ExATO – High Quality Term Extraction for Portuguese and English

Lucelene Lopes*, Paulo Fernandes*† and Renata Vieira*
*Computer Science Department – PUCRS University – Porto Alegre – Brazil
†UNDL Foundation – Geneva – Switzerland
Email: {lucelene.lopes,paulo.fernandes,renata.vieira}@pucrs.br

***Abstract*—This paper presents a novel version of ExATO, a term extractor originally designed to extract relevant terms from corpora in Portuguese. In this new version not only corpora in Portuguese can be handled, but also texts in English are accepted. This extension is likely to offer the same quality pattern already achieved for Portuguese. In this paper, we draw the analysis of results in parallel corpora with respect to the intrinsic differences between Portuguese and English languages, and also the environment of usage for ExATO for Portuguese and English corpora. A brief comparison of ExATO and other similar tool is presented to illustrate the higher quality of ExATO extraction from English corpora.**

I. Introduction

The importance of term extraction from corpora for several Natural Language Processing (NLP) tasks is acknowledged by the research community [1], [2], [3], [4]. Among applications of term extraction from corpora, it is possible to mention ontology learning [5], entity profiling [6], sentiment analysis [7] and many other web intelligence needs.

In many of such applications is of equal importance to effectively detect the meaningful terms, but also to determine their relevance to the target corpus. Many works focus on term extraction, either proposing techniques to locate meaningful terms [8], [9], or to establish term relevance [10], [11], [12]. Despite the abundance of theoretical work, few extractors are freely available, and even fewer are clear about the methods employed [13], [14], [15].

In this context, ExATO software tool presents an effective and efficient solution that has been successfully employed to extract relevant terms from Portuguese corpora [16], [17], [18]. ExATO basic idea is to adopt a term extraction based on linguistic approach, consequently, working over a previously parsed and Part-of-Speech (PoS) tagged version of the target corpus. Next, ExATO employs a statistical approach to estimate the relevance of each extracted term, and to chose terms that are relevant enough to be considered representative for the target corpus. Finally, ExATO provides several output formats for language resources containing selected extracted terms, *i.e.*, terms chosen according to their relevance.

Given the linguistic-based steps within ExATO, we propose in this paper (Section II) a extension to deal not only with Portuguese, but English corpora. We show with experiments over parallel corpora (Section III) that the extraction quality of ExATO applied to English is similar to the one achieved for Portuguese. The last section (Section IV) presents a brief comparison of the extracted terms by ExATO and other available term extractors to illustrate the quality of ExATO extraction. Finally, the conclusion summarizes this paper contributions and suggests future work.

II. ExATO for Portuguese and English

ExATO is a natural language processing (NLP) software originally designed for the extraction of relevant terms from corpora written in Portuguese language [19]. Originally, ExATO was called ExATOlp, since the last two letters ("lp") stands for "língua portugesa" (Portuguese Language). Considering the addition to deal with English corpora, we decided to short its name to ExATO thereafter.

ExATO implements several NLP techniques [20] in order to provide semantically representative terms (single and multiwords). Additionally, the relevance of each term with respect to a target corpus is computed, and it makes possible to present extracted terms as language resources in different formats. Figure 1 shows a schematic representation of ExATO dealing with Portuguese and English corpora.

A. Basic Steps

While dealing with Portuguese corpora, ExATO starts receiving texts parsed using PALAVRAS [21]. From PALAVRAS' TigerXML output format, ExATO locates Noun Phrases (NP) and it applies linguistic heuristics to improve the quality of located NPs. These linguistic heuristics include traditional operations as considering lemmatized versions of the terms, and removing determiners (articles and pronouns), but it also includes sophisticated ones as adjectives composition, *e.g.*, if the NP "dragão grande e feroz" ("big and fierce dragon" in English) is found, it considers as found also the terms "dragão grande" ("big dragon") and "dragão feroz" ("fierce dragon"). The application of these heuristics achieved an improvement of both Precision and Recall from 12% to 60% for bigrams and from 10% to 50% for trigrams, as detailed in a previous work describing the extraction effectiveness with these heuristics for a Portuguese domain corpora [22].

