Leveraging Concepts and Semantic Relationships for Language Model Based Document Retrieval

Lynda Said Lhadj[1], Mohand Boughanem[2], and Karima Amrouche[1]

[1] National High School for Computer Science (ESI), Algiers, Algeria
{l_said_lhadj,k_amrouche}@esi.dz
[2] IRIT, 118 route de Narbonne, 31062 Toulouse Cedex 9, France
bougha@irit.fr

Abstract. During the last decades, many language models approaches have been proposed to alleviate the assumption of single term independency in documents. This assumption leads to two known problems in information retrieval, namely polysemy and synonymy. In this paper, we propose a new language model based on concepts, to answer the polysemy issue, and semantic dependencies, to handle the synonymy problem. Our purpose is to relax the independency constraint by representing documents and queries by their concepts instead of single words. We consider that a concept could be a single word, a frequent collocation in the corpus or an ontology entry. In addition, semantic dependencies between query and document concepts have been incorporated into our model using a semantic smoothing technique. This allows retrieving not only documents containing the same words with the query but also documents dealing with the same concepts. Experiments carried out on TREC collections showed that our model achieves significant results compared to a strong single term based model, namely uni-gram language model.

Keywords: Information Retrieval, Language Modeling, semantic smoothing, Concept, Semantic Relatonships.

1 Introduction

Language models (LM) have shown so far more significant performances compared to some traditional Information Retrieval (IR) models, such as probabilistic and vector space model [6, 21]. This is particularly due to their simplicity and well-founded theoretical setting relying on probabilities. In fact, the relevance score of a document (D) to a query (Q) is simply given by the conditional probability $P(Q|D)$ [20, 13]. Several works have been proposed to estimate $P(Q|D)$ [6, 20, 13]. Most of them are based on the assumption that words are independent from each other. Such an assumption is in contrast with natural language where terms are related to each other. Thereby the two long-standing problems in IR, namely the synonymy and the polysemy phenomena are still arising in language models. To address these problems, a new generation of language modeling approaches based on n-grams or concepts has been developed [1, 9, 19, 23].

Y. Ait Ameur et al. (Eds.): MEDI 2014, LNCS 8748, pp. 100–112, 2014.

In this generation, two main underlying categories can be distinguished. The first one attempts to capture term dependencies directly from texts using statistical methods or learning techniques [1, 9, 20, 17, 18]. The second category use external semantic resources, such as ontologies to capture terms dependencies [3, 5, 8, 19].

The model presented in this paper is in the cross-road of both categories. It is intended as a novel language model that allows matching documents and queries at concept level to handle the polysemy issue. We assume that a concept can be a single word or a frequent collocation in the text. The latter can be either ontology entries or not. In addition, we exploit a semantic smoothing technique [6] to integrate semantic relationships between concepts in our retrieval model. This means that the model is capable to retrieve documents that contain not only the same concepts as a query but also those containing related concepts, such as synonyms, hypernyms, hyponyms.

The rest of this paper is organized as follows: Section 2 describes the general language model principle. Section 3 highlights previous LM approaches dealing with the issue of word independency assumption. In Section 4 we present our document retrieval model. Finally, we describe the experiment and the results in section 5. Section 6 summarizes the contribution and suggests some perspectives.

2 Related Work

The main idea behind LM is to assume that a document is generated by a statistical model. The relevance of a query with respect to a document is given by the conditional probability of the query to be generated by the language model of the document D. Therefore, the score of relevance is so given by the formula 1:

$$Score(Q, D) = P(Q|D) \tag{1}$$

In order to estimate $Score(Q, D)$, document words are assumed to be independent. This assumption has been widely adopted in IR, specifically in probabilistic models. Obviously, it simplifies the model estimation but there is a contradiction with the reality of natural language. Thus, two problems arise from this assumption. First, the problem of synonymy, for example, given a query containing the word "car", documents containing synonyms or related words as "automobile" or "vehicle" would not be retrieved even though they deal with the query concept. Second, the problem of polysemy, for queries containing ambiguous (polysemic) word such as "java" (programming language, island), irrelevant documents containing the same word but with a different meaning could be returned. These issues have been widely discussed in Information Retrieval. Particularly, in language models, most of works has attempted to address these issues by introducing some term dependencies into language models. According to [8], there are two kinds of dependencies can be considered, the ones within query words or within document ones, for example bi-grams, bi-terms or concepts. The intuition is that the combinations of words are less ambiguous than single words[16, 1, 17]. The second kind concerns dependencies between query words and document ones,

they are generally semantic such as synonyms [6, 8]. Both mentioned kinds are helpful for IR and many approaches have been proposed to integrate them into language models. They can be classified into tow categories : 1) the one capturing dependencies extracted directly within text and the approach and 2) the approaches integrating dependencies extracted from external resources

