Combining Probabilistic and Translation-Based Models for Information Retrieval based on Word Sense Annotations

Elisabeth Wolf, Delphine Bernhard, Iryna Gurevych
UKP Lab, Technische Universität Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

In this paper, we describe our experiments carried out for the robust word sense disambiguation (WSD) track of the CLEF 2009 campaign. This track consists of a monolingual and bilingual task and addresses information retrieval utilizing word sense annotations. We took part in the monolingual task only. Our objective was twofold. On the one hand, we intended to increase the precision of WSD by a heuristic-based combination of the annotations of the two WSD systems. For this, we provide an extrinsic evaluation on different levels of word sense accuracy. On the other hand, we aimed at combining an often used probabilistic model, namely the Divergence From Randomness BM25 model (DFR_BM25), with a monolingual translation-based model. Our best performing system with and without utilizing word senses ranked 1st overall in the monolingual task. However, we could not observe any improvement by applying the sense annotations compared to the retrieval settings based on tokens or lemmas only.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 [**Information Search and Retrieval**]: Performance evaluation

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Probabilistic Retrieval Model, Monolingual Translation-based Model, Word Sense Disambiguation

1 Introduction

The CLEF [1] robust word sense disambiguation (WSD) track aims at promoting the development and evaluation of textual document retrieval systems utilizing word sense disambiguated data. Participants to this task are provided with topics and documents from previous CLEF campaigns which were annotated by two different WSD systems. The WSD track consists of two independent tasks: a monolingual task with topics and documents in English and a bilingual task with Spanish topics and English documents. In both tasks, the goal is to analyze and compare the performance

[1] www.clef-campaign.org

of retrieval systems on the document collection with and without word sense information. We took part in the monolingual task only.
The CLEF robust WSD track was first introduced in 2008 and received submissions from eight groups plus two late submissions. The participants submitted runs by different systems varying in the pre-processing steps, indexing procedures, ranking functions, the application of query expansion methods, and the integration of word senses. The best performance could be achieved by a combination of different probabilistic models [6], namely the BM25 model [12], the I(n_e)C2 model - a Divergence From Randomness version of the BM25 model -, and a statistical language model introduced by Hiemstra [8]. Through their combination approach, they obtained the highest mean average precision (MAP) over all submitted retrieval systems. All participants always took only one of the two systems for WSD into account when selecting the word sense annotations. According to the MAP, most submitted systems obtain higher performance on the plain document collection than on the word sense annotated corpus. Only some participants were able to slightly improve the performance by utilizing word sense information in their system. However, it is questionable whether these improvements were significant. The WSD task using the same document collection was repeated in 2009 in order to further investigate the performance on WSD annotated data.
Our motivation in the robust WSD task is twofold. On the one hand, we intended to increase the precision of WSD by an heuristic-based combination of the annotations of the two WSD systems. We provide an extrinsic evaluation on different levels of word sense accuracy. On the other hand, we aimed at combining an often used probabilistic model with a monolingual translation-based model, which was trained on definitions and glosses provided by different lexical semantic resources, namely WordNet, Wiktionary, Wikipedia, and Simple Wikipedia. This translation-based model was successfully used for the task of answer finding by Bernhard and Gurevych [3]. We report all different index and retrieval settings and the performance on the training and test data.
The paper is organized as follows. In the next section, we describe the provided document collection and define our indexing and retrieval approach in detail. We define the applied probabilistic and translation-based models as well as the method to combine both of them. In Section 3, we report and discuss our evaluation results, and finally, in Section 4 we conclude our experiments.

2 Experiments

2.1 Document collection

The document collection consists of Los Angeles Times 1994 and Glasgow Herald 1995 English national newspaper articles which were used for CLEF 2001. It comprises around 169,000 documents (113,000 Los Angeles Times documents; 56,500 Glasgow Herald documents). The documents were annotated with two different word sense disambiguation systems provided by the IXA NLP Group at the University of the Basque Country (UBC) [1] and the Department of Computer Science at the National University of Singapore (NUS) [4]. The UBC system is based on a combination of k-nearest neighbor classifiers. Each classifier learns from a distinct set of features. The set of features comprises, e.g., syntactic, collocations, and bag-of-words features as well as features learned from a reduced space via Singular Value Decomposition. The NUS approach extracts similar features from English-Chinese parallel corpora, the SEMCOR, and the DSO corpus. Based on the extracted features, an SVM is trained for each open-class word. Both WSD systems were among the top performing systems in the lexical sample and all-words WSD subtasks of SemEval-2007 [11]. For the fine-grained all-words WSD task, the NUS system obtained an accuracy of 0.587, while the UBC system's accuracy was 0.544. The final collection contains three different corpora: (i) the plain corpus, (ii) a corpus where each token is annotated with a lemma as well as multiple senses and probability scores using the UBC system, and (iii) a corpus with the same annotations from the NUS system. Word sense annotations refer to synsets in WordNet version 1.6.
Training (150) and test (160) topics consist of a combination of CLEF topics from previous chal-

