GERAÇÃO SEMIAUTOMÁTICA DE UMA ONTOLOGIA GERAL DE LÍNGUA COMUM

1. Introdução

Sistemas computacionais como os de perguntas e respostas (P&R) têm a tarefa de responder automaticamente uma pergunta em linguagem natural, procurando por informações em fontes de dados, tais como um banco de dados ou documentos não estruturados em linguagem natural (por exemplo, jornais). Exemplos desses sistemas são Kaisser (2005), Lin (2005), Zheng (2002), Sarmento et al. (2008) e Amaral et al. (2006).

Um fator importante que tem grande impacto na performance de tais sistemas é a disponibilidade de ontologias dedicadas contendo os conceitos relevantes para a busca das informações. De modo resumido, ontologias são estruturas de conceitos com relações explícitas entre si, geralmente de ordem semântica. Um exemplo de sistema de P&R via telefone para o português é o projeto Comunica (XXXXX), que busca responder perguntas sobre receitas de municípios. Nesse projeto, tanto a pergunta do usuário quanto a resposta do sistema são em linguagem natural, visando a uma maior inclusão digital. Para realizar tal tarefa, são necessário quatro etapas básicas: reconhecer a voz e convertê-la em texto, identificar os conceitos presentes no texto, buscar a resposta e sintetizar a voz para repassar a resposta ao usuário. Nosso escopo é a identificação dos conceitos, que é feita por meio de duas ontologias, uma de língua comum e uma do domínio da aplicação (que, no caso, inclui informações sobre as receitas de municípios). Neste estudo, temos por objetivo mostrar a metodologia utilizada para gerar uma parte inicial da ontologia de língua comum: a parte dos sinônimos.

Essa parte da ontologia foi extraída de modo semiautomático a partir do PAPEL (Palavras Associadas Porto Editora Linguateca) (Oliveira et al., 2008) e convertida para o formato OWL. Os resultados da conversão foram validados por um linguista. O PAPEL apresenta um conjunto de relações entre palavras extraídas automaticamente das definições presentes em um tesauro eletrônico. Ele contém 199.672 entradas, distribuídas em 8 tipos de relações. Dentre elas, foram selecionadas as de sinonímia como o primeiro passo para a construção da ontologia de língua comum.

Este artigo é estruturado da seguinte maneira: na Seção 2, apresentamos brevemente o PAPEL; a metodologia para extração e validação é descrita na Seção 3; por fim, a Seção 4 apresenta nossas considerações finais sobre o trabalho.

2. O PAPEL

Entre diversos recursos lexicais existentes, a Wordnet (MacWhinney, 1995) destaca-se por apresentar relações entre palavras (como sinonímia e hiperonímia) organizadas manualmente. Contudo, o desenvolvimento desse recurso é extremamente trabalhoso e demorado. Nesse sentido, a sua extração automática apresenta uma opção atraente, apesar de necessitar de um processo de validação, pois o processo automático insere muito ruído. Um recursos amplamente utilizado em trabalhos como Oliveira, Costa e Santos (2012) e Oliveira, Pérez e Gomes (2012) é o PAPEL (Oliveira et al., 2008), o qual ainda não está completamente validado.

O PAPEL foi criado com o objetivo de prover uma ontologia geral da linguagem de grande abrangência, tendo sido construído através da extração semiautomática baseada em padrões de expressões que ocorrem nas definições do Dicionário da Língua Portuguesa (Porto Editora, 2005). Dessa forma, foram identificadas relações

composicionais, hierárquicas e de sinonímia. Exemplos dessas relações (retirados do PAPEL) seriam:

repartir	SINÔNIMO DE	partilhar
vasqueiro	PROPRIEDADE DE ALGO QUE CAUSA	vasca
vazar	AÇÃO QUE CAUSA	vazão
cabo	PARTE DE	vassoura
navio	HIPERÔNIMO DE	veleiro

Devido à sua abrangência (com 199.672 entradas), foi realizada uma avaliação por amostragem dos resultados da extração semiautomática (Oliveira et al., 2009), sendo que 50% das relações de sinonímia apresentam erros em potencial; por exemplo: deliberadamente SINÔNIMO DE peito. Para contornar esses problemas, apresentamos uma metodologia de extração e validação mais confiável inserida no projeto Comunica (XXXXX).

