Extração de Vocabulário Multilíngue a partir de Documentação de *Software*

Lucas Welter Hilgert, Renata Vieira, Rafael Prikladnicki

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

Abstract. *This work aims for extracting multilingual vocabulary from software documentation in order to create resources for improving machine translation assisted comunication in the context of software requirement meetings involving multilingual teams. The objective of this paper is to present the initial results obtained related to the research data (corpus) construction.*

Resumo. *Este trabalho tem por objetivo a extração de vocabulário multilíngue a partir de documentação de software, visando construir recursos para o melhoramento da comunicação assistida por tecnologias de tradução de máquina no contexto de reuniões de requisitos de software envolvendo times multilíngues. Este artigo tem por objetivo apresentar os resultados iniciais obtidos durante a construção do material de pesquisa (corpus).*

1. Introdução

O trabalho aqui apresentado encontra-se inserido no âmbito do projeto "*O Efeito do Processamento da Linguagem Natural no Desenvolvimento da Capacidade do Brasil no Mercado Global de Desenvolvimento de Software*", cujo principal objetivo é auxiliar na inclusão de equipes brasileiras no mercado global de desenvolvimento, mediante a investigação e utilização de métodos, técnicas e ferramentas da área de Processamento da Linguagem Natural (PLN).

Dentre as tecnologias de PLN, dá-se enfoque especial aos serviços de tradução de máquina, considerados como uma solução alternativa para as dificuldades linguísticas (diferentes idiomas) encontradas durante reuniões de equipes multilíngues de desenvolvimento de *software* [Calefato et al. 2012] [Yamashita and Ishida 2006].

Como apresentado em diferentes trabalhos [Calefato et al. 2011] [Calefato et al. 2012] [Yamashita and Ishida 2006], as tecnologias de tradução automática ainda estão longe da perfeição, possuindo uma série de questões a serem resolvidas, para as quais, uma das possíveis soluções é a construção de vocabulários multilíngues específicos do domínio [Nakatsuka et al. 2010].

Sendo assim, este trabalho tem como principal objetivo a construção de um vocabulário multilíngue referente às práticas de desenvolvimento distribuído de *software*, com a finalidade de auxiliar os serviços de tradução de máquina empregados durante as reuniões das equipes.

2. Contextualização do Trabalho

Dos trabalhos referenciados, destaca-se o experimento conduzido por Calefato *et al.* [Calefato et al. 2012], executado em uma parceria entre pesquisadores brasileiros (PUCRS) e italianos (Universidade de Bari), cujos registros (*logs*) foram utilizados como principal fonte para a investigação de problemas relacionados à tradução automática aplicada ao contexto de tarefas colaborativas.

Neste experimento, equipes multilíngues (formadas por 2 participantes brasileiros e 2 italianos) executaram, colaborativamente, tarefas relacionadas à engenharia de requisitos, utilizando, de forma alternada, o inglês como idioma comum, e seus idiomas nativos em conjunto com serviços de tradução de máquina.

A partir da análise dos registros do experimento, diferentes tipos de problemas foram encontrados sendo que, neste trabalho, optou-se por priorizar aqueles relacionados ao vocabulário, destacando-se: (1) traduções inconsistentes, (2) abreviações de termos, (3) erros de digitação.

Como exemplo de tradução inconsistente, pode-se mencionar o termo *"release"* , traduzido de diferentes formas (*"entrega"* e *"lançamento"*, por exemplo) para os mesmos contextos (motivo da inconsistência), ou mesmo mantido como *"release"*. Este tipo de inconsistência, pode induzir a problemas de compreensão entre os participantes da reunião [Nakatsuka et al. 2010].

A abreviação de termos se demonstrou um problema, principalmente quando aplicada à termos cujas abreviações possuem significado próprio. Como exemplos deste tipo de problemas, pode-se mencionar *"bluetooth"*, simplificado como *"blue"* gerando a tradução (inadequada ao contexto) *"azul"*, e *"ring tone"*, abreviado como *"ring"* e traduzido como *"anel"*.