2.1 Integrating Dependencies from Text

In this approach, terms dependencies are captured by mean of different statistical techniques such as, word collocations. The earliest work have been proposed by Song and Croft [16]. They extended the uni-gram (single word) model to the bi-grams (sequences of two ordered words) one. The results were not successful since bi-grams cannot cover all words dependencies; In addition, bi-grams introduce noise because of adjacency constraint [9]. Srikanth and Srihari [17] proposed a bi-term language model to relax the constraints of term order and term adjacency in the bi-gram model. In their work presented in [18], authors focus on a higher level of dependencies in queries: a query is seen as a sequence of concepts identified using a syntactic parser. Each concept is a sequence of words co-occurring in the documents. The performance of their concept based uni-gram language model has been consistently better than the bi-grams model and bi-terms models. However none relation between query concepts (words) and documents concepts (words) have been exploited. For their part, Hammache and al. [10] proposed to combine single words and filtered bi-grams into a language model. In their approach bi-grams are selected and weighted by considering both their own occurrence in the document and the occurrence of their component terms. The results of these approaches are better than the single word model and some state of the art models. In the same purpose, Gao and al. [9] modelled dependencies between query terms pairwise as a hidden linked variable. The latter is undirected acyclic graph which express the most distant and robust dependencies among query words. The results have shown that the incorporation of this type of dependency has a positive impact on retrieval performance. Moreover, the dependence model outperforms the uni-grams, bi-grams and the co-occurrence ones. Results of these approaches are mixed: the bi-gram language model has shown lower results than the uni-gram model. Nevertheless, this model has been further enhanced by relaxing the constraint order and adjacency in bi-grams or by considering more distant relations [17, 18, 9]. However there are implicit and important dependencies such as synonymy or any semantic relation which are not captured.

2.2 Integrating Dependencies Extracted from External Resources

A number of extension of the LM approach have attempted improve retrieval performance using semantic information a priori defined in external resources such as WordNet[8, 3], UMLS [23] and Wikipedia[**?**]. One of the main work in LM framework which incorporates relationships between query words and

document ones has been proposed by Cao and al. [8], they have proposed a language model framework which incorporates relationships between query words and document words. Accordingly, a relationship can be identified in two ways: direct connection when words are identical and/or indirect connection through co-occurrences and WordNet relations. The relationships between words are exploited in the smoothing ...of the proposed language model. Their results were better than the uni-gram model. In the same spirit, Bao and al [3] proposed a Language Sense Model (LSM) based on WordNet senses. They have also used the linear interpolation in order to integrate hyponyms and hypernyms. The experiments did not highlight strong conclusions [3]. However the combination of terms and their respective senses have improved retrieval effectiveness especially for long queries (having more than 20 words). Other works [23] proposed a semantic smoothing of the document model using topic signatures. The latter corresponds to sets of synonyms, senses, contextual information and collocations extracted from documents. For this purpose, MaxMatcher[1] and XTRACT[2] have been used. The model was tested on a domain collection (TREC Genomic) and results were significant. Xinhui and al. [19] used Wikipedia title articles as topic signature with the same smoothing model of [23] and results were also successful.