Once the terms were located and refined by the heuristics, ExATO considers other corpora, besides the target corpus, to compute a relevance index to each extracted term. This procedure follows the idea that contrasting corpora are an effective way to establish the relevance of terms to a target corpus as

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.91

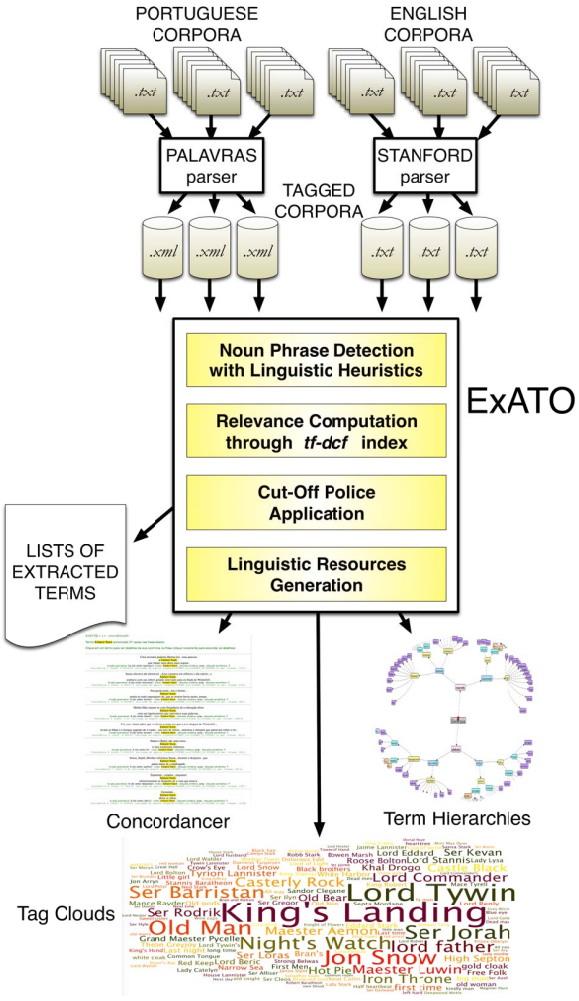

Fig. 1. ExATO overview.

a composition of their frequency and their uniqueness [23], [24]. Specifically, ExATO implements the computation of *tf-dcf* (term frequency, disjoint corpora frequency) index which is more effective than other similar options [25]. According to previous experiments, *tf-dcf* delivers an increase of precision to more than 90% with respect to the sole use of the term absolute frequency as the relevance index.

The third important technique implemented in ExATO is the automatic choice of cut-off policies in order to choose terms relevant enough to be considered representative for the target corpus [18]. Such choice of cut-off points is useful to adequately pick terms to generate specific language resources such the list of extracted terms dully sized. Such reduction of list of extracted terms allow us to call them the profile of the target corpus. In fact, the identification of representative terms may be a way to establish a profile [26]. Other language resources made available by ExATO are the generation of Tag Clouds and Term Hierarchies, as well as a Concordancer, *i.e.*, a program that locate all occurrences of a given term and its utilization contexts.

B. Adaptation for English

The more important adaptations to enable ExATO to deal with English corpora is the availability of a suitable parser to English texts, and, obviously, to provide similar heuristics to refine the extracted terms. Fortunately, there is an abundance of high quality parsers for English, in comparison with Portuguese. Among other quality options, we choose to work with the STANFORD parser, version 3.5.2 [27], which also provides NP annotation, as well as PoS-tagging.

STANFORD parser has a slightly different output information than PALAVRAS, since it does not independently provide morphological tagging, nor lemmatization, but a few lemmatization rules [28] was be incorporated in order to keep the same refinement heuristics employed for Portuguese. Among those heuristics, it is important to mention that few adaptations were required, *e.g.*, the adjective removal heuristic could be applied analogously to Portuguese by the simple consideration that in English the adjectives precede the noun. For instance, in Portuguese the term “homem velho” had the last word removed becoming “homem”, while the translated term “old man” has the first word removed becoming “man”.