Uma das possíveis soluções encontradas é a utilização de funcionalidades de autocomplementação para auxiliar os participantes durante a escrita. Estas funcionalidades podem ser alimentadas com um vocabulário inicial, a ser ampliado no decorrer da comunicação.

Em relação aos erros ortográficos, pode-se mencionar o termo *"bluetooth"* para o qual foram encontradas 5 diferentes grafias sendo 4 destas incorretas (*"blutooth"*, *"bluetoth"*, *"bluetooh"* e *"blutoofh"*).

Este tipo de erro é frequentemente tratado através de funcionalidades de correção ortográfica (*spellchecking*) sendo que essas dependem da existência de um vocabulário para a identificação das formas corretas das palavras.

Por fim, durante o levantamento dos recursos utilizados como base para a construção de um vocabulário multilíngue, identificou-se a necessidade desse tipo de recurso na construção automática de corpus, tarefa esta que pode ser empregada para ampliação do corpus existente.

3. Construção de Vocabulário Multilíngue

O processo de extração de vocabulário multilíngue, utilizado neste trabalho, encontra-se demonstrado, de forma simplificada, na Figura 1. Esse consiste em, dado um corpus multilíngue, realizar a extração do vocabulário equivalente aos idiomas envolvidos e,

posteriormente, avaliá-lo [Ha et al. 2008] [Daille and Morin 2005]. Uma descrição mais detalhada dessas etapas será apresentada a seguir.

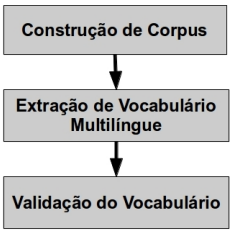

Figura 1. Processo genérico de extração

3.1. Construção do Corpus

Para a construção do corpus multilíngue foram considerados: a) textos paralelos (textos acompanhados por suas respectivas traduções [Ha et al. 2008]) e b)textos comparáveis (textos em diferentes idiomas que compartilham características comuns como tema, por exemplo [Daille and Morin 2005]).

Como fonte principal de material, optou-se pela utilização da documentação de *softwares open source* (código aberto), devido a sua disponibilidade para diferentes idiomas e ao tipo de licenciamento utilizado por esses. Posteriormente, foram incluídos à lista livros da Engenharia de *Software*.

Os materiais anteriormente mencionados foram coletados a partir das páginas oficiais dos projetos. Versões multilíngues do manual de usuário da ferramenta de versionamento TurtoiseSVN, por exemplo, foram extraídos a partir do sítio *"http://tortoisesvn.net/support.html"*. Como uma das colaborações deste trabalho, o *corpus* construído, relacionado a projetos *open source*, será futuramente disponibilizado em um repositório.

Conhecidas as fontes, a coleta dos documentos foi conduzida de forma manual. Abordagens automáticas para a construção do corpus foram cogitadas, no entanto, devido à falta de termos iniciais (*seeds*) e a complexidade associada a avaliação dos textos coletados por essas, optou-se por não utilizá-las (pelo menos inicialmente).

3.2. Extração de Vocabulário Multilíngue

O processo de extração de vocabulário multilíngue consiste na obtenção de equivalências entre palavras de diferentes idiomas. A execução desse varia de acordo com o tipo de corpus empregado (paralelo ou comparável).

Dado o trecho de sentença "...será a nova versão do *software*...", por exemplo, e seu trecho equivalente nos documentos em inglês "*...will be the new release of the software*", o método de extração deve estabelecer a relação entre as palavras *"release"* e *"versão"*.

Em relação a corpus paralelo, um dos procedimentos mais utilizados é o alinhamento textual, sentencial e/ou lexical, sendo que esse consiste na identificação de trechos correspondentes entre textos considerados como paralelos (texto e sua respectiva tradução)[Ha et al. 2008].

