Official URL

DOI : https://doi.org/10.1002/asi.23553

Enhancing Information Retrieval Through Concept-Based Language Modeling and Semantic Smoothing

Lynda Said Lhadj
Ecole nationale Supérieure d'Informatique ESI, P.O. Box 68 M, 16309 Algiers, Algeria. E-mail: l_said_lhadj@esi.dz

Mohand Boughanem
Institut de Recherche en Informatique de Toulouse IRIT, Université Paul Sabatier, Toulouse 31062, France. E-mail: bougha@irit.fr

Karima Amrouche
Ecole nationale Supérieure d'Informatique ESI, P.O. Box 68 M, 16309 Algiers Algeria. E-mail: k_amrouche@esi.dz

Traditionally, many information retrieval models assume that terms occur in documents independently. Although these models have already shown good performance, the word independency assumption seems to be unrealistic from a natural language point of view, which considers that terms are related to each other. Therefore, such an assumption leads to two well-known problems in information retrieval (IR), namely, polysemy, or term mismatch, and synonymy. In language models, these issues have been addressed by considering dependencies such as bigrams, phrasal-concepts, or word relationships, but such models are estimated using simple n-grams or concept counting. In this paper, we address polysemy and synonymy mismatch with a concept-based language modeling approach that combines ontological concepts from external resources with frequently found collocations from the document collection. In addition, the concept-based model is enriched with subconcepts and semantic relationships through a semantic smoothing technique so as to perform semantic matching. Experiments carried out on TREC collections show that our model achieves significant improvements over a single word-based model and the Markov Random Field model (using a Markov classifier).

Introduction

In most traditional information retrieval (IR) approaches, queries and documents are represented by single words or word stems (typically referred to as a bag of words representation). The relevance score of a document with respect to a query is estimated using the frequency distribution of query terms over the documents. The latter is carried out under the assumption that query words occur independently in documents (Salton, Buckley, & Yu, 1982). However, given the common knowledge about natural language, such an assumption might seem unrealistic (or even wrong), leading to the long-standing IR problems of synonymy and polysemy:

- The synonymy issue (also known as term mismatch issue) occurs when users and authors use different terms for the same concept (Gonzalo, Li, Moschitti, & Xu, 2014; Li & Xu, 2013; Wei, Hu, Tai, Huang, & Yang, 2007). More precisely, authors use a large vocabulary to express the same concepts while user queries are often a short or incomplete description. For example, given the query *"Auto Race,"* a document containing related concepts such as *"Automobile Race," "Grand Prix,"* or *"Rally"* would be not retrieved if both *"Auto"* and *"Race"* do not occur within this document.
- The polysemy issue concerns the ambiguity of single words, viz., that a word could have several meanings (Krovetz & Croft, 2000). For example, given a user query *"Dog Bark"* and a single word-based model, irrelevant documents dealing with *"The Bark of a Dog"* or *"The Bark of a Tree"* might be returned if the words *"Bark"* and *"Dog"* occur frequently therein.

These issues have been largely discussed in IR and have given rise to semantic IR approaches (Alvarez, Langlais, & Nie, 2004; Baziz, Boughanem, & Aussenac-Gilles, 2005;

Boughanem, Mallak, & Prade, 2010; Cao, Nie, & Bai, 2005; Li & Xu, 2013). Generally, these approaches rely on additional sources of evidence such as semantic resources (e.g., dictionaries or ontologies) or corpus features (e.g., term co-occurrence) to represent queries and documents with specific meaning of terms rather than single words. However, research developed so far has led to mixed results. We present the following questions about how to improve the mismatches that result from polysemy and synonymy between query term and corpus: (a) how to identify query and document terms denoting concepts and (b) how to use a concept-based model to enhance retrieval. Aiming to go beyond bag of word issues, we focus in this paper on language-based models (LMs) (Ponte & Croft, 1998). Indeed, during the two last decades LMs have attracted increased interest in the IR community mainly due to their reliance on probabilities. In addition, LMs have shown successful results over the other models such as probabilistic and vector space models (Bennett, Scholer, & Uitdenbogerd, 2007; Zhai, 2008). More particularly, in the context of semantic IR, a wide range of LM approaches have been proposed to relax the word independency assumption. Most approaches have focused on considering term dependencies such as n-grams, phrases, co-occurrence, or ontological concepts (Cao et al., 2005; Gao, Nie, Wu, & Cao, 2004; Hammache, Boughanem, & Ahmed-Ouamar, 2013; Srikanth & Srihari, 2002; Zhou, Hu, & Zhang, 2007) in an attempt to capture user query and document semantics. Two main types of approaches can be distinguished, namely, *corpus-based approaches* and *external resource-based approaches*. The former attempts to capture term dependencies within a corpus using statistical measures or learning techniques (Berger & Lafferty, 1999; Lavrenko & Croft, 2001) while the second category relies on external semantic resources (ontologies, encyclopedia) to recognize query and document concepts (Meij, Trieschnigg, de Rijke, & Kraaij, 2008; Tu, He, Chen, Luo, & Zhang, 2010; Zhou et al., 2007) or single-word relationships (Cao et al., 2005).

In this paper, we propose a novel concept-based language model to address the problem of word dependence. The proposed model is inspired by previous approaches (Baziz et al., 2005; Bendersky & Croft, 2012; Tu et al., 2010; Zakos, 2005; Zhou et al., 2007) which have proven the usefulness of concept-based representation in IR. More specifically, we assume that documents and queries are represented as a bag of concepts, instead of a bag of words, to meet the polysemy problem. Document and query concepts are identified using two sources of evidence: a semantic resource (an ontology) and the document collection. We view both sources as complementary since important concepts, such as neologisms or proper names, generally, are not found in semantic resources. We consider that such concepts might be frequent collocations in the document collection. Thus, a concept can be viewed as a single word, a frequent collocation, or an ontology entry. The idea of combining multiword phrases and ontological concepts has also been proposed (Zhou et al., 2007). However, the concept model they propose is based on counting only ontological concepts or frequently found multiword phrases. From our point of view, the estimation of concept-based models in IR can go beyond concept counting. Indeed, semantic information a priori defined in semantic resources such as subconcepts (component concepts corresponding to ontology entries) and semantic relatedness (e.g., hypernymy, hyponymy, etc.) could be integrated into the model through a semantic smoothing technique (Berger & Lafferty, 1999; Lafferty & Zhai, 2001; Zhai, 2008) to perform semantic matching. The intuition is that the authors tend to use concept relationships or subconcepts to avoid repetition or to refer to a concept they have previously used. In the same spirit, Cao et al. (2005) proposed an approach to incorporate individual word relationships into a language model framework without considering word relationships. In this work, we combine single words, frequent collocations, ontological concepts, subconcepts, and semantic relatedness into a unified language model so as to perform query and document matching at a semantic (concept) level. Accordingly, for a given query the retrieval model should be able to retrieve documents that contain both the same terms as those of the query and those that contain related concepts such as synonyms, hypernyms, and hyponyms. We evaluated various scenarios of our model on two TREC collections to demonstrate the effect of these elements on retrieval. Our results show significant improvements over state-of-the-art models: the unigram language model (Dirichlet) and the Markov Random Field model (Metzler & Croft, 2005).

The remainder of the paper is organized as follows. In the next section we describe the general language model principle and review previous works dealing with word independence. Concept-Based Retrieval Model follows and then Experimental Evaluation and Results performed on two TREC collections are presented. The final section summarizes the contributions and suggests some future research directions.

Related Work

Language models have steadily grown in popularity since their introduction in IR (Ponte & Croft, 1998). They have been successfully applied in various applications such as ad hoc retrieval (Ponte & Croft, 1998), expert finding (Macdonald & Ounis, 2008), and social retrieval (Zhou, Bian, Zheng, Lee, & Zha, 2008). Specifically, in the context of ad hoc retrieval, LMs have significant performance compared to some traditional IR models, such as probabilistic and vector space models (Bennett et al., 2007; Zhai, 2008). The main idea behind LM in IR is to view a document as a language sample generated according to some probability distribution of word sequences (Zhai, 2008). The relevance of a document D with respect to a query Q is seen as a probability of producing the query from the document language model. Several variations have been proposed to estimate this probability (Berger & Lafferty, 1999; Ponte & Croft, 1998; Song & Croft, 1999). The most common one is the query likelihood approach (Ponte & Croft, 1998) where the relevance score is given by the conditional probability P(Q|D) of Q to be generated by the underlying language model D expressed as follows:

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

To more easily estimate the above probability, query terms are assumed to occur independently. Thus, *D* is commonly estimated using the unigram model, so Equation 1 becomes the product of individual query term probabilities given the document model:

$$Score(Q, D) = \prod_{t_i \in Q} P(t_i|D) \tag{2}$$

Probability $P(t_i|D)$ is calculated using the maximum likelihood $P_{ML}(t_i|D)$ of individual terms estimated as follows:

$$P_{ML}(t_i|D) = \frac{f(t_i D)}{|D|}$$

where $f(t_i, D)$ is the frequency of term t_i within D, and $|D|$ is the total number of terms in that document.

According to this model, when a query term t_i does not occur in D, $P_{ML}(t_i|D)$ is zero, making the overall $Score(Q, D)$ zero, even if the remainder of query terms $\{t_j\}_{j \neq i}$ are seen in D. To cope with this problem, a number of smoothing techniques have been proposed (Chen & Goodman, 1996; Zhai & Lafferty, 2001). The general principle is to assign a nonzero probability to unseen query terms in documents and adjust low probabilities.

Using bag of word representations that assume word independence is prevalent within language model approaches. It is commonly accepted in IR that such an assumption is a matter of mathematical convenience because terms in natural language are often dependent. Therefore, much research goes beyond the bag of word representation. Earlier works captured dependencies between words by using units longer than single words such as bigrams and concepts (Song & Croft, 1999; Srikanth & Srihari, 2002; Zhou et al., 2007). The intuition is the following: the longer an indexing unit is than a single word, the less is its ambiguity (Shi & Nie, 2009). More advanced works have attempted to deal with the synonymy and polysemy issues using smoothing techniques, for instance, the translation model (Berger & Lafferty, 1999; Cao et al., 2005). These works rely on a specific resource, which may be either the document collection viewed as an internal resource in the IR process or a semantic resource viewed as an external resource. Accordingly, we classify these works into two categories of approaches depending on the source of word dependencies: (a) *corpus-based approaches* integrating dependencies extracted from a corpus such as bigrams or concepts defined on the basis of co-occurrence information and (b) *external resource-based approaches* enabling the capture of semantic dependencies from external resources (such as dictionaries, thesauri, or ontologies).

