexander Nikitin. |

A parallel corpus contains a direct translation of each sentence, embedding the same meaning at the sentence level. On the other hand, a comparable corpus does not have a direct translation of sentences. Thus, instead of sentences containing the same meaning, the meaning is at document or collection level. For example, Wikipedia's articles written in two or more languages are connected by interlanguage links, indicating that they have the same meaning or both articles are about the same subject. Unlike Wikipedia's articles or sentence-aligned documents, at a collection level the meaning is spread over the whole collection of documents. All corpora used in this thesis are briefly described below – note that `EN` in the name of the corpus means that documents are written in English, `PT` means that documents are written in Portuguese, and `EN-PT` means that one corpus is written in English and the other one is written in Portuguese. The corpora described below include not only the ones to generate taxonomic relations, but also the ones to contrast terms in concept extraction.

**`Europarl (EN-PT)`**: Europarl parallel corpus [66] is a collection of the proceedings of the European Parliament, comprising of about 30 million words for each of the 11 official languages of the European Union: Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese and Swedish. This bilingual corpus containing sentence aligned texts is freely available at the Statistical Machine Translation site[1]. In this work we use the English-Portuguese sentence-aligned version of the parallel corpus, *i.e.*, each sentence of the English corpus has its translation into Portuguese.

**`Geology (EN-PT)`**: The corpus of the Geology domain in Portuguese is composed of a set of 236 scientific texts collected from the public databases of theses, dissertations, technical reports, conference and journal papers. Domain specialists performed a shallow analysis of the texts and documents relevant to the domain were selected. The selected texts were transformed into plain textual format and then parsed using PALAVRAS [9]. The whole process of selecting, cleaning and parsing the corpus is presented by Lopes and Vieira [83].

The English version of the Geology corpus is a collection of 1,657 documents gathered from Geology.com website[2]. The set of documents includes subsets categorized into diamonds, earthquakes, gemstones, general geology, igneous rocks, metamorphic rocks, meteorites, *inter alia*. All pages of the website were crawled and the content extracted into plain text files.

**`TED Talks (EN-PT)`**: TED is a nonprofit organization and its website[3] makes available the video recordings together with subtitles provided in many languages of the best TED talks. Almost all talks have been translated by volunteers into about 70 other languages. The collection containing sentence aligned documents is provided by the Web inventory named WIT[3], an acronym for Web Inventory of Transcribed and Translated Talks [17]. The collection contains 1,112 transcribed and translated talks containing topics that span the entirety of human knowledge.

**`Brown Corpus (EN)`**: The Brown Corpus of Standard American English[4] is a general corpus

[1] http://www.statmt.org/

[2] http://geology.com

[3] http://www.ted.com

[4] http://icame.uib.no/brown/bcm.html

containing about 1 million words of various types of texts, being limited to written American English. It is provided in the Natural Language Toolkit (NLTK)[5] and contains 15 subsets of texts, separated into textual genres: adventure, belles lettres, editorial, fiction, government, hobbies, humor, learned, lore, mystery, news, religion, reviews, romance and science fiction.

`Conferences (EN)`: The corpus in the conference organisation domain [44] is a corpus constructed to support ontology-related tasks, such as multilingual ontology matching, extension, automatic ontology learning and population. It contains documents resulting from the crawling of Web pages using multilingual ontology concept labels as seeds in a search engine.

`Euronews (EN)`: This corpus was created by Saad *et al.* [126] as a comparable data set containing documents downloaded from the multilingual and pan-European television news channel EuroNews[6]. The corpus was extracted by parsing the HTML of each English news article and transforming into plain text files.

`FOOTIE (EN)`: Football in Europe (FOOTIE) corpus [129] was constructed from the transcription of the English press conferences scheduled before and after every game played by Italy's national team during the 2008 European football championships (UEFA EURO 2008) held in Switzerland and Austria. This corpus contains a set of 1,725 documents.

`Ohsumed (EN)`: Ohsumed corpus [55] was created to assist information retrieval research. It is a clinically-oriented MEDLINE subset[7], consisting of 348,566 references in English (out of a total of over 7 million), covering all references from 270 medical journals over a five-year period (1987-1991).

`CETEN-Folha (PT)`: CETEN-Folha[8] (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo) is a corpus containing about 24 million words in Brazilian Portuguese extracted from Folha de São Paulo newspapers. The corpus includes text published in 1994 (all the 365 editions). The texts are organized into one or more categories such as politics, sports, economy, culture, opinion, agriculture, vehicles, technology and not determined.

`Computer Science (PT)`: As the Geology corpus in Portuguese, the Computer Science corpus was constructed by collecting documents in Portuguese from the public databases of theses, dissertations, technical reports, conference and journal papers. The process of selecting, cleaning and parsing the corpus containing 203 documents is presented by Lopes and Vieira [83].

`Pediatrics (PT)`: The Pediatrics corpus was built by Coulthard [25] and is composed of 281 texts in Portuguese extracted from Jornal de Pediatria[9]. The corpus together with a list of domain terms is freely available at the TEXTECC (Textos Técnicos e Científicos) project website[10].

Table 4.1 presents the statistics of all described corpora, where "`Language`" is the language in which the corpus is written, "`|D|`" is the number of documents in which the corpus is divided, "`|S|`"

[5] http://www.nltk.org/

[6] http://www.euronews.com

[7] http://ir.ohsu.edu/ohsumed/ohsumed.html

[8] http://www.linguateca.pt/cetenfolha/

[9] http://www.jped.com.br

[10] http://www.ufrgs.br/textecc/

is the number of sentences of the corpus, "|W|" is the number of content words and "|V|" is the size of the vocabulary (number of different words) of "|W|". "Type" refers to the nature of the corpus, being "P" for parallel corpus and "C" for comparable corpus. The Europarl collections have just one document because it does not contain document borders, but a sentence aligned document. Thus, the English version of the corpus contains a single document with all sentences of the proceedings, while the Portuguese version contains a single document with all aligned sentences of the English version translated into Portuguese.

| **Language** | **Type** | **Corpus** | **\|D\|** | **\|S\|** | **\|W\|** | **\|V\|** |
|---|---|---|---|---|---|---|
| PT | – | CETEN-Folha | 340,798 | 1,696,496 | 10,658,452 | 123,126 |
| | – | Computer Science | 203 | 128,082 | 3,387,988 | – |
| | P | Europarl | 1 | 1,960,407 | 20,792,400 | 689,593 |
| | C | Geology | 236 | 71,418 | 711,166 | 96,149 |
| | – | Pediatrics | 281 | 27,724 | 835,412 | – |
| | P | TED Talks | 1,112 | 214,395 | 1,379,163 | 43,161 |
| EN | – | Brown corpus | 500 | 51,717 | 30,815 | 84,895 |
| | – | Conference | 27,643 | 354,419 | 4,728,144 | 93,069 |
| | – | Euronews | 48,300 | 235,432 | 2,872,907 | 68,971 |
| | P | Europarl | 1 | 1,960,407 | 22,159,518 | 86,367 |
| | – | FOOTIE | 1,725 | 25,273 | 298,703 | 46,798 |
| | C | Geology | 1,657 | 50,539 | 517,987 | 31,651 |
| | – | Ohsumed | 50,215 | 290,575 | 3,420,215 | 70,192 |
| | P | TED talks | 1,112 | 214,395 | 1,608,041 | 43,019 |

Table 4.1: Statistics about the corpora

In this work we used different sets of corpora for automatic and manual evaluation. The automatic evaluation is performed using the English and Portuguese versions of Europarl and Ted Talks. For the manual evaluation, corpora are separated into two parts: testing corpora and contrasting corpora. Testing corpora are used to generate taxonomic structures for the manual evaluation, and contrasting corpora are used to rank terms of the testing corpora. Testing corpora are composed by both English and Portuguese versions of Europarl parallel corpora and the Geology comparable corpora. In order to obtain relevant terms to each testing corpora (by means of $tf\text{-}dcf$ scores [84] – Section 4.2) the Portuguese versions of the testing set were contrasted with the Pediatrics, Computer Science, TED Talks and CETEN-Folha corpora. The English versions of the testing set were contrasted with the Conference, FOOTIE, TED Talks, Ohsumed, Euronews, and Brown corpora.

### 4.1.2 Gold standard

Assuming that we could find an ideal structure containing all interesting terms and relationships for each domain, the automatic evaluation task would become fairly easy – we would only need to compare the relations extracted or the taxonomic structure with the gold standard, and the quality of the structure would be computed by the overlap between the two. Unfortunately, in practice, such gold standards do not exist. Thus, for the automatic evaluation, we employ the widely accepted, but not perfect, taxonomies as gold standards:

`WordNet (EN)`: Princeton Wordnet [34] is considered the standard model of a lexical ontology for the English language, combining the traditional lexicographic information with modern computation. Words in WordNet are divided into nouns, verbs, adjectives, adverbs and functional words. The basic structure in WordNet is the synset, *i.e.* a set of synonyms that can be used to represent one concept. All synsets are organised in a network of semantic relations (*e.g.*, hyponymy for nouns and troponymy for verbs). According to Tudhope *et al.* [145], WordNet is the most widespread lexical database, containing 147,278 unique terms (nouns, verbs, adjectives, and adverbs) organized into 206,941 synsets[11]. Although WordNet is a widespread lexical database, it does not include much domain-specific terminology. In this work we use the version 3.0 of WordNet which is embedded into Natural Language Toolkit (NLTK)[12]. NLTK provides a WordNet interface containing methods to access synsets, lemmas and relations such as hypernymy and hyponymy.

`Onto.PT (PT)`: Onto.PT is a lexical ontology for Portuguese, structured in a similar fashion to WordNet, but unlike WordNet it was not manually built. As WordNet, it also groups synonymous words in synsets which are lexicalizations of a concept, and semantic relations are held between synsets. In order to build a system capable of automatically creating a semantic knowledge resource from other available resources, Oliveira [107, 109] takes advantage of available NLP tools. Onto.PT[13] integrates the lexical-semantic network PAPEL [108], the electronic dictionaries Dicionário Aberto [132], and Wiktionary.PT[14], and three public synset-based thesauri namely: TeP 2.0 [93], OpenWordNet-PT[15] and OpenThesaurus.PT[16]. In this work we use the version 0.6 of Onto.PT in a WordNet RDF/OWL Basic[17] model which contains 67,873 instances of NomeSynset (noun-based synsets), 26,451 instances of VerboSynset (verb-based synsets), 20,760 instances of AdjectivoSynset (adjective-based synsets) and 2,366 instances of AdverbioSynset (adverb-based synsets). Synsets in the lexical ontology are connected in a network containing 341,506 relations in which 79,425 belong ot the type `hiperonimoDe` (hypernym-of). Other relations found in Onto.PT are `parteDe` (part-of), `membroDe` (member-of), `temQualidade` (has-quality) *etc.*.

In the lack of an interface for the Onto.PT RDF/OWL file, we identified all synsets marked in RDF as `NomeSynset` meaning that all our taxonomic relations occur only between nouns. For each identified synset we extracted its id, its lemmas (marked in RDF as `formaLexical`) and its taxonomic relations (marked in RDF as `hiperonimoDe` and `hiponimoDe`). A directed graph (digraph) is created where each synset is a node and each taxonomic relation is an edge that connects two nodes. The digraph allows an easier access to hypernyms in higher levels such as the hypernym of a hypernym. An excerpt of the generated structure can be seen in Figure 4.2.

[11] http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[12] http://www.nltk.org/

[13] http://ontopt.dei.uc.pt/

[14] http://pt.wiktionary.org/

[15] https://github.com/arademaker/wordnet-br

[16] http://openthesaurus.caixamagica.pt/

[17] http://www.w3.org/2006/03/wn/wn20/

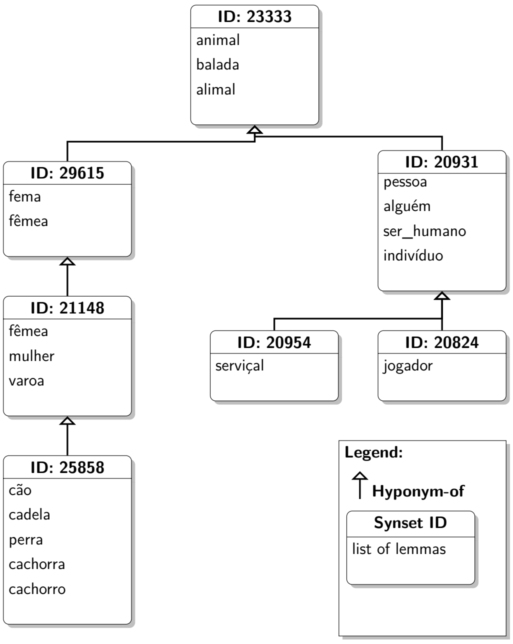


Figure 4.2: Excerpt of the extracted structure from Onto.PT.

As Onto.PT is constructed grouping resources automatically, we can observe that there are some errors such as the word "fêmea" (female) which repeats in two synsets (ID: 21615 and ID: 21148). In these cases we eliminate self-cycles, *i.e.*, a term that is hypernym of itself. As this resource is automatically generated, it also may contain errors, such as the word "mulher" (womam) ID: 21148 which should not an hypernym of "cão" (dog).

## 4.2 Preprocessing

### 4.2.1 Parsing

Term extraction is an important part of taxonomy extraction and ontology construction [87], since it extracts relevant terms of the domain. Usually term extraction requires a pre-processing in order to identify Part-of-Speech (PoS) tags or relations between words in a syntactic level (chunking). In a syntactic tree (Figure 4.3), individual words are leaves of the tree, and the internal nodes are tags that determine the grammatical function of the words or the relations between words. In this work we use the trees generated by syntactic parsers to extract noun-phrases.

The English corpora were parsed using the Stanford Lexicalized Parser [65] (version 3.3.1), a well known parser and widely used in relation extraction [79, 97, 161]. The options were set to "wordsAndTags,penn,typedDependencies", where wordsAndTags generates the part-of-speech for each word in the text, penn generates a context-free phrase structure grammar representation and typedDependencies generates a typed dependency representation. For example, in the parsed phrase "Parliament rejected the request from the president." presented below, WordAndTags is represented by the first line with the words and PoS tags, penn is represented by the syntactic tree where we can identify noun phrases marked as NP, and typedDependencies is represented by

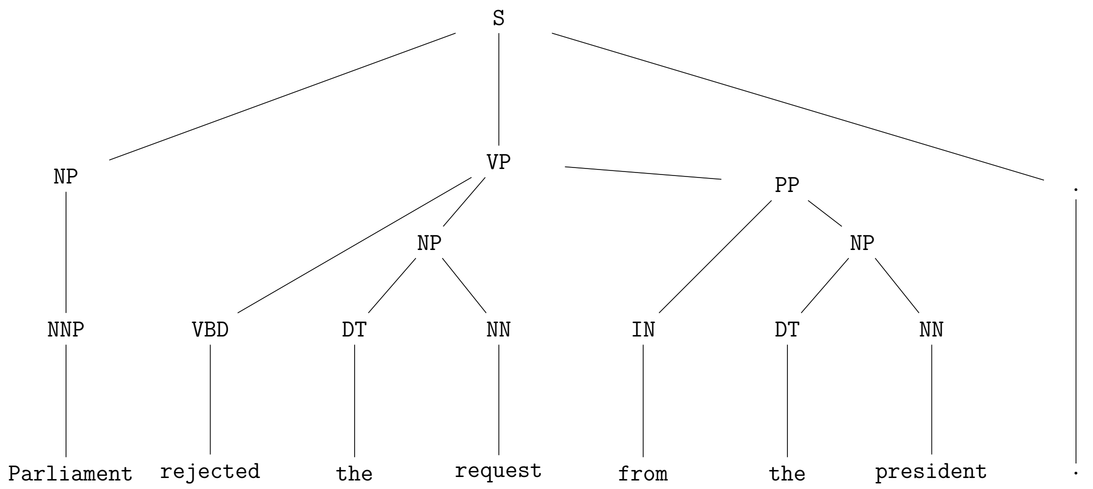


Figure 4.3: Syntactic tree for the phrase "Parliament rejected the request from the president.".

dependencies where we can identify relations such as the subject of a verb `nsubj` or its direct object `dobj`. Further details about the parser and an online version are available in the site of the tool[18].

```
Parliament/NNP rejected/VBD the/DT request/NN from/IN the/DT president/NN ./.

(ROOT
  (S
    (NP (NNP Parliament))
    (VP (VBD rejected)
      (NP (DT the) (NN request))
      (PP (IN from)
        (NP (DT the) (NN president))))
    (. .)))

nsubj(rejected-2, Parliament-1)
root(ROOT-0, rejected-2)
det(request-4, the-3)
dobj(rejected-2, request-4)
prep(rejected-2, from-5)
det(president-7, the-6)
pobj(from-5, president-7)
```

For the Portuguese corpora we applied the PALAVRAS [9] parser, which has been used in many works [4, 8, 87, 142], and one of the most adopted for Portuguese language parsing. For each phrase in the input of the parser, it generates a syntactic tree representation containing terminal nodes (leaves of the tree) and non-terminal nodes containing the grammatical structure of the phrase. Figure 4.4 presents the tree generated to the phrase "O parlamento rejeita o pedido do presidente.". In this example we can see the identification of functions of words such as the subject of the sentence marked with the `S` tag, the direct object marked with the `Od` tag *etc.*, as well as the forms of the

[18] http://nlp.stanford.edu:8080/parser/index.jsp

words as n for nouns, prp for prepositions and so on. Noun phrases and prepositional phrases are marked visually as g. In Figure 4.4, the node marked as subject (S) and group (g) is selected showing the type of group above the syntactic tree (line SUBJECT group (np)).

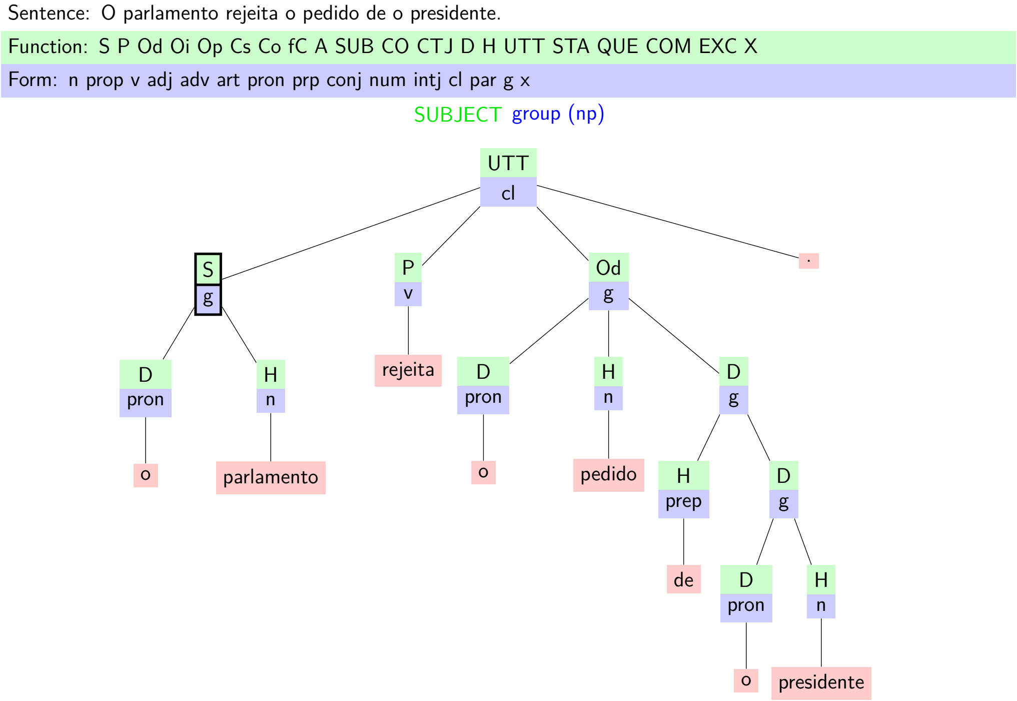

Sentence: O parlamento rejeita o pedido de o presidente.
Function: S P Od Oi Op Cs Co fC A SUB CO CTJ D H UTT STA QUE COM EXC X
Form: n prop v adj adv art pron prp conj num intj cl par g x
SUBJECT group (np)
UTT
cl
S
g
P
v
Od
g
.
D
pron
H
n
rejeita
D
pron
H
n
D
g
o
parlamento
o
pedido
H
prep
D
g
de
D
pron
H
n
o
presidente

Figure 4.4: Tree structure generated by PALAVRAS parser to the phrase "O parlamento rejeita o pedido do presidente."

The PALAVRAS output has the format of a Tiger XML [98] (a XML-based file), presenting morphological information such as word class and inflection, syntactic information such as subject and the objects of a phrase, and semantic information for some nouns, proper names, verbs and adjectives, such as <Hprof> (professional) or <Hnat> (Nationality human) in a Human prototype class. Although the parser generates semantic labels to some words, we do not use them in this work. Further details about the parser and an online version are available in the site of the tool [19].

Although we used Stanford parser for English and PALAVRAS parser for Portuguese, it would also be possible to use other parsers or even a PoS tagger and a chunker with a finite-state machine as presented in Figure 4.3 to extract noun phrases.

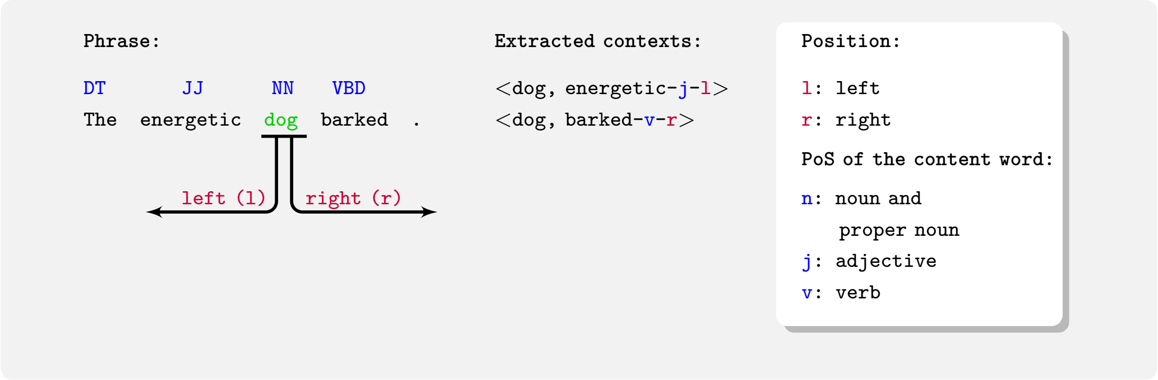
### 4.2.2 Context extraction

Context extraction is used in this thesis to select terms and contexts for the automatic evaluation process. As our evaluation is performed using WordNet [34] and Onto.PT [107, 109] as gold standards, all models for the automatic evaluation use single nouns instead of noun phrases, since the number of noun phrases in both resources is very limited. Also, WordNet cannot distinguish that "small dog" is a "dog" modified by an adjective, and thus "small dog" is a "dog" or that "small

[19] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php

dog" is a kind of "animal", as well as Onto.PT cannot understand that "cachorro pequeno" is the noun "cachorro" modified by an adjective.

The first step of the context extraction collects all co-occurrences between a target word (a noun or a proper noun in the corpus) and content words, *i.e.*, nouns, proper nouns, verbs and adjectives. The co-occurrences are extracted by sliding a window with a pre-defined length $n$ along each phrase. The size of the window was set to 5, *i.e.*, every content word that occurs in the same sentence and within a window of two words before or after is counted as a co-occurrence for the target word. Each context is marked with its relative position in relation to the target word. For example, consider the phrase "The energetic dog barked." and the relations defined as <target word, context> in the extraction below.



The contexts <dog, energetic-j-l> and <dog, barked-v-r> are extracted, where the two characters at the end of the context means that the term contains determined Part-of-Speech (PoS) tag and is placed at determined position of the target noun. Thus, v-r in <dog, barked-v-r> means that the verb "barked" is placed on the right of the target noun "dog". The outcome of the this process is a list containing nouns and their contexts followed by the frequency of these contexts.

For models that use documents as contexts intead of the window of size=5, the process of extracting contexts is different. For these models, all nouns and pronouns are extracted and their context is determined as the name of the file where they were extracted from. Thus, consider a file named "doc_1.txt" containing the words "dog" and "cat" and another file named "doc_2.txt" containing the words "dog" and "fish". The context of each word is: <dog, doc_1.txt>, <cat, doc_1.txt>, <dog, doc_2.txt> and <fish, doc_2.txt>. The outcome of the this process is a list containing nouns and their contexts followed by the frequency of the word in each contexts.

As generating models containing all terms is a time and resource consuming task, we decided to reduce our list of target words. To reduce our list we took into account only target words that appear into the gold standard and share the highest number of contexts. Thus, for each target word existing in the corpus and in the gold standard we counted the number of contexts and selected the top $n$ target words. The evaluation was performed using $n$=1,000 and $n$=10,000.

### 4.2.3 Concept extraction

Concept extraction is used in this thesis to select terms for the manual evaluation process. Selecting terms that are relevant to the domain is important because the manual evaluation is performed using domain corpora. As manual evaluation is a time consuming task and we could not evaluate all terms generated by all methods, we have to select some of these terms to be assessed by domain specialists. In order to identify the concept candidates we used the methodology developed by Lopes [84] which is divided into three phases, namely: NP extraction, heuristics application and term weighting. Consider as input of the process a parsed corpus as explained in Section 4.2.1. The process of concept extraction starts by extracting all noun phrases (NPs) from the parsed files, since NPs are well known for containing conceptual information [71]. In the next step heuristics are applied on each noun phrase in order to refine the set of extracted NPs. Finally, a measure to weight the relevance of each NP is applied, generating a list of concept candidates sorted by their relevance to the domain. Each step of the process is explained as follows:

- NP extraction

Considering English files parsed by Stanford parser as input, each syntactic tree is read and all terms inside a NP tag are extracted[20]. For example, the parser generates the following syntactic structure for the phrase “The man would firstly like to point out Mr Poettering’s lack of logic.”:

```
(ROOT
  (S
    (NP (DT The) (NN man))
    (VP (MD would)
      (ADVP (RB firstly))
      (VP (VB like)
        (S
          (VP (TO to)
            (VP (VB point)
              (PRT (RP out))
              (NP
                (NP
                  (NP (NNP Mr) (NNP Poettering) (POS ’s))
                  (NN lack))
                (PP (IN of)
                  (NP (NN logic))))))))
    (. .)))
```

Performing the extraction of all content within NP tags generates the following noun phrases from the English corpus:

[20] Stanford tag set containing all tags used in this work are presented in Appendix A

```
(NP (DT The) (NN man))
(NP (NNP Mr) (NNP Poettering) (POS 's))
(NP (NNP Mr) (NNP Poettering) (POS 's) (NN lack))
(NP (NNP Mr) (NNP Poettering) (POS 's) (NN lack) (IN of) (NN logic))
(NP (NN logic))
```

From a Portuguese corpus containing Tiger XML parsed files by PALAVRAS, we extract all content within tags marked as np (inside group g). For example, consider the phrase "Primeiramente o homem gostaria de realçar a ausência de lógica do senhor deputado Poettering." (translation of the phrase: "The man would firstly like to point out Mr Poettering's lack of logic.") and its parsed result represented in Figure 4.5 as a syntactic tree.

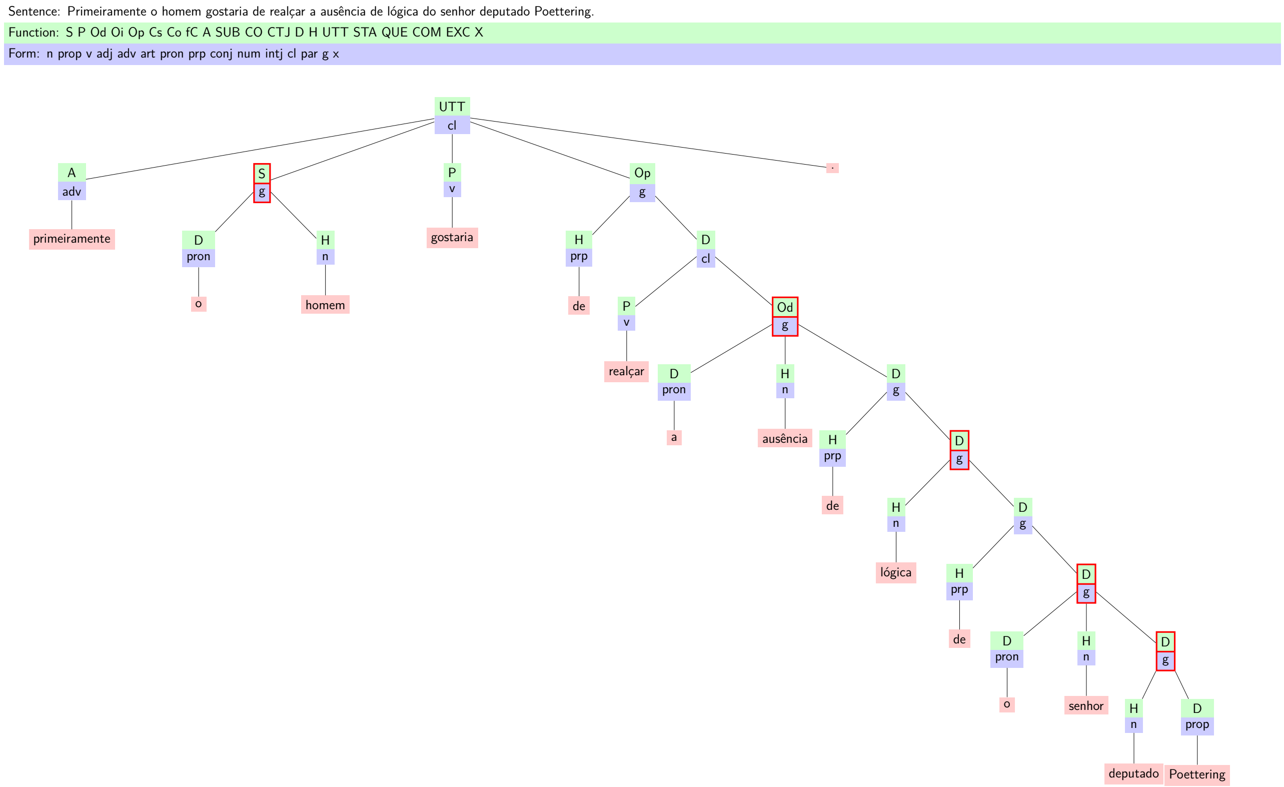

Sentence: Primeiramente o homem gostaria de realçar a ausência de lógica do senhor deputado Poettering.
Function: S P Od Oi Op Cs Co fC A SUB CO CTJ D H UTT STA QUE COM EXC X
Form: n prop v adj adv art pron prp conj num intj cl par g x
UTT
cl
A
adv
primeiramente
S
g
D
pron
o
H
n
homem
P
v
gostaria
Op
g
H
prp
de
D
cl
P
v
realçar
Od
g
D
pron
a
H
n
ausência
D
g
H
prp
de
D
g
H
n
lógica
D
g
H
prp
de
D
g
D
pron
o
H
n
senhor
D
g
H
n
deputado
D
prop
Poettering
.