3. Organização e Validação

Dividimos esta seção em dois passos. Primeiro, apresentamos o trabalho computacional realizado para as relações de sinonímia. Em seguida, apresentamos os passos utilizados para fazer a validação manual realizada por um linguista.

3.1. Organização dos Sinônimos

A base de sinônimos do PAPEL possui relações expressas da seguinte maneira: palavra 1 SINONIMO <classe> DE palavra 2, onde em <classe> ocorrem as seguintes tags, que indicam classes gramaticais: N = substantivo; V = verbo; ADJ = adjetivo; e ADV = advérbio. Ao todo, são 80.429 relações de sinonímia.

Dado o contexto deste trabalho, foram extraídas as relações de sinonímia apenas entre substantivos (contudo, a abordagem aqui apresentada poderia ser utilizada para as sinonímias das outras classes). Assim, das 80.429 relações, foram utilizadas as 36.504 que se referiam à sinonímia entre substantivos. A extração de sinônimos foi realizada através das seguintes etapas:

1. Criação de uma lista (inicialmente vazia) de conjuntos (inicialmente vazios); cada conjunto armazena palavras que são sinônimas entre si.
2. Para cada uma das entradas, realiza-se a identificação das duas palavras sinônimas presentes.
3. Verificação de se uma dessas palavras já se encontra em algum conjunto de sinônimos existente:
    (a) Se já existe, a outra palavra é inserida nesse conjunto.
    (b) Se as palavras não estavam em nenhum conjunto existente, um novo conjunto é criado com ambas as palavras.

Dada a ampla abrangência do PAPEL, a polissemia das palavras e a transitividade da relação de sinonímia (se A é sinônimo de B e B é sinônimo de C, então A é sinônimo de C), ao final do processo, quase todas as palavras foram reconhecidas como sinônimas entre si. Porém, para criar uma ontologia de alta precisão e cobertura, foi incluída uma validação manual no passo 3:

3. Verificação de se uma dessas palavras já se encontra em algum conjunto de sinônimos existente:

(a) Se já existe, passar a ambiguidade para avaliação manual, permitindo a visualização dos dois conjuntos em que as palavras seriam inserida (nesse passo, é possível que o avaliador humano edite os conjuntos manualmente, separando as palavras de modo adequado).
(b) Se as palavras não estavam em nenhum conjunto existente, um novo conjunto é criado com ambas as palavras.

Ao final do processo, obtém-se uma série de grupos de sinônimos validados por um humano. A seguir mostramos como se deu esse processo de validação.

3.2. Validação das Relações de Sinonímia

A validação manual foi realizada através de uma interface criada para a visualização dos dois grupos aos quais os sinônimos poderiam ser alocados (Figura 1).

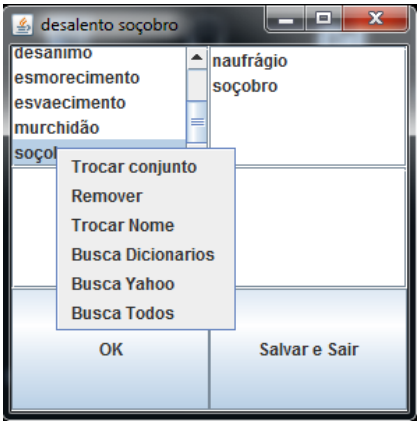

Figura 1. Interface de Validação de Sinônimos

Como se pode ver na Figura 1, a interface apresenta, no cabeçalho, as duas palavras que estão sendo avaliadas como sinônimas (no caso, *desalento* e *soçobro*); nos quatro quadros (dois inicialmente em branco), é possível mover as palavras, criando novos conjuntos, removê-las ou editá-las.

