Effective Term Weighting for Sentence Retrieval

Saeedeh Momtazi[1], Matthew Lease[2], Dietrich Klakow[1]

[1] Spoken Language Systems, Saarland University, Germany
[2] School of Information, University of Texas at Austin, USA

Abstract. A well-known challenge of information retrieval is how to infer a user's underlying information need when the input query consists of only a few keywords. Question Answering (QA) systems face an equally important but opposite challenge: given a verbose question, how can the system infer the relative importance of terms in order to differentiate the core information need from supporting context? We investigate three simple term-weighting schemes for such estimation within the language modeling retrieval paradigm [6]. While the three schemes described are ad hoc, they address a principled estimation problem underlying the standard word unigram model. We also show these schemes enable better estimation of a state-of-the-art class model based on term clustering [5]. Using a TREC QA dataset, we evaluate the three weighting schemes for both word and class models on the QA subtask of sentence retrieval. Our inverse sentence frequency weighting scheme achieves over 5% *absolute* improvement in mean-average precision for the standard word model and nearly 2% absolute improvement for the class model.

1 Introduction

Information Retrieval (IR) addresses a critical user need to be able to find relevant information in vast digital libraries. However, IR systems typically treat documents as the atomic unit of retrieval, and it is often the case that only a portion of any given document is actually relevant to the user's information need. Another shortcoming of standard IR systems is their emphasis on putting the burden on the user to formulate short keyword queries. Such formulation becomes increasingly difficult as information needs become more complex and can often lead to iterative query reformulation and search abandonment.

In contrast to typical IR, Question Answering (QA) both supports focused retrieval and allow people to easily express their information needs as natural language questions. While it is a laudable goal to shift the burden of effort from users in query formulation to systems in query interpretation, a clear challenge lies in developing QA systems capable of effectively interpreting such queries. Addressing this challenge represents an important direction for long-term research, and this paper presents an early step toward this overarching goal.

A standard QA system architecture incorporates several subtasks: (i) retrieving relevant documents (ii) performing Sentence Retrieval (SR) from those documents, and (iii) extracting answers from sentences. In this work, we focus on improving accuracy of the SR component. In comparison to document

retrieval, SR poses several distinct challenges: (1) the brevity of sentences vs. documents exacerbates the usual term-mismatch problems, and (2) the verbosity of questions can lead to critical query terms being obscured by supporting terms. Various research has been done to improve SR performance, such as the ones proposed by Balasubramanian [2] and Allan [1]; however, to the best knowledge of the authors, there was no focus on the above problems. In this paper, regarding to (1), we build on recent work addressing term-mismatch via class-based modeling [5]. As for (2), we investigate three simple term-weighting strategies for approximating the relative importance of query terms: Inverse Document Frequency (IDF), Inverse Collection Frequency (ICF) [7], and a novel Inverse Sentence Frequency (ISF) scheme. While more elaborate and principled estimation schemes can be envisioned for inferring such relative importance, the above schemes are simple, efficient, and as results show, remarkably effective.

2 Method

In word unigram Language Model (LM) retrieval [6] and class model based on term clustering [5] sentences S are ranked by :

$$P_{word}(Q|S) = \prod_{q \in Q} P(q|S) \tag{1}$$

$$P_{class}(Q|S) = \prod_{q \in Q} P(q|C_q, S)P(C_q|S) \tag{2}$$

where $Q = \{q_i \dots q_{|Q|}\}$ denotes a query of length $|Q|$. C_q is the cluster that contains q, $P(q|C_q, S)$ is the emission probability of q given its cluster and the sentence, and $P(C_q|S)$ is estimated based on clusters instead of terms.

Significant work has explored methods like Dirichlet-smoothing for better estimating the unigram models underlying observed documents, and the correlate here is the unigram model $P(Q|S) = \Theta^S$ underlying S. Complimenting this work, KL-divergence ranking was described in which a latent unigram Θ^Q is assumed to represent the user's information need underlying Q, and sentences are ranked via minimal KL-divergence between distributions [8]:

$$-D(\Theta^Q||\Theta^S) \stackrel{rank}{=} \Theta^Q \cdot \Theta^S \tag{3}$$

Of note here is that the standard LM approach is equivalent to KL-ranking only if Θ^Q is estimated from Q via Maximum Likelihood (ML). Thus the standard unigram model can be understood as implicitly assuming a uniform distribution over query terms, meaning all query terms are inferred to be equally important to the underlying information need. While this assumption is reasonable for short keyword queries, it is increasingly problematic as query length increases due to large variance of relative term importance in natural language [4]. While not described in this way, Smucker and Allan [7] introduced ICF-weighting as a simple (and admittedly ad hoc) alternative to ML for better estimating Θ^Q.

