Improving Sentence Retrieval with an Importance Prior

Leif Azzopardi
Department of Computing Science
University of Glasgow, United Kingdom
leif@dcs.gla.ac.uk

Ronald T. Fernández, David E. Losada
Dept. of Electronics and Computer Science
University of Santiago de Compostela, Spain
{ ronald.teijeira, david.losada } @usc.es

ABSTRACT

The retrieval of sentences is a core task within Information Retrieval. In this poster we employ a Language Model that incorporates a prior which encodes the importance of sentences within the retrieval model. Then, in a set of comprehensive experiments using the TREC Novelty Tracks, we show that including this prior substantially improves retrieval effectiveness, and significantly outperforms the current state of the art in sentence retrieval.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Experimentation, Performance

Keywords: Sentence Retrieval, Language Models

1. INTRODUCTION

Sentence retrieval (SR) is a challenging problem area that has received a significant amount of attention recently [1, 4, 5, 7]. The main SR task consists of finding relevant sentences from a document base given a query. This task is very useful in a wide range of Information Retrieval (IR) applications, such as summarization, question answering, and opinion mining. However, the task has usually been approached by taking a document retrieval model and adapting it for SR. In fact, the model that is the state of the art in SR is known as term frequency-inverse sentence frequency (TF.ISF), which is analogous to the traditional TF.IDF method used in document retrieval [1, 4]. While, numerous attempts to develop more sophisticated models that employ techniques such as Natural Language Processing and Clustering have been proposed [2, 3, 8], they have failed to significantly and consistently outperform the TF.ISF method. Consequently, little progress has been made in terms of improving sentence retrieval effectiveness.

In this poster we posit that a relevant sentence needs to be indicative of the query, but also representative and important within the context of the document; i.e. we assume that key statements within a document are more likely to be relevant, if they are on topic. With this aim, we adopt the Language Modeling framework and include a sentence based prior to encode the importance of a sentence in a document within the model. In a set of experiments performed over several TREC test collections, we compare the proposed models against existing SR models and show that using an importance prior within a LM framework delivers retrieval performance that significantly outperforms the current state of the art.

2. SENTENCE RETRIEVAL MODEL

The SR task consists of estimating the relevance of each sentence s in a document d in a given document set D, and supply the user with a ranked list of sentences which satisfy his/her need (expressed as a user query q). Using a language modeling framework to address this problem has been previously performed by Murdock [5] and Losada and Fernández [4]. The standard Language Modeling approach to SR estimates the probability of a query given a sentence language model (for specific details see [4, 5]). However, an unexplored extension is the inclusion of a sentence prior encoding the importance of the sentence within the context of the document.

To include a prior of importance of a sentence in a document, here we explicitly include the document in the sentence model and treat SR as a problem of estimating the probability of the query and the document given the sentence i.e. $p(q,d|s)$. This probability tells us how likely the sentence is to produce both the query and the document, i.e. is it relevant to the query and central to the document? Using Bayes' Theorem, we can re-write it to become:

$$p(q,d|s) \propto p(q|s,d)\,p(d|s) \tag{1}$$

where $p(d|s)$ is the probability of the document given the sentence and $p(q|s,d)$ is the probability of the query given the sentence and document:

$$p(q|s,d) = \prod_{t \in q} \left(\alpha p(t|s) + \beta p(t|d) + \gamma p(t) \right)^{c(t,q)} \tag{2}$$

where $\alpha + \beta + \gamma = 1$. In [5] Eq. 2 is used (we shall refer to this as 3MM), while in [4] either Jelinek-Mercer (JM) or Dirichlet (DIR) smoothing is employed by setting the parameters appropriately[1]. These three models provide the standard sentence language modeling baselines. For the proposed extension shown in Eq. 1 we need to estimate $p(d|s)$ which can be regarded as the importance of a sentence in a document[2]. To facilitate the estimation, Bayes Theorem can be employed, and then the components can be expressed as language models, so that:

$$p(d|s) = \frac{p(s|d)\,p(d)}{p(s)} \propto \frac{p(s|d)}{p(s)} = \frac{\prod_{t \in s} p(t|d)^{c(t,s)}}{\prod_{t \in s} p(t)^{c(t,s)}} \tag{3}$$

where $p(s|d)$ is the probability of a sentence given a document, the $p(s)$ the probability of a sentence, $p(d)$ is the prior probability of a document, $p(t|d)$ is the probability of generating t from the maximum likelihood estimator of the document, $c(t,\cdot)$ is the number of times the term appears in the sentence/document/query. Here, we assume that there is no a priori preference towards any

[1] For JM, $\beta = 0$. For DIR, $\beta = 0$, $\gamma = \mu/(c(d)+\mu)$ and $\alpha = 1-\gamma$, where $c(d)$ is the number of terms in the document.

[2] In the standard models $p(d|s)$ is assumed to be constant and is thus ignored.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

of the documents, and treat $p(d)$ as a constant. The $p(s|d)$ represents how likely the sentence is to be generated from the document, whereas $p(s)$ represents how likely the sentence is to be generated randomly. The ratio between the two expresses the importance of the sentence. Observe that $p(d|s)$ will give preference to those sentences that are central to the document's topics (i.e. high $p(s|d)$) but also rare within the collection (i.e. low $p(s)$). It should also be noted that this prior will implicitly tend to favor longer sentences because $p(t|d)$ is greater than $p(t)$[3]. With the importance prior, in our experiments we shall refer to the extended Language Models as 3MM.IP, JM.IP, and DIR.IP.

