Building Domain Specific Bilingual Dictionaries

L. Hilgert, L. Lopes, A. Freitas, R. Vieira, D. N. Hogetop, A. A. Vanin
PUCRS University – Porto Alegre – Brazil
lucashilgert@gmail.com, lucelene.lopes@pucrs.br, artur.freitas@acad.pucrs.br
renata.vieira@pucrs.br, dhogetop@terra.com.br, aline.vanin@ymail.com

Abstract

This paper proposes a method to build bilingual dictionaries for specific domains defined by a parallel corpora. The proposed method is based on an original method that is not domain specific. Both the original and the proposed methods are constructed with previously available natural language processing tools. Therefore, this paper contribution resides in the choice and parametrization of the chosen tools. To illustrate the proposed method benefits we conduct an experiment over technical manuals in English and Portuguese. The results of our proposed method were analyzed by human specialists and our results indicates significant increases in precision for unigrams and muli-grams. Numerically, the precision increase is as big as 15% according to our evaluation.

Keywords: Automatic dictionary construction, Term extraction and alignment, Domain specific information extraction

1. Introduction

The availability of domain specific bilingual vocabulary is a valuable and rare resource. There are some domains whose existence of parallel corpora is frequent, as every written material available in many languages, *e.g.*, multilingual manuals and product specifications. Considering that, it is interesting to develop a method to automatically extract bilingual vocabulary from such resources. The applicability of domain specific bilingual dictionaries for translation purposes is clear (Zhang, 2009).

This paper proposes a method to automatically extract bilingual vocabulary from parallel corpora. From a practical point of view, the proposed method is an extension of a pre-existent method based on the successive application of natural language processing tools available. Such method is exemplified by the construction and evaluation of an English-Portuguese bilingual vocabulary present in software manuals. Therefore, the proposed method combines some common steps of bilingual extraction processes (Ha et al., 2008) with terminology extraction tools (Lopes et al., 2009).

The result of this experiment was manually evaluated to illustrate the effectiveness of the proposed method. The evaluation was conducted with three human judges to estimate whether a pair of terms (one in English, one in Portuguese) were not only correctly related, *i.e.*, they are a correct translation, but also if the terms are relevant to the domain.

This paper is organized as follows: The next section describes the original method and the natural processing language tools employed. Section 3 describes the proposed method with special emphasis on the newly included step responsible to perform the choice of domain relevant terms to the outputted vocabulary. Section 4 describes the experiments with both the original and proposed methods to illustrate the benefits of our approach over a practical case. Finally, the conclusion summarizes the paper contribution and suggest future works.

2. Original Bilingual Vocabulary Extraction Method

The proposed method is based on a generic process proposed by Caseli (Caseli, 2007) and instantiated by Hilgert (Hilgert, 2013). The original method steps are described in Fig. 1. This process starts by a bilingual corpus and it is composed of four steps that are performed until the generation of a bilingual vocabulary, *i.e.*, a dictionary.

The first step of this process is called Sentence Alignment, and it is responsible for establishing equivalences between sentences of the parallel texts, since there is not always an one-to-one equivalence between parallel texts in different languages (Tiedemann, 2003). A possible tool to perform this step is the Bilingual Sentence Aligner (Moore, 2002), which is used in the experiment described in this paper.

The second step, called Morphological Analysis, consists in the assignment of morphological information, *e.g.*, part-of-speech (POS), number and gender, to the words of the corpus in order to enable lexical disambiguation. For the experiments in this paper this step is performed by the Lttoolbox, one of the components of the Apertium translation platform (Forcada et al., 2011). This specific tool performs the identification of sentences elements, *i.e.*, words and multiwords expressions, as well as a complete pos-tagging (attribution of word class, number and gender) for these elements. The generic dictionaries of the Lttoolbox are enhanced by the inclusion of lists of specific terms from the input corpora. The Portuguese corpus term list is generated using ExATOlp extraction tool (Lopes et al., 2009), and the English corpus term list is generated using the term extraction tool of the TTC Termsuite (Rocheteau and Daille, 2011).