We present the first investigation of this strategy for the SR task. In addition to considering ICF, we also evaluate IDF and a similar ISF scheme which differs from IDF by counting sentences rather documents. We evaluate these three

schemes for estimating Θ^Q in the context of two methods for modeling Θ^S: directly (the standard word-based model [6]) and via the class-based approach described above. In all cases, Dirichlet-smoothing is used to estimate Θ^S.

3 Evaluation

We evaluate our SR models using questions from the TREC[3] 2006 QA track with the TREC 2005 set was used for development. Documents come from the AQUAINT corpus[4] of 450 million tokens of English newswire text. Because original TREC relevance judgments were only made at the coarser document level, we used the *Question Answer Sentence Pair* corpus of Kaisser and Lowe [3].

To evaluate the SR component of our QA system independent of the document retrieval component, we adopted the following experimental setup. A separate sentence collection was first created for each *question-series* (TREC QA data specifies each questions in the context of a series of related questions). For each series, we identify all documents known to be relevant to *any* question in the series, and we add all sentences from that document to the sentence collection. The average size of this collection is 270 sentences per question while the average number of relevant sentences per question is only 4. Moreover, the non-relevant sentences in each collection exemplify exactly the sort of typical QA system false alarms we want our SR system to avoid: non-relevant sentences coming from (1) documents relevant to similar yet different questions and (2) non-relevant sentences found in relevant documents.

IDF, ISF, and ICF statistics were taken from AQUAINT. We did the similar experiments with the larger Gigaword corpus[5] and achieved similar trends which are not reported further. Following Momtazi and Klakow [5], we used the same approach to build the class model.

Table 1. Mean-average precision of different weighting methods for word and class models. * marks statistical significance at $p<0.01$ for 2-tailed paired t-test.

Model	**Baseline**	**Weighted**		
		IDF	ICF	ISF
Word LM	0.3696	0.4211*	0.4096*	0.4244*
Class LM	0.4174	0.4233	0.4336*	0.4353*

Table 1 shows results for mean-average precision; similar improvements were seen with mean reciprocal rank. For both word and class models, ISF-weighting is seen to consistently perform best and yield significant improvement. While both ISF and IDF-weighting of the word model exceed accuracy of the baseline class model, IDF-weighting fails to improve the class model. While ICF and ISF-weighting yield similar improvements for the class model, ICF-weighting under-performs ISF-weighting for the word model.

[3] http://trec.nist.gov
[4] Linguistic Data Consortium corpus LDC2002T31
[5] Linguistic Data Consortium corpus LDC2003T05

Comparing the results of ISF- and IDF-weighting, our intuition is that the more specific term contexts used by ISF is more informative. That is, ISF-weighting compiles its frequency statistics from narrower text segments in comparison to IDF-weighting, which uses a far wider context and does not consider the number of time a word appears in a specific context.

While the weighting schemes approximate term importance via simple frequency statistics, the class-based model clusters frequent terms into a single class which implicitly decreases their effect in a similar fashion. Thus it is not too surprising that the weighting schemes are somewhat less effective with the class-based model. Nevertheless, statistically significant improvement is still achieved over the baseline class-based model.

4 Summary

This paper showed a simple and effective way for integrating several alternative term weighting strategies with word or class models for sentence retrieval. While far more work will be needed to bring us closer to our long-term goal of supporting QA for rich, complex natural language questions, we believe this work represents a simple first step in this direction and provides a new, useful baseline to which more sophisticated methods can be later compared.

Acknowledgements Saeedeh Momtazi is funded by the German research foundation DFG through the International Research Training Group (IRTG 715).

References

1. J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of ACM SIGIR International Conference*, pages 314–321, 2003.
2. N. Balasubramanian, J. Allan, and W. Croft. A comparison of sentence retrieval techniques. In *Proceedings of ACM SIGIR International Conference*, pages 813–814, 2007.
3. M. Kaisser and J. Lowe. Creating a research collection of question answer sentence pairs with Amazon's mechanical turk. In *Proceedings of the LREC International Conference*, 2008.
4. M. Lease, J. Allan, and W. B. Croft. Regression Rank: Learning to Meet the Opportunity of Descriptive Queries. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 90–101, 2009.
5. S. Momtazi and D. Klakow. A word clustering approach for language model-based sentence retrieval in question answering systems. In *Proceedings of ACM CIKM International Conference*, pages 1911–1914, 2009.
6. J. Ponte and W. Croft. A language modeling approach to information retrieval. In *Proceedings of ACM SIGIR International Conference*, pages 275–281, 1998.
7. M. Smucker and J. Allan. Lightening the load of document smoothing for better language modeling retrieval. In *Proceedings of ACM SIGIR International Conference*, pages 699–700, 2006.
8. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):214, 2004.