Corpus-Based Approaches

Most of these approaches have examined different ways to exploit text features, such as bigrams or longer collocations to relax the word independency assumption (Bai, Song, Bruza, Nie, & Cao, 2005; Berger & Lafferty, 1999; Gao et al., 2004; Petrovic, Snajder, Dalbelo-Basic, & Kolar, 2006). Generally, these dependencies are learned from the document collection. Berger and Lafferty (1999) proposed one of the earliest works in the LM framework that exploits dependencies between query and document words. Word dependencies are explicitly expressed through a translation probability estimating the degree of link between query and document words using word co-occurrence in a training corpus. Song and Croft (1999) integrated ordered and adjacent dependencies between pairwise words in the so-called bigram model. Empirical results were not successful for two reasons: (a) bigrams cannot cover all useful word dependencies, given that there are more distant dependencies than bigrams. Indeed, terms occurring in a specific context can be related even though they are not adjacent. In addition, the bigram-based model assigns higher probabilities to documents containing query bigrams (e.g., "*Information Retrieval*") than those where component terms "*Information*" and "*Retrieval*" occur separately; and (b) bigrams introduce noise in the retrieval process. For example, the query "*Computer-aided Crime*" (TREC topic 94) contains the following bigrams: "*Computer-aided*" and "*aided Crime*." For this query, documents containing "*Computer-aided Design*" with higher frequencies are likely to be top-ranked by the bigram model as relevant. To overcome these limitations, Srikanth and Srihari (2002) extended the bigram model to a biterm one. They define biterms as bigrams in which the constraint of word ordering is relaxed. Given a document containing the phrase "*Information Retrieval*" and another containing "*Retrieval of Information*," the biterm language model would assign the same probability of generating the query "*Information Retrieval*" for both documents, unlike the bigram model. The results of this approach were slightly better than the unigram and bigram models. In Srikanth and Srihari (2003), the authors extended their model and integrated higher-level dependencies in queries seen as a sequence of concepts that are themselves viewed as sequences of single words. These concepts are identified using a syntactic parser, and their probabilities are estimated using a smoothed bigram model. Experiments have shown that the concept-based unigram model provides better results than both bigram and biterm models (Song & Croft, 1999; Srikanth & Srihari, 2002). With a similar intuition, Metzler and Croft (2005) proposed a dependency-based framework using Markov Random Field (MRF). They exploited text features to detect longer dependencies between query terms. Thus, they examined three types of dependencies: the Full Independence (FI), where query terms are assumed to be independent of each other, the Sequential Dependence (SD), where query terms are neighboring, and the Full Dependence (FD), where query terms are dependent on each other. The results have shown that modeling such dependencies significantly improves retrieval effectiveness. In particular, the authors noticed that the SD model is more effective on smaller collections with longer queries, while the FI model is best for larger collections with shorter queries. Similarly, Shi and Nie (2009) proposed a phrase-based model in which phrases and single words are used to estimate a document model. Only phrases having a higher inseparability factor than their component words have been used (Shi & Nie, 2009). The inseparability of a phrase is defined as a discriminant factor based mainly on IDF (Inverse Document Frequency). The experimental results stressed the usefulness of integrating phrase's inseparability in the language model. Along these lines, Hammache et al. (2013) combined single words and filtered bigrams into a language model. They proposed an approach for filtering bigrams and weighting them by considering both their frequencies and their component terms in the document collection. Gao et al. (2004) considered only dependencies among pairs of terms. Thus, a query is modeled by an undirected acyclic graph that expresses the most distant and robust dependencies among query words.

The results of the above approaches show that exploiting dependencies extracted from the corpus may have a positive influence on retrieval effectiveness. However, the bigram language model has not shown consistent performance beyond the unigram model because many dependencies are incorrect and introduce noisy dependencies into the retrieval process.

Although the bigram model has been extended by relaxing order and adjacency constraints or by considering more distant dependencies (Bai et al., 2005; Gao et al., 2004; Srikanth & Srihari, 2003), the retrieval effectiveness can be further enhanced by considering implicit dependencies such as synonymy or any semantic relationship (e.g., hypernymy). More recently, a number of LM-based approaches have attempted to use semantic information a priori defined in external resources such as WordNet (Bao et al., 2006; Cao et al., 2005), Unified Medical Language System (UMLS) (Zhou et al., 2007), and Wikipedia (Tu et al., 2010). We give an overview of this category of models in the next section.

External Resource-Based Approaches

Most of the underlying approaches are built upon the translation model (Berger & Lafferty, 1999), which provides a straightforward way to incorporate a kind of semantic smoothing by mapping query and document terms (Berger & Lafferty, 1999; Lafferty & Zhai, 2001; Zhai, 2008). Ontological relationships have also been exploited to estimate dependencies between query and document terms. In particular, Cao et al. (2005) extended the translation model (Berger & Lafferty, 1999) by considering word relationships in order to match query and document terms at a semantic level. For this purpose, they assumed that dependencies between query and document words are generated through two sources: (a) direct connection (matching) using a unigram model and/or (b) indirect connection using a link model expressed throughout co-occurrences and WordNet relationships. These two sources are used to smooth the document model. This approach has shown consistent improvement over the unigram model on TREC collections, although word ambiguity had not been discussed. In contrast, Bao et al. (2006) proposed a Language Sense Model (LSM) where the single-word unigram model is smoothed with WordNet synsets (Miller, 1995). More precisely, the appropriate sense of each single word in both document and query is selected a priori using the Word Sense Disambiguation System they developed. Afterward, the document model is smoothed with hyponyms and hypernyms found in WordNet. However the experiments did not show strong conclusions except for long queries (having more than 20 words). Other works (Tu et al., 2010; Zhou et al., 2007) have followed another direction to exploit word dependencies. Typically, Zhou et al. (2007) proposed a semantic smoothing approach to enhance the document model and address the problem of word ambiguity using the so-called topic signatures. The latter is a set of synonyms, senses, and collocations (word pairs) corresponding to named entities. These topic signatures are recognized in documents using both MaxMatcher[1] and XTRACT.[2] Experiments on a domain collection (TREC Genomic[3]) have shown the effectiveness of such an approach, with significant improvement over the unigram model. With this aim, Tu et al. (2010) proposed a semantic smoothing approach based on Wikipedia. More precisely, they consider the titles of Wikipedia articles instead of topic signatures to smooth the document model proposed in Zhou et al. (2007). They have shown that a Wikipedia article title has the same syntactic structure (generally a nominal phrase) as an ontological concept. Meij, Trieschnigg, Kraaij, and de Rijke (2009) proposed generative concept models to improve the query model. More precisely, the query is first translated into a conceptual representation obtained using ontology and feedback documents (issued from an initial retrieval run). Then the conceptual query model is translated into a textual query model. The intuition is that the textual representation is more detailed than the conceptual representation. Thus, retrieving with a textual query representation translated from a conceptual form yields a better performance than a strict concept-based matching.

The approach we propose in this paper is at the confluence of both described categories. We propose a concept-based model by considering two sources of word dependencies: frequent collocations and ontology entries. A similar proposition has been described previously (Tu et al.,

[1]Extraction tool of UMLS concepts.
[2]Collocation extraction tool.
[3]trec.nist.gov.

<DOC>
<DOCNO> AP890125-0124 </DOCNO>
<TEXT>
Attorneys for fired National Security Council Aide Oliver L. North disclosed today they have issued subpoenas in the **Iran-contra** case to three house committees and two members of the house. North attorney Barry Simon saia tne subpoenas were sent to the house intelligence, foreign affairs and armed services committees. North's subpoenas require ``surrender of the broadest imaginable range of privileged documents,'' including ``the whole set of records'' of the house intelligence panel's investigation of the resupply operation for the **Nicaraguan contra** rebels, the brief said. The committee began the investigation after the October 1986 shoot down of a contra resupply plane in Nicaragua carrying American Eugene Hasenfus as one of its crew members. At the court hearing, Simon said the notebooks might be applied to the portion of the case alleging destruction of documents by North in November 1986 as the **Iran-contra** scandal was unfolding. Simon said prosecutors might try to link references in the Notebooks to documents that north allegedly destroyed before leaving The NSC. Gesell said he would decide by Friday whether north should be compelled to produce the notebooks, portions of which were turned over to congress for the **Iran-contra** hearings in 1987. Simon declined to acknowledge that the notebooks existed or were in North's possession, but he said North is entitled to Fifth Amendment protection against self-incrimination. The notebooks also delve into North's assistance to the **Nicaraguan** rebels and his involvement in the Reagan Administration's secret arms shipments to Iran. North took the last of the notebooks with him when he was dismissed on Nov. 25, 1986 after the diversion of Iran arms sale proceeds to the contras was uncovered. Independent counsel Lawrence Walsh has been trying to get the Notebooks since well before North's indictment last year. A subpoena for the material was issued through the Federal Grand Jury investigating the **Iran-contra affair** but was withdrawn after North was charged, Walsh said in a court filing Monday. Walsh said the notes are presidential records over which the United States has complete ownership and control. North's notebooks describe successful efforts to obtain false End-user certificates from Guatemala stating that arms actually destined for the contras were instead for the exclusive use of the **Central American** country to which they were being sent. The false certificates enabled one of North's co-defendants, Richard Secord, to ship more than 90,000 pounds of east European munitions by chartered aircraft from a European arms dealer to the **contras**. Official learned before January 1986 that hawk missiles had been shipped to **Iran** in November 1985. In notebook entries for Nov. 20, 1985, several days before the Hawk missile shipment, north detailed a plan for shipping hawks to **Iran** in exchange for a four-phase release of hostages.
</TEXT>
</DOC>

FIG. 1. An example of a TREC document containing frequent collocations and ontology entries.

2010; Zhou et al., 2007) where frequent collocations and domain concepts are combined in a language model. In Bendersky and Croft (2012) and Zhang et al. (2007), several types of terms (bigrams, noun phrases, and named entities) denoting concepts and modeling high-order dependencies have been considered in a retrieval framework. However, none of those works exploited concepts and their relationships in the same model. We think that both are important sources of relevance for estimating the concept-based language model. Indeed, Cao et al. (2005) have shown the effectiveness of integrating different types of word relationships into LMs at the word level without considering word meanings. With the same spirit, we exploit the translation model but with a different formulation to incorporate concepts, their relationships, and subconcepts to enhance the concept-based language model.

