Extração de Vocabulário Multilíngue para Tradução em Domínios Especializados

Lucas Welter Hilgert, Renata Vieira

Faculdade de Informática (FACIN) – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brazil

lucaswhilgert@gmail.com, renata.vieira@pucrs.br

Abstract. *This paper presents a process for multilingual vocabulary extraction which aims to auxiliate machine translation tools during the intermediation of the communication between multilingual teams.*

Resumo. *Este trabalho apresenta um processo para a extração de vocabulário multilíngue proposto com o objetivo de auxiliar ferramentas de tradução de máquina durante a intermediação da comunicação entre equipes multilíngues.*

1. Introdução

Serviços de tradução de máquina em tempo real tem sido considerados como alternativas promissoras para o auxílio à comunicação entre equipes multilíngues durante a execução de tarefas colaborativas [Calefato et al. 2011].

No entanto, como demonstrado em diferentes trabalhos [Calefato et al. 2011] [Yamashita and Ishida 2006] [Yamashita et al. 2009], estes serviços apresentam problemas que interferem negativamente no processo de comunicação, causando atrasos durante o estabelecimento de um conhecimento comum (*common ground* entre as equipes distribuídas [Calefato et al. 2011] [Yamashita and Ishida 2006].

Dentre os principais problemas apontados na literatura (e identificados a partir da análise de registros de comunicação [Hilgert et al. 2012]), optou-se neste trabalho por abordar aqueles relacionados a traduções inadequadas ou inconsistentes (múltiplas traduções para um mesmo contexto), para os quais a construção de um vocabulário específico de domínio é apontado como uma das possíveis soluções [Nakatsuka et al. 2010].

Sendo assim, este artigo apresenta um processo para a extração de vocabulário multilíngue (português-inglês) a partir de um corpus paralelo composto por manuais de *software*, focado na extração de equivalências multilíngues de palavras mais específicas do domínio.

Inicialmente, a Seção 2 apresenta o corpus bilíngue empregado na pesquisa, seguido pela apresentção do processo de extração proposto (Seção 3), do método de avaliação utilizado e dos resultados obtidos (Seção 4). Por fim, a Seção 5 apresenta as principais conclusões e trabalhos futuros.

2. Corpus

Um corpus, no contexto deste trabalho, pode ser definido com um conjunto de documentos construído de acordo com um objetivo específico (extração de vocabulário

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 194–198,
Fortaleza, CE, Brazil, October 21–23, 2013.

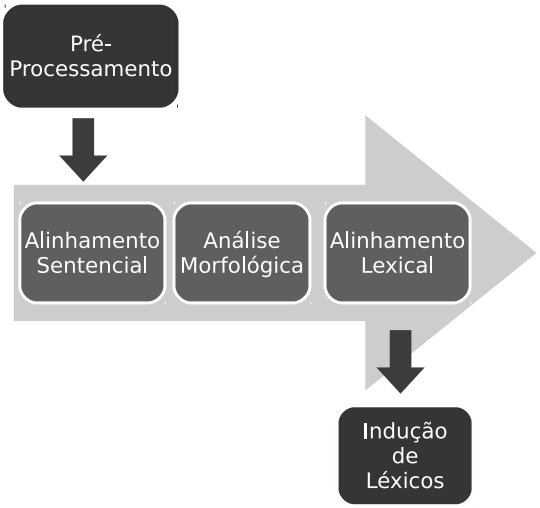

Figure 1. Processo proposto

bilíngue, por exemplo). Corpus paralelo, por sua vez, consiste em um corpus formado por documentos acompanhados de suas respectivas traduções para um segundo idioma [Ha et al. 2008].

O corpus utilizado no decorrer deste trabalho foi construído de forma manual a partir de manuais de *softwares* de projetos de código livre (*open source*), tendo o sítio desses como repositório. O corpus construído foi composto por 553.333 palavras (29.770 sentenças) para o português e 556.027 palavras (28.619 sentenças) para inglês, sendo considerado como de tamanho mediano de acordo com padrões linguísticos.

