Extração e Validação de Ontologias a partir de Recursos Digitais

Kassius Prestes[1], **Rodrigo Wilkens**[1], **Leonardo Zillio**[2], **Aline Villavicencio**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul(UFRGS)

[2]Instituto de Letras – Universidade Federal do Rio Grande do Sul(UFRGS)

Abstract. *This paper aims at presenting a methodology for semi-automatic validation of an wide-coverage ontology based on an existing electronic resource, PAPEL. From the existing relations, we choose those of synonymy and hypernymy to generate the ontology. The resulting output was converted to OWL format e manually validated by a lexicographer. As result, we have a wide-coverage ontological resource that can be used in different subareas of computer science. The resource displays concepts organized according to their hypernymy and validated synonymy relations.*

Resumo. *O objetivo desse trabalho é apresentar uma metodologia para validação semiautomática de uma ontologia de ampla cobertura, com base em um recurso eletrônico existente, o PAPEL. Das relações disponíveis, foram usadas as de sinonímia e de hiperonímia para construção da ontologia. Os resultados foram convertidos para o formato OWL e manualmente validados por um lexicógrafo. O resultado obtido foi um recurso ontológico de ampla cobertura que pode ser empregado em diversas áreas da computação. O recurso apresenta termos organizados a partir de suas relações de hiperonímia e de sinonímia, sendo estas validadas.*

1. Introdução

Sistemas computacionais como sistemas de perguntas e respostas (P&R) têm a tarefa de responder automaticamente uma questão em linguagem natural, procurando por informações em fontes de dados, tais como um banco de dados estruturado ou documentos não-estruturados em linguagem natural (p. ex., jornais). Esse tipo de sistema normalmente realiza quatro passos: análise da pergunta; identificação dos documentos candidatos; geração das respostas candidatas; e pontuação das respostas. Exemplos são [Kaisser 2005], [Lin 2005], [Zheng 2002], [Sarmento et al. 2008] e [Amaral et al. 2006].

Um exemplo de sistema de P&R para o português é o projeto Comunica [Wilkens et al. 2010], que busca responder perguntas sobre transferências constitucionais de municípios via telefone. Nele, tanto a pergunta do usuário quanto a resposta do sistema são em liguagem natural, visando a uma maior inclusão digital. O projeto se divide em quatro módulos-chave: reconhecimento de voz, processamento de texto, acesso a banco de dados e síntese de voz. O módulo de reconhecimento de voz realiza a conversão de áudio para texto. O processamento de texto tem a função de identificar dados relevantes informados pelo usuário a partir da frase transcrita (pelo módulo de reconhecimento de voz). A identificação dos conceitos é feita por meio de duas ontologias que validam as palavras da pergunta do usuário: uma de propósito geral e uma do domínio da aplicação. Os conceitos identificados são então buscados pelo módulo de acesso a banco de dados e a resposta gerada é sintetizada pelo módulo de síntese de voz. Para tal aplicação, é

necessária uma ontologia de ampla cobertura que possa auxiliar tanto o módulo de reconhecimento de voz a validar as palavras reconhecidas quanto o módulo de processamento de texto na identificação de conceitos.

O objetivo deste trabalho é apresentar a metodologia de extração e validação da ontologia de propósito geral usada no âmbito do projeto Comunica. A 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 lexicógrafo. O PAPEL é um conjunto de relações entre palavras extraídas automaticamente das definições em um tesauro eletrônico. Ele contém 199.672 entradas, distribuídas em 8 tipos de relações. Dentre essas, foram selecionadas as relações de hiperonímia [1] e de sinonímia [2] como base para a ontologia. Neste trabalho, são descritos os processos de conversão dos dois tipos de relações, com uma discussão da validação das relações de sinonímia.