Other heuristics, as the removal of determiners (articles and pronouns), were directly applied. However, some intrinsic differences between the languages are more difficult to tackle, as for instance the English NPs with the possessive ending like “King’s”. Such NP is not directly translated to Portuguese, unless it plays an adjective role as, for instance, a synonym to “Royal”, “Kingly”, or “Regal”. It is important to notice that the existence of possessive ending in the middle of a NP does not presents a problem, since the translation is usually unambiguous, as, for instance, “Night’s Watch” that is translated as a whole to the Portuguese NP “Patrulha da Noite” (literally: “Watch of the Night”).

As for the next steps of ExATO that are based on statistical approaches, very few adaptation were needed, concerning mostly ASCII extended character representation and other implementation details. As result, all available term relevance computation and linguistic resources generation were made available for English corpora.

III. Extraction in English versus Portuguese

In order to examine the effectiveness for English corpora with respect to Portuguese corpora, we have run ExATO over parallel corpora. Specifically, we took the five popular “A Song of Ice and Fire” books in the original English version [29] and the Portuguese translation [30] that we did not made public available, since it is copyright protected. Each book was considered as a corpus named respectively as Book1, Book2, Book3, Book4 and Book5. For some experiments within this paper another corpus will be used as contrastive, in this case we choose to consider the European Parliament transcripts in English and Portuguese [31] available at http://www.statmt.org/, named corpus EuroParl. Table I summarizes the information regarding these corpora stating: number of tokens, number of sentences, and extracted terms for English and Portuguese

versions submitted to ExATO in conjunction with STANFORD and PALAVRAS parsers, respectively.

	ExATO for English with STANFORD		
corpora	# tokens	# sentences	# terms
Book1	23,133,643	24,921	27,259
Book2	24,931,649	27,360	31,945
Book3	31,550,597	36,912	37,409
Book4	21,897,600	26,341	28,301
Book5	31,255,787	36,986	39,128
EuroParl	248,516,987	72,023	196,517
	ExATO for Portuguese with PALAVRAS		
corpora	# tokens	# sentences	# terms
Book1	346,539	25,018	21,022
Book2	381,494	27,413	24,204
Book3	497,720	37,787	28,922
Book4	346,949	25,804	21,965
Book5	479,924	37,691	29,606
EuroParl	2,431,920	88,621	141,179

TABLE I
PARALLEL CORPORA EXTRACTION RESULTS.

The simple observation of Table I allow us to see some important differences between STANFORD and PALAVRAS parsers. It is noticeable the difference between the number of tokens obtained for all corpora, since the English version has much more tokens. Despite the expected larger number of words in English in comparison with Portuguese, such large difference is mostly explained by the large chunks produced by PALAVRAS that considers, for instance, compounds with proper nouns as a single token. For example, the equivalents terms "Ser Barristan Selmy Lord Commander of the Kingsguard" and "Sor Barristan Selmy Senhor Comandante da Guarda Real" will be considered with eigth tokens by STANFORD parser, while PALAVRAS parser will agglutinate all words in one single token.

The number of sentences is more similar, and the only difference observed concerns the different policies to break the text in sentences. Typically, sentences containing the colon symbol (:) followed by a list of items ending by a semicolon symbol (;) will be considered one single sentence in STANFORD parser, while PALAVRAS will consider it a distinct sentence for the part before the colon symbol, and another sentence for each item.

Despite the parser differences, the essential aspects of the annotation for both parsers are similar enough with respect to tagging NPs, since the number of extracted terms is fairly similar. Given the ranking technique implemented for ExATO, we observe the similarity between terms extracted in the English and Portuguese versions.

To illustrate the terms similarity, we have made five experiments considering at each one of the books as target corpus and the other four books as contrasting corpora. The resulting ranked list of terms for each experiment represents, therefore, the more relevant terms found in each corpus.

