The Name of the Game is Comparable Corpora

Ruslan Mitkov
Research Group in Computational Linguistics
Research Institute in Information and Language Processing
University of Wolverhampton
R.Mitkov@wlv.ac.uk

Abstract

Comparable corpora are the most versatile and valuable resource for multilingual Natural Language Processing. The speaker will argue that comparable corpora can support a wider range of applications than has been demonstrated so far in the state of the art. The talk will present completed and ongoing work conducted by the speaker and colleagues from his research group where comparable corpora are employed for different tasks including but not limited to the identification of cognates and false friends, validation of translation universals, language change and translation of multiword expressions.

Corpora have long been the preferred resource for a number of NLP applications and language users. They offer a reliable alternative to dictionaries and lexicographical resources which may offer only limited coverage. In the case of terminology, for instance, new terms are coined on a daily basis and dictionaries or other lexical resources, however up-to-date they are, cannot keep up with the rate of emergence of new terms. As a result, terminologists (or term extraction programs) seek to analyse the use and/or identify the translation of a specific term using corpora.

Ideally, parallel data would be the best resource both for multilingual NLP applications such as Machine Translation systems and for users such as translators, interpreters or language learners. However, parallel corpora or translation memories may not be available, they may be time-consuming to develop or difficult to acquire as they may be expensive or proprietary. An alternative and more promising approach would be to benefit from comparable corpora which are easier to compile for a specific purpose or task.

Comparable corpora, whether strictly comparable by definition or 'loosely' comparable, have already been used in applications such as Machine Translation (Rapp, Sharoff and Zeigenbaum 2016) and term extraction and have been used by translators (Corpas and Seghiri 2009). The good news is that comparable corpora can facilitate almost any multilingual application and can beneficial to almost any language user. The view of the speaker is that comparable corpora are the most versatile, valuable and practical resource for multilingual NLP. The invited talk at the BUCC workshop at LREC'2016 will show that comparable corpora can offer more in terms of value and can support a wider range of applications than has been demonstrated so far in the state of the art. The talk will present completed and ongoing work conducted by the speaker and his colleagues at the Research Group in Computational Linguistics at the University of Wolverhampton in the domain of comparable corpora.

The talk will start with a discussion of the notion of comparable corpora and issues related to their use and compilation, and will briefly outline work by the speaker and his colleagues on the methodology related to the extraction of comparable documents and the building of purpose-specific comparable corpora.

Next the work carried out by the author on the automatic identification of cognates and false friends using comparable data will be presented. This will be followed by the presentation of three novel approaches developed by the speaker which use comparable data but do not resort to any dictionaries or parallel corpora, together with extensive evaluations of their performance.

The speaker will then focus on the use of purpose-built comparable corpora and NLP methodology in a project whose objective was to test the validity of so-called translation universals. In particular, the experiments on validating the universals of simplification, convergence and transfer will be detailed.

Following from this study, the speaker will outline the work on the use of comparable corpora to track language change over time, in particular the recent changes in lexical density and lexical richness in two consecutive thirty-year time periods in British English (1931–1961 and 1961–1991) and in American English from the 1960s to the 1990s (1961–1992).

Finally, the speaker will share the latest results from his work with colleagues on the use of comparable corpora for extracting and translating multiword expressions. The methodology developed does not rely on any dictionaries or parallel corpora, nor does it use any (bilingual) grammars. The only information comes from comparable corpora, inexpensively compiled with the help of the ACCURAT toolkit (Su and Babych 2012a) where only documents above a specific threshold were considered for inclusion. The presentation will conclude with the results of an interesting experiment as part of this study which sought to establish whether large loosely comparable data would yield better results than smaller but strictly comparable corpora.

Bibliographical References

Corpas, G. 2008. Investigar con corpus en traducción: los retos de un nuevo paradigma. Frankfurt: Peter Lang.

Corpas, G. and Seghiri M. 2009. "Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish)". In Beeby, A., Sánchez, P. and Rodríguez P. (Eds) Corpus Use and Learning to Translate. Proceedings from the CULT Conference, Barcelona, Spain, John Benjamins, 75-107.

Corpas, G., Mitkov R., Afzal, N. and Garcia Moya, L.

Proceedings of the Ninth Workshop on Building and Using Comparable Corpora, pages 1–2,
LREC 2016, Portorož, Slovenia, May 23, 2016.