Este artigo é estruturado da seguinte maneira: na Seção 2, são discutidos alguns trabalhos relacionados; a metodologia para extração e validação é descrita nas Seções 3 e 4; por fim, a Seção 5 apresenta conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Buitelaar [Buitelaar et al. 2005] define o aprendizado de ontologias como a aquisição de conhecimento a partir de textos. Grande parte desse trabalho utiliza como base áreas da computação como processamento de linguagem natural, inteligência artificial e aprendizado de máquina. Existem diversas abordagens para a extração automática de ontologias. Algumas abordagens são probabilísticas, como em [Venant 2008] [Chotimongkol and Rudnicky 2002] [Salton and Buckley 1988]. Contudo, uma das abordagens mais exploradas é a identificação de relações a partir de expressões textuais, como mostrado em [Hearst 1992]. A identificação baseada em expressões apresenta um melhor resultado na extração de documentos que possuem uma estrutura mais ou menos fixa, como dicionários, por isso, essa abordagem foi utilizada para a criação do PAPEL.

Ontologias, especialmente as de ampla cobertura, são recursos de grande valor para sistemas que empregam tecnologias de linguagem. Por exemplo, a WordNet [Miller 1995] é um recurso eletrônico que contém relações semântico-conceptuais e lexicais entre as palavras. Ela foi originalmente desenvolvida pela Universidade de Princeton para o inglês, e posteriormente estendida para outras línguas, inclusive para o português [Marrafa et al. 2005]. Nela, os termos são agrupados em synsets, onde todos os sinônimos de um termo estão no mesmo grupo que ele, contendo uma definição e um conjunto de relações linguísticas. A Wordnet é utilizada em aplicações como tradução automática, sistemas de busca e extração de informação, entre outros. A WordNet do português (WordNet.PT [3]) contém cerca de 19.000 termos, distribuídos em vários campos semânticos. O fragmento disponível é composto por termos de diversos domínios, como arte, saúde, transportes e vestuário.

[1] hiperonímia é uma relação entre palavras que dá idéia de um todo, da qual se originam diversas ramificações, por exemplo, veículo é hiperônimo de carro, barco e avião

[2] sinonímia é uma relação entre palavras da mesma categoria gramatical, com sentido parecido e com forma diferente, como por exemplo carro e automóvel

[3] Desenvolvida pelo Centro de Linguística da Universidade de Lisboa pelo CLG - Grupo de Computação do Conhecimento Léxico-Gramatical.

Outro importante recurso léxico para o português é o PAPEL, desenvolvido com o objetivo de prover uma ontologia geral da linguagem, de grande abrangência [Oliveira et al. 2008]. Esse recurso foi construido através de extração semiautomática baseada em padrões de expressões que ocorrem nas definições do Dicionário da Língua Portuguesa [Editora 2005]. Dessa forma, foram identificadas relações composicionais, hierárquicas e de sinonímia. Exemplos dessas relações seriam:

repartir SINONIMO_DE partilhar
vasqueiro PROPRIEDADE_DE_ALGO_QUE_CAUSA vasca
vazar ACCAO_QUE_CAUSA vazão
cabo PARTE_DE vassoura
navio HIPERONIMO_DE veleiro

Devido à sua abrangência (com 199.672 entradas), foi realizada uma avaliação por amostragem dos resultados da extração semiautomatica [Oliveira et al. 2009], sendo que 50% das relações de sinonímia apresentam erros em potencial. Por exemplo: deliberadamente SINONIMO_DE peito. Para contornar estes problemas apresentamos uma metodologia de extração e validação mais confiáveis (pela validação manual) inseridas no projeto Comunica[Wilkens et al. 2010].

3. Metodologia

Para a construção de uma ontologia de alta cobertura e precisão a partir dos dados disponibilizados no PAPEL, foi necessária a definição de uma metodologia para conversão e validação sistemática do recurso com a construção de um sistema para identificação automática de conflitos nas entradas definidas. Neste trabalho, é abordada a conversão de duas relações para o format OWL: sinônimos e hiperônimos.

3.1. Relações de Hiperonímia

O PAPEL contém 61.263 entradas com relações de hiperonímia expressas no seguinte formato: palavra_1 HIPERONIMO_DE palavra_2. Para a conversão, foi utilizada a linguagem de programação Java e o framework Jena. A metodologia de conversão prevê duas etapas: conversão passiva e conversão ativa.