It is important to notice that the ranked terms contrasting each book with the other four books reflect as relevance of a term its frequency and its uniqueness. Therefore, terms relevance does not represent their importance for the books generally speaking. For instance, very important terms as "Iron Throne", "House Lannister" and "Jon Snow" are not particularly relevant to any book, since they are very frequent in all five books.

term	*translation*	S	#En	P	#Pt
Ser Vardis	*Sor Vardis*	37	1^{st}	29	2^{nd}
Qotho	*Qotho*	35	2^{nd}	30	1^{st}
khas	*khas*	26	3^{rd}	23	3^{rd}
Haggo	*Haggo*	16	4^{th}	8	13^{th}
Cohollo	*Cohollo*	16	5^{th}	7	18^{th}
Desmond	*Desmond*	12	6^{th}	10	8^{th}
bride gift	*presente de noiva*	11	7^{th}	5	35^{th}
Halder	*Halder*	32	8^{th}	25	9^{th}
godswife	*esposa de deus*	10	9^{th}	15	6^{th}
men of khas	*homens de khas*	10	10^{th}	7	19^{th}

TABLE II
TOP TEN EXTRACTED TERMS FOR BOOK1.

Table II presents the top ten ranked terms for Book1 corpus in the English version, its number of occurrences found in STANFORD-based (S) and PALAVRAS-based (P) extraction, plus its rank according to ExATO processing in English (#En) and in Portuguese (#Pt).

Observing Table II, we notice that 9 of the 10 top ranked terms for English extraction are within the top 20 for Portuguese extraction (terms in bold). Only the term "bride gift" was somewhat badly ranked (35^{th} in Portuguese extraction), but such difference between ranks is explained by a particularity of translation choice of words. The term "bride gift" was translated by three different Portuguese terms: "presente de noivado" (literally "bridal gift"), "presente de casamento" (literally "wedding gift"), and "presente da noiva" (literally "bride gift").

It is important to remember that the rank of a term does not take into account solely the term frequency in the corpus, but its uniqueness as well. For instance, the term "Halder", ranked at the 8^{th} position, has 32 occurrences in Book1, which is more than the 26 occurrences of term "khas", ranked in the 3^{rd} position. However, *tf-dcf* takes into account occurrences in other corpora, and Book3 has 3 occurrences of "Halder".

For the other corpora the analysis of the top ten ranked terms also shows similarity between English and Portuguese extraction. Table III shows, analogously to Table II, how the top ten ranked terms in English were ranked in Portuguese for Book2 to Book5. In these tables, the terms that are in the top 20 more relevant terms according to Portuguese extraction are marked in bold. These terms marked in bold illustrate the similarity between Portuguese and English ranking. Therefore, since 44 of the 50 terms in Tables II and III are within the top twenty relevant terms extracted in Portuguese of its respective corpus, we state that ExATO in Portuguese and English behaves quite similarly.

Book2					Book3					Book4					Book5				
term	S	#En	P	#Pt	term	S	#En	P	#Pt	term	S	#En	P	#Pt	term	S	#En	P	#Pt
Lommy	54	1^{st}	41	5^{th}	**Tom Sevenstrings**	28	1^{st}	22	6^{th}	**Ser Hyle**	59	1^{st}	42	1^{st}	**Hizdahr**	146	1^{st}	92	2^{nd}
Ser Cortnay	22	2^{nd}	13	9^{th}	**Lem**	73	2^{nd}	58	3^{rd}	**Nimble Dick**	45	2^{nd}	29	5^{th}	**Griff**	112	2^{nd}	98	1^{st}
Black Loren	21	3^{rd}	12	10^{th}	Greenbeard	23	3^{rd}	23	162^{nd}	**Septon Meribald**	37	3^{rd}	33	3^{rd}	**Haldon**	75	3^{rd}	52	8^{th}
Esgred	18	4^{th}	12	11^{th}	**Anguy**	42	4^{th}	36	1^{st}	**Ser Creigthon**	27	4^{th}	18	13^{th}	**Reznak**	52	4^{th}	53	7^{th}
Weese	52	5^{th}	41	2^{nd}	**Lady Smallwood**	19	5^{th}	19	7^{th}	**Xhondo**	26	5^{th}	19	11^{th}	**Shavepate**	52	5^{th}	60	4^{th}
Alebelly	15	6^{th}	8	61^{th}	**Lame Lothar**	16	6^{th}	10	20^{th}	**Arianne**	108	6^{th}	96	8^{th}	**Hizdahr zo Loraq**	46	6^{th}	48	9^{th}
Ser Imry	15	7^{th}	11	13^{th}	**Jarl**	40	7^{th}	28	10^{th}	**Crabb**	24	7^{th}	17	15^{th}	Skahaz	42	7^{th}	29	23^{th}
Cressen	57	8^{th}	52	1^{st}	**Kraznys**	15	8^{th}	25	5^{th}	**Alayne**	121	8^{th}	115	4^{th}	**Green Grace**	41	8^{th}	35	16^{th}
Lady Selyse	13	9^{th}	13	8^{th}	**Lem Lemoncloak**	15	9^{th}	14	11^{th}	**Taena**	46	9^{th}	38	2^{nd}	Tattered Prince	41	9^{th}	20	34^{th}
Chiswyck	12	10^{th}	8	22^{th}	**Lothar**	15	10^{th}	15	9^{th}	**Mollander**	22	10^{th}	16	18^{th}	**Gerris**	36	10^{th}	48	10^{th}

