Declaração de Pesquisa: Extração Automática de Ontologias da Wikipédia

Clarissa Castellã Xavier[1], Vera Lúcia Strube de Lima[1]

[1] Faculdade de Informática – PUCRS, Av. Ipiranga, 6681 – Prédio 32, Porto Alegre, Brazil
{clarissa.xavier, vera.strube}@pucrs.br

Resumo. Apresentamos um apanhado sobre o trabalho sendo desenvolvido no grupo de pesquisa em PLN da PUCRS e, entre as pesquisas em curso, selecionamos a extração de ontologias da Wikipédia para enfoque. Nesta pesquisa propomos um *framework* para extração automática de estruturas ontológicas a partir da Wikipédia.

1 Declaração de Pesquisa

O grupo de pesquisa em Processamento da Linguagem Natural (PLN) da PUCRS tem como foco a pesquisa em PLN e tem buscado voltar suas pesquisas para a língua portuguesa, embora abra espaço para a pesquisa em outras línguas. Nos últimos anos, as aplicações de recuperação de informações reconhecimento de entidades nomeadas e categorização de textos em português, bem como as ontologias, passaram a ser foco dos trabalhos do grupo. As ontologias são estudadas especialmente no que se refere à construção e enriquecimento, cálculo da similaridade, extração de conceitos e extração semiautomática e automática de estruturas semânticas a partir de corpora e grandes bases de dados, como a Wikipédia.

Uma definição usualmente utilizada para ontologia[1] e aqui adotada é a de Gruber [1]: "uma ontologia é uma especificação formal e explícita de uma conceitualização de um domínio".

Construir manualmente ontologias [2] é um processo oneroso, tedioso e propenso a erros. Além disso, o número de ontologias de domínio disponíveis é extremamente pequeno, sendo em número menor ainda em língua portuguesa [3].

A Wikipédia, de acordo com o pelo projeto que a mantém[2], é uma enciclopédia multilíngue, *online*, livre, sem fins lucrativos, colaborativa, criada em 15 de janeiro de 2001. A enciclopédia *online* tem se mostrado uma fonte muito interessante para extração de estruturas ontológicas, visto que conta com milhões de entradas, centenas

[1] Utilizaremos os termos "ontologia" e "estrutura ontológica" intercambiadamente, e adotaremos, para ontologia, a abordagem mais aberta, que pode remeter a uma terminologia dotada de relações semânticas simples.

[2] Definição disponível em http://pt.wikipedia.org/wiki/Wikipédia

de milhares de colaboradores e milhões de artigos revisados [4], cobrindo uma extensa faixa de assuntos, sendo uma das mais importantes coleções de conteúdo geradas por usuários [5] disponível livremente.

Os documentos da Wikipédia estão organizados em uma hierarquia de categorias que indexa os artigos [6], que são páginas que contêm informações sobre algum assunto. O corpo do artigo pode conter informações suplementares como tabelas, imagens, mensagens em outras línguas, mensagens para os outros contribuintes da Wikipédia, referências a outros artigos da enciclopédia, *hyperlinks* para sítios externos [7] e *infoboxes*. *Infobox* é um modelo padrão da Wikipédia, que contém uma tabela com informações básicas sobre a entidade descrita no artigo em que está inserido. Por exemplo, *infoboxes* de artigos que descrevem países costumam conter informações como o nome do país na língua nativa, sua capital e área. Artigos mais longos geralmente são divididos em seções ou subseções e possuem um índice.

O uso crescente, a dificuldade de criação manual e a carência de ontologias disponíveis, particularmente em língua portuguesa, bem como a riqueza de conteúdo disponibilizado pela Wikipédia, motivaram uma investigação a respeito das possibilidades da exploração da Wikipédia para a extração de ontologias na língua portuguesa. Dentro deste contexto, iniciamos com uma primeira questão de pesquisa: é possível extrair estruturas ontológicas de domínio em português a partir da estrutura de categorias da Wikipédia?

Experimentos intermediários, detalhados em [8] e [9], efetuando a extração de uma estrutura taxonômica e de relações de localização a partir da categoria Turismo da Wikipédia em português, levaram à proposta de um método semiautomático para extração de estruturas ontológicas de domínio da Wikipédia [10] e [11].

O método foi validado através de um estudo de caso criando-se um protótipo que gerou uma estrutura ontológica do domínio de Turismo tal como organizado na Wikipédia em língua portuguesa. A avaliação dos resultados obtidos foi conduzida comparando a estrutura gerada pelo protótipo com um mapeamento de referência elaborado por uma linguista, também baseado na estrutura da categoria Turismo. Os resultados foram promissores, obtendo-se 79.51% de Precisão e 91.95% de Cobertura, comparáveis aos encontrados na literatura para outros idiomas.

Iniciamos, assim, a busca pela mais completa automatização da extração de ontologias da Wikipédia. Neste ponto levantamos uma nova questão: é possível gerar automaticamente estruturas ontológicas de domínio a partir da Wikipédia? Também ampliamos nosso olhar para as línguas portuguesa e inglesa.

