Abs-149

Silvia Bernardini[1], Sara Castagnoli[1], Adriano Ferraresi[1, 2], Federico Gaspari[1] and Eros Zanchetta[1]
[1] University of Bologna (Italy)
[2] University of Naples “Federico II” (Italy)

Turning Wikipedia into Comparapedia: Towards a new type of comparable corpus for language professionals

Special-purpose comparable corpora are among the most valuable resources for translators. Typically, however, they are not publicly available and have to be constructed as the need arises (Varantola 2003). This solution is not ideal, since the resulting corpora are likely to be small if constructed manually and rather low-quality if an automatic procedure is used, e.g. the BootCaT method (Baroni and Bernardini 2004), or else absorb more time and effort than most translators are willing to spend on the task.

Many translation professionals and students today resort to the Web and to Web-based resources for documentation about a specialised domain, for solving content-related problems, and for finding translation equivalents. Wikipedia is one of the most popular choices, thanks to features such as its multilingual nature, size, variety of domains covered, and up-to-dateness. In terms of linguistically-sophisticated searching and handling of results, though, it suffers from the well-known drawbacks often pointed out with reference to the Web itself (Fletcher 2004).

The paper describes the method we used to tap the potential of Wikipedia for corpus construction, and compares it with other attempts along similar lines (e.g. Gamallo Otero and González López 2010). Comparapedia (En-It) is a large bilingual corpus (over 270 million words in English and almost 140 million words in Italian), allowing on-the-fly consultation of theme-restricted comparable sub-corpora. Adapting tools and methods developed for Web-as-corpus construction (Baroni et al. 2009), all the bilingual data (i.e. explicitly linked entries) are extracted from a given Wikipedia dump, cleaned, lemmatised, part-of-speech tagged, and indexed using the Corpus WorkBench (Christ 1994). Keywords are obtained from human-inserted categories and recorded as structural attributes with each text. Other structural attributes include the article id (corresponding to its title) and the article target (the title of the matching article in the other language). Thus, Comparapedia allows users not only to search (the English and/or Italian) Wikipedia as a corpus, but also to search single texts and sub-corpora covering exactly the same topics in two languages. At the moment this is achieved through the use of keywords, even though in future work we intend to investigate the potential of Wikipedia-derived ontologies for this purpose (e.g. Nastase et al. 2010).

Current work, that the paper will also address, focuses on investigating the potential of Comparapedia as a hybrid com-parallel corpus. Since some of the entries are likely to have been translated from their matching entry (or from a third text), it should be possible to align (parts of) them, and search them as a parallel (sub-)corpus. This raises methodological and theoretical issues concerning the current status of established notions such as translated vs./and original text, parallel vs./and comparable corpus, collaborative vs./and conventional authoring.

References

Baroni, M. and S. Bernardini (2004) “BootCaT: Bootstrapping corpora and terms from the web”. In *Proceedings of LREC 2004*. Lisbon: ELDA. 1313-1316.

Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta (2009) “The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora”. *Journal of Language Resources and Evaluation* 43 (3): 209-226.

Christ, O. (1994) “A modular and flexible architecture for an integrated corpus query system”. In *Proceedings of COMPLEX’94*, Budapest, 1994. Online: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench.

Fletcher, W. (2004) “Making the web more useful as a source for linguistic corpora”. In Connor, U. and T. Upton (eds.) *Corpus Linguistics in North America 2002*. Amsterdam: Rodopi. 191-205.

Gamallo Otero, P. and González López, I. (2010) “Wikipedia as Multilingual Source of Comparable Corpora”. In *Proceedings of the third BUCC Workshop, LREC 2010.* La Valetta, Malta, 2010. 21-25.

Nastase, V., M. Strube, B. Börschinger, C. Zirn, and A. Elghafari (2010) “WikiNet: A very large scale multi-lingual concept network”. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, La Valetta, Malta, 2010. 1015-1022.

Varantola, K. (2003) “Translators and disposable corpora”. In Zanettin, F., S. Bernardini and D. Stewart (eds.) *Corpora in translator education*. Manchester: St. Jerome. 55-70.