lenges and are annotated with word sense information as well. Each topic consists of a brief *title*, a one-sentence *description*, and a more detailed *narrative* field specifying the relevance assessment criteria. Participants were instructed to create their queries only from the title and description fields. Documents and topics are provided in XML format.

2.2 Indexing

We used Terrier (TERabyte RetrIEveR) [9], version 2.1 for indexing the documents. This framework provides state-of-the-art retrieval and query expansion models, such as the commonly used Divergence From Randomness (DFR) BM25 probabilistic model. During the training phase, we determined the best performing combination of retrieval and query expansion method.
Each document is represented by its tokens. Each token is assigned a lemma and multiple word senses. The accuracy of word sense annotations can highly influence the retrieval performance when utilizing word senses (see e.g. Sanderson [13]). Therefore, we extrinsically analyzed the automatically annotated word senses based on information retrieval experiments. The original document collection consists of approximately 100 Mil. tokens including stop words. The NUS annotated corpus comes with around 199 Mil. sense annotations including the sense probability scores, i.e. on average 2 senses per token. The UBC annotated corpus even consists of around 275 Mil. sense annotations and probability scores, i.e. on average 2.75 senses per token. Preliminary experiments on the training topics have shown that restricting the incorporated senses to the highest scored sense for each token increases the MAP of retrieval.
Further, we hypothesize that combining the NUS and UBC sense assignments increases the precision of annotated word senses. Therefore, we created several indices for our experiments. Each index consists of three fields, namely token, lemma, and sense. The indexed senses vary in the way they are selected. Four different indices were created: (i) an index with the highest scored UBC sense for each token (**UBCBest**), (ii) an index with the highest scored NUS sense for each token (**NUSBest**), (iii) an index with senses that were assigned by both systems and have the greatest sum of scores (**CombBest**), and finally (iv) an index with senses as in (iii), but where we chose the sense with the highest score from the UBC or NUS corpus when the set of senses that were assigned by both systems is empty (**CombBest**$^+$). The construction of **CombBest** can be formally described by:

$$sense(t) = \underset{s \in S(t)}{\operatorname{argmax}} \quad score^{UBC}(s) + score^{NUS}(s) \tag{1}$$

with $S(t) = S^{UBC}(t) \cap S^{NUS}(t)$, where $S^{UBC}(t)$ is the set of senses of token t obtained from the UBC system and $S^{NUS}(t)$ is the sense set accordingly obtained from the NUS system. Thus, $S(t)$ is the intersection of the senses of token t annotated from the UBC and NUS systems. Further, $score^{UBC}(t)$ and $score^{NUS}(t)$ is the probability score assigned to sense s from the UBC and NUS system, respectively. Accordingly, **CombBest**$^+$ is defined as:

$$sense(t) = \begin{cases} \operatorname{argmax}_{s \in S(t)} & score^{UBC}(s) + score^{NUS}(s) & \text{if } S \neq \emptyset \\ \operatorname{argmax}_{s \in S^+(t)} & score^{UBC,NUS}(s) & \text{otherwise} \end{cases} \tag{2}$$

where $S^+(t) = S^{UBC}(t) \cup S^{NUS}(t)$ is the union of the sense sets of token t from the UBC and NUS systems.
We created multi field indices including fields for tokens, lemmas, and the word senses. Prior to indexing, we applied standard stopword removal. Without stopwords, all indices consists of approximately 40.7 Mil. tokens. As shown in the third column of Table 1 the UBCBest index contains around 34.1 Mil. senses, the NUSBest index contains around 34.5 Mil. senses, i.e. 6.6 Mil. and 6.2 Mil. tokens are not annotated with any sense in the UBCBest and NUSBest index, respectively. The CombBest index contains only 31.7 Mil. senses, while the CombBest$^+$ index consists of 35.1 Mil. senses.

The queries were automatically constructed from the topic fields *title* and *description*. The stopword list used for queries was the same as the one used for the documents, plus the following terms: *find*, *describing*, *discussing*, *document*, and *report*.

