Knowledge based word-concept model estimation and refinement for biomedical text mining

Antonio Jimeno Yepes[a], Rafael Berlanga[b]

[a]*Department of Computing and Information Systems, The University of Melbourne, VIC 3010, Australia, phone: +61 405 096 629, antonio.jimeno@gmail.com*
[b]*Departamento de Lenguages y Sistemas Informáticos, Universitat Jaume I, Castellón de la Plana, 12071, Spain*

Abstract

Text mining of scientific literature has been essential for setting up large public biomedical databases, which are being widely used by the research community. In the biomedical domain, the existence of a large number of terminological resources and knowledge bases (KB) has enabled a myriad of machine learning methods for different text mining related tasks. Unfortunately, KBs have not been devised for text mining tasks but for human interpretation, thus performance of KB-based methods is usually lower when compared to supervised machine learning methods. The disadvantage of supervised methods though is they require labelled training data and therefore not useful for large scale biomedical text mining systems. KB-based methods do not have this limitation.

In this paper, we describe a novel method to generate word-concept probabilities from a KB, which can serve as a basis for several text mining tasks. This method not only takes into account the underlying patterns within the descriptions contained in the KB but also those in texts available from large unlabelled corpora such as MEDLINE. The parameters of the model have been estimated without training data. Patterns from MEDLINE have been built using MetaMap for entity recognition and related using co-occurrences.

The word-concept probabilities were evaluated on the task of word sense disambiguation (WSD). The results showed that our method obtained a higher degree of accuracy than other state-of-the-art approaches when evaluated on the MSH WSD data set. We also evaluated our method on the task of document ranking using MEDLINE citations. These results also showed an increase in performance over existing baseline retrieval approaches.

Keywords: word-concept probability, text mining, word sense disambiguation, information retrieval, biomedical literature

1. Introduction

Text mining of biomedical literature has supported the development of biomedical knowledge bases (KB), which are actively used by the research community [23]. These databases have contributed as well in the development of methods to perform text mining related tasks like entity recognition and relation extraction. There are a large number of KBs available for biomedical text mining purposes. Some of these resources are integrated into the Unified Medical Language System® (UMLS®) [12] and many resources are available from the Open Biological and Biomedical Ontologies (OBO) foundry [39][1]. Unfortunately, since these resources were not developed to

[1] OBO foundry: http://www.obofoundry.org

perform text mining tasks, knowledge based methods usually exhibit lower performance compared to ad hoc supervised methods (e.g., supervised classifiers) [20]. Despite this limitation, knowledge based approaches become crucial when either there is a scarcity of labelled data to train supervised methods. Due to the heterogeneity and large scale of biomedical resources, knowledge based methods are becoming more popular.

Estimating word-concept probabilities from KBs provides an effective way to support a large range of text mining tasks in the biomedical domain [40]. Unlike supervised methods, the absence of manually labelled data can be alleviated by defining statistical approximations from either the existing data in the KBs (e.g., names, relations and descriptions) or external data such as MEDLINE® abstracts [20]. Other approaches are aimed at building statistical models directly from corpora, like Latent Dirichlet Allocation (LDA) [11], but it is not clear how to interpret or integrate these models within the KB structures [15].

Word sense disambiguation (WSD) and information retrieval (IR) are two tasks that benefit from word-concept probability models. Given an ambiguous word with its context, WSD attempts to select the proper sense given a set of candidate senses. An example of ambiguity is the word *cold* which could either refer to *low temperature* or the *viral infection.* The context in which *cold* appears is used to disambiguate it. WSD is an intermediate task that supports other tasks such as: information extraction [5], information retrieval and summarization [33]. WSD in the biomedical domain is mostly based on either supervised learning or knowledge based approaches [37]. As previously mentioned, the scarcity of training data makes knowledge based methods preferable to supervised ones.

In IR, KB based methods have been proposed for either expanding queries or for performing semantic searches [14, 25]. However, these methods do not provide a proper way to combine the expanded words, and just use the KB for defining improved IR queries as we have shown in [25].

