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

EXTRAÇÃO DE CONTEXTOS DEFINITÓRIOS A PARTIR DE TEXTOS EM LÍNGUA PORTUGUESA

IGOR DA SILVEIRA WENDT

Dissertação de Mestrado apresentada como requisito à obtenção do título de Mestre em Ciência da Computação pelo Programa de Pós-graduação da Faculdade de Informática - Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Renata Vieira
Co-orientadora: Vera Lucia Strube de Lima

Porto Alegre, Brasil
2011

W473a Wendt, Igor da Silveira
Extração de contextos definitórios a partir de textos em língua portuguesa / Igor da Silveira Wendt. – Porto Alegre, 2010.
69 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Renata Vieira.

1. Informática. 2. Heurística - Informática.
3. Engenharia de Software. I. Vieira, Renata. II. Título.

CDD 005.1

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

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

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada **"Extração de Contextos Definitórios a partir de Textos em Língua Portuguesa"**, apresentada por Igor da Silveira Wendt, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional , aprovada em 15/03/2011 pela Comissão Examinadora:

Profa. Dra. Renata Vieira– Orientadora — PPPGCC/PUCRS

Prof. Dr. Paulo Henrique Lemelle Fernandes - — PPGCC/PUCRS

Profa. Dra. Maria José Bocorny Finatto – — UFRGS/USP

Homologada em 22/06/2012., conforme Ata No. 013 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

“Se eu soubesse
o que estava fazendo,
não seria chamado pesquisa.”
Albert Einstein

“Eu diria que, bem antes de servir para comunicar,
a linguagem serve para viver.”
Benveniste

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer e dedicar este trabalho à minha mãe Rosângela, a meu pai Adão, à minha irmã Carolina e à minha tia Iolanda, por sempre terem batalhado para me dar uma educação de qualidade e por terem me dado, apoio, incentivo e amor.

À minha orientadora Renata, primeiramente por me conceder a oportunidade de cursar o mestrado, mas também pela amizade, pela paciência, pela atenção e pela dedicação. Ainda gostaria de agradecer à professora Vera, minha co-orientadora, que assim como a Renata, sempre me deu apoio no desenvolvimento do trabalho.

A todos que conheci durante o período do mestrado, em especial aos colegas e amigos Aline, Fabiana, Humberto, Christian, Vinicius, Tiago, Roger e Lucelene, pelo apoio e pelo companheirismo.

À meus amigos Mario e Aurea, pelo apoio, incentivo e confiança durante o período de graduação e mestrado.

Aos professores do PPGCC da PUCRS Paulo e Marcelo.

Aos professores do ICMC da USP São Carlos, Thiago, Sandra e Graça.

E aos professores do DL e do DC da UFSCar Ariani, Gladis, Helena e Lúcia, pelas sugestões.

Por fim, agradeço à CAPES pelo apoio financeiro durante o mestrado.

EXTRAÇÃO DE CONTEXTOS DEFINITÓRIOS A PARTIR DE TEXTOS EM LÍNGUA PORTUGUESA

RESUMO

O contexto definitório é a parte de um texto ou de um enunciado que fornece informação sobre um conceito, com base em seu uso.

A extração de contextos definitórios a partir de textos é uma tarefa importante em várias aplicações. Diversos trabalhos usam este recurso como auxílio na construção de ontologias, no desenvolvimento de material de auxílio à tradução, na elaboração de sistemas de perguntas e respostas, na criação de glossários, dicionários, entre outros.

Nesse sentido, este trabalho propõe um conjunto de heurísticas para fazer a extração automática de contextos potencialmente definitórios em textos de língua portuguesa.

Os resultados dessas heurísticas foram avaliados por terminólogos. Os resultados mostram 35,1% de F-measure quando o conjunto de heuristicas foi aplicado a um *corpus* de Geologia Geral e 51,7% de F-measure quando aplicado a um *corpus* de Química Geral. Isso proporcionou uma redução, em ambos os *corpus*, de mais de 50% da quantidade de contextos para serem analisados pelo especialista, comparando-se com a extração de contextos em um concordanciador.

Palavras-chave: Extração; Contexto Definitório; Língua Portuguesa.

DEFINING CONTEXTS EXTRACTION FROM PORTUGUESE TEXTS

ABSTRACT

The definitory context is part of a text or utterance that provides information about a concept based on its use.

The extraction of definitions from texts is an important task in various applications. Several papers present this feature as an aid in the construction of ontologies, in the development of material for aid in translation, in question answering systems, in creation of glossaries, dictionaries, among others.

Thus, this study proposes set of heuristics to make the automatic extraction of potentially definitory contexts contained in Portuguese texts.

The results of these heuristics were evaluated by terminologists and obtained 35.1 % F-measure when applied in a General Geology *corpus* and 51.7 % F-measure when applied in a General Chemistry *corpus*, reducing in both more than 50 % of the amount of contexts to be examined by a specialist compared with the contexts extraction through a concordancer.

Keywords: Extraction; Definitory Context; Portuguese Language.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

PLN	*Processamento de Linguagem Natural*
IA	*Inteligência Artificial*
SIG	*Sistema de Informação Geográfica*
QA	*Question Answering Systems*
XML	*Extensible Markup Language*
SN	*Sintagma Nominal*
PUCRS	*Pontifícia Universidade Católica do Rio Grande do Sul*
UnB	*Universidade de Brasília*
MINEROPAR	*Minerais do Paraná*
AEQ	*Área de Educação Química*
UFRGS	*Universidade Federal do Rio Grande do Sul*

SUMÁRIO

1. INTRODUÇÃO

Neste capítulo, apresentamos o contexto em que este texto foi escrito, o que nos motivou a escreve-lo e o objetivo e a organização desta dissertação.

1.1 Contexto e Motivação

Atualmente, diversas tecnologias que utilizam a linguística como base, vêm sendo desenvolvidas com o objetivo de dar suporte a diversas tarefas, entre elas a busca de informações, a elaboração de sumários, como apoio a tradutores, e a elaboração de dicionários e glossários.

Uma das etapas da elaboração de dicionários e glossários, foco deste trabalho, consiste, primeiramente, na identificação dos conceitos de um domínio e de sua descrição. A descrição de um conceito, quando extraída a partir de textos, é chamada de contexto definitório ou explicatório e tem como função contribuir para a determinação do seu significado [7].

Para desenvolver esta etapa manualmente, é necessário que especialistas do domínio des-pendam grande quantidade de tempo na busca desses contextos, o que implica em custos elevados. Devido a isso, ultimamente tem se prestado mais atenção à tarefa de automatizar o processo de extração de contextos definitórios.

Essa tarefa também é recorrente no processo de construção de ontologias, tema que tem sido foco de diversas pesquisas na área de Processamento da Linguagem Natural (doravante PLN).

Ontologia é definida por Gruber [10] como “uma especificação explícita de conceitos”. Pode-se considerar que a definição textual do conceito é uma parte dessa especificação.

Porém o processo de construção de ontologias é difícil, exigindo alternativas que auxiliem a agilizar as etapas deste processo.

Uma das alternativas para auxiliar o processo de construção de ontologias é fazer sua extração a partir de textos em que o conhecimento está representado de forma não estruturada.

Um dos elementos essenciais das ontologias são os conceitos, os quais são representados por termos. Obter uma descrição desses conceitos é uma importante etapa na construção de ontologias, e essa descrição pode ser capturada em textos, utilizando técnicas de PLN .

O processo de construção de uma ontologia envolve diversas etapas conforme apresentado na Figura 1.1 a seguir.

Ainda não existem ferramentas que realizem, automaticamente, todas as etapas de

Figura 1.1: Etapas para a construção de ontologias (adaptado de [4])

construção de uma ontologia. No entanto, diversos trabalhos têm tentado avançar em cada uma dessas etapas.

No contexto de construção de ontologias, uma das contribuições apresentadas pelo grupo de PLN da PUCRS é o extrator de termos [12]. Essa ferramenta tem a função de realizar a análise de textos, e então, extrair deles os termos mais relevantes.

Neste sentido, uma motivação para o desenvolvimento deste trabalho é que a etapa de extração de contextos definitórios complementa o trabalho de construção automática de ontologias, que está em desenvolvimento pelo grupo de pesquisa em PLN da PUCRS. A extração de contextos definitórios pode ainda auxiliar atividades desenvolvidas por linguistas e por terminólogos como, por exemplo, o desenvolvimento de material de auxílio à tradução, a elaboração de glossários específicos de um domínio e a elaboção de dicionários. Além disso, essa tarefa de extração pode também ser utilizada para sistemas de perguntas e respostas.

Outro fator que motivou este estudo é a necessidade de diferentes grupos de pesquisa, de diferentes especialidades, trabalharem em um mesmo projeto. Como nem todos possuem conhecimentos sobre um determinado domínio, é necessário que se disponibilize materiais que descrevam os conceitos da área em que estão trabalhando. Para isso, a extração de contextos definitórios agiliza a geração de um glossário, sendo este um recurso bastante importante para o esclarecimento de dúvidas sobre um domínio específico.

1.2 Objetivo da Dissertação

Nos tópicos abaixo são explicitados os objetivos deste trabalho e as atividades realizadas para seu desenvolvimento.

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é propor, implementar e avaliar um conjunto de heurísticas para extrair contextos potencialmente definitórios dos conceitos de um domínio. As heurísticas são aplicadas sobre textos em língua portuguesa anotados com informações morfossintáticas.

1.2.2 Atividades Realizadas para Atingir o Objetivo Geral

Para atingir o objetivo geral descrito acima, foram realizadas as seguintes atividades:

- Aprofundamento dos estudos sobre extração de informação (Contextos definitórios);
- Estudo de métodos e de *parsers* para lidar com XML;
- Estudo e desenvolvimento de técnicas para a identificação e extração de definições;
- Pesquisa e construição de material de referência (glossários) para avaliação dos resultados; e
- Realização de experimentos com base no material de referência e refinamento através da análise da avaliação.

1.3 Organização da Dissertação

Esta dissertação está organizada da seguinte forma: No capítulo 2, há uma revisão da literatura que trata da extração de contextos definitórios. No capítulo 3, são apresentados os detalhes sobre o material e os métodos empregados no desenvolvimento deste trabalho, bem com as medidas utilizadas para avaliar o trabalho. O capítulo 4 descreve as heurísticas desenvolvidas para fazer a identificação e a recuperação dos contextos definitórios. Além disso, esse capítulo mostra a fórmula de ranqueamento utilizada para selecionar os contextos. No capítulo seguinte, apresentamos o funcionamento do protótipo implementado que contém as heurísticas apresentadas, assim como exemplos de contextos extraídos pelo protótipo. No capítulo 6, relatamos como foi conduzida a avaliação e expomos os resultados obtidos. No capítulo 7, apresentamos quais as contribuições deste trabalho para a área, as considerações finais e a perspectiva futura de trabalho a partir deste estudo. Finalmente, nos capítulos 8, 9, 10 e 11, são apresentadas as listas de termos que foram utilizadas na avaliação. Junto há exemplos de contextos avaliados.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo, são apresentados alguns estudos realizados sobre o tema extração de definições. Além disso, incluímos aqui trabalhos desenvolvidos que abordam o tema de forma semelhante à apresentada neste trabalho.

2.1 Contextos definitórios

Geralmente, o desenvolvimento de um dicionário ou de um glossário exige que um grande volume de textos especializados em um domínio específico seja analisado para que, a partir desta análise, possam ser identificados os possíveis termos mais representativos do domínio.

Esses termos, que podem vir a ser entradas de um glossário ou dicionário, costumam ser explicados nos textos. Os trechos que contêm a explicação dos termos, chamados de contextos definitórios, podem ser identificados por um conjunto de critérios [14].

O processo de identificação de contextos definitórios, quando realizado por profissionais e sem a ajuda de *software*, demanda grande custo, visto que é necessário uma grande quantidade de profissionais e gasta-se uma grande quantidade de horas para completar a tarefa. Além disso, os contextos definitórios identificados ainda passam por outras etapas de avaliação e de preparação até constituírem o dicionário ou glossário final.

Essa dificuldade, representa uma das motivações para o desenvolvimento de métodos que agilizem esse processo. O uso de técnicas de Processamento da Linguagem Natural (PLN) e de Inteligência Artificial (IA) são indispensáveis no processamento automático ou semi-automático de *corpora*.

De acordo com J. L De Lucca [5], os contextos definitórios aparecem em três circunstâncias: a primeira é quando o autor cita um termo técnico; a segunda é quando o autor, em uma publicação científica, introduz um novo termo ou um termo pouco conhecido pela área; a terceira ocorre quando o termo é conhecido somente em uma língua e o autor informa o equivalente em sua língua nativa. Com exceção do último caso, os termos são seguidos de sua definição (mesmo que essa não siga os rigores da definição lexicográfica encontrada nos dicionários).

2.2 Trabalhos Relacionados

Diversos autores propõem o uso de padrões para a identificação de contextos definitórios. Swales [20] sugere que o padrão mais comum de definição é:

"Um/uma x,y é um/uma palavra de classe geral + palavra ... onde x é um substantivo contável, onde y é um substantivo incontável."

