Language Modeling and Document Re-Ranking: Trinity Experiments at TEL@CLEF-2009

Dong Zhou[1,2] Vincent Wade[1]

1. Centre for Next Generation Localisation, University of Dublin, Trinity College, Dublin 2, Ireland
2. School of Computer and Communication, Hunan University, Changsha, Hunan, China, 410082

dongzhou1979@hotmail.com Vincent.Wade@cs.tcd.ie

Abstract

This paper presents a report on our participation in the CLEF-2009 monolingual and bilingual *ad hoc* TEL@CLEF tasks involving three different languages: English, French and German. Language modeling is adopted as the underlying information retrieval model. While the data collection is extremely sparse, smoothing is particular important when estimating a language model. The main purpose of the monolingual task is to compare different smoothing strategies and investigate the effectiveness of each alternative. This retrieval model is then used alongside a document re-ranking method based on Latent Dirichlet Allocation (LDA) which exploits the implicit structure of the documents with respect to original queries for the monolingual and bilingual tasks. Experimental results demonstrated that three smoothing strategies behave differently across testing languages while LDA-based document re-ranking method should be considered further in order to bring significant improvement over the baseline language modeling systems in the cross-language setting.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods; Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Experimentation, Measurement

Keywords

Language Modeling, Smoothing, Document Re-Ranking, Latent Dirichlet Allocation, Cross-Language Information Retrieval

1 Introduction

This year's participation in the CLEF 2009 *ad hoc* monolingual and bilingual track was motivated by a desire to compare different smoothing strategies applied to language modeling for library data retrieval as well as test and extend a newly developed document re-ranking method.

Language modeling has been successfully applied to the problem of ad hoc retrieval [4, 5, 6]. It provides an attractive information model due to its theoretical foundations. The basic idea behind this approach is extremely simple - estimate a language model for each document and/or a query, and rank documents by the likelihood of the query (with respect to the document language model) or by the distance between the two models. The main object of smoothing is to adjust the maximum likelihood estimator of a language model so that it will be more accurate [7].

However, previous success over news collection data does not necessarily mean it will be efficient over the library data. Firstly the data is actually multilingual: all collections to a greater or lesser extent contain records pointing to documents in other languages. However this is not a major problem because the majority of documents in the test collection are written in main languages of those test collections. Furthermore, documents written in different languages tend not to match the queries in main languages. The main characteristic of the data is that it is very different from the newspaper articles and news agency dispatches previously used in the CLEF. The data tends to be very sparse. Many records contain only title, author and subject heading information; other records provide more detail (see the experiment section on what fields are chosen to include). The average document lengths are 14.66 for BL and 24.19 for BNF collections after pre-processing, respectively.

Recently, there is a trend of exploring the hidden structure of documents to re-rank results [2, 3, 8]. We addressed in the previous work that there are two important factors that should be taken into account when designing any re-ranking algorithms: the original queries and initial retrieval scores. Based on this observation, we introduce a new document re-ranking method based on Latent Dirichlet Allocation (LDA) [1, 9] which exploits implicit structure of the documents with respect to original queries. Rather than relying on graph-based techniques as in [5, 6] to identify the internal structure, the approach tries to directly model the latent structure of ”topics” or ”concepts” in the initial retrieval set. Then we can compute the distance between queries and initial retrieval results based on latent semantic information inferred. Experiments in [9] demonstrated the effectiveness of the proposed method in monolingual retrieval. In this experiment, we try to extend the approach to cross-language information retrieval.

2 Methodology

2.1 Language Modeling

Smoothing a data set typically means creating an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena. In language modeling, the simplest reason to use smoothing is to not assign a zero probability to unseen words. The accuracy of smoothing is directly related to the retrieval performance.

