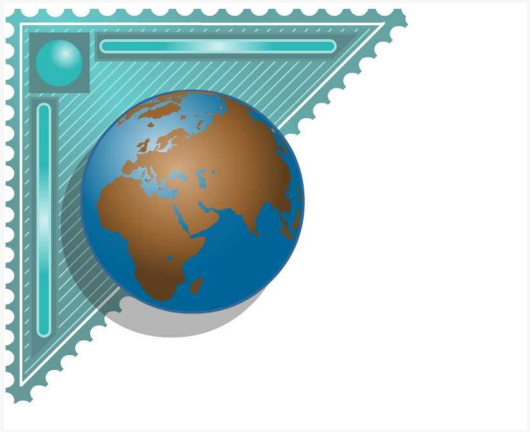
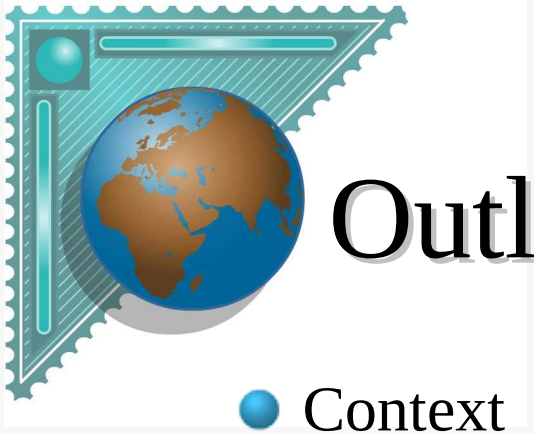


Multilingual Vocabulary Extraction from Software Documentation

Lucas Welter Hilgert



Outline

- Context
- Problems
- Research Work
- Related Work
- Corpus Construction
- Vocabulary Extraction
- Bibliography

Context

- **Project:** "*The Effect of Natural Language Processing in the Development of Brazil Competence in the Global Software Development Market*"
  - Investigation of NLP methods, techniques and tools aiming to contribute in the inclusion of Brazilian development teams in the global software development market
  - Focus on Real Time Machine Translation services

Problems

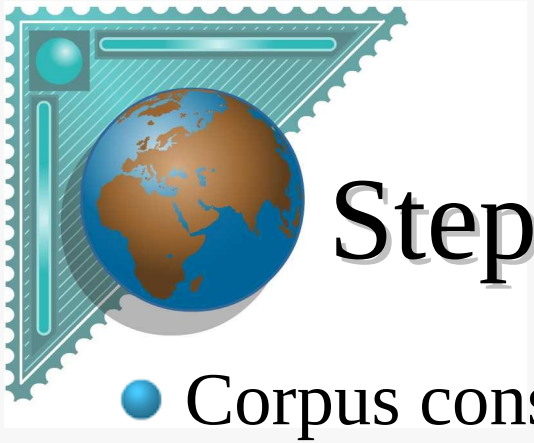
- Experiments involving Brazilian (PUCRS) and Italian (University of Bari) participants [Calefato *et al.* 2012]
  - Similar performance using English as a common language and Real Time Machine Translation during meetings [Calefato *et al.* 2012]
  - Comunication delays [Calefato *et al*., 2012][Yamashita and Ishida, 2006]
  - Difficulties establishing a common ground [Yamashita *et al.* 2009]

Purposed Work

- *"Multilingual vocabulary extraction from software documentation"*
  - Open source software user guides;
  - Parallel corpus (English – Portuguese);
  - Enrichment of machine translation services dictionaries (*Apertium*);
  - Utilization in writing aid technologies (auto-complete, spell-checking, etc.);

Related Work

- Lexicon Induction [Caseli and Nunes, 2007]
- OPUS Project [Tiedemann, 2009]
- Uplug [Tiedeman, 2003]
- Bilingual Terminology Extraction [Ha *et al.*, 2008]