3. Processo Proposto

O processo proposto (baseado em [Caseli and Nunes 2007] [Tiedemann 2012] [Ha et al. 2008] e [Zhang 2009]) é apresentado na Figura 1 e descrito a seguir.

3.1. Pré-Processamento

Durante o pré-processamento os documentos componentes do corpus foram convertidos para texto puro (sem formatação e ilustrações) e filtrados para a remoção de símbolos considerados como sujeira (marcadores, caracteres não codificáveis, etc.).

Posteriormente, os documentos foram organizados no formato de uma sentença por linha, como exigido pelas ferramentas de alinhamento sentencial (discutido na próxima seção).

3.2. Alinhamento Sentencial

O alinhamento sentencial consiste na identificação de sentenças equivalentes entre os documentos paralelos do corpus, ou seja, em identificar possíveis traduções para essas.

Dentre as ferramentas de alinhamento sentencial encontradas ([Caseli 2003] [Moore 2002] e [Varga et al. 2005]), optou-se pela utilização do *Bilingual Sentence Aligner* [Moore 2002] (melhor desempenho em testes anteriores) para a realização dos alinhamentos.

3.3. Análise Morfológica

A etapa de análise morfológica consiste na atribuição de informações morfológicas (classe gramatical, flexões de número e gênero, etc.) às palavras formadoras dos documentos, tendo como objetivo a eliminação de ambiguidades.

Durante esta etapa é conduzida, ainda, a identificação de expressões multipalavras (conjunto de palavras) [Ramisch et al. 2010] e a unificação de seus componentes (mediante inserção do símbolo "_"). Esta unificação se faz necessária durante o alinhamento lexical.

Nesta etapa foram utilizadas ferramentas da plataforma de tradução *Apertium* [Forcada et al. 2011] que, para um maior desempenho tanto na identificação de expressões multipalavras quanto na atribuição dos rótulos morfológicos, tiveram seus dicionários morfológicos ampliados.

Esta ampliação foi realizada a partir das listas de palavras (e expressões multipalavras) extraídas pelas ferramentas de extração terminológica *ExATOlp* [Lopes et al. 2012] (português) e *TTC TermSuite* [Rocheteau and Daille 2011] (inglês). Essas, além de fornecerem as informações morfológicas necessárias, realizam a identificação das palavras mais relevantes de um determinado domínio, permitindo a especialização do vocabulário.

3.4. Alinhamento Lexical

O alinhamento lexical consiste na identificação de equivalências entre palavras e expressões multipalavras de sentenças consideradas paralelas (previamente alinhadas), logo, na definição de traduções para essas.

Para a realização desta etapa foi utilizada a ferramenta *Giza++* (versão 2.0) [Och and Ney 2003]. O alinhamento foi conduzido em ambos os sentidos (português-inglês e inglês-português) e os resultados desses foram unidos utilizando o algoritmo de união (em conjunto com heurísticas) proposto em [Och and Ney 2003].

3.5. Indução do Vocabulário Bilíngue

Por fim, a partir dos resultados dos alinhamentos anteriores, foram induzidos os vocabulários bilíngues mediante a utilização da ferramenta *ReTraTos* [Caseli and Nunes 2007].

O *ReTraTos* extrai, a partir da lista de equivalências gerada na etapa anterior, uma segunda lista, composta pelas entradas consideradas como mais relevantes. Esta segunda lista (composta pelo vocabulário bilíngue) é então formatada de acordo com o formalismo utilizado nos dicionários bilíngues da plataforma de tradução *Apertium* [Forcada et al. 2011].

4. Avaliação e Resultados

Os dicionários gerados ao final do processo foram avaliados de forma intrínseca (tradução das palavras avaliada fora da sentença) e manual. Esta estratégia de avaliação foi selecionada devido a padrões de referência (*golden standards*) não terem sido encontrado.