3.1.1. Conversão Passiva

A abordagem passiva consiste em armazenar as classes ontológicas apenas quando estas são apresentadas pelo PAPEL (as classes vão sendo criadas enquanto o arquivo de relações é lido). Para cada entrada do PAPEL, definida em termos de duas palavras (palavra_1 e palavra_2):

1. Para cada palavra:
   (a) Verificar se a palavra já existe como classe (dadas as várias palavras repetidas no PAPEL),
   (b) Se não existe, criar a classe.
2. Adicionar à ontologia a relação onde a classe palavra_1 é superclasse da classe palavra_2.

Esse processo é repetido para cada uma das relações de hiperonímia existentes. Devido ao número de entradas do PAPEL e às várias comparações necessárias, esta abordagem se mostrou muito custosa em termos de processamento e alto consumo de memória [4].

[4] Apenas cerca de 3% das relações em um período de aproximadamente 24 horas.

3.1.2. Conversão Ativa

Para tornar o processo mais eficiente em ontologias de alta cobertura, a segunda abordagem proposta cria inicialmente uma nova ontologia esquemática, e as informações já validadas somente são inseridas na nova ontologia ao final de cada passo de validação. Esse processo é composto por duas etapas. Primeiro, define-se uma ontologia básica com todas as classes necessárias, mas sem as relações entre elas. Para tanto, extraem-se do PAPEL todas as palavras sem repetição e cria-se uma classe para cada uma delas. A partir dessa ontologia básica com as classes necessárias, adicionam-se as relações entre as classes. Dada uma definição no PAPEL no formato *palavra_1 HIPERONIMO_DE palavra_2*, procura-se na ontologia básica as classes relativas a palavra_1 e palavra_2 e adiciona-se a relação à ontologia. Como resultado, obtem-se a conversão da estrutura de hiperonimos do PAPEL para o formato OWL.

3.2. Relações de Sinonímia

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. Dado o contexto deste trabalho, foram extraídas as relações de sinonímia entre substantivos [5]. A extração de sinônimos foi realizada através das seguintes etapas:

1. Criar uma lista (inicialmente vazia) de conjuntos (inicialmente vazios). Cada conjunto armazenará palavras que são sinônimas entre si.
2. Para cada uma das entradas, identificar as duas palavras sinônimas presentes.
3. Verificar se uma dessas palavras já se encontra em algum conjunto de sinônimos existente.
    (a) Se sim, insere-se a outra palavra nesse conjunto.
    (b) Se as palavras não estavam em nenhum conjunto existente, cria-se um novo conjunto contendo ambas.

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. Para criar uma ontologia de alta precisão e cobertura, a aplicação da metodologia foi semiautomática, e esses casos foram supervisionados por um lexicógrafo, com a seguinte modificação no passo 3:

3. Verificar se uma dessas palavras já se encontra em algum conjunto de sinônimos existente.
    (a) Se sim, realizar a verificação manual da ambiguidade, analisando os dois conjuntos em que as palavras seriam inseridas. Caso necessário, editar os conjuntos manualmente, separando as palavras de modo adequado.
    (b) Se nenhuma estava em um conjunto existente, criar um novo conjunto com ambas.

Ao final do processo obtem-se um amplo conjunto de grupos de sinônimos, que pode ser integrado à estrutura da ontologia gerada no passo anterior.

[5] Porém, essa abordagem pode ser, em princípio, aplicada diretamente aos outros tipos de relação.

4. Validação das Relações de Sinonímia

As relações de sinonímia da ontologia resultante foram manualmente validadas. Essa validação ocorreu com consulta a dicionários de língua portuguesa e a contextos reais de ocorrência dos pares de sinônimos analisados [6]. Caso as definições dos dicionários não aclarassem o problema, utilizou-se o buscador on-line do Yahoo! para se observarem também os contextos de ocorrência.