Figure 4.5: Tree structure generated by PALAVRAS parser to the phrase "Primeiramente o homem gostaria de realçar a ausência de lógica do senhor deputado Poettering."

From this parsing tree, all multi-token terms tagged in group g as np are extracted as well as single tokens when they play a role of subject (tagged as S), object (tagged as Od, Oi or Op) or their complements (tagged as Cs or Co)[21]. The extraction of single tokens is also performed due to the fact that the parser does not tag them as np when they play the role of subjects, objects or complements. Thus, the following terms are the result of the extraction of NPs (highlighted in red in Figure 4.5), where each word is followed by its PoS tag.

[21] The tag set of the parser is presented in Appendix B

```
np: o/det homem/n
np: a/det ausência/n de/prp lógica/n de/prp o/det senhor/n deputado/n Poettering/prop
np: lógica/n de/prp o/det senhor/n deputado/n Poettering/prop
np: o/det senhor/n deputado/n Poettering/prop
np: deputado/n Poettering/prop
```

- Heuristics application

After extracting all NPs from the corpus, a refining process is performed where heuristics are applied on each NP. These heuristics were adapted from the ones applied in Portuguese by Lopes [84, 87] and separated into three groups: Heuristics for adjustment (HA), Heuristics for discarding (HD) and Heuristics for inclusion (HI). Heuristics for adjustment aim at removing elements from the NP that do not carry any information, such as determiners, pronouns, adverbs and possessive marks. Phrases below show examples of how NPs become after applying the heuristic for removing determiners (DT) and predeterminers (PDT).

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (DT The) (NN man)) | (NP (NN man)) |
| 2. | (NP (DT the) (NNP Member) (NNPS States)) | (NP (NNP Member) (NNPS States)) |
| 3. | (NP (PDT all) (DT the) (JJ technical) (NNS measures)) | (NP (JJ technical) (NNS measures)) |
| 4. | (NP (PDT all) (DT these) (NNS agreements)) | (NP (NNS agreements)) |
| 5. | (NP (PDT such) (DT the) (JJ European) (NNS institutions)) | (NP (JJ European) (NNS institutions)) |

Heuristics for discard intend to remove NPs when they are not relevant to the domain, *i.e.*, NPs that are composed by non-informative terms such as symbols, numbers *etc.*. Unlike the heuristics for adjustment, the heuristic for discard remove the entire NP instead of some of its terms. The example below shows the application of the heuristic to remove empty NPs or NPs without nouns. This heuristic is applied after all heuristics for adjustment, where a verification of the consistence of the NPs is required. This verification intends to check whether the NP still contains any noun.

| # | Original NP: | After applying the heuristic | Reason to remove |
|---|---|---|---|
| 1. | (NP (PRP we) (MD may) (VB correct) (PRP them)) | (NP (MD may) (VB correct)) | Non-informative |
| 2. | (NP (RB even) (JJR worse)) | (NP (JJR worse)) | Non-informative |
| 3. | (NP (PRP us) (DT both)) | (NP ) | Empty NP |
| 4. | (NP (PRP We)) | (NP ) | Empty NP |

Heuristics for inclusion intend to detect inner NPs, *i.e.*, NPs that are implicit inside a greater NP. The example below shows a heuristic to create NPs for adjectives and nouns when they are linked by a conjunction into a single NP.

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (JJ technical) (CC and)<br>(JJ industrial) (NNS developments)) | (NP (JJ technical) (NNS developments))<br>(NP (JJ industrial) (NNS developments)) |
| 2. | (NP (JJ national) (, ,) (JJ regional)<br>(CC and) (JJ local) (NNS obstacles)) | (NP (JJ national) (NNS obstacles))<br>(NP (JJ regional) (NNS obstacles))<br>(NP (JJ local) (NNS obstacles)) |
| 3. | (NP (JJ relevant) (JJ American) (, ,)<br>(JJ Canadian) (CC and)<br>(JJ Japanese) (NNS authorities)) | (NP (JJ relevant) (JJ American) (NNS authorities))<br>(NP (JJ Canadian) (NNS authorities))<br>(NP (JJ Japanese) (NNS authorities)) |

The result of heuristics application is a list containing all valid NPs from the corpus. A valid NP is a resultant NP after applying all heuristics. A detailed description of each heuristic with examples of its application is presented in Appendix C.

- Term weighting

The step of term weighting consists in sorting the extracted NPs according to their relevance to the domain. Although traditional approaches use the frequency of the term in a corpus [111,124,127] in order to decide its relevance, in this work we used an approach that uses contrastive corpora, *i.e.*, uses corpora from different domains in order to give the relevance for terms in the target domain [63, 85, 113, 157].

The weighting scheme used in this work was proposed by Lopes [84] and intends to estimate the term relevance to a domain following the idea of contrasting corpora. It is called Term Frequency - Disjoint Corpora Frequency ($tf\text{-}dcf$) and considers the absolute frequency ($tf$) as the primary indication of term relevance, while the disjoint corpora frequency ($dcf$) intends to penalize terms that appear in contrasting corpora proportionally to its number of occurrences. $dcf$ penalizes terms that appear in the contrasting corpora by dividing its absolute frequency in the domain corpus by a geometric composition of its absolute frequency in each of the contrasting corpora. The $tfp\text{-}dcf$ index is mathematically expressed in Equation 4.1, where $t$ is a term in corpus $c$ and $\mathcal{G}$ is a set of contrasting corpora.

$$tf\text{-}dcf_t^{(c)} = \frac{tf_t^{(c)}}{\prod\limits_{\forall g \in \mathcal{G}} 1 + log(1 + tf_t^{(g)})} \tag{4.1}$$

This equation preserves an intuitive comprehension since $tf\text{-}dcf$ will be equal to $tf$ if the term does not appear in the contrastive corpora (the term is not be penalized at all), or smaller than $tf$ if the term appears in the contrastive corpora. If the term appears in many corpora is more likely to be irrelevant to the domain corpus. As the frequency of the terms are distributed according to a Zipf law [168], the logarithm is applied on the frequency of each term of the contrastive corpora in order to correctly estimate its importance. Finally, the product of the occurrences is applied decreasing geometrically the importance of the term when it appears in other corpora. Thus, a term tends to be more relevant to the domain when it appears many times in the target corpus and few times in constrastive corpora.

We decided to use $tf\text{-}dcf$ because it obtained better results when comparing with other metrics that use contrastive corpora ($tds$ [113], $tf\text{-}idf$ [159], $thd$ [63] and $TF\text{-}IDF$ [62]) in a domain concept ranking task. This metric is also chosen because it is incorporated into E$\chi$ATO$_{lp}$ [86] which was used in order to extract terms for Portuguese.

## 4.3 Relation extraction methods

### 4.3.1 Models for automatic evaluation

For the automatic evaluation we developed seven methods for English and Portuguese languages. The only method that cannot use the same implementation in both languages is the pattern-based method (Patt) because it is a language-oriented model. It means that the method has one implementation for English and another one for Portuguese. The other methods use the same implementation for both languages, changing only the process of parsing (see Section 4.2.1). After parsing, lists are created containing nouns and their contexts (see Section 4.2.2). Having these lists, we built the following models:

**DSim**: The model based on Directional Similarity takes into account the Distributional Inclusion Hypothesis, according to which the contexts of a narrow term are also shared by the broad term (see Section 2.1.4). This model uses the list of terms and contexts extracted using a window of size=5. The degree of association between terms and contexts is determined by a weight function. Thus, the value of the frequency of a term with a context is replaced by its Positive Pointwise Mutual Information (PPMI) [20] value, where all negative values are set to zero. For computing the directional similarity we tested the measure proposed by Weeds *et al.* [156] (Equation 2.6) and the measure proposed by Clarke [24] (Equation 2.9, hereafter ClarkeDE). These measures can identify taxonomic relations between terms using the notion of precision and recall of a term (see 2.1.4) instead of defining a threshold. As the results using both measures were almost the same in almost all corpora, we decided to use in the evaluation process only the values generated by ClarkeDE measure. The code implementing these measures are freely available by Weeds[22] [154].

**SLQS**: The model based on entropy was developed by Santus *et al.* [130] and relies on the idea that superordinate terms are less informative than their hyponyms. This model also uses the list of terms and contexts extracted using a window of size=5. The difference when compared with DSim model is that SLQS model employs Local Mutual Information (LMI) to weight co-occurrences as well as uses entropy as an estimate of context informativeness. After extracting co-occurrences and weighting them using LMI, the $N$ most associated contexts are identified using the Shannon entropy measure [131], where $N$ is set to 50. The resulting values of entropy are normalized using the Min-Max-Scaling in a range 0–1. Finally, the entropy of a word is defined as the median entropy of its $N$ contexts as presented in equation 2.14.

[22] https://github.com/SussexCompSem/learninghypernyms

`TF`: The model based on the frequency takes into account the number of times a word occur in the whole collection as an indicative of generalization-specialization. The idea in this model is that the more general a word, the higher its frequency. This model uses the list of terms extracted using documents as contexts. For each word, the resulting frequency is the sum of all individual frequencies in documents.

`DocSub`: The model based on document subsumption uses the probability of the distribution of the words across shared documents in order to identify a taxonomic relation between them. According to this model, a word that appears in more documents tends to be more general than a word that appears in a subset of these documents. It is important to note that in this model a word subsumes another word only if it appears in a subset of the documents that the other word appears, as presented in equation 2.17. A threshold indicating the percentage of shared documents may be set. In the evaluation process we vary this threshold to see the effect in results. This model uses the list of terms extracted using documents as contexts, where we can verify the intersection of documents shared by two words.

`DF`: The model based on the document frequency takes into account the number of documents in which a word appears as an evidence of taxonomic relation. Thus, a word that occurs in more documents tends to be more general than a word that appears in few documents. This model is different from `DocSub` because it takes into account only the number of documents in which a word appears and not the number of shared documents. This model uses the list of terms extracted using documents as contexts, where the frequency is represented by the number of documents in which a word occurs.

`HClust`: The model based on hierarchical clustering uses contexts to group similar words together and the document frequency to identify the taxonomic relations between them. This model uses both lists of terms extracted using a window of size=5 and documents as contexts. The first list is used to hierarchically cluster similar words together. Similar words are words that share similar contexts. The second list is used to identify the taxonomic relation between terms based on the document frequency as it occurs in `DF` model. The difference of this model when compared with `DF` is that the former refines the latter verifying whether the taxonomy exists only for semantically related terms. The botton up clustering, grouping the most similar terms together is performed using the freely available scripts of Fastcluster[23].

`Patt`: The model based on patterns extracts taxonomic relations between words using rules developed by Hearst [52, 53] and its adapted rules for Portuguese [7]. This model is the unique language dependent and different rules run on English and Portuguese. For automatic evaluation we performed two approaches to extract nouns. The first one uses patterns to extract relations between NPs. The comparison against the gold standard uses the head of the NP. The second approach uses patterns to extract taxonomic relations between nouns. Thus, adjectives or nouns that belong to a multi-word term are discarded.

[23] http://danifold.net/fastcluster.html

### 4.3.2 Models for manual evaluation

As manual evaluation is a laborious and time consuming task, we decided to limit the number of methods for the manual evaluation. Also, models that use distributional inclusion hypothesis are well known for performing better in hyponymy detection, *i.e.*, given a pair of words, determine whether one word is hyponym of the other [6, 67] and poorly in hyponym acquisition, *i.e.*, extract all possible hyponyms given a single word as input, or hyponym generation, *i.e.*, return a list of all possible hyponyms given only a single word as input [122]. Thus, we generated results for 4 methods: Patterns (`Patt`), Head-modifier (`HMod`), Document subsumption (`DocSub`) and Hierarchical Clustering (`HClust`).

Before applying each method, a list containing the domain terms for each corpus was generated. This list is built extracting terms (all noun phrases of the corpus filtered by heuristics) that are ranked according to their relevance to the domain. In this work, $tf$-$dcf$ score [85] is used as the measure of domain relevance. A threshold on the $tf$-$dcf$ scores is applied, selecting the domain concepts. As demonstrated by Lopes and Vieira [88], the first 15% of the best $tf$-$dcf$ scores are considered a good sample of domain concepts. Thus, our list of domain concepts contains the best of the 15% terms ranked by $tf$-$dcf$ score. For statistical methods we also apply a threshold on the term frequency ($tf$), since terms that occur few times or have few contexts are not significant and may induce noise in the process. In this work, terms occurring less than 5 times ($tf < 5$) are filtered out of the process. The methods tested in this work are as follows:

**`Patt`**: Although there are works [64, 102, 144] presenting a list of patterns greater than the one used by Hearst [52, 53], we decided to use the same patterns presented by Hearst as presented in Table 2.1 because they proved to have good results [135]. We kept in the manual evaluation the same patterns used in the automatic evaluation. For the Portuguese corpora we applied the same patterns developed by Basegio [7] (as presented in Table 2.1).

**`HMod`**: The model based on the head-modifier explores the idea that a noun phrase may contain inner noun phrases when extracting its head and modifiers, as explained in Section 2.1.2. This model uses the list of noun phrases generated by the concept extraction (see Section 4.2.3). For each term in this list we split the n-gram and search for inner terms. An inner term is composed by the head of the noun phrase or by the head of the noun phrase plus its modifiers. The taxonomic relation between two terms is generated when two terms have the same head but at least one modifier less than the other. The term containing less words is the hypernym of the relation.

**`DocSub`**: The document subsumption model (or coocurrence model) is almost equal to the one developed for the automatic evaluation, differing only by the fact that the model for the manual evaluation contains noun phrases instead of single nouns. These noun phrases are terms that belong to the domain and were generated by the concept extraction step (see Section 4.2.3). The threshold (parameter $\lambda$ in Equation 2.17) was set to 0.8 as performed by Sanderson and Croft [128].

**`HClust`**: The hierarchical clustering model also starts by using the list of concept terms in order to filter out terms not related to the domain. For each term of this list we extracted the content words in the corpus as contexts, using a window equal to 5 (as performed in model DSim for the automatic evaluation, with the difference that in models for manual evaluation we use noun phrases instead of single nouns). Thus, for each term in the list of concepts a vector of contexts is created, where the cell value contains the Pointwise Mutual Information between the concept and the context. These vectors serve as input to an agglomerative clustering, where terms that occur in similar contexts are clustered together. For each new generated cluster, we verify which term is the hypernym to set it as the head of the cluster (as explained in Section 2.1.5). It is important to note that when the document frequency is not different between terms in a cluster, the head of the cluster is not identified. For defining the number of clusters to generate the taxonomies we used the list containing domain terms. A clustering process must stop when at least 10 noun phrases (corresponding to the number of hierarchies to be evaluated) from the top 100 domain terms are included into the clusters.

## 4.4 Evaluations

### 4.4.1 Automatic evaluation

This evaluation intends to automatically compare the relations extracted from each model with the entries in the gold standards. A gold standard structure is a taxonomy that serves as reference. Thus, the quality of the extracted relations are expressed by their similarity to the gold standard hierarchy. In this thesis we use WordNet [34] as gold standard for English and Onto.PT [109] as gold standard for Portuguese.

The quality of the extracted relations may be measured in terms of Precision, Recall and F-measure when comparing with the gold standard. In order to achieve such measures, we first define the common relations (CR) as the relations between a term and its super- and sub-terms that appear in both the extracted taxonomy and the gold standard:

$$CR(c, O_1, O_2) = \{c_i \in C_1 \cap C_2 | rel\} \qquad rel = \begin{cases} (c_i, c) & \text{if } c_i <_{c_1} c \\ (c, c_i) & \text{if } c <_{c_1} c_i \end{cases} \tag{4.2}$$

where $O_1$ and $O_2$ are two taxonomies, $c$ is the term being analyzed, $c_i$ is a term common to both taxonomies, $C_1$ is the set of terms in $O_1$, $C_2$ is the set of terms in $O_2$, and $<_{c_1}$ is the partial order induced by the relationship in $O_1$. Take for instance the taxonomies in Figure 4.6, assuming that the left taxonomy ($O_1$) was extracted by some method and the taxonomy on the right ($O_2$) is the gold standard, the common relations for the term car in the taxonomy $O_1$ are {(vehicle, car), (car, cab), (car, tram)}. As CR only takes into account terms shared by both taxonomies, the set of relations for car also contains the relation (car, tram) even if this relation does not exist in the gold standard. It is also important to note that even if the gold standard does not contain the direct relation between car and vehicle, this relation is inherited by transitivity.

Using the common relations (CR), we can define precision ($\mathcal{P}$) and recall ($\mathcal{R}$) as:

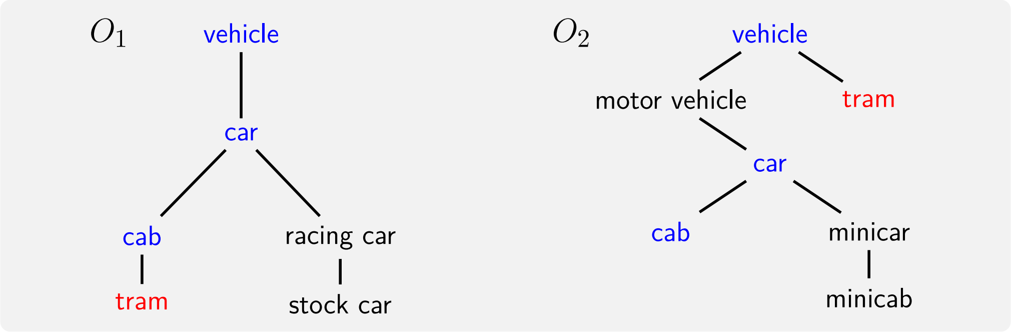


Figure 4.6: Taxonomies for common semantic cotopy example.

$$\mathcal{P} = \frac{\sum_{c \in C_T \cap C_{GS}} |CR(c, O_T, O_{GS}) \cap CR(c, O_{GS}, O_T)|}{\sum_{c \in C_T \cap C_{GS}} |CR(c, O_T, O_{GS})|} \tag{4.3}$$

$$\mathcal{R} = \frac{\sum_{c \in C_T \cap C_{GS}} |CR(c, O_T, O_{GS}) \cap CR(c, O_{GS}, O_T)|}{\sum_{c \in C_T \cap C_{GS}} |CR(c, O_{GS}, O_T)|} \tag{4.4}$$

where $c$ is the term being analyzed, $C_T$ is the set of terms from the generated taxonomy, $C_{GS}$ is the set of terms from the gold standard, $O_T$ is the generated taxonomy, and $O_{GS}$ is the gold standard. Thus, the precision score is the result of dividing the amount of relations suggested by the system and contained in the gold standard by the total amount of relations suggested. The recall score is the result of dividing the amount of relations suggested by the method and contained in the gold standard by the total amount of relations existing in the gold standard taxonomy. The F-measure score can be interpreted as a weighted average of the values corresponding to the two parameters precision and recall:

$$\mathcal{F} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{4.5}$$

### 4.4.2 Manual evaluation

As gold standards are limited to a real evaluation of the results, we decided to perform also a manual evaluation with human experts. A manual evaluation usually demands domain specialists in order to assess terms and relations. These terms compose a subset of all domain terms and are selected according to the relevance to the domain. This type of evaluation measures only the precision of a method. Thus, the purpose of this procedure was to obtain judgements on whether a term is taxonomic related to another. *Europarl* and *Geology* corpora were chosen because *Europarl* has a parallel corpora, allowing us to identify changes using the same method in different languages, and *Geology* because it is a domain comparable corpora, having domain specialists to assess the results.

For each domain we selected 3 evaluators. In case of a tie between two evaluators, the third one would decide the analysis. Terms in the geology domain were assessed by experts with degree in Geology or Oceanography. Terms generated using the Europarl corpus were assessed by undergraduate students of the International Relations course. All evaluators are Portuguese natives and English proficient.

The evaluation process consists in assessing triples by domain experts, considering only a taxonomic relation (hypernymy - hyponym), *i.e.*, "is a" or "is a kind of" relation. A triple is composed as $<w_1,\ r,\ w_2>$, where $w_1$ and $w_2$ are two terms and $r$ is the relation between both terms. The whole taxonomy is not evaluated directly by the experts, but the relations between pairs of terms. As result, the values of Precision for each method is computed based on the number of correct relations. For each evaluator a list containing pairs of terms and their relation is presented for judgement[24].

The first step to generate the list of relations is the selection of the terms to be assessed (hereafter named seeds). As not all methods can generate taxonomies for the more relevant terms of the domain (defined by their $tf\text{-}dcf$ score, as presented in Section 4.2.3), we decided to select the top 10 most relevant terms generated by each method and use them as seeds to extract relations. For example, the most relevant term for the corpus in the geology domain in Portuguese is "arenito". The head-modifier (Hmod model) and the document subsumption (DocSub model) generated taxonomies for this term. On the other hand, there was no pattern (Patt model) in the corpus that could be extracted with this term or the term was not much similar to others to belong to a cluster (HClust model). Hence, this term is evaluated only on the taxonomies generated by Hmod and DocSub models.

For each seed, we extracted all relations that are directly or indirectly connected to the branches of the generated taxonomy, *i.e.*, all hypernyms and hyponyms related to the seed in the taxonomy. Elements that are in the same taxonomy but not in the same branch of the seed are discarded. For instance, using the Figure 4.7, consider the taxonomy for the root "formação de solos" and the term "carbonatos" as seed. As we are not evaluating co-hyponyms, *i.e.*, terms that share the same hypernym, the terms "apatita" and "cimentos" are discarded, as well as the co-hyponyms of "minerais" ("decomposição química" and "desintegração física") that are not directly related to the hierarchy of the seed. Thus, the relations (<hypernym, hyponym>) <minerais, carbonatos> and <formação de solos, minerais> are extracted.

As statistical method tends to generate a large number of relations for a seed (*e.g.*, the tree generated to the seed "matéria orgânica" contains 1,813 relations), we decided to evaluate a subset of all these relations. Thus, we limited to the 50 closest relations for each seed in each method. Hence, the number of relations evaluated per method is up to 500.

### 4.4.3 Metrics for characterizing taxonomies

According to Vrandecic and Sure [151], measuring ontologies is necessary to evaluate them both during engineering and application, being a necessary precondition to perform quality assurance and control the process of improvement. Metrics allow the fast and simple assessment of an ontology

[24] All instructions given for evaluators are included in Appendix D

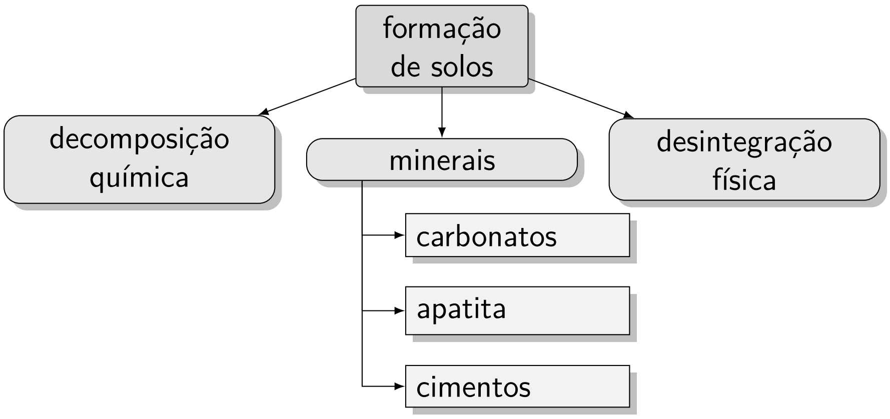


Figure 4.7: Tree generated for the root "formação de solos"

and also to track their subsequent evolution. Indeed, they are expected to give some insight for ontology developers to help them to design ontologies, improve ontology quality, anticipate and reduce future maintenance requirements, as well as help ontology users to choose the ontologies that best meet their needs [165]. A compilation of metrics for ontology evaluation is performed by Freitas [38, 39].

Although these metrics are applied to ontologies, they also can be applied to other structures such as taxonomies. In this thesis, besides interpreting methods in terms of automatic and manual evaluations, we also intend to describe methods in terms of the structure of the generated taxonomy. Thus, we analyse the subset of the metrics that are applied to taxonomies or Rooted Directed Acyclic Graphs (Rooted DAG), that is, a structure having a single highest node (Root) and all other nodes are connected by means of `is-a` links, generating a chain of links to the Root.

In order to better understand the metrics, consider that each applied method may generate one or more taxonomies, and each taxonomy has a root term and at least one leaf term. Examples presented in some metrics use Figure 4.8 where two taxonomies are represented as directed graphs (digraphs). Nodes represent words being identified by their ids, and edges represent the relation "is hypernym of". Thus, the connection between nodes "`ID: 1`" and "`ID: 2`" means that the word identified by "`ID: 1`" is hypernym of the word identified by "`ID: 2`".

A set containing the main metrics are described below, where "`terms`" refers to a set of terms, "$\mathtt{term}_{\mathtt{ij}}$" refers to the $i$th term in the taxonomy $j$ and "$\mathtt{term}_{\mathtt{ij}}$" $\in$ "`terms`", "`Count()`" is a function that counts the number of occurrences, "`Max()`" determines the maximum value in a set, "`Min()`" determines the minimum value in a set, "`isRoot()`" and "`isLeaf()`" verify whether the argument is a root or a leaf term in the taxonomy, "`Depth()`" returns the depth of a term, "`Width()`" returns the number of siblings of a term, and "`hasSiblings()`" returns terms that have siblings. The list contained all metrics is presented in Appendix E.