Given a text sequence (either a query or a document), the probability distribution can be regarded as a probabilistic language model M_d or M_q from each document d or each query q. In other words, it assumes that there is an underlying language model which ”generates” a term (sequence) [1]. The unigram language model is utilized here. There are several ways to estimate the probabilities. Let $g(w \in d)$ denotes the number of times the term w occurs in a document d (same idea can be used on a query). The Maximum-likelihood estimation (MLE) of w with respect to d is defined as:

$$MLE_d w = \frac{g(w \in d)}{\sum_{w'} g(w' \in d)} \tag{1}$$

We choose to use three representative methods that are widely used in previous research and relatively efficient to implement. The Jelinek-Mercer method, the Bayesian smoothing using Dirichlet priors, and the absolute discounting method. The description of the methods could be found in [7]

2.2 Document Re-Ranking

The intuition behind the document re-ranking method is the hidden structural information among the documents: *similar documents are likely to have the same hidden information with respect to a query*. In other words, if a group of documents are talking about the same topic which shares a strong similarity with a query, in our method they will get allocated similar ranking as they are more likely to be relevant to the query. In addition, the refined ranking scores should be relevant to the initial ranking scores, which, in the experiments conducted in this paper, are combined together with the re-ranking score either using a linear fashion.

Table 1: Indexing and searching fields

Fields	English	French	German
dc:language	√	√	√
dc:identifier	√	√	√
dc:rights	√	√	√
dc:type	√	√	√
dc:creator	√	√	√
dc:publisher	√	√	√
dc:date	√	√	√
dc:relation	√		
dc:contributor	√	√	√
dcterms:issued	√		√
dcterms:extent	√		
dcterms:spatial			√
dcterms:isPartOf			√
dcterms:edition			√
dcterms:available			√
mods:location	√	√	√

The distance between a query and a document in this method adopts the *KL* divergence between the query terms and document terms and through a linear combination of the re-ranking scores based on initial ranker and the latent document re-ranker.

This method could be found in great detail in [9]. We apply this method into the cross-language re-ranking by concatenating texts from different languages into several dual-language documents and a single dual-language query. An LDA analysis of these texts results in a multilingual semantic space in which terms from both languages are presented. Hence force the re-ranking process can be carried out by directly model the latent structure of multilingual ”topics” or ”concepts” in this enriched initial retrieval set. The similarity of ”contexts” in which the terms appear is guaranteed to capture the inter-relationship between texts in different languages.

3 Experimental Setup

3.1 Overview of the Experimental Process

All of the documents in the experiment were indexed using the Lemur toolkit[1]. Prior to indexing, Porter's stemmer and a stopword list[2] were used for the English documents. We use a French analyzer[3]and a German analyzer to analyze French and German documents. The query sets consist of 50 topics, all of which were used in the experiment. Each topic is composed of several parts such as: *Title*, *Description*, *Narrative*. We chose to conduct *Title+Description* runs as queries. The queries are processed similarly to the treatment in the test collections. The chosen fields used in the indexing and searching are shown in the Table 1.

3.2 Experimental Runs

In order to investigate the effectiveness of various techniques, we performed a retrieval experiment with several permutations. These experimental runs are denoted as follows:

For monolingual retrieval:

LM-DIR: This part of the experiment involved retrieving documents from the test collection using language modeling with Bayesian smoothing method using Dirichlet prior.

LM-ABS: as above, except that the absolute discounting smoothing method was used.

LM-JM: as above, except that the Jelinek-Mercer smoothing method was adopted.