Concept-Based Language Model

In this section, we describe the concept-based language model we propose. Our goal is to address term independency with two contributions:

- We consider documents and queries as a bag of concepts instead of words. We assume that a concept might be a single word or a multiple words and in both cases it might be an ontology entry or a frequent word collocation in the document but having no entry in the ontology. We assume that a frequent collocation can refer to a neologism or a proper name that has not been recorded in the ontology. A collocation as defined in Petrovic et al. (2006, p. 321) refers to ". . . a set of words occurring together more often than by chance. . .." Thus, our definition of a concept is roughly equivalent to the ones given in Bendersky and Croft (2012, p. 941) where ". . . concepts may model a variety of linguistic phenomena, including n-grams, term proximities, noun phrases, and named entities."

In our approach both types of concepts (i.e., frequent collocations and ontology entries) are combined in a unique language model framework. The example shown in Figure 1 is a TREC[4] document taken from the AP (Associated Press) data set. We see in this document that *"Iran-Contra"* is a frequent collocation in the document and is indeed an important concept because the document deals with

[4]trec.nist.gov.

"Iran-Contra scandal." The terms *"Iran"* and *"Contra"* are entries in the WordNet ontology.

- We use a semantic smoothing method based on the translation model (Berger & Lafferty, 1999) to map query concepts to document concepts through semantic relationships defined in the ontology. We consider such relationships as dependencies and additional sources of importance in estimating concept models. The intuition is to exploit concept relationships during query evaluation to retrieve documents dealing with the same or related query concepts. Thus, the concept-based language model is estimated according to occurrences of both concepts and their related ones, unlike previous approaches (Tu et al., 2010; Zhou et al., 2007) where concept models are only estimated by counting concepts whether they are ontology entries or not. For example, for the TREC query *"Iran-Contra Affair,"* the retrieval model should take into account of the presence of the concept *"Nicaraguan"* occurring twice in the document illustrated in Figure 1. It is a related concept to query concept *"Contra"* (in WordNet, *"Nicaraguan"* is a hypernym of *"Contra"*). Indeed, the document in Figure 1 deals with the *"Nicaraguan scandal,"* a related concept to *"Iran-Contra Affair."* Concept relationships are measured according to their semantic similarity (Resnik, 1995). In this context, most concept-based IR approaches merge generic and specific concept relationships even though they have a different effect on retrieval performance (Baziz et al., 2005; Cao et al., 2005; Zakos, 2005). We will show the positive effect of considering concepts and semantic relationships into a unified language model using smoothing techniques.

Overview

Given query Q, document D and ontology O. In our approach, we assume that D and Q are represented by concepts, $Q = \{c_1, c_2, \ldots c_m\}$ and $D = \{c_1, c_2, \ldots c_n\}$ respectively, where c_i is a concept that can be either an ontology entry or a frequent collocation in document collection, and in both cases it can be a single word or multiple words. Therefore, the relevance score $RSV(Q, D)$ of D with respect to query Q is given by the probability of query concepts to be generated by the document model described below.

$$RSV(Q,D) = P(Q|D)) = \prod_{i}^{n} P(c_i|D) \tag{3}$$

The estimation of $P(Q|D)$ attempts to abstract the unigram model described in Equation 1 (Related Work, above) at the concept level to which we refer as a concept-based document model, where $P(c_i|D)$ is the probability of concept c_i in D estimated as follows:

$$P(c_i|D) = \begin{cases} P(c_i, \bar{O}|D) & if\ c_i \notin O \\ P(c_i, O|D) & otherwise \end{cases} \tag{4}$$

For better clarity, we can rewrite probability $P(c_i|D)$ as:

$$P(c_i|D) = P(c_i, \bar{O}|D) + P(c_i, O|D) \tag{5}$$

where

- $P(c_i, \bar{O}|D)$ corresponds to the probability of c_i in D given the information that c_i is a frequent collocation having no entry in the ontology O.
- $P(c_i, O|D)$ is the probability that c_i has an ontology entry in the document model.

Assuming that concept c_i is an ontological entry, its probability $P(c_i, O|D)$ is estimated using the translation model (Berger & Lafferty, 1999), which is the most appropriate one to take into account semantic dependencies.

$$P(c_i, O|D) = \sum_{c_j \in D} P(c_i, O|c_j) P_{sem}(c_j|D)) \tag{6}$$

Equation 6 can also be seen as semantic smoothing and highlights the intuition that it allows incorporating semantic relationships between query and document concepts. Thus, when a query concept c_i is effectively seen in the document, its probability is adjusted, or more precisely, enhanced with probabilities of its related concepts within the document. Moreover, the estimation of $P(c_i, O|D)$ is carried out proportionally to the relationship degree estimated through the probability $P(c_i, O|c_j)$. Thus, the probability of query concept c_i is estimated by highlighting concept centrality. This concept centrality is a factor which "... measures how much a query concept is related to a document concept" (Boughanem et al., 2010, p. 2).

Substituting probability $P(c_i, O|D)$ (Equation 6) in 5, the latter becomes:

$$P(Q|D) = \prod_{c_i \in Q} \left[P(c_i, \bar{O}|D) + \sum_{c_j \in D} P(c_i, O|c_j) P_{sem}(c_j|D) \right] \tag{7}$$

The above equation shows the general principle of our approach, which combines frequent collocations and ontology entries into a unified language model framework. In our contribution, we exploit relationships such as synonymy, hypernymy, and hyponymy because they have been shown to be useful for IR (Cao et al., 2005). We notice that the synonymy is taken into account when representing documents and queries by ontological concepts viewed as sets of synonyms. Besides, Baziz et al. (2005) and Cao et al. (2005) showed empirically that each relationship has a specific impact on retrieval effectiveness (Cao et al., 2005). Accordingly, probability $P(c_i, O|c_j)$ is estimated by combining and weighting concept relationships differently. Thus, $P(c_i, O|c_j)$ is expressed by:

$$P(c_i, O|c_j) = \alpha P_{hyper}(c_i, O|c_j) + (1-\alpha) P_{hypo}(c_i, O|c_j) \tag{8}$$

where α and $(1 - \alpha)$ are the mixture weights of each relationship, $P_{hyper}(c_i, O|c_j)$ is the hypernymy model, and $P_{hypo}(c_i, O|c_j)$ is the hyponymy one. According to Equation 8, we end up with the following ranking function:

$$P(Q|D) = \prod\nolimits_{c_i \in Q} \Big[P(c_i, \bar{O}|D) + \sum\nolimits_{c_j \in D} [\alpha P_{hyper}(c_i, O|c_j) + (1-\alpha) P_{hypo}(c_i, O|c_j)] P_{sem}(c_j|D) \Big] \tag{9}$$

In what follows, we define the three conditional probabilities $P(c_i, \bar{O}|D)$, $P_{sem}(c_i|D)$, and $P(c_i, O|c_j)$.

Probability Estimation

Probability $P(c_i, \bar{O}|D)$. Probability $P(c_i, \bar{O}|D)$ of query concept c_i in document D given the information that c_i is not an ontology entry is smoothed using Dirichlet smoothing, which has shown good results in previous studies (Zhai & Lafferty, 2001). Indeed, c_i may not be effectively seen in the document. Thus, its probability $P(c_i, \bar{O}|D)$ is given by:

$$P(c_i, \bar{O}|D) = \frac{count(c_i, D) + \mu P_{ML}(c_i|C)}{\sum_{c_k \in D} count(c_k, D) + \mu} \tag{10}$$

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

$$P_{ML}(c_i|C) = \frac{count(c_i, C)}{\sum_{c_k} count(c_k, C)} \tag{11}$$

Semantic probability $P_{sem}(c_j|D)$. This probability can be estimated in different ways. For instance, using the maximum likelihood estimator based on a simple concept counting such as in Tu et al. (2010). In our work, probability $P_{sem}(c_j|D)$ is estimated by smoothing the maximum likelihood (ML) model of concept c_j with the ML of its component concepts called subconcepts and also corresponding to ontology entries.

Baziz et al. (2005) and Hammache et al. (2013), in the same way, estimate concept weights. The intuition is that the authors tend to use subconcepts to refer to a given concept they have previously used in the document. More precisely, when a multiterm concept, for instance, *"Military Coup,"* occurs in a document, the concept *"Coup"* which appears later is very likely used to refer to a *"Military Coup"* than to another sense, for example, "a brilliant and notable success" (see the example in Figure 2). Thus, $P_{sem}(c_j|D)$ is given by:

$$\begin{aligned} P_{sem}(c_j|D) &= \theta P_{ML}(c_j|D) \\ &+ (1-\theta) \sum_{sc \in sub-Concepts(c_j)} \frac{length(sc)}{length(c_j)} P_{ML}(sc|D) \end{aligned} \tag{12}$$

where $length(sc)$ is the number of words of subconcept sc that corresponds to an ontology entry. The ratio $\frac{length(sc)}{length(c_j)}$ is a factor that adjusts the subconcept model according to the relative length of sc and c_j (Baziz et al., 2005). The intuition is to strengthen long subconcepts. Smoothing parameter θ is $\in$ [0, 1]. The probabilities $P_{ML}(c_j|D)$ and $P_{ML}(sc|D)$ are respectively estimated using the Dirichlet smoothing as in Equation 10.

Probability P(c$_i$, O|c$_j$). $P(c_i, O|c_j)$ is the probability translation into the strength of the association between concepts c_i and c_j. Most of methods proposed in the literature for estimating word relationships are based on variants of co-occurrence in a training corpus (Bai et al., 2005; Berger & Lafferty, 1999; Cao et al., 2005). For instance, Cao et al. (2005) estimated such a probability at a word level (i.e., $P(w_i|w_j)$) by counting word co-occurrences in the collection and checking whether both words (w_i and w_j) are linked in WordNet. Bai et al. (2005) exploited other information such as Information Flow degree between single words within a certain context (a text passage, the whole of the document, or a window of fixed length). In this work, we estimate $P(c_i|c_j, O)$ using the relationship degree between c_i and c_j in the ontology relative to whole relationships between c_i and all document concepts. Formally, $P(c_i, O|c_j)$ is estimated according to two possible cases summarized as follows:

$$P(c_i, O|c_j) = \begin{cases} \frac{Rel(c_i, c_j)}{\sum_{c_k} Rel(c_i, c_k)} & if\ c_i \neq c_j \\ 1 & otherwise \end{cases} \tag{13}$$

where c_k, $k = \{1..n\}$, is a document concept and n is the number of concepts within the document.

$Rel(c_i, c_j)$ is a function estimating the c_i and c_j relationship. We use here a variant of Resnik Semantic Similarity (Resnik, 1995) based on the Information Content (IC) metric revisited by Zakos (2005). This metric highlights the specificity of concepts (Resnik, 1995; Zakos, 2005). Indeed, it has been shown in previous work (Baziz et al., 2005; Boughanem et al., 2010; Cao et al., 2005; Zakos, 2005) that specific concepts are more likely to be useful in IR than generic ones.

