Collection and Preparation of Multilingual Data for Multiple Corpus-based Approaches to Translations

Hernani Costa
hercos@uma.es

LEXYTRAD, University of Malaga
Malaga, Spain

Background

- Literature review on
  - Computational linguistics
  - Concepts of corpus and typology
  - Techniques and applications to semi-automatically compile corpora
  - Amongst other related topics
- Current limitations of corpora compilation tools
  - compilation tools are scarce or proprietary
  - simplistic with limited features
  - built to compile one monolingual corpus at a time
  - do not cover the entire compilation process

Research Goals

- Manually and semi-automatically exploit multilingual textual resources to compile parallel and multilingual comparable corpora
- Design tools that
  - fulfil not only translators and interpreters' needs, but also professionals and ordinary people
  - cover the entire compilation process, i.e. capable of compiling, managing and exploring both parallel and comparable corpora
  - are publicly available for being used by anyone, both in a research or in a commercial setting

Semi-automatically Compile Multilingual Corpora

Compilation

- Monolingual corpora independently in multiple languages
- Multilingual comparable corpora through CLIR techniques
- Parallel corpora

Management

- Give a similarity coefficient to the documents in a corpus
- Analyse the representativeness of a corpus
- Manage comparable and parallel corpora

Exploration

- Automatic extraction of terminology
- Manage terminology
- Concordance

Published Work

Technology-assisted Interpreting [1]

- Offers a tentative catalogue of current language technologies for interpreters
  - terminology tools for interpreters
  - note-taking apps for consecutive interpreting
  - apps for voice recording
  - training tools

A comparative User Evaluation of Terminology Management Tools for Interpreters [2]

- Reviews several terminology management tools
- Summarises the interpreters' most required features
- Proposes a set of specific and measurable features to assess and distinguish these systems, which allowed us to
  - make a comparative analysis
  - highlight some of the features that interpreters can expect from these systems
  - help interpreters choose a specific tool for a given service
  - give hints to the designers of such systems

Currently Working On

- Literature review on similarity measures [3, 4, 5]
- Adaptation and implementation of several document-document similarity measures
  - Spearman's Rank Correlation Coefficient
  - χ^2
  - LSA
  - Cocitation
  - amongst others
- Evaluation of several approaches
  - statistical
  - statistical with linguistic

First Secondment - University of Wolverhampton

- Giving similarity scores to documents in a corpus
  - results analysis
  - methodology improvements
- ReCor
  - implementation of new features
  - conversion from standalone to web-based
- Semi-automatic compilation of comparable corpora tool
  - user interface requirements and design
  - server requirements and implementation
  - semantic CLIR method implementation

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been (partially) carried out in the framework of the research group LEXYTRAD.

References

[1] H. Costa, G. Corpas Pastor, and I. Durán Muñoz, "Technology-assisted Interpreting," *MultiLingual #143, April/May*, vol. 25, no. 3, pp. 27–32, 2014.

[2] H. Costa, G. Corpas Pastor, and I. Durán Muñoz, "A comparative User Evaluation of Terminology Management Tools for Interpreters," in *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, (Dublin, Ireland), p. 9, August 2014.

[3] A. Kilgarriff, "Comparing Corpora," *Int. Journal of Corpus Linguistics*, vol. 6, no. 1, pp. 97–133, 2001.

[4] S. Sharoff, "Measuring the Distance Between Comparable Corpora Between Languages," in *Building and Using Comparable Corpora* (S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung, eds.), pp. 113–130, Springer, 2013.

[5] R. Köhler, "Statistical Comparability: Methodological Caveats," in *Building and Using Comparable Corpora* (S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung, eds.), pp. 77–91, Springer, 2013.