[1] http://www.lemurproject.org

[2] ftp://ftp.cs.cornell.edu/pub/smart/

[3] http://lucene.apache.org/

Table 2: Retrieval results for monolingual task

Run ID	source	target	description	MAP	bpref	P@5	P@10
TCDENRUN1	EN	EN	LM-DIR	0.2905	0.3001	0.4560	0.4140
TCDENRUN2	EN	EN	LM-ABS	0.4035	0.4054	0.6160	0.5640
TCDENRUN3	EN	EN	LM-JM	0.3696	0.3658	0.5680	0.5060
TCDFRRUN1	FR	FR	LM-DIR	0.1451	0.1570	0.2000	0.1740
TCDFRRUN2	FR	FR	LM-ABS	0.1745	0.1767	0.2320	0.2380
TCDFRRUN3	FR	FR	LM-JM	0.1723	0.1765	0.2520	0.2280
TCDDERUN1	DE	DE	LM-DIR	0.2577	0.2615	0.4480	0.3760
TCDDERUN2	DE	DE	LM-ABS	0.2397	0.2397	0.4280	0.3540
TCDDERUN3	DE	DE	LM-JM	0.2686	0.2653	0.4520	0.3840

For bilingual retrieval:

GOOGLETRANS: In this part of the experiment, documents were retrieved from the test collection using the Google Translator for translating the queries. (It is worth to note that due to the submission restrictions this is an unofficial measurement.)

GOOGLETRANS-LDA: Here we retrieved documents from the document collection using query translations suggested by the Google Translator, then directly re-rank the retrieval results using the translated query with the proposed LDA based document re-ranking method.

GOOGLETRANS-SLDA: Here we retrieved documents from the document collection using query translations suggested by the Google Translator, then we conducted a multilingual corpus with documents written in both query and document languages. Re-ranking was performed by apply the LDA based method on this multilingual space (with the translated and the original query).

4 Results and Discussion

4.1 Monolingual Task

In this section we compare three smoothing methods across different languages in the library search (Table 2). As we conducted queries using the title and description fields, they could be considered as long informative queries. Previous research on news and web data suggested that on average, Jelinek-Mercer is better than Dirichlet and absolute discounting in metrics like non-interpolated average precision, precision at 10 and 20 documents while both Jelinek-Mercer and Dirichlet clearly have a better average precision than absolute discounting. The German monolingual runs demonstrated the same observation where Jelnek-Mercer is better than Dirichlet, which is subsequently better than absolute discounting.

English and French runs showed a different behaviour. Absolute discounting is a clear winner of three smoothing methods while Jelinek-Mercer still perform better than Dirichlet. If explained by the two different roles in the query likelihood retrieval method [7], Dirichlet method performs better as it is good for the estimation role (as for shorter queries). So that it consistently demonstrate the worst performance across all the languages. However, Jelinek-Mercer performs best for longer queries and should be good for the role of query modeling. This is the case for the German runs while not for English and French runs where absolute discounting substitute the Jelinek-Mercer's role in the modeling process. The results suggest that smoothing methods tend to be sensitive for distinct languages and different test collections.

4.2 Bilingual Task

We now turn to the bilingual tasks to study the LDA-based re-ranking method. The main experimental results are presented in Table 3, for all three languages. The first question we are interested in is how the re-ranking method performs directly over the bilingual retrieval results (taken as a whole). It is shown that our methods bring improvements upon the Google translator baselines in all of the 6 relevant comparisons. Another observation is that in many cases, the method can outperform the baselines for all the evaluation metrics.