Quanto a corpus comparável, o estabelecimento de equivalências costuma ser realizado mediante a utilização de vetores de contexto (que levam em consideração as palavras próximas ao termo a ser traduzido) em conjunto com dicionários multilíngues de domínio geral [Daille and Morin 2005].

O processo de extração de vocabulário empregado até momento, baseia-se na extração de *n*gramas, e é realizado de acordo com os seguintes passos:

1. Os documentos extraídos são normalizados quanto a seu formato, sendo convertidos para documentos de texto sem formatação (*plain text*);
2. Os textos são separados em sentenças (*Sentence detection*) que, por sua vez, são separadas em seus símbolos formadores (palavras, sinais de pontuação, etc.);
3. Numerais, sinais de pontuação, símbolos especiais (marcadores, por exemplo) e palavras muito comuns do idioma (*stopwords*) são removidos;
4. As palavras restantes são submetidas a um processo de lematização para a obtenção de suas respectivas formas canônicas;
5. Listas de *n*gramas (sequências *n* palavras) são construídas e posteriormente contabilizadas. Neste trabalho foram considerados unigramas, bigramas e trigramas;

Com a exceção da lematização dos textos em português (conduzida com o lematizador da ferramenta CoGroo [Kinoshita et al. 2007]), as demais etapas foram realizadas com o conjunto de ferramentas disponibilizadas pelo NLTK [Bird et al. 2009].

3.3. Avaliação do Vocabulário

A avaliação do vocabulário construído pode ser realizada tanto de forma manual quanto automática. A validação manual consiste na revisão do vocabulário por um tradutor profissional ou por um especialista da área, enquanto na validação automática o vocabulário construído é comparado com um padrão de referência (*golden standard*) previamente criado [Daille and Morin 2005] [Ha et al. 2008].

4. Resultados Parciais

Até o momento, como recurso selecionado para pesquisa, tem-se um corpus multilíngue composto por 567.458 palavras (unigramas) em inglês e 331.626 palavras em português.

Exemplos de palavras (unigramas) extraídas a partir dos textos em português do corpus, mediante o processo apresentado na seção 3.2, são: “Tela”, “Contato”, “lista”, “agenda”, *“wi-fi”*, *“bluetooth”*, “calculadora” e *“sms”*.

Nesses, são encontrados tanto termos de domínio como *“Contato”* (dispositivos móveis), quanto termos que designam tecnologias como *“bluetooth”*, por exemplo. A lista de palavras (assim como a de bigramas e trigramas) será melhor investigada em busca de palavras e termos comuns a diferentes domínios, com enfoque principal na terminologia de Engenharia de *Software*.

Quanto aos materiais auxiliares, buscou-se por vocabulários já compilados, relevantes ao domínio. A Tabela 1 demonstra os principais vocabulários encontrados juntamente com a quantidade de termos contidos em cada um desses. Vale ressaltar que apesar desses serem monolíngues (inglês), são compostos por termos diretamente relacionados ao domínio, sendo que traduções para os mesmos serão buscadas.

Vocabulário	Termos
System and Software Engineering–Vocabulary [ISO/IEC/IEEE 2010]	3.349
Standard Glossay of Software Engineering Terminology [IEEE 1990]	1.300
Glossário "*Software Engineering*" [Sommerville 2010]	167
Lista de Assuntos "*Software Engineering*" [Sommerville 2010]	1.600

Tabela 1. Vocabulários obtidos

Por fim, foram obtidos registros (*logs*) de comunicação entre equipes de desenvolvedores. A partir dos registros do experimento de Calefato *et al.* [Calefato et al. 2012] foram extraídas mensagens em português (449), italiano (694) e inglês (874). Posteriormente, foram obtidos registros de comunicação entre desenvolvedores da fundação *Mozilla* compostos por 161.316 mensagens (aproximadamente 1.669.000 palavras).

4.1. Aplicabilidade dos Recursos

Uma análise inicial dos recursos levantados demonstrou a aplicabilidade desses na solução de problemas identificados durante a análise dos registros (seção 2).