In this paper, we propose to go beyond single words representation by assuming that document and queries are represented by mean of concepts which may be a single word, or frequent word collocations. The latter might be an ontology entry. Indeed, documents as well as queries should contain potential concepts which are not available in the ontologies, such as neologisms or proper names. Our definition of a concept joins the one described in [5]. Our belief is that a robust IR model should take into account all these elements and integrates them in a weighting schema proportionally to their importance in the text. Unlike the models proposed in [8, 3], both kinds of dependencies mentioned previously are integrated into our model, namely : 1) dependencies within the document and ones within throughout frequent collocation and ontology concepts to answer the polysemy issue. Indeed, multi-words are less ambiguous than single words. 2) Higher level dependencies are integrated in LM throughout relationships between query concepts and document ones. There are two intuitions behind integrating concepts relationships into the language model: the first one is to retrieve relevant document containing the same concepts whereas they are written with different words. The second one is to increase the concept weight with its related concepts such as hyponyms, hypernyms. The translation model of Berger and Lafferty [6], also known as a semantic smoothing model, seems to be the well adapted one to take into account our intuitions. Accordingly, the centrality of a concept, viewed as an important factor of relevance [7], is taken into account in a retrieval model.

[1] UMLS concepts extraction tool.
[2] Collocation extraction tool.

3 Concept Based Language Model

Let us consider query Q, document D and Ontology O where query Q and document D are respectively represented by concepts based modeling: $Q = c_1, c_2, ...c_m$ and $D = c_1, c_2, ...c_n$, where c_j is a concept which may be a single word or a multi words which corresponds also to an ontology entry or to a frequent collocation in the document collection. The relevance score RSV(Q,D) of a document to a query is estimated as mentioned in the formula 2.

$$RSV(Q, D) = P(Q|D)) = \prod_{i}^{n} P(c_i|D) \tag{2}$$

Formula 2 can be considered as an abstraction of the Ponte and Croft unigram model [13], where $P(c_i|D)$ is the probability of concept c_i in document D expressed as:

$$P(c_i|D) = \begin{cases} P(c_i|D, \bar{O}) & if c_i \notin O \\ \\ P(c_i|D, O) & otherwise \end{cases} \tag{3}$$

For a better clarity, probability $P(c_i|D)$ can be reformulated as:

$$P(c_i|D) = P(c_i|D, \bar{O}) + P(c_i|D, O) \tag{4}$$

$P(c_i|D, \bar{O})$ corresponds to the probability of c_i in D given the information that c_i is ont an ontology entry.
$P(c_i|D, O)$ is the probability of c_i an ontology entry in the document model.
These probabilities are complementary since events O and $\bar{O}$ are complementary.
We assume that a concept c_i is an ontological entry, its probability is estimated by taking into account its related concepts in the ontology and in the document. Therefore, it should model the fact that c_i should be seen effectively in the document model or represented by its related concepts.

$$P(c_i|D, O) = \sum_{c_j \in D} P(c_i|c_j) \;\; P_{sem}(c_j|D)) \tag{5}$$

This formulation is based on the translation model[6]. Our aim, here, is to integrate semantic relationships between query and document concepts. This can be seen as "Semantic Smoothing". Thus, the weight of query concepts are enhanced with those of related concepts (Synonyms, hypernyms or hyponyms). Therefore, estimating $P(c_i|D, O)$ in this way highlights the centrality of query concepts in the documents. The centrality of a concept c_i is defined by the number of its related concepts in D [7]. Notice that in our model, the centrality is implicitly taken into account and corresponds to the number sequences of the sum in formula (5).
The global score of relevance is expressed by the following formula:

$$P(Q|D) = \prod_{c_i \in Q} \left[P(c_i|D, \bar{O}) + \sum_{c_j \in Q} P(c_i|c_j, O) \;\; P_{sem}(c_j|D) \right] \tag{6}$$

where the three probabilities $P(c_i|D,\bar{O})$,$P_{sem}(c_j|D,O)$ and $P(c_i|c_j,O)$ are described below

The Probability $P(c_i|D,\bar{O})$ of c_i Given D
In the case where a query concept c_i is not in the ontology, its probability in D is given by using the Dirichlet prior smoothing.