The third step is called Lexical Alignment and it is the step responsible to identify the equivalences between single and multi word expressions in parallel sentences previously aligned. In this paper experiments this step is performed using Giza++ tool (Och and Ney, 2003).

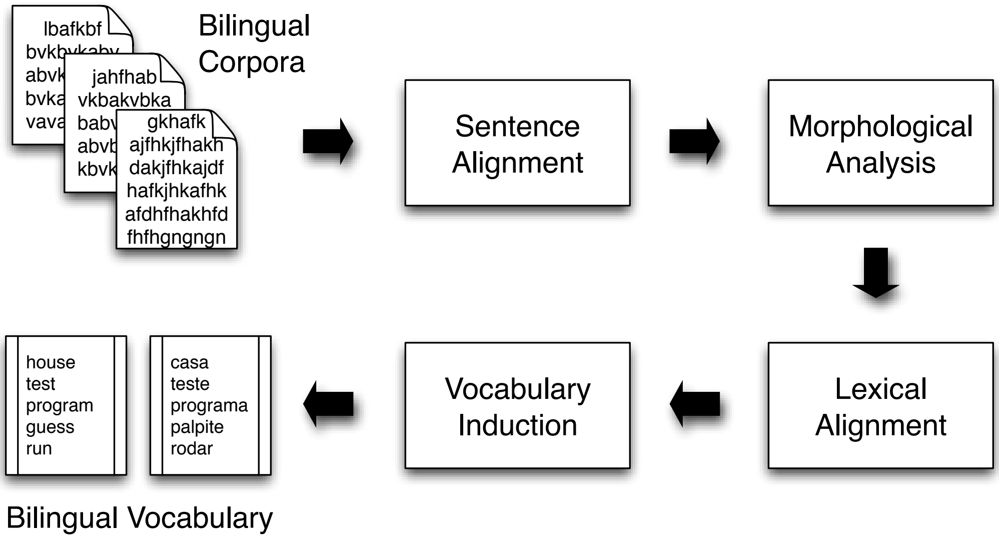

Figure 1: Original method steps.

A particularity of this step implementation is a two-way alignment, *i.e.*, the alignment from Portuguese to English, and also from English to Portuguese, followed by a union algorithm to resolve possible conflicts between these two processes.

The fourth step is responsible to select the more relevant entries for the vocabulary. For this paper experiments, the Vocabulary Induction step was implemented by the ReTraTos tool (Caseli, 2007). The selection in this tool takes into account frequency of occurrence and co-reference metrics to chose the bilingual vocabulary entries to be kept.

3. Proposed Method

Considering the original method described, the proposed method consists in the addition of a fifth step as shown in Fig. 2. This fifth step, called Domain Filtering, aims to reduce the number of selected entries to keep only those that are specific to the domain.

The Domain Filtering step starts with the identification of specific terms for the domain represented by the input corpora. In this paper experiments this step is performed with the ExATOlp extraction tool (Lopes, 2012), a term extractor for Portuguese corpora that is capable to generate a list of relevant terms of a domain specific corpus using linguistic and statistical approaches.

ExATOlp software tool (Lopes et al., 2009) thus select domain significant terms from an annotated domain *corpus*. From a linguistic point of view, the extraction is based on the syntactic annotation performed by the parser PALAVRAS (Bick, 2000). The candidate terms are terms annotated as noun phrases, subjects or objects by the parser according to an extra set of discard and transformation rules (Lopes and Vieira, 2012). These transformation rules include:

- Adjustment rules that considers terms in the canonical form with all determinants (articles and verbs) removed;
  - Ex.: "*the starts of the movie*" becomes "*star of movie*"
- Discard rules that ignores terms with numerals and symbols, but also terms with an inadequate head (pronouns or adverbs);
  - Ex.: "*Code 46*" is discarded,
  - Ex.: "*my dream*" are discarded
- Inclusion rules that considers non explicit mentioned terms as implicit terms by use of conjunctions, adjectives removal and multiple predicates.
  - Ex.: "*good and bad boys*" becomes "*good boy*" and "*bad boy*",
  - Ex.: "*good old-fashioned boy*" becomes also "*old-fashioned boy*" and "*boy*"

From a statistical point of view, those candidate terms are subject to frequency analysis, *i.e.*, in order to select the more frequent ones. However, the frequency computation is not a trivial one, but it considers *tf-dcf* index (Lopes et al., 2012) based on the use of contrasting corpora. Once this relevance index is computed, the more relevant terms are considered assuming cut-off points to keep approximatively 15% of extracted terms, since previous works indicate that this amount delivers a good balance between precision and recall (Lopes et al., 2010).