Allen *et al.* [1] citam exemplos de duas formas de definições comuns no discurso científico. Porém, esses autores, não apresentam maiores explicações sobre como as definições são expressas. As duas formas de definições comuns no discurso científico são, segundo os autores:

1. A [é] B que C

   A [pode ser definido como] B que C

2. B que C [é chamado] A

   B que C [é conhecido como] A

Ainda neste sentido, Flowerdew [8] diz que a presença de uma definição pode ser sinalizada tanto por artifícios sintáticos como por artifícios léxicos. Flowerdew aponta o uso de expressões léxicas sinalizadoras como "chamamos", "é chamado", "são chamados", "chamado" e "conhecido como" para fazer a identificação de contextos definitórios.

Uma vez identificados os padrões, alguns trabalhos realizam a extração de definições, de acordo com esses padrões.

Sousa *et al.* [19] propõem a utilização de uma ontologia desenvolvida para atuar juntamente com um SIG (Sistema de Informação Geográfica). Esse SIG é um sistema utilizado por profissionais de diversas áreas para fazer estudos do impacto ambiental causado pela extração de petróleo. Como alguns termos utilizados nesses estudos podem não ser conhecidos por esses profissionais, a ontologia fornece descrições desses termos, ajudando, assim, os usuários a entenderem o significado deles. A geração desse dicionário é feita através do processamento de um conjunto de conceitos presentes em ontologias do domínio de geologia, as quais foram desenvolvidas por especialistas em geologia e representam a base de conhecimento do SIG.

Existem trabalhos que visam à extração de definições de termos a partir de textos. Neste sentido, busca-se o termo seguido de sua definição, como apresentado por Del Gaudio e Branco [9]. Esses autores apresentam um sistema baseado em regras gramaticais, em que se procura identificar definições em documentos anotados com informações morfossintáticas. Os textos utilizados são escritos em língua portuguesa e pertencem a três domínios diferentes; E-learning, Tecnologia da Informação e Sociedade da Informação. As regras adotadas são baseadas em expressões regulares e têm como objetivo realizar uma busca nos documentos anotados através dos padrões linguísticos impostos pelas regras. São apresentados três grupos de regras gramaticais que abrangem a maior quantidade de definições: "*Copula definitions*", "*Verbs definitions*" e "*Punctuation definitions*". O "*Copula definition*" é utilizado para encontrar definições em que o verbo que segue o substantivo é

o verbo “ser”. O “*Verbs definition*” procura por definições em que os verbos que seguem o substantivo são diferentes do verbo “ser”. Por fim, o “*Punctuation definition*” considera somente as definições introduzidas por dois pontos (:).

Um outro trabalho é o apresentado por Przepiórkowski *et al.* [15]. Nesse trabalho, a busca de definições é feita de forma semelhante a feita no trabalho de Del Gaudi e Branco, porém essas definições estão presentes em textos do domínio de *e-learning* e são textos escritos em língua búlgara, polonesa e tcheca. Os textos são anotados com informações linguísticas e, então, para cada língua é utilizado um conjunto de regras diferentes adaptadas à linguagem em que serão aplicadas. Para a língua búlgara são aplicadas 8 regras, para a língua polonesa são utilizadas 34 regras e para a língua tcheca 147 regras. O trabalho foi desenvolvido com o propósito de apoiar o especialista na construção de um glossário. Portanto, a abrangência torna-se mais importante que a precisão, pois é mais rápido e fácil para o especialista avaliar o que foi extraído do que buscar as definições nos documentos.

Há também o trabalho de Iftene *et al.* [11], que apresenta um grupo de regras gramaticais desenvolvidas para extrair definições de textos em língua romena. Nesse trabalho foi utilizado um *corpus* de 56 documentos que tem como tema o uso de computadores na educação. Esse *corpus* contém aproximadamente 700.000 palavras e cada documento foi anotado com informações linguísticas. Com o *corpus* anotado, foi aplicado um conjunto de regras que definem um padrão específico de definições. As definições foram divididas em cinco categorias: *is_def* (definições contendo o verbo “ser”, em romeno “esta”), *verb_def* (definições contendo verbos romenos específicos, diferentes do verbo “esta”), *punct_def* (definições que usam pontuação como travessão, parênteses e vírgula), *layout_def* (definições que podem ser deduzidas pelo layout), *pron_def* (definições anafóricas) e *other_def* (definições que não podem ser incluídas nas outras categorias). As regras utilizadas para cada uma das categorias fazem uma busca por esses padrões nos documentos anotados e, então, extraem as frases que correspondem a essas regras. Os autores sugerem a utilização deste método de extração de definições para aperfeiçoar sistemas que fornecem respostas a perguntas (QA) ou, então, para a aquisição de documentos relevantes em *corpus*.

Por fim, um outro estudo é o trabalho proposto por Sclano e Velardi [17], que apresenta uma ferramenta, a TermExtract, para a extração de definições da Internet. O processo utilizado visa a construir um *corpus* sobre o domínio de interesse e, a partir dele, extrair candidatos a termos utilizando um processo híbrido que leva em consideração informações linguísticas e que decide quais os melhores candidatos a partir de critérios estatísticos. Esses candidatos a termos são submetidos a buscas genéricas na Internet, procurando por estruturas frasais que contenham o termo (T) seguido de verbos que sugiram definições como por exemplo, “*T is a*”, “*T is an*”, “*T are the*”, “*T defines*”, “*T refers to*”, “*T concerns*”, “*T is the*” e “*T is any*”. Posteriormente, as definições recuperadas passam por dois filtros. Um

é chamado de Filtro Estilístico (*Stylistic filter*), em que se procura encontrar definições bem formadas, o outro filtro é chamado de Filtro de Domínio (*Domain filter*) e visa a remover as definições que não são pertinentes ao domínio utilizado.

2.3 Considerações sobre os Trabalhos Relacionados

Analisando os trabalhos descritos no item anterior, percebe-se que a extração de definições se dá a partir de diferentes fontes como ontologias, *corpus* de diferentes domínios e ferramentas de busca na internet. Além disso, a extração de definições é utilizada em diversas aplicações e como material de apoio em diversas atividades.

Verifica-se que o uso de anotação linguística torna-se indispensável para os trabalhos que fazem extração de definições a partir de textos. Isso ocorre porque a anotação linguística fornece o suporte necessário para o desenvolvimento de regras gramaticais que façam a identificação e recuperação de contextos potencialmente definitórios.

Nesse sentido, o trabalho apresentado propõe o uso de um extrator de contextos definitórios a partir de textos anotados com informações linguísticas. O extrator se baseia em regras gramaticais previamente apresentadas por diferentes autores. Essas regras foram expandidas de acordo com uma análise realizada sobre o *corpora*.

3. MATERIAIS E MÉTODOS

Neste capítulo, são descritos os materiais, os métodos e as ferramentas utilizadas para o desenvolvimento desse trabalho.

A extração de contextos potencialmente definitórios tem várias etapas. Primeiro, o *corpus* precisa ser anotado com informações linguísticas para somente então serem extraídos os termos mais relevantes. Em seguida, são recuperados os contextos definitórios desses termos através do uso de um conjunto de heurísticas, conforme ilustrado na Figura 3.1.

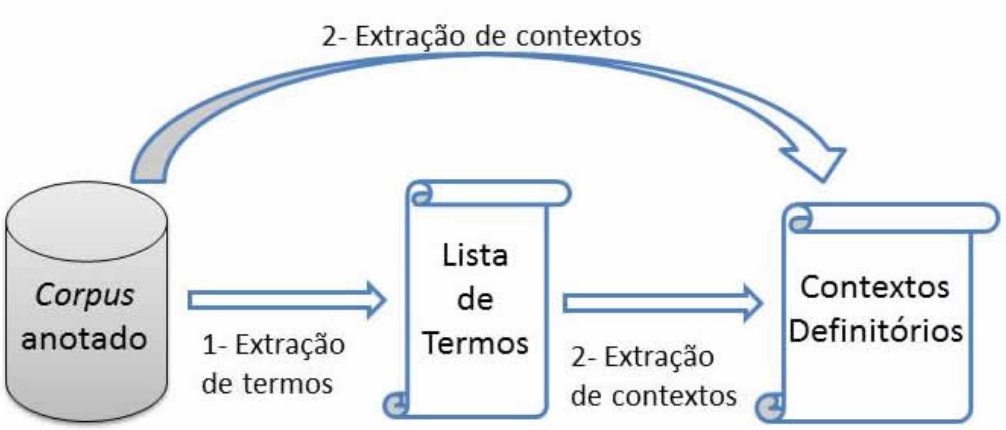


Figura 3.1: Etapas para a extração de contextos definitórios

Os detalhes sobre as ferramentas e os recursos utilizados no desenvolvimento deste trabalho são apresentados a seguir.

3.1 *Parser* PALAVRAS

A anotação linguística do *corpus* foi feita com o *parser* PALAVRAS [6]. Esse *parser* recebe um texto em formato `ASCII` e, então, faz uma análise sintática, produzindo uma árvore, em que as folhas são as palavras do texto e os terminais representam as categorias estruturais da frase.

O resultado da análise é armazenado em um arquivo em formato `XML` que contém todas as palavras do documento analisado junto com suas características morfológicas.

Um exemplo do arquivo gerado através da anotação do *parser* PALAVRAS é apresentado na Figura 3.2 mais abaixo.

3.2 Extrator de Termos

Para fazer a extração da lista de termos do *corpus* foi utilizada a ferramenta . Essa ferramenta recebe um *corpus* anotado e extrai automaticamente todos os sintagmas nominais

(SN) classificando-os segundo o número de palavras (*tokens*) que o compõem.

A Figura 3.2, a seguir, apresenta uma frase de um texto anotado pelo PALAVRAS. A frase é decomposta em *tokens* entre as *tags* "<terminals>" e "</terminals>". Cada *token* carrega consigo informações morfológicas. Entre as *tags* "<nonterminals>" e "</nonterminals>" são apresentados os SN. Os SN são identificados com o atributo cat='NP' da *tag* "<nt ...>". Em seguida, ele indica através do "id" das *tokens*, quais os *tokens* que formam o SN. Por exemplo, a *tag* <nt id='s6_503' cat='NP'> indica que os *tokens* de id='s6_1', id='s6_2' e id='s6_3' seguidos de id='s6_4' e id='s6_6' formam um SN.

Baseado em informações linguísticas, o utiliza algumas heurísticas para refinar o processo de extração de termos. Algumas destas heurísticas abaixo. Mais detalhes são apresentados no trabalho de Lopes *et al.* [12].

- SN que terminam com preposição são armazenados sem a preposição. Por exemplo, o sintagma "rocha acrescida de" é armazenado como "rocha acrescida";
- são eliminados SN que contêm números. Por exemplo, sintagmas como "década de 50", "dois estudos";
- são excluídos os SN cujo núcleo não for substantivo, uum nome próprio, ou um adjetivo. Por exemplo, sintagmas como "valor superestimado", "observado por outros";
- são excluídos os SN que iniciam com pronomes. Por exemplo, sintagmas como "estas condições", "seus acompa-nhantes", "esses dados"; e
- SN que começam com artigos são armazenados sem o artigo. Por exemplo, o sintagma "a rocha magmática" é armazenado apenas como "rocha magmática";

Desta forma, são identificados e recuperados termos com maior grau de significado no domínio em questão.

```
<s id="s6" ref="6" source="Running text" forest="1" text="O elemento lamoso FF (finos da planície de inundação)
                                                          ocorre em toda área estudada.">
    <graph root="s6_500">
        <terminals>
            <t id="s6_1" word="O" lemma="o" pos="art" morph="M S" sem="--" extra="--"/>
            <t id="s6_2" word="elemento" lemma="elemento" pos="n" morph="M S" sem="ac" extra="--"/>
            <t id="s6_3" word="lamoso" lemma="lamoso" pos="adj" morph="M S" sem="--" extra="DERS np-close"/>
            <t id="s6_4" word="FF" lemma="FF" pos="prop" morph="M S" sem="--" extra="org np-long"/>
            <t id="s6_5" word="(" lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
            <t id="s6_6" word="finos" lemma="fino" pos="adj" morph="M P" sem="--" extra="np-close"/>
            <t id="s6_7" word="de" lemma="de" pos="prp" morph="--" sem="--" extra="sam-"/>
            <t id="s6_8" word="a" lemma="o" pos="art" morph="F S" sem="--" extra=" -sam"/>
            <t id="s6_9" word="planície" lemma="planície" pos="n" morph="F S" sem="Ltop" extra="--"/>
            <t id="s6_10" word="de" lemma="de" pos="prp" morph="--" sem="--" extra="np-close"/>
            <t id="s6_11" word="inundação" lemma="inundação" pos="n" morph="F S" sem="event" extra="--"/>
            <t id="s6_12" word=")" lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
            <t id="s6_13" word="ocorre" lemma="ocorrer" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="fmc mv"/>
            <t id="s6_14" word="em" lemma="em" pos="prp" morph="--" sem="--" extra="--"/>
            <t id="s6_15" word="toda" lemma="todo" pos="pron-indef" morph="DET F S" sem="--" extra="quant"/>
            <t id="s6_16" word="área" lemma="área" pos="n" morph="F S" sem="L" extra="--"/>
            <t id="s6_17" word="estudada" lemma="estudado" pos="adj" morph="F S" sem="--" extra="np-close"/>
            <t id="s6_18" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
        </terminals>

        <nonterminals>
            <nt id="s6_500" cat="s">
                <edge label="UTT" idref="s6_501"/>
            </nt>
            <nt id="s6_501" cat="x">
                <edge label="fA" idref="s6_505"/>
                <edge label="H" idref="s6_7"/>
                <edge label="DP" idref="s6_506"/>
            </nt>
            <nt id="s6_502" cat="fcl">
                <edge label="S" idref="s6_503"/>
            </nt>
            <nt id="s6_503" cat="np">
                <edge label="DN" idref="s6_1"/>
                <edge label="H" idref="s6_2"/>
                <edge label="DN" idref="s6_3"/>
                <edge label="DN" idref="s6_504"/>
            </nt>
            <nt id="s6_504" cat="np">
                <edge label="H" idref="s6_4"/>
                <edge label="DNc" idref="s6_6"/>
            </nt>
            <nt id="s6_505" cat="pp">
            </nt>
            <nt id="s6_506" cat="np">
                <edge label="DN" idref="s6_8"/>
                <edge label="H" idref="s6_9"/>
                <edge label="DN" idref="s6_507"/>
            </nt>
            <nt id="s6_507" cat="pp">
                <edge label="H" idref="s6_10"/>
                <edge label="DP" idref="s6_11"/>
                <edge label="P" idref="s6_13"/>
                <edge label="fA" idref="s6_508"/>
            </nt>
            <nt id="s6_508" cat="pp">
                <edge label="H" idref="s6_14"/>
                <edge label="DP" idref="s6_509"/>
            </nt>
            <nt id="s6_509" cat="np">
                <edge label="DN" idref="s6_15"/>
                <edge label="H" idref="s6_16"/>
                <edge label="DN" idref="s6_17"/>
            </nt>
        </nonterminals>
    </graph>
</s>
```