$$P_{Dir}(c_i|D,\bar{O}) = \frac{count(c_i,D) + \mu P_{ML}(c_i|C)}{\sum_{c_k count(c_k,D)} + \mu}$$

where $Count(c_i,D)$ is c_i frequency in the document D, μ is the Dirichlet smoothing parameter and $P_{ML}(c_i|C)$ corresponds to the background collection language model by the maximum likelihood estimator as follows:

$$P_{ML}(c_i|C) = \frac{count(c_i,C)}{\sum_{c_k} count(c_k,C)} \tag{7}$$

The Semantic Probability $P_{sem}(c_j|D,O)$ of c_j Given D
This probability is called "semantic probability" since it estimates the likelihood of the ontological entry c_j and the one of its component sub concepts (sc) corresponding to WordNet entries. The intuition is the following : usually, authors tend to use sub-concepts to refer to the multi-term concepts that they have previously used in the same document[4]. For example, in TREC documents, the concept "coup" occurs after the multi-word concept "military coup d'etat" used more than once. Therefore, the component sub-concept "coup" is very likely to refer to "Military coup d'etat" than to another one. Therefore, $P_{sem}(c_j|D,O)$ is expressed as follow:

$$P_{sem}(c_j|D,O) = \theta P(c_j|D) + (1-\theta) \sum_{sc \in sub_C oncepts(c_j)} \frac{length(sc)}{length(c_j)} P(sc|D) \tag{8}$$

where $length(sc)$ is the size of sc in words, Sc is a sub-concept of c_j that corresponds to an ontology entry. $\theta is \in [0,1]$. The probabilities $P_{ML}(c_j|D)$ and $P_{ML}(sc|D)$ are respectively estimated using the Dirichlet smoothing like formula (7). We notice that formula (8) is equivalent to the CF-IDF weighting formula proposed by Baziz and al. [**?**]

The Probability $P(c_i|c_j,O)$ of c_i Given c_j
This probability estimates the relationship degree between concepts c_i and c_j. Several ways have been proposed to estimate the probability of a term given another. In general, they are based on relationships such as co-occurrences, mutual information, fuzzy relationships [1], or ontological relationships like synonymy, hypernymy or hyponymy [8, 23]. For our part, we estimate $P(c_i|c_j,O)$ using the degree of relation between c_i and c_j in the ontology compared to the whole relationship degrees between query concepts and document ones. Thus, $P(c_i|c_j,O)$ est estimated as follow

$$P(c_i|c_j, O) = \begin{cases} \dfrac{Rel(c_i, c_j)}{\sum_{c_k \in Q} Rel(c_k, c_j)} & if \quad c_i \neq c_j \\ \\ 1 \qquad\qquad otherwise \end{cases} \tag{9}$$

where $Rel(c_i, c_j)$ is relationship degree between concepts c_i and c_j, estimated using Resnik Semantic Similarity [14] as:

$$Rel(c_i, c_j) = sim_{res^*}(c_i, c_j)$$

Resnik Similarity is based on the is-a relationship and Information Content (IC) metric proposed in Seco and al.[15].

$$sim_{res^*}(c_i, c_j) = max_{c \in S(c_i, c_j)} IC_{wn}(c))$$

$S(c_i, c_j)$ is the set of concepts that subsumes c_i and c_j. The Information Content IC_{wn} is based on the following principle: the more a concept has descendants, the less is its Information Content. Therefore, concepts that are leaves (specific concepts) have more IC than ones situated up in the hierarchy. In fact, Resnik Similarity based on IC highlights specificity defined in [7].

$$IC_{wn}(c) = 1 - \frac{\log(hypo(c)) + 1}{\log(max_{wn})}$$

where $hypo(c)$ returns the hyponyms number of concept c and max_{wn} is a constant corresponding generally to the maximal number of concepts in the taxonomy. In the version 2.1 of WordNet, $max_{wn} = 117659$.

4 Experimental Evaluation

To evaluate the effectiveness of our retrieval model, we used two datasets issued from TREC [3] collections and WordNet 2.0 [11] a linguistic ontology to detect ontological concepts and relationships. We carried out a threefold objective-based experiment:

a) Evaluating the impact of combining ontological concepts and the non-ontological ones in a language modeling approach.
b) Highlighting the impact of integrating concept relationships (hyponymy, hyponymy) in the retrieval model.
c) Comparing our model with a strong single word language model, namely the uni-gram model.

[3] trec.nist.gov

4.1 Datasets and Experimental Setting

As mentioned above, experiments were carried out on two datasets of TREC (issued from disk 1 & 2) namely WSJ 86-87 (Wall Street Journal) and AP 89 (Associated Press News). Each dataset is a collection of news. In addition, a set of topics constituting the queries and the relevance judgement are provided for each dataset.