**TotalTerms**: Total number of terms takes into account all unique terms generated by all taxonomies

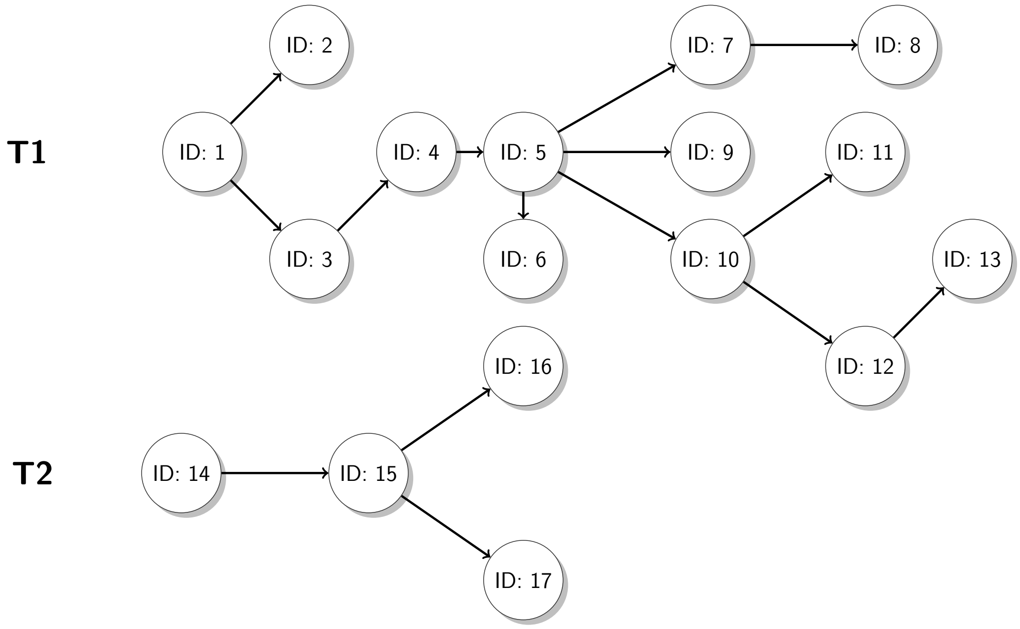


Figure 4.8: Examples of taxonomies represented as direct graphs

using a specific method. Example: The total number of terms in Figure 4.8 is: 17 (ID: 1 to ID: 17).

$$\texttt{TotalTerms} = Count(\texttt{terms})$$

**TotalRoots**: Total number of roots indicates the number of upper terms of all taxonomies, *i.e.*, terms without hypernyms in a taxonomy. This means also the number of taxonomies generated by the method. Example: The total number of roots in Figure 4.8 is: 2 (ID: 1 and ID: 14).

$$\texttt{RootTerms} = Count(isRoot(\texttt{term}_{\texttt{ij}}))$$

**MaxDepth**: Maximum depth extracts the longest path between a root and a leaf for each taxonomy and selects the maximum value. Example: The maximum depth in Figure 4.8 is: 6 (passing by IDs: 1, 3, 4, 5, 10, 12 and 13 in T1)

$$\texttt{MaxDepth} = Max(Depth(isLeaf(\texttt{term}_{\texttt{ij}})))$$

**MinDepth**: Minimum depth extracts the shortest path between a root and a leaf for each taxonomy and selects the minimum value. Example: The minimum depth in Figure 4.8 is: 1 (path between ID:1 and ID:2 in T1)

$$\texttt{MinDepth} = Min(Depth(isLeaf(\texttt{term}_{\texttt{ij}})))$$

**AvgDepth**: Average depth is the ratio between the sum of all depths and the total number of taxonomies. Example: The average depth in Figure 4.8 is: (1+4+5+4+5+6+2+2)/2 = 14.5

$$\texttt{AvgDepth} = \frac{\sum_j Depth(isLeaf(\texttt{term}_{\texttt{ij}}))}{\texttt{TotalRoots}}$$

**DepthCoesion**: The coesion of a taxonomy is indicated by the maximum depth divided by its average depth. Example: The coesion of the taxonomy in Figure 4.8 is: (6/14.5) ≈ 0.41

$$\texttt{DepthCoesion} = \frac{\texttt{MaxDepth}}{\texttt{AvgDepth}}$$

**MaxWidth**: Maximum width is the maximum number of term siblings in all taxonomies, *i.e.*, the maximum number of hyponyms of a term. Example: The maximum width in Figure 4.8 is: 4 (formed by IDs: 6, 7, 9 and 10 in T1)

$$\texttt{MaxWidth} = Max(Width(\texttt{term}_{\texttt{ij}}))$$

**MinWidth**: Minimum width is the minimum number of term siblings in all taxonomies, *i.e.*, the minimum number of hyponyms of a term. Example: The minimum width in Figure 4.8 is: 1 (single IDs: 4, 5, 8, 13, or 15)

$$\texttt{MinWidth} = Min(Width(\texttt{term}_{\texttt{ij}}))$$

**AvgWidth**: Average width is the ratio between the sum of widths (*i.e.*, the sum of hyponyms) and the total number of siblings (*i.e.*, the total number of terms that have hyponyms), and the number of taxonomies, where "TaxWidth$_j$" measures the width of the taxonomy $j$. Example: The average width in Figure 4.8 is: (((2+1+1+4+1+2+1)/7) + ((1+2)/2))/2 ≈ 1.61

$$\texttt{TaxWidth}_{\texttt{j}} = \frac{\sum_i Width(\texttt{term}_{\texttt{ij}}))}{Count(hasSiblings(\texttt{term}_{\texttt{ij}}))}$$

$$\texttt{AvgWidth} = \frac{\sum_j \texttt{TaxWidth}_{\texttt{j}}}{\texttt{TotalRoots}}$$

## 4.5 Summary

We introduced the resources and the methodology used in this thesis for developing and evaluating models that extract taxonomic relations from text corpora in Portuguese and English. Six corpora in Portuguese and eight corpora in English were used to generate models. For the manual evaluation, the corpora was divided into two groups, testing corpora and contrasting corpora. The testing corpora are composed by four sets of documents, namely Europarl and Geology, both in English and Portuguese, and are used to generate models. Constrasting corpora are composed by ten sets of documents, namely Brown corpus, Conference, Euronews, Footie, Ohsumed and TED Talks for English, and CETEN-Folha, Computer Science, Pediatrics and TED Talks for Portuguese, and are used to rank terms in Testing corpora by contrasting noun phrases and thus, identifying domain terms.

The automatic evaluation is performed using widely accepted, but not perfect, taxonomies as gold standards. For English, the NLTK[25] version of Princeton Wordnet [34] version 3.0 was used. NLTK provides a WordNet interface containing methods to access synsets, lemmas and relations such as hypernymy and hyponymy. For Portuguese, the RDF file containing the lexical ontology Onto.PT [109] version 0.6 was transformed into a directed graph where each noun synset is a node and the edge that connects two nodes represents the taxonomic relation (hypernym-of/hyponym-of) between them.

Before generating models, corpora pass through a preprocessing step. In this step, each corpus is parsed using Stanford Parser [65] (version 3.3.1) for English and PALAVRAS parser [9] for Portuguese. Contexts are extracted from parsed corpora for models that are automatically evaluated, and a concept extraction is applied for all corpora. Concept extraction applies heuristics on terms in order to refine the term extraction process and to weight each term according to its relevance to the domain. The weighting process uses Term Frequency - Disjoint Corpora Frequency ($tf\text{-}dcf$) [84] measure that uses contrasting corpora in order to penalize terms that appear in more than one corpus, indicating that is more likely to be an irrelevant term to the domain corpus.

Having the corpora preprocessed models for the automatic and manual evaluations are generated. For the automatic evaluation, seven models for each language are built, namely `Patt`, `DSim`, `SLQS`, `TF`, `DF`, `DocSub`, `Hclust`. `Patt` is a pattern-based model, *i.e.*, a model uses patterns in text to extract taxonomic relations between terms. `DSim` is a model based on Directional Similarity and takes into account the Distributional Inclusion Hypothesis, which says that two terms have a taxonomic relation if the contexts of one term are shared by the other term. `SLQS` is a model based on entropy and relies on the idea that superordinate terms are less informative than their hyponyms. `TF` model is based on the term frequency and takes into account the number of times a word occur in the whole collection as an indicative of generalization-specialization. `DF` is a model based on the document frequency and takes into account the number of documents in which a word appears as an evidence of taxonomic relation. `DocSub` is a model based on document subsumption and uses the probability of the distribution of the words across shared documents in order to identify a taxonomic relation between them. `Hclust` is a model based on hierarchical clustering and uses contexts to group similar words together and the document frequency to identify the taxonomic relations between them.

For the manual evaluation, four models for each language are developed, namely `Patt`, `DocSub`, `Hclust` and `Hmod`. `Patt`, `DocSub` and `Hclust` are similar to the ones developed for the automatic evaluation, differing only in the type of term used. While models for the automatic evaluation contain relations between nouns, models for the manual evaluation contain relations between noun phrases. Nouns are applied in models for the automatic evaluation due to the fact that the gold standards does not contain a great number of noun phrases. `Hmod` is a model based on the head-modifier, *i.e.*, a noun phrase may contain inner noun phrases when extracting the head and its modifiers.

In order to assess the quality of the extracted relations and indirectly the quality of the applied methods to generate correct relations between terms, automatic and manual evaluations are performed. The former is based on the comparison of the extracted relations against a gold standard,

[25] http://www.nltk.org/

qantifying relations in terms of precision, recall and f-measure, and the latter is performed by human experts that assess a subset of the relations generated by each method. Taxonomies generated by each method are also characterized using metrics, where we can see if a method generate deep of flat taxonomies, with multiple relations of with few relations *etc.*.

# 5. Evaluation of Taxonomic Relation Extraction Methods

This chapter presents a series of experiments aiming to evaluate the methods presented in Section 4.3. In order to verify the quality of the extracted relations and indirectly the quality of the method that generated such relations, we performed automatic and manual evaluations. Automatic evaluation is based on the comparison of the relations extracted from each model with a gold standard, resulting in precision, recall and f-measure for each method. Precision is the fraction of extracted relations that are relevant, while recall is the fraction of relevant instances that are extracted. F-measure is the weighted average of the values corresponding to the two parameters precision and recall. Manual evaluation is based on human judgements on whether a term is taxonomic related to another. As manual evaluation is a time-consuming task, only a subset of all relations are evaluated manually, as explained in Section 4.4.

Section 5.1 presents the automatic evaluation which is performed on relations extracted from Europarl and TED Talks corpora in English and Portuguese, using 7 models: `Patt`, `DSim`, `SLQS`, `TF`, `DF`, `DocSub` and `HClust`. As `DocSub` model may generate different values of precision, recall and f-measure according to the threshold, a deeper analysis on the results generated is performed, as well as for `HClust` model, since this model can generate different results according to the selected number of clusters. Manual evaluation is performed on results extracted from 4 methods (`Patt`, `DocSub`, `HClust` and `HMod`) using Europarl and Geology corpora in both English and Portuguese as described in Section 5.2. Section 5.3 presents an analysis on the characteristics of the taxonomies generated by each method. These characteristics include the number of generated relations, number of taxonomies, depth of the taxonomy, *etc.*. Finally, we analyze the complementarity of the developed models in Section 5.4.

## 5.1 Results for the automatic evaluation

### 5.1.1 All methods

Zipf's law [168] states that the relationship between a word's frequency and the rank order of its frequency is roughly a reciprocal curve, *i.e.*, if we count up how often each word occurs in a large corpus, and then sort the words by their frequency of occurrence, there exists a relationship between the frequency of a word and its position in the sorted list (called rank). For instance, according to this law, the 50th most common word in the corpus should occur with three times the frequency of the 150th most common word.

Because of this Zipfian distribution of words, cutting out low frequency words will greatly reduce our space (as well as the memory requirements of the system), while not considerably affecting the model quality. Thus, we decided to reduce our vocabulary. The first experiment reduces the vocabulary to 1,000 terms, a second experiment considers 10,000 terms and a third experiment considers 1,000 terms but applying a filtering algorithm to induce a taxonomy where each term contains only one hypernym. As in our experiments the word frequency is directly related to the number of contexts, due to the fact that words are extracted in a window, our vocabulary was reduced taking into account only terms that share the highest number of contexts and appear into the gold standard. Thus, for each word existing in the corpus and in the gold standard we counted the number of contexts and selected the top $N$ terms. Using the new vocabulary we applied all models described in Section 4.3.1 and compared the result of each method with the gold standard, generating values of precision, recall and f-measure for each model in each corpus.

- Experiment 1: Using 1,000 terms

In this first experiment we reduced the vocabulary to the top 1,000 terms with the highest number of contexts. Table 5.1 presents the general overview of these values for each method in each corpus, where values in bold are the highest values. As `DocSub` and `HClust` can generate a range of values of precision and recall according to the threshold, the table contains the values with the highest value of f-measure (`DocSub`: $\lambda$=0.1 using all corpora but TED Talks in English which has $\lambda$=0.3, `HClust`: 1,000 clusters for all corpora). `Patt` contains values of precision and recall and f-measure considering all rules for the limited vocabulary.

| | Language | Corpus | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{P}$ | EN | Europarl | **0.1173** | 0.0366 | 0.0503 | 0.0554 | 0.0548 | 0.0443 | 0.0761 |
| | | Ted Talks | **0.1125** | 0.0301 | 0.0382 | 0.0425 | 0.0441 | 0.0710 | 0.0664 |
| | PT | Europarl | 0.5163 | 0.3330 | 0.5257 | 0.6109 | 0.5984 | **0.7311** | 0.5676 |
| | | Ted Talks | 0.5387 | 0.2907 | 0.5300 | 0.6117 | 0.6159 | **0.6533** | 0.5656 |
| $\mathcal{R}$ | EN | Europarl | 0.0396 | 0.3999 | 0.5499 | **0.6045** | 0.5887 | 0.0023 | 0.0017 |
| | | Ted Talks | 0.0018 | 0.4442 | 0.5377 | 0.5657 | **0.6077** | 0.2666 | 0.0019 |
| | PT | Europarl | 0.0111 | 0.3554 | 0.5795 | **0.6727** | 0.5184 | 0.0053 | 0.0012 |
| | | Ted Talks | 0.0004 | 0.3142 | 0.5484 | **0.6877** | 0.5515 | 0.4706 | 0.0011 |
| $\mathcal{F}$ | EN | Europarl | 0.0591 | 0.0671 | 0.0922 | **0.1015** | 0.1003 | 0.0044 | 0.0033 |
| | | Ted Talks | 0.0035 | 0.0564 | 0.0713 | 0.0791 | 0.0822 | **0.1121** | 0.0037 |
| | PT | Europarl | 0.0217 | 0.3438 | 0.5513 | **0.6403** | 0.5555 | 0.0105 | 0.0024 |
| | | Ted Talks | 0.0008 | 0.3020 | 0.5390 | **0.6475** | 0.5819 | 0.5471 | 0.0022 |

Table 5.1: Precision, recall and F-measure for methods using the top 1,000 words with the highest number of contexts.

Analyzing Table 5.1, we can see that all values of precision using the Portuguese corpora have higher scores when compared with the English corpora. This is due to the fact that the Portuguese gold standard, Onto.PT, has more connections between synsets. For example, the word "`dog`" in WordNet is present in a total of 7 synsets, its direct hypernyms (first level of hypernyms) are distributed into 8 synsets and the second level of hypernyms (hypernyms of hypernyms of the word "`dog`") contains 16 synsets. On the other hand, the word "`cachorro`" (dog) in Onto.PT appears in 4 synsets, its direct hypernyms are distributed into 21 synsets and the second level of hypernyms contains 33 synsets. A higher number of terms associated in hypernyms tends to increase the precision. Another aspect to be considered is the fact that as Onto.PT is automatically constructed, there are relations that would not exist if it was manually constructed or revised. For instance, a synset containing the word "homem" (man) is hypernym of a synset containing the words "cara" (face), "face" (face), "fronte" (forehead) and "testa" (forehead). Although the polyssemy of the word "cara" (face, but also guy) may make the taxonomic relation with "homem" correct, since "cara" (guy) is a kind of "homem" (man), the other words that belong to the same synset will make this relation wrong, since "cara" (face) is not a kind of "homem" (man). Assuming that the relation between both synsets is correct because we have a polyssemic word will make relations such as "testa" (forehead) is a kind of "homem" (man) also correct, even if they are not.

As we can observe in Table 5.1, Patt has the best values of precision for the English corpora while DocSub has the best values for the Portuguese corpora. TF has the best values of recall and f-measure for all corpora but the English version of TED Talks which has in DF the best value of recall and in DocSub the best value of f-measure. It was expected quite similar values of precision, recall and f-measure between TF and DF using the Europarl corpora since the size of each document was set to the size of the phrase because Europarl does not have document borders. Thus, terms that appear only once in a phrase have the same value of TF and DF. This value differs when a term appears more than once in a phrase, and thus, having a higher value of TF than DF. In some cases it seems to make difference in results, *e.g.*, Europarl in Portuguese which increased the precision from $\mathcal{P}$=0.5984 in DF to $\mathcal{P}$=0.6109 in TF, as well as the recall from $\mathcal{R}$=0.5184 in DF to $\mathcal{R}$=0.6727 in TF, resulting in an increase of f-measure from $\mathcal{F}$=0.5555 in DF to $\mathcal{F}$=0.6403 in TF. On the other hand, TED Talks corpora, which have document borders and each document contains dozens phrases, have similar values when compared with Europarl corpora. It makes sense since we are using terms that share a the highest number of contexts and thus, terms that should appear in a greater number of documents.

When comparing DF model which takes into account only the number of documents that the word occurs, with DocSub which considers the number of shared documents between two words, DocSub achieved better values of precision, but lower values of recall. In fact, DocSub had worse results in precision only when using Europarl corpus in English, where DF reached best values of precision and f-measure. As DocSub uses the shared documents, it seems reasonable that it has lower recall when compared with DF. In Section 5.1.3, a further analysis on the distribution of precision, recall and f-measure using the DocSub model shows that the highest values of f-measure were obtained using very low thresholds (when the number of documents shared are close to 10%), and thus, approximating to the values of DF.

Another interesting observation is to compare the results obtained by DF with the results achieved by HClust. This comparison is interesting because HClust uses the values of document frequency over semantically clustered terms. By clustering semantically related terms, the HClust model intends to increase the precision of the extracted relations with a detrimental effect on the recall. As we can observe, it seems that clustering semantically related terms will increase the precision (at least for the top 1,000 terms in the English corpora used in this experiment) as expected. On the other hand, the problem of clustering similar terms is that terms that occur in the same contexts tend to be synonyms or co-hyponyms instead of hypernyms. Thus, HClust may cluster also synonyms or co-hyponyms that differ in terms of document frequency and thus, the model classify them as a hypernym-hyponym relation. For instance, the cluster containing the highest value of similarity using the TED Talks corpus in English is composed by the terms "consumer" which occurs in 196 documents, and "employee" which occurs in 149 documents. Thus, according HClust model the relation <consumer, is-a, employee> holds. Analyzing their relation in WordNet, they both are hyponym of the synset "person.n.01" but there is not a taxonomic relation between them.

Low values of precision were expected for methods that use the distribution of the words across documents due to the fact that such methods are good to indicate a semantic relation between terms, but not really good to identify the type of semantic relation. On the other hand, values of precison were expected to be high for the method based on patterns (Patt). Patterns are well known for having high precision and low recall values. In fact, compared to the other methods, Patt achieved better precision and lower recall, although it is still a low precision when compared with other works. Patterns are analyzed in details in Section 5.1.2. DSim obtained the lowest scores of precision, recall and f-measure, lower than the usual baseline TF. This low result might be because many words do not share any context with other words and, even if their relation exists in WordNet, it can not be detected by ClarkeDE measure. As SLQS takes into account the median entropy of all its most related contexts as an estimate of word informativeness, it does not need to share any context with the other word to identify their taxonomic relation. Also, it might work well in small data sets, since it uses a limited number of contexts to generate the median entropy. In our work, we used the same number of most associated contexts as the original work ($N = 50$), being the contexts ranked by their value of Local Mutual Information (LMI) with the target word.

- Experiment 2: Using 10,000 terms

Observing results from HClust, which obtained the best f-measure in a cluster containing 1,000 terms (the f-measure was still rising), we decide to perform a second experiment, increasing the number of terms, and consequently clusters. Adding more terms and clusters could increase the number of semantic relations and maybe taxonomic relations would be more evident. As the number of terms may influence the results we decided to perform the experiment using up to 10,000 words in the dictionary. This number of words is higher than the maximum number of words in the Portuguese version of TED Talks after filtering terms with Onto.PT (7,066 words). The other corpora have more terms than this threshold (TED Talks in English contains 19,601 words, Europarl in Portuguese 13,139 words and Europarl in English 32,007 words). Table 5.2 presents the values of precision and recall for all models using a vocabulary containing up to 10,000 words, where DocSub and HClust contain results when the best f-measure was achieved, and Patt considers all patterns with the limited number of words.

As we can see, values of precision were lower for most methods, with exception of Patt and DocSub, which increased for most corpora. When increasing the number of terms to 10,000, the

| | Language | Corpus | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{P}$ | EN | Europarl | **0.1192** | 0.0083 | 0.0137 | 0.0150 | 0.0150 | 0.0445 | 0.0326 |
| | | Ted Talks | **0.1022** | 0.0069 | 0.0060 | 0.0092 | 0.0090 | 0.0356 | 0.0162 |
| | PT | Europarl | 0.5710 | 0.1948 | 0.3855 | 0.5474 | 0.4485 | **0.8052** | 0.4058 |
| | | Ted Talks | **0.6304** | 0.1870 | 0.3250 | 0.5312 | 0.4576 | 0.6064 | 0.3698 |
| $\mathcal{R}$ | EN | Europarl | 0.0037 | 0.3278 | 0.5941 | 0.6486 | **0.6490** | 0.0017 | 0.0003 |
| | | Ted Talks | 0.0002 | 0.1486 | 0.4332 | **0.6467** | 0.6332 | 0.0967 | 0.0003 |
| | PT | Europarl | 0.0002 | 0.1562 | 0.5157 | **0.7255** | 0.5932 | 0.0032 | 0.0001 |
| | | Ted Talks | $2.10^{-5}$ | 0.0507 | 0.4492 | **0.7000** | 0.5887 | 0.1390 | 0.0002 |
| $\mathcal{F}$ | EN | Europarl | 0.0073 | 0.0162 | 0.0268 | **0.0293** | **0.0293** | 0.0033 | 0.0006 |
| | | Ted Talks | 0.0004 | 0.0132 | 0.0118 | 0.0181 | 0.0179 | **0.0520** | 0.0005 |
| | PT | Europarl | 0.0005 | 0.1733 | 0.4412 | **0.6240** | 0.5109 | 0.0064 | 0.0002 |
| | | Ted Talks | $4.10^{-5}$ | 0.0798 | 0.3771 | **0.6040** | 0.5149 | 0.2261 | 0.0004 |

Table 5.2: Precision, recall and F-measure for methods using the top 10,000 words with the highest number of contexts.

DocSub models using Europarl corpora performed better than when using TED Talks corpora. It seems that the higher the number of documents, the more accurate relations become. Although decreasing the values of precision, TF and DF increased the values of recall, but decreasing the values of f-measure. As occured in the experiment using the top 1,000 words, this experiment also kept TF with the highest values of f-measure for most methods. TF and DF achieved almost the same values of precision, recall and f-measure using the English corpora, achieving the same value of precision ($\mathcal{P}$=0.0150) and f-measure ($\mathcal{F}$=0.0293) when using the Europarl corpus in English.

When comparing DF with HClust, it seems a good approach in English to verify the hierarchical relation only for terms that are semantically related instead of considering all terms. As occurred in the Experiment 1, DF performed better than HClust using the Portuguese corpora. The lowest values of precision are achieved by DSim model, and the lowest recalls are obtained by HClust and Patt models. These models are known for having lower levels of recall since HClust tries to increase precision in detriment of recall when clustering similar words, and Patt model uses patterns that are very sparce in texts. As these models have very low values of recall, they also contain the lowest values of f-measure. Observing the increase of terms for hierarchical clustering, all values of precision, recall and f-measure decreased when comparing with the ones obtained using 1,000 terms.

- Experiment 3: Selecting the best parent using 1,000 terms

The third experiment intends to remove multiple hypernyms when it occurs to a term, maintaining the taxonomy with a tree structure. The decision for the correct hypernym of a term is based on a score calculated for each potential hypernym. The score is defined in Equation 2.15, as explained in Section 2.1.5. Table 5.3 presents the values of precision, recall and f-measure for the methods in all corpora. The filtering on multiple hypernyms is applied in relations extracted using 1,000 terms in the dictionary. Thus, Table 5.1 can be compared with Table 5.3.

| | Language | Corpus | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{P}$ | EN | Europarl | **0.1038** | 0.0170 | 0.0490 | 0.0641 | 0.0641 | 0.0613 | 0.0761 |
| | | Ted Talks | **0.1282** | 0.0291 | 0.0410 | 0.0270 | 0.0270 | 0.1154 | 0.0661 |
| | PT | Europarl | 0.6185 | 0.3744 | 0.4144 | 0.4394 | 0.4394 | **0.7553** | 0.5676 |
| | | Ted Talks | 0.6308 | 0.4124 | 0.4404 | 0.4515 | 0.4945 | **0.8609** | 0.5295 |
| $\mathcal{R}$ | EN | Europarl | **0.0021** | 0.0004 | 0.0011 | 0.0014 | 0.0014 | 0.0013 | 0.0017 |
| | | Ted Talks | 0.0011 | 0.0008 | 0.0011 | 0.0008 | 0.0008 | **0.0030** | 0.0018 |
| | PT | Europarl | 0.0012 | 0.0008 | 0.0009 | 0.0010 | 0.0010 | **0.0016** | 0.0012 |
| | | Ted Talks | 0.0003 | 0.0009 | 0.0009 | 0.0010 | 0.0010 | **0.0017** | 0.0011 |
| $\mathcal{F}$ | EN | Europarl | **0.0041** | 0.0007 | 0.0021 | 0.0027 | 0.0027 | 0.0026 | 0.0033 |
| | | Ted Talks | 0.0022 | 0.0016 | 0.0022 | 0.0015 | 0.0015 | **0.0058** | 0.0036 |
| | PT | Europarl | 0.0024 | 0.0016 | 0.0018 | 0.0019 | 0.0019 | **0.0031** | 0.0023 |
| | | Ted Talks | 0.0005 | 0.0018 | 0.0018 | 0.0020 | 0.0021 | **0.0034** | 0.0022 |

Table 5.3: Precision, recall and F-measure for methods using the top 1,000 words with the highest number of contexts and selecting the best parent.

Analyzing Table 5.3 we observe that the Patt method achieves again the best precision values of all methods for the English corpora. On the other hand, choosing the best hypernym worked very well for DocSub which obtained the best precision for the Portuguese corpora. As filtering out multiple hypernyms might remove also correct relations, the recall values for all corpora are very low. Comparing the values achieved by methods containing all relations (Table 5.1) and the reduced taxonomies, we can see that all recalls and f-measures decreased, as expected. The values of precision increased for most corpora of the Patt and DocSub models. Using f-measure as reference, it seems not worth applying this kind of filtering because the loss in recall is much greater than the gain in precision.

Comparing the values achieved by HClust using 1,000 terms and the ones obtained when reducing to one parent, we can see that they are almost the same. Only TED Talks corpora obtained a decrease in precision (from $\mathcal{P}$=0.0664 to $\mathcal{P}$=0.0661 in English and from $\mathcal{P}$=0.5656 to $\mathcal{P}$=0.5295 in Portuguese). It seems that when applying the hierarchical clustering the taxonomies are reduced as they are when the algorithm to one parent is applied.

Analyzing all results generated for each method in each corpora, the algorithm to choose the best parent seems to work mainly with Patt and DocSub models. Increasing the number of terms form 1,000 to 10,000 only increased the recall in TF and DF models. Clustering semantically related terms also seems a good strategy to reduce the number of relations and increase the precision, when not considering the recall. Regarding the corpora, the highest precision obtained by corpora in English was achieved by Patt models, while the highest recall by statistical models such as TF and DF. The highest f-measure for corpora in English was achieved by TF model in Europarl and DocSub model in TED Talks. For the corpora in Portuguese, the highest precision was achieved by DocSub model, while the highest recall and f-measures were achieved by TF models.

In the next sections we analyze in details the results of some models presented here. In Section 5.1.2 we analyze the pattern-based models (Patt), observing the values of precision, recall and f-measure of each pattern applied in English and Portuguese., Document subsumption-based models (DocSub) is analyzed in Section 5.1.3, where we vary the threshold values and verify the difference in terms of precision, recall and f-measure. Section 5.1.4 analyzes hierarchical clustering-based models HClust, where we verify the impact in precision, recall and f-measure when adding new clusters to the model.

### 5.1.2 Results from Pattern-based model (Patt)

After analyzing results for all patterns together, using some threshold limiting the number of terms in the dictionary, we decided to analyze the results using all possible terms and also verify what is the contribution of each applied pattern. Results for precision ($\mathcal{P}$), recall ($\mathcal{R}$) and f-measure ($\mathcal{F}$) of each pattern using the English version of the corpora are presented in Table 5.4, where the last line containing "All patterns" means that the values consider all patterns together, and values presented in bold are the highest scores.

| Patterns | **Europarl** | | | **TED Talks** | | |
|---|---|---|---|---|---|---|
| | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ |
| 1. NP such as LNP | 0.1832 | $38.10^{-4}$ | $77.10^{-5}$ | **0.3200** | $11.10^{-6}$ | $22.10^{-6}$ |
| 2. such NP as LNP | **0.2363** | $35.10^{-6}$ | $70.10^{-6}$ | 0.1429 | $10.10^{-7}$ | $20.10^{-7}$ |
| 3. NP including LNP | 0.1375 | $12.10^{-4}$ | $23.10^{-5}$ | 0.2545 | $50.10^{-7}$ | $10.10^{-6}$ |
| 4. NP especially LNP | 0.1820 | $25.10^{-6}$ | $50.10^{-6}$ | 0.1667 | $10.10^{-7}$ | $20.10^{-7}$ |
| 5. NP and/or other LNP | 0.2195 | $26.10^{-4}$ | $52.10^{-5}$ | 0.2851 | $23.10^{-6}$ | $46.10^{-6}$ |
| 6. NP is NP | 0.0861 | $28.10^{-4}$ | $54.10^{-5}$ | 0.0978 | $50.10^{-6}$ | $10.10^{-5}$ |
| 7. LNP are LNP | 0.0844 | $79.10^{-6}$ | $16.10^{-5}$ | 0.0935 | $10.10^{-6}$ | $20.10^{-6}$ |
| 8. All patterns | 0.1228 | $\mathbf{95.10^{-5}}$ | $\mathbf{19.10^{-4}}$ | 0.1527 | $\mathbf{81.10^{-6}}$ | $\mathbf{16.10^{-5}}$ |

Table 5.4: Results for each pattern using Europarl and TED Talks corpora in English.

As observed by Snow *et al.* [135], the patterns with the highest precision scores were pattern "NP such as LNP" and "such NP as LNP". The lowest values of precision were obtained by the called "generic patterns" [112], *i.e.*, patterns that would probably have low values of precision and high values of recall. Analyzing the obtained results, generic patterns obtained low values of precision but average values of recall. The pattern used by Caraballo [15] to select hyponyms in clusters also have values of precision equivalent to the ones with the highest values. As methods based on patterns are well known for having high precision and low recall scores, the recall scores obtained by each pattern were very low. The best recall and f-measure were obtained using all possible patterns.

Using the TED Talks corpus we found a cycle, *i.e.*, a hypernym points to a hyponym and this hyponym points to the hypernym. The occurrence contains the cycle: services > firms > communications > services, and were discovered in a phrase of the talk "One Laptop per Child, two years on"[1] by Nicholas Negroponte. The phrase says: "If you look at our professional services, including search firms, including communications, including legal services, including banking, they're all pro bono.". From this phrase we can see the problem that is to work with transcribed talks.

[1] https://www.ted.com/talks/nicholas_negroponte_on_one_laptop_per_child_two_years_on/

Hardly such sentence would be found in written texts, since it is not very common to write the word "including" so many times. As a spoken sentence, it is more common to repeat words in order to reinforce the idea.