2.3 Retrieval Models

We carried out several retrieval experiments using a probabilistic model, a monolingual translation-based model, and their combination.

2.3.1 Probabilistic Model

Terrier provides a set of different probabilistic ranking models. We report the performance on the training and test topics applying the term-weighting model I(n)OL2, which is called DFR_BM25 in Terrier. The DFR_BM25 model is the Divergence From Randomness (DFR) version of the BM25 model [12]. According to Ounis et al. [10] the DFR version of the BM25 model infers the informativeness of a term t in the document d by the divergence between its within-document term-frequency and its frequency within the whole collection. Multiple TREC experiments have shown that this is a competitive model in various retrieval settings. The DFR_BM25 model is defined by

$$weight(t|d) = \frac{tfn}{tfn+1} \cdot \log_2 \Big(\frac{|D| - df + 1}{df + 0.5} \Big) , \tag{3}$$

where

$$tfn = tf(t|d) \cdot \log_2(1 + c \cdot \frac{adl}{dl}) , \tag{4}$$

$|D|$ is the size of the document collection, df is the document frequency, tf is the term frequency, dl is the document length, and adl is the average document length. We set the parameter c to the default value of 1. The probabilistic model can be applied on indexed tokens, lemmas, and senses.

2.3.2 Relevance Feedback

It is often observed that probabilistic models have problems dealing with synonymy. This problem, also called *lexical gap*, arises from alternative ways of expressing a concept using different terms. Query expansion models try to overcome the lexical gap problem by reformulating the original query to increase the retrieval performance. Terrier provides different query expansion models, namely the Bose-Einstein 1, the Bose-Einstein 2, and the Kullback-Leibler (KL) model [5]. We chose the KL query expansion model, since it performed best on preliminary experiments on the training data. The Kullback-Leibler Divergence term weighting model is defined by:

$$weight(t) = P_R(t) \cdot log \Big(\frac{P_R(t)}{P_C(t)} \Big) , \tag{5}$$

where $P_R(t)$ is the probability of the term t in the top ranked documents and $P_C(t)$ is the probability of the term t in the whole collection. Terms with a high probability in the top ranked documents and a low probability in the whole collection are likely to be the expansion terms. In our experiments the original query is expanded by up to 10 most informative (highest weighted) terms from the 3 top ranked documents.

2.3.3 Translation Model

A further solution to the lexical gap problem is the integration of monolingual statistical translation models first introduced by Berger and Lafferty [2]. These models encode statistical word associations which are trained on parallel monolingual document collections such as question-answer pairs. Recently, Bernhard and Gurevych [3] successfully applied monolingual translation models for the task of answer finding. In order to automatically train the translation models, they used the definitions and glosses provided for the same term by different lexical semantic resources,

index type	# tokens	# senses	MAP (training)	MAP (test)
UBCBest	40.1 Mil.	34.1 Mil.	0.2514	0.2636
NUSBest		34.5 Mil.	0.2930	0.3473
CombBest		31.7 Mil.	0.2921	0.3313
CombBest$^+$		35.1 Mil.	**0.3011**	**0.3551**

Table 1: Number of indexed word senses and MAP on retrieval for different index types (retrieval model: DFR_BM25 + KL).

namely WordNet, Wiktionary, Wikipedia, and Simple Wikipedia. The usage of these resources yields domain-independent monolingual translation models. The authors have shown that their models significantly perform better than baseline approaches for answer finding.
We employed the model defined by Xue et al. [14] and used by Bernhard and Gurevych [3] in our experiments:

$$P(q|D) = \prod_{w \in q} P(w|d)\,, \tag{6}$$

where

$$P(w|d) = (1-\lambda)P_{mx}(w|d) + \lambda P(w|D)\,, \tag{7}$$

$$P_{mx}(w|d) \quad = \quad (1-\beta)P_{ml}(w|d) + \beta \sum_{t \in d} P(w|t)P_{ml}(t|d)\,, \tag{8}$$

q is the query, d the document, λ the smoothing parameter for the document collection D and $P(w|t)$ is the probability of translating a document term t to the query term w. The parameter β was set to 0.8 and λ to 0.5.
We applied the translation-based model trained for the answer finding task on the newswire document collection, though it was not particular trained for this task. As the translation-based model was trained on tokens, we apply it on the indexed token field exclusively.