This work proposes a novel method for generating word-concept statistical models from KBs that can be used directly for both IR and WSD. As mentioned earlier, this method is also able to take advantage of existing data in MEDLINE to produce a model with improved performance. These models can be integrated into IR language models to resolve ambiguity.

An implementation of the presented method is available from
https://bitbucket.org/ajjimeno/wkpropability.

2. Related Work

In the biomedical domain, there have been several big projects and initiatives to build comprehensive knowledge resources such as OBO and UMLS. At the same time, during the last decade researchers have devised automatic text mining techniques to find new knowledge from the scientific literature [9]. In this paper, we are interested in developing a general purpose probabilistic model that can be used in several text mining tasks, such as WSD and Document Ranking.

WSD methods are based on supervised learning or KB-based approaches [37]. Supervised methods are trained on examples for each one of the senses of an ambiguous word. A trained model is used to disambiguate previously unseen examples. This approach requires a large set of training examples, which is usually not available. For example, the 2009AB version of the UMLS contains approximately 24 thousand ambiguous words, based on the exact match of the words in the UMLS Metathesaurus. Preparing such training examples would be very expensive to build and maintain [44].

In the biomedical domain, KB-based methods for WSD either build a concept profile [29, 28, 20], develop a graph-based model [2, 3] or rely on the semantic types assigned to each concept for disambiguation [19]. These derived models are compared to the context of the ambiguous word being disambiguated to select the most likely sense. In these approaches, candidate senses of the ambiguous word are UMLS concepts.

KB-based methods have been complemented with information available from existing resources like MEDLINE. An example is the use of MeSH indexing®[2] as additional information [41]; although this approach is dependent on the availability of MeSH indexing. In previous work, we collected training data from MEDLINE citations for each sense of an ambiguous word [20]. PubMed queries used to retrieve these citations were generated using English monosemous relations [27] of the candidate concepts which, po-

[2] NLM's controlled vocabulary used to index MEDLINE: https://www.nlm.nih.gov/mesh

tentially, have an unambiguous use in MEDLINE. This approach has shown good performance compared to other KB-based methods. In a subsequent study, we extended the work in [20] by considering all of MEDLINE instead of the top 100 recovered citations by PubMed and by generating concept profiles that can be easily estimated on large number of examples [21]. Using a large number of examples showed an improvement over previous methods.

Semi-supervised algorithms could be used to obtain additional examples of contexts for ambiguous words. We explored this in [22], where the initial disambiguation predictions provided by an unsupervised method were used as a seed to identify better concept profiles. This method showed a significant improvement.

There are several approaches in WSD that utilize the graph structure of the resources [30, 1], e.g. by applying adaptations of the page rank algorithm. Unfortunately, these methods cannot be re-used for other tasks like IR, because the generated models are only able to rank senses for given contexts, and not documents for given concepts. Conversely, approaches for IR that take into account the KB (e.g. [25]) are aimed at generating IR queries but not statistical models for other purposes.

In this paper, we claim that the generation of statistical models from both the KB and existing external corpora can provide a very valuable resource for effectively performing various text mining tasks. Furthermore, we show that the presented model generates word-concept probabilities that produce good results on these tasks.

3. Methods

In this section, we present the word-concept statistical model. The estimation of the model based on the knowledge base is presented in Section 3.1. The model estimates weights to combine probabilities from concepts at different traversal steps. In this work, the model is adjusted using it for disambiguation, which is introduced in Section 3.2. The adjustment is based on Expectation-Maximization as explained in Section 3.3. Once the model is trained, it can be refined based on existing corpora in an unsupervised way as explained in Section 3.4. The word-concept probabilities obtained from this model can be used in other tasks such as IR as explained in Section 3.5. Lastly, experimental set up and data sets used in this work are presented in Section 3.6.