Para a avaliação, cada avaliador recebeu um conjunto composto por 200 entradas do dicionário (palavra acompanhada de sua tradução) referentes a palavras simples e 50

entradas referentes a expressões multipalavras. Estas entradas foram classificadas (de acordo com a equivalência bilíngue) em: Válidas (V), Parcialmente Válidas (PV), e Não-Válidas (NV) (de acordo com [Caseli and Nunes 2007]). Ao final, entradas nas categorias V e PV foram classificadas como corretas.

Como resultados, obteve-se uma precisão de 81% na identificação de equivalências para palavras simples e 39% na identificação de expressões multipalavras. Estes valores são próximos aos apresentados em [Caseli and Nunes 2007] (80% para palavras simples e 38% para multipalavras), apresentando uma vantagem em relação às expressões multipalavras.

5. Conclusões e Trabalhos Futuros

Como pode ser observado, mediante os resultados apresentados, o processo proposto apresentou resultados próximos aos trabalhos de referência [Caseli and Nunes 2007], tendo apresentado melhor desempenho na identificação de expressões multipalavras.

Estima-se que esta melhora tenha sido causada pelo enfoque dado ao reconhecimento deste tipo de estrutura durante a etapa de análise morfológica, com a ampliação dos dicionários mediante a inserção de palavras mais específicos de domínio.

A investigação desta hipótese, bem como da relação do vocabulário extraído com os conceitos de domínio estão previstos como trabalhos futuros. Ainda como trabalhos futuros, prevê-se a otimização de cada uma das etapas envolvidas com a utilização de ferramentas como, por exemplo, o *mweToolkit* [Ramisch et al. 2010] em conjunto com o *ExATOlp* [Lopes et al. 2012] durante a análise morfológica.

6. Agradecimentos

Agradecimentos à FAPERGS pelo incentivo financeiro concedido ao projeto.

Referências

Calefato, F., Lanubile, F., and Prikladnicki, R. (2011). A controlled experiment on the effects of machine translation in multilingual requirements meetings. In *Global Software Engineering (ICGSE), 2011 6th IEEE International Conference on*, pages 94 –102.

Caseli, H. M. (2003). Alinhamento sentencial de textos paralelos português-inglês. Master's thesis, ICMS-USP.

Caseli, H. M. and Nunes, M. (2007). Automatic induction of bilingual lexicons for machine translation. In *International Journal of Translation*, volume 19, pages 29–43.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Ha, L. A., Fernandez, G., Mitkov, R., and Corpas, G. (2008). Mutual bilingual terminology extraction. In *Proceedings of LREC*, Marrakesh, Morocco.

Hilgert, L., Calefato, F., Lanubile, F., Prikladnicki, R., Vieira, R., Finatto, M. J., and Termignoni, S. (2012). Real-time machine translation for software development teams. Technical report, PUCRS.

Lopes, L., Vieira, R., Fernandes, P., and Couto, G. (2012). Exatolp: extraction of language resources from portuguese corpora. In *International Conference on Computational Processing of the Portuguese Language - PROPOR*.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK. Springer-Verlag.

Nakatsuka, M., Yasunaga, S., and Kuwabara, K. (2010). Extending a multilingual chat application: Towards collaborative language resource building. In *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*, pages 137 –142.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). mwetoolkit: a Framework for Multiword Expression Identification. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta. European Language Resources Association.

Rocheteau, J. and Daille, B. (2011). Ttc termsuite: A uima application for multilingual terminology extraction from comparable corpora. In *5th International Joint Conference on Natural Language Processing*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1246.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.

Yamashita, N., Inaba, R., Kuzuoka, H., and Ishida, T. (2009). Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 679–688, New York, NY, USA. ACM.

Yamashita, N. and Ishida, T. (2006). Effects of machine translation on collaborative work. *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work CSCW 06*, page 515.

Zhang, C. (2009). Extracting chinese-english bilingual core terminology from parallel classified corpora in special domain. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '09, pages 271–274, Washington, DC, USA. IEEE Computer Society.