Figura 3.2: Frase anotada pelo *parser* PALAVRAS

3.3 *Corpus* de Geologia Geral

Esse *corpus*, composto por 137 textos em português da área de Geologia Geral, contém aproximadamente um milhão de palavras. O *corpus* foi construído em um projeto de doutorado inserido no grupo de PLN da PUCRS.

A Tabela 3.1 apresenta os detalhes sobre a composição do *corpus*.

Tabela 3.1: Composição do *corpus* de Geologia Geral

Tipo	Nº de Textos	Nº de Palavras
Artigo	119	815.381
Tese	9	110.788
Dissertação	9	88.528

Para a construção deste *corpus* foram adotados alguns critérios a fim de que fossem coletados apenas textos científicos (artigos, teses e dissertações), os quais geralmente arquivos `.pdf`, são livres e estão disponíveis na internet. Após coletados, esse textos foram encaminhados para especialistas do domínio, que avaliaram e selecionaram os textos que eram adequados e de qualidade dentro da área.

Após a análise e seleção dos especialistas, os textos aprovados foram convertidos para arquivos no formato `.txt` com o auxílio da ferramenta Entrelinhas [18]. Em seguida, uma revisão manual do texto foi realizada para corrigir possíveis problemas que ocorrem quando se converte um arquivo de um formato para outro. O objetivo era verificar se as frases estavam coerentes. Foram retirados os títulos das seções, as referências bibliográficas, as tabelas, os gráficos e algumas seções irrelevantes para o domínio como, por exemplo, a dedicatória, os agradecimentos, os sumários, etc. Por fim, o *corpus* foi anotado com o *parser* PALAVRAS.

3.3.1 Extração de Termos

Para verificar quais termos extraídos pela ferramenta podem ser considerados conceitos do domínio, foram utilizados dois glossários de referência e uma enciclopédia *online*: o glossário da MINEROPAR[1] com 3.078 termos, o glossário da UnB[2] com 1.447 termos e a Wikipédia.

Para recuperar os termos dos glossários (*definiendum*) e suas definições (*definiens*), foram desenvolvidas expressões regulares específicas para cada *site*, pois as formas de apresentação das informações não são padronizadas.

[1] http://www.mineropar.pr.gov.br
[2] http://www.unb.br/ig/glossario/

A outra fonte de recursos, a Wikipédia[3], mesmo não sendo uma fonte específica de Geologia, contém grande quantidade de informações úteis para o domínio em questão.

Visto que a Wikipédia não é um repositório específico do domínio, alguns critérios foram adotados para selecionar os resultados mais relevantes, os quais são apresentados na seção a seguir.

3.3.2 Identificação dos conceitos

O processo utilizado no desenvolvimento deste experimento divide-se em três etapas descritas a seguir:

Etapa 1

A meta principal dessa etapa é a extração de termos candidatos à obtenção de contexto definitórios.

O primeiro passo do processo consiste em extrair do *corpus* de Geologia Geral uma lista de termos relevantes através do .

Este processo gerou uma lista com 4.889 termos. Desses, 1.556 são unigramas (termos com 1 *token*), 2.237 são bigramas (termos com 2 *tokens*) e 1.096 são trigramas (termos com 3 *tokens*).

Etapa 2

Nesta etapa, foram utilizados dois glossários específicos do domínio, os quais estão disponíveis publicamente na internet, um pertencente a MINEROPAR e o outro a UnB.

Como nem todos os termos da lista extraída do *corpus* foram encontrados nesses glossários, foi realizada uma busca específica desses termos na Wikipédia. Dessa forma, foi possível encontrar diversas definições para o mesmo termo. Logo, a terceira e última etapa consiste em escolher o termo com a definição mais adequada ao domínio.

Etapa 3

Inicialmente, os dois glossários específicos são considerados corretos. Assim, para cada termo da lista extraída que possui uma definição em um dos glossários, não são consideradas as definições da Wikipédia. O glossário da UnB, por ser um trabalho acadêmico, é considerado mais adequado. Por isso, suas definições são priorizadas às do glossário da MINEROPAR.

No caso de definições que existem somente na Wikipédia, faz-se uma escolha entre as definições disponíveis através de um cálculo do índice de pertinência ao domínio. Este

[3] http://pt.wikipedia.org

índice é calculado pelo o número de termos extraídos do *corpus* e presentes no texto de cada uma das definições, dividido pelo total de palavras do texto da definição. Desta forma, escolhe-se como melhor definição aquela em que outros termos também extraídos do *corpus* são mais frequentes.

Finalmente, caso um termo extraído não possua nenhuma definição, ele é descartado.

A lista final de termos extraídos do *corpus*, com definição em algum dos glossários consultados, tem 926 unigramas, 458 bigramas e 142 trigramas.

Procurou-se, na Wikipédia, definições dos termos extraídos que não tinham definições nos glossários. Com isso, o número de unigramas, bigramas e trigramas para os quais havia uma definição na wikipedia subiu para 1.367, 488 e 268 termos, respectivamente.

A partir desta etapa, são empregados um conjunto de heurísticas desenvolvidas para realizar a extração de contextos potencialmente definitórios. As heurísticas desenvolvidas são apresentadas em detalhes no capítulo 4.

Como subproduto deste trabalho, foi gerado um glossário de referência específico do *corpus* de Geologia Geral. Um artigo sobre esta etapa foi publicado no 3º Seminário de Pesquisa em Ontologia no Brasil (ONTOBRAS) [21].

3.4 *Corpus* de Química Geral

Este *corpus* é composto por 8 textos da área de Química Geral, sendo 4 da obra de ATKINS [2] e 4 da obra de RUSSEL [16]. Esses textos são compostos por uma seleção dos capítulos mais relevantes para o conhecimento da área de Química Geral, e esse *corpus* foi desenvolvido pela equipe TEXTQUIM do Instituto de Letras e pela equipe da Área de Educação Química (AEQ) da Universidade Federal do Rio Grande do Sul .

Esse *corpus* foi anotado com o *parser* PALAVRAS.

3.4.1 Lista de Termos

O projeto TEXTQUIM oferece um banco de expressões e de termos técnicos para auxiliar as tarefas de tradução, redação, revisão e ensino de tradução. O banco disponibilizado no *site* TextQuim[4] possui atualmente 513 termos técnicos. Juntamente com os termos é apresentada a parte do texto em que o mesmo se encontra. Desses 513 termos, 295 destes apresentam uma definição.

Esses 295 termos estão divididos em 215 bigramas, 78 trigramas e 2 quadrigramas.

3.5 Medidas de avaliação

Nessa avaliação, utilizamos as métricas de precisão, abrangência e F-measure.

[4] http://www6.ufrgs.br/textquim/

O cálculo da Precisão se dá pela intersecção entre a Lista de Referência (LR), ou seja, pelos contextos presentes no *corpus*, e a Lista Extraída (LE), os contextos recuperados através das heurísticas, dividida pela Lista Extraída (LE), conforme a fórmula abaixo.

$$P = \frac{|LR \cap LE|}{|LE|}$$

O cálculo da Abrangência é definido pela intersecção entre a Lista de Referência (LR) e a Lista Extraída (LE) dividida pela Lista de Referência (LR), conforme apresentado abaixo:

$$A = \frac{|LR \cap LE|}{|LR|}$$

A F-Measure é dada pelo dobro da multiplicação entre o resultado da Precisão (P) e da Abrangência (A), dividido pelo resultado da Precisão (P) mais o resultado da Abrangência (A), conforme a fórmula abaixo. O resultado da F-Measure indica uma medida harmônica entre os resultados obtidos pela Precisão e Abrangência.

$$F = \frac{2 \times P \times A}{P + A}$$

4. EXTRAÇÃO DE CONTEXTOS DEFINITÓRIOS

Neste capítulo, são descritos os métodos desenvolvidos para realizar a extração de contextos definitórios a partir dos termos previamente identificados.

4.1 Pré-Processamento

Inicialmente, o *corpus* é anotado com o *parser* PALAVRAS. Então para realizar a leitura do documento XML gerado pelo *parser* é utilizada a API JDOM, que faz o processamento de XML através do parser JAXP.

O parser JAXP só admite documentos XML bem formados, sem erros em TAGS. Porém, o *parser* PALAVRAS, em algumas situações insere símbolos que causam erros de formação no arquivo, conforme exemplo abaixo:

Exemplo: <edge label="NUM<" idref="s215_506"/>

No exemplo acima, o símbolo "<", inserido ao lado de "NUM", causa erro de formação, pois ele duplica o símbolo de início de uma TAG.

Ainda existem casos em que alguns termos são descartados pelo *parser*. Nesses casos, eles são inseridos dentro de uma TAG nomeada "lixo", que não possui marcação de final, ocasionando problemas de formação também.

Exemplo: <lixo uranona,>

<lixo 3,3-bis(4-hidroxifenil)-1-(3H)-isobenzof>

Devido aos problemas citados (que são os mais comuns) e ainda outros, foi desenvolvido um script que varre os documentos em busca desses erros de formação e os altera de forma que o documento tenha uma estrutura consistente.

Após estruturar os documentos anotados corretamente pelo *parser* PALAVRAS, são pesquisados os termos candidatos a serem entradas de um glossário ou dicionário, nos documentos. Após identificá-los, verifica-se, através das heurísticas, se estão em contextos potencialmente definitórios, conforme apresentado na seção 4.2.

Ao pesquisar por termos compostos (bigramas e trigramas) no documento anotado, é necessário verificar:

- Primeiro: se o termo não foi anotado pelo *parser* como um único termo, ou seja, se duas ou mais palavras deste termo estão unidas por "_". Em alguns casos, o *parser* une os termos acrescentando "_".

  Exemplo: <t id="s20_21" word="rochas_ígneas" lemma="rocha_ígnea" pos="n" morph="FP" sem="mat" extra="cjt-Od"/>

- Segundo: se o termo foi anotado pelo *parser* separadamente, ou seja, em dois termos individuais.

  Exemplo: <t id="s292_66" word="rochas" lemma="rocha" pos="n" morph="F P" sem="cc-stone" extra="–"/>

  <t id="s292_67" word="vulcânicas" lemma="vulcânico" pos="adj" morph="F P" sem="–" extra="np-close"/>

4.2 Padrões de Contextos Definitórios

Esta etapa do trabalho tem como finalidade identificar padrões de contextos definitórios através de heurísticas baseadas em [1], [8], [9], [11], [15], [17], [20] e discutidas no capítulo 2.

Para a identificação dos padrões, foi desenvolvido um concordanciador para apresentar os contextos que continham os termos utilizados. O concordanciador é uma ferramenta utilizada para listar as ocorrências de uma palavra ou frase.

Em seguida, foi realizada uma análise manual, através de uma leitura sistemática dos contextos recuperados. Os padrões identificados nessa análise foram divididos em quatro grupos, apresentados a seguir. As heurísticas descritas nestes padrões foram implementadas em linguagem Java, utilizando o *parser* JAXP, conforme explicado na seção 4.1.

- Padrões sintáticos:

  Os padrões sintáticos apresentam apenas uma forma sintática. O predicado verbal utilizado foi o verbo "ser" e suas flexões.

  Este padrão recupera somente contextos em que o termo seja diretamente seguido do verbo "ser" ou de suas flexões, ou seja, verifica-se no documento anotado se os atributos desse verbo são: lemma="ser" e pos="v-fin".