2.4 Combination of Retrieval Models

Our hypothesis is that translation-based models retrieve different documents for some queries than probabilistic models. Therefore, we compute a combined relevance score to improve the retrieval performance.
First, we normalize the scores resulting from each model applying standard normalization:

$$r_{norm}(i) = \frac{r_{orig}(i) - r_{min}}{r_{max} - r_{min}}\,, \tag{9}$$

where $r_{orig}(i)$ is the original score, r_{min} is the minimum, and r_{max} is the maximum occurring score for a query.
Second, we combine the normalized relevance scores computed for individual models into a final score using the CombSUM method introduced by Fox and Shaw [7]. This method ranks the documents based on the sum of the (normalized) similarity scores of individual runs. Each run can be assigned a different weight.

3 Results

In the following subsections we describe all our results carried out during our experiments and discuss them in detail. We chose the five best performing experiments with and without utilizing word senses for submission based on the MAP values obtained for the training set. In addition to the officially submitted runs we report some further experiments on the test topics.

retrieval model	training data				test data			
	token	lemma	sense		token	lemma	sense	
			Comb Best	Comb Best$^+$			Comb Best	Comb Best$^+$
translation model	0.3045	-	-	-	0.3616	-	-	-
DFR_BM25	0.3374	0.3425	0.2565	0.2557	0.3741	0.4054	0.2867	0.3096
DFR_BM25 + KL	0.3760	**0.3829**	0.2921	0.3011	0.4223	**0.4451**	0.3313	0.3551

Table 2: MAP values of the different retrieval models and fields

3.1 Preliminary Experiments on Word Senses

As stated in Section 2.2 we created four indices which differ in the way word senses assigned by the UBC and NUS systems are selected. Table 1 shows the number of indexed word senses and the MAP values of different retrieval experiment applying the DFR_BM25 ranking model with the Kullback-Leibler query expansion model. Retrieval on the UBCBest index shows a MAP value of 0.2514 for the training and 0.2636 for the test topics. For retrieval based on the NUSBest index the MAP value increases by 14.2% and 24.1% for training and test topics, respectively. According to this extrinsic evaluation, the NUS system clearly outperforms the UBC system. While CombBest does not increase the retrieval performance measured by MAP (0.2921), we were able to increase the MAP value using the CombBest$^+$ index up to 0.3551.
In the remainder of this paper, we use the indices CombBest and CombBest$^+$ as our intention was to analyze the performance of the heuristic-based combination approach. Each index consists of three fields: token, lemma, and sense. The runs that we officially submitted are based on the CombBest index only.

3.2 Retrieval Experiments

We report experiments applying the probabilistic retrieval model DFR_BM25 (with and without query expansion), and the monolingual translation-based model on both the training and test data. The translation-based model is always restricted to the indexed tokens; the probabilistic model can use all different fields. We did not perform any fine-tuning on the parameters.
Table 2 shows the MAP of the different models. For the training data the DFR_BM25 model on tokens outperforms the translation model approach, even without any query expansion. The translation-based model shows a MAP value of 0.3045, while the DFR_BM25 model achieves a MAP value of 0.3374 without and 0.3760 with query expansion. Retrieval on lemmas even increases the MAP value further to 0.3829. Retrieval on senses shows the lowest MAP values ranging from 0.2557 up to 0.3011. Applying query expansion on the CombBest$^+$ index outperforms the according runs on the CombBest index.
For the test data, the translation model and the DFR_BM25 model without any query expansion show similar MAP values. However, when applying query expansion the DFR_BM25 approach outperforms the translation-based model.
An interesting aspect is that the difference between the performance on lemmas compared to tokens is much higher on the test topics than on the training topics. The DFR_BM25 model with query expansion on tokens yields a MAP value of 0.4223 while we get a MAP value of 0.4451 on lemmas, which is an improvement of 5.1%. Again, experiments on senses achieve the lowest performance. Again, retrieval on the CombBest$^+$ index performs better than on the CombBest index.
For the probabilistic model, we additionally conducted several experiments applying multi field queries. We have submitted one run querying the token, lemma, and sense field at the same time, which achieved a MAP value of 0.4380 on the CombBest and 0.4456 on the CombBest$^+$ index, respectively. However, as they do not have any significant different outcomes we do not report the figures here.

token		lemma	sense	MAP	
trans-lation	proba-bilistic			Comb Best	Comb Best$^+$
single retrieval models					
+	-	-	-	0.3616	
-	+	-	-	0.4223	
-	-	+	-	0.4451	
-	-	-	+	0.3313	0.3551
combinations without word senses					
-	0.5	0.5	-	0.4409*	
0.2	0.8	-	-	0.4316*	
0.2	-	0.8	-	**0.4509***	
0.2	0.4	0.4	-	0.4500*	
combinations with word senses					
-	0.8	-	0.2	0.4303	0.4327
-	-	0.8	0.2	0.4461	0.4473
-	0.4	0.4	0.2	0.4458*	0.4462
0.1	0.8	-	0.1	0.4330*	0.4331
0.1	-	0.8	0.1	0.**4500***	**0.4507**
0.1	0.4	0.4	0.1	0.4481*	0.4480

Table 3: MAP values and weights for the combination of different models, using the CombBest and CombBest$^+$ indices. The settings marked with a '*' were submitted.