A second issue in this phrase refers to the pattern applied. The phrase, the pattern used and the matched patterns from the phrase are:

```
Phrase:

If you look at our professional services, including search firms, including communications, including
legal services, including banking, they're all pro bono.

Pattern:

NP including LNP

Identified Patterns:

[NP professional services], including [NP search firms]
[NP search firms], including [NP communications]
[NP communications], including [NP legal services]
[NP legal services], including [NP banking]
```

Although the pattern erroneously extracts the relations <search firms, is-a, professional services>, <communications, is-a, search firms>, and <legal services, is-a, communications>, they do not have a cycle. However, the way that we generate terms to maximize the number of relations to be evaluated creates the cycle. As explained in Section 4.2.2, WordNet does not cover a great number of noun phrases, thus we decided to use the head of the noun phrase instead of the whole noun phrase in all models. Extracting nouns, we generate the relations: `<firms, is-a, services>`, `<communications, is-a, firms>`, `<services, is-a, communications>`, creating the cycle. On the other hand, using the head of the noun phrase instead of the whole noun phrase in order to find a taxonomic relations works for many cases. For example, from the phrase:

> "We all know that nothing will prevent possible partnership **countries**, including **China**, from importing timber into Europe illegally via third countries anyway."

We can identify the relation `<China, is-a, possible partnership countries>` using the pattern "`NP, including NP`". Looking for the entire noun phrases in WordNet, we would not find the relation, since the term "`possible partnership countries`" does not exist. When assuming that the head of the noun phrase is the word that carries the main part of the meaning, we can find the relation `<China, is-a, countries>`, where "`China`" is a kind of/is an instance of "`countries`". Using the function that generates the lemma of a word in NLTK–WordNet API we can find out that the relation `<China is-a country>` exists.

Analyzing the results, we also identified relations that are correct but missing in the gold standard and wrong relations provided by the structure of the phrase. For example, consider the phrase:

> "This must be achieved by taking advantage of opportunities provided by modern technology, including radio and television."

Using the pattern "`NP, including LNP`", we would extract the relations <`radio, is-a, modern technology`> and <`television, is-a, modern technology`>. As WordNet does not have an entry for "modern technology", we use the head of the noun phrase "technology" to verify the taxonomic relation with "radio" and "television". Although"television" is hyponym of "telecom", "telecommunication system", "electronic equipment" and "physical object", it is not hyponym of, or taxonomically related to "technology". Other relations that are not found in WordNet: "gas (is a kind of) energy", "wind (is a kind of) energy source", "solar energy (is a kind of) energy source", *etc.*.

Considering the structure of the phrase that may lead to errors, consider the following phrase:

> "All the states in crisis in Europe, all of those which blocked the new moves, including Hungary and Spain, have socialist governments."

The pattern "`NP, including LNP`" would erroneously identify the relations <`Hungary, is-a, new moves`> and <`Spain, is-a, new moves`>. However the phrases "All the states in crisis in Europe" and "all of those which blocked the new moves" are in apposition, *i.e.*, the second element is serving to identify the first one in a different way. As in this case, the pattern may erroneously identify a taxonomic relation when the correct relation must be meronymy (part-of). This is the case in the phrase:

> "The European Union, including Hungary, can only maintain its leading position by introducing new economic solutions."

where the pattern identifies the taxonomic relation <`Hungary, is-a, European Union`>, when the correct relation would be "Hungary" is a part of "European Union". In order to have more evidences of these errors a manual evaluation is performed.

Patterns using the Portuguese corpora are also analyzed. Not all patterns from Table 2.1 were found in texts using the Portuguese corpus. For example, there is no occurrence of the patterns "`NP principalmente LNP`", "`NP de maneira espercial LNP`", "`NP sobretudo LNP`", *etc.*. Patterns that returned taxonomic relations using the Portuguese corpora are presented in Table 5.5. As we can observe, using TED Talks corpus, the model could not find any occurence that matchs the patterns "`NP em especial LNP`" and "`NP e/ou outro(s) LNP`".

The highest precision values were obtained by patterns that are the translations of the patterns with the highest precision in English. Using Europarl corpus, the highest precision was obtained by the pattern "`NP e/ou outro(s) LNP`" which is the translation to the pattern "`NP and/or other LNP`" in English. For TED Talks, the highest precision was achieved by the pattern "`tal(is) NP como LNP`" which is the translation to the pattern "`such NP as LNP`". It is interesting to notice that the highest precision using the TED Talks achieved the lowest precision using Europarl corpus.

| Patterns | Europarl | | | TED Talks | | |
|---|---|---|---|---|---|---|
| | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ |
| 1. NP tal(is) como LNP | 0.58 | $30.10^{-7}$ | $60.10^{-7}$ | 0.50 | $26.10^{-8}$ | $10.10^{-7}$ |
| 2. tal(is) NP como LNP | 0.47 | $40.10^{-6}$ | $80.10^{-6}$ | 0.58 | $20.10^{-6}$ | $39.10^{-6}$ |
| 3. NP incluindo LNP | 0.50 | $11.10^{-4}$ | $22.10^{-6}$ | 0.56 | $10.10^{-7}$ | $20.10^{-6}$ |
| 4. NP em especial LNP | 0.62 | $10.10^{-7}$ | $20.10^{-7}$ | 0.00 | 00.00 | 00.00 |
| 5. NP e/ou outro(s) LNP | 0.66 | $32.10^{-6}$ | $64.10^{-6}$ | 0.00 | 00.00 | 00.00 |
| 6. NP é um NP | 0.55 | $17.10^{-5}$ | $35.10^{-5}$ | 0.33 | $10.10^{-6}$ | $30.10^{-7}$ |
| 7. All rules | 0.55 | $17.10^{-5}$ | $35.10^{-5}$ | 0.58 | $20.10^{-5}$ | $39.10^{-6}$ |

Table 5.5: Results for each pattern using Europarl and TED Talks corpora in Portuguese.

### 5.1.3 Document subsumption-based analysis (DocSub)

According to Sanderson and Croft [128] a term $u$ subsumes another term $v$ if the documents which $v$ occurs in are a subset of the documents which $u$ occurs in, and their conditional probability is over a threshold $\lambda$, as explained by Equation 2.17. Initially the value of $\lambda$ was set to 1, and thus, all documents of $v$ must be shared with $u$. Although there were terms that satisfied the condition, Sanderson and Croft noticed that many terms were not included in relations because a few occurrences of the term $v$ did not co-occur with $u$. Thus, through informal analysis of the subsumption term pairs they decided to relax the threshold to $\lambda$=0.8.

Instead of using a fixed threshold, in automatic evaluation we decided to vary the threshold ($\lambda$) and observe the impact in each corpus. The model is evaluated with threshold varying in a range between 0.1 and 1 with a step of 5% of the maximum threshold, on the sets containing 1,000 and 10,000 terms for each corpus. Figure 5.1 presents the values of precision, recall and f-measure in each corpus, where the x-axis contains the variation of the threshold.

As we can observe, the highest values of precision were achieved by the model using TED Talks with the top 1,000 terms. This model achieved the highest precision with a threshold $\lambda$=1 ($\mathcal{P}$=0.0833). On the other hand, only 5 correct relations were found using such a restricted threshold. The second highest precision was achieved with a threshold $\lambda$=0.45 ($\mathcal{P}$=0.0802), which found 4,795 correct relations. The highest f-measure was achieved by the threshold $\lambda$=0.3 ($\mathcal{F}$=0.1121), which obtained the precision $\mathcal{P}$=0.0710. Europarl containing the top 1,000 terms achieved good values of precision up to a threshold $\lambda$=0.5. Increasing the threshold, *i.e.*, restricting more the document subsumption, the model starts to decrease the value of precision, achieving $\mathcal{P}$=0.0 when using a threshold $\lambda$=0.8 or more.

The zero numbers in all precision, recall and f-measure indicate that the threshold is too strict to find terms that share at least 80% of the documents. For a model using using a limited number of terms and the Europarl corpus where the document is represented by a sentence, it is understandable. For example, 1,236 is the minimum number of documents for a term in the model with this configuration. Thus, to exist a relation with a threshold $\lambda$=0.8, two terms must share at least 989 documents. TED Talks corpora are much more flexible in this case, since this corpus contains 1,112 documents. Using a configuration of the top 1,000 terms, the minimum number of documents that

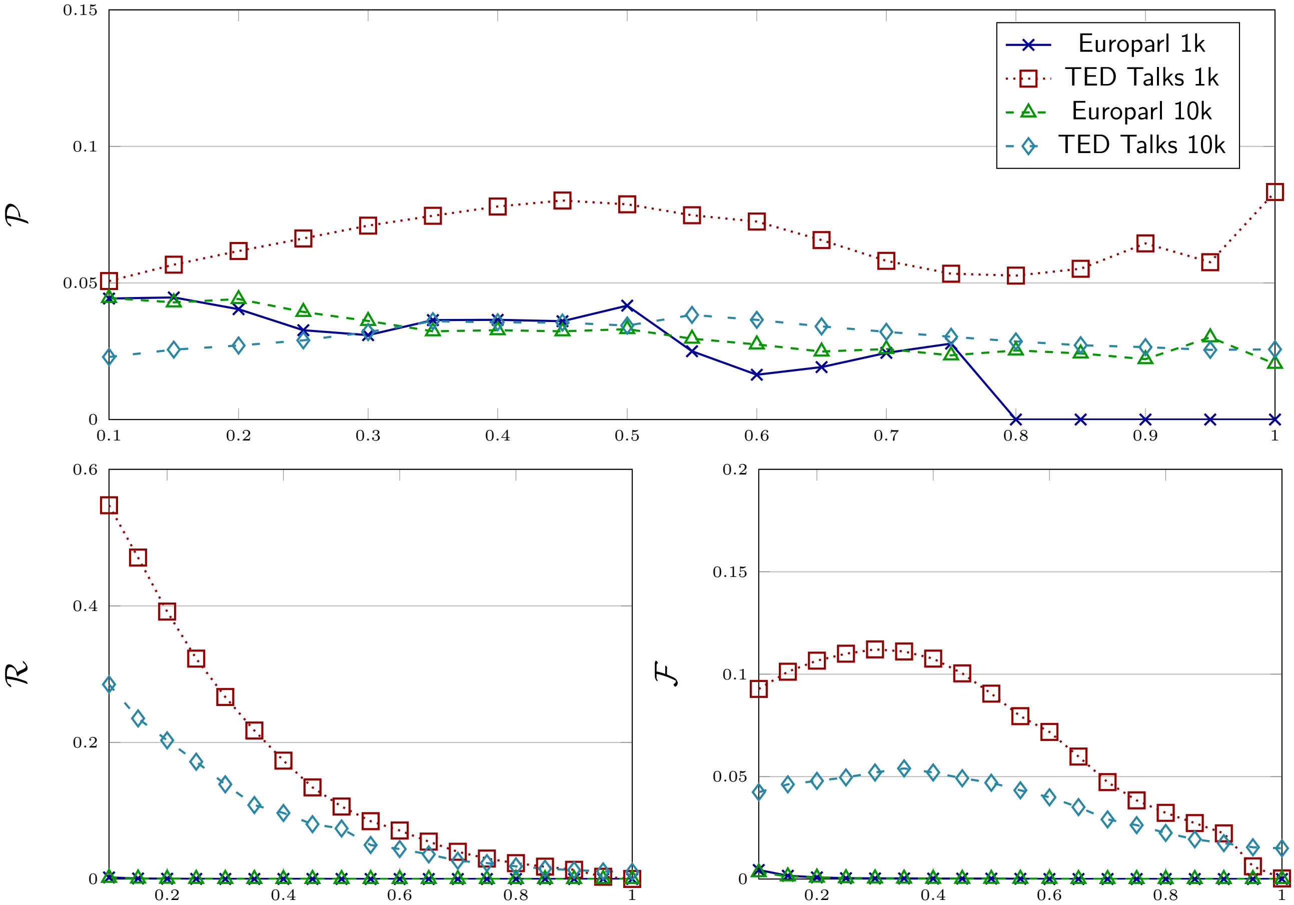


Figure 5.1: Precision, recall and F-measure of `DocSub` for data sets in English using 1,000 terms (1k) and 10,000 terms (10k)

a term contains is 20. Thus, two terms must share at least 16 documents if we want the model to infer a taxonomic relation between them.

When adding new terms to the model, *i.e.*, changing the range from the top 1,000 terms to the top 10,000 terms, the minimal number of documents also decreases, facilitating the model to discover taxonomic relations between terms. For instance, the minimal number of documents that a term must appear decreases from 1,236 to 5 when increasing the number of terms from the top 1,000 to the top 10,000 in the Europarl corpus. Although adding new terms helps the model to discover new relations since the number of shared documents decreases, it also decreases the precision since many terms that share few documents are not hierarchically related. This can be better observed in the model using TED Talks corpus that decreased the precision from $\mathcal{P}$=0.0833 in a model with 1,000 terms to $\mathcal{P}$=0.0257 in a model with 10,000 terms when using $\lambda$=1.

Figure 5.2 presents the values of precision, recall and f-measure in each threshold using the Portuguese corpora. As we can observe, the values of precision were quite constants for the models using corpora containing the top 10,000 terms, meaning that even increasing the threshold value, the model keeps finding the correct relations. For models using the top 1,000 terms the precision tended to increase as the threshold value did so. Thus, the less terms in the model, the better the model

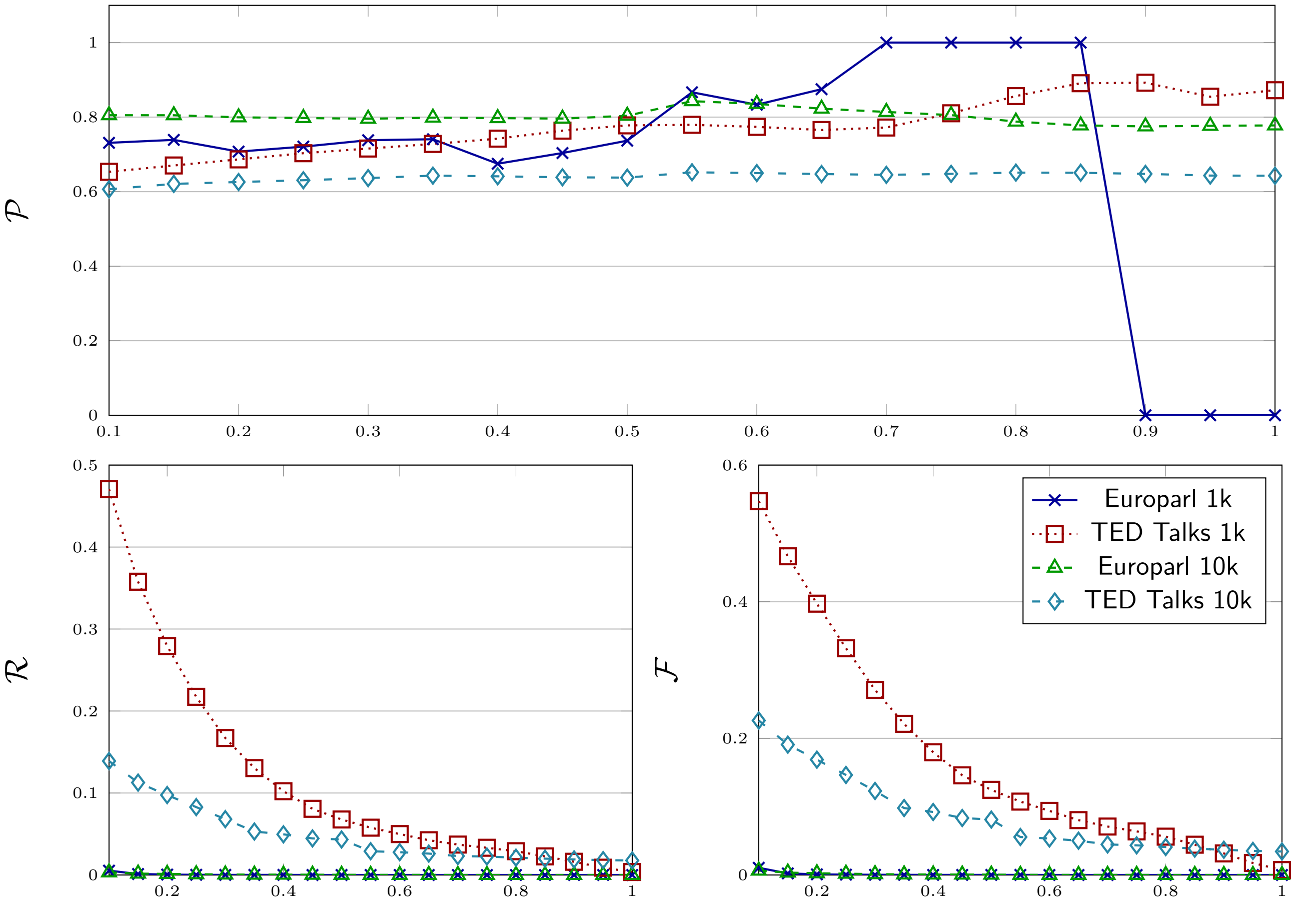


Figure 5.2: Precision, recall and F-measure of `DocSub` for data sets in Portuguese using 1,000 terms (1k) and 10,000 terms (10k)

performs. For similar reasons, the model using Europarl suffered the same problem as the one of the English version, *i.e.*, the higher the threshold, the harder to find terms that share a great number of documents. Thus, when the threshold is set to $\lambda$=0.9 in the model the minimum number of shared documents must be 801. The high precision in the thresholds lower than $\lambda$=0.9 is due to the low number of relations found by the model. From the threshold $\lambda$=0.7 to $\lambda$=0.85 the model found two correct relations, achieving a precision of $\mathcal{P}$=1.0. But, due to the small number of relations found by the model, the recall is very low ($\mathcal{R}$=4.10$^{-6}$), and consequently the f-measure is also very low ($\mathcal{F}$=9.10$^{-6}$). Although the low values of recall ($\mathcal{R}$=0.0032) and f-measure ($\mathcal{F}$=0.0064) with the lowest threshold ($\lambda$=0.1), the model achieved a high value of precision ($\mathcal{P}$=0.8052), identifying in total 119,295 correct relations.

As occurred using models in English, the model using TED Talks obtained greater values of f-measure, finding relations between terms with high values of threshold (above $\lambda$=0.8). As a comparative, 7 was the minimum number of documents that must be shared when using the model with the top 1,000 terms in TED Talks with the highest threshold $\lambda$=1. Thus, it is easier to find relations in TED Talks than in Europarl where two terms must share at least 890 documents with the highest threshold.

The highest f-measure achieved by the model using the top 1,000 terms in TED Talks was

$\mathcal{F}$=0.5471 with the lowest value of threshold $\lambda$=0.1. Due to the lower number of documents when compared with the models using Europarl, the model using the top 10,000 terms of TED Talks also achieved a great value of f-measure ($\mathcal{F}$=0.2261), with a precision of $\mathcal{P}$=0.6064 using the threshold $\lambda$=0.1.

### 5.1.4 Hierarchical clustering-based analysis (HClust)

Unlike most clustering approaches where terms are grouped together creating clusters with several terms, in hierarchical clustering, terms are grouped together incrementally: every term is set in a distinct (singleton) cluster, and successively clusters are merged together until a stopping criterion is satisfied or until all terms belong to one cluster [57]. As we have no clue about a stopping criterion, we decided to generate clusters incrementally up to the maximum number of terms. Calculating precision recall and f-measure for each cluster is a time and resource consuming task, thus, we decided to generate such values using a step of 10% of the maximum number of terms. For models using the top 1,000 terms that share the highest number of contexts we use a step of 100, and for models using the top 10,000 terms we use a step of 1,000. Note that this step refers to the number of clusters and not to the number of terms in clusters.

Figure 5.3 presents values of precision, local recall, local f-measure, global recall and global f-measure for HClust using the corpora in English with 1,000 terms. For the sake of visualization, the $x$ axis containing the number of clusters was converted from linear to logarithmic scale ($\log_{10}$). Global recall and f-measure are interesting to understand how the general scores are impacted by the inclusion of clusters since it takes into account the full gold standard, while the local recall and f-measure can show how the scores are impacted in each clusters instead of globally.

Local recall and local f-measure are calculated considering the gold standard as the maximum number of correct relations in the local cluster, while global recall and global f-measure are calculated considering the gold standard with the maximum number of correct relations in the model. For example, consider that we are calculating recall for HClust model with the top 1,000 terms and the gold standard contains 40,000 correct relations for these 1,000 terms. In the first step of the hierarchical clustering, which contains 100 clusters, the model found 200 correct relations, while the gold standard contains 2,000 correct relations for these 100 terms. Thus, the local recall is calculated as $(\frac{200}{2000}) = 0.1$, while the global recall is calculated as $(\frac{200}{40000}) = 0.005$. Local recall and local f-measure allow us to observe the impact in the model as we add new clusters, while global recall and global f-measure allow us to see how the model grows globally. Precision of the models using Europarl perfomed better in general. Better results may be obtained with a larger size of the corpus, and consequently a greater number of contexts for each term. As hierarchical clustering uses the contexts of each term to cluster them together, a great number of contexts tend to improve the quality of such clusters. We can also observe that the values of precision are increasing when adding new clusters to the model. Nevertheless, when the range is changed from the top 1,000 terms to 10,000 terms, the precision decreases to less than 0.01. It seems that when changing the range and consequently the number of terms, the model also included terms that are not semantically related to others, but still containing similar context vectors.

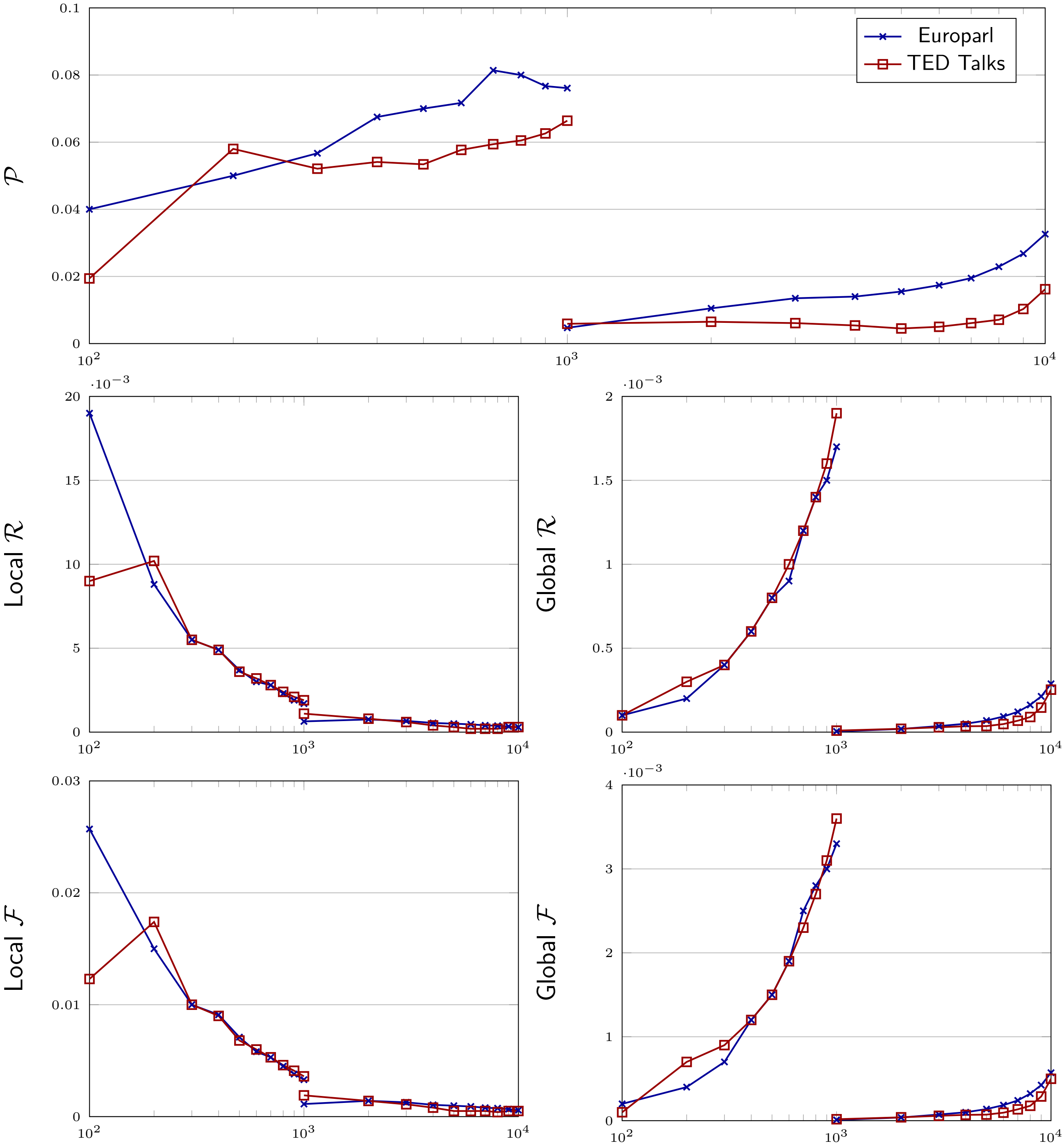


Figure 5.3: Precision, local and global recall, local and global F-measure of `HClust` for data sets in English using 1,000 terms and 10,000 terms.

Adding more terms to the system might help to cluster together terms containing small vectors. For example, consider that when using the top 1,000 terms with the highest number of contexts, the mininum number of contexts to a term is 100. When changing the range to the top 10,000 terms, the minimum number of contexts decays to 10. Thus, terms that share at least 10 contexts may be clustered before the terms that share 100 contexts.

Local recall decreases in each new clustering, indicating that the more terms are included in clusters, the more taxonomic semantic relations are not inferred by the model. On the other hand, the global recall increases when a new cluster is inserted, as expected. As global recall takes into account the maximum number of correct relations in the model, it may be divided into two parts, where the first contains the 1,000 clusters and take into account a gold standard with 1,000 terms, and the second contains the 1,000th cluster to the 10,000th cluster and takes into account the maximum number of correct relations for the 10,000 terms.

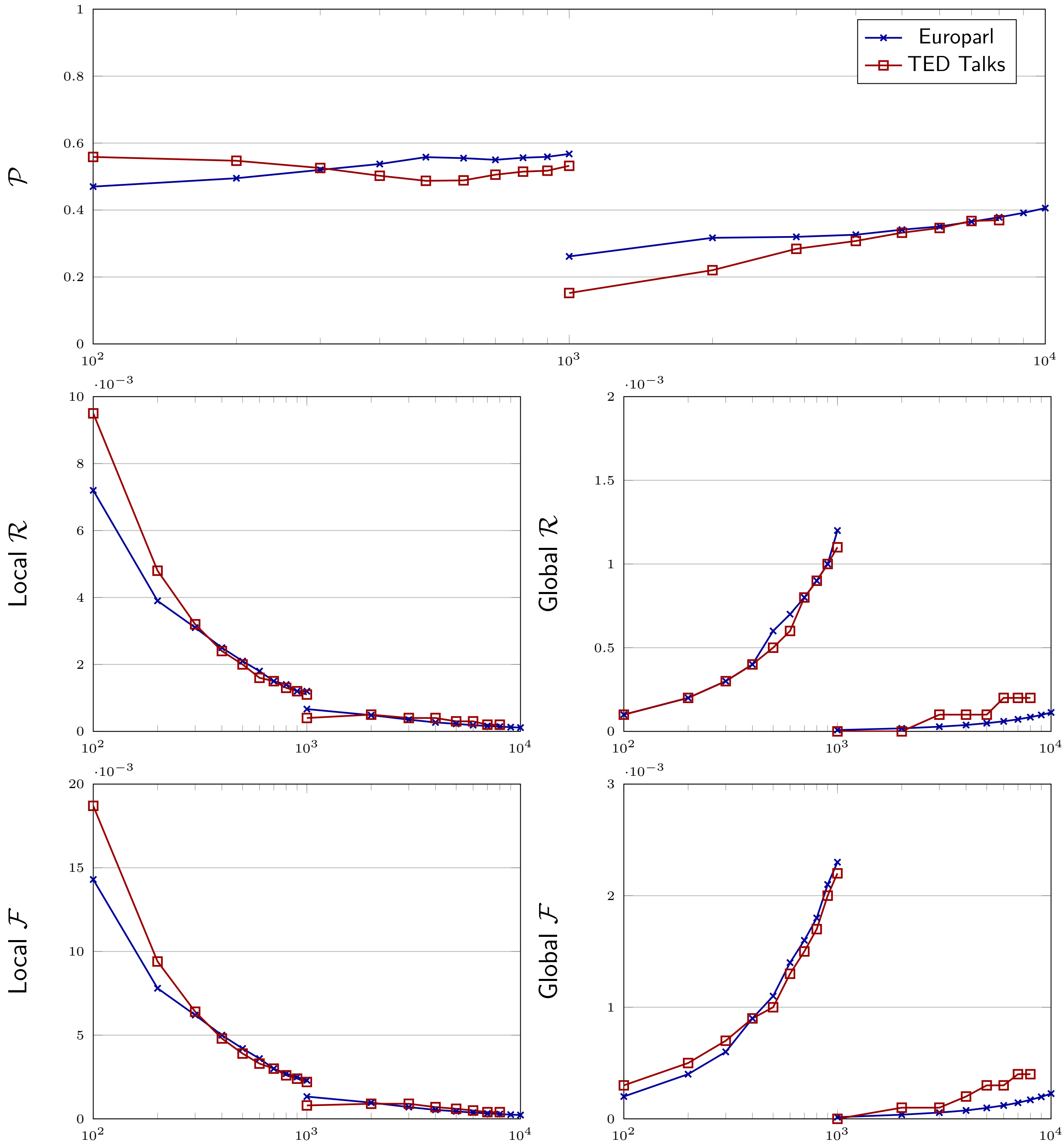


Figure 5.4: Precision, local and global recall, local and golbal F-measure of `HClust` for data sets in Portuguese using 1,000 terms and 10,000 terms.