  Heurística 1: Verbo "Ser" e suas flexões:

  Exemplo: "A sismo estratigrafia é um método estratigráfico de análise e interpretação de dados sísmicos, utilizado no estudo e compreensão da evolução tectono-sedimentar de uma bacia, visando subdividir, correlacionar e mapear pacotes de rochas sedimentares(...)"

- Padrões tipográficos:

  Neste padrão, é verificado no documento anotado se o atributo da palavra que segue o termo é word=":" ou word="(".

  Este padrão recupera somente os contextos em que o termo seja diretamente seguido de dois pontos ou de parênteses.

Heurística 2: ":"

Exemplo: "Granulometria: Medição do tamanho dos grãos que compõem uma rocha sedimentar."

Heurística 3: "( )"

Exemplo: "Elementos com baixas eletronegatividades (tais como, os metais do bloco s) são freqüentemente chamados de eletropositivos."

Caso o termo seja seguido de ":", o contexto apresentado é a frase seguinte, pois quando o *parser* encontra ":", ele automaticamente quebra a linha, mesmo que no documento original não haja quebra de linha.

- Padrões verbais:

Este padrão tem por finalidade utilizar verbos que indiquem a presença de um possível contexto definitório. Nesse padrão, não é necessário que o termo seja diretamente seguido de verbos, basta que o contexto contenha o termo e um dos verbos apresentados abaixo anotados no atributo "lemma".

Heurística 4: Verbo "Chamar" e suas flexões

Exemplo: "Kps é também chamado de constante do produto de solubilidade ou simplesmente de constante de solubilidade."

Heurística 5: Verbo "Formar" e suas flexões

Exemplo: "A fácies Stb é formada por estratos de 10 a 20 cm de espessura, compostos por areia fina bem selecionada, com estratificação cruzada tangencial na base e truncamento no topo por superfícies erosivas, normalmente planares, com direção de mergulho para WNW, S e SSE."

Heurística 6: Verbo "Compor" e suas flexões

Exemplo: "As fácies de praia são compostas por areias quartzosas claras, finas, bem selecionadas, apresentando estratificações bem desenvolvidas que são truncadas eventualmente por tubos de ophiomorpha (Callichirus sp)."

Heurística 7: Verbo "Constituir" e suas flexões

Exemplo: "A sequência pelítica é constituída por granada-muscovita / biotitaxistos, ricos em veios e/ ou lentes de quartzo relativamente homogêneos, por vezes feldspáticos, chegando a apresentar camadas de paragnaisses."

Heurística 8: Verbo "Denotar" e suas flexões

Exemplo: "A deposição da SEQ-B4 se dá de maneira ampla e abrangente por toda a bacia, e sua não-ocorrência em determinadas regiões é atribuída a posteriores

erosões, denotando um padrão de marcante preenchimento e transbordamento dos sistemas de meio-gráben criados na fase rifte."

Heurística 9: Verbo "Mostrar" e suas flexões

Exemplo: "Vale destacar que as rochas com andaluzita mostram como acessórios, geralmente como inclusões, zircão, monazita, rutilo e grafita."

Heurística 10: Verbo "Representar" e suas flexões

Exemplo: "Cada par compartilhado conta como uma ligação covalente e é representado por uma linha entre os dois átomos."

Heurística 11: Verbo "Definir" e suas flexões

Exemplo: "A estratigrafia de seqüências pode ser definida como o estudo dos estratos sedimentares geneticamente relacionados, situados entre duas superfícies crono estratigraficamente relevantes."

Heurística 12: Verbo "Consistir" e suas flexões

Exemplo: "De acordo com a teoria da valência (VB), a ligação covalente consiste num par de elétrons compartilhados em dois átomos ligados."

Heurística 13: Verbo "Indicar" e suas flexões

Exemplo: "A estratigrafia da Barreira III, indica uma seqüência progradante (regressiva) composta por sedimentos praiais quartzosos, finos e claros, bem selecionados e estratificados recobertos por areias eólicas."

Heurística 14: Verbo "Significar" e suas flexões

Exemplo: "Uma entalpia de ligação alta significa que o poço de energia é profundo e que uma grande quantidade de energia é necessária para quebrar a ligação."

Heurística 15: Verbo "Simbolizar" e suas flexões

Exemplo: "O número quântico de Spin, é simbolizado pela letra S."

Heurística 16: Verbo "Caracterizar" e suas flexões

Exemplo: "Sedimentos mineralogicamente maturos, são caracterizados por o elevado teor de quartzo e, normalmente, tiveram sua composição modificada a partir de a sua fonte original, causando perdas substanciais das informações inerentes à proveniência."

Heurística 17: Verbo "Conter" e suas flexões

Exemplo: "Cristais de plagioclásio contêm freqüentes inclusões de zircão, apatita e mi-nerais opacos e mostram feições de recristalização como extinção ondulante e lamelas de geminação acunhadas ou recurvadas."

Heurística 18: Verbo “Apresentar” e suas flexões

Exemplo: “A fácies h apresenta aspecto geral opaco, com refletores fortes, semi-contínuos, delineando superfícies onduladas paralelas.”

- Padrões indicativos:

Neste padrão são utilizadas expressões que indicam uma explicação prévia de determinado termo como, por exemplo, “Conhecido como”, “Reconhecido como” e “Isto é”, que indica uma explicação ou introduz o termo sobre o qual se discute.

Não é necessário que o termo seja diretamente seguido destas expressões, basta que o contexto contenha o termo e a expressão para que possa ser recuperado.

Nesse padrão, verifica-se no documento anotado se o contexto recuperado contém o termo buscado e as palavras anotadas como lemma=“conhecer” seguido de lemma=“como” ou lemma=“reconhecer” seguido de lemma=“como” ou word=“isto” seguido de word=“é”.

Heurística 19: Expressão “Conhecido como”

Exemplo: “O Membro Outeiro da Formação Macaé reúne, além de calcilutito creme, marga cinza-clara e folhelhos cinza, are-nitos turbiditos informalmente conhecidos como arenitos Namorado, que por vezes ocorrem em camadas isoladas ou confinados em calhas deposicionais, as quais subsidiram diferencialmente em resposta à halocinese Arenitos Namorado pertencente ao Membro Outeiro da Formação Macaé.”

Heurística 20: Expressão “Isto é”

Exemplo: Durante as épocas mais úmidas, isto é, aquelas em que o balanço hídrico é positivo e existe grande aporte sedimentar trazido por as correntes fluviais, o potencial para a taxa de sedimentação superar a taxa de subsidência é grande, ocorrendo a progradação na maior parte da bacia.”

Heurística 21: Expressão “Reconhecido como”

Exemplo: “Em a análise de testemunhos efetuada em dois poços na região do Campo de Merluza, foram reconhecidos, como principal litofáciesreservatório, os arenitos maciços de granulometria fina a grossa e seleção pobre a moderada do Membro Ilhabela.”

4.3 Tratamento de pronomes

Frases iniciadas por pronome demonstrativo geralmente indicam que algo foi dito anteriormente. Assim, essas frases podem auxiliar a complementar um contexto definitório previamente recupe-rado por alguma das heurísticas apresentadas na seção 4.2.

Para tal, foi implementado, no protótipo (capítulo 5), uma opção para apresentar a frase anterior ao contexto definitório recuperado, caso esse contexto iniciasse por pronome demonstrativo.

Quando um contexto é identificado, verifica-se no documento anotado, se o atributo da primeira palavra é pos="pron-dem". Se sim, então é apresentada a frase anterior.

Exemplo:

Contexto extraído: Esta **fácies** pode representar porções proximais dos lobos, um pouco confinadas devido a íntima associação com as Fácies L2 (laminações e truncamentos) e L5.

Frase anterior: Arenitos Maciços e Estratificados Sucedem a deposição de fácies conglo-meráticas ao longo de calhas ou depressões.

No exemplo acima, o contexto extraído para "fácies" foi identificado através da heurística 10. Visto que esse contexto inicia por pronome demonstrativo, é possível apresentar a frase anterior.

Este procedimento não foi avaliado junto com as outras heurísticas, pois para isso seria necessário fazer uma avaliação à parte para esses casos.

4.4 Ranqueamento

Considerando que os contextos definitórios mais relevantes geralmente apresentam o *definiendum* como sujeito da frase, uma fórmula foi desenvolvida para pontuar a posição do termo na frase.

Essa fórmula leva em consideração a posição do termo na frase. Assim, quanto mais no começo da frase estiver o termo, maior será seu peso, visto que a estrutura frasal mais comum é SVO (Sujeito, Verbo, Objeto) ou SOV (Sujeito, Objeto, Verbo) [3].

A pontuação dada pela fórmula varia de 0,0000 a 1,0000.

Abaixo, temos a fórmula, onde "*A*" é o total de termos da frase e "*B*" a quantidade de termos presentes antes do termo em questão.

$$\chi = \frac{A - B}{A}$$

O uso desta fórmula permite fazer um *ranking* das definições através da pontuação que o *definiendum* recebe, facilitando a visualização e a seleção de contextos mais relevantes.

5. PROTÓTIPO

Neste capítulo, é apresentado o protótipo chamado ExContext. Esse protótipo contém as heurísticas apresentadas, um concordanciador e duas medidas de posição do termo.

5.1 Interface

O objetivo deste protótipo é facilitar extração de contextos definitórios. Para isso, foram implementadas as heurísticas, previamente apresentadas, em uma interface de simples interação, conforme a Figura 5.1.

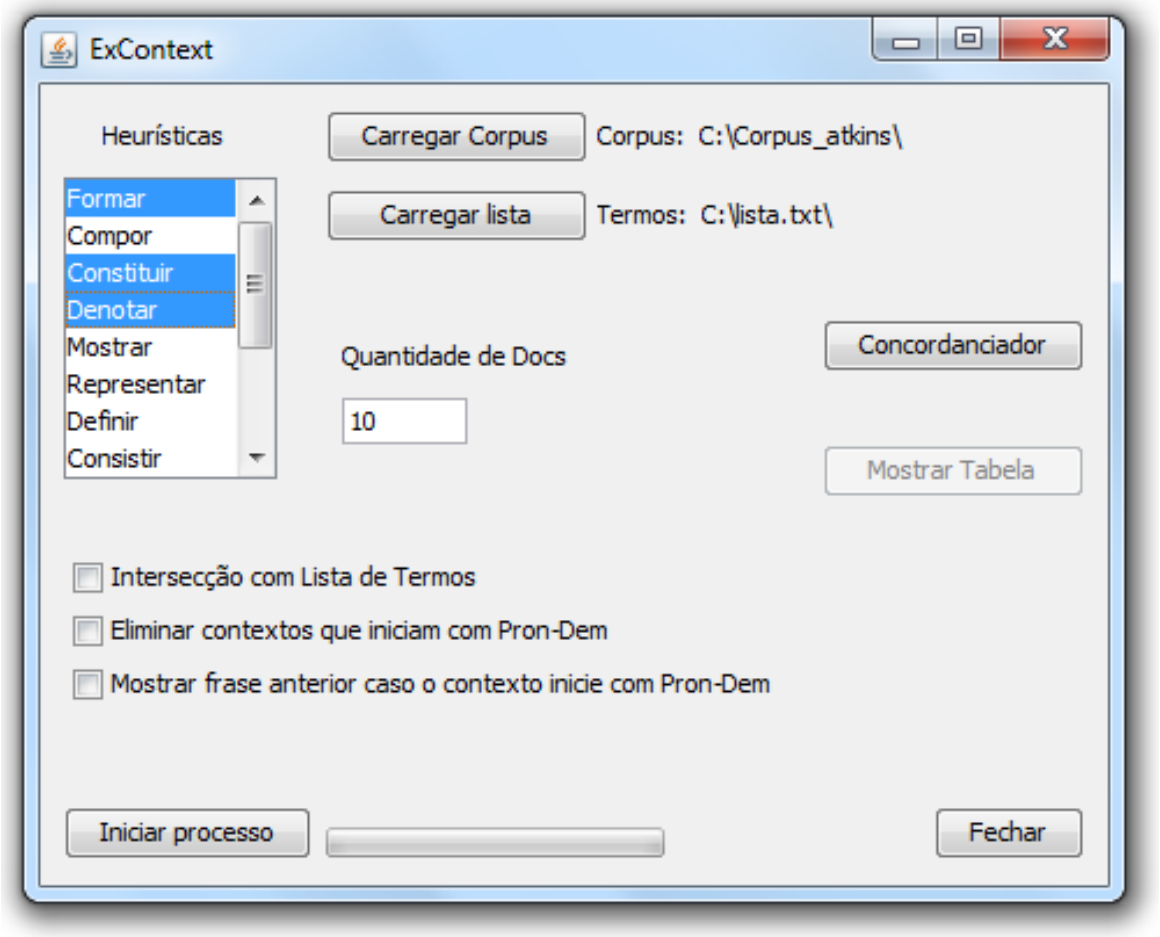


Figura 5.1: Tela inicial do protótipo ExContext

O protótipo ainda pode ser útil para apoiar algumas atividades desenvolvidas por linguistas e por terminólogos, como, por exemplo, a elaboração de glossários e de dicionários.

5.2 Funções

Na interface principal, o usuário pode selecionar o *corpus* em que serão aplicadas as heurísticas e a lista dos termos para os quais se pretende extrair os contextos definitórios.