A interface também disponibiliza três opções de busca. Essas opções foram incluídas para que o trabalho de validação não fosse algo totalmente subjetivo, de modo que elas permitem ao avaliador acessar ferramentas de apoio, como dicionários de língua portuguesa e contextos reais de ocorrência dos pares de sinônimos. Assim, a interface de validação permite acesso rápido ao buscador do Yahoo! (www.yahoo.com) e também à WordNet.Br (Dias da Silva et al., 2008), além dos dicionários Dicionário Moderno da Língua Portuguesa, disponível *on-line* (http://michaelis.uol.com.br/moderno/portugues/index.php) e Dicionário da Língua Portuguesa da Porto Editora, que também se encontra disponível *on-line* (http://www.infopedia.pt/). O acesso a esses recursos *on-line* é feito por intermédio da própria interface, que pode ser vista na Figura 2 a seguir.

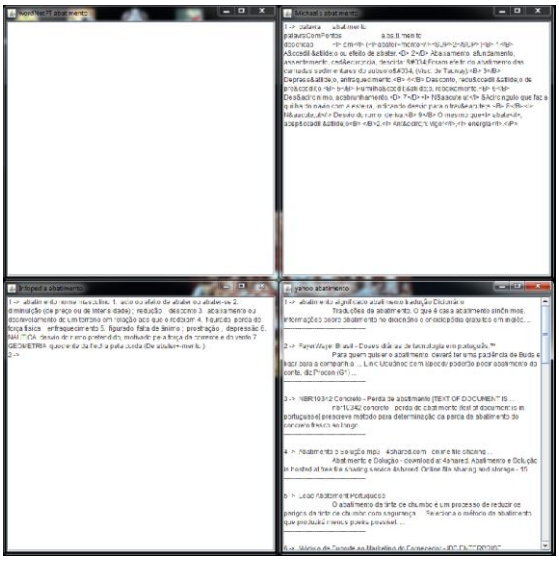

Figura 2. Interface de acesso aos recursos linguísticos *on-line*

Nessa interface, vemos quatro janelas distintas: em cima, à esquerda, temos a WordNet.Br; em cima, à direita, temos as informações do Michaelis *on*-line; embaixo, à esquerda, aprecem as informações do Dicionário da Porto Editora; por fim, embaixo, à direita, são mostradas as informações do buscador do Yahoo!. Além dos recursos acessados diretamente via interface de validação, também foi utilizado o dicionário Houaiss Eletrônico versão 1.0 (Houaiss, 2001), tendo em vista que suas informações são bastante ricas no que diz respeito aos sinônimos.

Durante o processo de validação, geralmente os dicionários eram os primeiros recursos a serem acessados, porém, caso as definições dos dicionários não aclarassem o problema, utilizavam-se o buscador *on-line* do Yahoo! para se observarem também os contextos de ocorrência.

Dada a magnitude do recurso gerado, considerou-se, nessa primeira fase da validação, que cada palavra poderia ocorrer em apenas um conjunto de sinônimos, de modo que, para palavras polissêmicas, a validação foi realizada com base no significado mais frequente verificado nos contextos de ocorrência. Para ficar mais fácil de compreender o processo empregado, tomemos como exemplo a avaliação da relação de sinonímia proposta entre *abatimento*, *diminuição* e *desânimo*. Apesar de *desânimo* e *diminuição* não parecerem, à primeira vista, sinônimas, a palavra *abatimento* pode ser considerada sinônima de ambas, entre outras. A consulta aos dicionários retornou informações sobre *abatimento de animais*, *abatimento de preços* (*diminuição*) e *abatimento emocional* (*desânimo*). No buscador do Yahoo!, entre as primeiras 20 ocorrências de "abatimento", havia 12 ocorrências *abatimento de preços*, 4 de *abatimento emocional* e 1 de *abatimento de animais*; as demais eram irrelevantes (definições de dicionários *on-line* etc.). Assim, a palavra *abatimento* foi incluída no conjunto da palavra *diminuição* e excluída do conjunto onde estava a palavra *desânimo*.

A adoção da metodologia apresentada permitiu apoiar o trabalho lexicográfico através de uma interface para avaliação que disponibiliza recursos eletrônicos de forma integrada. Além disso, se reduz um pouco a subjetividade envolvida em todo o processo de decidir que palavras são sinônimas entre si e quais não devem ser. Isso se torna importante para a replicabilidade do processo de validação, tendo em vista a natureza inerentemente subjetiva e dependente do vocabulário do avaliador no momento de decidir quais palavras possuem uma semelhança de significado.

