Natural Language in Information Retrieval

Elżbieta Dura

Lexware Labs, Göteborg, Sweden
elzbieta@lexwarelabs.com

Abstract. It seems the time is ripe for the two to meet: NLP has grown out of prototypes and IR is having hard time trying to improve precision. Two examples of possible approaches are considered below. Lexware is a lexicon-based system for text analysis of Swedish applied in an information retrieval task. NLIR is an information retrieval system using intensive natural language processing to provide index terms on a higher level of abstraction than stems.

1 Not Much Natural Language in Information Retrieval so Far

Problems of finding the right data in a big data collection had been addressed long before NLP and are still addressed without NLP. The two fields hardly meet: "There is (…) an inherent granularity mismatch between the statistical techniques used in information retrieval and the linguistic techniques used in natural language processing." [8]. The results obtained in attempts of using NLP in information retrieval were so poor that the title of an article describing yet another test in 2000 is meant to surprise: "Linguistic Knowledge Can Improve Information Retrieval" [1]. The tenet of SMART seems to be still generally valid in IR: "good information retrieval techniques are more powerful than linguistic knowledge" [2].

When NLP-track was introduced in TREC in the nineties, several experiments proved that language resources can actually help. The gain in recall and precision is not negligible even if far from a dramatic breakthrough. For instance, adding simple collocations to the list of available terms could improve precision by 10%. [2] More advanced NLP techniques remain too expensive for large-scale applications: "the use of full-scale syntactic analysis is severely pushing the limits of practicality of an information retrieval system because of the increased demand for computing power and storage." [6].

2 NLIR – a Natural Language Information Retrieval

NLIR and Lexware are examples of projects which pursue improvement in IR by incorporation of NLP, each in a different way. The conviction behind the Natural Language Information Retrieval system – NLIR, is that "robust NLP techniques can help to derive better representation of text documents for indexing and search purposes than any simple word and string-based methods commonly used in statistical full-text retrieval." [6] The system is organized into a "stream model". Each stream provides an index representing a document in one special aspect. Various streams have been tried and reported in TREC, from 1995 on. Streams are obtained from different NLP methods which are run in parallel on a document. Contribution of each stream is optimised during merging the results of all streams.

All kinds of NLP methods are tested in NLIR. In TREC-5 a Head-Modifier Pairs Stream involves truly intensive natural language processing: part of speech tagging, stemming supported with a dictionary, sentence analysis with Tagged Text Parser, extraction of head-modifier pairs from the parse trees, corpus-based disambiguation of long noun phrases. Abstract index terms are obtained from the stream, in which paraphrases like *information retrieval*, *retrieval of information*, *retrieve more information*, etc can be linked together. In TREC-7 the streams are yet more sophisticated, e.g. a functional dependency grammar parser is used, which allows linking yet more paraphrases, e.g. *flowers grow wild* and *wild flowers*. The conclusions are positive but cautious: "(…) it became clear that exploiting the full potential of linguistic processing is harder than originally anticipated." [7] The results prove also that it is actually not worth the effort because the complex streams turn out to be the less effective than a simple Stems Stream, i.e. content words.

The approach of NLIR is a traditional statistical IR backbone with NLP support in recognition of various text items, which in turn is supposed to provide index terms on a higher level of abstraction than stems. The approach of Lexware is almost opposite: an NLP backbone plus support from statistics in assigning weights to abstract index terms. These are constituted primarily by lexemes.

3 Rich Resources and Shallow Analysis in Lexware

Lexicon and the concept of lexeme are central in Lexware approach. This means that word forms are associated with content from the beginning, which in its turn opens up for adding content information dependent on a specific task. In the information retrieval task described below Lexware performance is boosted by integration of its lexicon with a thesaurus specifically developed for the domain of the documents to be retrieved.

Text-analysis is shallow and it is not demanding in terms of computing power and storage. [3] The strength of the system is its rich lexicon and the possibility to expand the lexicon with external information without negative impact on access times. [4] Lexware has about 80 000 lexical items represented with features and relations of their forms and senses. Complex items are represented in terms of components. Kernel vocabulary items are separated, which is important when weights are calculated - occurrences of kernel items are less relevant than occurrences of more specific items.

4 Lexware Applied in Indexing of Swedish Parliamentary Debates

"Djupindexering" is the Swedish name of an application which assigns keywords to documents of Swedish parliamentary debates. Keywords are chosen from a thesaurus of about 4000 terms specially created for parliamentary debates. Each document is assigned from 2 to 10 keywords that best represent its content. The indexing is performed manually at the moment. The task for automatic indexing is the same as for human indexer: choose such terms for keywords that not only are representative of the subject of the document but also have proper level of specificity. For instance, when a document takes up university education the term *university education* and not *education* should be picked from the thesaurus.