The model using TED Talks corpus with the top 1,000 terms obtained the best local f-measure ($\mathcal{F}$=0.0174) in the 200th cluster, achieving a precision of $\mathcal{P}$=0.058 and a local recall of $\mathcal{R}_L$=0.0102 (the best local recall). Such high precision in the model was achieved again only in the 700th cluster, meaning that the first 200 clusters contain a great number of correct taxonomic relations.

Figure 5.4 shows the results of precision, local and global recall, and local and global f-measure for HClust models in Portuguese corpora. As clusters in the English corpora, the $x$ axis is also presented in logarithm scale ($\log_{10}$) for a better visualization. As discussed in Section 5.1, the values of precision achieved using Portuguese corpora are very high when comparing with the ones achieved by models with English corpora. Also, values in TED Talks end at the 8,000th cluster because the maximum number of terms that exist in the corpus and in the gold standard is 7,066.

As occurred in HClust using TED Talks in English, the model using the TED Talks in Portuguese also achieved higher values of precision in the first 200 clusters. In fact the behavior of the models is similar in both languages, discarding the difference in scores. In both languages, TED Talks grows the precision up to 200 clustes, and then the precision decreased in the next clusters and start to grow again close to the 600th cluster. In general, local recall tends to decrease as the number of clusters grow, meaning that the more terms in the model, the less the model can infer correct relations. As the precision and global recall grow after every new cluster inserted, the global f-measure also grows.

## 5.2 Results for the manual evaluation

In order to verify the quality of the relations extracted by each method, we also performed a manual evaluation. This type of evaluation is based on human judgements on whether a term is taxonomic related to another. In this evaluation a smaller subset of all relations is evaluated, as explained in Section 4.4. The methods evaluated here are based on patterns (Patt), based on document subsumption (DocSub), based on hierarchical clustering (HClust) and based on the head-modifier (HMod).

The subset of the relations is composed by 10 terms (seeds) with the highest $tf$-$dcf$ scores and their taxonomic related terms. As not all methods can generate taxonomies for the top 10 most relevant terms of the domain, we decided to evaluate 10 taxonomies generated by each method when the seeds occur up to the 500th position. For instance, the HClust model starts to generate at least 10 terms that belong to the top 500 domain terms, using 3,000 clusters. The complete list of terms with their positions according to the $tf$-$dcf$ scores is presented in Table 5.6.

Having the list of seed terms (Table 5.6) we extracted the relations to be evaluated by domain specialists. As it was expected, methods based on patterns (Patt) and based on head-modifier HMod contained the smallest taxonomies, while for the model based on document subsumption DocSub we had to limit the number of relations for some seeds. As described in Section 4.4.2, the evaluation is performed with 3 subjects for each corpus where each evaluator should follow the guidelines (see Appendix D). Thus, using an online form each subject should annotate relations answering the

| | # | Patterns | # | Head-modifier | # | Document Subsumption | # | Hierarchical clustering |
|---|---|---|---|---|---|---|---|---|
| Geology PT | 7 | matéria_orgânica | 1 | arenito | 1 | arenito | 8 | sedimentação |
| | 9 | grupo_itararé | 2 | bacia | 2 | bacia | 14 | sistema_deposicional |
| | 12 | bacia_do_paraná | 3 | granito | 3 | granito | 20 | hidrocarbonetos |
| | 25 | litofácies | 4 | fácies | 5 | feições | 22 | areias |
| | 29 | bacia_de_campos | 5 | feições | 7 | matéria_orgânica | 26 | granito_cinza |
| | 30 | feldspato | 7 | matéria_orgânica | 8 | sedimentação | 33 | planície |
| | 36 | óleo | 8 | sedimentação | 9 | litologia | 35 | lagoa |
| | 38 | eletrofácies | 10 | sedimentos | 11 | folhelhos | 50 | subsidência |
| | 42 | litologia | 12 | grupo_itararé | 12 | grupo_Itararé | 58 | dunas |
| | 57 | carbonatos | 13 | crosta | 13 | crosta | 76 | clasto |
| Geology EN | 2 | natural_gas | 2 | natural_gas | 1 | USGS | 22 | major_physical_features |
| | 4 | limestone | 4 | limestone | 2 | natural_gas | 28 | major_physical_features_of_state |
| | 6 | quartz | 5 | US_geological_survey | 3 | interstate | 29 | elevation_trends |
| | 8 | shale | 6 | quartz | 4 | limestone | 31 | generalized_topographic_map |
| | 9 | gemstones | 7 | eruptions | 6 | quartz | 90 | detailed_map |
| | 12 | marcellus_shale | 8 | shale | 7 | eruptions | 95 | county_boundaries_on_map |
| | 17 | fossils | 9 | gemstones | 8 | shale | 102 | major_streams |
| | 34 | feldspar | 10 | diamonds | 9 | gemstones | 149 | clay_minerals |
| | 36 | sandstone | 11 | magma | 10 | diamonds | 152 | collision |
| | 37 | utica_shale | 12 | marcellus_shale | 11 | magma | 182 | erupting |
| Europarl PT | 1 | conselho | 5 | sector | 1 | conselho | 8 | direitos_humanos |
| | 2 | parlamento | 7 | directiva | 2 | parlamento | 12 | acção |
| | 5 | sector | 16 | projecto | 3 | estado_membro | 18 | deputado |
| | 6 | presidência | 42 | orador | 5 | sector | 23 | controlo |
| | 13 | tratado | 65 | mandato | 6 | presidência | 31 | imigração |
| | 14 | comissão | 76 | contribuinte | 7 | directiva | 34 | despesas |
| | 16 | projecto | 94 | imigrante | 8 | direitos_humanos | 37 | tribunais |
| | 38 | presidente | 95 | referendo | 11 | democracia | 41 | mercado_interno |
| | 53 | obrigação | 97 | debate | 12 | acção | 43 | política_agrícola |
| | 55 | cimeira | 104 | contributo | 13 | tratado | 47 | proposta_de_resolução |
| Europarl EN | 1 | Mr_president | 1 | Mr_president | 1 | Mr_president | 33 | european_people |
| | 2 | member_states | 2 | member_states | 2 | member_states | 44 | european_people_party |
| | 3 | rapporteur | 3 | rapporteur | 3 | rapporteur | 86 | european_people_party_christian_democrats |
| | 4 | Madam_president | 4 | Madam_president | 5 | european_parliament | 87 | growth_pact |
| | 5 | european_parliament | 5 | european_parliament | 6 | member_state | 100 | group_of_european_people_party_christian_democrats |
| | 6 | member_state | 6 | member_state | 7 | amendments | 284 | FI_Mr_president |
| | 7 | amendments | 7 | amendments | 8 | internal_market | 285 | mass_destruction |
| | 8 | internal_market | 8 | internal_market | 9 | third_countries | 315 | constitutional_affairs |
| | 9 | third_countries | 9 | third_countries | 10 | commission_proposal | 357 | party_of_european_socialists |
| | 11 | president-in-office | 10 | commission_proposal | 11 | president-in-office | 427 | joint_motion |

Table 5.6: Lists of terms for manual evaluation

question "Is $\mathtt{w}_1$ a (kind of/form of) $\mathtt{w}_2$?", where $\mathtt{w}_1$ and $\mathtt{w}_2$ are two words in a relation. The valid answers were "Yes" (the relation is correct), "No" (the relation is wrong) and "Not applicable" (a term is ill-formed or does not belong to the domain). In order to verify how reliable the annotations are as well as the difficulty of this task we compute the degree of inter-annotator agreement among the evaluators using the Fleiss' kappa coefficient $\kappa$ [36] which is a generalization of Scott's kappa for more than two annotators. According to Landis and Koch [73], the relative strength of agreement associated with kappa statistics can be understood by the following labels:

| $\kappa$ | Strength of agreement |
|---|---|
| $< 0$ | Poor agreement |
| 0.01 - 0.20 | Slight agreement |
| 0.21 - 0.40 | Fair agreement |
| 0.41 - 0.60 | Moderate agreement |
| 0.61 - 0.80 | Substantial agreement |
| 0.81 - 1.00 | Almost perfect agreement |

We measure the inter-annotator agreement for each method and for each corpus. For a better understanding on how difficult the task of evaluation is, we also measured the percentage of equal votes, *i.e.*, when all three evaluators give the same vote ("Yes", "No" or "Not applicable"). Table 5.7 presents the values of Fleiss' kappa ($\kappa$) and the percentage of equal votes (Votes) for each model using each corpus, where AllM contains the values for all methods together and AllC contains the values for all corpora together. Thus, the overall values of $\kappa$ and the general percentage of equal votes are obtained by the conjuction of AllM with AllC. It is important to note that due to the unbalanced number of subjects in the evaluations, values of AllM and AllC do not represent the arithmetic mean of the scores.

| Language | Corpus | | Patt | HMod | DocSub | HClust | AllM |
|---|---|---|---|---|---|---|---|
| EN | Europarl | $\kappa$: | 0.73 | 0.39 | 0.93 | 0.62 | 0.68 |
| | | Votes: | 80% | 65% | 96% | 74% | 77% |
| | Geology | $\kappa$: | 0.90 | 0.78 | 0.59 | -0.04 | 0.75 |
| | | Votes: | 94% | 91% | 82% | 37% | 80% |
| PT | Europarl | $\kappa$: | 0.50 | 1.00 | 0.87 | 0.72 | 0.87 |
| | | Votes: | 70% | 100% | 91% | 81% | 91% |
| | Geology | $\kappa$: | 0.11 | 0.17 | 0.18 | 0.11 | 0.44 |
| | | Votes: | 33% | 90% | 65% | 41% | 61% |
| AllC | | $\kappa$: | 0.73 | 0.57 | 0.69 | 0.45 | 0.72 |
| | | Votes: | 80% | 85% | 82% | 60% | 78% |

Table 5.7: Fleiss' Kappa inter-annotator agreement and percentage of equal votes.

The overall agreement is substantial ($\kappa$=0.72) with a high percentage of equal votes (78%). Votes for relations generated by HClust using the Geology corpus in English obtained a poor agreement ($\kappa$=-0.04) and a low percentage of equal votes (37%). Analyzing the relations generated by the model we observe that most relations (60%) contain at least one vote for "Not applicable", and none of them contains 3 equal votes. In the case of disagreement between annotators we selected the relation annotated by the majority (there were no three-way ties in the evaluation). Among terms that received two "Not applicable" votes are "Geologist Information Geology Jobs", "Visit Geology" and "MyTopo affiliates". These ill-formed terms come from the problem of extracting terms from HTML pages, where terms that belong to a list of links when transformed into plain text files stand side-by-side, for example. This is the case of the first term "Geologist Information Geology Jobs" that were links to "Geologist Information" and to "Geology Jobs". The other corpora that were not extracted from HTML files contain very few ill-formed terms.

Relations generated using the Geology corpus in Portuguese achieved a moderated agreement ($\kappa$=0.44). When analyzing the evaluations we observe that there was almost no term evaluated as "Not applicable", showing how hard is to evaluate domain terms. There is only an exception for relations containing the term "origem diversa" (diverse origin) that was marked by one of the evaluators as "Not applicable". When asked about the evaluation, one of the subjects told that it was hard to evaluate these types of relations even for him that works in the field, and in some cases he had to search about terms on the internet.

Table 5.8 presents the values of precision for each method. This ratio is obtained by dividing the number of correct relations by the total evaluated relations. Line "Both" considers the corpus gathering all documents of the language, AllM is the precision considering the relations of all methods in the same corpus, AllC represents the precision for all relations considering only the method. The overall precision is obtained by the conjunction of AllM with AllC.

| Language | Corpus | Patt | HMod | DocSub | HClust | AllM |
|---|---|---|---|---|---|---|
| EN | Europarl | 0.51 | 0.83 | 0.22 | 0.40 | 0.50 |
| | Geology | 0.77 | 0.84 | 0.09 | 0.05 | 0.37 |
| | Both | 0.66 | 0.83 | 0.13 | 0.30 | 0.43 |
| PT | Europarl | 0.28 | 1.00 | 0.44 | 0.64 | 0.63 |
| | Geology | 0.53 | 0.98 | 0.09 | 0.22 | 0.27 |
| | Both | 0.35 | 0.99 | 0.27 | 0.32 | 0.47 |
| AllC | | 0.57 | 0.90 | 0.20 | 0.30 | 0.45 |

Table 5.8: Precision obtained by manual evaluation for each model in each corpus.

Results show that the model based on head-modifier (HMod) achieve the best precision ($\mathcal{P}$=0.90) over all corpora. Although this method was not discussed in the automatic evaluation due to the low number of noun phrases in the gold standards, HMod has a very low recall, being very limited since the relation between terms consider only the head of the terms and its modifiers. Thus, terms composed by different words can not be detected by this method. Observing the evaluations, terms that are assigned as not taxonomically related include "Loud applause Madam President" as hyponym of "Madam President" and "Votes Mr. President" as hyponym of "Mr. President".

The model based on patterns (Patt), which is well known for also having a high precision, achieved lower precision in the Portuguese corpora ($\mathcal{P}$=0.28 in Europarl and $\mathcal{P}$=0.53 in Geology). Analyzing the evaluations, we observe that many relations extracted with the pattern "NP como LNP" (line 2 in Table 2.1) were incorrect. For example, consider the phrases and the relations extracted:

Sentence 1: A proibição de pesca foi imposta por a Comissão como castigo por a Polônia exceder a quota anual de capturas de bacalhau.
Relation: <castigo, is-a, Comissão>

Sentence 2: Creio que enfraquecemos a nossa credibilidade como Parlamento
Relation: <Parlamento, is-a, credibilidade>

Sentence 3: Estas piadas são degradantes para o Parlamento como instituição e para a União Europeia enquanto processo político.
Relation: <instituição, is-a, Parlamento>

Sentence 4: O Reino Unido trabalhou incansável e literalmente tanto a nível nacional como em outras instâncias que representavam a União Europeia como Presidência.
Relation: <União Europeia, is-a, Presidência>

Although the pattern "NP como LNP" may extract correct relations, such as in the phrases "desejo-lhe uma longa e distinta carreira como deputado." where "deputado" (representative) is a kind of "carreira" (career), and "Quatrocentas mil crianças equivalem a uma cidade como Estrasburgo ou a uma cidade maior do que Granada" where "Estrasburgo" (Strasbourg) is a kind of "cidade" (city), it also may induce to errors.

Analyzing the English versions of the same phrases, we have:

Sentence 1: The fishing ban was imposed by the Commission as a punishment for exceeding the annual quota for cod catches.

Sentence 2: I believe that we have undermined our credibility as a parliament.

Sentence 3: Such jibes degrade Parliament as an institution and the European Union as a political process.

Sentence 4: United Kingdom has worked tirelessly and literally to the last minute both nationally and in other fora representing the European Union as Presidency.

where the pattern "NP as a LNP" is not an evidence for taxonomic relation in English. Thus, many of the wrong relations are not extracted. On the other hand, the relation <representative, is-a, career> is not extracted from the phrase "I wish him a long and distinguished career as a representative.". Observing the corpora in both languages, most of the correct relations extracted with the pattern "NP como LNP" in Portuguese are expressed with the pattern "NP such as LNP" in English, as presented in the phrases: "A investigação tem demonstrado que são outros os motivos que justificam o aparecimento de valores elevados de ozono em cidades como Barcelona, Atenas ou Milão e em muitas zonas turísticas costeiras." and "Research shows that the reasons why ozone affects cities such as Barcelona, Athens and Milan and many of the tourist resorts along the coast must be sought elsewhere."

Models based on document subsumption (DocSub) and hierarchical clustering (HClust) obtained low values of precision for most of the corpora. Hierarchical clustering obtained good results in Europarl corpora. Analyzing the results obtained by HClust, we observe a great number of relations where one term contains a inner noun phrase of the other term, *i.e.*, one term is a substring of the other (*e.g.*, <Democratic People Republic of Korea, is-a, Democratic People Republic> and <European Institute of Technology, is-a, European Institute>). This is almost the same principle used in head-modifier detection. This type of relations were privileged not only by the clustering process, but also by the document borders (set to the size of the phrase, since Europarl does not have document border). As the number of documents increased, any difference in terms would make a term belong to the cluster. As the Geology corpora have a limited number of documents, a variation of the term must occur in other documents instead of in a phrase. For example, in order to find the relation <european parliament, is-a, parliament> the Europarl corpora should have phrases "The european parliament has the opportunity to discuss the directives." and "The parliament has given its consent.", while the Geology corpora should have them in different documents instead of phrases.

On the other hand, unlike the head-modifier model, the HClust model using a small document border may generate erroneous relations based on noun phrases that include preposition. For example, considering that the term "child" occurs in more documents that the term "rights of child", the erroneous relation <rights of child, is-a, child> is extracted, instead of the correct relations <rights of child, is-a, right>. Analyzing separately the results generated by HClust using the Europarl corpus in English, the precision for relations containing inner noun phrase between the terms achieved $\mathcal{P}$=0.38, while the precision for relations without occuring inner noun phrases achieved $\mathcal{P}$=0.52. Examples of wrong extracted relations include <Council of Agriculture Ministers, is-a, Agriculture Ministers> and <weapons of mass destruction, is-a, mass destruction>. Examples of correct extracted relations without inner noun phrases include <Xenophobia, is-a, Racism> and <wheeled agricultural, is-a, forestry tractors>.

Analyzing the relations obtained by HClust using the Europarl in Portuguese which achieved the precision $\mathcal{P}$=0.64 we observed the same problem that occurred in English. For example, the relations <imunidade de Deputado, is-a, Deputado> and <matéria de Direitos humanos, is-a, Direitos humanos> do not have a correct taxonomic relation. On the other hand, the number of evaluated relations containing preposition were much smaller. Although the majority of the extracted relations contain inner noun phrases, they do not contain preposition (*e.g.*, <Acção externa, is-a, Acção> and <Deputado europeu, is-a, Deputado>). Thus, when verifying the precision of the relations that contain inner noun phrases, it increases to $\mathcal{P}$=0.86, while the precision for terms without inner noun phrases decreases to $\mathcal{P}$=0.14.

Models based on document subsumption (DocSub) also obtained low values of precision. As HClust models, such models also detected inner noun phrases and due to the high threshold ($\lambda$=0.8) the model using the Europarl corpora generated almost all relations based on inner noun phrases. Using the English version of Europarl the model also generated wrong relations when a preposition appeared, and according to the results, the model also generated wrong relations when there was no preposition. For example, the wrong relation <Member States budgets, is-a, Member States> are generated since "Member States" is more frequent than "Member States budgets" and the latter is present in at least 80% of the documents in which the former appears.

## 5.3 Metrics Analysis

Another way to see learned taxonomies is on the basis of their characteristics. The characteristics can be translated into a group of metrics that were initially developed for ontology evaluation [38,39]. In this thesis we adapted all metrics that can be used for taxonomies and applied them in all learned taxonomies in order to see what types of taxonomies each model generates. For the sake of space, we selected some of all developed metrics to analyze in detail. The selected metrics are based on what we think would be interesting to analyze in each taxonomy. A description of the selected metrics is presented in Section 4.4.3, and the full list containing all metrics is presented in Appendix E.

Analyzing values generated in each metric, specifically the maximum depth of models that use the statistics between words to generate the taxonomy (DSim, SLQS, TF, *etc.*), we observe that the distance between a root term, *i.e.*, a term that does not have a parent term, and a leaf term, *i.e.*, a terms that does not have a child term, was equal to 1, meaning that a direct edge connects the root term and the leaf term. For example, consider the digraph $D_1$ in Figure 5.5, where all relations between terms are shared using only one edge (*e.g.*, the distance between the root term A and the leaf term F is equal to 1 because there is a direct edge connecting them).

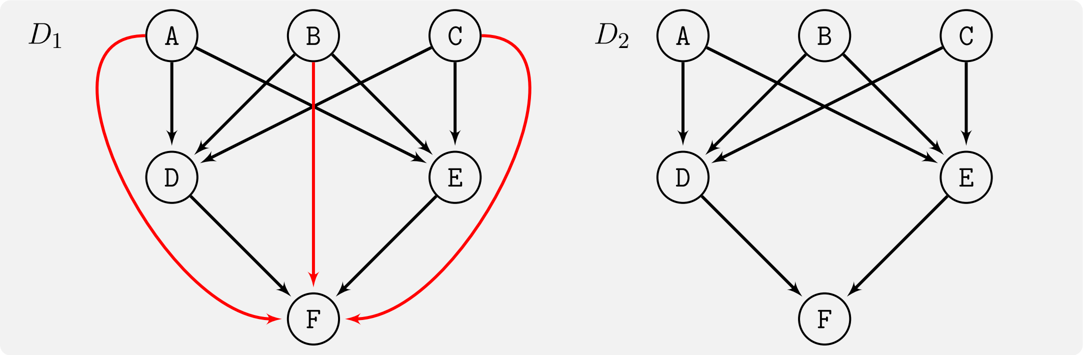


Figure 5.5: Transitive reduction from digraph D1 to D2.

To calculate the maximum distance between a root term and a leaf term we decided to remove edges using the transitive reduction of a finite directed graph. The transitive reduction generates the minimum equivalent graph, *i.e.*, a graph with the fewest possible edges that has the same reachability relation as the original graph. Thus, the minimum graph must contain all paths, but not necessarily all edges, between nodes as the original graph. For example, the digraph $D_2$ presented in Figure 5.5 shows the transitive reduction of the original digraph $D_1$, where the edges connecting A→F, B→F and C→F are removed, and all paths between terms are kept. For instance, the direct edge connecting terms A and F is removed, but the path between them passes through D or E (path

A→D→F or A→E→F), as well as the path between C and F passes through D or E (path C→D→F or C→E→F). Besides reducing the number of relations, removing transitivity creates a structure as it is usually presented in semantic relation networks.

Using the direct graph with transitive reduction for each model, we generated the metrics presented in Table 5.9. The results are generated for models using the top 1,000 tems of each corpus in English generated for the automatic evaluation, *i.e.*, models using nouns instead of noun phrases. It is also important to notice that as we generated several taxonomies in DocSub with different thresholds, and several taxonomies in HClust with different numbers of clusters, we selected the ones that achieved the highest f-measure value. Thus, the taxonomy generated by DocSub model using the TED Talks corpus is obtained with the threshold $\lambda$=0.3 and the taxonomy generated using the Europarl corpus corresponds to threshold $\lambda$=0.1). For the HClust models, the taxonomy contains 1,000 terms and the maximum number of clusters. Patt model contains relations of all patterns for the limited number of terms.

| Corpus | Metric | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | TotalTerms: | 957 | 1,000 | 1,000 | 1,000 | 1,000 | 836 | 1,000 |
| | TotalRoots: | 44 | 1 | 1 | 1 | 1 | 43 | 1 |
| | NumberRels: | 1,588 | 1,025 | 1,028 | 1,185 | 1,103 | 1,184 | 999 |
| | MaxDepth: | 21 | 921 | 901 | 788 | 835 | 8 | 15 |
| | MinDepth: | 1 | 921 | 901 | 788 | 835 | 1 | 1 |
| | AvgDepth: | 11.82 | 921 | 901 | 788 | 835 | 3.05 | 8.46 |
| | DepthCohesion: | 1.78 | 1 | 1 | 1 | 1 | 2.62 | 1.77 |
| | MaxWidth: | 20 | 2 | 3 | 4 | 3 | 88 | 41 |
| | MinWidth: | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | AvgWidth: | 1.99 | 1.03 | 1.03 | 1.19 | 1.10 | 4.20 | 2.38 |
| TED Talks | TotalTerms: | 476 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| | TotalRoots: | 164 | 2 | 1 | 1 | 1 | 1 | 1 |
| | NumberRels: | 521 | 1,029 | 1,331 | 3,025 | 3,438 | 3,802 | 1,009 |
| | MaxDepth: | 16 | 915 | 658 | 454 | 395 | 118 | 12 |
| | MinDepth: | 1 | 913 | 658 | 454 | 395 | 110 | 1 |
| | AvgDepth: | 5.82 | 914 | 658 | 454 | 395 | 112.24 | 5.95 |
| | DepthCohesion: | 2.75 | 1 | 1 | 1 | 1 | 1.05 | 2.02 |
| | MaxWidth: | 25 | 2 | 77 | 13 | 12 | 66 | 98 |
| | MinWidth: | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | AvgWidth: | 1.83 | 1.03 | 1.36 | 3.03 | 3.44 | 6.64 | 2.35 |

Table 5.9: Metrics for taxonomies generated by models using the top 1,000 terms of each corpus in English.

As we can observe in Table 5.9, limiting the number of terms to 1,000, Patt and DocSub do not generate relations for all terms (number of TotalTerms inferior to 1.000). Patt model could not generate relations for all terms because terms must be in a pattern in order to have their taxonomic relation identified. As explained before, patterns are very sparce and not all terms would fit in a pattern. The size of the corpus also impacts the number of terms that belong to a relation, thus, the larger the corpus, the easier to find terms that match a pattern. For DocSub model, the limited threshold may have influenced in results. Some terms would not have a relation because our threshold expects that two terms must share at least 30% of the documents to exist a taxonomic relation between them.

An interesting characteristic of DSim, SLQS, TF and DF models is the high number of generated relations as explained before, where almost all terms are connected before applying the transitive reduction. When the transitive reduction is applied on these models, they transform the taxonomy into a deep taxonomy, where small differences may indicate a taxonomic relation. For example, using relations generated by TF model using the Europarl corpus, we can understand the MaxDepth as having 789 terms with different values of term frequency, while having 211 that share the same value of term frequency with other terms. Using the SLQS model, we can understand that 902 terms share different values of entropy, while 98 share the same value with other terms. For such models, the number of generated relation is very high before applying the transitive reduction (*e.g.*, DSim contained a total of 499,101 relations before applying the transitive reduction, meaning that almost all terms are interconnected by taxonomic relations).

The maximum width (MaxWidth) for these models can be understood as the number of terms that share the same characteristic. For instance, the MaxWidth for TF model using the Europarl corpus is equal to 4, meaning that there are at least one term that share its value of frequency with other 4 terms. Comparing the values of both corpora we observe that a small corpus tends to have less difference in these characteristics. Thus, in a small corpus the maximum depth (MaxDepth) tends to be smaller and the maximum width (MaxWidth) tends to be greater when comparing with a larger corpus.

As hierarchical clustering (HClust) groups semantically similar terms before verifying the taxonomic relation between them, then the number of relations (NumberRels) tends to be smaller than the number of relations generated by DF model. Such clustering also transforms the generated taxonomy into a dense taxonomy since the number of maximum depth and width are closer than the values presented by DF.

Table 5.10 contains the results for each metric after applying the transitive reduction for relations generated by models using the top 1,000 terms of each corpus in Portuguese. As DocSub generated a range of taxonomies with differente threshold ($\lambda$) values, the results presented use the taxonomies generated with the highest value of f-measure ($\lambda$=0.1 for both taxonomies). For HClust model the taxonomies contain all 1,000 terms, since both models achieved the highest f-measure with the maximum number of clusters. The Patt models contain all patterns but a limited number of terms.

The results for the Portuguese corpora are quite similar to the ones generated by the English corpora, having terms without relations in Patt and DocSub, and DSim, SLQS, TF and DF generating deep taxonomies, affirming the characteristics of each method. For Portuguese, the number of relations found by Patt model using the TED Talks corpus were smaller than the one found using the English corpus, impacting the maximum depth. But, the number of siblings for a term was greater.