Once the list of relevant terms of the domain is known, the bilingual vocabulary selected by the Vocabulary Induction step are compared to this list and only the entries with relevant terms are kept. In other words, selected entries with terms that are not presented in the ExATOlp term list are discarded. It is important to call

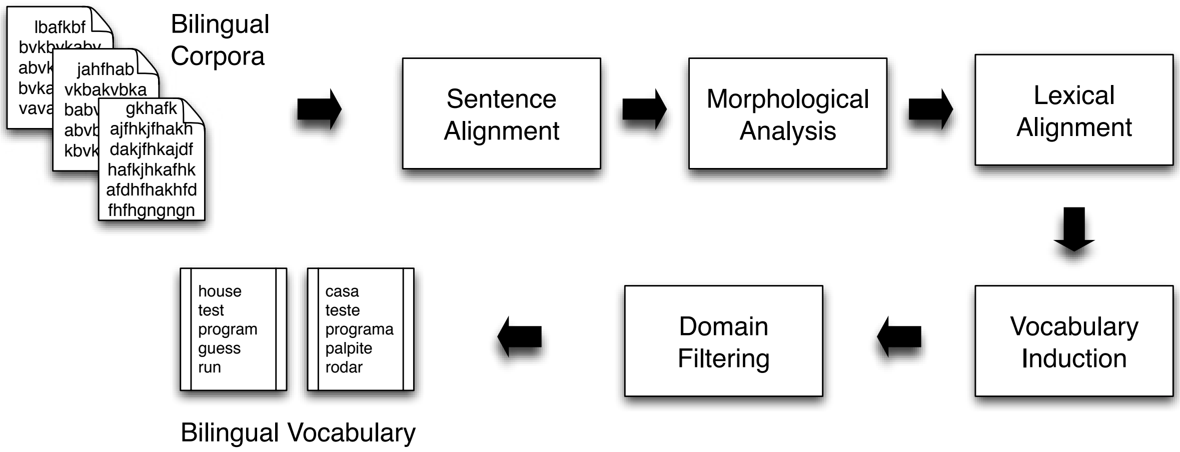

Figure 2: Proposed method steps.

attention to the fact that the filtering step is performed only through the comparison of the Portuguese relevant terms.

The final result of this fifth step is, therefore, a bilingual vocabulary focused on the corpora domain, since it discards the entries that may not be relevant. As a consequence, the resulting bilingual vocabulary tends to present a better precision than the bilingual vocabulary produced by the original process. In fact, this assumption is verified through a practical example described in the next section.

4. Experiments with a Software Manual Corpora

To illustrate the proposed process we conduct an experiment over bilingual corpora composed by 8 pairs of software manuals (8 texts in English, and 8 equivalent texts in Portuguese). The English corpus totalized 481,068 words, distributed over 25,734 sentences, and the Portuguese corpus totalized 491,718 words, distributed in 24,946 sentences.

The bilingual corpora was submitted to the original process as described in Fig. 1. After the first step, Sentence Alignment, the sentences in both corpora that could not be associated to a counterpart in the other language were discarded. As a result, both English and Portuguese corpora were reduced to 21,818 sentences, with 392,804 English words and 417,381 Portuguese words, respectively.

The Morphological Analysis, Lexical Alignment, and Vocabulary Induction steps were also applied as described in the previous section and, as a result, 18,268 bilingual entries were selected.

Applying the fifth step, Domain Filtering, ExATOlp delivered a list with 2,793 relevant terms for the domain. Comparing these relevant terms with the selected entries only 1,041 were kept. It is important to call the reader attention that this represents a significant reduction to less than 6%, *i.e.*, the last step of the proposed method (Domain Filtering) reduced the 18,268 possibly generic bilingual entries to 1,041 specific ones.