Document Processing
In our approach each document in both datasets is processed using the following approach:

a) Terms and multi-terms of different sizes are detected using Text-NSP tool [2] and saved in a list.
b) Detected terms of that list are then processed to remove all terms beginning or ending with a stop word. Terrier [12] stop word list is used for this purpose. We underline that unlike almost related work, we avoid pretreatment of the text before detecting multi-terms in order to keep potential concepts such as "Ministry of justice", "Bureau of investigation "[4]. This type of concepts is called "complex phrases"[22] and are frequently monosemic.
c) For validating that a given multi-term concept, we only keep those occurring at least twice.
d) We check whether a concept occur or not in WordNet. For those having an entry, they are selected and represented by their Synset Number. For instance, the concepts "coup", "coup d'etat", "takeover" and "putch" are grouped in the synset with number 01147528.
e) When a given concept has several entries (polysemy) in WordNet, the default first sense is selected.
f) The remainder of concepts is kept as the non WordNet entries and weighted with simple count of occurrences.

TREC Topics have been used as queries. Each topic is composed of three parts: Title part which is a short query, Description part, which is a long (verbose) query and Narrative part, describing in detail previous parts and precisely what relevant documents should contain. In this evaluation, we used the Title part of topics as queries since they are as short as user queries.
The value of Dirichlet prior parameter μ is set to 2500 for all datasets and models.

4.2 Results and Evaluation

In this section, we present the results obtained throughout our experimental evaluation. For this purpose, we used the following metrics $P@x$ precision at the point $x \in \{10, 20\}$ [5] and the MAP.

[4] These examples are taken from TREC documents.
[5] P@x, is the ratio evaluating the number of relevant document at the top x retrieved documents and the Mean average precision

Analysing the impact of combining WordNet terms and non WordNet terms. We aim here at evaluating the impact of combining terms (single words, phrases non corresponding to WordNet entries, and those belonging to WordNet). First, we test the retrieval model corresponding to formula (2) using all detected collocations and without filtering WordNet concepts. Thus $P(c_i|D)$ is estimated using Dirichlet smoothing (see formula (7)). Second, the model was tested by considering only WordNet concepts and finally the combined model of WordNet concepts and frequent collocations (the whole fromula (6)). Fig.1 illustrates changes in MAP for the three tests mentioned above. It can be clearly seen

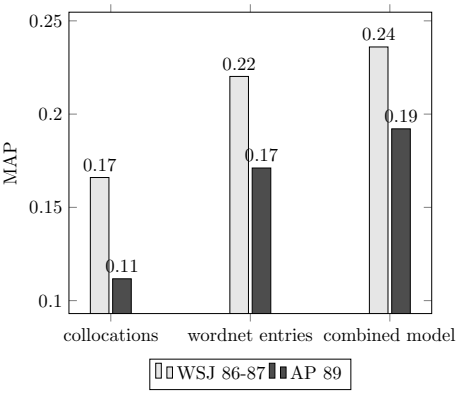

Fig. 1. The variation of MAP

in Fig.1 that the combined model outperforms the two other models. It achieves a MAP value of $0,24$ for WSJ 86-87 dataset and $0,19$ for AP 89 dataset. Accordingly, the combined model noted CLM_0 is used in the remaining experiments and compared to the uni-gram model noted ULM.

Table.1 compares the obtained results of the evaluation of CLM_0 and ULM. Gain (%) line denotes the percentage of improvement regarding the ULM. The reported precisions show that CLM_0 generally outperforms ULM with significant improvement equal to the value of $+3,21\%$ for WSJ 86-87 dataset. For AP 89 dataset, the improvement is not as significant as for WSJ 86-87. Nevertheless, at $P@10$ and $P@20$, the improvements are statistically significant with the p-value is lower to 0,5 according to t-test.

Indeed, we have performed a more in-depth analysis on some queries such as " Military Coup d'Etat" (topic-62). We observed that ULM achieved an Average Precision (AP) of $0,1446$ while our model CLM_0 reached an AP of $0,3956$. To show the reason of this improvement in AP for that query, we have checked how both models have ranked some relevant documents[6]. For example, document "WSJ870526-0068" was promoted from rank 114 with ULM to rank 22 with

[6] The list of relevant documents is taken from Relevance judgement file.