Ao final do processo de validação, a ontologia resultante apresentou 20.096 relações de sinonímia, partindo das 36.504 relações iniciais. O recurso apresenta alta abrangência e pode ser utilizado em uma grande variedade de sistemas de tecnologia de linguagem.

4. Considerações Finais

Este artigo propôs uma metodologia para validação de relações de sinonímia obtidas automaticamente por meio de padrões observados em um dicionário. Foi apresentado em detalhes o procedimento de avaliação de sinonímia realizado manualmente e com decisões auxiliadas por outros recursos linguísticos. Apesar da magnitude do recurso original, essa metodologia possibilitou uma validação mais ampla do recurso, em vez da avaliação por amostragem proposta em (Oliveira et al. 2009).

Além da etapa apresentada neste trabalho, há ainda a validação das relações de hiperonímia, também presente no PAPEL, que é a segunda parte da ontologia. Outro trabalho importante que precisa ser desenvolvido é observar os conjuntos de sinônimos e adicionar informações de polissemia. Apesar desses trabalhos importantes que precisam ser levados adiante, a parte já validada do recurso permite a sua utilização em diversas áreas da computação, tais como tradução automática, sistemas de busca e extração de informação, sistemas conversacionais, sistemas de inferência e extração automática de ontologias.

Referências Bibliográficas

AMARAL, C.; FIGUEIRA, H.; MARTINS, A.; MENDES, A.; MENDES, P.; PINTO, C. (2006). Priberams question answering system for portuguese. *CLEF*.

DIAS DA SILVA, B.C.; DI FELIPPO, A., NUNES, M.G.V. (2008) The automatic mapping of Princeton WordNet lexical-conceptual relations onto the Brazilian Portuguese WordNet database. In: LREC, 6, 2008. Marrakech, Morocco. *Proceedings*. Marrakech, 2008.

HOUAISS (2001) Dicionário eletrônico Houaiss da língua portuguesa. Editora Objetiva. Versão 1.0.

KAISSER, M. (2005). Qualim at trec 2005: Web-question answering with framenet. *TREC*.

LIN, J. (2005). Evaluation of resources for question answering evaluation. *Technical report*, University of Maryland, College Park.

MACWHINNEY, B. (1995). The CHILDES project: Tools for analyzing talk (2nd ed.). Lawrence Erlbaum Associates.

OLIVEIRA, H.G.; COSTA, H.; SANTOS, D. (2012). Folheador: browsing through Portuguese semantic relations. *Conference of the European Chapter of the Association for computational Linguistics (EACL)*

OLIVEIRA, H.G.; PÉREZ, L.; GOMES, P. (2012) Exploring Onto.PT. Demo Session of PROPOR 2012, *10th International Conference on the Computational Processing of the Portuguese Language (PROPOR)*

OLIVEIRA, H. G.; SANTOS, D.; GOMES, P. (2009). Extracção de relações semânticas entre palavras a partir de um dicionário: o papel e sua avaliação. STIL 2009, *Linguamática*, p. 77–93.

OLIVEIRA, H. G.; SANTOS, D.; GOMES, P.; SECO, N. (2008). Papel: A dictionary-based lexical ontology for portuguese. In: TEIXEIRA, A.; DE LIMA, V. L. S.; DE OLIVEIRA, L. C.; QUARESMA, P. (Eds.) *Proceedings of Computational Processing of the Portuguese Language* (PROPOR), volume 5190 of LNAI, p. 31–40. Springer.

PORTO EDITORA. (2005). *Dicionário da Língua Portuguesa*. Porto.

SARMENTO, L.; TEIXEIRA, J. F.; OLIVEIRA, E. (2008). Experiments with query expansion in the raposa (fox) question answering system. The Cross-Language Evaluation Forum (CLEF).

XXXXX *OMITIDO PARA REVISÃO CEGA*

ZHENG, Z. (2002). Answerbus question answering system. *Proceeding of HLT Human Language Technology Conference* (HLT 2002).