Tab. 1 summarizes the results splitting the total number of terms/entries in single words (unigrams) and multiwords (multi-grams) according to the Portuguese version. Note that the last row of this table indicates the size of the bilingual vocabulary produced after the Domain Filtering step, *i.e.*, the result of the proposed method. In opposition, the first row indicates the size of bilingual vocabulary at the end of the original method.

	unigrams	multi-grams	total
Selected entries after Vocabulary Induction step	4,788	13,480	18,268
Relevant terms according to ExATOlp	398	2,395	2,793
Intersection between selected entries and relevant terms	268	773	1,041

Table 1: Number of entries and terms found in the experiment.

To illustrate the terms analyzed during the Domain Filtering step, Table 2 presents some randomly chosen terms outputted by the Vocabulary Induction step. To each term of Table 2 presents the term extracted from the Portuguese and English corpora in the first and second columns, respectively. In the third column there is an indication of its correctness according to the judges opinion (column *correct*) and in the fourth column there is an indication of its pertinence to the

from the Portuguese corpus	from the English corpus	*correct*	*domain*
aba	**tab**	√	∈
abrir	open	√	
abrir	request	**open**	
alocação	lease	**allocation**	
avião	airplane	√	
clique	**click**	√	∈
configuração manual	**manual setup**	√	∈
favorita pasta	hidden folder	**favorite folder**	
fechado	todos	**closed**	
ferramenta de preenchimento	**fill tool**	√	∈
fixo	dialing number	**fixed**	
fonte desconhecida	unknown source	√	
gráfico vetorial	vectorial graphics	√	
imagem anterior	next image	**previous image**	
infelizmente	unfortunately	√	
lista de agrupamentos	**grouping list**	√	∈
lista de reprodução	playlist	√	
máquina sem atraso	restart	**machine without delay**	
marca verde	**green check mark**	√	∈
miniaplicativo indicador de mensagem	turn green	**message indicator applet**	
navegador firefox	**firefox web browser**	√	∈
navegador mozilla	mozilla browser	√	
preencher	enter	**fill**	
primeiro dvd	dvd	**first dvd**	∈
primeiro nome	last name	**first name**	
segurar	hold	√	
senha	**password**	√	∈
unidade de medida	measurement	**measurement unit**	
usuário doméstico normal	normal domestic user	√	

Table 2: Sample of entries outputted by the Vocabulary Induction step.

domain according to the ExATOlp output.

If the term in English is the correct translation of the Portuguese one, column *correct* presents the √ symbol, otherwise it has the correct English translation. Column *domain* is marked with ∈ symbol if the term in Portuguese was outputted as relevant by ExATOlp tool. In fact, entries kept after the Domain Filtering step, *i.e.*, entries outputted as correct by the proposed method are indicated in bold for both the Portuguese and English versions.

For instance, the Portuguese term “clique” was correctly associated to the English term “click”, and it is a relevant term to the software manual domain. Entries like this one are considered true positives.

The Portuguese term “infelizmente” was correctly translated to English as “unfortunately”, but this is not a relevant term to the domain. Entries like this one were discarded by the Domain Filtering step, and they can be considered as true negatives.

Another possible situation is the Portuguese term “imagem anterior” which was incorrectly translated, since its correct English translation is “previous image”. However, since the Portuguese term was not considered relevant by ExATOlp output, entries like this one were also discarded by the Domain Filtering step, and they can also be considered as true negatives.

A problem of the method output can be noticed in the entry “navigator mozilla”, which was correctly translated to “mozilla browser”, but it was not indicated as relevant by the ExATOlp tool. Therefore this entry was discarded by the Domain Filtering step. Despite of that, this entry is likely to be as relevant as the entry “navigator firefox” which is also correctly translated to “firefox web browser” and was considered relevant by ExATOlp. Possibly, it was caused because the “firefox” entry was frequent enough to be considered relevant, unlike the “mozilla” entry. Nevertheless, entries like this one can be considered as false negatives.