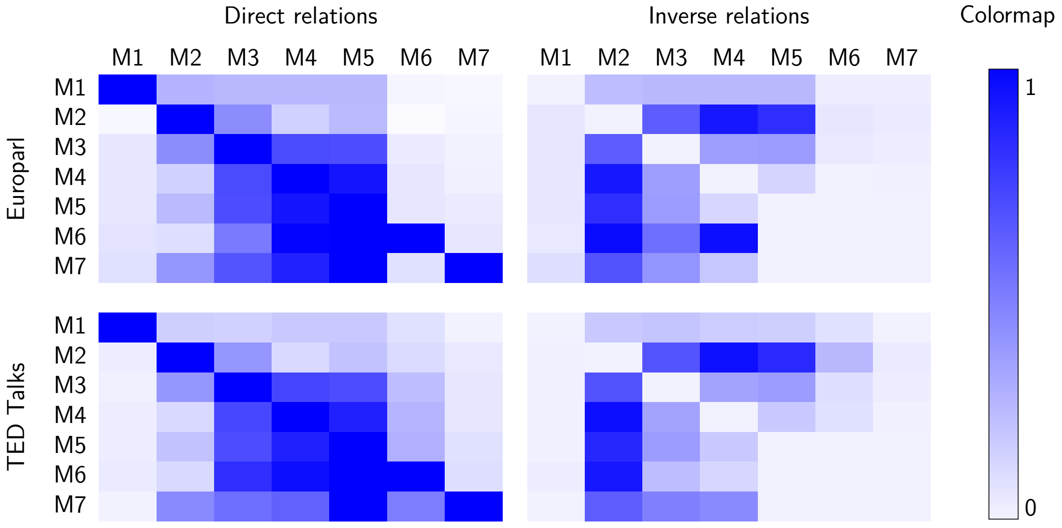
| Corpus | Metric | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | TotalTerms: | 980 | 1,000 | 1,000 | 1,000 | 1,000 | 996 | 1,000 |
| | TotalRoots: | 79 | 1 | 1 | 1 | 1 | 1 | 1 |
| | NumberRels: | 1,527 | 1,031 | 1,049 | 1,185 | 1,093 | 1,644 | 999 |
| | MaxDepth: | 19 | 902 | 894 | 784 | 849 | 6 | 10 |
| | MinDepth: | 1 | 902 | 894 | 784 | 849 | 1 | 1 |
| | AvgDepth: | 9.43 | 902 | 894 | 784 | 849 | 2.73 | 4.29 |
| | DepthCohesion: | 2.02 | 1 | 1 | 1 | 1 | 2.19 | 2.33 |
| | MaxWidth: | 27 | 3 | 3 | 4 | 3 | 201 | 58 |
| | MinWidth: | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | AvgWidth: | 1.98 | 1.03 | 1.05 | 1.19 | 1.09 | 6.25 | 2.55 |
| TED Talks | TotalTerms: | 296 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| | TotalRoots: | 101 | 1 | 1 | 1 | 1 | 1 | 1 |
| | NumberRels: | 291 | 1,045 | 1,229 | 3,637 | 4,284 | 2,875 | 999 |
| | MaxDepth: | 10 | 860 | 727 | 388 | 354 | 252 | 17 |
| | MinDepth: | 1 | 860 | 727 | 388 | 354 | 249 | 1 |
| | AvgDepth: | 3.94 | 860 | 727 | 388 | 354 | 250.43 | 6.16 |
| | DepthCohesion: | 2.54 | 1 | 1 | 1 | 1 | 1.01 | 2.76 |
| | MaxWidth: | 37 | 3 | 79 | 18 | 13 | 9 | 41 |
| | MinWidth: | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | AvgWidth: | 1.79 | 1.05 | 1.23 | 3.64 | 4.29 | 2.94 | 2.37 |

Table 5.10: Metrics for taxonomies generated by models using the top 1,000 terms of each corpus in Portuguese.

## 5.4 Complementarity Analysis

Observing all generated taxonomies, it is interesting to analyze how models are complementary, *i.e.*, verify whether the same relations are generated by more than one model. It is also interesting to observe whether relations are generated in opposite directions (*e.g.*, `Patt` model generates the relation `A`→`B` and `DSim` model generates the inverse relation `B`→`A`). To illustrate the effect of mixing models, Figure 5.6 presents a color map with the ratios of direct relations, *i.e.* relations that are equal in both models, and inverse relations. All ratios are generated by models using the top 1,000 words of the English corpora. The ratio is computed as the number of taxonomic relations shared by the models in the row and the model in the column divided by the number of relations generated by the model in the row. For example, `Patt` model using Europarl corpus in English generated a total of 15,797 taxonomic relations from which 4,014 were shared with `DSim`. Thus, the value of the (`Patt`, `DSim`) cell is ($\frac{4{,}014}{15{,}797}$)=0.2541. It is important to notice that the ratio takes into account the order of the relation, thus, the cell (`M5`, `M7`) is not equal to the cell (`M7`, `M5`). Each matrix element is indicated with a color, going chromatically from a dark blue for higher values (starting with 1 and found in the diagonal of the direct matrix) to a light blue for lower values (ending with 0 and found in the diagonals of the inverse matrix). All values that generated the color map are presented in Appendix F.

Such a colorful representation let us observe that some models share a high number of relations, while other models share a high number of inverse relations. Models that share low number of direct relations and low number of inverse relations tend to be complementary, since relations in one model



**M1**: Patt **M2**: DSim **M3**: SLQS **M4**: TF **M5**: DF **M6**: DocSub **M7**: HClust

Figure 5.6: Ratios for relations shared by models in the English corpora.

are not generated in the other model and vice versa. Observing the patterns generated by colors in 5.6, the common point in both corpora is the fact that most models share a high number of relations with SLQS (M3), TF (M4) and DF (M5) and low number of relations with Patt (M1), DocSub (M6) and HClust (M7). Also, a high number of inverse relations are shared with DSim (M2).

Although the number of relations shared with Patt are very low (close to zero) for DSim, SLQS, TF and DF models, the number of shared relations for Patt model is almost 20% for each model. The number of shared relations with Patt by DocSub and HClust is very low in both ways. Thus, these models tend to be complementar to the Patt model. This relation between models is more explicit in the Europarl corpus.

Almost 60% of the relations in SLQS model are shared directly with TF and DF and almost 30% of its relations are shared inversely with these models. It means that these models generated relations for almost the same terms, and only few relations between terms were not discovered. This can be seen by the number of relations found by each model (494,871 for SLQS, 498,222 for TF and 497,905 for DF using the TED Talks corpus). As term frequency and context frequency are somehow related, using the top 1,000 terms that share most contexts will create relations for almost all terms.

HClust model have a high number of similar relations with DF and TF models. It totally makes sense for relations shared with DF because HClust is driven by relations generated by this model, *i.e.*, HClust keep only DF relations that are semantically related. The high number of shared relations with TF, mainly in the Europarl corpus, is because DF moves towards TF when the size of the documents decrease. As HClust is driven by DF, it is also affected by the size of the document and consequently it approaches TF.

Figure 5.7 shows the color map for models using the Portuguese corpora to generate taxonomic relations. As occured when analysing metrics, the results generated by these models are quite similar

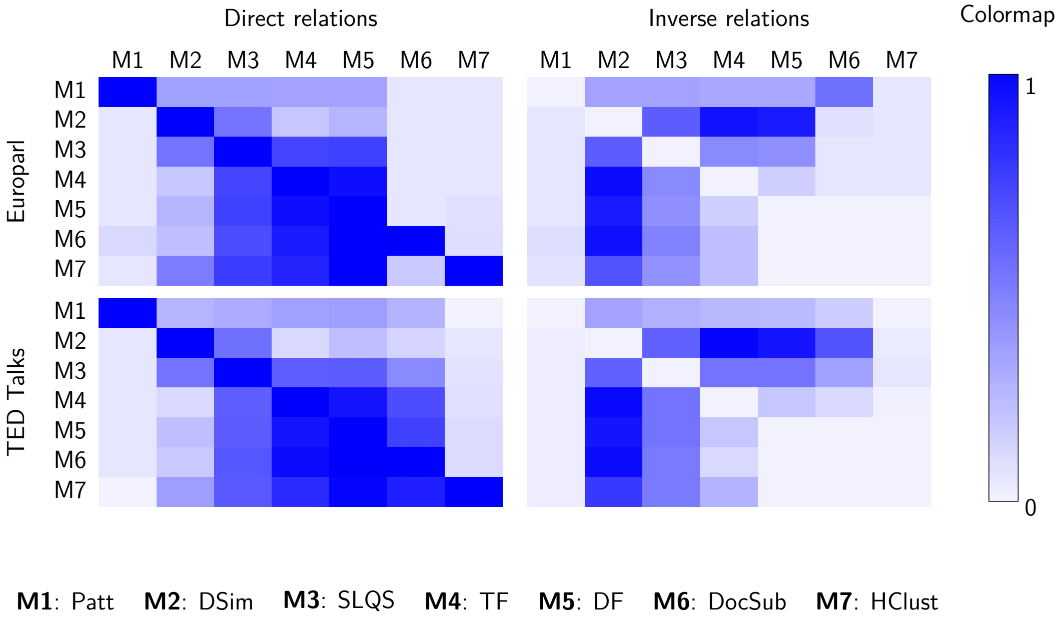


Figure 5.7: Ratio of relations shared by models in the Portuguese corpora.

to the results generated using the English corpora. The most notable difference between English and Portuguese results rely on the high number of inverse relations generated by TF when the source is DocSub, as presented in cell (M6, M4), when using the Europarl corpus in English. Although values in cells are quite similar for the most results, the value in cell (M6, M4) using the Europarl in Portuguese is very low when comparing with the results in English.

Minor differences may be also observed in Patt model (M1) which has more shared relations with other models in English than in Portuguese. On the other hand, the inverse relation of this model with DocSub model (M6) using Europarl in Portuguese is higher when compared with the relation using the English version of the corpus.

Observing the relations shared between models, we believe that it would be also interesting to verify the impact in precision when combining models. The relative precision ($\mathcal{P}_R$) indicates what is the gain or loss of the model when using only terms that are shared by another model. Thus, we calculate the relative precision as the precision of the relations in the intersection of the models divided by the precision of the original model. For instance, consider that the precision in the original Patt model containing the top 1,000 terms is equal to $\mathcal{P}$=0.53, and the precision taking into account only terms shared with DocSub model is equal to $\mathcal{P}$=0.65. The relative precision is the ratio between both values $\mathcal{P}_R$=1.2. Values above 1 of relative precision indicate that the precision of the relations in the intersection of the models is higher than the precision of the original model. Values between 0 and 1 indicate that the intersection between models affect negatively the original model.

Figure 5.8 presents a color map with the relative precisions achieved by each intersection of the model in each corpus for English and Portuguese. All $\mathcal{P}_R$ are generated by models using the top 1,000 word of each corpus. It is important to notice that the relations of the original model are the

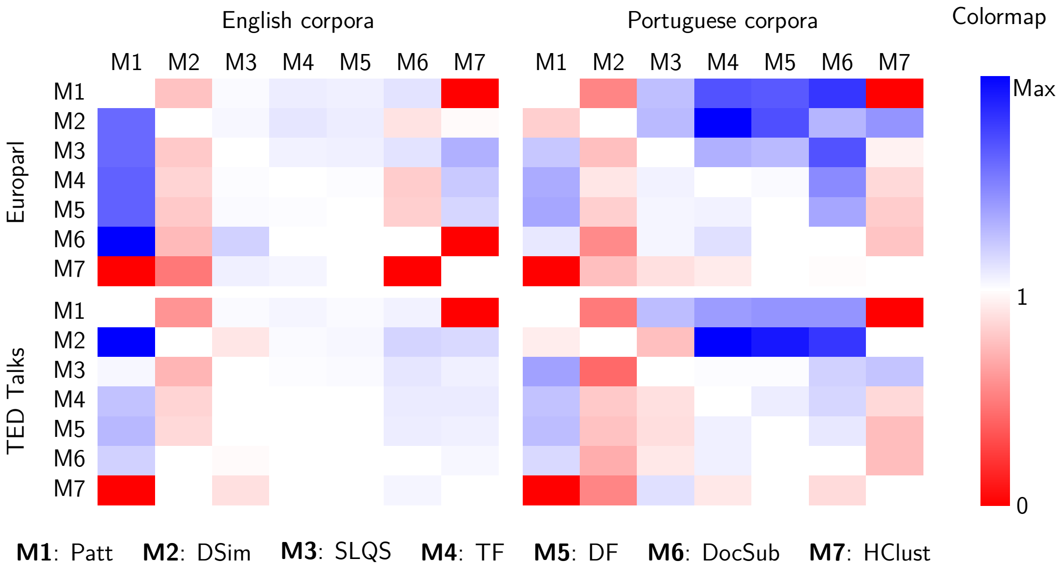


Figure 5.8: Relative precision using the intersection of the models.

rows of the matrix and the models used in the intersection are the columns. Thus, the cell (M5, M7) contains the precision of the relations in the intersection divided by the precision of the original model M5. Each matrix element is indicated with a color, going chromatically from a dark blue for higher relative precision values, passing through white for neutral value of relative precision (values close to 1), to a dark red for lower values of relative precision (values close to zero). Thus, if the cell contains a shade of blue the intersection achieves a higher precision when compared to the original and if it contains a shade of red the precision in the intersection is lower than the original. Values that generated the color map are presented in Appendix F.

As observed in the previous experiment, the number of shared relations between SLQS, TF and DF is very high. Having this high number of shared relations, the intersection of these models do not significantly change the precision when compared to the their original scores. The precision increased for most methods when they are combined with Patt. The unique exception is the HClust model which decreased the precision. This low value of precision is due to the fact that there are no relations in the intersection of both methods. As the intersection is equal to zero, the value of precision of the intersection is also equal to zero.

Most methods also increased the precision using the intersection with the DocSub model. On the other hand, most methods decreased when combined with DSim model. Methods using corpora in English tend to increase the precision when combined with HClust, meaning that it seems interesting to generate taxonomic relations only for semantically related terms. DSim model benefits from the filtering of relations when combining almost all other models. This benefit might be due to the low value of precision that DSim achieved, and thus, when filtering relations that are incorrect, the model increase the precision. Observing generally all results, Patt and DocSub are models that may serve as filter in order to increase the precision score. On the other hand, when a model is combined with DSim the values of precision tend to decrease.

# 6. Conclusions and further work

This thesis presented a detailed comparative evaluation of different methods for automatic taxonomic relation extraction from text corpus, considering variations in genre and language. This chapter summarises the work presented in previous chapters, as well as our major conclusions, and describes the current and future directions of our research. It is organised in three sections: thesis overview (Section 6.1), main results (Section 6.2) and directions for further work (Section 6.3).

## 6.1 Thesis overview

We started our work by presenting its motivations in Chapter 1, rising questions that we would like to answer in this thesis. Our four main questions were: is there a method that outperforms all other methods? If changing the language, do the methods perform equally? Do all methods generate similar taxonomies? Are results generated by different methods complementary or dissimilar? In order to answer these questions, we developed, evaluated and characterized in terms of hierarchy metrics a set of methods that are the state of the art according to the literature in the area.

As presented in Chapter 2, ontology learning from texts is an active research field that has produced over the last decades a large body of proposals covering the extraction process under different perspectives. Approaches were presented following a broad classification on methods. While methods that use little or no supervised algorithms require none or a small amount of examples to identify taxonomic relations in texts, supervised algorithms claim for training data sets. Methods used in the state of the art were presented and their pros and cons according to the literature were discussed.

Manual and automatic evaluation strategies are discussed in Chapter 3. The manual evaluation by domain experts is highly dependent on the view that the specialist has on the domain, as well as being a time consuming task, and the automatic evaluation uses classical measures such as Precision, Recall and F-measure, or alternative measures as taxonomic overlap, which requires the presence of a gold standard. The chapter also discusses the difficulties in reproducing the results reported in the literature due to the use of different corpora or gold standards. While some proposals report improvements on previous approaches, others apply similar strategies on different corpora, making hard to draw a comprehensive comparison between them.

The methodology used in this thesis for developing and evaluating models that extract taxonomic relations from text corpora in Portuguese and English is described in Chapter 4. The chapter also presents all the resources used during the evaluation process to build models, such as testing corpora and contrasting corpora, as well as the resources to evaluate them, such as gold standards. The methodology described includes the preprocessing of the corpora and how models for automatic and manual evaluation are constructed. The design of the automatic and manual evaluation, as well as guidelines presented to each evaluator are detailed.

Chapter 5 examines characteristics of the methods to extract taxonomic relations from text corpora by using automatic and manual evaluations. The characteristics can be measured in terms of precision, recall and f-measure using the automatic evaluation and the quality of the relations when compared with human judgement in a manual evaluation. The learned taxonomies also are characterized in terms of hierarchy metrics such as depth, width *etc.*. The intersections of the relations generated by the methods are analysed in terms of complementarity or similarity. Thus a method is complementar to another when relations of one method are not generated by the other, and similar when the same relations are generated by both methods.

## 6.2 Main results

The major contribution of this thesis is a better understanding of the different methods for automatic taxonomy contruction from texts. Starting from four questions, we tried to understand how each method works and its characteristics, as well as whether the model works when changing the language.

From the automatic evaluation perspective, we observed that methods that use the distribution of words in contexts tend to have low values of precision while having higher recall scores. It was also observed that methods using patterns have high values of precision but their recall is very low due to the scarcity of patterns in texts. Pattern "NP such as LNP" and its translation to Portuguese achieved the highest values of precision, while general patterns such as "NP is/are NP/LNP" achieved lower precision but higher recall scores. Using the term frequency as indicative of taxonomic relation seems to be a good option when intending to have a balanced result of precision and recall.

In order to improve results, it seems a good option to cluster terms before identifying the taxonomic relation. On the other hand, increasing the number of terms in the model decreases values of precision for most methods with exception of the method that uses patterns and the method that uses document subsumption. It also seems a good option to select the best hyperym for each term using the algorithm proposed by De Knijff *et al.* [27] since it reduces substantially the taxonomy and improves the precision. On the other hand, using this algorithm the recall and f-measure decrease significantly for most methods.

Observing results from the document subsumption model in the automatic evaluation, we could observe that low values of threshold achieved the highest values of f-measure. Also, models using the top 1,000 terms with the highest number of contexts achieved the best values of precision. Due to the document borders, the method using Europarl corpus could not identify any taxonomic relation using strict thresholds.

For hierarchical clustering, we note that although grouping semantically related terms improves the identification of the taxonomic relations between them, when changing the range from 1,000 terms to 10,000 terms the values of precision dropped indicating that not taxonomically related terms are belonging to the clusters. While increasing the precision when adding new terms using the top 1,000 terms, we observed that the local recall decreased, meaning that the more terms are included in clusters, the less taxonomic semantic relations are inferred by the model.

In the manual evaluation we observed how difficult it is to humans to evaluate domain relations. This could be observed by some low values of inter-annotator agreement. In general, these low values were obtained when an annotator selected "Not applicable" while the others selected "No" to score the taxonomic relation. Manual evaluation showed that the method based on head-modifier achieved the best results. On the other hand, this method is very limited, only finding relations between terms that share the same head of the noun phrase. Methods based on patterns also achieved good results, while methods based on the document subsumption achieved a lower number of correct relations. Wrong and correct relations are discussed and examples are presented.

Taxonomies generated by models in the automatic evaluation using the top 1,000 terms with the highest number of contexts are analysed in terms of hierarchy metrics. We observed that models that use the distribution of words generate taxonomies with much more relations than other methods. Almost all terms were related between themselves and the maximum distance between a root term (*i.e.*, a term without hypernym) and a leaf term (*i.e.*, a term without hyponym) was usually equal to one. Having a direct connection between the root and the leaf, it was difficult to measure the maximum distance between them. Thus, we applied a transitive reduction in the taxonomy and observed the characteristics of these new taxonomies. Hence, the taxonomies generated by distributional methods are much deeper than the other taxonomies. They also are very narrow, while taxonomies generated by methods such as document subsumption and hierarchical clustering are wider but not so deep.

Finally, methods are analyzed in terms of complementarity, where we can observe that most methods generate the same relations learned by models that use entropy between terms (`SLQS`), models that use term frequency `TF` and models that use document frequency (`DF`). Also, the method based on patterns `Patt` seems to be complementary to methods that use document subsumption `DocSub` and hierarchical clustering `HClust`. A high number of inverse relations generated by `DSim` also was noted, *i.e.*, when a model generates the relation between a term and its hypernym and `DSim` model generates the inverse relation. When observing the precision of the mixed models, `DSim` seems to benefit when joined to another method, and `Patt` model seems to benefit when joined to the `DocSub` model more than when joined to the `HClust` model.

## 6.3 Directions for further work

There are a number of major directions in which this work can be extended. First, the set of methods that we have evaluated is not exhaustive. There are a number of other methods that we could not include in this thesis due to time and resources issues. For example, it would be interesting to include machine learning methods to compare with the developed methods, all using the same test corpora.

Second, in this thesis we applied hierarchical clustering on a method that uses the document frequency to indicate a taxonomic relation. As semantically grouping terms with hierarchical clustering seems to improve results when compared with a method without the clustering process, it would be interesting to apply the hierarchical clustering in other methods and observe whether it improves the results.

Third, a deeper analysis on the complementarity of the methods would also be interesting. In this thesis we analyzed the intersection of pairs of methods, but a deeper analysis using the intersection of more than two methods, or the union of these methods might generate new insights.

Fourth, performing evaluations using manually generated taxonomies as gold standard for both languages using parallel corpora. This type of evaluation would help to understand the nuances of the methods in different languages. As we used two different gold standards in this work, it was difficult to identify the reasons why a method generates a high precision in one language and not such high precision in the other.

Finally, consider alternative dimensions for the evaluation. In this thesis we considered only one dimension of evaluation, performing an automatic evaluation against a gold standard and manual evaluation using the human judgements. Other interesting dimensions of the evaluation could be an in-use evaluation where we measure whether the taxonomy is more relevant for a document collection semantic annotation, for machine translation or any other task. Also, performing evaluations using corpora and gold standards in a domain specific task. As we used general gold standards, results might not be truly applicable when using a domain specific corpora.

# Bibliography

[1] Hiyan Alshawi. Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics*, 13(3-4):195–202, 1987.

[2] Robert A. Amsler. A taxonomy for english nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pages 133–138. ACL, 1981.

[3] Maya Ando, Satoshi Sekine, and Shun Ishizaki. Automatic extraction of hyponyms from japanese newspapers using lexico-syntactic patterns. In *Proceedings of the 4th international conference on Language Resources and Evaluation*, LREC'04, pages 387–390. European Language Resources Association (ELRA), 2004.

[4] Denis A de Araujo, Sandro J Rigo, Carolina Muller, and Rove Chishman. Automatic information extraction from texts with inference and linguistic knowledge acquisition rules. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 151–154. IEEE, 2013.

[5] Marco Baroni and Alessandro Lenci. Distributional memory: a general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[6] Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10. Association for Computational Linguistics, 2011.

[7] Túlio Lima Basegio. Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do brasil. Master's thesis, Computer Science Department, Pontifícia Universidade Católica do Rio Grande do Sul, 2007.

[8] Daniel Emilio Beck. Syntax-based statistical machine translation using tree automata and tree transducers. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, pages 36–40. Association for Computational Linguistics, 2011.

[9] Eckhard Bick. *The parsing system PALAVRAS*. PhD thesis, University of Arhus, 2000.

[10] Charles Blundell, Yee Whye Teh, and Katherine A. Heller. Bayesian rose trees. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, UAI-2010, 2009.

[11] Theodore L. Brown, H. Eugene LeMay, Bruce E. Bursten, and Julia R. Burdge. *Chemistry: The central science*. Prentice Hall, 2003.

[12] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*, NAACL-2001, pages 29–34, 2001.

[13] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Applications and Evaluation*, pages 3–12. IOS Press, 2005.

[14] Paul Buitelaar, Daniel Olejnik, and Michael Sintek. A protégé plug-in for ontology extraction from text based on linguistic analysis. In *Proceedings of the 1st First European Semantic Web Symposium*, ESWS 2004, 2004.

[15] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL '99, pages 120–126. Association for Computational Linguistics, 1999.

[16] Scott Cederberg and Dominic Widdows. Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CoNLL-2003, pages 111–118. Association for Computational Linguistics, 2003.

[17] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit$^3$: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, EAMT, pages 261–268, 2012.

[18] Eugene Charniak, Sharon Goldwater, and Mark Johnson. Edge-based best-first chart parsing. In *Proceedings of the 6th Workshop on Very Large Corpora*, ACL, pages 127–133. Association for Computational Linguistics, 1998.

[19] Shui-Lung Chuang and Lee-Feng Chien. A practical web-based approach to generating topic hierarchy for text segments. In *Proceedings of the 13th ACM international conference on Information and knowledge management*, CIKM'04, pages 127–136. ACM, 2004.

[20] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[21] Philipp Cimiano. *Ontology learning and population from text: algorithms, evaluation and applications*, volume 27. Springer-Verlag New York, Inc., 2006.

[22] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI-2004, pages 435–439, 2004.

[23] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24(1):305–339, 2005.

[24] Daoud Clarke. Context-theoretic semantics for natural language: An overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 112–119. Association for Computational Linguistics, 2009.

[25] Robert James Coulthard. *The application of corpus methodology to translation: the JPED parallel corpus and the pediatrics comparable corpus*. PhD thesis, Universidade Federal de Santa Catarina, 2005.

[26] Carolyn J. Crouch and Bokyung Yang. Experiments in automatic statistical thesaurus construction. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 77–88, 1992.

[27] Jeroen De Knijff, Flavius Frasincar, and Frederik Hogenboom. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. *Data & Knowledge Engineering*, 83:54–69, 2013.

[28] Scott Deerwester, Susan T. Dumais, George W. Furmas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[29] Melania Degeratu and Vasileios Hatzivassiloglou. An automatic method for constructing domain-specific ontology resources. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC-2004, pages 2001–2004. European Language Resources Association (ELRA), 2004.

[30] Emmanuelle-Anna Dietz, Damir Vandic, and Flavius Frasincar. Taxolearn: A semantic approach to domain taxonomy learning. In *Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence*, pages 58–65. IEEE Computer Society, 2012.

[31] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: an experimental study. 165(1):91–134, 2005.

[32] Stefan Evert. *The statistics of word cooccurrences: Word pairs and collocations*. PhD thesis, Institut fur maschinelle Sprachverarbeitung, University of Stuttgart, 2005.

[33] Robert Fano. Trasmission of information. Technical report, MIT press, 1961.

[34] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT Press, 1998.

[35] John Rupert Firth. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, pages 1–32, 1957.

[36] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[37] Richard Forsyth and Roy Rada. *Machine learning, applications in expert systems and information retrieval*. Ellis Horwood Limited, 1986.

[38] Larissa Astrogildo Freitas. Métricas para ontologias: Revisão sistemática e aplicação ao portal ontolp. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 2010.

[39] Larissa Astrogildo Freitas and Renata Vieira. Revisão sistemática sobre métricas para ontologias. In *Proceedings of Joint III Seminar on Ontology Research in Brazil*, ONTOBRAS 2010, 2010.

[40] P. Gamallo and S. Bordag. Is singular value decomposition useful for word similarity extraction? *Language Resources and Evaluation*, 45(2):95–119, 2011.

[41] Bernhard Ganter, Rudolf Wille, and Cornelia Franzke. *Formal concept analysis: Mathematical foundations*. Springer-Verlag New York, Inc., 1997.

[42] Maayan Geffet and Ido Dagan. Feature vector quality and distributional similarity. In *Proceedings of the 20th international Conference on Computational Linguistics*, Coling-04, pages 247–253. Association for Computational Linguistics, 2004.

[43] Maayan Geffet and Ido Dagan. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL-2005, pages 107–114. Association for Computational Linguistics, 2005.

[44] Roger Granada, Lucelene Lopes, Carlos Ramisch, Cassia Trojahn, Renata Vieira, and Aline Villavicencio. A comparable corpus based on aligned multilingual ontologies. In *Proceedings of the 1st Workshop on Multilingual Modeling*, MM '12, pages 25–31. Association for Computational Linguistics, 2012.

[45] Roger L. Granada, Renata Vieira, and Vera L.S.D. Lima. Evaluating co-occurrence order for automatic thesaurus construction. In *Proceedings of the IEEE 13th International Conference on Information Reuse and Integration*, IRI 2012, pages 474–481, 2012.

[46] Gregory Grefenstette. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers Norwell, 1994.

[47] Grolier. *Academic American encyclopedia*. Grolier Electronic Publishing, Danbury, Conneeticut, 1990.

[48] Maya R. Gupta and Yihua Chen. Theory and use of the em algorithm. *Foundations and Trends in Signal Processing*, 4(3):223–296, 2011.

[49] Patrick Hanks. Definitions and explanations. In *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. Collins ELT, 1987.

[50] Zellig S. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[51] Zellig Sabbettai Harris. *Mathematical structures of language*. Wiley, 1968.

[52] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, 1992.

[53] Marti A. Hearst. Automated discovery of wordnet relations. *WordNet: An electronic lexical database and some of its applications*, pages 131–153, 1998.

[54] Aurelie Herbelot and Ann Copestake. Acquiring ontological relationships from wikipedia using rmrs. In *Proceedings of Workshop on Web content Mining with Human Language Technologies*, ISWC '06, 2006.

[55] William Hersh, Chris Buckley, T.J. Leone, and David Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 192–201. Springer-Verlag, 1994.

[56] Andrew Hippisley, David Cheng, and Khurshid Ahmad. The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, 11(2):129–157, 2005.

[57] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[58] Jing Jiang. Information extraction from text. In *Mining Text Data*, pages 11–41. Springer-Verlag New York, Inc., 2012.

[59] Xing Jiang and Ah-Hwee Tan. Crctol: A semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology*, 61(1):150–168, 2010.

[60] D. Jurafsky and J.H. Martin. *Speech And language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2009.

[61] Hiroyuki Kaji, Yasutsugu Morimoto, Toshiko Aizono, and Noriyuki Yamasaki. Corpus dependent association thesauri for information retrieval. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 404–410, 2000.

[62] Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. An unsupervised approach to domain-specific term extraction. In *Proceedings of the 2009 Australasian Language Technology Association Workshop*, pages 94–98. Australasian Language Technology Association, 2009.

[63] Chunyu Kit and Xiaoyue Liu. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2):204–229, 2008.

[64] Carmen Klaussner and Desislava Zhekova. Lexico-syntactic patterns for automatic ontology building. In *Proceedings of the Student Research Workshop associated with RANLP 2011*, pages 109–114, 2011.

[65] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430. Association for Computational Linguistics, 2003.

[66] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT summit*, volume 5, 2005.

[67] Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389, 2010.

[68] Sotiris Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31(3):249–268, 2007.

[69] Zornitsa Kozareva and Eduard Hovy. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1110–1118. Association for Computational Linguistics, 2010.

[70] Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL-2008, pages 1048–1056. Association for Computational Linguistics, 2008.

[71] Hélio Kuramoto. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. *Ciência da Informação*, 25(2):182–192, 1996.

[72] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240, 1997.

[73] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[74] Ming-Che Lee, Ding Yen Ye, and Tzone I Wang. Java learning object ontology. In *Proceedings of the 5th IEEE international conference on advanced learning technologies*, ICALT 2005, pages 538–542. IEEE Computer Society, 2005.