Formally, Resnik Semantic Similarity is based on an "is-a" relation and the Information Content (IC) metric proposed in (Seco, Veale, & Hayes, 2004) given as:

$$\begin{aligned} Rel(c_i, c_j) &= sim_{res*}(c_i, c_j) \\ &= max_{c \in S(c_i, c_j)} IC_o(c) \end{aligned} \tag{14}$$

where $S(c_i, c_j)$ is the set of concepts subsuming c_i and c_j.

IC_O has the particularity to be estimated according to the hierarchical structure of the ontology O, unlike the basic IC metric relying on word occurrences in a given corpus (Seco et al., 2004). Its principle is the following: the more descendants a concept has, the less information it expresses. Moreover, concepts that are leaves,[5] namely, specific ones, have more Information Content than the ones located higher in the hierarchy. Accordingly, the IC_O metric highlights concept specificity and is defined in Seco et al. (2004) as:

$$IC_O = 1 - \frac{\log(hypo(c)) + 1}{\log(max_O)} \tag{15}$$

[5]Located at the bottom of the ontology hierarchy.

<DOC>
<DOCNO> WSJ870519-0018 < /DOCNO>
<HL> WHAT'S NEWS -- WORLD-WIDE</HL>
<DD> 05/19/87</DD>
<SO> WALL STREET JOURNAL (J)</SO>
<TEXT>

The pentagon said Iraq attacked a U.S. frigate despite radio warnings. Officials said the Iraqi airplane's missile assault Sunday in the Persian gulf, in which 28 crew members were killed, followed two radio warnings to the mirage f-1 fighter jet from the ship, the uss stark. Nonetheless, the frigate, part of a u.s. naval force assigned to the gulf, never activated an array of defense systems it carried to thwart such a raid. The air strike creates new strains in U.S.-Iraq relations, and is expected to subject the navy to criticism for apparently being caught off guard. Reagan expressed "concern and anger" and ordered naval ships in the gulf to assume a heightened state of alert. Iraq's president, in a letter to Reagan, stated "deepest regret" over what he called the "unintentional" incident. Thousands of demonstrators battled with south Korean riot police as violence erupted throughout the country during protests marking the seventh anniversary of an uprising in the southern city of Kwangju. Mobs of students, workers and religious activists called for immediate elections and the removal of president chun doo hwan. The leader of FIJI'S **military coup** said he was installed as chairman of a caretaker government, but acknowledged that the south pacific nation's governor-general hasn't recognized the **coup** as legal. The governor-general, however, said he had received assurances that military rule would be ended "as soon as it is possible." A hijacker was overpowered by cabin crew members of an air new Zealand jet at nadi airport after the FIJI indian had commandeered the jet and threatened to blow up the plane with dynamite. The man, who had released all but three of the 126 people aboard, was negotiating with the control tower when the hijacking was thwarted. Queen Elizabeth ii dissolved Britain's parliament, officially opening a three-week general election campaign in which prime minister thatcher is seeking a third term. Britain's third party, the alliance of the liberals and social democrats, published a campaign platform, pledging a reduction in unemployment.

</TEXT>
</DOC>

FIG. 2. An example of a TREC document illustrating the intuition of considering subconcepts: "Coup" is a subconcept of the concept "Military Coup."

where *hypo*(c) is a function that returns the number of hyponyms of concept c. max_O is a constant. It generally takes the value of the maximum number of concepts in the ontology hierarchy.

Experimental Evaluation

To evaluate the retrieval effectiveness of our model, we used standard IR collections issued from TREC[6] evaluation campaigns. Our objectives were:

a) Evaluating the combination of ontological and the nonontological concepts in a language model.
b) Highlighting the impact of incorporating semantic information such as subconcepts and concept relationships (hyponyms, hypernyms).
c) Assessing the importance of disambiguating concepts in retrieval.
d) Comparing our model to two language models, namely, the unigram model smoothed with Dirichlet prior and the MRF language model.

Experimental Setting

We used two data sets issued from disk 1&2 of the TREC ad-hoc collection: the Associated Press 1989 (AP 89) and the Wall Street Journal 1986–1987 (WSJ 86–87) subcollections. For each data set, queries and relevance judgments are provided.

[6]trec.nist.gov.

Document indexing and query processing. For both data sets we indexed all documents with single words, frequent collocations, and ontological concepts. For this purpose we used the WordNet[7] v. 2.1 as ontology, its depth is $max_{wn} = 117659$. Thus, each document is processed using the following approach:

a) Terms (single words) and multiterms are identified using a collocation extractor, the Text-NSP tool (Banerjee & Pedersen, 2003). The latter is a software tool for extracting n-grams (sequences of n-tokens in text corpora) and provides statistics to detect frequent collocations such as frequencies and Mutual Information. The multiterm size is limited to three or less.
b) Detected terms are then processed in order to remove *"not valid terms,"* mainly those beginning or ending with a stopword defined in the Terrier stopword list (Ounis et al., 2005). We avoid pretreatment of the text before detecting multiterms in order to retain all potential concepts. Indeed, terms such as *"Chief of State"* or *"Obstruction of Justice"* could be important concepts.[8] They are generally called *"complex phrases"* (Zhang et al., 2007) and are frequently monosemic.
c) For validating that a given multiterm is a concept, we keep only those occurring at least twice.
d) We check whether a concept (a single word or a frequent collocation) occurs in WordNet. Those having an entry are selected and represented by their Synset Number. The latter is a set of synonyms having a unique identifier. For instance, the concepts *"Coup," "Coup d'etat," "Takeover,"* and *"Putsch"* are grouped in the Synset, whose identifier is 01147528. Therefore, the synonymy relationship is automatically incorporated in the model.
e) When a given concept has several entries (polysemy), the first sense in WordNet is selected as the default.
f) The remainder of concepts (including single words) are retained as non-WordNet entries (single words and frequent collocations) and weighted with simple count of occurrences.

A set of search topics (numbers 51–100) were used as queries. Each topic is composed of three parts: Title part, which is the short query, Description part, which is a long query, and Narrative part, describing previous parts and precisely what relevant documents should contain.

In our experiments, for all data sets, we used only the Title part of topics as queries for two main reasons: (a) Title parts are as short as user queries and have the same syntactic form, usually nominal phrases such as the query "Information Retrieval System" (TREC topic 65), and (b) Terms of Title part are more important than the remainder of the topic parts (Metzler & Croft, 2005). During querying, the extraction of topic concepts is the same as the process of document indexing. Notice that some topics have no relevant documents: topic 63 on the WSJ 86–87 data set and topics 65, 66, and 69 on the AP 89.

[7]wordnet.princeton.edu.

[8]These examples are taken from TREC documents.

Evaluation Metrics

Our model is compared with baseline models using the standard Text Retrieval Conference method. It is reported in Buckley and Voorhees (2000) that the mean average precision (MAP) and the precision at the rank x noted $P@x$, $x \in \{10, 20\}$ (the ratio evaluating the number of relevant documents in the top x retrieved documents) are the most used metrics to evaluate the overall effectiveness of an IR system. As MAP is estimated over all the queries, some details about the performance of our model can be hidden. For this reason we conducted a per query analysis (by comparing average precision per query). All performance measures are obtained by evaluating our retrieval runs using the trec_eval[9] standard tool. In addition, to show the consistency of our results, we perform statistical significance testing. We use the Student's *t*-test shown in Hull (1993) to be suitable for information retrieval systems.

Baseline and Evaluation Scenarios

The baseline we used to compare our model is the unigram model-based Dirichlet prior smoothing (Smucker & Allan, 2005; Zhai & Lafferty, 2004) available on the Terrier system (Ounis et al., 2005). The relevance score is given by:

$$RSV(Q,D) = \prod_{q_i \in Q} P(q_i|D) = \prod_{q_i \in Q} \frac{tf(q_i, D) + \mu P(q_i|C)}{\sum_{w \in D} tf(w, D) + \mu} \tag{16}$$

where $tf(q_i, D)$ is the query term frequency in document D, μ is a smoothing parameter, and probability $P(q_i|C)$ corresponds to the background collection language model estimated by the maximum likelihood estimator as follows.

$$P_{ML}(q_i|C) = \frac{count(q_i, C)}{\sum_{w \in V} count(w, C)} \tag{17}$$

where V is the vocabulary.

For comparison purposes, we performed experiments using three variants of our model considering the concept unigram model and the smoothed one with concept relationships:

- *CLM_I* is the individual concept model which corresponds to Equation 7 without considering concept relationships, thus the ranking function becomes:

$$P(Q|D) = \prod_{c_i \in Q} [P(c_i, \bar{O}|D) + P(c_i, O|D)] \tag{18}$$

where $P(c_i, O|D)$ is simply estimated using Dirichlet smoothing without considering relationships (see Equation 10).

- *CLM_R* is the concept-based model integrating concept relationships corresponding to Equation 9.

[9]See http://trec.nist.gov/trec_eval/.

- *CLM_R** is the concept-based model integrating concept relationships in which query concepts are disambiguated according to their centrality in the document. This disambiguation method has shown powerful results (Boughanem et al., 2010). The centrality of a concept is equivalent to the number of concepts related to it in document *D*. We notice that relationships are taken from WordNet.

$$Centrality(c_i) = \# R(c_i, c_k) \quad \forall c_k \in D \tag{19}$$

where $\#R(c_i, c_k)$ is the number of c_i relationships c_k in *D*.

In this experiment, the value of μ is set to 2,500 for all Dirichlet smoothing models and data sets. This value corresponds to the optimal value recommended in the literature (Zhai & Lafferty, 2001).

Results and Evaluation

In the following sections, we discuss the results obtained in the experimental evaluation of the three variants of our model. We show results in MAP and precision at the 10-document level (P@10), which is comparable to MAP and easier to score. At the end, we compare the variant achieving the best MAP to the MRF model (Metzler & Croft, 2005).

Impact of combining frequent collocations and WordNet concepts. We aim here at evaluating the impact of combining frequent collocations and WordNet entry concepts. We performed a four-stage evaluation of *CLM_I*.

a) *CLM_IF* using only frequent collocations.
b) *CLM_IC* using only WordNet concepts (we recall that a WordNet concept is a set of synonyms) without considering their subconcepts as detailed in Equation 12 where θ value is set at 1.
c) *CLM_IS* using WordNet concepts with considering their subconcepts. This variant returns the best performances when $\theta = 0.6$. This value has been set by tuning $\theta \in [0, 1]$ with increments of 0.1. Therefore, we kept the optimal value of θ for the remainder of the experiments.
d) *CLM_I* considering both frequent collocations and WordNet concepts with their subconcepts.