Em relação aos problemas de tradução inconsistente, partindo do termo *"release"*, previamente apresentado, buscou-se nos textos em inglês do corpus por ocorrências desse. Dentre os contextos nos quais o termo foi encontrado, 6 foram selecionados, sendo que seus trechos equivalentes foram buscados nos textos em português. Para os 6 contextos avaliados o termo foi coerentemente traduzido como *"versão"*, indicando uma tendência na utilização desta tradução no domínio em questão.

Referente aos problemas de abreviação e erros ortográficos, para ambos os exemplos apresentados (*"bluetooth"* e *"ring tone"*), os termos foram encontrados no material compilado, indicando que estes poderiam ser utilizados em funções de correção ortográfica (erros ortográficos) e auto-complementação (abreviação), ambos recursos de auxílio a escrita.

Mais exemplos de aplicabilidade do material compilado na solução dos problemas mencionados serão obtidos durante a execução das próximas etapas do processo de extração.

O vocabulário bilíngue extraído pode vir a ser empregado, ainda, na tradução de ontologias existentes na área de Engenharia de *Software*. Conhecida a natureza multilíngue do desenvolvimento global de *software*, torna-se importante que estas ontologias encontrem-se disponíveis em mais de um idioma.

5. Conclusões

A utilização de serviços de tradução simultânea de máquina, considerada como uma solução alternativa ao inglês durante reuniões de equipes distribuídas, tem seu desempenho comprometido devido à problemas de tradução de máquina, dentre os quais destacam-se os apresentados na seção 2.

Entre os diferentes tipos de problemas observados, optou-se por priorizar aqueles relacionados ao vocabulário, para os quais a solução proposta consiste na construção de um vocabulário multilíngue das práticas usuais do processo de desenvolvimento de *software*.

No entanto, como apresentado na seção 3.2, a construção de um vocabulário multilíngue depende da existência de um corpus multilíngue previamente compilado. Neste trabalho, identificamos os recursos disponíveis para a construção deste vocabulário.

Em relação à primeira etapa deste trabalho, o principal resultado obtido foi um conjunto de materiais formado por um corpus multilíngue, composto por manuais de *software*, conjuntos de vocabulários referentes ao domínio, e registros (*logs*) de comunicação entre desenvolvedores.

Uma vez o corpus compilado, as próximas etapas consistem na aplicação de métodos de extração de vocabulário multilíngue (baseados em *corpus* paralelo), sobre o *corpus* construído, seguida da avaliação manual do vocabulário extraído.

Referências

Bird, S., Loper, E., and E., K. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

Calefato, F., Lanubile, F., Conte, T., and Prikladnicki, R. (2012). Assessing the impact of real-time machine translation on requirements meetings: A replicated experiment. In *6th Int'l Symposium on Empirical Software Engineering and Measurement (ESEM'12) (to appear)*, page 19–20.

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.

Daille, B. and Morin, E. (2005). French-english terminology extraction from comparable corpora. In Dale, R., Wong, K.-F., Su, J., and Kwong, O. Y., editors, *IJCNLP*, volume 3651 of *Lecture Notes in Computer Science*, pages 707–718. Springer.

Ha, L. A., Fernandez, G., Mitkov, R., and Pastor, G. C. (2008). Mutual bilingual terminology extraction. In *LREC*. European Language Resources Association.

IEEE (1990). Ieee standard glossary of software engineering terminology std 610.12-1990.

ISO/IEC/IEEE (2010). Systems and software engineering – vocabulary.

Kinoshita, J., Salvador, L. N., and Menezes, C. E. D. (2007). Cogroo - an openoffice grammar checker. In *Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications*, ISDA '07, pages 525–530, Washington, DC, USA. IEEE Computer Society.

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.

Sommerville, I. (2010). *Software Engineering*. Addison-Wesley, Harlow, England, 9. edition.

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.