Sense Discrimination for Physics Retrieval

Christina Lioma
Institute for Natural Language Processing
Stuttgart University, Germany
liomaca@ims.uni-stuttgart.de

Alok Kothari
Institute for Natural Language Processing
Stuttgart University, Germany
kotharak@ims.uni-stuttgart.de

Hinrich Schütze
Institute for Natural Language Processing
Stuttgart University, Germany
hs999@ifnlp.org

ABSTRACT

Information Retrieval in technical domains like physics is characterised by long and precise queries, whose meaning is strongly influenced by term context and domain. We treat this as a disambiguation problem, and present initial findings of a retrieval model that posits a higher probability of relevance for documents matching disambiguated query terms. Preliminary evaluation on a real-life physics test collection shows promising performance improvement.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Information Retrieval, Sense Discrimination

1. INTRODUCTION

User queries in technical domains like physics tend to be long and precise, e.g. `I am looking for a general expression for the heat transfer coefficient in a liquid solid interface`. For such queries, standard bag of word approaches may fail to capture the dependence between the query constituents, or the modification relationships of their components. For a shorter query this may not harm retrieval, and often the collocation of the query terms in the collection suffices. However, for longer queries, a similiar approach would result in notably more collocations in the collection, causing accidental matches with nonrelevant documents. An additional problem of technical domains like physics is that words may acquire special domain-specific meanings, e.g. in the query above, `expression` denotes a formula. This sense is not only different to other senses that `expression` may convey, but also central to the user need.

We present initial findings of a retrieval model that posits a higher prior probability of relevance for documents matching disambiguated query terms. Disambiguation is realised with Schütze & Pedersen's [3] unsupervised approach that disambiguates query terms by considering their context representations in the retrieval collection. This approach outputs a disambiguated sense for a query term, and a cluster of documents that contain the query term in its disambiguated sense. We treat this output as evidence of relevance of those documents to the query, and we embed it into the retrieval model as a prior probability. Initial findings on a real-life physics test collection are presented.

2. SENSE DISCRIMINATION MODEL

Let t' be a query term that we wish to disambiguate. We use Schütze & Pedersen's [3] algorithm to build a context vector for each occurrence of t' in the collection and in the query. This context vector consists of features occurring within $n = 10$ words of t' (following [3]), and uses as features sequences of alphanumeric characters. Context vectors are weighted using inverse context frequency and cosine-normalised: $v_{ij} = \frac{\delta(i,j)\log\frac{N}{N(j)}}{\sqrt{\sum_k(\delta(i,k)\log\frac{N}{N(k)})^2}}$, where v_{ij} is component j of context vector $\vec{v}_i$ of term t'; $\delta(i,j)$ is 1 if feature j occurs in context i, 0 otherwise; $N(j)$ is the number of contexts of t' in which feature j occurs; and N is the total number of contexts of t'. The weighted context vectors in the collection are then clustered into $k = 20$ *sense clusters* using k-means clustering. Disambiguating t' consists of assigning the query context vector t' to the closest centroid of the sense clusters computed for t' in the collection.

The single sense cluster assigned to t' consists of contexts that occur in documents in the collection. These documents are deemed by our method to contain t' with the same sense as used in the query. Hence, boosting the ranking of these documents may benefit retrieval performance. We implement this boosting as a prior probability that these documents are relevant, which we incorporate into the basic query likelihood model used for retrieval.

Given a query Q and a document D, the query likelihood model [1] ranks D by $P(D|Q)$. By Bayes rule, $P(D|Q) \propto P(Q|D)P(D)$, where $P(Q|D)$ is the probability that D generates Q, and $P(D)$ is the document's prior probability. There are several ways of estimating $P(Q|D)$, for instance using Jelinek-Mercer or Dirichlet smoothing. $P(D)$ is usually assumed to be uniform, in which case documents are ranked solely by $P(Q|D)$. Alternatively, $P(D)$ can be viewed as prior knowledge about the relevance of D [4]. Various types of prior evidence about document relevance have been implemented as $P(D)$, e.g. document quality [4]. In this work, we measure the degree of match between the sense of

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

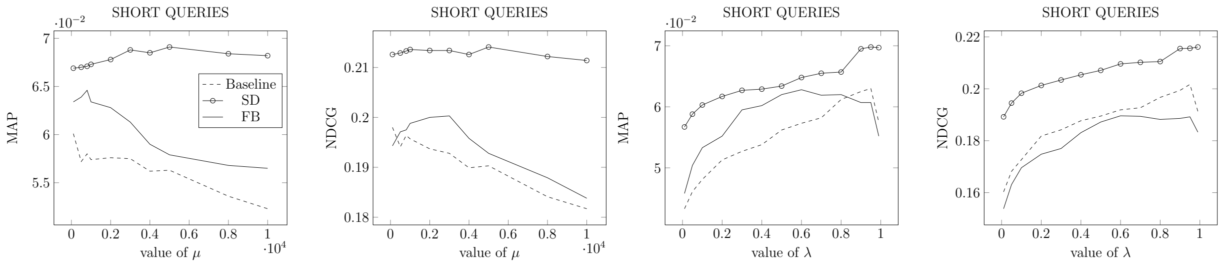

Figure 1: Retrieval performance vs. parameter μ (Dirichlet)/λ (Jelinek-Mercer) for short queries.

a term in the document with its sense used in the query, and we implement it as $P(D)$.