Figure 3 shows clearly that the combined model (*CLM_I*) outperforms all of the individual models based on frequent collocations (*CLM_IF*) and the one based on WordNet concepts (*CLM_IC* and *CLM_IS*) in MAP and *P*@10. Indeed, for the WSJ 86–87 data set, *CLM_I* achieves a MAP value of 0.2376 and *P*@10 value of 0.3640. As for the AP 89 data set, the MAP and P@10 are 0.1908 and 0.3280. This result could be explained by the fact that *CLM_IC*, *CLM_IF,* and *CLM_IS* used individually do not cover all document content, whereas *CLM_I* captures more semantic content by considering both frequent collocations and concepts. Moreover, the concept model includes synonyms and subconcepts. Accordingly, *CLM_I* was used in the remaining experiments.

We compare *CLM_I* and ULM performance in Table 1, where the line gain (%) denotes the percentage of noticed improvements. The reported precisions show that *CLM_I* gives significant improvement over ULM (Unigram Model) for both data sets. More precisely, for the WSJ 86–87 data set we notice a significant improvement at *P*@10 than for MAP, with the respective values of +10.98% and +3.21%. For the AP 89 data set, we notice the most important improvement on MAP with a value of +6.41%. However, the *P*@10 and *P*@20, improvements are less important than the ones noticed for the WSJ 86–87.

In what follows, we perform a deep analysis per topic to highlight the factors that contribute to this improvement and to illustrate through examples of queries how relevant documents are promoted with our model (*CLM_I*).

Per-topic analysis. In this analysis, queries are separated depending on whether or not they contain concepts. Accordingly, we have 34 concept queries and 16 nonconcept queries for the WSJ 86–87 data set, while we have 27 concept queries and 20 nonconcept queries for the AP 89 data set. Concept queries correspond wholly to a WordNet entry (there are eight) or a frequent collocation, or contain a

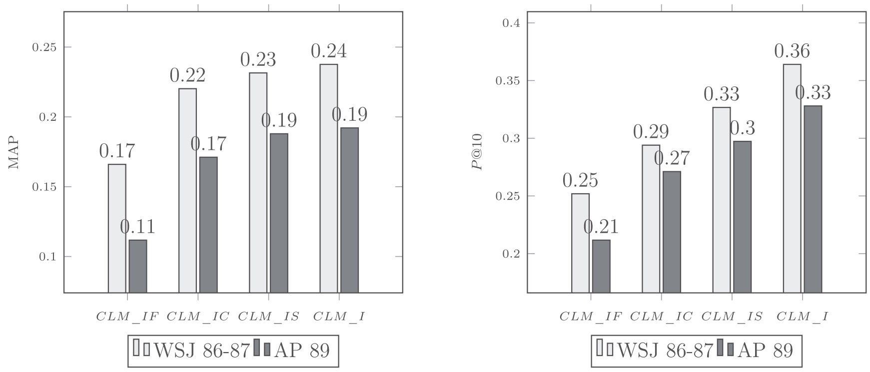

FIG. 3. Overview of performances (MAP and P@10) of *CLM_IF*, *CLM_IC*, *CLM_IS*, and *CLM_I*.

TABLE 1. Comparison between performances of the unigram model (ULM) and the combined concept-based model *CLM_I*. Signs $^+$ and $^{++}$ indicate that the difference is statistically significant according to *t*-test at *p*-value < .05 and *p*-value < .01, respectively.

		Performance evaluation		
Collection	Model	*P@10*	*P@20*	*MAP*
WSJ 86–87	ULM	0.3280	0.2790	0.2302
	CLM_I	0.3640	0.3030	0.2376
	Gain over ULM(%)	$+10.98^{++}$	$+8{,}60^{++}$	$+3{,}21^{+}$
AP 89	ULM	0.3160	0.2810	0.1809
	CLM_I	0.3280	0.2910	0.1925
	Gain over ULM(%)	$+3{,}80^{+}$	$+3{,}56^{+}$	$+6{,}41^{++}$

TABLE 2. TREC-query examples of concept and nonconcept queries (OE and FC indicate ontology entry and frequent collocation, respectively).

Collection	Concept queries	Nonconcept queries
WSJ 86–87	Query 51: **Airbus Subsidies** (FC)	Query 85: Official corruption
	Query 65: Military **Coups D'etat** (OE)	Query 94: Computer-aided crime
AP 89	Query 64: **Hostage-Taking** (FC)	Query 59: Weather-related Fatalities
	Query 79: **FRG** Political Party Positions (OE)	Query 95: Computer-aided crime detection

TABLE 3. Per-topic analysis of ranking models (*CLM_I* vs. ULM).

		CLM_I vs. ULM			
Collection	Query category	–	=	+	Change (%)
WSJ 86–87	Concept queries (35)	13	3	19	+38,61
	Nonconcept queries (14)	6	2	6	+21,63
AP 89	Concept queries(27)	8	3	16	+4,79
	Nonconcept queries (20)	12	2	6	–18,00

TABLE 4. Ranking of relevant document TREC topic 62 with ULM and *CLM_I*.

	Document ranking		
Document number	ULM	*CLM_I*	Change in ranking
WSJ871016-0026	88	34	–54
WSJ870928-0164	29	8	–21
WSJ870831-0105	2	1	–1
WSJ870828-0026	13	9	–4
WSJ870828-0019	80	45	–35
WSJ870518-0113	98	19	–53
WSJ870514-0016	25	67	+42
WSJ870128-0030	8	6	–2
WSJ870127-0005	103	132	+29
WSJ870526-0068	114	22	–92

long concept (more than one word). Examples of concept and nonconcept queries are given in Table 2.

Table 3 compares performance of our model *CLM_I* with respect to ULM for the two underlying categories. Columns noted (–, =, +) indicate the number of queries for which *CLM_I* achieved a worse, as equal as, or better average precision (AP) than the ULM. The column change (%) presents the rate of improvement in MAP over each category of query.

For both data sets, our model outperforms ULM mainly on queries containing concepts. Particularly for the WSJ 86–87 data set, we observed the best performance by yielding an improvement rate of (+38.61%) in MAP. For instance, for the query *"Iran-contra Affair"* (topic 99), *CLM_I* and ULM achieved respectively an average precision of 0.2706 and 0.069. This result could be explained by the fact that the query itself and *"Iran-contra"* are concepts (frequent collocations which occur more than twice in the data set). We also notice for nonconcept queries an important improvement of (+21.63%). For example, the query *"Computer-aided crime"* does not contain any collocation or long concepts, but the single words *"Computer"* and *"Crime"* are WordNet entries. For this query, our model achieves an AP value of 0.0526, whereas ULM achieves 0.0035. This enhancement is mainly due to the representation of query words *"Computer"* and *"Crime"* by their Synsets in WordNet (a synset is essentially a set of synonyms) because it does not contain any long concept.

We focused our analysis on the ranking of relevant documents[10] with ULM and *CLM_I*. The query *"Military Coup d' etat"* (Topic-62) is an example of a concept query for which our model achieves an AP of 0.2911, whereas ULM achieves 0.1407. This query contains the following concepts: two WordNet entries *"Military"* and *"Coup d'etat"* and two terms that are not WordNet entries *"Military Coup"* (a frequent collocation) and *"Etat."* The WSJ 86–87 data set contains 13 relevant documents according to the relevance judgment file of the TREC Adhoc Collection. These documents are ranked with ULM and *CLM_I* as showed in Table 4.

We notice that most documents have their ranks promoted with *CLM_I*. In particular, the respective ranks of

[10]Taken from the relevance judgments file.

TABLE 5. Ranking of AP 89 relevant document to TREC topic 61 with ULM and *CLM_I*.

Document number	Document ranking		
	ULM	*CLM_I*	Change in ranking
AP890105-0053	26	22	–4
AP890106-0010	41	1	–40
AP890125-0124	98	71	–27
AP890328-0075	54	16	–48
AP890329-0168	11	31	+20
AP890408-0099	65	53	–12
AP890413-0132	6	18	+12
AP890504-0157	144	60	–84

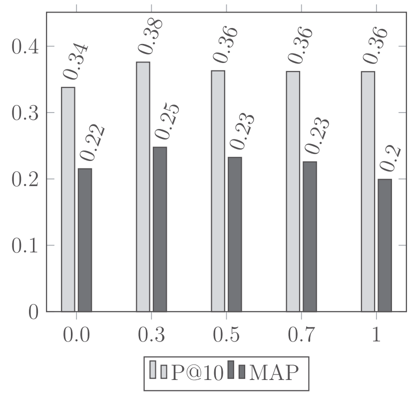

FIG. 4. Variation of P@10 and MAP with α using *CLM_R* for the WSJ 86–87 data set.

documents WSJ870518-0113 and WSJ870526-0068 have been promoted from 98 and 114 with ULM to 19 and 22 with *CLM_R*. We analyzed the content of the document WSJ870526-0068 and observed that the term *"Coup d'etat"* does not occur in that document, but synonyms (which belong to the same synset) such as *"Coups"* and *"takeover"* occur respectively seven and six times.

However, for queries corresponding to a WordNet entry and composed of a unique term, our model performs equally or nearly as well as ULM. For instance, on topic 78 *"Greenpeace,"* our model and ULM achieve respectively an AP of 0.2951 and 0.2929.

For the AP 89 data set, the results are slightly lower than those achieved for the WSJ 86–87 data set. For nonconcept queries, we notice a degradation of performance over ULM with the value of −18.00%. However, we notice improvement on some queries. For instance, for the query *"Hostage-taking"* (topic 64) our model returns a better AP than ULM, contrary to what we observed for the WSJ 86–87, where ULM is better. The reason is that the query is wholly a frequent collocation in the AP data set, in contrast to WSJ 86–87.

The same point has been observed for the query *"Israeli Role in Iran-Contra Affair"* (topic 61). *CLM_I* and ULM achieved respectively an AP of 0.1041 and 0.0733. The AP 89 data set contains eight relevant documents. Table 5 reports the ranking of these documents to this query (topic 61) with ULM and our model. This table shows that the ranks of all relevant documents have been promoted with *CLM_I*. For example, the rank of document AP890106-0010 has been promoted from 41 to 1. This is mainly due to the fact that this document contains almost all of the query concepts. On the one hand, *"IRAN," "Contra," "Affair,"* and *"Israeli"* are WordNet entries that occur respectively 25, 14, 3, and 2 times. In addition, the related concept *"matter,"* which is a synonym of query concept *"Affair,"* occurs once. On the other hand, *"Iran-contra"* and *"Iran-contra Affair"* are frequent collocations in the document and occur respectively two and six times.