À esquerda, é apresentada a lista de heurísticas disponíveis. Assim, é possível escolher uma ou mais, segurando a tecla “CTRL” e clicando sobre a heurística. O usuário pode utilizar os verbos *formar, compor, constituir, conter, denotar, mostrar, representar, apresentar, caracterizar, definir, consistir, indicar, significar* e *simbolizar* sem utilizar uma lista de termos, obtendo, desse modo, uma tabela de contextos com esses verbos, conforme aparece na Figura 5.2

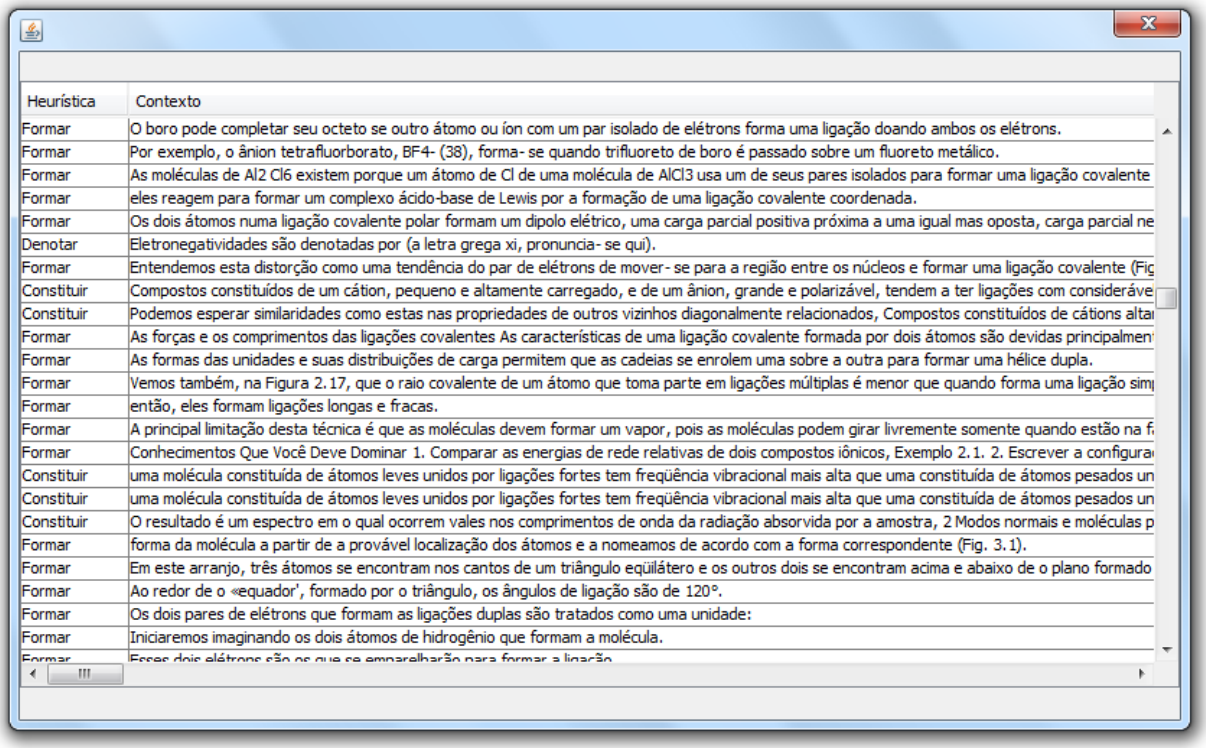
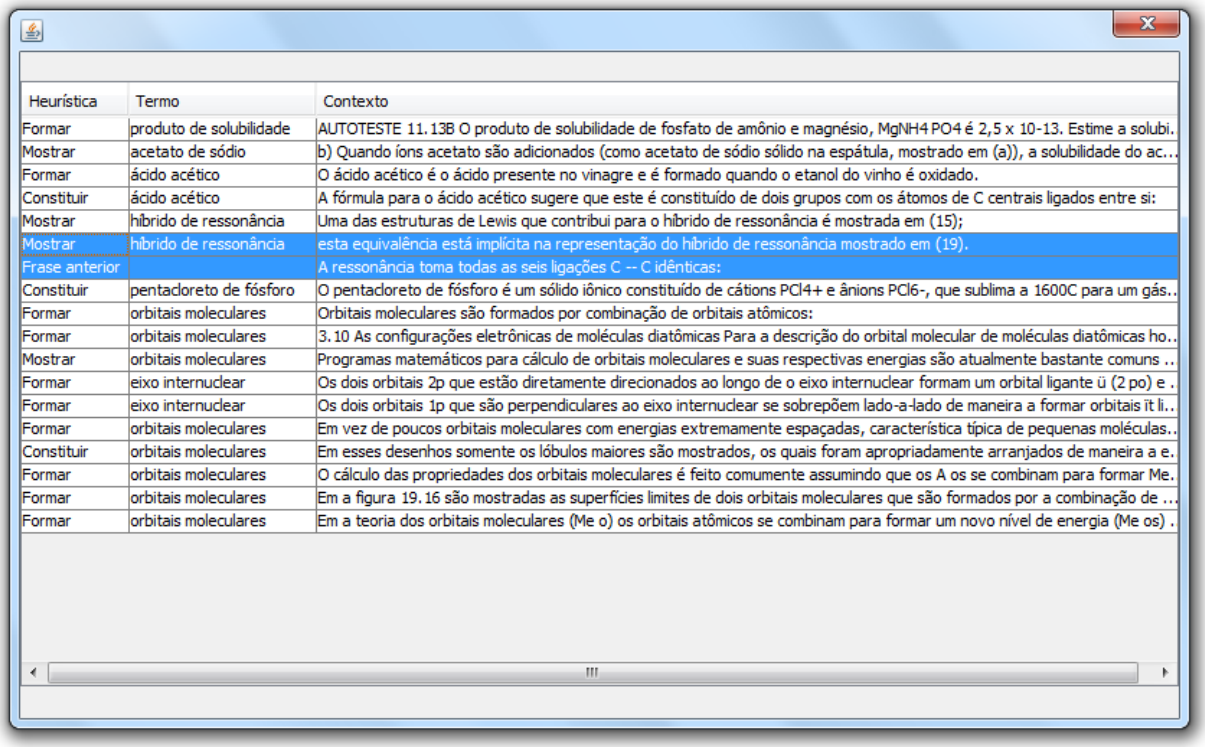


Figura 5.2: Tabela de contextos sem utilizar uma lista de termos

Marcando a opção *“Intersecção com Lista de Termos”* são extraídos somente os contextos que se enquadram em alguma das heurísticas selecionadas junto com algum dos termos da lista. As heurísticas “Ser”, “:” e “()”, só podem ser utilizadas com essa opção selecionada.

Ainda é possível exibir a frase anterior, caso o contexto extraído inicie por pronome demonstrativo, selecionando a opção *“Mostrar frase anterior caso o contexto inicie com Pron-Dem”*, conforme apresentado na Figura 5.3. Além disso, pode-se excluir o

contexto iniciado por pronome demonstrativo, marcando a opção *"Eliminar contextos que iniciam com Pron-Dem"*.



Heurística	Termo	Contexto
Formar	produto de solubilidade	AUTOTESTE 11.13B O produto de solubilidade de fosfato de amônio e magnésio, MgNH4 PO4 é 2,5 x 10-13. Estime a solubi.
Mostrar	acetato de sódio	b) Quando íons acetato são adicionados (como acetato de sódio sólido na espátula, mostrado em (a)), a solubilidade do ac...
Formar	ácido acético	O ácido acético é o ácido presente no vinagre e é formado quando o etanol do vinho é oxidado.
Constituir	ácido acético	A fórmula para o ácido acético sugere que este é constituído de dois grupos com os átomos de C centrais ligados entre si:
Mostrar	híbrido de ressonância	Uma das estruturas de Lewis que contribui para o híbrido de ressonância é mostrada em (15);
Mostrar	híbrido de ressonância	esta equivalência está implícita na representação do híbrido de ressonância mostrado em (19).
Frase anterior		A ressonância toma todas as seis ligações C -- C idênticas:
Constituir	pentacloreto de fósforo	O pentacloreto de fósforo é um sólido iônico constituído de cátions PCl4+ e ânions PCl6-, que sublima a 1600C para um gás..
Formar	orbitais moleculares	Orbitais moleculares são formados por combinação de orbitais atômicos:
Formar	orbitais moleculares	3.10 As configurações eletrônicas de moléculas diatômicas Para a descrição do orbital molecular de moléculas diatômicas ho..
Mostrar	orbitais moleculares	Programas matemáticos para cálculo de orbitais moleculares e suas respectivas energias são atualmente bastante comuns ..
Formar	eixo internuclear	Os dois orbitais 2p que estão diretamente direcionados ao longo de o eixo internuclear formam um orbital ligante ü (2 po) e .
Formar	eixo internuclear	Os dois orbitais 1p que são perpendiculares ao eixo internuclear se sobrepõem lado-a-lado de maneira a formar orbitais it li..
Formar	orbitais moleculares	Em vez de poucos orbitais moleculares com energias extremamente espaçadas, característica típica de pequenas moléculas..
Constituir	orbitais moleculares	Em esses desenhos somente os lóbulos maiores são mostrados, os quais foram apropriadamente arranjados de maneira a e.
Formar	orbitais moleculares	O cálculo das propriedades dos orbitais moleculares é feito comumente assumindo que os A os se combinam para formar Me.
Formar	orbitais moleculares	Em a figura 19.16 são mostradas as superfícies limites de dois orbitais moleculares que são formados por a combinação de ...
Formar	orbitais moleculares	Em a teoria dos orbitais moleculares (Me o) os orbitais atômicos se combinam para formar um novo nível de energia (Me os) .

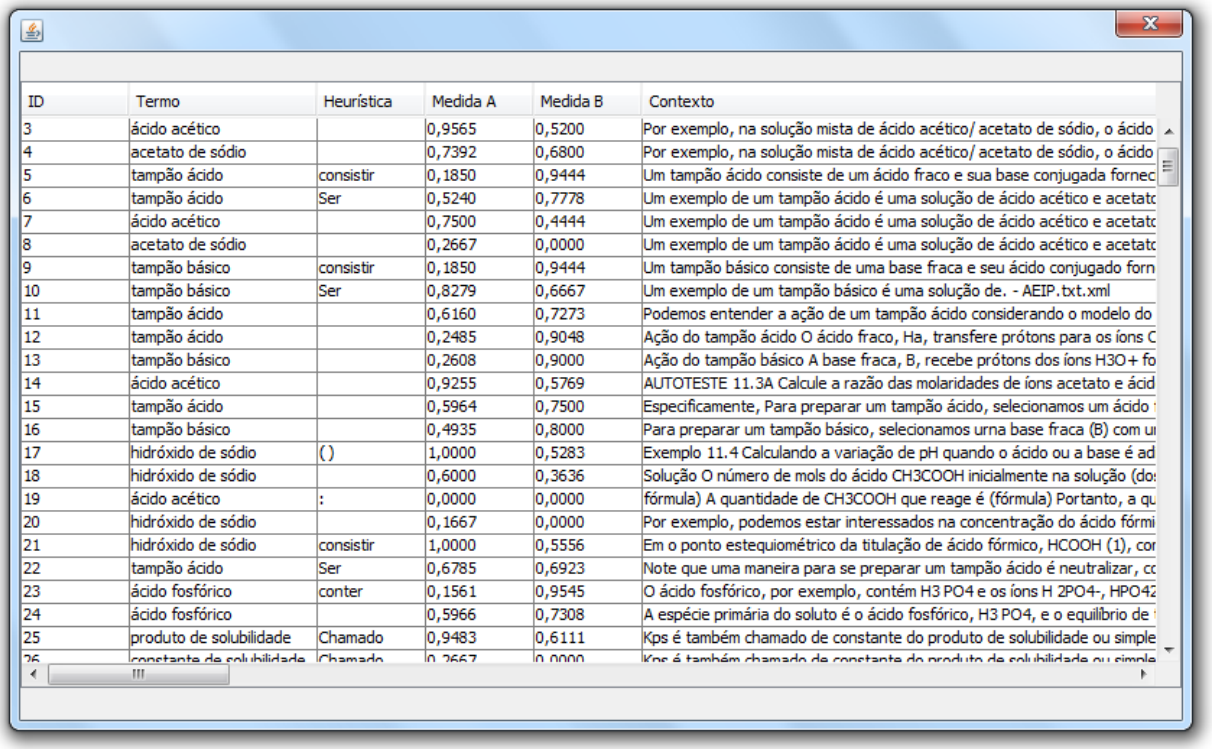
Figura 5.3: Tabela de contextos extraídos para uma lista de termos

O protótipo ainda possui um concordanciador, que só funciona com a utilização de uma lista de termos. Além disso, no concordanciador, o usuário pode escolher a quantidade de documentos que deseja analisar. Ele não precisa trabalhar com o corpus inteiro; pode selecionar alguns documentos do corpus com os quais quer trabalhar e proceder a pesquisa.