TABLE III
TOP TEN EXTRACTED TERMS FOR ENGLISH VERSIONS OF BOOK2, BOOK3, BOOK4 AND BOOK5 CORPORA.

IV. EXATO AND SIMILAR EXTRACTORS

ExATO delivers extraction quality similar for Portuguese and English corpora, but in order to illustrate the quality of ExATO extraction in comparison with similar extractors freely available this section shows the extracted terms for a small text, actually the 39^{th} chapter of Book1.

The extractors chosen to be compared with ExATO are:

- NSP - N-gram Statistical Package, a popular extractor proposed by Banerjee and Pedersen [13] that is solely based on statistical occurrence of terms, but can have the results filtered by a stopword list, *i.e.*, a list of common words that must be ignored;
- AntConc - an concordancer tool that is capable of term extraction based on statistical techniques [14], and that provides the relevance of terms according to an index called "keyness" that takes into account co-occurrence, besides a stopword list;
- Termo Stats - a software tool capable of sophisticated term extraction taking into account both linguistic and statistical information based on maximum likelihood [15].

The target text for this term extraction comparison is the 39^{th} chapter of Book1. This chapter is composed by 3,213 words distributed in 264 sentences, and it was chosen by being a semantically important chapter in the context of the history told in the books. Therefore, we want to observe the twenty more relevant terms extracted by each tool. Table IV presents the top 20 terms according to each extractor in decrescent order of relevance.

For the results in Table IV the term extractors needed some additional specifications. For ExATO the EuroParl corpus was employed as contrastive corpus. For NSP and AntConc the stopword list was the BNC wordlist provided with AntConc distribution. Termo Stat need no additional specification.

Observing the top 20 relevant term extracted presented in Table IV, the first observation is the predominance of unigrams. However, it is noticeable that NSP and AntConc only deliverd unigrams in the top 20 terms. Probably this is a side effect of the untempered influence of term frequency to estimate the relevance. Termo Stats terms, on the contrary, are more prone to deliver complex terms with four bigrams and three trigrams.

Observing subjectively the extracted terms we notice that the term "Ned" (marked in bold) is correctly present in ExATO, NSP and AntConc lists, since this is the name of the point-of-view character narrating this chapter ("Eddard" is also present in NSP and AntConc). It is also remarkable that Termo Stats list fails to include any variations of the chapter central character's name in the top twenty terms. In fact, Termo Stats does not seems to extract proper nouns at all, which is a important handicap for Termo Stats aiming to extract relevant terms for profiling. Actually, for the examined corpora the proper names seems to be absolutely relevant to their profile.

Pushing the subjective analysis a little further, the other terms marked in bold are semantically important characters for this chapter, according to the book author. For these characters we notice that ExATO succeeds to locate them in four occurrences ("Lyanna", "Kingsguard", "Ser Arthur Dayne" and "Howland Reed"), which is only approached by AntConc list that delivers two out of these four important character.