3.3 Combination of Retrieval Models

We have manually analyzed the documents retrieved for some topics by the probabilistic and the translation model. We observed that the sets of retrieved documents by the two models are often different from each other. Therefore, we combined both models in order to improve the overall performance. We extensively experimented on the training data with different combination weights for the two retrieval models using the CombSUM method described in Section 2.4. The conclusion was that the combination achieves best performance when the probabilistic models based on tokens and lemmas were assigned a higher weight (due to their higher MAP values) than the model based on senses or the translation-based model. Table 3 illustrates the results obtained on the test topics by different combinations, with and without the integration of word senses. The weight combinations were determined during the training phase. They yielded best performance on the training data.

Two combinational aspects are of particular interest. The combination of the probabilistic models based on tokens and lemmas yields no improvement. In contrast, the combinations of the probabilistic model with the translation-based model always leads to an improvement. Even if the impact of the translation model, i.e. its weight, is low (here: 0.2), the MAP values increase when compared to the results obtained by the probabilistic model alone, on the token and lemma index fields. This fact corroborates our hypothesis that the probabilistic and the translation-based models retrieve different sets of relevant documents for some queries and that those different sets are effectively combined applying the CombSUM approach. The best performance could be obtained by a combination of the probabilistic model based on lemmas and the translation model based on tokens with weights 0.8 and 0.2, respectively. This combination yields a MAP value of 0.4509 and ended up with the 1st rank in the official challenge (see Section 3.4).

The second interesting aspect concerns the integration of word sense information. Retrieval based on senses from the CombBest index yields a MAP of 0.3313, while retrieval based on senses of the CombBest$^+$ index shows a MAP of 0.3551. We attribute the difference to the fact that CombBest looses information about the documents due to the smaller amount of indexed senses. However,

all combinations either with the CombBest or the CombBest$^+$ senses end up with a very similar performance. The reason could be that the loss of information when using the CombBest index is compensated by querying the tokens or lemmas as well.
In some combinational variations, the integration of word senses could achieve a higher MAP value than retrieval settings without word senses. For example, the MAP value corresponding to the retrieval based on tokens alone is 0.4223, while the combination with senses obtains a MAP value of 0.4303 for the CombBest index and even 0.4327 for the CombBest$^+$ index. For the combinations based on lemmas and senses, the difference is not significant. Overall, the best performance is obtained by the combination of the translation model and the probabilistic model based on lemmas and senses, applying weights of 0.1, 0.8, and 0.1, respectively. For the officially submitted run on the CombBest index a MAP value of 0.4500 was achieved, while the run on the CombBest$^+$ index achieves a slightly better MAP value of 0.4507.

3.4 Discussion

In the previous section we described all our experiments carried out on the document collection disambiguated with word senses. We submitted five runs without the integration of word senses and five further runs utilizing the annotated word senses. According to the MAP values our runs without word senses ended up in the 1st place out of 10 participants. Our highest MAP value could be achieved with the combination of the translation-based and the probabilistic model based on lemmas and the assigned weights of 0.2 and 0.8, respectively.
When utilizing word senses, the combination of the translation model based on tokens and the probabilistic model based on both lemmas and senses obtains the 1st place according to the MAP in the official challenge. We mistakingly submitted runs on the CombBest index, even though we planned to focus on the CombBest$^+$ index. However, we have shown that the differences between the combinational approaches are minimal. Our best performing submitted retrieval setting achieved a MAP value of 0.4500, whereas the second top scoring system in the official challenge obtains a MAP value of 0.4346.
Overall, we could not observe any significant improvement by applying the sense annotations compared to the retrieval settings based on tokens or lemmas only. This observation is consistent with the conclusion of last years' challenge. Participants of last years' challenge proposed several different methods for utilizing word senses, but could not achieve a significant improvement. We increased the precision of WSD annotations through a heuristic-based combination of the UBC and NUS annotated senses, which we evaluated extrinsically. This evaluation has shown that the accuracy of annotated word senses highly influences the outcome of retrieval systems utilizing these disambiguated data (see Table 1).
Regarding the performance of the translation-based model, the results on the combination is promising given that we merely applied a translation model built for a previous application in the field of answer finding. The main drawback of the straightforward use is the discrepancy in the tokenization scheme. The tokenization of the document collection is not always compatible with the tokenization of the parallel corpora used for training the translation model. In addition, the translation model we used contains only tokens and thus cannot deal with indexed multi word expressions. For instance, the phrase "public transport" is indexed as "public_transport". In the translation model the two terms "public" and "transport" appear, but not the phrase "public_transport". We quickly analyzed the amount of multi word expressions in the test topic collection. In fact, 61 queries out of the 160 test queries contain at least one multi word expression. This analysis shows that the translation model was not particularly trained for this task and motivates further improvements. In addition, further translation models could be trained on lemmas and senses. The latter option, however, requires a word sense disambiguated monolingual parallel corpus.