Ao clicar no botão *"concordanciador"* será processada a quantidade de documentos escolhida e será apresentada uma tabela com os termos presentes na lista, seguida da heurística (caso se enquadre em alguma), da medida de centralização do termo (Medida A), da fórmula de ranqueamento (Medida B), apresentada no capítulo 4, e do contexto que o termo está inserido. A Medida A é utilizada para auxiliar tradutores. Ela pontua a relevância do uso do termo em uma frase e avalia a posição do termo na frase. Quanto mais centralizado o termo estiver na frase, maior é o seu valor, o qual, varia de 0,000 a 1,000.

A tabela gerada é apresentada pela Figura 5.4.

Ao clicar na descrição de qualquer coluna, a tabela é ordenada de forma crescente ou decrescente, permitindo que diferentes combinações de ordenação sejam executadas para que o usuário possa filtrar o que deseja.



ID	Termo	Heurística	Medida A	Medida B	Contexto
3	ácido acético		0,9565	0,5200	Por exemplo, na solução mista de ácido acético/ acetato de sódio, o ácido
4	acetato de sódio		0,7392	0,6800	Por exemplo, na solução mista de ácido acético/ acetato de sódio, o ácido
5	tampão ácido	consistir	0,1850	0,9444	Um tampão ácido consiste de um ácido fraco e sua base conjugada fornec
6	tampão ácido	Ser	0,5240	0,7778	Um exemplo de um tampão ácido é uma solução de ácido acético e acetatc
7	ácido acético		0,7500	0,4444	Um exemplo de um tampão ácido é uma solução de ácido acético e acetatc
8	acetato de sódio		0,2667	0,0000	Um exemplo de um tampão ácido é uma solução de ácido acético e acetatc
9	tampão básico	consistir	0,1850	0,9444	Um tampão básico consiste de uma base fraca e seu ácido conjugado forn
10	tampão básico	Ser	0,8279	0,6667	Um exemplo de um tampão básico é uma solução de. - AEIP.txt.xml
11	tampão ácido		0,6160	0,7273	Podemos entender a ação de um tampão ácido considerando o modelo do
12	tampão ácido		0,2485	0,9048	Ação do tampão ácido O ácido fraco, Ha, transfere prótons para os íons C
13	tampão básico		0,2608	0,9000	Ação do tampão básico A base fraca, B, recebe prótons dos íons H3O+ fo
14	ácido acético		0,9255	0,5769	AUTOTESTE 11.3A Calcule a razão das molaridades de íons acetato e ácid
15	tampão ácido		0,5964	0,7500	Especificamente, Para preparar um tampão ácido, selecionamos um ácido
16	tampão básico		0,4935	0,8000	Para preparar um tampão básico, selecionamos urna base fraca (B) com u
17	hidróxido de sódio	()	1,0000	0,5283	Exemplo 11.4 Calculando a variação de pH quando o ácido ou a base é ad
18	hidróxido de sódio		0,6000	0,3636	Solução O número de mols do ácido CH3COOH inicialmente na solução (do
19	ácido acético	:	0,0000	0,0000	fórmula) A quantidade de CH3COOH que reage é (fórmula) Portanto, a qu
20	hidróxido de sódio		0,1667	0,0000	Por exemplo, podemos estar interessados na concentração do ácido fórmi
21	hidróxido de sódio	consistir	1,0000	0,5556	Em o ponto estequiométrico da titulação de ácido fórmico, HCOOH (1), cor
22	tampão ácido	Ser	0,6785	0,6923	Note que uma maneira para se preparar um tampão ácido é neutralizar, cc
23	ácido fosfórico	conter	0,1561	0,9545	O ácido fosfórico, por exemplo, contém H3 PO4 e os íons H 2PO4-, HPO42
24	ácido fosfórico		0,5966	0,7308	A espécie primária do soluto é o ácido fosfórico, H3 PO4, e o equilíbrio de
25	produto de solubilidade	Chamado	0,9483	0,6111	Kps é também chamado de constante do produto de solubilidade ou simple
26	constante de solubilidade	Chamado	0,2667	0,0000	Kps é também chamado de constante do produto de solubilidade ou simple

Figura 5.4: Tabela de contextos extraídos através do concordanciador

Não é possível minimizar a janela contendo os resultados. No entanto, ao fechá-la, o botão *"Mostrar tabela"* é habilitado, permitindo que o usuário re-abra a última extração realizada.

O usuário pode alterar o conteúdo extraído clicando duas vezes sobre a célula que deseja alterar. Também é permitido copiar o conteúdo da tabela, basta selecionar as células desejadas e pressionar a tecla "CTRL" mais a tecla "C".

6. AVALIAÇÃO

Neste capítulo, é apresentada a avaliação deste trabalho e os resultados obtidos para cada *corpus* utilizado.

A avaliação do conjunto de heurísticas apresentadas, envolveu:

1. Avaliação manual, por especialistas em terminologia, dos contextos definitórios extraídos.

   Os contextos extraídos do *corpus* de Química foram avaliados por um terminólogo do projeto TextQuim. Esse profissional analisou os contextos, extraídos através de um concordanciador e das heurísticas apresentadas, sem saber quais as ferramentas utilizadas para extraí-los. Em seguida, marcou quais contextos eram bons, quais eram potenciais e quais eram ruins para constituir uma definição.

   Os contextos extraídos do *corpus* de Geologia Geral foram avaliados da mesma forma. No entanto, o profissional que avaliou os contextos de Geologia não foi o mesmo que avaliou os contextos de Química. Em Geologia, foi um mestrando em Terminologia da UFRGS.

2. Avaliação dos resultados obtidos individualmente pelas heurísticas.

   Os contextos foram classificados pelos especialistas como “Bom”, “Potencial” e “Ruim”. Aqueles contextos classificados como “Bom” claramente apresentavam a definição do termo. Os classificados como “Potencial” são os que possuíam elementos que ajudavam a compor uma definição. Por fim, os contextos assinalados como “Ruim”, não apresentavam elementos que ajudassem a compor uma definição.

6.1 *Corpora*

Os experimentos descritos abaixo são divididos em dois momentos. O primeiro é conduzido utilizando o *corpus* de Geologia Geral e o segundo é conduzido utilizando o *corpus* de Química Geral.

6.1.1 *Corpus* de Geologia Geral

O experimento aqui descrito, foi realizado utilizando parte do *corpus* de Geologia Geral: 9 dissertações e 9 teses.

Teses e dissertações tem como característica serem textos mais explicativos do que artigos, que são documentos científicos escritos de especialistas para especialistas e têm como característica serem menos explicativos sobre a terminologia utilizada [13].

A partir dos documentos de Geologia foram extraídos, através da ferramenta , os unigramas, os bigramas e os trigramas candidatos a possuírem definições. Desta lista, foram escolhidos os 10 termos mais frequentes de cada categoria (unigramas, bigramas e trigramas) e aqueles que estavam presentes nos glossários de referência, apresentados na seção 3.3.2, totalizando 30 termos. Esse corte foi necessário porque, se usássemos toda a lista, seria muito grande a quantidade de contextos extraídos o que inviabilizaria a avaliação do especialista, pois seria necessário uma a grande quantidade de tempo para a avaliação.

Foram extraídos todos contextos em que os 30 termos selecionados apareciam, gerando, assim, um total de 1.498 contextos.

Exemplos desses contextos são apresentados na Tabela D.1 no Apêndice.

Resultados da extração

Os contextos classificados como “Bom” e “Potencial” somam 152 contextos. Desses, 45 foram marcados como “Bom” e 107 como “Potencial”.

Analisando todos os contextos extraídos (1.498), verifica-se que somente 10,1% (152) desses são considerados úteis como contextos definitórios.

Através do uso das heurísticas, a quantidade de contextos extraídos reduziu de 1498 para 552 contextos. Desses, 37 foram classificados como “Bom” e 48 como “Potencial”, totalizando 85 contextos relevantes. Os valores de Precisão, Abrangência a F-Measure são apresentados na Tabela 6.1.

Tabela 6.1: Resultado da extração de contextos a partir do *corpus* de Geologia Geral

–	# Bom	# Potencial	B & P/Total	P	A	F
Sem Heurísticas	45	107	152/1498	10,1%	100%	18,3%
Com Heurísticas	37	48	85/552	15,4%	55,9%	24,1%

Analisando a Tabela 6.1, constata-se que através das heurísticas foram extraídos 552 contextos de um total de 1498. Desses 552 contextos, foram extraídos 85 contextos válidos (Bom / Potemcial), o que dá uma precisão de 15,4%.

Através das heurísticas, foram extraídos 85 dos 152 contextos válidos, o que resulta em 55,9% de abrangência.

A Tabela 6.2 a seguir, apresenta as heurísticas que coletaram os contextos classificados como “Potencial” e “Bom”.

Tabela 6.2: Resultado detalhado da extração de contextos a partir do *corpus* de Geologia Geral

Nro - Heurística	Bom	Potencial	Ruim	P	A	F
5 - Formar	18	3	41	3,8%	13,8%	6%
1 - Ser	6	12	62	3,3%	11,8%	5,2%
11 - Definir	4	6	30	1,8%	6,6%	2,8%
18 - Apresentar	3	7	90	1,8%	6,6%	2,8%
16 - Caracterizar	2	4	35	1,2%	4%	1,8%
7 - Constituir	4	0	35	0,7%	2,6%	1,1%
10 - Representar	0	4	55	0,7%	2,6%	1,1%
6 - Compor	0	4	29	0,7%	2,6%	1,1%
13 - Indicar	0	3	37	0,5%	2%	0,8%
12 - Consistir	0	2	6	0,4%	1,3%	0,6%
3 - ( )	0	2	16	0,4%	1,3%	0,6%
2 - :	0	1	5	0,1%	0,7%	0,2%
9 - Mostrar	0	0	15	0%	0%	0%
8 - Denotar	0	0	1	0%	0%	0%
Total	37	48	457	15,4%	55,9%	24,1%

Verificando a lista de contextos classificados como “Bom”, foi constatado que três contextos não foram recuperados devido à ausência dos bigramas que estavam sendo definidos na lista de termos. Nessa lista, estava presente o termo “Fácies”, porém não continha os termos “Fácies i”, “Fácies f” e “Fácies Sísmica”, os quais seriam recuperados através de heurísticas presentes nos padrões sintáticos.

Cabe ressaltar que 6 contextos assinalados como “Ruim” apresentavam o termo como constituinte da definição e não como *Definiendum*, fato que, embora prejudique a precisão, é válido para indicar a presença de novos conceitos do domínio.

Para amenizar este tipo de ocorrência, foi utilizada a fórmula de ranqueamento dos contextos definitórios, apresentada na seção 4.4.

Através dos valores obtidos por essa fórmula, foi feito a média para o conjunto de contextos assinalados como “Bom”, “Potencial” e “Ruim”. O resultatado foi que grande parte dos contextos com valor abaixo de 0,7000 foram considerados “Ruim” pelo avaliador.

Partindo dos resultados obtidos, foi aplicado um ponto de corte em 0,7000. Assim, somente os contextos que possuíam o termo com valor acima de 0,7000 foram mantidos. Deste modo, foram extraídos 264 contextos, sendo 37 avaliados como “Bom”, 36 avaliados como “Potencial” e 191 avaliados como “Ruim”. Esses números resultam em

uma precisão geral de 27,7% e em uma abrangência de 48%, conforme apresentado na Tabela 6.3

Tabela 6.3: Resultado da extração de contextos a partir do *corpus* de Geologia Geral com uso da fórmula de ranqueamento

–	# Bom	# Potencial	B & P/Total	P	A	F
Sem Heurísticas	45	107	152/1498	10,1%	100%	18,3%
Com Heurísticas	37	48	85/552	15,4%	55,9%	24,1%
Com Heurísticas & Ranqueamento	37	36	73/264	27,7%	48%	35,1%

Nota-se que, ao aplicar o ponto de corte, duas heurísticas não aparecem nos resultados. Uma delas é a heurística 2 (:), que não aparece porque o termo recuperado aparece no final de uma frase, sendo o valor dado pela fórmula de ranqueamento próximo de zero. A outra é a heurística 8 (Denotar), que não aparece por não ser suficientemente expressiva no *corpus*.

Outro ponto a ser observado é que, são removidos mais de 50% dos contextos extraídos e classificados como “Ruim”, aumentando a precisão para 27,7%, enquanto a abrangência diminui em menor proporção, para 48%.

Esses dados são apresentados pela Tabela 6.4.

Tabela 6.4: Resultado detalhado da extração de contextos do *corpus* de Geologia Geral, utilizando ponto de corte

Nro - Heurística	Bom	Potencial	Ruim	P	A	F
5 -Formar	18	2	22	7,6%	13,2%	9,7%
1 - Ser	6	12	45	6,8%	11,8%	8,6%
18 - Apresentar	3	6	26	3,4%	5,9%	4,3%
11 - Definir	4	4	14	3,0%	5,3%	3,8%
16 - Caracterizar	2	2	13	1,5%	2,6%	1,9%
7 - Constituir	4	0	16	1,5%	2,6%	1,9%
6 - Compor	0	3	13	1,1%	2%	1,4%
10 - Representar	0	2	14	0,8%	1,3%	1%
13 - Indicar	0	2	14	0,8%	1,3%	1%
12 - Consistir	0	2	2	0,8%	1,3%	1%
3 - ( )	0	1	8	0,4%	0,7%	0,5%
9 - Mostrar	0	0	4	0%	0%	0%
Total	37	36	191	27,7%	48%	35,1%

Comparando os resultados obtidos através do uso das heurísticas e da fórmula de ranqueamento com o resultado obtido a partir da extração de todos os contextos dos termos, nota-se que a precisão aumenta de 10,1% para 27,7%, ou seja, quase triplica. Da mesma forma, também houve o aumento da F-measure, de 18,3% para 35,1%.