Based on this subjective analysis, it is fair to claim some observations:

- NSP relevance solely based on term frequency is too simple approach;
- AntConc improves NSP result, but still fails to locate important compound terms;
- Termo Stats, on the contrary pushes too hard the effort to locate relevant terms, since it misses unavoidable terms as the point-of-view character that must be present in the relevant terms;
- Finally, ExATO balances well the relevance between frequency and importance.

A. Common Nouns Extraction

In order to be fair, the comparison between ExATO and Termo Stats must take into account the more relevant terms casting out proper nouns. ExATO offers the possibility to extracted only terms within specific PoS classes, for instance, only common nouns. Table V presents the top 20 terms extracted by ExATO and Termo Stats only considering common nouns, as well as the term frequency (*tf*) according to each extractor.

Observing Table V we notice that both extractors deliver similar results, since many terms occur in both top 20 lists. Eight of the top twenty terms extracted from ExATO are also in the top twenty extracted from Termo Stats (terms marked in boldface). The majority of those coincident terms is within the

rank	ExATO	NSP	AntConc	Termo Stats
1	**Ned**	**Ned**	**Ned**	Kingsroad
2	Robert	said	Ser	armor
3	Cersei	Robert	Cersei	**crannogman**
4	King	King	**Eddard**	white cloak
5	Alyn	Ser	Robert	wraith
6	**Lyanna**	Lord	King	whorehouse
7	steward	Cersei	Gerold	morrow
8	King's	Queen	Jory	bedchamber
9	Ser Gerold	Hand	Jaime	cloak
10	cup	now	Lord	Queen
11	Jaime	**Eddard**	Lannister	flagon
12	Jory	cup	**Lyanna**	King
13	**Kingsguard**	Jaime	Catelyn	Lord
14	morrow	dream	Dayne	clasp
15	Poole	wine	**Kingsguard**	sad smile
16	Rhaegar	seven	Oswell	icy courtesy
17	**Ser Arthur Dayne**	face	Rhaegar	flagon of wine
18	Lord	three	Vayon	eye of death
19	Queen	Poole	Alyn	shadow sword
20	**Howland Reed**	leg	Poole	blood-streaked sky

TABLE IV
TOP TWENTY EXTRACTED TERMS FOR ENGLISH VERSION OF CHAPTER 39 OF BOOK1 FOR EXATO AND SIMILAR EXTRACTORS.

top 10. For these eight terms the absolute frequency expressed by *tf* is the same for seven of them. However, the term "cloak" was detected 5 times by ExATO, and 4 times by Termo Stats. This indicates that ExATO detection was slightly more precise, since its occurrences in the text were in the following five excerpts:

- "*... of three knights in white cloaks ...*";
- "*... their white cloaks blowing ...*";
- "*... mantle with a cloak of black and gold squares.*";
- "*... bring the gold cloaks before the fighting began ...*";
- "*... a pocket in the lining of his cloak and tossed it ...*".

Another observation from Table V are terms as "queen", "king" and "Hand" that were quite frequent, but not necessarily well ranked. It is important to recall that the target of extraction was a small text (39^{th} chapter of Book1) with only 3,213 words. Consequently, it is hard to estimate the more relevant terms based on frequency, even thou tempered by other statistic information. Nevertheless, both extractors were similarly effective for common nouns, and of course, ExATO is also able to deal with proper nouns, that may be important to some applications as corpora profiling.

	ExATO		Termo Stat	
rank	term	*tf*	term	*tf*
1	steward	4	kingsroad	2
2	cup	7	**armor**	2
3	leg	6	**crannogman**	2
4	**morrow**	3	**white cloak**	2
5	**queen**	9	wraith	2
6	**armor**	2	whorehouse	2
7	**bedchamber**	2	**morrow**	3
8	blade	2	**bedchamber**	2
9	**crannogman**	2	**cloak**	4
10	**flagon**	2	**queen**	9
11	grandfather	2	**flagon**	2
12	Imp	2	king	17
13	king's peace	2	lord	8
14	swallow	2	clasp	2
15	**white cloak**	2	sad smile	1
16	Grace	5	icy courtesy	1
17	**cloak**	5	flagon of wine	1
18	dream	6	eye of death	1
19	lip	3	shadow sword	1
20	Hand	11	blood-streaked sky	1

TABLE V
TOP 20 EXTRACTED COMMON NOUNS.