2008. "Translation universals: do they exist? A corpus-based and NLP approach to convergence". Proceedings of the LREC'2008 Workshop on Building and Using Comparable Corpora.

Corpas, G., Mitkov R., Afzal, N. and Pekar, V. 2008. "Translation universals: do they exist? A corpus-based NLP study of convergence and simplification". Proceedings of the AMTA'2008 conference, Honolulu, Hawaii, 75-81.

Costa, H., Corpas, G., Mitkov, R. and M. Seghiri. 2015. "Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora". In Proceedings of the 7th International Conference of the Iberian Association of Translation and Interpreting Studies (AIETI'2015). Malaga, Spain

Costa, H., Corpas, G. and R. Mitkov. 2015. "Measuring relatedness between documents in comparable corpora". In Proceedings of the 11th International Conference on Terminology and Artificial Intelligence (TIA'15), Granada, Spain, 29-37.

Fung, P. and Cheung, P. 2004. "Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus". In Proceedings of the 20th international conference on Computational Linguistics (COLING), Geneva, Switzerland

Ilisei, I., Inkpen, D., Corpas, G., and Mitkov, R. 2012. "Romanian Translational Corpora: Building Comparable Corpora for Translation Studies". In Proceedings of the 5th Workshop on Building and Using Comparable Corpora (5th BUCC), held in conjunction with LREC 2012, Istanbul, Turkey, 56-61.

Kilgarriff , A. 2010. "Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project". In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC'2010, Malta.

Mendoza Rivera, O., Mitkov R. and G. Corpas Pastor. 2013. "A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora" In Proceedings of the International Workshop on Multiword units in Machine Translation and Translation Technology. Nice, France.

Mitkov R., Pekar V., Blagoev D. and Mulloni A. 2008. "Methods for extracting and classifying pairs of cognates and false friends ". Machine Translation. 21 (1), 29-53.

Mitkov, R. 2016. "The benefit of comparable corpora: automatic translation of multiword expressions without translation resources" (forthcoming). In Corpas, G. and Seghiri, M. (Eds). Corpus-based approaches to translation and interpreting: from theory to applications. Peter Lang

Pekar V., Mitkov R., Blagoev D. and Mulloni A. 2008. "Finding Translations for Low-Frequency Words in Comparable Corpora". Machine Translation, 20 (4), 247-266.

Pekar V., Mitkov R., Blagoev D. and Mulloni A. 2007. "Finding Translations for Low-Frequency Words in Comparable Corpora. " Proceedings of the CONTEXT-07 Workshop on "Contextual Information in Semantic Space Models" (CoSmo-2007), Roskille, Denmark, 17-25.

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., and Babych, B. 2012. "ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. Proceedings of the ACL 2012 System Demonstrations, Jeju, Korea, 91-96.

Rapp, R, Sharoff, S. and Zweigenbaum, P. (Eds). 2016. Special Issue on using comparable corpora for Machine Translation. Journal of Natural Language Engineering, 22(4). (forthcoming).

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. and Pinnis, M. 2012. "Collecting and Using Comparable Corpora for Statistical Machine Translation". Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 438–445.

Štajner, S and Mitkov, R. 2012. Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness, in Proceedings of the 5th Workshop on Building and Using Comparable Corpora (5th BUCC), held in conjunction with LREC 2012, Istanbul, Turkey, 88-97.

Stambolieva, E. 2012. Compiling Comparable Corpora: A Machine Learning Approach. MSc Dissertation, University of Wolverhampton.

Su, F. and Babych, B. 2012a. "Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents". Proceedings of the EACL'12 Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), Avignon, France, 10-19.

Su, F. and Babych, B. 2012b. "Development and Application of a Cross-language Document Comparability Metric". Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey. 3956-3962.

Taslimipoor, S., Mitkov, R., Corpas Pastor, G. and Fazly, A. 2016. "Bilingual Contexts from Comparable Corpora to Mine for Translations of Collocations." In A. Gelbukh (Ed.): CICLing 2016, LNCS vol. 9623. Springer, Heidelberg.

Yapomo, M., Corpas. G., Mitkov, R. 2012. "CLIR- and ontology-based approach for bilingual extraction of comparable documents Proceedings of the 5th Workshop on Building and Using Comparable Corpora, LREC 2012, Istanbul, Turkey, 121-125.