The value of $P(D)$ is typically derived from measurements about the evidence being modelled as the document prior. In this work we choose a fixed value ($P(D) = 0.5$) to boost those documents that disambiguate query terms. The performance of our approach reported in this work may further improve if prior values are tuned, however, at this stage, our aim is to test whether our approach is beneficial to retrieval, and not to optimise its performance.

	SHORT QUERIES			LONG QUERIES		
	MAP	BPREF	NDCG	MAP	BPREF	NDCG
Dir	0.0601	0.1954	0.1980	0.0552	0.1813	0.1892
JM	0.0630	0.1944	0.2017	0.0579	0.1844	0.1920
Dir+SD	**0.0691***	**0.1965***	**0.2141***	**0.0682***	**0.1874**	**0.2086**
JM+SD	**0.0698**	**0.1945**	**0.2161**	**0.0708**	**0.1855**	**0.2106**
Dir+FB	**0.0646**	**0.1959**	**0.2003**	**0.0585**	0.1802	0.1780
JM+FB	**0.0628**	0.1936	0.1896	**0.0610**	0.1819	0.1846

Table 1: Retrieval performance. * = stat. significance at $p < 0.05$ (2-tailed t-test). Bold = better than the baseline.

3. EVALUATION

We use the iSearch real-life physics test collection [2], which contains approx. half a million physics documents and 65 queries with graded relevance assessments, created by physicists. Queries contain 5 fields: *1. information need, 2. task, 3. background, 4. ideal answer, 5. keywords.* We use Indri for indexing and retrieval without removing stopwords or stemming, because these settings give the highest baseline performance. Our baseline runs use the Kullback-Leibler language model [1] with Dirichlet and Jelinek-Mercer smoothing (Dir, JM resp.). Our sense disambiguation (SD) runs use Dir & JM, enhanced with the SD priors presented above. We also include runs with pseudo-relevance feedback (FB) using Indri's default implementation, in order to compare our method to a more competitive approach. We use short queries (fields 1&5)[1] and long queries (all fields), to check the effect of query length on SD. We measure performance with: mean average precision (MAP), binary preference (BPREF), and normalised discounted cumulative gain (NDCG). For each measure, we tune: (for Dir) $\mu \in \{100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10000\}$; (for JM) $\lambda \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$; (for FB) the number of feedback documents $fbD \in \{1, 2, 5, 10, 20\}$ and the number of feedback terms $fbT \in \{3, 5, 10, 20, 40\}$.

Table 1 shows that for MAP & NDCG, SD performs the best, with large relative improvements over the baseline and FB. Note that MAP uses binary relevance, whereas NDCG uses graded relevance; SD is shown superior on both of these measures at all times, which is a good indication of its usefulness to IR. For BPREF, the improvement of SD over the baseline and FB is not large, but the best run for both types of queries is again a SD run. This result may be affected by a bias in iSearch: its relevance assessments contain a much larger proportion of documents judged as nonrelevant than of documents judged as relevant [2]. BPREF is affected by this because it depends heavily on the number of judged non-relevant documents that are retrieved at higher ranks than relevant documents (whereas MAP & NDCG do not distinguish between non-relevant and non-judged documents). Regarding query length, the longer the query, the higher the improvement brought by SD, most likely because longer queries have lower baseline scores, so there is more room for improvement. The improvement of SD is not heavily reliant on smoothing optimisation, as Figure 1 shows (the plots of long queries and BPREF are similar to these): the improvement of SD is more stable across different parameter values than the baseline and FB, especially for short queries.

4. CONCLUSION

This poster presented initial work on a retrieval model that includes a word sense disambiguation component, the output of which is embedded into the ranking function as a prior probability. The motivation was to improve retrieval performance for technical domains like physics, where words often have domain-specific senses. Preliminary experiments on a real-life physics test collection gave positive findings; more testing on other technical collections are needed to further explore our so-far promising approach.

5. REFERENCES

[1] W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval.* Kluwer, 2003.

[2] M. Lykke, B. Larsen, H. Lund, and P. Ingwersen. Developing a test collection for the evaluation of integrated search. In *ECIR*, pages 627–630, 2010.

[3] H. Schütze and J. O. Pedersen. Information retrieval based on word senses. In *SDAIR*, pages 161–175, 1995.

[4] Y. Zhou and W. B. Croft. Document quality models for web ad hoc retrieval. In *CIKM*, pages 331–332, 2005.

[1] We use fields 1&5 because they give higher baseline performance than field 5 alone.