[75] Benoît Lemaire and Guy Denhiére. Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters*, 18(1):1–12, 2006.

[76] Douglas B. Lenat and Ramanathan V. Guha. *Building large knowledge-based systems: representation and inference in the CYC project*. Addison-Wesley, Reading, MA, 1990.

[77] Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics - Volume 1*, pages 75–79. Association for Computational Linguistics, 2012.

[78] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM, 1986.

[79] Chuan-Xi Li, Ru-Jing Wang, Peng Chen, He Huang, and Ya-Ru Su. Interaction relation ontology learning. *Journal of Computational Biology*, 21(1):80–88, 2014.

[80] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 2*, COLING '98, pages 768–774, 1998.

[81] Dekang Lin. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems – 1st International Conference on Language Resources and Evaluation*, LREC-1998. European Language Resources Association (ELRA), 1998.

[82] Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD'12, pages 1433–1441. ACM, 2012.

[83] L. Lopes and V. Vieira. Building domain specific corpora in portuguese language. Technical Report TR 062, PUCRS, Porto Alegre, Brasil, 2010.

[84] Lucelene Lopes. *Extração automática de conceitos a partir de textos em língua portuguesa*. PhD thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil, 2012.

[85] Lucelene Lopes, Paulo Fernandes, and Renata Vieira. Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence*, ICAI 2012, pages 1001–1007. CSREA Press, 2012.

[86] Lucelene Lopes, Paulo Fernandes, Renata Vieira, and Guilherme Fedrizzi. Exatolp – an automatic tool for term extraction from portuguese language corpora. In *Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, LTC-2009, pages 427–431, 2009.

[87] Lucelene Lopes and Renata Vieira. Improving portuguese term extraction. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, PROPOR'12, pages 85–92. Springer-Verlag, 2012.

[88] Lucelene Lopes and Renata Vieira. Aplicando pontos de corte para listas de termos extraídos. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 1–9. SBC, 2013.

[89] Lucelene Lopes and Renata Vieira. Construção automática de hierarquias de conceitos a partir de corpus: abordagens existentes e limitações. Technical Report TR 074, PUCRS, Porto Alegre, Brasil, 2013.

[90] Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 251–263. Springer-Verlag, 2002.

[91] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[92] Melvin E. Maron and Lary Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960.

[93] Erick G. Maziero, Thiago A. S. Pardo, Ariani Di Felippo, and Bento C. Dias-da Silva. A base de dados lexical e a interface web do tep 2.0: Thesaurus eletrônico para o português do brasil. In *Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, WebMedia '08, pages 390–392, 2008.

[94] Tony McEnery and Richard Xiao. Parallel and comparable corpora: What is happening? *Incorporating Corpora: The Linguist and the Translator*, pages 18–31, 2007.

[95] Paul McNamee, Rion Snow, Patrick Schone, and James Mayfield. Learning named entity hyponyms for question answering. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, IJCNLP, pages 799–804, 2008.

[96] Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. Similarity involving attributes and relations: judgments of similarity and difference are not inverses. *Psychological Science*, 1(1):64–69, 1990.

[97] Kevin Meijer, Flavius Frasincar, and Frederik Hogenboom. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems*, 62:78–93, 2014.

[98] Andreas Mengel and Wolfgang Lezius. An xml-based representation format for syntactically annotated corpora. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, LREC'00. European Language Resources Association, 2000.

[99] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[100] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.

[101] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[102] Verginica Barbu Mititelu. Hyponymy patterns. In Petr Sojka, Ales Horák, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 5246 of *Lecture Notes in Computer Science*, pages 37–44. Springer Berlin Heidelberg, 2008.

[103] Emmanuel Morin and Christian Jacquemin. Projecting corpus-based semantic links on a thesaurus. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, ACL '99, pages 389–396. Association for Computational Linguistics, 1999.

[104] Vivi Nastase, Preslav Nakov, Diarmuid O Seaghdha, and Stan Szpakowicz. *Semantic relations between nominals*. Morgan & Claypool Publishers, 2013.

[105] Roberto Navigli, Paola Velardi, and Stefano Faralli. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the 22nd international joint conference on Artificial Intelligence - Volume 3*, IJCAI'11, pages 1872–1877. AAAI Press, 2011.

[106] Hermine Njike-Fotzo and Patrick Gallinari. Learning «generalization/specialization» relations between concepts: application for automatically building thematic document hierarchies. In *Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval*, RIAO-2004, pages 143–155, 2004.

[107] Hugo Gonçalo Oliveira. *Onto.PT: Towards the automatic construction of a lexical ontology for Portuguese*. PhD thesis, University of Coimbra, 2013.

[108] Hugo Gonçalo Oliveira, Diana Santos, and Paulo Gomes. Relations extracted from a portuguese dictionary: Results and first evaluation. In *Proceedings of the 14th Portuguese Conference on Artificial Intelligence*, EPIA, pages 541–552, 2009.

[109] Hugo Gonçalo Oliveira. The creation of onto.pt: a wordnet-like lexical ontology for portuguese. In *Proceedings of 11th International Conference on Computational Processing of the Portuguese Language*, PROPOR 2014, pages 161–169. Springer-Verlag, 2014.

[110] Patrick Pantel. *Clustering by committee*. PhD thesis, Department of Computing Science, University of Alberta, 2003.

[111] Patrick Pantel and Dekang Lin. A statistical corpus-based term extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, AI '01, pages 36–46. Springer-Verlag, 2001.

[112] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-2006, pages 113–120. Association for Computational Linguistics, 2006.

[113] Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. An empirical analysis of word error rate and keyword error rate. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2070–2073, 2008.

[114] Marius Pasca. Acquisition of categorized named entities for web search. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 137–145. ACM, 2004.

[115] Marius Pasca and Benjamin Van Durme. Weakly supervised acquisition of open-domain classes and class attributes from web documents and query logs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL-2008, pages 19–27. Association for Computational Linguistics, 2008.

[116] Mari-Sanna Paukkeri, Alberto Pérez García-Plaza, Víctor Fresno, Raquel Martínez Unanue, and Timo Honkela. Learning a taxonomy from a set of text documents. *Applied Soft Computing*, 12(3):1138–1148, 2012.

[117] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL-1993, pages 183–190. Association for Computational Linguistics, 1993.

[118] Simone Paolo Ponzetto and Michael Strube. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, pages 1440–1445. AAAI Press, 2007.

[119] Simone Paolo Ponzetto and Michael Strube. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 9:1737–1756, 2011.

[120] Kevin L. Priddy and Paul E. Keller. *Artificial neural networks: An introduction*, volume 68. SPIE Press – International Society for Optical Engineering, 2005.

[121] Andrew Radford. *Syntax: A minimalist introduction*. Cambridge University Press, 1997.

[122] Marek Rei and Ted Briscoe. Looking for hyponyms in vector space. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, CoNLL-14, pages 68–77. Association for Computational Linguistics, 2014.

[123] Ana B. Rios-Alvarado, Ivan Lopez-Arevalo, and Victor J. Sosa-Sosa. Learning concept hierarchies from textual resources for ontologies construction. *Expert Systems with Applications*, 40(15):5907–5915, 2013.

[124] Stephen E. Robertson and S. Walker. On relevance weights with little relevance information. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 16–24. ACM, 1997.

[125] María Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. Using context-window overlapping in synonym discovery and ontology extension. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP 2005, 2005.

[126] Motaz Saad, David Langlois, and Kamel Smaïli. Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia-Social and Behavioral Sciences*, 95(0):40–47, 2013.

[127] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[128] Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 206–213. ACM, 1999.

[129] Annalisa Sandrelli. Introducing footie (football in europe): Simultaneous interpreting in football press conferences. In *Breaking Ground in Corpus-Based Interpreting Studies*, pages 119–154. Berna: Peter Lang, 2012.

[130] Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. Chasing hypernyms in cector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2*, pages 38–42. Association for Computational Linguistics, 2014.

[131] Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[132] Alberto Simões, ÁlvaroIriarte Sanromán, and JoséJoão Almeida. Dicionário-aberto: A source of resources for the portuguese language processing. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, PROPOR'12, pages 121–127. Springer Berlin Heidelberg, 2012.

[133] Michael Sintek, Paul Buitelaar, and Daniel Olejnik. A formalization of ontology learning from text. In *Proceedings of the Workshop on Evaluation of Ontology-based Tools*, EON 2004, 2004.

[134] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[135] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of Advances in Neural Information Processing Systems 17*, pages 1297–1304, 2005.

[136] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 801–808. Association for Computational Linguistics, 2006.

[137] Steffen Staab, Christian Braun, Ilvio Bruder, Antje Düsterhöft, Andreas Heuer, Meike Klettke, Günter Neumann, Bernd Prager, Jan Pretzel, Hans-Peter Schnurr, Rudi Studer, Hans Uszkoreit, and Burkhard Wrenger. Getess: searching the web exploiting german texts. In *Cooperative Information Agents III*, pages 113–124. Springer-Verlag, 1999.

[138] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706. ACM, 2007.

[139] Asuka Sumida and Kentaro Torisawa. Hacking wikipedia for hyponymy relation acquisition. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 883–888, 2008.

[140] Asuka Sumida, Naoki Yoshinaga, and Kentaro Torisawa. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in wikipedia. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC'08. European Language Resources Association, 2008.

[141] Idan Szpektor and Ido Dagan. Learning entailment rules for unary templates. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 849–856. Association for Computational Linguistics, 2008.

[142] Leonardo Sameshima Taba and Helena Caseli. Automatic semantic relation extraction from portuguese texts. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC'14, pages 2739–2746. European Language Resources Association, 2014.

[143] Erik Tjong Kim Sang and Katja Hofmann. Automatic extraction of dutch hypernym-hyponym pairs. *LOT Occasional Series*, 7:163–174, 2007.

[144] Mireya Tovar, David Pinto, Azucena Montes, Gabriel González, Darnes Vilariño, and Beatriz Beltrán. Use of lexico-syntactic patterns for the evaluation of taxonomic relations. In *Pattern Recognition*, volume 8495 of *Lecture Notes in Computer Science*, pages 331–340. Springer International Publishing, 2014.

[145] Douglas Tudhope, Traugott Koch, and Rachel Heery. Terminology services and technology: Jisc state of the art review. Technical report, University of Bath - UKOLN Research Centre, 2006.

[146] Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, ECML 2001, pages 491–502, 2001.

[147] Peter D. Turney and Patrick Pantel. From frequency to meaning. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[148] Paola Velardi, Paolo Fabriani, and Michele Missikoff. Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, FOIS '01, pages 270–284. ACM, 2001.

[149] Paola Velardi, Stefano Faralli, and Roberto Navigli. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3), 2013.

[150] Piek Vossen. Extending, trimming and fusing wordnet for technical documents. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, pages 125–131. The Association for Computational Linguistics, 2001.

[151] Denny Vrandečić and York Sure. How to design better ontology metrics. In *Proceedings of the 4th European Conference on The Semantic Web: Research and Applications*, ESWC '07, pages 311–325. Springer-Verlag, 2007.

[152] Vishnu Vyas and Patrick Pantel. Semi-automatic entity set refinement. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 290–298. Association for Computational Linguistics, 2009.

[153] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. A phrase mining framework for recursive construction of a topical hierarchy.

In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD-2013, pages 1433–1441. ACM, 2013.

[154] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, COLING 2014, pages 2249–2259. Dublin City University and Association for Computational Linguistics, 2014.

[155] Julie Weeds and David Weir. A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 81–88. Association for Computational Linguistics, 2003.

[156] Julie Weeds, David Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING-2004, pages 1015–1021, 2004.

[157] Joachim Wermter and Udo Hahn. You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 785–792. Association for Computational Linguistics, 2006.

[158] Dominic Widdows. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 197–204. Association for Computational Linguistics, 2003.

[159] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, 1999.

[160] Fei Wu and Daniel S. Weld. Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 635–644. ACM, 2008.

[161] Feiyu Xu, Hong Li, Yi Zhang, Hans Uszkoreit, and Sebastian Krause. Parse reranking for domain-adaptative relation extraction. *Journal of Logic and Computation*, 24(2):413–431, 2014.

[162] Quang Xuan Do and Dan Roth. Exploiting the wikipedia structure in local and global classification of taxonomic relations. *Natural Language Engineering*, 18(02):235–262, 2012.

[163] Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond, and Asuka Sumida. Hypernym discovery based on distributional similarity and hierarchical structures. In *Proceedings of the 2009 Conference on Empirical*

*Methods in Natural Language Processing - Volume 2*, pages 929–937. Association for Computational Linguistics, 2009.

[164] Dongqiang Yang and David M. Powers. Automatic thesaurus construction. In *Proceedings of the 31st Australasian Conference on Computer Science*, ACSC-2008, pages 147–156, 2008.

[165] Haining Yao, Anthony M. Orme, and Letha Etzkorn. Cohesion metrics for ontology design and application. *Journal of Computer science*, 1(1):107–113, 2005.

[166] Tugba Yildiz and Savas Yildirim. Association rule based acquisition of hyponym and hypernym relation from a turkish corpus. In *Proceedings of the IEEE International Symposium on Innovations in Intelligent Systems and Applications*, INISTA, pages 1–5. IEEE Computer Society, 2012.

[167] Maayan Zhitomirsky-Geffet and Ido Dagan. Bootstrapping distributional feature vector quality. *Computational linguistics*, 35(3):435–461, 2009.

[168] George Kingsley Zipf. *The psycho-biology of language*. Oxford, England: Houghton, Mifflin, 1935.

# Appendix A. Tagset of Stanford Parser

List of tags used by Stanford parser followed by their description.

Source: `http://nlp.stanford.edu/software/tagger.shtml`

| **Tag** | **Description** |
|---|---|
| $ | dollar |
| “ | opening quotation mark |
| ” | closing quotation mark |
| ( | opening parenthesis |
| ) | closing parenthesis |
| , | comma |
| – | dash |
| . | sentence terminator |
| : | colon or ellipsis |
| CC | conjunction, coordinating |
| CD | numeral, cardinal |
| DT | determiner |
| EX | existential there |
| FW | foreign word |
| IN | preposition or conjunction, subordinating |
| JJ | adjective or numeral, ordinal |
| JJR | adjective, comparative |
| JJS | adjective, superlative |
| LS | list item marker |
| MD | modal auxiliary |
| NN | noun, common, singular or mass |
| NNP | noun, proper, singular |
| NNPS | noun, proper, plural |
| NNS | noun, common, plural |
| PDT | pre-determiner |
| POS | genitive marker |
| PRP | pronoun, personal |
| PRP$ | pronoun, possessive |
| RB | adverb |
| RBR | adverb, comparative |
| RBS | adverb, superlative |
| RP | particle |
| SYM | symbol |
| TO | “to” as preposition or infinitive marker |
| UH | interjection |
| VB | verb, base form |
| VBD | verb, past tense |
| VBG | verb, present participle or gerund |
| VBN | verb, past participle |
| VBP | verb, present tense, not 3rd person singular |
| VBZ | verb, present tense, 3rd person singular |
| WDT | WH-determiner |
| WP | WH-pronoun |
| WP$ | WH-pronoun, possessive |
| WRB | Wh-adverb |

# Appendix B. Tagset of PALAVRAS parser

List of word class tags and group forms used by PALAVRAS parser followed by their description. Source: `http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html`

## Word class tags

| Tag | Description |
|---|---|
| N | Nouns |
| PROP | Proper nouns (names) |
| SPEC | Specifiers (defined as non-inflecting pronouns, that can't be used as prenominals) |
| DET | Determiners (defined as inflecting pronouns, that can be used as prenominals) |
| PERS | Personal pronouns (defined as person-inflecting pronouns) |
| ADJ | Adjectives (including ordinals, excluding participles which are tagged V PCP) |
| ADV | Adverbs (both 'primary' adverbs and derived adverbs ending in -mente) |
| V | Verbs (full verbs, auxiliaries) |
| NUM | Numerals (cardinals) |
| PRP | Preposition |
| KS | Subordinating conjunctions |
| KC | Coordinationg conjunctions |
| IN | Interjections |
| EC | Hyphen-separated prefix ("elemento composto", category being phased out) |

## Group forms

| Tag | Description |
|---|---|
| NP | noun phrase |
| AP | adjective phrase |
| ADVP | adverb phrase |
| VP | verb phrase |
| PP | prepositional phrase |

# Appendix C. Heuristics application

After extracting all NPs from the corpus, a refining process is performed where heuristics are applied on each NP. These heuristics were adapted from the ones applied in Portuguese by Lopes [84,87] and separated into three groups: Heuristics for adjustment (HA), Heuristics for discard (HD) and Heuristics for inclusion (HI). Heuristics for adjustment aim at removing elements from the NP that do not carry any information, such as determiners, pronouns, adverbs and possessive mark. Heuristics for discard intend to remove NPs when they are not relevant to the domain, *i.e.*, NPs that are composed by non-informative terms such as symbols, numbers *etc.*. Unlike the heuristics for adjustment, the heuristic for discard remove the entire NP instead of some of its terms. Heuristics for inclusion intend to detect inner NPs, *i.e.*, NPs that are implicit inside a greater NP. Each heuristic is explained as follows:

1. HA: Removing determiners and predeterminers

   As determiners express the reference of a noun or noun phrase in the context, indicating whether the noun is referring to a definite or indefinite element of a class, to a particular number or quantity, to a closer or more distant element, *etc.*, they do not play an important role to the domain identification. In English, determiners (DT) include articles (*e.g.*, “the”, “a” and “an”), demonstrative pronouns (*e.g.*, “this” and “that” and their respective plural forms “these” and “those”) and quantifiers (*e.g.*, “some”, “any”, “many”). A predeterminer (PDT) is a type of determiner that precedes other determiners in a noun phrase. Thus, this heuristic removes all determiners and predeterminers of an NP independently of its position. Consider the following examples below of removing determiners and predeterminers:

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (DT The) (NN man)) | (NP (NN man)) |
| 2. | (NP (DT the) (NNP Member) (NNPS States)) | (NP (NNP Member) (NNPS States)) |
| 3. | (NP (PDT all) (DT the) (JJ technical) (NNS measures)) | (NP (JJ technical) (NNS measures)) |
| 4. | (NP (PDT all) (DT these) (NNS agreements)) | (NP (NNS agreements)) |
| 5. | (NP (PDT such) (DT the) (JJ European) (NNS institutions)) | (NP (JJ European) (NNS institutions)) |

   Following the work by Lopes and Vieira [87], in Portuguese only the articles (definite and indefinite) are removed from NPs. It includes articles in the beggining of an NP as well as articles generated by the splitting of a preposition plus article, such as “do=de+o”. Thus, considering the NPs extracted from Figure 4.5, the resulting NPs would be:

```
np: homem/n
np: ausência/n de/prp lógica/n de/prp senhor/n deputado/n Poettering/prop
np: lógica/n de/prp senhor/n deputado/n Poettering/prop
np: senhor/n deputado/n Poettering/prop
np: deputado/n Poettering/prop
```

2. HA: Removing pronouns

   This heuristic removes all personal pronoun (PRP) such as "I", "you", "we", *etc.* and possessive pronoun (PRP$) such as "mine", "yours", "ours", *etc.*. Unlike PALAVRAS parser, Stanford parser does not identify the head of a noun phrase. Thus, the removal of pronouns when they play the role of the head of an NP can lead to empty or non-informative NPs (*i.e.*, an NP containing only adjectives). These NPs are further removed by heuristics of discard. Examples of the application of this heuristic are shown below, where we can observe that the heuristic applied on the example 5 resulted in an empty NP and examples 6 and 7 resulted in non-informative NPs.

```
#    Original NP:                                      After applying the heuristic
-----------------------------------------------------------------------------------------------
1.   (NP (PRP$ your) (NN meeting))                     (NP (NN meeting))
2.   (NP (PRP$ its) (JJ diplomatic) (NNS contradictions))  (NP (JJ diplomatic) (NNS contradictions))
3.   (NP (NNP Europe) (PRP itself))                    (NP (NNP Europe))
4.   (NP (PRP we) (NNPS Socialists))                   (NP (NNPS Socialists))
5.   (NP (PRP We))                                     (NP )
6.   (NP (PRP us) (DT both))                           (NP (DT both))
7.   (NP (RB then) (PRP I))                            (NP (RB then))
```

   For Portuguese, the heuristic removes all pronouns (tagged as `pron`) when they do not play the role of the head of the NP. For example, consider the NP "Nosso/pron parlamento/n" in the phrase "Nosso parlamento irá interpor." and the NP "Eu/pron" in the phrase "Eu irei interpor.". In the former NP the pronoun is removed because "parlamento" plays the role of the head of the NP. Contrary, in the latter the pronoun is not removed because it plays the role of the head of the NP.

3. HA: Removing adverbs

   This heuristic removes any adverb of the NP, since most of the adverbs tell when, where, how, in what manner, or to what extent an action is performed, and hence, not carrying any meaning of the domain. The application of this heuristic on NPs are presented below, where we can see that examples 4 and 5 left empty NPs. As empty NPs left from the heuristic of removing pronouns, the empty NPs generated in this heuristic are further removed by heuristics for discard.

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (RB even) (JJ non-speculative) (NNS streams)) | (NP (JJ non-speculative) (NNS streams)) |
| 2. | (NP (RB mainly) (NNS guillemots)) | (NP (NNS guillemots)) |
| 3. | (NP (RB only) (RB ever) (NN man)) | (NP (NN man)) |
| 4. | (NP (RB yesterday)) | (NP ) |
| 5. | (NP (RB Not) (RB only)) | (NP ) |

For Portuguese, Lopes and Vieira [87] located this heuristic in "heuristics for discard". Thus, when an adverb occurs in the NP, the whole NP is removed.

4. HA: Removing possessive ('s)

   This heuristic is developed only for English since it was noted that Stanford Parser keeps the possessive (`'s POS`) at the end of certain noun phrases. Thus, this heuristic removes the possessive mark when it appears at the end of the NP. Examples of the application of this heuristic in NPs containing the possessive mark are presented below.

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (NNP Parliament) (POS 's)) | (NP (NNP Parliament)) |
| 2. | (NP (NNP Parliament) (POS 's) (NN regret)) | (NP (NNP Parliament) (POS 's) (NN regret)) |
| 3. | (NP (NNP President) (NNP Clinton) (POS 's)) | (NP (NNP President) (NNP Clinton)) |
| 4. | (NP (NNP President) (NNP Clinton) (POS 's) (NN visit)) | (NP (NNP President) (NNP Clinton) (POS 's) (NN visit)) |

5. HD: Removing NPs composed by numbers

   As numbers alone do not carry any semantic information about the domain, NPs containing numbers are removed. Numbers can be in written form (*e.g.*, "seven") or in numeric characters (*e.g.*, "7"). Thus, this heuristic also removes NPs containing dates, years, hours, ranks, positions, *etc.*.

6. HD: Removing NPs containing symbols

   This heuristics discard NPs that contain any type of symbol but alphabetic characters or hyphen. Thus, the heuristic discards malformed NPs as well as email addresses which contain "@", internet addresses which contain "/", *etc.*. Terms containing symbols are probably uninteresting typos and garbage, and as terms containing numbers, they are not meaningful terms to the domain.

7. HD: Removing NPs containing a pronoun as the head

   Since pronouns do not contain important information about a domain and heuristics for adjustment do not remove pronouns when they play the role of the head of an NP, this heuristic removes such NPs. Refusing NPs containing the pronoun as the head, the resulted NPs must have common nouns, proper nouns, adjectives or verbs in the past participle tense playing the role of its head.

8. HD: Removing NPs that begin with adverb

   This heuristic removes NPs that do not explicitly refer to a term. According to [84] sometimes an NP references to a term previously mentioned. Usually when it occurs, the first word of the NP is an adverb and the head of the NP is an adjective. For example, the NP "mais/adv frequente/adj" refers to something previously mentioned in the text, and thus, do not carry any meaning to the domain. Hence, this heuristic removes NPs that begin with adverb.

9. HD: Removing empty NPs or NPs without nouns

   After applying all the heuristics for adjustment a verification of the consistence of the NPs is required. This verification intends to check whether the NP still contains any noun. As explained before, Stanford parser does not tag the head of the NP, hence removing pronouns can lead to empty or non-informative NPs, which must be removed. We consider a non-informative NP when it does not have a noun. Examples of the aplication of a heuristic for adjustment and the resulting NPs are presented below.

| # | Original NP: | After applying the heuristic | Reason to remove |
|---|---|---|---|
| 1. | (NP (PRP we) (MD may) (VB correct) (PRP them)) | (NP (MD may) (VB correct)) | Non-informative |
| 2. | (NP (RB even) (JJR worse)) | (NP (JJR worse)) | Non-informative |
| 3. | (NP (PRP us) (DT both)) | (NP ) | Empty NP |
| 4. | (NP (PRP We)) | (NP ) | Empty NP |

10. HI: Removing the conjunction of adjectives

    This heuristic intends to create NPs for adjectives and nouns when they are linked by a conjunction into a single NP. The heuristic works also with a list of adjectives separated by comma containing a conjunction in the end. Examples of the application of the elimination of the conjunction of NPs are presented below.

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (JJ technical) (CC and)<br>(JJ industrial) (NNS developments)) | (NP (JJ technical) (NNS developments))<br>(NP (JJ industrial) (NNS developments)) |
| 2. | (NP (JJ national) (, ,) (JJ regional)<br>(CC and) (JJ local) (NNS obstacles)) | (NP (JJ national) (NNS obstacles))<br>(NP (JJ regional) (NNS obstacles))<br>(NP (JJ local) (NNS obstacles)) |
| 3. | (NP (JJ relevant) (JJ American) (, ,)<br>(JJ Canadian) (CC and)<br>(JJ Japanese) (NNS authorities)) | (NP (JJ relevant) (JJ American) (NNS authorities))<br>(NP (JJ Canadian) (NNS authorities))<br>(NP (JJ Japanese) (NNS authorities)) |

11. HI: Removing the conjunction of nouns

    Unlike the heuristic for removing the conjunction of adjectives, this heuristic creates separated noun phrases when two or more nouns that are linked by comma or/and conjuntion into the same NP. Thus, instead of keeping the head of the NP and creating new NPs with their modifiers, this heuristic splits the NP into two or more NPs, each one containing a noun. It is important to note here that the elimination of conjunctions only occurs when the NP does not contain another NP within. Examples of the elimination of the conjunction for nouns are presented below. Note that example 4 contains an NP inside another NP and thus it is not eliminated by this heuristic.

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (NN supply) (CC and) (NN demand)) | (NP (NN supply))<br>(NP (NN demand)) |
| 2. | (NP (NNS Subsidies) (, ,)<br>(NNS monopolies) (CC or) (NNS barriers)) | (NP (NNS Subsidies))<br>(NP (NNS monopolies))<br>(NP (NNS barriers)) |
| 3. | (NP (JJ European) (NNP Liberal) (, ,)<br>(NNP Democrat) (CC and)<br>(NNP Reform) (NNP Party)) | (NP (JJ European) (NNP Liberal))<br>(NP (NNP Democrat))<br>(NP (NNP Reform) (NNP Party)) |
| 4. | (NP (NP (JJ perfect) (NN competition)) (CC and)<br>(NP (JJ optimum) (NN distribution))) | --- |

12. HI: Removing adjectives, nouns and proper nouns

This heuristic intends to find new NPs by sucessively removing adjectives, nouns and proper nouns. For English, post-positive adjectives are not removed by this heuristic, but adjectives placed before the noun. Additionally, this heuristic also remove nouns and proper nouns. It is important to note that this heuristic does not separate proper nouns (NNP and NNPS), as presented in the line 4 of the examples below.

| # | Original NP: | After applying the heuristic |
|---|---|---|
| 1. | (NP (JJ fundamental) (JJ legal) (NN principle)) | (NP (JJ legal) (NN principle))<br>(NP (NN principle)) |
| 2. | (NP (JJ three-speed) (JJ European) (NN territory)) | (NP (JJ European) (NN territory))<br>(NP (NN territory)) |
| 3. | (NP (JJ small) (NN project) (NNS partners)) | (NP (NN project) (NNS partners))<br>(NP (NNS partners)) |
| 4. | (NP (NNP Sri) (NNP Lankan) (NN president)) | (NP (NNP Sri) (NNP Lankan))<br>(NP (NN president)) |

According to Lopes [84], in Portuguese this heuristic is applied only to post-positive adjectives or verbs in the past participle tense. For example, from the phrase "Este é o papel da nova ordem democrática internacional." original NPs are extracted from the NP tag. The augmented NPs are extracted by sucessively removing the adjectives as presentend below. It is important to note that as only post-positive adjectives are removed, the adjective "nova" in line 1 is not removed since it is placed before the noun.

```
1.  Original NP:    nova/adj ordem/n democrática/adj internacional/adj
    Augmented NP:   nova/adj ordem/n democrática/adj
    Augmented NP:   nova/adj ordem/n
2.  Original NP:    papel/n de/prp nova/adj ordem/n democrática/adj internacional/adj
    Augmented NP:   papel/n de/prp nova/adj ordem/n democrática/adj
    Augmented NP:   papel/n de/prp nova/adj ordem/n
```

13. HI: Multiplicate phrases when occurring multiple predicates

    This heuristic was only developed for Portuguese and intends to multiplicate a phrase when it has multiple predicates. Multiple predicates occur when the predicate has two or more verbs that modify the subject of the phrase. For example, the phrase “O parlamento respeita e aprova o projeto de lei.” has the verbs “aprova” and “respeita” for the subject “O parlamento”. Thus, this phrase is separated into two, generating the phrases:

```
Original phrase:    O parlamento respeita e aprova o projeto de lei.
First predicate:    O parlamento respeita o projeto de lei.
Second predicate:   O parlamento aprova o projeto de lei.
```

    When duplicating NPs, this heuristic increases the frequency of each NP when it is playing the role of subject or object, as well as the number of verbs in the phrase. The heuristic also creates two different relations for subjects and objects.