V. CONCLUSION

The extension of ExATO to deal with English corpora was a challenge from a linguistic point of view, since the extraction heuristics had to be adapted. The benefits of the linguistic heuristics is more pronounced for the extraction of common nouns, as may be observed with the specific comparison between ExATO and Termo Stats at the end of Section IV. Nevertheless, the relevance of proper nouns is clear to profile corpora as the ones experimented in this paper.

Experiments with the five parallel corpora ("A Song of Ice and Fire" books) indicate that, despite differences between languages and parsers, the quality of extracted terms and their respective relevance from English corpora remains quite similar as the one encountered for Portuguese corpora.

The experiments with similar extractors applied to a single chapter (39^{th} chapter in Book1) analyzes subjectively the top relevant extracted terms for each extractor indicating the quality of ExATO. The concordancer, the tag clouds and the

hyperbolic trees generation made by ExATO illustrates the capabilities of ExATO output options.

Further studies with other English corpora and a deeper analysis of the effectiveness of ExATO to specific NLP applications are the natural future works for our research. Nevertheless, the experiments presented here suggest a behavior similar enough to encourage the practical application of ExATO term extraction for tasks related to machine translation. Since ExATO follows the same implementation steps for both English and Portuguese texts, it can be a reliable test bed for bilingual experiments.

Naturally, another interesting future work is the adaptation of ExATO to other languages, which can be technically facilitated by the availability of STANFORD parser for other languages than English. That being said, we believe ExATO dealing with English and Portuguese corpora already offers many options for both quality term extraction for web intelligence applications, and a fertile ground for NLP research.

VI. Acknowledgements

Lucelene Lopes is funded by CAPES and FAPERGS (Grant DOCFIX #09/12). Paulo Fernandes is partially funded by CNPq (Grants #307602/2013-3 and #459725/2014-9). Renata Vieira is partially funded by CNPq (Grant #455114/2014-5).

References

[1] K. Kageura and B. Umino, "Methods of automatic term recognition: A review," *Terminology*, vol. 3, no. 2, pp. 259–289, 1996.

[2] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 1999.

[3] L. Teixeira, G. Lopes, and R. Ribeiro, "Automatic extraction of document topics," in *Technological Innovation for Sustainability*, ser. IFIP Advances in Information and Communication Technology, L. Camarinha-Matos, Ed. Springer Boston, 2011, vol. 349, pp. 101–108.

[4] J. Wermter and U. 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*, ser. ACL-44. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 785–792.

[5] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology learning from text: methods, evaluation and applications*. IOS Press, 2005.

[6] X. Liu and H. Fang, "Entity Profile based Approach in Automatic Knowledge Finding," in *Proceedings of Text Retrieval Conference, TREC 2012*, 2012.

[7] Y. Kim and O. Zhang, "Credibility adjusted term frequency: A supervised term weighting scheme for sentiment analysis and text classification," *CoRR*, vol. abs/1405.3518, 2014. [Online]. Available: http://arxiv.org/abs/1405.3518

[8] P. Pantel and D. 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*. New York, USA: ACM Press, 2001, pp. 36–46.

[9] R. Navigli and P. Velardi, "Glossextractor: A web application to automatically create a domain glossary," in *AI*IA*, 2007, pp. 339–349.