Table 1. Comparison between performances of the uni-gram model(ULM) and the combined concept based model CLM_0. Test significance : + for p-value < 0.05 and ++ p-value < 0.01.

Collection	Evaluated model	Performance evaluation		
		$P@10$	$P@20$	MAP
WSJ (86-87)	ULM	$0,3280$	$0,2790$	$0,2302$
	CLM_0	$0,3640$	$0,3030$	$0,2376$
	Gain (%)	$+10,98^{++}$	$+8,60^{++}$	$+3,21^{+}$
AP (89)	ULM	$0,3160$	$0,2810$	$0,1924$
	CLM_0	$0,3280$	$0,2910$	$0,1925$
	Gain (%)	$+3,80^{+}$	$+3,56^{+}$	$+0,05$

our model. When examining manually the document content, we found that it does not contain exactly the concept "coup dtat", but synonyms such as "coup" and "takeover occur respectively 7 and 6 times. The same has been observed for query "Iran-contra affair" (Topic 99) for which our model reached an AP value of $0,2832$ and ULM achieved an AP of 0.0069. This enhancement in AP is due to the fact that the query itself is a frequent collocation in WSJ 86-87 dataset. However, for AP 89 dataset, the noticed enhancement is less important. However, both models perform nearly equally for AP 89 dataset. They achieved respectively AP values of 0.3212 and 0.3209. This result is due to the fact that "Iran contra affair" and "Iran-contra" are frequent collocations in WSJ 86-87 dataset in contrast with AP 89 dataset. This highlights the importance of considering various concepts (single words, collocations and WordNet entries) in the retrieval model.

Analysing the Impact of Incorporating Concepts Relations. We perform a further analysis evaluation in order to show the impact of integrating concepts relations into the retrieval model. This evaluation concerns the model named CLM_1 expressed in the formula 6 where we only integrated "is a" relationships namely the hypernymy and hyponymy.

We can see in table2 that for WSJ dataset, we notice that CLM_1 overpasses generally CLM_0 and the ULM. As of AP 89 collection, the improvements are not as significant as those achieved for WSJ 86-87 collection but they are statistically significant on P@10 and P@20. Indeed, t-test shows that our results are significant since the p-value $< 0,01$.

We performed the same analysis as the previous experiment with the same query (topic-62). We notice that ULM, CLM_0 and CLM_1 achieve respectively average precisions with value of $0,1446$,$0,2911$ and $0,4342$. The analysis of a relevant document, for example WSJ870526-0068, showed that it is promoted from ranks 8 and 6 under ULM and CLM_0 respectively to the rank 2 because it contains a direct hypernym of "military" which is "forces" and it occurs 4 times

Table 2. Comparison between uni-gram model ULM and concept based models CLM_1. Test significance : + for p-value < 0.05 and ++ for p-value < 0.01 .

Collection	Evaluated model	Performance evaluation		
		$P@10$	$P@20$	MAP
WSJ (86-87)	ULM	$0,3280$	$0,2790$	$0,2302$
	CLM_0	$0,3640$	$0,3030$	$0,2376$
	Gain (%)	$+10,976^{++}$	$+8,60^{++}$	$+3,21^{+}$
	CLM_1	$0,3642$	$0,3266$	$0,2380$
	Gain (%)	$+11,04^{++}$	$+17,06^{++}$	$3,39^{++}$
AP (89)	ULM	$0,3160$	$0,2810$	$0,1924$
	CLM_0	$0,3280$	$0,2910$	$0,1925$
	Gain (%)	$+3,80^{+}$	$+3,56^{+}$	$+0,05$
	CLM_1	$0,33140$	$0,2900$	$0,1932$
	Gain (%)	$+8,23^{++}$	$+4,63^{+}$	$+0,51$

in the document. So the probability of "military" is boosted with that of its related concept "forces". These statements lead us to conclude that integrating concepts, semantic relations in the retrieval model enhanced document retrieval.