Cabe ressaltar que a quantidade de contextos reduziu em 82% (de 1498 para 264) o que reduz a quantidade de 5 para 1 os contextos a serem analisados.

6.1.2 *Corpus* de Química Geral

Um outro momento deste trabalho é a extração de contextos definitórios a partir de um corpus de Química Geral. Nessa etapa do trabalho, contamos com 295 termos do banco de expressões e de termos técnicos disponibilizado pelo *site* do projeto TextQuim, os quais já possuíam uma definição, conforme apresentado na seção 3.4.1.

Para a extração de contextos potencialmente definitórios foram selecionados aleatoriamente 10 bigramas e 10 trigramas da lista de 295 termos previamente comentada. Isso foi necessário, visto que a quantidade de contextos extraídos para a lista completa de 295 termos era muito grande, impossibilitando que o terminólogo avaliasse todos os contextos no curto espaço de tempo que tínhamos disponível.

Utilizando os 10 bigramas e os 10 trigramas, foram localizados 246 contextos. Esses contextos foram analisados e classificados por um terminólogo como “Bom”, “Muito Bom”, “Ótimo”, “Mais ou menos” e “Ruim”. Essa classificação foi padronizada para “Bom” (Bom, Muito Bom e Ótimo), “Potencial” (Mais ou menos) e “Ruim”. Exemplos desses contextos são apresentados na Tabela B.1 no Apêndice.

Na análise, foram assinalados que 122 dos 246 contextos eram “Bons” ou “Potenciais”. Isso significa que, 49,6% dos contextos são úteis para constituir uma definição, o que mostra que o *corpus* de Química Geral é rico em definições, sendo esse, mais adequado para a extração de contextos.

Com o uso das heurísticas, foram recuperados 102 contextos. Desses, 58 foram avaliados como válidos (Bom ou Potencial), o que gera a Precisão de 56,9%, a Abrangência de 47,5% e a F-measure de 51,7%, conforme pode ser observado na Tabela 6.5 a seguir.

Tabela 6.5: Resultado da extração de contextos a partir do *corpus* de Química Geral

–	# Contextos Extraídos	# Contextos Válidos	P	A	F
Sem Heurísticas	246	122	49,6%	100%	66,3%
Com Heurísticas	102	58	56,9%	47,5%	51,7%

Inicialmente, utilizamos a fórmula de ranqueamento com ponto de corte de 0,7000, como no experimento anterior. Porém, a queda de abrangência e F-measure foi muito acentuada. Portanto, foram averiguados diferentes pontos de corte, variando de 0,7000 até 0,4000.

A Tabela 6.6 apresenta os resultados obtidos para os diferentes pontos de corte, utilizando os contextos válidos (Bom e Potencial).

Tabela 6.6: Resultado da extração de contextos a partir do *corpus* de Química Geral com o uso da fórmula de ranqueamento

–	# Contextos Extraídos	# Contextos Válidos	P	A	F
Sem Corte	102	58	56,9%	47,5%	51,7%
Corte 0,7	37	27	72,9%	22,1%	33,9%
Corte 0,6	55	37	67,2%	30,3%	41,7%
Corte 0,5	69	43	62,3%	35,2%	44,9%
Corte 0,4	73	45	61,6%	36,9%	46,2%

Nota-se que quanto maior o ponto de corte, maior a precisão. No entanto, como o *corpus* de Química Geral é rico em contextos definitórios, com o aumento do ponto de corte são removidos os contextos bons, diminuindo a abrangência e a F-measure.

Os resultados detalhados por heurísticas e sem corte são apresentados na Tabela 6.7, e os resultados detalhados com corte são apresentados na Tabela 6.8.

Comparando os resultados obtidos sem corte e com corte, observa-se que, após aplicar o corte, duas heurísticas não aparecem, a 2 (:) e a 20 (Isto é). Outro ponto observado é que o uso da fórmula de ranqueamento no *corpus* de Química Geral gera um ganho satisfatório de precisão, porém, também diminui a abrangência. Isso demonstra que a fórmula de ranqueamento contribui bastante quando aplicada sobre documentos pobres em definições, visto que reduz bastante a quantidade de contextos a serem analisados pelo especialista e retorna resultados mais precisos, como pode ser analisado na Tabela 6.6.

Tabela 6.7: Resultado detalhado da extração de contextos do *corpus* de Química Geral, sem ponto de corte

Nro - Heurística	Bom	Potencial	Ruim	P	A	F
1 - Ser	12	5	7	16,6%	13,9%	15,1%
5 - Formar	4	6	7	9,8%	8,2%	8,9%
4 - Chamar	6	1	1	6,8%	5,7%	6,2%
12 - Consistir	5	0	2	4,9%	4,1%	4,4%
9 - Mostrar	1	2	4	2,9%	2,6%	2,7%
10 - Representar	1	2	5	2,9%	2,6%	2,7%
7 - Constituir	2	0	0	2%	1,6%	1,8%
16 - Caracterizar	2	0	3	2%	1,6%	1,8%
3 - ( )	1	1	8	2%	1,6%	1,8%
14 - Significar	1	1	0	2%	1,6%	1,8%
19 - Conhecido como	1	0	0	1%	0,8%	0,9%
11 - Definir	1	0	1	1%	0,8%	0,9%
20 - Isto é	1	0	0	1%	0,8%	0,9%
2 - :	0	1	1	1%	0,8%	0,9%
18 - Apresentar	0	1	2	1%	0,8%	0,9%
13 - Indicar	0	0	3	0%	0%	0%
Total	38	20	44	56,9%	47,5%	51,7%

Tabela 6.8: Resultado detalhado da extração de contextos do *corpus* de Química Geral, com ponto de corte

Nro - Heurística	Bom	Potencial	Ruim	P	A	F
1 - Ser	11	5	7	23,2%	13,1%	16,7%
12 - Consistir	5	0	1	7,3%	4,1%	5,2%
5 - Formar	1	3	5	5,8%	3,4%	4,2%
9 - Mostrar	1	2	2	4,4%	2,5%	3,2%
10 - Representar	1	2	2	4,4%	2,5%	3,2%
4 - Chamar	2	0	0	2,9%	1,6%	2,1%
16 - Caracterizar	2	0	1	2,9%	1,6%	2,1%
3 - ( )	1	1	3	2,9%	1,6%	2,1%
14 - Significar	1	1	0	2,9%	1,6%	2,1%
19 - Conhecido como	1	0	0	1,4%	0,8%	1%
7 - Constituir	1	0	0	1,4%	0,8%	1%
18 - Apresentar	0	1	2	1,4%	0,8%	1%
11 - Definir	0	1	1	1,4%	0,8%	1%
13 - Indicar	0	0	2	0%	0%	0%
Total	27	16	26	62,3%	35,2%	44,9%

6.2 Análise de erros

Após o desenvolvimento dos dois experimentos, é possível averiguar que, embora tenham sido empregadas 21 heurísticas, ainda existe a possibilidade se serem aplicadas outras a fim de recuperar contextos com outras características, aumentando assim, a sua abrangência.

Uma opção seria utilizar outros verbos indicativos, como no caso do *corpus* de Química Geral, em que o uso do verbo “Encontrar” e “Preparar” é frequente em contextos definitórios. Entretanto, essa situação não é válida para o *corpus* de Geologia Geral. Esta opção teria que ser aperfeiçoada com experimentos que utilizassem *corpora* de diferentes domínios para, então, determinar as opções que melhor se adequam ao conjunto.

Foi possível notar que parte dos contextos recuperados foram avaliados como “Ruim”, porque não se verificou se o termo que está sendo observado é o sintagma nominal da frase.

Por exemplo, o contexto: “Este tipo de ligação é chamada ligação covalente coordenada.”

O contexto acima foi recuperado pela heurística 4 (Chamar) quando se buscou por contextos para o termo “ligação covalente”. Entretanto, contexto refere-se ao termo “ligação covalente coordenada”.

Outra questão a ser tratada é a nominalização de verbos. Ao anotar os documentos, o *parser* anota alguns verbos como substantivos, fazendo com que algumas heurísticas não encontrem esse contexto.

Por exemplo, o contexto: “A ligação formada numa reação ácido-base de Lewis é uma ligação covalente coordenada.”

No exemplo acima o *parser* anotou o termo “ligação” como substantivo e “formada” como adjetivo.

Por fim, é possível apontar a questão de contextos que se remetem a frase anterior, como contextos iniciados por pronome demonstrativo. Essa questão é tratada no protótipo, porém não foi possível avaliá-la devido ao curto prazo de tempo disponível e à necessidade de desenvolver um experimento específico para este trabalho.

7. CONSIDERAÇÕES FINAIS

Neste capítulo, são apresentadas as contribuições deste trabalho para área, as conclusões finais e os possíveis trabalhos futuros.

7.1 Contribuições do Trabalho

Para o presente trabalho, é possível apontar as seguintes contribuições:

1. Um conjunto de heurísticas para a extração de contextos potencialmente definitórios;
2. Uma fórmula de ranqueamento, para pontuar a posição do termo na frase;
3. Um protótipo contendo um concordanciador, as heurísticas, a fórmula de ranqueamento e a medida de centralização do termo;
4. Publicação do artigo “Geração automática de glossários de termos específicos de um corpus de geologia” no 3º SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL - ONTOBRAS (2010).
5. Publicação do artigo “Extração de Contextos Definitórios a partir de Textos em Língua Portuguesa” no 8 º Simpósio Brasileiro em Informação e Tecnologia da Linguagem Humana - STIL 2011 [22].

7.2 Conclusão

De acordo com os resultados obtidos, pode-se concluir que os métodos empregados neste trabalho fornecem uma grande redução de material a ser analisado. Além disso, fornece um aumento de precisão na extração de contextos potencialmente definitórios, em comparação com a utilização de um concordanciador. A redução de material para ser analisado foi observada em ambos os *corpus*. Entretanto, a medida de ranqueamento mostrou-se mais relevante no *corpus* de Geologia Geral, composto por teses e dissertações. Estima-se que esse *corpus* contenha menos contextos definitórios do que o de Química Geral.

O conjunto de heurísticas desenvolvido, apresenta resultados satisfatórios, principalmente quando as heurísticas são aplicadas sobre um *corpus* pobre de contextos definitórios. Esse ponto pode ser observado através do experimento realizado com o *corpus* de Geologia Geral. Desse *corpus*, foi obtido 1498 contextos, sendo 152 úteis. Com o uso das heurísticas e da fórmula de ranqueamento, a quantidade de

contextos a serem analisados passa para 264, sendo 73 úteis. Esse resultado representa um aumento de precisão de 10,1% para 27,7%. O resultado da F-measure também aumentou de 18,3% para 35,1%, quando comparado a um concordanciador.

Utilizando o *corpus* de Química Geral, um *corpus* rico em contextos definitórios, foi obtido um aumento de precisão de 49,6% para 56,9%. Nesse *corpus*, a F-measure cai, visto que a quantidade de contextos recuperados passam de 246 (sem heurísticas) para 102 (com heurística). E, visto que esse *corpus* é rico em definições, a abrangência também diminui. Através da fórmula de ranqueamento foi possível obter valores de até 72,9% de precisão, porém a abrangência também reduziu, prejudicando a média harmônica (F-measure). Esse aumento de precisão torna-se interessante para o terminólogo ou para o linguísta, que procura identificar os contextos definitórios mais relevantes de um determinado termo. Caso seja necessário obter maior número de definições, basta pesquisar por valores inferiores de ponto de corte.

Del Gaudio e Branco apresentam os resultados que obteram ao extrair contextos definitórios a partir de três *corpus* em diferentes áreas. Em seu trabalho, eles obteram 32% de F-measure para o *corpus* de *Information Society*, 51% para o *corpus* de Information Technology e 24% para o *corpus* de *e-Learning*. Os resultados demonstram a influência do domínio, mesmo esses sendo compostos principalmente por tutoriais, que possuem a característica de fornecer maiores explicações sobre os conceitos.

O trabalho apresentado por Przepiórkowski *et al.*, utilizando três *corpus* no domínio de *e-Learnig*, dividido em língua búlgara, tcheca e polonesa obteve como resultado de F-measure respectivamente 11,1%, 33,9% e 28,4%. Os autores argumentam que o resultado poderia ser melhorado caso fossem extraídos contextos definitórios anafóricos, característica fortemente presente no *corpus* em língua búlgara.

Essa observação não é uma característica predominante no *corpora* utilizado neste trabalho, porém é um recurso que poderia agregar bons resultados em um *corpus* com características diferentes.

Em uma análise geral, é possível concluir que o presente trabalho obteve resultados satisfatórios em comparação com trabalhos que utilizam métodos semelhantes, conforme pode ser visto na Tabela 7.1.

