Dynamic Lexica for Query Translation

Jussi Karlgren, Magnus Sahlgren, Timo Järvinen, Rickard Cöster
SICS and Connexor
jussi@sics.se mange@sics.se timo.jarvinen@connexor.com rick@sics.se

1 Lexical Resources Should be Dynamic

Multilingual information access applications, which are driven by modeling lexical correspondences between different human languages, are obviously reliant on lexical resources to a high degree — the quality of the lexicon is the main bottleneck for quality of performance and coverage of service. While automatic text and speech translation have been the main multilingual tasks for most of the history of computational linguistics, today the recent awareness within the information access field of the multilingual reality of information sources has made the availability of lexica an all the more critical system component.

Machine readable lexica in general, and machine readable multilingual lexica in particular, are difficult to come across. Manual approaches to lexicon construction vouch for high quality results, but are time- and labour-consuming to build, costly and complex to maintain, and inherently static as to their nature: tuning an existing lexicon to a new domain is a complex task that risks compromising existing information and corrupting usefulness for previous application areas. As a specific case, human-readable dictionaries, even if digitized and made available to automatic processing, are not vectored towards automatic processing. Dictionaries originally designed for human perusal leave much information unsaid, and belabor fine points that may not be of immediate use for the computational task at hand.

Automatic lexicon acquisition techniques promise to provide fast, cheap and dynamic alternatives to manual approaches, but have yet to prove their viability. In addition to this, they typically require sizeable computational resources. This experiment utilises a simple and effective approach to using distributional statistics over parallellized bilingual corpora – text collections of material translated from one language to another – for automatic multilingual lexicon acquisition and query translation. The approach is efficient, fast and scalable, and is easily adapted to new domains and to new languages. We evaluate the proposed methodology by first extracting a bilingual lexicon from aligned Swedish-French data, translating CLEF topics from Swedish to French, and then retrieving documents using the resulting French queries and a mono-lingual retrieval system from the French section of the CLEF document database. The results clearly demonstrate the viability of the approach.

2 Cooccurrence-based Bilingual Lexicon Acquisition

Cooccurrence-based bilingual lexicon acquisition models typically assume something along the lines "... If we disregard the unassuming little grammatical words, we will, for the vast majority of sentences, find precisely one word representing any one given word in the parallel text. Counterterms do not necessarily constitute the same part of speech or even belong to the same word class; remarkably often, corresponding terms can be identified even where the syntactical structure is not isomorphic."[3] or alternatively formulated "... words that are translations of each other are more likely to appear in corresponding bitext regions than other pairs of words" [6]. These models, first implemented by Brown and colleagues [1] use aligned parallel corpora, and define a translational relation between terms that are observed to occur with similar distributions in corresponding text segments.

Our approach, the *Random Indexing* approach, by contrast with most other approaches to distributionally based algorithms for bilingual lexicon acquisition, takes the context – an utterance, a window of adjacency, or when necessary, an entire document – as the primary unit. Rather than building a huge vector space of contexts by lexical item types, we build a vector space which is large enough to accommodate the occurrence information of tens of thousands of lexical item types in millions of contexts, yet compact enough to be tractable; constant in size in face of ever-growing data sizes; and designed to model association between distributionally similar lexical items without compilation or explicit dimensionality reduction. Our approach is described in detail in several publications [2, 4, 7, 8]; this paper describes experiments made on this year's CLEF data.

3 Experiment

We use the document-aligned Europarl corpus [5] which consists of parallel texts from the proceedings of the European Parliament, and is available in 11 European languages, freely from http://www.isi.edu/~koehn/europarl. From this multilingual corpus we extracted the Swedish-French section which we then lemmatized and normalized using the commercially available FDG tools from Connexor. The resulting data consist of several tens of thousands of sentence-level aligned document pairs. These data were used to extract a bilingual Swedish-French lexicon.

The topic texts were lemmatized and normalized using the same morphological analysis tools from Connexor as were used for the Swedish corpus.

The queries were translated word-by-word from Swedish to French using the extracted lexicon.

The text retrieval engine used for our experiments is a system being developed at SICS, The system is described in more detail in our CLEF paper [9] from last year. The French target collection was indexed by the system and the translated French queries were used to retrieve texts from the French collection without manual intervention.

4 Results

p@100 12 best 26 on or above median 9 near but below median 8 worst p@1000 26 best 34 on or above median 4 near but below median 4 worst ap 2 best 19 on or above median 10 near but below median 2 worst

The results were reasonably good with 34 of fifty queries on or above median, whereof 26 queries at top score for the "precision at 1000 documents" recall oriented score. For the other established two scoring schemes ("average precision" and "precision at 100 documents") the results were slightly lower, but the majority of queries in each case on, above or near median submitted scores. A more precision oriented evaluation scheme where average precision is calculated at 5 retrieved documents gives a satisfying score of 30 per cent.

Closer result analysis is still in progress, but some of the failed queries can be observed as having retrieved no documents at all. This is typically due to mistranslation or missing translation of some crucial query term, most often a name.

Acknowledgements

The work reported here is partially funded by the European Commission under contracts IST-2000-29452 (DUMAS) which is hereby gratefully acknowledged.

References

[1] P. Brown, S. Cocke, V. Della Pietra, F. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. A statistical approach to language translation. In *Proceedings of the 12th Annual Conference*

on Computational Linguistics (COLING 88). International Committee on Computational Linguistics, 1988.

[2] P. Kanerva, J. Kristofersson, and A. Holst. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036. Erlbaum, 2000.

[3] Hans Karlgren. Term-tuning, a method for the computer-aided revision of multi-lingual texts. *International Forum for Information and Documentation*, 13(2):7–13, 1988.

[4] J. Karlgren and M. Sahlgren. From words to understanding. In Y. Uesaka, P. Kanerva, and H. Asoh, editors, *Foundations of Real-World Intelligence*, pages 294–308. CSLI Publications, 2001.

[5] P. Koehn. Europarl: A multilingual corpus for evaluation of machine translation. net resource, 2002.

[6] D. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

[7] M. Sahlgren. Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data. In *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*, 2004.

[8] M. Sahlgren and J. Karlgren. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, forthcoming.

[9] Magnus Sahlgren, Jussi Karlgren, Rickard Cöster, and Timo Järvinen. Automatic query expansion using random indexing. In *Proceedings of CLEF 2002*, 2002.