Evoluindo nessa pesquisa, visualizamos a construção de uma solução para a extração automática de estruturas ontológicas da Wikipédia, que tenha as seguintes características:

- Ser totalmente automática.
- Funcionar para qualquer domínio do conhecimento cadastrado na enciclopédia.
- Utilizar apenas a Wikipédia como fonte de dados (sem uso de recursos externos).
- Ser multilíngue (com validação realizada em inglês e português).
- Possuir módulos que funcionem integrados ou de maneira independente.
- Ter como entrada de dados o banco de dados da Wikipédia.
- Ter como saída uma estrutura descrita em OWL.

Para viabilizar a solução proposta, buscamos ampliar o uso dos recursos oferecidos pela enciclopédia e explorar diferentes métodos de extração de informação. Este trabalho está sendo conduzido de forma gradual, gerando inicialmente estruturas semânticas simples e gradativamente iremos refinar os métodos utilizados, de modo a qualificar a estrutura obtida no final.

2 Estado Atual do Projeto

Foi realizado um levantamento do estado da arte das iniciativas que realizam extração de dados semânticos da Wikipédia, em combinação, ou não, com outras fontes de dados. Estes trabalhos foram estudados e contrastados, focando-se no modo como o conteúdo e a estrutura da enciclopédia são explorados, métodos e tecnologias utilizados, forma e teor dos dados extraídos e resultados obtidos.

A partir deste estudo algumas tendências nos trabalhos que realizam extração de estruturas ontológicas da Wikipédia foram identificadas:

- Uso da Wordnet[3] em combinação da Wikipédia [12], [13], [14], [15], [16]
- Exploração dos *infoboxes* [12], [13], [15], [16]
- Exploração da estrutura de categorias [12], [13], [14], [16], [17], [18]
- Exploração dos títulos dos artigos e categorias [12], [14], [16], [17], [18]
- Exploração da conexão entre os artigos [14], [17], [18]
- Exploração das páginas de desambiguação e redirecionamento [13], [16], [17], [18]

A partir dos dados apurados neste estudo, e tendo em vista nosso objetivo de extrair automaticamente estruturas ontológicas de domínio da Wikipédia, foi idealizado o desenvolvimento de um primeiro *Framework* intitulado *Et cetera*[4], composto por quatro módulos, representado na Figura 1.

[3] wordnet.princeton.edu

[4] *et cetera*: "e outras coisas da mesma espécie" – segundo Claudio Moreno em http://wp.clicrbs.com.br/sualingua/2009/05/13/pontuacao-do-etc/

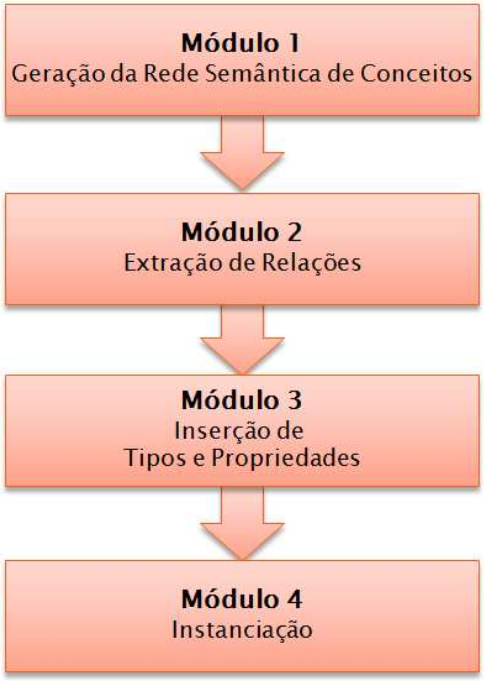

Fig. 1. Arquitetura do *framework Et cetera*.

O primeiro módulo deste *framework* extrairá uma rede semântica de conceitos na forma de um grafo direcionado. A partir da categoria que apresenta o domínio do conhecimento a ser representado, serão extraídas informações dos seguintes recursos da Wikipédia: estrutura de categorias, nome dos artigos pertencentes à categoria, páginas de redirecionamento e ligação entre artigos e categorias.

O segundo módulo irá rotular as relações da rede semântica gerada no primeiro módulo, explorando a conexão entre artigos e categorias, páginas de redirecionamento e o texto dos artigos.

O terceiro módulo adicionará tipos e propriedades aos conceitos da rede semântica, gerando uma ontologia de domínio contendo classes, relações, propriedade e tipos. Nesta tarefa as informações são extraídas do texto dos artigos e *infoboxes*.

O quarto módulo irá popular a ontologia com instâncias.

Os próximos passos de nossa pesquisa são a construção de um protótipo que implemente os dois primeiros módulos do *framework*, dando início à especificação, desenvolvimento e testes das estratégias a serem empregadas.