In the light of these results, we can conclude that incorporating diversified sources of concepts, such as frequent collocations, ontology concepts, and subconcepts may be effective.

Impact of incorporating concept relations. We evaluated integrating concept relationships into the retrieval model. This evaluation concerns *CLM_R*, which integrates hypernymy and hyponymy ("IS-A") relationships.

It has been shown that retrieval performance is sensitive to smoothing parameter values (Zhai & Lafferty, 2001). Thus, in our experiment we tuned parameter α in Equation 9 (corresponding to *CLM_R*), where α is varied in [0, 1] in increments of 0.1. Figure 4 illustrates the variation of retrieval effectiveness mainly in P@10 and MAP according to α. It shows that the best performance is achieved at α value 0.3. We have also observed that the same α value corresponds to the best retrieval performance for the AP 89 data set. This shows that combining the concept relationships hypernymy and hyponymy is more effective than exploiting each one individually (when $\alpha = 0$ or $\alpha = 1$). Therefore, $\alpha = 0.3$ is used throughout the remainder of the experiments. We recall that α and $(1 - \alpha)$ are respectively the mixture weights of the hypernym and hyponym models.

Table 6 summarizes the retrieval performance of *CLM_R* over *CLM_I* and ULM. We notice that for both data sets *CLM_R* significantly outperforms ULM, mainly on P@10 and MAP. It also achieves better performance than *CLM_I*. In particular, we observe significant improvement over ULM on P@10 for both data sets. However, the improvement achieved by *CLM_R* over *CLM_I* is less important for the AP 89 data set. Nevertheless, these observations lead us to conclude that integrating concepts and semantic relationships are effective for IR. In what follows, we will further show this on a per-topic analysis.

Per-topic analysis. Here we refine the analysis by topic with a focus on AP. The results of comparison of the three models ULM, *CLM_I,* and *CLM_R* overall long and short queries are summarized in Table 7.

TABLE 6. Comparison between ULM and concept-based models *CLM_I* and *CLM_R*. Signs $^{+}$ and $^{++}$ indicate that the difference is statistically significant according to *t*-test at *p*-value < .05 and *p*-value < .01, respectively.

Collection	Model	Performance evaluation		
		P@10	*P@20*	*MAP*
WSJ 86–87	ULM	0.3280	0.2790	0.2302
	CLM_I	0.3640	0.3092	0.2376
	CLM_R	0.376	0.317	0.2477
	Gain over ULM (%)	+14,63^{++}	+13,62^{++}	+7,60^{++}
	Gain over *CLM_I* (%)	+3,23^{+}	+2,52^{+}	+4,25^{+}
AP 89	ULM	0.3160	0.2810	0.1809
	CLM_I	0.3280	0.2910	0.1925
	CLM_R	0.3420	0.2940	0.1932
	Gain over ULM (%)	+8,23^{++}	+4,63^{+}	+6,79^{+}
	Gain over *CLM_I* (%)	+4,27^{++}	+1,03	+0.36

TABLE 7. Per-topic analysis of ranking models (*CLM_R* vs. ULM, and *CLM_I*).

Collection	Query category	*CLM_R* vs. ULM				*CLM_R* vs. *CLM_I*			
		−	=	+	Change	−	=	+	Change
WSJ 86–87	Concept queries (35)	8	4	23	+34,06	15	2	18	+7,15
	Nonconcept queries (14)	6	1	8	+21,00	6	2	6	−4,5
AP 89	Concept queries (27)	7	2	18	+15,40	11	3	13	+10.12
	Nonconcept queries (20)	11	2	9	−2,72	10	3	7	−12,04

TABLE 8. Ranks of relevant documents to TREC topic 62 with ULM, *CLM_I,* and *CLM_R*.

Document number	Document ranking				
	ULM	*CLM_I*	Change	*CLM_R*	Change
WSJ871016-0026	88	34	−84	28	−6
WSJ870928-0164	29	8	−11	3	−5
WSJ870831-0105	2	1	−1	1	0
WSJ870828-0026	13	9	−5	4	−1
WSJ870828-0019	80	45	−35	125	+122
WSJ870518-0113	98	19	−79	14	−5
WSJ870514-0016	25	67	+42	121	+54
WSJ870128-0030	8	6	−2	2	−4
WSJ870127-0005	103	132	+31	95	−37
WSJ870526-0068	114	22	−92	18	−4

TABLE 9. Ranking of relevant documents to TREC topic 61 with ULM, *CLM_I,* and *CLM_R*.

Document number	Document ranking				
	ULM	*CLM_I*	Change	*CLM_R*	Change
AP890105-0053	26	22	−4	20	−2
AP890106-0010	41	1	−40	1	0
AP890125-0124	98	71	−27	59	−12
AP890328-0075	54	16	−38	13	−3
AP890329-0168	11	31	+20	11	−20
AP890408-0099	65	53	−12	37	−16
AP890413-0162	53	18	−35	6	−12
AP890504-0157	144	60	−84	43	−17

This comparison shows that *CLM_R* is enhancing retrieval over ULM and *CLM_I*. For both data sets, *CLM_R* shows the best performance on concept queries. This enhancement in MAP is mainly due to long concepts in queries that are often monosemic (a concept having a unique sense). Therefore, their probabilities are enhanced with those of their hyponyms and hypernyms in returned documents. To illustrate our statement, we give in Table 8 a focus on ranks of relevant documents for the query *"Military Coup d'etat"* (topic 62). The achieved AP with ULM, *CLM_I,* and *CLM_R* are respectively 0.1446, 0.2911, and 0.3985.

It can be seen that most relevant documents were promoted. The analysis of some document content confirms that these contain, as expected, hyponyms and (or) hypernyms of query concepts. For example, document WSJ870128-0030 is promoted from 6 to 2 because it contains the concept *"Forces,"* which is a direct hypernym of the query concept *"Military"* (it occurs four times just within that document).

The same observation has been noticed for AP 89 results. Table 9 presents in more detail the results of the reranking achieved per query *"Israeli Role in Iran-Contra Affair"* (TREC topic 61).

Table 9 highlights that most documents have been promoted with *CLM_R*. For example, document AP890504-0157, after being ranked at 144 with ULM, has been

TABLE 10. Comparison of *CLM_R**, *CLM_R,* and ULM performances. Signs + and ++ indicate that the difference is statistically significant according to *t*-test at the level of *p*-value < .05 and *p*-value < .01.

Collection	Model	Performance evaluation		
		P@10	*P@20*	*MAP*
WSJ 86–87	ULM	0.3280	0.2790	0.2302
	CLM_R	0.376	0.317	0.2477
	*CLM_R**	0.3833	0.3102	0.2827
	Gain over ULM (%)	+16,85++	+15,05++	+22,81++
	Gain over *CLM_R* (%)	+1,94	−2,14	+6,12++
AP 89	ULM	0.3160	0.2810	0.1809
	CLM_R	0.3420	0.2940	0.1932
	*CLM_R**	0.3740	0.3090	0.1935
	Gain over ULM (%)	+14,63++	+13,62++	6.97++
	Gain over *CLM_R* (%)	+9,36++	+5,10++	+0.16

TABLE 11. Per-topic analysis of ranking models (ULM, *CLM_R,* and *CLM_R**).

Collection	Query category	*CLM_R** vs. ULM				*CLM_R** vs. *CLM_R*			
		−	=	+	Change	−	=	+	Change
WSJ 86–87	Concept queries (35)	12	1	26	+41,27	11	10	15	+6,15
	Nonconcept queries (14)	6	2	7	+21,00	7	2	5	+1,08
AP 89	Concept queries (27)	8	4	15	+14,50	4	3	16	+30.35
	Nonconcept queries (20)	7	2	11	+1,09	6	4	10	+2,35

promoted to 43 with our model. This occurs because query concepts appear in that document as follows: concepts that are not WordNet entries, such as *"Iran-Contra"* occurs three times, *"Iran-contra Affair"* occurs twice, and WordNet entries *"Israeli," "Affair,"* and *"Nicaraguan"* (hypernym of Contra) occur respectively 1, 5, and 3 times in that document. The probability of *"contra"* is boosted the presence of its hypernym *"Nicaraguan"* as shown in Equation 9.

Another example illustrating concept relationships is the query *"Crude Oil Price Trends"* (topic 88). For this query, *CLM_R* reached an AP value of 0.0654, whereas ULM achieved 0.0332. This enhancement in AP is explained by the presence of the query concept *"Crude Oil"* (also a WordNet entry) occurring in relevant documents. In addition, the weight of that concept is boosted with its hyponyms (*"Residual Oil"*) and hypernyms (*"Fossil Fuel"*). However, for nonconcept queries, *CLM_R* has not shown as strong an improvement as ULM and *CLM_I*. For most of these queries, ULM outperforms our model. One can explain this by the fact that terms of these queries are sometimes wrongly disambiguated with their first sense in WordNet, making the relationships wrong also. Another possible explanation is that there are no hypernyms and hyponyms of query concepts in relevant documents. We can conclude in general that incorporating some semantic relationships between concepts and weighting them proportionally to their importance (see Equation 13) yields significant improvements in retrieval effectiveness.

Impact of incorporating disambiguated concept relations. We evaluate our retrieval model by disambiguating query and document concepts by concept centrality described above (Boughanem et al., 2010). We recall that for previous experiments we selected the first sense of the concepts.

Table 10 recapitulates the performance effectiveness of *CLM_R**,*CLM_R,* and ULM. The reported precisions show that the improvement achieved with *CLM_R** over ULM is significant. However, the comparison over *CLM_R* indicates mixed results. For WSJ 86–87, we notice that *CLM_R** slightly outperforms *CLM_R* on *P*@10 with an improvement of +1.94%, while the improvement +9.36% is higher for the AP 89 data set. In contrast, the noticed improvement in MAP is clearly significant for WSJ 86–87. Although improvements achieved by *CLM_R** are variable, we can conclude that integrating correct relationships into the retrieval model increases to some extent the retrieval effectiveness.

Per-topic analysis. To show how the disambiguation of concepts enhances retrieval effectiveness, we perform in what follows a per-topic analysis.

Table 11 shows that *CLM_R** clearly outperforms ULM, while it slightly outperforms *CLM_R*, particularly for concept queries. Nevertheless, there are some nonconcept queries for which *CLM_R** has achieved better results than did *CLM_R*. For instance, for the query *"Information*

TABLE 12. Ranking of relevant documents to TREC topic 65 with ULM, *CLM_R,* and *CLM_R**.