Dada a magnitude do recurso gerado, para palavras polissêmicas, a validação foi realizada com base no significado mais frequente apropriado para o grupo de sinônimos, sendo que cada substantivo poderia estar presente em apenas um grupo de sinônimos. Por exemplo: dada a avaliação da relação de sinonimia proposta entre abatimento, diminuição e desânimo, apesar de desânimo e diminuição não parecerem sinônimas sem um contexto muito específico, a palavra abatimento pode ser considerada sinônima de ambas, entre outras. A consulta a 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 outras eram irrelevantes (definições de dicionários on-line etc.). Como resultado, abatimento foi incluída com diminuição e excluída do conjunto de desânimo.

A adoção da metodologia proposta permite reduzir a subjetividade envolvida em todo o processo de decidir que palavras são sinônimas entre si e quais não devem ser [7]. Isso se torna importante para a replicabilidade do processo de validação, tendo em vista a natureza inerentemente subjetiva e depedente do vocabulário do avaliador da decisão de quais palavras possuem uma semelhança de significado.

Ao final desse processo, a ontologia resultante contém 46.904 entradas, com 40.614 relações de hiperonímia e 20.096 de sinonímia. O recurso apresenta alta abrangência e pode ser utilizado em uma grande variedade de sistemas de tecnologia de linguagem.

5. Conclusões e Trabalhos Futuros

Este artigo propôs uma metodologia para validação de uma ontologia com relações de hiperonímia e sinonímia a partir de um recurso lexical eletrônico. Foi apresentada em detalhes a avaliação de sinonímia realizada manualmente e com decisões auxiliadas por outros recursos. Apesar da magnitude do recurso original e das relações a serem adicionadas, essa metodologia possibilitou uma validação mais ampla do recurso, em vez da avaliação por amostragem proposta em [Oliveira et al. 2009]. Após a etapa apresentada neste trabalho, resta ainda a validação da estrutura de hiperônimos. Contudo, 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.

Agradecimentos

Esta pesquisa tem apoio dos projetos COMUNICA (FINEP/SEBRAE 1194/07), CAPES-COFECUB (707/11) e CNPq (479824/2009-6 e 309569/2009-5).

[6] Os dicionários utilizados foram o Houaiss Eletrônico (disponível em CD), o Michaelis On-line e o Dicionário da Língua Portuguesa da Porto Editora.

[7] Tal procedimento pode ser suficiente para os casos que envolvem mais de um significado frequente, mas, dada a abrangência do recurso, tal avaliação manual se torna impraticável.

References

Amaral, C., Figueira, H., Martins, A., Mendes, A., Mendes, P., and Pinto, C. (2006). Priberams question answering system for portuguese. *CLEF*.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12.

Chotimongkol, A. and Rudnicky, A. (2002). Automatic concept identification in goal-oriented conversations. In *Seventh International Conference on Spoken Language Processing*.

Editora, P. (2005). *Dicionário PRO da Língua Portuguesa*. Porto.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

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.

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., and Mendes, S. (2005). Wordnet.pt uma rede léxico-conceptual do português on-line. *XXI Encontro da Associação Portuguesa de Linguística*, pages 28–30.

Miller, G. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Oliveira, H. G., Santos, D., and Gomes, P. (2009). Extracção de relações semânticas entre palavras a partir de um dicionário: o papel e sua avaliação. *STIL 2009, Linguamática*, pages 77–93.

Oliveira, H. G., Santos, D., Gomes, P., and Seco, N. (2008). Papel: A dictionary-based lexical ontology for portuguese. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Proceedings of Computational Processing of the Portuguese Language (PROPOR)*, volume 5190 of *LNAI*, pages 31–40. Springer.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval* 1. *Information processing & management*, 24(5):513–523.

Sarmento, L., Teixeira, J. F., and Oliveira, E. (2008). Experiments with query expansion in the raposa (fox) question answering system. *The Cross-Language Evaluation Forum (CLEF)*.

Venant, F. (2008). Semantic visualization and meaning computation. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*, pages 185–188. Association for Computational Linguistics.

Wilkens, R., Villavicencio, A., Muller, D., Wives, L., da Silva, F., and Loh, S. (2010). Comunica - a question answering system for brazilian portuguese. *Coling 2010*.

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