The task is performed by Lexware as follows. In the preprocessing phase lexemes are identified in thesaurus terms. A document is analyzed in order to establish its index terms. Both thesaurus terms and lexemes are identified in text occurrences. Independent occurrences of components of complex thesaurus terms are also recorded if semantically heavy. Relevance weights of index terms in a document can be very precisely calculated thanks to the possibility of taking into consideration thesaurus relations. For instance, if a term occurs in a document together with its parent term or with majority of its children its weight can be increased.

Lexware does not use parallel sources like in NLIR but it operates on index terms of high level of abstraction from the beginning. When relevance of an index term of a document is to be decided Lexware can invoke all information present in its lexicon besides corpus statistics.

5 Evaluation

The Swedish parliament library – Riksdagsbiblioteket, designed and conducted tests of software in order to determine whether manual indexing of the parliament documents could be supplemented or even substituted by automatic indexing. Software from Connexor, Lingsoft, Kungliga Tekniska Högskola and LexWare Labs participated in the tests. The evaluation was based on a comparison of keywords assigned manually and automatically by the tested programs. The overlap in keywords assigned manually by two different indexers was only 34%, which is not astonishing given a high detail level of the thesaurus. Lexware proved to obtain the best F-value (2*precision*recall) / (precision + recall)): 36%, Kungliga Tekniska Högskola 32%, Connexor 22%, Lingsoft 19%.

Recent tests of the fully developed Lexware application for indexing of parliament documents proves to have surprisingly high coverage with full precision. Lexware automatic indexing was compared with manual indexing for 1400 documents from Riksdagsbiblioteket. 64.61% of keywords from Lexware are the same as those assigned manually, 22.99% are closely related in the thesaurus to those assigned manually. Thus 87.60% of keywords selected from the thesaurus are relevant. 9.84% of Lexware keywords not found among manually provided keywords are significant proper names. Only 2.56% of keywords are really different from the ones chosen in manual indexing. These require manual inspection in order to determine whether they are different but relevant or different and irrelevant.

6 Conclusions

Natural language processing may not be of assistance in all information retrieval applications but there are clear cases in which it leads to better results. For instance, NLIR tests show that query building clearly gains from NLP. Lexware indexing system based on a thesaurus performs very well: it is both fast and precise. Considering the fine results of Lexware Djupindexering it seems that the limitation to a specific language is not a major drawback..

The reluctance of IR people is not astonishing at all. They equate NLP with costly syntactic analysis which helps them very little if at all. Language resources rather than NLP techniques proved so far to have some impact on effectiveness in document retrieval. The Meaning-Text Theory advocating enormous size lexicons and multitude of paraphrasing rules in a description of any natural language may be the proper inspiration for natural language processing in information retrieval tasks. Now that language resources are built for many languages, it is not necessary that information retrieval should be limited to methods which do not involve comprehensive knowledge of a specific language. As a matter of fact, it is hard to see how precision can be hoped to improve otherwise.

References

1. Bookman, L.A., Green, S., Houston, A., Kuhns, R.J., Martin, P. and Woods, W.A. 2000. Linguistic Knowledge can Improve Information Retrieval. Proceedings of ANLP-2000, Seattle, WA, May 1-3, 2000.
2. Buckley, C., Singhal, M., Mitra, M. 1997. Using Query Zoning and Correlation within SMART: TREC-5 Report. Online at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html
3. Dura, E. 1998. Parsing Words. Data linguistica 19. Göteborg: Göteborgs universitet.
4. Dura, E. 2000. Lexicon-based Information Extraction with Lexware. In: PALC99 Proceedings.
5. Strzalkowski, T., Lin, F. J. Wang, J., Guthrie, L., Leistensnider, J. Wilding, J., Karlgren, J., Straszheim, T., Perez-Carballo, J. 1997 Natural Language Information Retrieval: TREC-5 Report. Online at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html
6. Strzalkowski, T., Stein, G., Bowden Wise, G., Perez-Carballo, J., Tapanainen, P., Jarvinen, T., Voutilainen, A., Karlgren, J.1999. Natural Language Information Retrieval: TREC-7 Report. Online at: http://trec.nist.gov/pubs/trec7/t7_proceedings.html
7. Strzalkowski, T., Perez-Carballo, J., Karlgren, J., Hulth, A., Tapanainen, P., T. Lahtinen, T. 2000. Natural Language Information Retrieval: TREC-8 Report. Online at: http://trec.nist.gov/pubs/trec8/t8_proceedings.html
8. Survey of the State of the Art in Human Language Technology. 1996. Cole, R.A, Mariani J., Uszkoreit, H., Zaenen, A., Zue, V. (eds.). Online at: http://cslu.cse.ogi.edu/HLTsurvey/ch7node4.html