Finally, there was another situation where the Vocabulary Induction step delivered a wrong translation, but the ExATOlp output has validated the entry. An example of such situation in Table2 is the entry “primeiro dvd” which was associated with “dvd”, but the correct translation was “first dvd”. Situations like this one, called false positives, were found in 265 of the 1,041 kept entries.

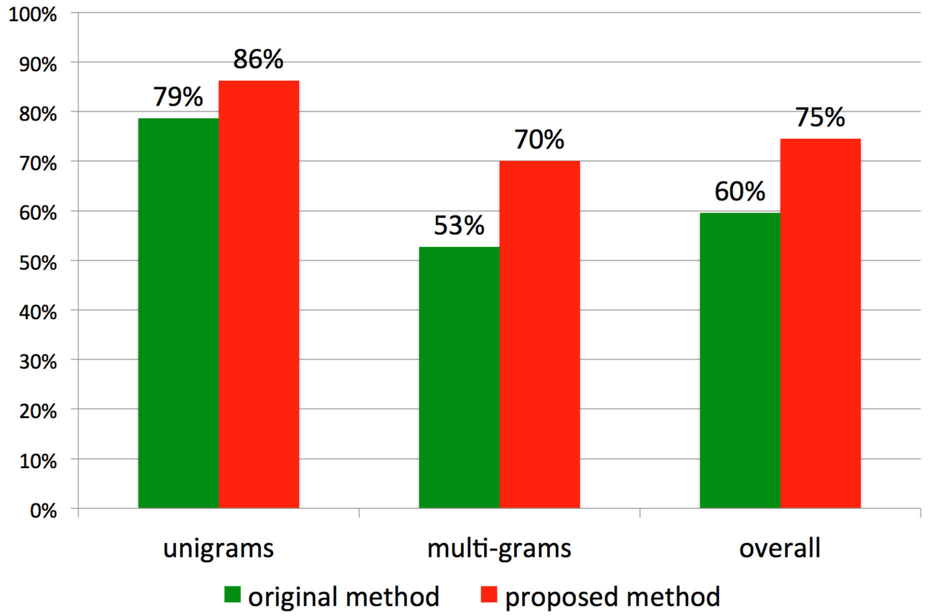

Figure 3: Precision achieved by each method.

4.1. Quality Evaluation

The main objective for the original and the proposed methods was the extraction of bilingual vocabularies. Unfortunately, there were no golden standards, *e.g.*, reference lists, bilingual dictionaries, to verify the precision of the result. Hence, an intrinsic evaluation (words evaluated without context, *i.e.*, which were not used in sentences) was conducted by three human judges, who considered the correctness of the assigned bilingual equivalent entries.

For the original method, the result of 18,268 bilingual entries was too large to be fully analyzed. Thus, 750 entries, chosen randomly, were submitted to the judges that considered a term correct if the judges were unanimous to admit it as a correct translation. Additionally, the judges estimate the relevance of the translated terms for the software manual domain.

Specifically, 600 unigrams and 150 multi-grams were manually analyzed. From those, 473 unigrams entries out of the sample set of 600 were considered correct, and 79 out of the 150 multi-grams set were considered correct. Those numbers indicate as precision:

$$\textit{precision for unigrams} = \frac{473}{600} = 78.83\%$$

$$\textit{precision for multi-grams} = \frac{79}{150} = 52.67\%$$

Assuming that these percentages of correct entries remain the same for the rest of entries that were outputted in the original process, the precision of the original process may be estimated.

Considering the number of 4,788 unigrams and 13,480 multi-grams (see the first row of Tab. 1), and the percentage of correct entries for the 750 sampled entries (600 unigrams and 150 multi-grams), the estimated number of correct entries for unigrams and multi-grams will be:

$$4,788 \;\; \textit{unigrams} \times 78.83\% = 3,774 \;\; \textit{unigrams}$$

$$13,480 \;\; \textit{multi-grams} \times 52.67\% = 7,100 \;\; \textit{multi-grams}$$