Steps

- Corpus construction
- Vocabulary extraction
- Evaluation

Corpus Construction

- **Source:** *open source* project repositories

User Guide	Portuguese	English
Android 2.3.4	72.371	81.412
Blender 2.6	31.661	25.671
Debian	149.709	140.367
LibreOffice 3.3	100.978	108.450
Slackbook	65.549	64.744
TortoiseSVN 1.7	79.008	79.696
Ubuntu	54.057	55.687
Total	**553.333**	**556.027**

Corpus Construction

- Parallel corpus
- Manually constructed
- Corpus characteristics:
  - Noisy corpora
  - Cooperatively translated

Bilingual Vocabulary Extraction

- Corpus preprocessing:
  - Conversion to plain text (pdf, html);
- Sentence alignment
- Morphological analysis
- Lexical alignment

Sentence Alignment

- Tools:
  - Bilingual Sentence Aligner [Moore, 2002]
  - TCAaligner [Caseli, 2003]
  - Uplug [Tiedeman, 2003]

Morphological Analysis

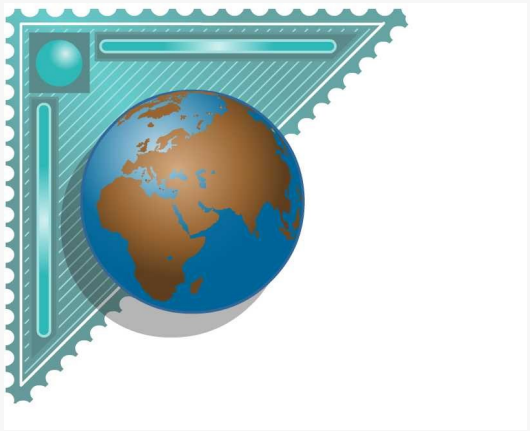
- Information:
  - Part-of-speech
  - Lemma
  - Inflections (gender, number,...)
- Tools:
  - Lttoolbox (*Apertium*) [Forcada *et al.*, 2011]
  - FreeLing 3.0
  - LX-Center

Lexical Alignment

- Statistical approach [Tiedemann, 2003]
  - Giza++ (version 2.0) [Och & Ney, 2003]
  - Trained using the whole corpus
  - Standard configuration
- Aligned in both directions (English → Portuguese, Portuguese → English)
- Union algorithm applied [Och & Ney, 2003]
- ReTraTos [Caseli & Nunes, 2003]

Vocabulary Extraction

- Evaluation:
  - Comparation with term lists:
    - ExATOlp [Lopes *et al.*, 2012] - pt
    - TTC Term Suite - en
  - Intrinsic



Questions?

Bibliography

Calefato, F.; Lanubile, F.; Conte, T.; Prikladnicki R. "**Assessing the Impact of Real-Time Machine Translation on Requirements Meetings: A Replicated Experiment**".6th Int'l Symposium on Empirical Software Engineering and Measurement (ESEM'12), Lund, Sweden, Sept. 19–20, 2012 (to appear). 2012.

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

Caseli, Helena M. "**Indução de léxicos bilíngues e regras para a tradução automática**". PhD thesis, ICMCUSP, São Paulo, Brazil, 2007.

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

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

Lopes, L. ; Vieira, R. ; Fernandes, P. ; Couto, G. . "**ExATOlp: extraction of language resources from Portuguese corpora**". In: International Conference on Computational Processing of the Portuguese Language - PROPOR, 2012, Coimbra. Proceedings of PROPOR 2012, 2012.

Bibliography

Moore Robert C.. "**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, 2002.Springer-Verlag.

Och Franz Josef and Ney Hermann . "**A systematic comparison of various statistical alignment Models".** In Comput. Linguist, *29(1):19–51*, March *2003.*

*Yamashita N. et al. "**Difficulties in establishing common ground in multiparty groups using machine translation**". Proceedings of the 27th international conference on Human factors in computing systems. 2009.*

Yamashita, N.; Ishida. T. "**Effects of machine translation on collaborative work**". Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work CSCW 06. ACM Press. 2006.

Tiedemann Jörg . "**Parallel data, tools and interfaces in opus"**. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1246. 2012.

Tiedemann Jörg. "**Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing"**. PhD thesis, Uppsala University, Uppsala, Sweden, 2003.