	Document ranking				
Document number	ULM	*CLM_R*	Change	*CLM_R**	Change
WSJ870304-0091	137	250	–113	36	–137
WSJ870331-0042	98	71	–27	59	–12
WSJ870429-0078	48	4	–44	1	–43
WSJ871103-0021	17	16	–1	11	–5
WSJ871202-0145	311	19	–292	8	–11

Retrieval System" (topic 65), *CLM_R** and *CLM_R* achieved high APs of 0.3652 and 0.3519, respectively, while the AP achieved by ULM is 0.0361. The WSJ 86–87 data set contains five documents relevant for this query and they are ranked with the three compared models as follows.

The example shown in Table 12 confirms the enhancement observed on average precision, notably with the document WSJ870304-0091, whose rank jumps from 250 with *CLM_R* to 36 with *CLM_R**. WSJ870429-0078 has also seen its rank promoted from 4 to 1. This enhancement is mainly due to the right disambiguation of query terms as follows.

a) The term *"Information,"* which corresponds in this query to the fourth sense in WordNet (i.e., "Data: a collection of facts from which conclusions may be drawn"). For the remaining words, the first sense is indeed the right one. Relevant documents contain the term *"Data"* as a synonym for *"Information."*
b) The right sense of the term *"Retrieval"* is still the first one in WordNet (Computer Science). This sense is also related to the concept *"Storage,"* which is omnipresent in relevant documents such as WSJ870304-0091, where it occurs five times.

These statements confirm that incorporation of correct concept relationships into the concept-based language model improves document retrieval.

Comparison with the MRF model. We also compare our model *CLM_R** to a language model named MRF (Metzler & Croft, 2005). We used the Sequential Dependency (MRF-SD) variant of MRF (see above). The value of free parameters of the MRF-SD are taken from (Metzler & Croft, 2005). Table 13 compares MAP achieved by these models for both data sets.

The change (%) column in Table 13 highlights that our model achieves the best MAP over MRF-SD for both data sets. The enhancement for the WSJ 86–87 is significant (+13.71%) while the improvement is less noticeable for the AP 89 data set (+3.81%) but statistically significant. However, on P@10, MRF-SD is marginally more effective than our model. This can be explained by the fact that the collocations of MRF-SD are better filtered than ours, which are somewhat noisy (adjacent and order constraint).

TABLE 13. Comparison between *CLM_R** and MRF-SD performances. Signs $^{+}$ and $^{++}$ indicate that the difference is statistically significant according to *t*-test at *p*-value < .05 and *p*-value < .01, respectively.

	Performance	*MRF – SD*	*CLM_R**	Change (%)
WSJ 86–87	MAP	0.2486	0.2827	+13,71^{++}
	P@10	0.3857	0.3833	–0.62
AP 89	MAP	0.1864	0.1935	+3,81^{+}
	P@10	0.3851	0.3740	–2,88^{+}

TABLE 14. Per-topic analysis of ranking models (*CLM_R** vs. MRF-SD).

		*CLM_R** vs. MRF-SD			
Collection	Query category	–	=	+	Change (%)
WSJ 86–87	Concept queries(35)	12	3	20	+4,93
	Nonconcept queries (14)	8	1	6	–2,54
AP 89	Concept queries (27)	12	2	13	+2,08
	Nonconcept queries (14)	10	2	10	–3,15

To further clarify results achieved by *CLM_R** and MRF-SD, we give below an analysis per topic.

Per-topic analysis. We also perform an analysis per topic to show how our model outperforms the MRF-SD. The results of the comparison are given in Table 14, which shows that *CLM_R** is better than MFR-SD especially for concept queries in both data sets.

Take the example of the concept query *"Attempts to Revive the SALT II Treaty"* (topic 69) which contains an ontological concept *"SALT II."* For this query, our model achieved a better AP (with a value of 0.6003) than MRF-SD and ULM, which achieved respectively APs of 0.4725 and 0.4709. We show in Table 15 the ranking of relevant documents under the three models. We can observe that ranks of some documents have decreased markedly, for instance, WSJ861201-0004, WSJ861209-0002, WSJ870507-0018, and WSJ870120-0047.

When examining document WSJ870120-0047 ranked at 175 with ULM, it appears that it does not contain some query words such as *"Attempt"* and *"Treaty."* However, it is matched to the query and is promoted from 175 to 32 and 24 with MRF-SD and *CLM_R**, respectively. The explanation is that this document contains concepts related to *"Treaty"* such as *"Accord"* (synonym) and *"Agreement"* (hypernyms) and *"SALT II"* (hyponym). These concepts occur once, twice, and once, respectively. Thus, the frequency of *"Treaty"* is not null, and is enhanced with related concept frequencies. The same has been observed in the document WSJ870130-0003 which has been promoted from 44 and 40 with ULM and MRF-SD to 19 with our model. Although the concept *"Treaty"* only occurs once, its probability is enhanced with its related concepts occurring in the document, namely, *"Pact"*(synonym),

TABLE 15. Ranking of relevant documents to TREC topic 69 with ULM, MRF-SD and *CLM_R**.

Document number	Document ranking				
	ULM	MRF-SD	Change	*CLM_R**	Change
WSJ861201-0004	53	40	−13	17	−27
WSJ861202-0040	3	2	−1	2	0
WSJ861204-0019	52	44	−12	36	−8
WSJ861205-0001	40	58	+18	35	−23
WSJ861209-0002	105	104	−1	39	−65
WSJ861216-0141	7	15	+8	13	−2
WSJ861218-0172	2	1	−1	1	0
WSJ861222-0149	4	6	+2	4	−2
WSJ861229-0047	27	31	+4	28	−5
WSJ870106-0081	25	30	+5	21	−9
WSJ870120-0047	175	32	−143	24	−8
WSJ870130-0003	44	40	−4	19	−21
WSJ870203-0101	8	4	−4	7	+3
WSJ870305-0116	1	9	−8	9	0
WSJ870506-0144	15	25	−10	16	−9
WSJ870507-0018	82	67	−15	34	−33

"Accord" (synonym), *"Agreement"* (hypernym), and *"SALT II"* (hyponym).

For nonconcept queries, we can see in Table 14 that for both data sets, MRF-SD outperforms our model in MAP with the value of (+2.54%) for the WSJ 86–87 data set and (+3.15%). Let us examine for example query *"1988 Presidential Candidate Platforms"* (topic 80). Our model performs nearly as well as ULM, with an AP of 0.0625 (AP achieved by ULM is 0.0655), whereas the MRF-SD registered an AP with value of 0.1047. This can be explained by two main reasons: the first, the numerical date *"1988,"* which is an important information in this query. The second is related to noisy collocations such as *"Candidate Platforms"* kept as a frequent collocation in our approach since (its frequency >2) but appears also in documents irrelevant to this query (topic 80).

Overall, this study allows us to confirm that our model outperforms the MRF model particularly for queries containing concepts (WordNet entries as well as frequent collocations). For nonconcept queries, the improvement for both data sets is marginal or even worse. Thus, we can expect further improvements by carrying out deeper analysis to capture semantic information using proximity features, for example.

Discussion

Our research is mainly related to harnessing concepts extracted from both a semantic resource and documents (frequent collocations) to enhance document retrieval via concept queries. We have shown through the experiments reported in this paper that estimating an accurate concept-based document model by considering concepts, subconcepts, and their relationships (synonyms, hypernyms, hyponyms) can be effective for IR.

Previous works (Baziz et al., 2005; Tu et al., 2010; Zhou et al., 2007) have shown the advantage of using units (phrases or concepts) larger than single words. They exploit either an external resource (ontology) or a resource built upon a corpus to capture document and query concepts (Bendersky & Croft, 2012). In the context of LMs, most research has relied exclusively on concept frequencies to estimate the concept-based document models, whereas LM variants such as the translation model provide a way to exploit semantic information such as synonymy. In addition, it is known that terminology is continuously enriched with new concepts including neologisms, named entities, or acronyms that are not yet integrated into semantic resources. That is why researchers have focused on looking for more effective approaches to capture the semantics of a document.

As mentioned previously (see the Introduction), this paper deals with two main research questions that aim to investigate the effectiveness of our proposed concept-based language model.

First, we have been answering the first question and showed that a concept can be an ontological entry or a frequently found collocation (occurring more than twice in the text collection). Indeed, the latter may denote important concepts such as neologisms, proper names, or specific concepts that have no entry in the ontology. The results of the experiment showed that combining both types of concepts is effective. We believe that concept queries are predominant in domain-specific retrieval such as medical or biological queries where users use specific concepts to express their queries (Meij et al., 2009; Zhou et al., 2007). However, on the web we rely on a study of Bendersky and Croft (2009) where they showed that long queries (containing more than four words) also called verbose queries are more likely to contain concepts.

Second, to estimate the concept-based model, we have shown the possibility of integrating concepts from the ontology as well as from the document collection into a unified language model framework. Our intuition is similar to the one used in Zhou et al. (2007), which also considered frequently found collocations and ontological concepts (a concept is a set of synonyms). However, in the Zhou et al. (2007) approach, both frequently found collocations and ontological entries are used equally; that is, the document model is smoothed with concept frequencies, which are either ontology entries or not. In our approach, the concept-based model is estimated by considering subconcepts and synonyms (since a concept is a set of synonyms). The results we obtained by comparing the model combining these elements to single word and concept unigram models have shown improvements in terms of MAP. Furthermore, we exploit semantic relationships between concepts to semantically match query and document concepts in such a way that generic concepts (hypernyms) are separated from specific concepts. Therefore, more distant dependencies than word proximity (Bai et al., 2005; Gao et al., 2004; Zhao & Yun, 2009) are integrated into the retrieval language model. Cao et al. (2005) also integrated more distant dependencies between query and document words through word relationships (cooccurrence, synonymy, hypernymy, and hyponymy). However, they have not considered the meaning of single words, which are often ambiguous, so the problem of polysemy is still present. We demonstrated here that concept relationships are effective to improve IR performance in comparison to state-of-the-art models in cases in which the queries are concepts.

Our work has some limitations:

- The proposed concept-based model, in particular the collocation-based one, is mainly estimated using frequencies. We think that it can be further enhanced by considering frequent subconcepts as the approach proposed in Hammache et al. (2013).
- Our approach of recognizing concepts is somehow limited by order because we extract first the collocations that vary from one to three words. We can overcome this limitation by relaxing concept word adjacency and recognizing concepts in a larger context such as a text passage.
- The number of trials for this experiment are the total number of queries in the WSJ data set (49) plus the total number from the AP data set (47). The number of trials is acceptable, but we need to perform further experiments using a larger set of queries (more than 100) to better assess the robustness of our model.
- Word collocation would need to be performed upon document ingest, and whenever new items are added to the collection (rather than at query time) in order for the system to retrieve at decent speed. Web-wide, this method would require massive item indexing.