Therefore, it is possible to estimate that 10,874 terms (3,774 plus 7,100) were correct out of 18,268 (4,788 plus 13,480). This estimated number of correct terms corresponds to an overall precision of:

$$\textit{original method precision} = \frac{10,874}{18,268} = 59.52\%$$

For the proposed process output, on the contrary, all 1,041 resulting entries (see the last row of Tab. 1) were analyzed in the same way as the 750 sampled entries of the original process output. This analysis resulted in 231 correct entries out of 268 unigrams and in 545 correct entries out of 773 multi-grams:

$$\textit{precision for unigrams} = \frac{231}{268} = 86.19\%$$

$$\textit{precision for multi-grams} = \frac{545}{773} = 70.05\%$$

Thus, the overall precision of the proposed method was of:

$$\textit{original method precision} = \frac{776}{1,041} = 74.54\%$$

The numerical comparison of precision achieved by each method is depicted in Fig. 3.

5. Conclusion

These experiments have shown the benefits brought by the proposed method over the original one (15% overall precision increase). For multi-grams particularly, the precision was increased from 53% to 70%. For unigrams as well, the precision improvement from 79% to 86% was non-negligible.

These numbers let us be encouraged by the quality of the proposed method. Since it is not necessarily true that a reduction of the number of terms would increase the precision. In fact, this is only true when the procedure to discard terms is somehow correlated with the quality of the terms. In other words, the choice of relevant terms to the domain correspond to choose the terms that are correctly translated.

For instance, a similar experiment conducted by Caseli (Caseli, 2007) delivered precision values of 86% for unigrams and only 38% for multi-grams. Even though Caseli's experiment was conducted over different corpora, many of the tools employed in each step were the same as the ones used in our experiment.

Therefore, our achieved precision of 86% for unigrams, and specially the precision of 70% for multigrams, indicates a clear advantage of our method. As previously mentioned, this also indicates that our choice of consider relevant terms to the domain was correct.

The bilingual vocabularies generated by the original and the proposed methods, as well as the term lists generated for the Domain Filtering step are available in electronic format at:

`http://www.inf.pucrs.br/~linatural/`

6. Acknowledgments

The authors wish to thank the research support received, namely: project PRONEM/FAPERGS/CNPq no. 003/2011 funding Lucas Hilgert and Renata Vieira; project DOCFIX/FAPERGS/CAPES no. 09/12 funding Lucelene Lopes; project PNPD/CAPES no. 388/09-0 funding Denise N. Hogetop and Aline A. Vanin.

7. References

Bick, E. (2000). *The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework*. Ph.D. thesis, Arhus University, Arhus, Danemark.

Caseli, H. (2007). *Indução de léxicos bilíngues e regras para a tradução automática*. Ph.D. thesis, ICMC-USP, São Paulo, Brazil.

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. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, June.

Ha, L., Fernandez, G., Mitkov, R., and Corpas, G. (2008). Mutual bilingual terminology extraction. In et al., N. C., editor, *Proceedings of the Sixth Int. Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. ELRA.

Hilgert, L. W. (2013). Extração de vocabulário multilíngue a partir de documentação de software. Master's thesis, PUCRS.

Lopes, L. and Vieira, R. (2012). Heuristics to improve ontology term extraction. In *PROPOR 2012 – International Conference on Computational Processing of Portuguese Language*, LNCS vol. 7243, pages 85–92.

Lopes, L., Fernandes, P., Vieira, R., and Fedrizzi, G. (2009). ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora. In *Proceedings of the 4th Language & Technology Conference (LTC '09)*, pages 427–431, Poznan, Poland. Faculty of Mathematics and Computer Science, Adam Mickiewicz University.

Lopes, L., Vieira, R., Finatto, M. J., and Martins, D. (2010). Extracting compound terms from domain corpora. *Journal of the Brazilian Computer Society*, 16:247–259. 10.1007/s13173-010-0020-4.

Lopes, L., Fernandes, P., and Vieira, R. (2012). Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)*, pages 1001–1007, Las Vegas, USA. CSREA Press.

Lopes, L. (2012). *Extração automática de conceitos a partir de textos em língua portuguesa*. Ph.D. thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil.

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.

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

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. (2003). *Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Uppsala University, Uppsala, Sweden.

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*, pages 271–274, Washington, DC, USA. IEEE Computer Society.