5 Conclusion and Future Work

In this paper we introduced a concept-based language model for enhancing document retrieval. The intuition is to build a rich document representation through single words, ontological concepts, their relationships available in an ontology and collocations which are not ontological concepts. The latter can be either proper names or neologisms. Moreover, through integrating relationships between query and document concepts, our model allows to take into account of related concepts to those of the query. This is carried out through smoothing part of the proposed language model. The empirical results on TREC collections show that our model outperforms the uni-gram one. Indeed, this highlights the effectiveness of combining statistical collocations and WordNet concepts and their relationships namely "is-a" relation in a language modeling approach. These results are also encouraging to mix further evidence sources of concepts to estimate richer and more precise document model. Our model could be further improved by integrating additional NLP rules for filtering collocations and other resources such as Wikipedia. We also plan to test the impact of other semantic relationships on retrieval.

References

[1] Bai, J., Song, D., Bruza, P., Nie, J.Y., Cao, G.: Query expansion using term relationships in language models for information retrieval. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM 2005, pp. 688–695. ACM (2005)

[2] Banerjee, S., Pedersen, T.: The design, implementation, and use of the ngram statistics package. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 370–381. Springer, Heidelberg (2003)

[3] Bao, S., Zhang, L., Chen, E., Long, M., Li, R., Yu, Y.: LSM: Language sense model for information retrieval. In: Yu, J.X., Kitsuregawa, M., Leong, H.-V. (eds.) WAIM 2006. LNCS, vol. 4016, pp. 97–108. Springer, Heidelberg (2006)

[4] Baziz, M., Boughanem, M., Passi, G., Prade, H.: An information retrieval driven by ontology from query to document expansion. In: Large Scale Semantic Access to Content (Text, Image, Video, and Sound), RIAO 2007, pp. 301–313 (2007)

[5] Bendersky, M., Croft, W.B.: Modeling higher-order term dependencies in information retrieval using query hypergraphs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012, pp. 941–950. ACM (2012)

[6] Berger, A., Lafferty, J.: Information retrieval as statistical translation. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, pp. 222–229. ACM (1999)

[7] Boughanem, M., Mallak, I., Prade, H.: A new factor for computing the relevance of a document to a query. In: Proceedings of the International Conference on Fuzzy Systems, pp. 1–6. IEEE (2010)

[8] Cao, G., Nie, J.Y., Bai, J.: Integrating word relationships into language models. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, pp. 298–305. ACM (2005)

[9] Gao, J., Nie, J.Y., Wu, G., Cao, G.: Dependence language model for information retrieval. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004, pp. 170–177. ACM (2004)

[10] Hammache, A., Boughanem, M., Ahmed Ouamar, R.: Combining compound and single terms under language model framework. In: Knowledge and Information Systems, pp. 329–349 (2013)

[11] Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM 38(11), 39–41 (1995)

[12] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 517–519. Springer, Heidelberg (2005)

[13] Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998, pp. 275–281. ACM (1998)

[14] Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, pp. 448–453. Morgan Kaufmann Publishers Inc. (1995)

[15] Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in wordnet. In: ECAI, vol. 4, pp. 1089–1090 (2004)

[16] Song, F., Croft, W.B.: A general language model for information retrieval. In: Proceedings of the Eighth International Conference on Information and Knowledge Management, CIKM 1999, pp. 316–321. ACM (1999)
[17] Srikanth, M., Srihari, R.: Biterm language models for document retrieval. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2002, pp. 425–426. ACM (2002)
[18] Srikanth, M., Srihari, R.: Incorporating query term dependencies in language models for document retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR 2003, pp. 405–406. ACM (2003)
[19] Tu, X., He, T., Chen, L., Luo, J., Zhang, M.: Wikipedia-based semantic smoothing for the language modeling approach to information retrieval. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 370–381. Springer, Heidelberg (2010)
[20] Victor, L., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 120–127. ACM (2001)
[21] Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 334–342. ACM (2001)
[22] Zhang, W., Liu, S., Yu, C., Sun, C., Liu, F., Meng, W.: Recognition and classification of noun phrases in queries for effective retrieval. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, pp. 711–720. ACM (2007)
[23] Zhou, X., Hu, X., Zhang, X.: Topic signature language models for ad hoc retrieval. IEEE Trans. on Knowl. and Data Eng. 19(9), 1276–1287 (2007)