4 Conclusions

We have described a combinational approach to information retrieval on word sense disambiguated data, which combines a probabilistic and a monolingual translation-based model. For the probabilistic model we have used the Divergence From Randomness (DFR) BM25 model with the Kullback-Leibler Divergence as the query expansion method. For the translation-based model we have applied a model which was already trained for an answer finding task.
Our aim was to assess the benefits of the combination of both models. We have shown that the combinational approach always achieves better performance than the stand-alone models. Our second goal was to analyse different methods for selecting word senses from annotated corpora in order to increase their accuracy. We have discovered that our heuristic-based approach CombBest$^+$ increases the retrieval performance based on word senses by 2.2% when compared to NUSBest and even 25.8% when compared to UBCBest. The huge difference between NUSBest and UBCBest demonstrates that WSD accuracy is essential for utilizing word sense information. However, the experiments on the CombBest$^+$ index have shown that we could only increase the retrieval performance in one specific case: by combining the probabilistic model based on tokens with the same model based on senses. Nevertheless, other combinations without word senses outperformed this setting easily.
In conformance with our results, the best run out of all participants of last years' challenge was conducted without any word sense annotations. However, some participants were able to improve the performance of their own retrieval system slightly when utilizing word sense annotations, but it is questionable whether the improvements were significant. The UniNE group [6] has manually analyzed fifty queries of the test queries provided last year and figured out that some disambiguations were incorrect. Therefore, a manual evaluation of the current test queries would be of particular interest, which we leave for future work.
In summary, we agree with Sanderson [13] that first of all the accuracy of annotated word senses has to increase in order to improve the performance of retrieval based on word sense annotations.

5 Acknowledgements

This work has been supported by the Emmy Noether Program of the German Research Foundation (DFG) under the grant No. GU 798/3-1, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

References

[1] Eneko Agirre and Oier Lopez de Lacalle. UBC-ALM: Combining k-NN with SVD for WSD. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 342–345, Prague, Czech Republic, June 2007.

[2] Adam Berger and John Lafferty. Information Retrieval as Statistical Translation. In *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 222–229, 1999.

[3] Delphine Bernhard and Iryna Gurevych. Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 728–736, Suntec, Singapore, August 2009.

[4] Yee Seng Chan, Hwee Tou Ng, and Zhi Zhong. NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 253–256, Prague, Czech Republic, June 2007.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience, New York, NY, USA, 1991.

[6] Ljiljana Dolamic, Claire Fautsch, and Jacques Savoy. UniNE at CLEF 2008: TEL, Persian and Robust IR. In *Working Notes for the CLEF 2008 Workshop 17-19 September 2008*, Aarhus, Denmark, September 2008.

[7] Edward A. Fox and Joseph A. Shaw. Combination of Multiple Searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

[8] Djoerd Hiemstra. Term-specific Smoothing for the Language Modeling Approach to Information Retrieval: the Importance of a Query Term. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 35–41, New York, NY, USA, 2002.

[9] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[10] Vassilis Plachouras, Ben He, and Iadh Ounis. University of Glasgow at TREC2004: Experiments in Web, Robust and Terabyte tracks with Terrier. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, 2004.

[11] Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June 2007.

[12] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne. Okapi at TREC-4. In *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, pages 73–96, 1995.

[13] Mark Sanderson. Word Sense Disambiguation and Information Retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 142–151, New York, NY, USA, 1994.

[14] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval Models for Question and Answer Archives. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 475–482, New York, NY, USA, 2008.