Conclusion and Future Work

In this paper we introduce a concept-based language modeling approach to enhance information retrieval. In our approach, document and queries are represented through concepts. We consider concepts (that vary from one to two or three words) that can be an ontology entry or a frequent collocation having no entry in the ontology. This leads to a rich representation of document content and closes the semantic gap between query and document (brought about assuming the word independence). In addition, we integrate concept relationships a priori defined in the ontology WordNet to better model the document. Indeed, we consider that a document can be represented by concepts (synonyms, hypernyms, hyponyms, subconcepts, frequent collocations, and single words).

Our experimental results on TREC data sets showed that our model yields significant improvements over the unigram model and the MRF-SD by Metzler and Croft (2005). This has been noticed for queries containing concepts that are long queries or queries that are wholly ontology entries. We showed also that differentiating concept relationships such as the hypernymy and the hyponymy is promising for retrieval.

We plan in the short term to test our model on medical collections, where the notion of concept is important. In the long term, our model could be further improved by integrating additional Natural Language Processing rules for recognizing useful phrases. For nonconcept queries, it would be interesting to use an additional resource such as Wikipedia, Dbpedia, or Yago to identify concepts. This might also enhance the query model with term relationships through the graphical structures of these resources.

Acknowledgments

We thank the reviewers for their helpful comments and suggestions.

References

Alvarez, C., Langlais, P., & Nie, J.-Y. (2004). Word pairs in language modeling for information retrieval. In C. Fluhr, G. Grefenstette, & W.B. Croft (Eds.), Proceedings of the 7th International Conference on Computer Assisted Information Retrieval (RIAO'04) (pp. 26–28). Avignon: CID.

Bai, J., Song, D., Bruza, P., Nie, J.-Y., & Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In O. Herzog, H. Schek, N. Fuhr, A. Chowdhury, & W. Teiken (Eds.), Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM′05) (pp. 688–695). Bremen, Germany: ACM.

Banerjee, S., & Pedersen, T. (2003). The design, implementation and use of the n-gram statistics package. In G. Alexander (Ed.), Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (pp. 370–381). Mexico City: Springer-Verlag.

Bao, S., Zhang, L., Chen, E., Long, M., Li, R., & Yu, Y. (2006). LSM: Language sense model for information retrieval. In J.X. Yu, M. Kitsuregawa, & H.V. Leong (Eds.), Proceedings of 7th International Conference on Advances in Web-Age Information Management (WAIM′05) (pp. 97–108). Berlin, Heidelberg: Springer.

Baziz, M., Boughanem, M., & Aussenac-Gilles, N. (2005). A conceptual indexing based on document content representation. In F. Crestani, & I. Ruthven (Eds.), Context: Nature, impact, and role (pp. 171–186). Glasgow, Heidelberg: Springer.

Bendersky, M., & Croft, W.B. (2009). Analysis of long queries in a large scale search log. In N. Craswell, R. Jones, G. Dupret, & E. Viegas (Eds.), Proceedings of the of the 2009 Workshop on Web Search Click Data (WSCD′09) (pp. 8–14). Barcelona, Spain: ACM.

Bendersky, M., & Croft, W.B. (2012). Modeling higher-order term dependencies in information retrieval using query hypergraphs. In W.R. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 941–950). Portland: ACM.

Bennett, G., Scholer, F., & Uitdenbogerd, A. (2007). A comparative study of probabilistic and language models for information retrieval. In A. Fekete, & X. Lin (Eds.), Proceedings of the 9th Conference on Australasian Database (pp. 65–74). Gold Coast: Australian Computer Society.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In F. Gey, M. Hearst, & R. Tong (Eds.), Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 222–229). Berkeley: ACM.

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

Buckley, C., & Voorhees, E.M. (2000). Evaluating evaluation measure stability. In N.J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 33–40). Athens: ACM.

Cao, G., Nie, J.-Y., & Bai, J. (2005). Integrating word relationships into language models. In R.A. Baeza-Yates, N. Ziviani, G. Marchionini, A.

Moffat, & J. Tait (Eds.), Proceedings of the 28th Annual International ACM SIGIR Conference on research and development in information retrieval (pp. 298–305). Salvador: ACM.
Chen, S.F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In A. Joshi, & M. Palmer (Eds.), Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (pp. 310–318). Stroudsburg: Association for Computational Linguistics.
Gao, J., Nie, J.-Y., Wu, G., & Cao, G. (2004). Dependence language model for information retrieval. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza (Eds.), Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 170–177). Sheffield: ACM.
Gonzalo, J., Li, H., Moschitti, A., & Xu, J. (2014). Semantic matching in information retrieval. In S. Geva, A. Trotman, P. Bruza, C.L.A. Clarke, & K. Järvelin (Eds.), Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1296–1296). Gold Coast: ACM.
Hammache, A., Boughanem, M., & Ahmed-Ouamar, R. (2013). Combining compound and single terms under language model framework. Knowledge and Information Systems, 39(2), 329–349.
Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In R. Korfhage, E.M. Rasmussen, & P. Willett (Eds.), Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 329–338). Pittsburgh: ACM.
Krovetz, R., & Croft, W.B. (2000). Lexical ambiguity and information retrieval. ACM Transaction Information Systems, 10(2), 115–141.
Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 111–119). New Orleans: ACM.
Lavrenko, V., & Croft, W.B. (2001). Relevance based language models. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 120–127). New Orleans: ACM.
Li, H., & Xu, J. (2013). Semantic matching in search. Foundations and Trends in Information Retrieval, 7(5), 343–469.
Macdonald, C., & Ounis, I. (2008). Voting techniques for expert search. Knowledge and Information Systems, 16(6), 259–280.
Meij, E., Trieschnigg, D., de Rijke, M., & Kraaij, W. (2008). Parsimonious concept modeling. In D.W. Oard, F. Sebastiani, T. Chua, & M. Leong (Eds.), Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 815–816). Singapore: ACM.
Meij, E., Trieschnigg, D., Kraaij, W., & de Rijke, M. (2009). Conceptual language models for domain-specific retrieval. Information Processing and Management, 47(4), 448–469.
Metzler, D., & Croft, W.B. (2005). A Markov random field model for term dependencies. In R.A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 05) (pp. 472–479). Salvador: ACM.
Miller, G.A. (1995). WordNet: A lexical database for english. Communications of the ACM, 38(11), 39–41.
Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Johnson, D. (2005). Terrier information retrieval platform. In D.E. Losada, & J.M. Fernández-Luna (Eds.), Proceedings of the 27th European Conference on IR Research (ECIR'05) (pp. 517–519). Santiago de Compostela, Heidelberg: Springer.
Petrovic, S., Snajder, J., Dalbelo-Basic, B., & Kolar, M. (2006). Comparison of collocation extraction measures for document indexing. Journal of Computing and Information Technology, 14(4), 321–327.
Ponte, J.M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In W.B. Croft, A. Moffat, C.J.V. Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 257–281). Melbourne: ACM.
Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 448–453). Montreal: Morgan Kaufmann Publishers.
Salton, G., Buckley, C., & Yu, C.T. (1982). An evaluation of term dependence models in information retrieval. In G. Salton, & H. Schneider (Eds.), Proceedings of the 5th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 151–173). Berlin, Heidelberg: Springer.
Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In R. López de Mntaras, & L. Saitta (Eds.), Proceedings of the 16th Eureopean Conference on Artificial Intelligence ECAI (pp. 1089–1090). Valencia: IOS Press.
Shi, L., & Nie, J.-Y. (2009). Integrating phrase inseparability in phrase-based model. In J. Allan, J. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 708–709). Boston: ACM.
Smucker, M.D., & Allan, J. (2005). An Investigation of Dirichlet Prior Smoothing's Performance Advantage (Tech. Rep.). The University of Massachusetts, The Center for Intelligent Information Retrieval.
Song, F., & Croft, W.B. (1999). A general language model for information retrieval. In F. Gey, M. Hearst, & R. Tong (Eds.), Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM 99) (pp. 316–321). Berkeley: ACM.
Srikanth, M., & Srihari, R. (2003). Biterm language models for document retrieval. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, & S. Myaeng (Eds.), Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 425–426). Tampere: ACM.
Srikanth, M., & Srihari, R. (2003). Incorporating query term dependencies in language models for document retrieval. In C. Clarke, G. Cormack, J. Callan, D. Hawking, & A. Smeaton (Eds.), Proceedings of the 26th Annual International ACM SIGIR Conference on Research and development in Information Retrieval (pp. 405–406). Toronto: ACM.
Tu, X., He, T., Chen, L., Luo, J., & Zhang, M. (2010). Wikipedia-based semantic smoothing for the language modeling approach to information retrieval. In C.l. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S.M. Rüger, & K.V. Rijsbergen (Eds.), Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR 10) (pp. 370–381). Berlin, Heidelberg: Milton Keynes.
Wei, C.-P., Hu, P., Tai, C.-H., Huang, C.-N., & Yang, C.-S. (2007). Managing word mismatch problems in information retrieval: A topic-based query expansion approach. Journal of Management Information Systems: JMIS, 24(3), 269–295.
Zakos, J. (2005). A novel concept and context-based approach for Web information retrieval. Unpublished doctoral dissertation, School of Information and Communication Technology, Griffith University, Gold Coast.
Zhai, C. (2008). Statistical language models for information retrieval a critical review. Foundations and Trends in Information Retrieval, 2(3), 137–213.
Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 334–342). New Orleans: ACM.
Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems, 22(2), 179–214.
Zhang, W., Liu, S., Yu, C., Sun, C., Liu, F., & Meng, W. (2007). Recognition and classification of noun phrases in queries for effective retrieval. In M.J. Silva, A.H.F. Laender, R.A. Baeza-Yates, D.L. McGuinness, B. Olstad, Ø. Olsen, & A.O. Falcão (Eds.), Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (pp. 711–720). Lisbon: ACM.

Zhao, J., & Yun, Y. (2009). A proximity language model for information retrieval. In J. Allan, J. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 291–298). Boston: ACM.

Zhou, D., Bian, J., Zheng, S., Lee, G., & Zha, H. (2008). Exploring social annotations for information retrieval. In J. Huai, R. Chen, H. Hon, Y. Liu, W. Ma, A. Tomkins, & X. Zhang (Eds.), Proceedings of the 17th International World Wide Web Conference (pp. 715–724). Beijing: ACM.

Zhou, X., Hu, X., & Zhang, X. (2007). Topic signature language models for ad hoc retrieval. IEEE Transactions on Knowledge and Data Engineering, 19(9), 1276–1287.