NLP Track at TREC-5

Summarized by Tomek Strzalkowski from notes by Karen Sparck Jones and himself

ABSTRACT

NLP track has been organized for the first time at TREC-5 to provide a more focused look at how NLP techniques can help in achieving better performance in information retrieval. The intent was to see if NLP techniques available today are mature enough to have an impact on IR, specifically if and when they can offer an advantage over purely quantitative methods. This was also a place to try some more expensive and more risky solutions than those used in main TREC evaluations.

1. AIMS

More specifically, there were two principal aims of NLP track evaluations:

1. To see whether NLP has value in specific retrieval circumstances even if it has not hitherto been proven advantageous for routine document/text indexing and retrieval.
2. To see if NLP can be effectively used as a means to translate an NL text into whatever representation the search engine allows: this applies to either documents or queries, or both. In term-based systems, we have a representation that is basically: terms + weights + "=" (i.e., equivalence relation between terms). Can NLP help to get closer to the 'optimal' query.

2. PARTICIPANTS

Five teams participated in this NLP track: GE/Rutgers/NYU/Lockheed Martin, Xerox, Mitre, Claritech, and ISS Singapore. Results were submitted by the first four teams only. In addition, Chris Buckley supplied baselines for Sabir/SMART system. Other "baselines" were created by GE abd Xerox teams running their system in no-NLP mode.

3. EVALUATION SETUP

The evaluation was done in the ad-hoc retrieval mode only. Both automatic and manual modes were allowed. In an automatic run, no human intervention was permitted at any stage. In a manual run, queries could be expanded or modified manually, by adding or deleting terms or text, including from any documents in the test collection.

4. RESULTS

All systems did better than SMART statistical baseline, some substantially so (see attached recall-precision graphs). At least three out of the four systems used some kind of phrase extraction mechanism based on more or less elaborate syntactic analysis of text. This is worth noting particularly because the SMART baseline system extracts rudimentary statistical "phrases" (adjacent word bigrams) to expand word-only indexing. Thus, at least in this particular setup, linguistic phrases seem more effective than adjacency bigrams.

FIGURE 1. NLP Track Summary: Best Results

run id	GENLP4	CLARMC	xerox_nlp5	Mitre	SMART
type	manual	manual	manual	manual	auto. base
11pt prec %change	0.3176 +79	0.2842 +60	0.2320 +31	0.1896 +7	0.1771
R-prec. %change	0.3090 +70	0.2934 +61	0.2490 +37	0.1859 +2	0.1823

run id	xerox_nlp4	GENLP3	CLPHR1	SMART
type	automatic	automatic	automatic	auto. base
11pt prec %change	0.2280 +29	0.2220 +25	0.2010 +13	0.1771
R-prec. %change	0.2460 +35	0.2242 +23	0.2127 +17	0.1823

In addition to phrase-based indexing, full-text query expansion experiments performed by GE-led team showed very promising results. In this method, original search queries are expanded adding entire text passages from any documents containing related material. See Strzalkowski et al. paper for details.

Claritech team experimented with several alternative phrase extracting methods for document indexing. These included head-modifier pairs, adjacent subphrases, and full noun phrases. Phrases were obtained using very fast, shallow noun phrases parser. Further experiments included various combinations of phrase indexing methods and traditional single word indexing. Claritech results show the strongest gain from phrasal indexing. See Evans et al. paper for details.

GE/NYU/Rutgers/Lockheed Martin team used “stream-based” architecture to evaluate several phrase-indexing approaches, including head+modifier representation obtained via full syntactic parsing of entire data set. GE’s head+modifier pairs include verb+object and subject+verb combinations in addition to pairs obtained from noun phrases. Precision gains were less than for Clarit system, with unnormalized phrases slighly outperforming the more advanced head+modifer representation. In addition, manual and automatic full-text query expansion methods have been used, producing very encouraging results.

Mitre’s experiments were limited to using part-of-speech tagger and applying differential term weighting depending upon its part of speech. They noted only minimal gains over statistical SMART baseline. See Burger et al. paper for details.

Xerox group’s goal was to recreate on a larger scale Joel Fagan’s experiments in which he compared the effects of using syntactic and statistical phrases for document indexing. Statistical phrases were obtained using adjacent word pairs that occurred with certain frequencies in the data set. Syntactic phrases were derived with a “light-weight” phrasal parser, but no normalization (e.g., head-modifier) was performed. These experiments showed only very modest improvement over non-NLP baseline. For details please see Grefenstette et al. paper.

5. CONCLUSIONS

This NLP track demonstrated that natural language processing techniques have solid but limited impact on the quality of text retrieval, particularly precision. Techniques aimed at producing higher quality queries, e.g., query expansion, constraints, appear to be more effective than those aimed primarily at obtaining improved indexing of database documents. More work is needed before more substantial gains can be seen, including the use of more advanced, and therefore more expensive, semantic analysis techniques.

Figure 2 summarizes a rather subjective view of which NLP techniques have been tried in information retrieval, and what might be their potential for improving retrieval precision. This chart was discussed at the NLP track workshop on the last day of TREC-5 meeting. It was decided that NLP techniques that show particular promise in relatively smaller-scale track evaluations should be transferred to main evaluations as soon as practical.

ACKNOWLEDGEMENTS

We would like to thank Donna Harman, Ellen Voorhees, and Dawn Hoffman, of NIST for their support in organizing of NLP track evaluations. This project was supported in part by the Defense Advanced Research Projects Agency under the Tipster Phase 2 and Phase 3 contracts 94-FI59900-000, and 97-FI56800-000.

FIGURE 2. NLP results analysis: a subjective view

NL technique	class	%change precision
Full-text query expansion	query build	40 to ???
Term-based query expansion	query build	15 to 25
deleting extraneous text from queries	query build	0 to 5
hyphenated phrases	phrases	-15
word bi-grams	phrases	5 to 10
extended bi-grams (windows)	phrases	-5
FSA phrases (noun groups)	phrases	7 to 25
Head+Modifier Pairs (full parsing)	phrases	2 to 15
proper names	concepts	1 to 3
concept tagging for indexing	concepts	0 to ???
concept tagging for re-ranking	concepts	0 to 3
stylistics	discourse	0 to ???
lexical normalization	stemming	5 to 8