MediaLab @ CLEF-2003:
Using keyword disambiguation.

Peter van der Weerd
pweerd@medialab.nl

MediaLab BV, Schellinkhout, The Netherlands
http://www.medialab.nl

Abstract

This report describes the participation of MediaLab BV in the CLEF-2003 evaluations. This year we participated in the monolingual Dutch task, experimenting with a keyword disambiguation tool. MediaLab developed this tool to exploit human assigned keywords in the search engine in a better way than just blind searching with the keywords themselves. Although this tool was not planned to be used for CLEF-like applications it was fun to check if it could help boosting the search quality.

1. Disambiguation

In traditional search applications people are used to assign keywords to searchable items, and then let the user search via these assigned keywords. The main problem with this approach is that the searching user has to follow the same thoughts as the assigning people to let the correct keywords cross his/her mind. However, people invested lots of time in assigning keywords, so it is a pity not using this effort.

MediaLab developed a tool for ranking a list of keyword (or other items) given a full text query. This list is generated in 2 phases.

- During indexing we build a co-occurrence network of the used words in the item at one site and the assigned keywords at the other side.
- At retrieval time we split the query into words and for each query word its connected keywords from the co-occurrence network are collected.

Note that the tool uses the human assigned keywords, so it is fully dependant of the quality of the assigned keywords.

For example, using this tool in the library of Eindhoven querying for “jaguar” let to a top3 keyword list of:

- jaguar (auto); geschiedenis   [jaguar (car); history]
- katachtigen   [cat-likes]
- zuid-amerikaanse mythen   [South American myths]

With the same library data we experimented feeding the tool normalised author names instead of keywords. Results were surprisingly good. For instance querying for “pesten” (nagging) the top3 authors all wrote children’s books about nagging.

MediaLab plans to use the tool in search applications. Except presenting the standard result list we present also a list of best keywords, authors, etc which can be used by the user to re-order the results. In this way the user has the ability to view several cross-sections of the result list.

2. Approach

The CLEF data-collection contains some extra fields with keyword information (HTR) and with geographical information (GEO). We used both extra fields to feed the disambiguation tool. The searching process is done in the following way:

- doing a "normal" search giving result R1
- determine the top5 of disambiguation items and search all these items giving result R2
- than recomputed the weights in R1 by adding a fraction of the weights in R2
- the modified R1 is used as the submission

We submitted a base run, and for each field (HTR and GEO) we submitted 5 runs with different relative weights of the second result.
HTR-5 means: normal result combined with a 50% weight of the HTR-result.
HTR-2 means: normal result combined with a 20% weight of the HTR-result.

3. Results

The following table summarizes some measures of the runs.

Run	**Rel_ret**	**Precision at 10 docs**	**Average prec. (non-interp.)**	**R-precision**
base	1248	0.4071	0.3959	0.3695
HTR1 (10%boost)	1265	0.4018	0.4044	0.3809
HTR2 (20%boost)	1282	0.4000	0.4095	0.3903
HTR3 (30%boost)	1292	0.3946	0.3939	0.3749
HTR4 (40%boost)	1285	0.3893	0.3737	0.3507
HTR5 (50%boost)	1257	0.3804	0.3520	0.3396
GEO1 (10% boost)	1258	0.4036	0.3983	0.3739
GEO2 (20% boost)	1282	0.3929	0.3916	0.3661
GEO3 (30% boost)	1271	0.3750	0.3783	0.3479
GEO4 (40% boost)	1264	0.3536	0.3577	0.3309
GEO5 (50% boost)	1222	0.3268	0.3383	0.3185

It is clear that blind boosting the results with HTR or GEO data helps a little bit to retrieve more relevant documents. In case of boosting the results by the HTR data the optimum is about 10% to 20%, increasing the average precision with 3%. However, the effect is rather small.

The profit of boosting by the GEO data is less convincing, probably caused by the quality of the GEO-data. Looking at the data made us already hesitate about using it.

4. Conclusion

Although MediaLab's disambiguation tool was not intended for blind boosting search results, it might be used for it. Probably better results are achieved by using these fields in a normal blind relevance feedback procedure as used McNamee and Mayfield and others [1].

5. References

[1] *APL Experiments at CLEF: Translation Resources and Score Normalization*, Paul McNamee and James Mayfield, Johns Hopkins University, USA, in [1]