2 Áreas de Pesquisa do Grupo de Pesquisa

Embora o grupo exista desde o início da década de 1990, abordando diferentes temas da linguística computacional tais como aplicações de Recuperação de Informações e

Categorização de Textos em português, a partir dos anos 2000 as ontologias também passaram a ser estudadas, especialmente no que se refere ao cálculo da similaridade, portais de ontologias, extração de conceitos, construção semiautomática de ontologias a partir de textos, obtenção de métricas, visualização e extração de ontologias da Wikipédia. Para um detalhamento consultar http://www.inf.pucrs.br/~linatural/.

3 História do grupo e dos membros do grupo

O grupo de pesquisa foi criado em 1990 e tem voltado suas pesquisas para a língua portuguesa, sendo que os trabalhos desenvolvidos se destacam em duas linhas: formação de pessoas e produção científica. Na formação, destacamos a qualidade da formação dos egressos e a fixação dos mesmos em instituições de ensino superior, órgãos públicos e empresas privadas. Na vertente da geração de conhecimento, o grupo tem desenvolvido projetos com instituições nacionais (PUCRS, UFRGS, UFSC, UNICAMP, UFSCAR e USP) e estrangeiras. Os laços de cooperação internacional são estabelecidos principalmente com Portugal, Espanha e Uruguai. Nos últimos anos estreitamos a aproximação com empresas, como a HP e a Plugar. O grupo se envolve com atividades da comunidade científica da área de Processamento de Linguagem Natural, através da participação em sociedades científicas nacionais como a SBC (Sociedade Brasileira de Computação) e internacionais como a ACL (*The Association for Computational Linguistics*), participando tanto como membros de diversos comitês de programa, assim como na organização de eventos científicos na área.

Atualmente o grupo conta com 2 coordenadoras, 3 pesquisadores, 8 estudantes de doutorado, 6 de mestrado e 5 alunos de graduação.

Referências

1. Gruber, T. R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: International Journal of Human and Computer Studies, vol. 43–5–6, pp. 907--928 (1993)
2. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers (2002)
3. Lima, V. L. S., Nunes, M. G. V., Vieira, R.: Desafios do Processamento de Línguas Naturais. In: 34º Seminário Integrado de Software e Hardware (SEMISH 2007), pp. 2202--2216 (2007)
4. Spinellis, D., Louridas, P.: The collaborative organization of knowledge. In: Communications of the ACM, vol. 51-8, pp. 68--73 (2008)
5. Mika, P., Ciaramita, M., Zaragoza, H., Atserias, J.: Learning to Tag and Tagging to Learn: A Case Study on Wikipedia. In: IEEE Intelligent Systems, vol. 23-5, pp. 26--33 (2008)
6. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipédia. In: Proceedings of the 15th international Conference on World Wide Web, pp. 585--594 (2006)
7. Schönhofen, P., Benczur, A., Biro, I., Csalogany, K.: Performing Cross-Language Retrieval with Wikipedia. In: Cross-Language Evaluation Forum 2007 (CLEF 2007). Budapeste, Hungria (2007)

8. Xavier, C. C.; Lima, V. L. S. “Construção de uma Estrutura Ontológica de Domínio a partir da Wikipédia”. In: 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009) (2009)
9. Xavier, C. C., Lima, V. L. S.: Extração de Estruturas Ontológicas a partir da Wikipédia em Língua Portuguesa. In: 9º Workshop de Teses (2009)
10. Xavier, C. C.: Extração de estruturas ontológicas de domínio da Wikipédia em língua portuguesa. Diss. (Mestrado) – Fac. de Informática, PUCRS. Porto Alegre (2010)
11. Xavier, C. C., Lima, V. L. S.: A Semi-Automatic Method for Domain Ontology Extraction from Portuguese Language Wikipedia's Categories. In: 20th SBIA - Brazilian Symposium on Artificial Intelligence (SBIA 2010), (2010)
12. Syed, Z., Finin, T.: Unsupervised techniques for discovering ontology elements from Wikipedia article links. In: Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR '10), pp. 78—86. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
13. Suchanek, F. M., Kasneci, G., Weikum, G.Yago: A large ontology from wikipedia and wordnet. In: Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6(3), pp. 203--217 (2008)
14. Navigli R., Ponzetto S.P.: BabelNet: building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 216-225 (2010)
15. Szumlanski, S., Gomez, F.: Automatically acquiring a semantic network of related concepts. In: Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). ACM, New York, NY, USA, pp. 19--28 (2010)
16. Melo, G., Weikum G.: MENTA: inducing multilingual taxonomies from wikipedia. In: Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). ACM, New York, NY, USA, pp. 1099--1108 (2010)
17. Fogarolli, A.: Wikipedia as a Source of Ontological Knowledge: State of the Art and Application. In: Caballé, S., Xhafa, F., Abraham, A. (eds.) Intelligent Networking, Collaborative Systems and Applications. Studies in Computational Intelligence, vol.329, pp. 1—26. Springer Berlin / Heidelberg (2011)
18. Nastase, V., Strube, M., Boerschinger, B., Zirn, C., Elghafari, A. Wikinet: A very large scale multilingual concept network. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10), Valletta, Malta (2010)