14. HI: Multiplicate NPs when occurring conjunctions of adjectives

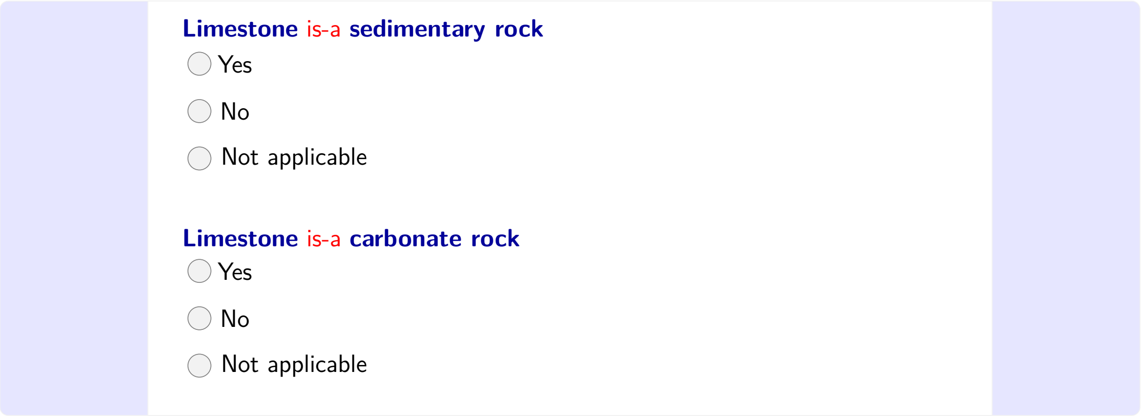
    This heuristic was developed only for Portuguese and intends to multiplicate an NP when it has a conjunction of adjectives, *i.e.*, when two or more adjectives of an NP are separated by comma or conjunction. For example, consider the phrase “Trabalharemos para ter um mercado competitivo e seguro.” containing the NP “mercado/n competitivo/adj e/conj-c seguro/adj”. The NP can be splitted into two NPs separated by the conjunction, as presented below.

```
1.  Original NP:    mercado/n competitivo/adj e/conj-c seguro/adj
    Augmented NP:   mercado/n competitivo/adj
    Augmented NP:   mercado/n seguro/adj
2.  Original NP:    mercado/n competitivo/adj ,/, produtivo/adj e/conj-c seguro/adj
    Augmented NP:   mercado/n competitivo/adj
    Augmented NP:   mercado/n produtivo/adj
    Augmented NP:   mercado/n seguro/adj
```

    This heuristic uses the conjunctions “ou” and “e” to split the NP into two or more NPs. For more conjunctions the heuristic can also split the NP where a comma occurs, as presented in the item 2 of the example.

# Appendix D. Guidelines for the manual evaluation of taxonomic relations

This chapter describes the guidelines presented to evaluators in order to perform the manual evaluation of the triples containing taxonomic relations between terms. The evaluation process consists of the assessment of triples by domain experts, considering only a taxonomic relation (hypernymy - hyponym), *i.e.*, "is a" or "is a kind of" relation. A triple is composed as <$\mathtt{w}_1$, r, $\mathtt{w}_2$>, where $\mathtt{w}_1$ and $\mathtt{w}_2$ are two terms and r is the relation between both terms. The whole taxonomy is not evaluated directly by the experts, but the relations between pairs of terms. For the sake of simplicity, a Google form[1] was generated containing all triples that should be evaluated by the domain experts. An excerpt of the Google form can be seen in Figure D.1.



**Limestone** is-a **sedimentary rock**

○ Yes

○ No

○ Not applicable

**Limestone** is-a **carbonate rock**

○ Yes

○ No

○ Not applicable

Figure D.1: Google form used in the manual evaluation

For each triple the evaluators should follow these guidelines:

1. Assess each triple while answering the question: "Is $\mathtt{w}_1$ a (kind of/form of) $\mathtt{w}_2$?". Using the examples of Figure D.1, the question should be "Is Limestone a (kind of/form of) sedimentary rock?" and "Is Limestone a (kind of/form of) carbonate rock?". The possible answers are:

    - "Yes" which means that the relation is correct
    - "No" which means that the relation is wrong
    - "Not applicable" that should be used when a term is ill-formed or does not belong to the domain. An example of ill-formed term is "decomposiça&#771;o". For terms that does not belong to the domain, consider the relation between the triple "disease" is-a "flu". This relation is taxonomically correct, but if you are evaluating relations in the Geology domain, this relation is not correct. Attention because acronyms may exist in relations.

2. If a phrase has multiple senses, consider all senses of the phrase in the domain, , "school" may refer to both the building and the institution. Thus, the triples "school is a building" and

[1] http://www.google.com/forms/

"school is a institution" are correct. In other words, assess the triple as correct if there is at least one sense of the word that holds the relation.

3. Disregard the number of the phrases (*i.e.*, singular or plural): , both triples "`natural gas is-a fluid`" and "`natural gases is-a fluid`" are correct, although "natural gases" refers to a set instead of single individual.

# Appendix E. Metrics for characterizing taxonomies

This chapters presents all metrics for characterizing taxonomies presented in Section 4.4.3. These metrics intend to characterize how taxonomies are generated by each method. They are applied to taxonomies or Rooted Directed Acyclic Graphs (Rooted DAG), *i.e.*, a structure having a single highest node (Root) and all other nodes are connected by means of is-a links, generating a chain of links to the Root.

Each metrics presented here is decribed below, where "terms" refers to a set of terms, "$\mathtt{term_{ij}}$" refers to the $i$th term in the taxonomy $j$ where "$\mathtt{term_{ij}}$" ∈ "terms", "rels" to a set of relations in the taxonomy, "$\mathtt{rel_{ij}}$" refers to the $i$th relation in the taxonomy $j$ where "$\mathtt{rel_{ij}}$" ∈ "rels", "Count()" is a function that counts the number of occurrences, "Max()" determines the maximum value of the set, "Min()" determines the minimum value of the set, "isRoot()" and "isLeaf()" verify whether the argument is a root or a leaf term in the taxonomy, "Depth()" returns the depth of a term, "Width()" returns the number of siblings of a term, and "hasSiblings()" returns terms that have siblings, "noRepetition()" returns relations without counting its repetitions, and "StdDev()" calculates the standard deviation of a set. In order to better understand the metrics, consider that each applied method may generate one or more taxonomies, and each taxonomy has a root term and at least one leaf term. Examples presented in some metrics use Figure E.1 where two taxonomies are represented as directed graphs (digraphs). Nodes represent words being identified by their ids, and edges represet the relation is hypernym of. Thus, the connection between nodes "ID: 1" and "ID: 2" means that the word identified by "ID: 1" is hypernym of the word identified by "ID: 2".

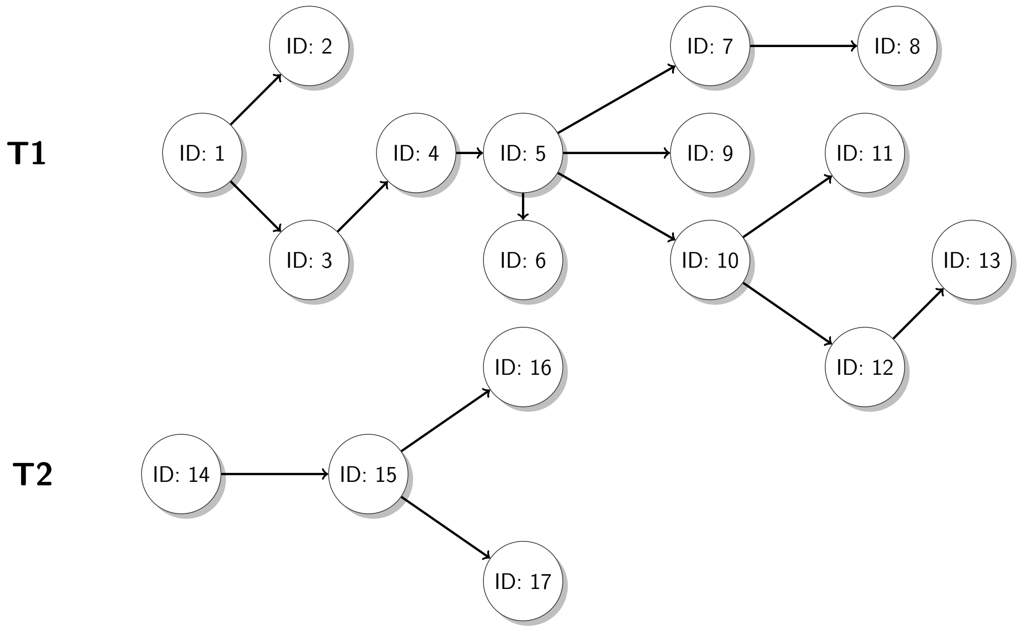


Figure E.1: Examples of taxonomies represented as direct graphs

**TotalTerms**: Total number of terms takes into account all unique terms generated by all taxonomies using a specific method. Example: The total number of terms in Figure 4.8 is: 17 (ID: 1 to ID: 17).

$$\texttt{TotalTerms} = Count(\texttt{terms})$$

**TotalRoots**: Total number of roots indicates the number of upper terms of all taxonomies, *i.e.*, terms without hypernyms in a taxonomy. This means also the number of taxonomies generated by the method. Example: The total number of root in Figure 4.8 is: 2 (ID: 1 and ID: 14).

$$\texttt{RootTerms} = Count(isRoot(\texttt{term}_{\texttt{ij}}))$$

**NumberRels**: Number of relations extracted from the corpus without count repetitions.

$$\texttt{NumberRels} = Count(noRepetition(\texttt{rel}_{\texttt{ij}}))$$

**RatioRoots**: Ratio of root terms. The ratio between the number of roots and the total number of terms. Example: The ratio of root terms in Figure 4.8 is: (2/17) ≈ 0.12.

$$\texttt{RatioRoots} = \frac{\texttt{RootTerms}}{\texttt{TotalTerms}}$$

**TotalLeafTerms**: Total number of leaf terms. The number of terms in the bottom considering all taxonomies, *i.e.*, terms without hyponyms. Example: The total number of leaf terms in Figure 4.8 is: 8 (IDs: 2, 6, 8, 9, 11, 13 in T1 and, 16 and 17 in T2).

$$\texttt{LeafTerms} = Count(isLeaf(\texttt{term}_{\texttt{ij}}))$$

**RatioLeafTerms**: Ratio of leaf terms. Ratio between the number of leaf terms of all taxonomies and the total number of terms. Example: The ratio of leaf terms in Figure 4.8 is: (8/17) ≈ 0.47.

$$\texttt{RatioLeafTerms} = \frac{\texttt{TotalLeafTerms}}{\texttt{TotalTerms}}$$

**MaxLeafTerms**: Maximum number of leaf terms. Extract the number of leaf terms for each taxonomy and select the highest number of leaf terms generated by a taxonomy. Example: The maximum number of leaf terms in Figure 4.8 is: 6 (in T1).

$$\texttt{MaxLeafTerms} = Max(Count(isLeaf(\texttt{term}_{\texttt{ij}}))$$

**MinLeafTerms**: Minimum number of leaf terms: Extract the number of leaf terms for each taxonomy and select the lowest number of leaf terms generated by a taxonomy. Example: The minimum number of leaf terms in Figure 4.8 is: 2 (in T2).

$$\texttt{MinLeafTerms} = Min(Count(isLeaf(\texttt{term}_{\texttt{ij}})))$$

**AvgLeafTerms**: Average number of leaf terms: Ratio between the number of leaf terms of all taxonomies and the number of taxonomies. Example: The average number of leaf terms in Figure 4.8 is: (8/2) = 4.

$$\texttt{AvgLeafTerms} = \frac{\texttt{TotalLeafTerms}}{\texttt{TotalRoots}}$$

**MaxDepth**: Maximum depth extracts the longest path between a root and a leaf for each taxonomy and select the maximum value. Example: The maximum depth in Figure 4.8 is: 6 (passing by IDs: 1, 3, 4, 5, 10, 12 and 13 in T1)

$$\texttt{MaxDepth} = Max(Depth(isLeaf(\texttt{term}_{\texttt{ij}})))$$

**MinDepth**: Minimum depth extracts the shortest path between a root and a leaf for each taxonomy and select the minimum value. Example: The minimum depth in Figure 4.8 is: 1 (path between ID:1 and ID:2 in T1)

$$\texttt{MinDepth} = Min(Depth(isLeaf(\texttt{term}_{\texttt{ij}})))$$

**AvgDepth**: Average depth is the ratio between the sum of all depths and the total number of taxonomies. Example: The average depth in Figure 4.8 is: (1+4+5+4+5+6+2+2)/2 = 14.5

$$\texttt{AvgDepth} = \frac{\sum_j Depth(isLeaf(\texttt{term}_{\texttt{ij}}))}{\texttt{TotalRoots}}$$

**DepthCohesion**: The cohesion of a taxonomy is indicated by the maximum depth divided by its average depth. Example: The cohesion of the taxonomy in Figure 4.8 is: (6/14.5) ≈ 0.41

$$\texttt{DepthCohesion} = \frac{\texttt{MaxDepth}}{\texttt{AvgDepth}}$$

**MaxWidth**: Maximum width is the maximum number of term siblings in all taxonomies, *i.e.*, the maximum number of hyponyms of a term. Example: The maximum width in Figure 4.8 is: 4 (formed by IDs: 6, 7, 9 and 10 in T1)

$$\texttt{MaxWidth} = Max(Width(\texttt{term}_{\texttt{ij}}))$$

**MinWidth**: Minimum width is the minimum number of term siblings in all taxonomies, *i.e.*, the minimum number of hyponyms of a term. Example: The minimum width in Figure 4.8 is: 1 (single IDs: 4, 5, 8, 13, or 15)

$$\texttt{MinWidth} = Min(Width(\texttt{term}_{\texttt{ij}}))$$

**AvgWidth**: Average width is the ratio between the sum of widths (*i.e.*, the sum of hyponyms) and the total number of siblings (*i.e.*, the total number of terms that have hyponyms), and the number of taxonomies, where "TaxWidth$_j$" measures the width of the taxonomy $j$. Example: The average width in Figure 4.8 is: (((2+1+1+4+1+2+1)/7) + ((1+2)/2))/2 ≈ 1.61

$$\texttt{TaxWidth}_\texttt{j} = \frac{\sum_i Width(\texttt{term}_\texttt{ij}))}{Count(hasSiblings(\texttt{term}_\texttt{ij}))}$$

$$\texttt{AvgWidth} = \frac{\sum_j \texttt{TaxWidth}_\texttt{j}}{\texttt{TotalRoots}}$$

**WidthCohesion**: The width cohesion of a taxonomy is indicated by the maximum width divided by the average width. Example: The width cohesion in Figure 4.8 is: 4/1.61 ≈ 2.48

$$\texttt{WidthCohesion} = \frac{\texttt{MaxWidth}}{\texttt{AvgWidth}}$$

**ExtractedRels**: Total number of relations extracted from the corpus by the method.

$$\texttt{ExtractedRels} = Count(\texttt{rels})$$

**MaxRels**: Maximum number of relations considering separated taxonomies. Example: The maximum number of relations in Figure 4.8 is: 12 (in T1)

$$\texttt{MaxRels} = Max(Count(\texttt{rel}_\texttt{ij}))$$

**MinRels**: Minimum number of relations verifies in every taxonomy the number of relations and returns the minimum number. Example: The minimum number of relations in Figure 4.8 is: 3 (in T2)

$$\texttt{MinRels} = Min(Count(\texttt{rel}_\texttt{ij}))$$

**AvgRels**: Average number of relations is the ratio between the number of relations and the number of taxonomies.

$$\texttt{AvgRels} = \frac{\texttt{ExtractedRels}}{\texttt{TotalRoots}}$$

**SDRels**: Standard deviation of relations indicates how distributed the relations are in taxonomies.

$$\texttt{SDRels} = StdDev(\texttt{rels})$$

# Appendix F. Relation tables

This chapters presents tables containing the intersection of direct and inverse relations for all each models described in Section 5.3. These tables are interesting to see how models are complementary or dissimilar, *i.e.*, verify whether the same relations are generated by more than one method. Tables presents the ratio of similar relations and inverse relations generated by models using the top 1,000 words for each corpus. The ratio presented in each cell is computed as the number of taxonomic relations shared by the models in the row and the model in the column divided by the number of relations generated by the model in the row. For example, `Patt` model using Europarl corpus in English generated a total of 15,797 taxonomic relations from which 4,014 were shared with `DSim`. Thus, the value of the (`Patt`, `DSim`) cell is $(\frac{4{,}014}{15{,}797})$=0.2541. Table F.1 presents all ratios the for direct relations, *i.e.*, when a relation `A`→`B` in model $x$ is also found in model $y$, using the top 1,000 terms of each corpus from the English corpora.

| Corpus | | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | Patt | 1.0000 | 0.2541 | 0.2362 | 0.2357 | 0.2339 | 0.0013 | 0.0006 |
| | DSim | 0.0080 | 1.0000 | 0.4078 | 0.1328 | 0.2221 | 0.0002 | 0.0007 |
| | SLQS | 0.0075 | 0.4078 | 1.0000 | 0.6608 | 0.6552 | 0.0023 | 0.0012 |
| | TF | 0.0075 | 0.1328 | 0.6610 | 1.0000 | 0.8898 | 0.0048 | 0.0017 |
| | DF | 0.0074 | 0.2222 | 0.6553 | 0.8897 | 1.0000 | 0.0048 | 0.0020 |
| | DocSub | 0.0083 | 0.0385 | 0.4717 | 0.9905 | 1.0000 | 1.0000 | 0.0298 |
| | HClust | 0.0100 | 0.3664 | 0.6206 | 0.8288 | 1.0000 | 0.0721 | 1.0000 |
| TED Talks | Patt | 1.0000 | 0.1494 | 0.1301 | 0.1652 | 0.1634 | 0.0703 | 0.0000 |
| | DSim | 0.0002 | 1.0000 | 0.3638 | 0.0716 | 0.1978 | 0.0290 | 0.0008 |
| | SLQS | 0.0001 | 0.3669 | 1.0000 | 0.6815 | 0.6543 | 0.2110 | 0.0011 |
| | TF | 0.0002 | 0.0717 | 0.6770 | 1.0000 | 0.8331 | 0.2431 | 0.0012 |
| | DF | 0.0002 | 0.1983 | 0.6503 | 0.8336 | 1.0000 | 0.2695 | 0.0020 |
| | DocSub | 0.0003 | 0.1077 | 0.7780 | 0.9026 | 1.0000 | 1.0000 | 0.0035 |
| | HClust | 0.0000 | 0.4172 | 0.5282 | 0.5798 | 1.0000 | 0.4609 | 1.0000 |

Table F.1: Ratio of relations shared by models in the English corpora.

Table F.2 presents all ratios for the inverse relations, *i.e.*, when a relation `A`→`B` in model $x$ is inversely found (`B`→`A`) in model $y$, using the top 1,000 terms of each corpus from the English corpora. While values of Table F.1 allows to see how similar the results of the models are, Table F.2 allows to verify how dissimilar models generate relations.

Table F.3 presents all ratios the for direct relations using the top 1,000 terms of each corpus from the Portuguese corpora.

Table F.4 presents all ratios for the inverse relations using the top 1,000 terms of each corpus from the Portuguese corpora.

Table F.5 presents the relative precision for all models using each corpus in English.

Table F.6 presents the relative precision for all models using each corpus in Portuguese.

| Corpus | | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | Patt | 0.0000 | 0.2135 | 0.2314 | 0.2317 | 0.2335 | 0.0008 | 0.0009 |
| | DSim | 0.0068 | 0.0000 | 0.5922 | 0.8669 | 0.7777 | 0.0047 | 0.0013 |
| | SLQS | 0.0073 | 0.5922 | 0.0000 | 0.3389 | 0.3447 | 0.0026 | 0.0008 |
| | TF | 0.0073 | 0.8672 | 0.3390 | 0.0000 | 0.1100 | 0.0000 | 0.0003 |
| | DF | 0.0074 | 0.7778 | 0.3447 | 0.1100 | 0.0000 | 0.0000 | 0.0000 |
| | DocSub | 0.0054 | 0.9615 | 0.5283 | 0.0095 | 0.0000 | 0.0000 | 0.0000 |
| | HClust | 0.0140 | 0.6336 | 0.3794 | 0.1702 | 0.0000 | 0.0000 | 0.0000 |
| TED Talks | Patt | 0.0000 | 0.1670 | 0.1845 | 0.1511 | 0.1494 | 0.0598 | 0.0000 |
| | DSim | 0.0002 | 0.0000 | 0.6269 | 0.9258 | 0.7990 | 0.2399 | 0.0012 |
| | SLQS | 0.0002 | 0.6323 | 0.0000 | 0.3159 | 0.3425 | 0.0595 | 0.0009 |
| | TF | 0.0002 | 0.9275 | 0.3138 | 0.0000 | 0.1637 | 0.0261 | 0.0008 |
| | DF | 0.0002 | 0.8009 | 0.3404 | 0.1638 | 0.0000 | 0.0000 | 0.0000 |
| | DocSub | 0.0003 | 0.8923 | 0.2194 | 0.0968 | 0.0000 | 0.0000 | 0.0000 |
| | HClust | 0.0000 | 0.5828 | 0.4529 | 0.4103 | 0.0000 | 0.0000 | 0.0000 |

Table F.2: Ratio of inverse relations shared by models in the English corpora.

| Corpus | | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | Patt | 1.0000 | 0.2775 | 0.2741 | 0.2683 | 0.2683 | 0.0069 | 0.0005 |
| | DSim | 0.0058 | 1.0000 | 0.4573 | 0.1248 | 0.1921 | 0.0011 | 0.0008 |
| | SLQS | 0.0057 | 0.4573 | 1.0000 | 0.6340 | 0.6531 | 0.0041 | 0.0013 |
| | TF | 0.0056 | 0.1248 | 0.6341 | 1.0000 | 0.9050 | 0.0057 | 0.0015 |
| | DF | 0.0056 | 0.1922 | 0.6531 | 0.9048 | 1.0000 | 0.0067 | 0.0020 |
| | DocSub | 0.0216 | 0.1595 | 0.6085 | 0.8510 | 1.0000 | 1.0000 | 0.0339 |
| | HClust | 0.0050 | 0.4154 | 0.6627 | 0.7528 | 1.0000 | 0.1131 | 1.0000 |
| TED Talks | Patt | 1.0000 | 0.1903 | 0.2355 | 0.2742 | 0.2839 | 0.2032 | 0.0000 |
| | DSim | 0.0001 | 1.0000 | 0.4674 | 0.0546 | 0.1523 | 0.0734 | 0.0006 |
| | SLQS | 0.0001 | 0.4685 | 1.0000 | 0.5374 | 0.5435 | 0.3693 | 0.0011 |
| | TF | 0.0002 | 0.0546 | 0.5355 | 1.0000 | 0.8693 | 0.5867 | 0.0015 |
| | DF | 0.0002 | 0.1522 | 0.5420 | 0.8698 | 1.0000 | 0.6481 | 0.0020 |
| | DocSub | 0.0002 | 0.1132 | 0.5682 | 0.9058 | 1.0000 | 1.0000 | 0.0025 |
| | HClust | 0.0000 | 0.2843 | 0.5508 | 0.7335 | 0.9862 | 0.7808 | 1.0000 |

Table F.3: Ratio of relations shared by models in the Portuguese corpora.

| Corpus | | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | Patt | 0.0000 | 0.2663 | 0.2697 | 0.2406 | 0.2405 | 0.0046 | 0.0008 |
| | DSim | 0.0055 | 0.0000 | 0.5427 | 0.8749 | 0.8077 | 0.0056 | 0.0012 |
| | SLQS | 0.0056 | 0.5427 | 0.0000 | 0.3657 | 0.3468 | 0.0026 | 0.0007 |
| | TF | 0.0050 | 0.8752 | 0.3658 | 0.0000 | 0.0949 | 0.0010 | 0.0005 |
| | DF | 0.0050 | 0.8078 | 0.3468 | 0.0949 | 0.0000 | 0.0000 | 0.0000 |
| | DocSub | 0.0144 | 0.8405 | 0.3909 | 0.1487 | 0.0000 | 0.0000 | 0.0000 |
| | HClust | 0.0080 | 0.5846 | 0.3353 | 0.2462 | 0.0000 | 0.0000 | 0.0000 |
| TED Talks | Patt | 0.0000 | 0.2677 | 0.2194 | 0.1839 | 0.1710 | 0.1387 | 0.0000 |
| | DSim | 0.0002 | 0.0000 | 0.5257 | 0.9419 | 0.8437 | 0.5744 | 0.0014 |
| | SLQS | 0.0001 | 0.5268 | 0.0000 | 0.4592 | 0.4524 | 0.2758 | 0.0009 |
| | TF | 0.0001 | 0.9407 | 0.4576 | 0.0000 | 0.1267 | 0.0600 | 0.0005 |
| | DF | 0.0001 | 0.8431 | 0.4511 | 0.1268 | 0.0000 | 0.0000 | 0.0000 |
| | DocSub | 0.0001 | 0.8857 | 0.4244 | 0.0926 | 0.0000 | 0.0000 | 0.0000 |
| | HClust | 0.0000 | 0.6999 | 0.4265 | 0.2498 | 0.0000 | 0.0000 | 0.0000 |

Table F.4: Ratio of inverse relations shared by models in the Portuguese corpora.

| Corpus | | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | Patt | 1.0000 | 0.7573 | 1.0487 | 1.1853 | 1.1667 | 1.2887 | 0.0000 |
| | DSim | 2.4261 | 1.0000 | 1.0745 | 1.2516 | 1.1737 | 0.8810 | 0.9701 |
| | SLQS | 2.4433 | 0.7814 | 1.0000 | 1.1348 | 1.1469 | 1.2718 | 1.7619 |
| | TF | 2.5112 | 0.8277 | 1.0320 | 1.0000 | 1.0261 | 0.7997 | 1.5270 |
| | DF | 2.4989 | 0.7847 | 1.0544 | 1.0373 | 1.0000 | 0.8083 | 1.3891 |
| | DocSub | 3.4146 | 0.7287 | 1.4465 | 1.0002 | 1.0000 | 1.0000 | 0.0000 |
| | HClust | 0.0000 | 0.4669 | 1.1661 | 1.1113 | 1.0000 | 0.0000 | 1.0000 |
| TED Talks | Patt | 1.0000 | 0.5790 | 1.2119 | 1.3330 | 1.2060 | 1.4003 | 0.0000 |
| | DSim | 8.9374 | 1.0000 | 0.8936 | 1.1609 | 1.2440 | 2.3821 | 2.2078 |
| | SLQS | 1.2541 | 0.7045 | 1.0000 | 1.1557 | 1.1894 | 1.8145 | 1.5221 |
| | TF | 2.9779 | 0.8232 | 1.0395 | 1.0000 | 1.0730 | 1.6675 | 1.6498 |
| | DF | 3.2975 | 0.8498 | 1.0307 | 1.0337 | 1.0000 | 1.6104 | 1.5059 |
| | DocSub | 2.4563 | 1.0106 | 0.9764 | 0.9976 | 1.0000 | 1.0000 | 1.3023 |
| | HClust | 0.0000 | 1.0016 | 0.8759 | 1.0555 | 1.0000 | 1.3926 | 1.0000 |

Table F.5: Relative precision values for models using the English corpora.

| Corpus | | Patt | DSim | SLQS | TF | DF | DocSub | HClust |
|---|---|---|---|---|---|---|---|---|
| Europarl | Patt | 1.0000 | 0.5188 | 1.1610 | 1.4361 | 1.4174 | 1.5008 | 0.0000 |
| | DSim | 0.8043 | 1.0000 | 1.1710 | 1.6330 | 1.4430 | 1.1853 | 1.2662 |
| | SLQS | 1.1404 | 0.7419 | 1.0000 | 1.1991 | 1.1724 | 1.4338 | 0.9454 |
| | TF | 1.2138 | 0.8903 | 1.0319 | 1.0000 | 1.0158 | 1.2939 | 0.8468 |
| | DF | 1.2229 | 0.8031 | 1.0298 | 1.0369 | 1.0000 | 1.2217 | 0.7945 |
| | DocSub | 1.0599 | 0.5399 | 1.0309 | 1.0811 | 1.0000 | 1.0000 | 0.7626 |
| | HClust | 0.0000 | 0.7430 | 0.8756 | 0.9114 | 1.0000 | 0.9823 | 1.0000 |
| TED Talks | Patt | 1.0000 | 0.4775 | 1.1714 | 1.2574 | 1.2773 | 1.2747 | 0.0000 |
| | DSim | 0.9272 | 1.0000 | 0.7450 | 1.6527 | 1.6003 | 1.5199 | 1.0032 |
| | SLQS | 1.2478 | 0.4087 | 1.0000 | 1.0097 | 1.0082 | 1.1193 | 1.1531 |
| | TF | 1.1605 | 0.7855 | 0.8748 | 1.0000 | 1.0487 | 1.1099 | 0.8405 |
| | DF | 1.1708 | 0.7554 | 0.8675 | 1.0416 | 1.0000 | 1.0608 | 0.7395 |
| | DocSub | 1.1014 | 0.6764 | 0.9080 | 1.0392 | 1.0000 | 1.0000 | 0.7392 |
| | HClust | 0.0000 | 0.5157 | 1.0805 | 0.9091 | 1.0000 | 0.8539 | 1.0000 |

Table F.6: Relative precision values for models using the Portuguese corpora.