Table 3: Retrieval results for bilingual task

Run ID	source	target	description	MAP	bpref	P@5	P@10
TCDFRENRUN1	FR	EN	GOOGLETRANS	0.3481	0.3526	0.5760	0.5220
TCDFRENRUN2	FR	EN	GOOGLETRANS-LDA	0.3488	0.3527	0.5720	0.5220
TCDFRENRUN3	FR	EN	GOOGLETRANS-SLDA	0.3500	0.3535	0.5760	0.5140
TCDDEENRUN1	DE	EN	GOOGLETRANS	0.3411	0.3500	0.5700	0.5040
TCDDEENRUN2	DE	EN	GOOGLETRANS-LDA	0.3500	0.3596	0.5760	0.5040
TCDDEENRUN3	DE	EN	GOOGLETRANS-SLDA	0.3505	0.3602	0.5880	0.5040
TCDENFRRUN1	EN	FR	GOOGLETRANS	0.1579	0.1572	0.2520	0.2320
TCDENFRRUN2	EN	FR	GOOGLETRANS-LDA	0.1591	0.1573	0.2520	0.2340
TCDENFRRUN3	EN	FR	GOOGLETRANS-SLDA	0.1576	0.1561	0.2560	0.2320
TCDDEFRRUN1	DE	FR	GOOGLETRANS	0.1618	0.1743	0.2680	0.2300
TCDDEFRRUN2	DE	FR	GOOGLETRANS-LDA	0.1633	0.1752	0.2600	0.2300
TCDDEFRRUN3	DE	FR	GOOGLETRANS-SLDA	0.1624	0.1739	0.2600	0.2260
TCDENDERUN1	EN	DE	GOOGLETRANS	0.1901	0.1923	0.3480	0.2900
TCDENDERUN2	EN	DE	GOOGLETRANS-LDA	0.1910	0.1922	0.3480	0.2920
TCDENDERUN3	EN	DE	GOOGLETRANS-SLDA	0.1935	0.1944	0.3480	0.2920
TCDFRDERUN1	FR	DE	GOOGLETRANS	0.1826	0.2053	0.3480	0.2700
TCDFRDERUN2	FR	DE	GOOGLETRANS-LDA	0.1840	0.2063	0.3520	0.2780
TCDFRDERUN3	FR	DE	GOOGLETRANS-SLDA	0.1839	0.2050	0.3560	0.2760

With respect to the bilingual re-ranking, the method showed some improvements over the Googe translator and direct re-ranking methods in the X2EN and X2DE runs in terms of mean average precision. The performance is somewhat disappoint in the X2FR runs. Furthermore, although there are some improvements, the difference are not large enough in terms of MAP. However, regarding to the precision at 5 documents the method could demonstrate higher performance than simple re-ranking. This shows that the method is a promising direction but need further investigation.

It is worth mentioning that the combination of methods used in this experiment could achieve very good overall performance as nearly all of our selected monolingual and bilingual runs are among top five participants in CLEF 2009 (except in the French monolingual task) such as:

TCDENRUN2 absolute discounting, English monolingual

TCDDERUN1 Dirichlet prior, German monolingual

TCDDEENRUN3 Google translator with SLDA, German-English bilingual

TCDDEFRRUN2 Google translator with LDA, German-French bilingual

TCDENDERUN3 Google translator with SLDA, English-German bilingual

5 Conclusion

In this paper we have described our contribution to the CLEF 2009 *ad hoc* monolingual and bilingual tracks. Our monolingual experiment involved the comparison of three different smoothing strategies applied to language modeling approach for library data retrieval. We also made a first attempt to extend the previously proposed document re-ranking method to cross-language information retrieval. Experimental results demonstrated that smoothing methods tend to behave differently in the library search and across testing languages. They also showed that LDA-based document re-ranking method should be considered further in order to bring significant improvement over the baseline language modeling systems in the cross-language setting.

6 Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University of Dublin, Trinity College.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. 944937.

[2] Oren Kurland and Lillian Lee. Pagerank without hyperlinks: structural re-ranking using links induced by language models. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 306–313, Salvador, Brazil, 2005. ACM. 1076087.

[3] Oren Kurland and Lillian Lee. Respect my authority!: Hits without hyperlinks, utilizing cluster-based language models. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90, Seattle, Washington, USA, 2006. ACM. 1148188.

[4] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, Melbourne, Australia, 1998. ACM Press. 291008.

[5] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, Seattle, Washington, USA, 2006. ACM. 1148204.

[6] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, Atlanta, Georgia, USA, 2001. ACM. 502654.

[7] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004. 984322.

[8] Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 504–511, Salvador, Brazil, 2005. ACM. 1076120.

[9] Dong Zhou and Vincent Wade. Latent document re-ranking. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. (EMNLP 2009)*, pages 1571–1580, Singapore, 2009. Association for Computational Linguistics.