3. EMPIRICAL STUDY AND RESULTS

In this paper, we adopt the same definition of the sentence retrieval problem as proposed in the TREC Novelty Tracks. Although these tracks are mostly focused on researching redundancy filtering, they also involve a SR task that enables research into how to retrieve sentences that are relevant to a given query. The SR problem is framed as follows: given a textual query that represents an information need, a ranked set of documents is supplied and the systems have to process this ranking to extract the sentences that are estimated as relevant to the information need.

Data: Along with this definition we used all three TREC Novelty Track collections 2002, 2003 and 2004[4]. Each collection was indexed using the Lemur toolkit[5], where standard stop words were removed but stemming was not applied. The corresponding set of topics for each collection was used, where short queries were constructed taking the title field of the TREC Topic[6]. The TREC 2002 collection was used to train and estimate the parameters of each model used, while the TREC 2003 and 2004 collections were used to test the sentence retrieval models.

Models: In this work, we used a number of baseline models: (i) the current state of the art, TF.ISF [1], (ii) BM25 [6], which closely matches the performance of TF.ISF but is parameterized [4], and (iii) the standard sentence language models, JM and DIR, as well as 3MM [4, 5]. These are compared against the extended sentence language models, JM.IP, DIR.IP and 3MM.IP.

Measures: For all of our experiments, we report the performance of each method using Mean Average Precision (MAP) and R-Prec. To compare the differences in performance between the different methods, statistical significance tests were applied using the t-test with a 95% confidence level. Here, we show the statistical comparisons between each model and TF.ISF and DIR (see Table 1).

Results: Table 1 shows the performance obtained for each of the different models tested. Firstly, we note that the standard sentence language models do not outperform the state of the art TF.ISF or BM25. And in fact, TF.ISF and BM25 are significantly better than DIR. However, when the prior on sentence importance is incorporated within the language modeling framework, we note that these models all significantly outperform both TF.ISF and DIR, with improvements of up to 20% in some cases. The model that performed the best overall was DIR.IP which resulted in gains of 5-8% over TF.ISF. This is a substantive gain making these extended models an attractive and stronger baseline.

	TREC 2003		TREC 2004	
Model	**MAP**	**R-Prec**	**MAP**	**R-Prec**
TF.ISF	0.3851†	0.4581†	0.2358†	0.3298†
BM25	0.3852†	0.4580†	0.2368⋆†	0.3300†
JM	0.3474	0.4406	0.2131	0.3010
3MM	0.3513	0.4419	0.2195	0.3060
DIR	0.3638	0.4457	0.2240	0.3146
JM.IP	0.4137⋆†	0.4800⋆†	0.2548⋆†	0.3520⋆†
3MM.IP	0.4104⋆†	0.4802⋆†	0.2527⋆†	0.3504⋆†
DIR.IP	**0.4144⋆†**	**0.4802⋆†**	**0.2549⋆†**	**0.3522⋆†**

Table 1: The Mean Average Precision (MAP) and R-Precision (R-Prec) for each model on TREC 2003 and 2004. ⋆ and † denote that the model is significantly better than TF.ISF and DIR, respectively, ($p < 0.05$). Parameters estimated on TREC 2002.

4. DISCUSSION AND FUTURE WORK

In this poster, we proposed and empirically evaluated an extension of the LM framework for SR to include sentence importance through a prior. It was found that by including the importance prior substantial improvements were obtained for all the different Language Models which resulted in significantly better performance. However, as the importance prior implicitly tends to favor longer sentences, it may be the case that the improvements witnessed are due to better length normalization (if longer sentences are more likely to be relevant). This work also suggests that the naive application of document retrieval models to other task may lead to non-optimal performance. This will be the focus of future investigation along with examining how the vector space and other probabilistic models can be extended to also incorporate sentence importance and potentially better length normalization.

Acknowledgments: This work was partially supported by FEDER and Xunta de Galicia under projects 07SIN005206PR and 2008/068.

5. REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the* 26^{th} *ACM SIGIR* , pages 314–321, Toronto, Canada, 2003.

[2] S. Kallurkar, Y. Shi, R. S. Cost, C. K. Nicholas, A. Java, C. James, S. Rajavaram, V. Shanbhag, S. Bhatkar, and D. Ogle. UMBC at TREC 12. In *Proceedings of the* 12^{th} *TREC 2003*, pages 699–706, 2003.

[3] X. Li and W. B. Croft. Novelty detection based on sentence level patterns. In *Proceedings of the* 14^{th} *CIKM 2005*, pages 744–751, Bremen, Germany, 2005.

[4] D. E. Losada and R. T. Fernández. Highly frequent terms and sentence retrieval. In *Proceedings of the* 14^{th} *SPIRE 2007*, pages 217–228, Chile, 2007.

[5] V. G. Murdock. *Aspects of sentence retrieval*. PhD thesis, University of Massachusetts Amherst, September 2006.

[6] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, VCL and interactive track. In *Proceedings of the* 7^{th} *TREC* , pages 253–264, Gaithersburg, USA, 1999.

[7] R. W. White, J. M. Jose, and I. Ruthven. Using top-ranking sentences to facilitate effective information access. *American Society for Information Science and Technology*, 56(10):1113–1125, 2005.

[8] M. Zhang, C. Lin, Y. Liu, L. Zhao, and S. Ma. THUIR at TREC 2003: Novelty, robust and web. In *Proceedings of the* 12^{th} *TREC*, pages 556–567, Gaithersburg, USA, 2003.

[3] So in the product in Eq. 3 the ratio for each term in the sentence is greater than one, and the more terms the greater the influence.

[4] See `http://trec.nist.gov` for track descriptions and reports.

[5] `http://www.lemurproject.org`

[6] It should be noted that most teams participating in the TREC novelty tracks used the whole topic, so our results are not directly comparable to the official TREC results, but instead are based on a more realistic scenario.