Tabela 7.1: Resultados obtidos por diferentes autores

Autor	*Corpus*	F-Measure
Wendt	Química Geral	51,7%
Del Gaudio e Branco	IS	51%
Przepiórkowski *et al.*	*e-Learning*	33,9%

8. TRABALHOS FUTUROS

Concluindo este trabalho, é possível apontar os seguintes trabalhos futuros:

8.1 Avaliação do protótipo ExContext

A avaliação do uso do protótipo no trabalho de terminólogos e linguistas demonstraria quão útil são as heurísticas no auxílio da extração de contexto definitórios e na redução de tempo para desempenhar essa tarefa.

8.2 Avaliação dos contextos definitórios anafóricos

A avaliação da extração de contextos anafóricos, implementado no protótipo, poderia demonstrar um ganho superior ao apresentado, visto que os contextos anafóricos também ocorrem no *corpora* utilizado.

8.3 Extração de Sintagmas Nominais a partir do protótipo ExContext

A extração do sintagma nominal a partir do contexto identificado por uma heurística reduziria a quantidade de contextos nos quais o termo identificado não é o termo que está sendo definido.

8.4 Avaliação da fórmula de ranqueamento

O *parser* PALAVRAS faz a anotação dos Sintagmas Nominais do texto e, através desta anotação, poderíamos verificar a eficiência da fórmula de ranqueamento e fazer os ajustes necessários para a obtenção de melhores resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ALLEN, J., WIDDOWSON, H., AND ALLEN, J. *English in physical science*. Oxford University Press, 1974.

[2] ATKINS, P., JONES, L., AND CARACELLI, I. *Princípios de Química: questionando a vida moderna e o meio ambiente*. Bookman, 2001.

[3] BOTELHO, J. A ordem dos termos em português e a topicalização. *Rio de Janeiro: CiFEFiL*, 47 (2010), 43–61.

[4] CIMIANO, P. *Ontology learning and population from text: algorithms, evaluation and applications*. Springer Verlag, 2006.

[5] DE LUCCA, J. L. Identificação de padrões recorrentes no discurso técnico e científico para a extração automática de candidatos a contextos definitórios em língua portuguesa. *Revista Intercâmbio XV* (2006), 9.

[6] ECKHARD, B. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[7] FINATTO, M. O papel da definição de termos técnico-científicos. *Revista da ABRALIN 1*, 1 (2002), 73–97.

[8] FLOWERDEW, J. Salience in the performance of one speech act: the case of definitions. *Discourse processes 15*, 2 (1992), 165–181.

[9] GAUDIO, R., AND BRANCO, A. Supporting e-learning with automatic glossary extraction: Experiments with Portuguese. In *RANLP Workshop: Natural Language Processing and Knowledge Representation for eLearning Environments* (2007), p. 7.

[10] GRUBER, T. A translation approach to portable ontology specifications. *Knowledge acquisition 5* (1993), 199–220.

[11] IFTENE, A., TRANDABĂ, D., AND PISTOL, I. Grammar-based automatic extraction of definitions and applications for Romanian. In *Proceedings of RANLP workshop. Natural Language Processing and Knowledge Representation for eLearning environments* (2007), Citeseer, pp. 978–954.

[12] LOPES, L., OLIVEIRA, L., AND VIEIRA, R. Portuguese term extraction methods: Comparing linguistic and statistical approaches. *International Conference on Computational Processing of Portuguese Language, PROPOR* (2010), 6.

[13] PEARSON, J. Comment accéder aux éléments définitoires dans les textes spécialisés. *Terminologie et intelligence artificielle (TIA'1999)* (1999), 21–38.

[14] PICHT, H. *Korpora als Ausgangspunkt für die Extraktion von terminologischen Daten*, vol. 8. Synaps, 2001.

[15] PRZEPIÓRKOWSKI, A., DEGÓRSKI, Ł., WOJTOWICZ, B., SPOUSTA, M., KUBOŇ, V., SIMOV, K., OSENOVA, P., AND LEMNITZER, L. Towards the automatic extraction of definitions in slavic. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies* (2007), Association for Computational Linguistics, pp. 43–50.

[16] RUSSEL, J. Química Geral, vol. 2. *São Paulo: Makron* (1994).

[17] SCLANO, F., AND VELARDI, P. Termextractor: a web application to learn the shared terminology of emergent web communities. *Enterprise Interoperability II* (2007), 287–290.

[18] SILVEIRA, F. P. Entrelinhas - uma ferramenta para processamento e análise de corpus. *Dissertação de Mestrado, PUCRS* (2008).

[19] SOUSA, L., LEITE, J., DE SOUZA, C., DE CASTRO, A., AND AMARO, V. Uma abordagem baseada em ontologias para melhorar a cooperação em estudos ambientais em áreas produtoras de petróleo e gás-natural. *3° Congresso Brasileiro de P&D em Petróleo e Gás* (2005), 6.

[20] SWALES, J. *Writing Scientific English*. Nelson, 1985.

[21] WENDT, I. S., LOPES, L., MARTINS, D., VIEIRA, R., AND STRUBE DE LIMA, V. L. Geração automática de glossários de termos específicos de um corpus de geologia. *3º SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL - ONTOBRAS* (2010), 10.

[22] WENDT, I. S., AND VIEIRA, R. Extração de contextos definitórios a partir de textos em língua portuguesa. *The 8th Brazilian Symposium in Information and Human Language Technology (STIL)* (2011), 9.

Apêndice A. LISTA DE TERMOS DE GEOLOGIA GERAL

1. Unigramas
    - bacia
    - litologia
    - fácies
    - granito
    - poços
    - ambientes
    - margens
    - reservatório
    - sedimentação
    - barreiras
    - granito cinza
2. Bigramas
    - sistema deposicional
    - Formação Sergi
    - plataforma continental
    - Formação Maricá
    - margem continental
    - dados sísmicos
    - planície costeira
    - seções sísmicas
    - ambiente deposicional
3. Trigramas
    - estratigrafia de seqüências
    - espaço de acomodação
    - superfície de inundação
    - planície de inundação
    - plataforma continental interna
    - granitos Nazaré Paulista
    - bordas de grão
    - feições de terraceamento

geoquímica de superfície

quebra de biotita

Apêndice B. CONTEXTOS DEFINITÓRIOS DE GEOLOGIA GERAL

Tabela B.1: Exemplos de contextos do *corpus* de Geologia Geral avaliados pelo terminólogo

Avaliação	Termo	Heurística	Ranking	Contexto
Bom	margem continental	1- Ser	0,9524	A margem continental é uma processo de abertura do Oceano Atlântico Sul e região propícia para a acumulação de recursos Equatorial.
Bom	estratigrafia de seqüências	11- Definir	0,9545	A estratigrafia de seqüências pode ser definida como o estudo dos estratos sedimentares geneticamente relacionados, situados entre duas superfícies cronoestratigraficamente relevantes.
Bom	espaço de acomodação	11- Definir	0,9565	O espaço de acomodação, de acordo com, pode ser definido como o espaço disponível para a acumulação ou deposição de sedimentos.
Bom	reservatório	1- Ser	0,9231	O reservatório é um domo alongado, parcialmente falhado por o alto estrutural.
Potencial	ambiente deposicional	10- Representar	0,9474	O ambiente deposicional da Formação Ponta Grossa é representado por uma plataforma marinha em rampa de águas pouco profundas.
Potencial	barreiras	1- Ser	0,8621	Os principais fatores que definem a evolução das barreiras são as respostas dos ambientes costeiros à velocidade de variação do nível relativo do mar, a relação entre o volume de sedimentos e a morfologia do substrato, que define o espaço de acomodação, e a relação entre a energia de ondas e a amplitude das marés.
Potencial	ambiente deposicional	10- Representar	0,9474	O ambiente deposicional da Formação Ponta Grossa é representado por uma plataforma marinha em rampa de águas pouco profundas.
Potencial	espaço de acomodação	13- Indicar	0,7879	Assim, a retrogradação indica que o espaço de acomodação criado é relativamente maior que o aporte sedimentar disponível para preenchimento, fazendo com que a linha de costa do lago seja transgressiva.
Potencial	feições de terraceamento	5- Formar	0,8214	Em o fotomosaico são observadas feições de terraceamento formadas a partir de a ação de ondas, esculpindo duas escarpas erosivas na margem lagunar ativa (C e D).
Ruim	planície costeira	5- Formar	0,75	O nível do mar subiu rapidamente e avançou sobre a ampla planície costeira formada na fase regressiva anterior, possibilitando a formação de um novo sistema de barreira, transgressiva, que evoluiu para progradante durante a fase posterior regressiva (Sistema Laguna– Barreira IV).
Ruim	estratigrafia de seqüências	11- Definir	0,7	Assim, podemos concluir que é possível a utilização dos conceitos da estratigrafia de seqüências de alta resolução na interpretação de feições no Holoceno, definindo, dessa forma, ciclos de alta freqüência dentro de o registro geológico recente.
Ruim	fácies	7- Constituir	0,7391	É bastante evidente, que as fácies mais porosas, que constituem os depósitos turbidíticos arenosos, estão associados as baixas amplitudes sísmicas.
Ruim	litologia	16- Caracterizar	0,8276	Um corpo de rocha caracterizado por uma combinação particular de litologia, estruturas físicas e biológicas, a qual diferência este corpo de rocha dos outros corpos acima, abaixo e lateralmente adjacentes», a qual representa diretamente o registro de condições de sedimentação singulares em relação a as fácies adjacentes (energia, tipo de fluxo, fluido).

Apêndice C. LISTA DE TERMOS DE QUÍMICA GERAL

1. Bigramas
    - ácido acético
    - ácido fosfórico
    - eixo internuclear
    - elétrons emparelhados
    - forma pura
    - ligação covalente
    - orbitais moleculares
    - reação negativa
    - tampão ácido
    - tampão básico
2. Trigramas
    - acetato de sódio
    - constante de solubilidade
    - diferença de eletronegatividade
    - energia potencial total
    - entalpia de ligação
    - híbrido de ressonância
    - hidróxido de sódio
    - hidróxido de cálcio
    - pentacloreto de fósforo
    - produto de solubilidade

Apêndice D. CONTEXTOS DEFINITÓRIOS DE QUÍMICA GERAL

Tabela D.1: Exemplos de contextos do *corpus* de Química Geral avaliados pelo terminólogo

Avaliação	Termo	Heurística	Ranking	Contexto
Bom	ácido acético	1- Ser	0,9524	O ácido acético é o ácido presente no vinagre e é formado quando o etanol do vinho é oxidado.
Bom	ligação covalente	12- Consistir	0,9545	De acordo com a teoria da valência (VB), a ligação covalente consiste num par de elétrons compartilhados em dois átomos ligados.
Bom	híbrido de ressonância	19- Conhecido como	0,7442	A estrutura da molécula de ozônio é conhecida como um híbrido de ressonância das formas I e II, ou seja, é uma forma intermediária entre I e II e não pode ser representada satisfatoriamente por uma simples estrutura de Lewis.
Bom	ligação covalente	10- Representar	0,7	Cada par compartilhado conta como uma ligação covalente e é representado por uma linha entre os dois átomos.
Bom	orbitais moleculares	4- Chamar	0,5	Por a teoria do orbital molecular, os elétrons ocupam orbitais chamados orbitais moleculares que se espalham por toda a molécula.
Potencial	tampão ácido	1- Ser	0,6923	Note que uma maneira para se preparar um tampão ácido é neutralizar, com uma base forte, a metade da quantidade de ácido fraco presente.
Potencial	ligação covalente	5- Formar	0,7073	As características de uma ligação covalente formada por dois átomos são devidas principalmente às propriedades destes átomos e variam somente um pouco com as identidades de outros átomos presentes na molécula.
Potencial	diferença de eletronegatividade	14- Significar	0,7632	Entretanto, uma boa regra diz que, se há uma diferença de eletronegatividade de cerca de 2 unidades, isto significa que um grande caráter iônico está presente numa ligação, e é melhor considerar a ligação como iônica.
Potencial	orbitais moleculares	10- Representar	0,8592	A energia relativa entre o orbital atômico original e os orbitais moleculares ligante e antiligante são representados na forma de diagramas de níveis de energia do orbital molecular, como o que está apresentado na Figura 3.31.
Ruim	ligação covalente	5- Formar	0,4048	As moléculas de Al2 Cl6 existem porque um átomo de Cl de uma molécula de AlCl3 usa um de seus pares isolados para formar uma ligação covalente coordenada com o átomo de Al de uma molécula de AlCl, vizinha (42).
Ruim	orbitais moleculares	1- Ser	0,5	Em o limite inferior das bandas, os orbitais moleculares são totalmente ligantes.
Ruim	hidróxido de sódio	12- Consistir	0,5556	Em o ponto estequiométrico da titulação de ácido fórmico, HCOOH (1), com hidróxido de sódio, a solução consiste de formiato de sódio, NaHCO2, e água.
Ruim	reação negativa	3- ( )	0,6923	Para uma reação exotérmica com uma entropia de reação negativa (logo AHr$^{\circ}$ e ASr$^{\circ}$ são negativos), a entalpia contribui num termo negativo para AGr$^{\circ}$.
Ruim	orbitais moleculares	10- Representar	0,72	As contribuições relativas dos orbitais atômicos aos orbitais moleculares estão representadas por o tamanho das esferas e por a posição horizontal das caixas.