[10] S. N. Kim, T. Baldwin, and M.-Y. Kan, "Extracting domain-specific words - a statistical approach," in *Proceedings of the 2009 Australasian Language Technology Association Workshop*, L. Pizzato and R. Schwitter, Eds. Sydney, Australia: Australasian Language Technology Association, December 2009, pp. 94–98. [Online]. Available: www.alta.asn.au/events/alta2009/proceedings/pdf/ALTA2009_12.pdf

[11] D. Wang and H. Zhang, "Inverse-category-frequency based supervised term weighting schemes for text categorization," *Journal of Information Science Engineering*, vol. 29, no. 2, pp. 209–225, 2013. [Online]. Available: http://www.iis.sinica.edu.tw/page/jise/2013/201303_02.html

[12] G. Bordea, P. Buitelaar, and T. Polajnar, "Domain-independent term extraction through domain modelling," in *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*, ser. TIA 2013. Paris, France: Université Paris Nord, 2013, pp. 61–68. [Online]. Available: http://www.insight-centre.org/sites/default/files/publications/tia2013.pdf

[13] S. Banerjee and T. Pedersen, "The design, implementation and use of the ngram statistics package," in *4th ITPCL*, 2003, pp. 370–381.

[14] L. Anthony, "Antconc (version 3.4.3) [computer software]," http://www.laurenceanthony.net/, Tokyo, Japan: Waseda University, 2014.

[15] P. Drouin, "Term extraction using non-technical corpora as a point of leverage," *Terminology*, vol. 9, no. 1, pp. 99 – 115, 2003.

[16] J. C. Reis, R. Bonacin, and M. C. Baranauskas, "Addressing universal access in social networks: An inclusive search mechanism," *Univers. Access Inf. Soc.*, vol. 13, no. 2, pp. 125–145, Jun. 2014. [Online]. Available: http://dx.doi.org/10.1007/s10209-013-0290-7

[17] M. Conrado, A. Felippo, T. Pardo, and S. Rezende, "A survey of automatic term extraction for brazilian portuguese," *Journal of the Brazilian Computer Society*, vol. 20, no. 1, p. 12, 2014. [Online]. Available: http://www.journal-bcs.com/content/20/1/12

[18] L. Lopes and R. Vieira, "Evaluation of cutoff policies for term extraction," *Journal of the Brazilian Computer Society*, vol. 21, no. 1, p. 9, 2015. [Online]. Available: http://www.journal-bcs.com/content/21/1/9

[19] L. Lopes, P. Fernandes, R. Vieira, and G. Fedrizzi, "ExATO lp – 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 '09)*. Poznan, Poland: Faculty of Mathematics and Computer Science of Adam Mickiewicz University, November 2009, pp. 427–431.

[20] L. Lopes, "Extração automática de conceitos a partir de textos em língua portuguesa," Ph.D. dissertation, PUCRS University - Computer Science Department, Porto Alegre, Brazil, 2012.

[21] E. Bick, "The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework," Ph.D. dissertation, Arhus University, 2000.

[22] L. Lopes and R. Vieira, "Heuristics to improve ontology term extraction," in *PROPOR 2012 – International Conference on Computational Processing of Portuguese Language*, ser. LNCS vol. 7243, 2012, pp. 85–92.

[23] T. M. Chung, "A corpus comparison approach for terminology extraction," *Terminology*, vol. 9, pp. 221–246, 2003.

[24] P. Drouin, "Detection of domain specific terminology using corpora comparison," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC) 2004*, M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, Eds., ELRA. Lisbon, Portugal: European Language Resources Association, May 2004, pp. 79–82.

[25] L. Lopes, P. Fernandes, and R. Vieira, "Estimating term domain relevance through term frequency, disjoint corpora frequency - tf-dcf," *Knowledge-Based Systems*, pp. –, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705115004979

[26] L. Lopes and R. Vieira, "Building and applying profiles through term extraction," in *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, ser. STIL 2015. Natal, RN, Brazil: IEEE Press, 2015, pp. 193–196.

[27] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars." in *ACL (1)*. The Association for Computer Linguistics, 2013, pp. 455–465. [Online]. Available: http://dblp.uni-trier.de/db/conf/acl/acl2013-1.html#SocherBMN13

[28] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[29] G. R. R. Martin, *A Song of Ice and Fire*. New York, USA: Bantam Books, 2013.

[30] ——, *As Crônicas de Gelo e Fogo*. São Paulo, Brazil: LeYa Brasil, 2014.

[31] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, 2005.