In this work, a KB is defined as an inventory of concepts $\mathcal{C}$, where each concept $c \in \mathcal{C}$ is associated to a list of lexical forms $lex(c)$ (i.e., strings of text that are synonyms, variants, and so on), and a set of relations to other concepts, denoted with $r(c, c')$. These relations can be of any kind, from taxonomic *is-a* relations to other specific biomedical domain relationships (e.g. treats). Resources like the UMLS Metathesaurus fit this KB definition (see Section 3.6). Strings of text consist of tokens, that are their model primitives. Tokens may be punctuation or words, which are the minimal semantic tokens in the text. Terms are words or multi-word expressions denoting a concept(e.g. the synonyms and lexical variants linked to concepts in the UMLS).

3.1. Word-concept probability estimation

We propose estimating the probability $P(w_j|c_i)$ by selecting a word w_j given a concept c_i in a KB. This is done by selecting a word from the concept c_i, step 0, or from any of the related concepts at any specific step k while traversing the KB relations. The method described below provides a way to estimate this probability at different traversal steps.

The models obtained at different steps are combined using a linear combination. The weights of the linear combination are defined in the vector $\overrightarrow{\beta}$ (from equation 2), whose dimension is the number of traversal steps as shown in equation 1.

$$P(w_j|c_i) = \sum_{k=0...l} \beta_k P_k(w_j|c_i) \tag{1}$$

$$\beta_0 \ldots \beta_l > 0, \sum_{k=0...l} \beta_k = 1 \tag{2}$$

At step 0, the probability of a word w_j given a concept of interest c_i is given in equation 3. The equation considers the relative frequency of the word in the context of the lexical forms of the concept. The function *count* returns the number of times the word w_j is linked to concept c_j by any of the synonyms associated to the concept.

$$P_0(w_j|c_i) = \frac{count(w_j, c_i)}{\sum\limits_{w_j \in lex(c_i)} count(w_j, c_i)} \tag{3}$$

When estimating the probability for a step k larger than 0, the probability of the word w_j for the concept c_i is derived from equation 4-8, considering all the concepts at k steps from c_i. The concept of interest is at step 0 and referred as c_0. The final concept of a path is denoted as c_k. The probabilities are summed for all possible paths with length k linking word w_j and concept c_0.

In these equations, $R_k(c_0)$ returns the concepts reached after k steps from concept c_0. $P(c_{l+1}|c_l)$ is the traversal probability estimated using equation 9. The concepts in the paths of length k can be obtained by traversing the KB relations using breadth-first search starting at concept c_0.

Equation 8, shows how to estimate the probability $P_k(w_j|c_0)$ for word w_j and concepts at step k from c_0. The final equation depends on traversal probabilities from c_0 to concept at step k (c_k) and the conditional probability of the word w_j with the concept c_k ($P_0(w_j|c_k)$), which can be estimated as shown in equation 3.

$$P_k(w_j|c_0) = \tag{4}$$

$$\sum_{c_k \in R_k(c_0)} P_k(w_j, c_k, ..., c_1|c_0) = \tag{5}$$

$$\frac{\sum_{c_k \in R_k(c_0)} P_k(w_j, c_k, ..., c_0)}{P(c_0)} = \tag{6}$$

$$\frac{\sum_{c_k \in R_k(c_0)} P_0(w_j|c_k) \prod_{l=0..k-1} P(c_{l+1}|c_l) P(c_0)}{P(c_0)} = \tag{7}$$

$$\sum_{c_k \in R_k(c_0)} P_0(w_j|c_k) \prod_{l=0..k-1} P(c_{l+1}|c_l) \tag{8}$$

When estimating $P(c_{l+1}|c_l)$, as shown below, the function $r(c_1, c_2)$ returns the relations in which concepts c_1 and c_2 are related in the knowledge base. The numerator is the count of relations in which concepts c_{l+1} and c_l are related. The denominator is the count of relations in which concept c_l is in.

$$P(c_{l+1}|c_l) = \frac{|r(c_{l+1}, c_l) \in KB|}{|r(\cdot, c_l) \in KB|} \tag{9}$$

We have set the initial β weights to $\beta_i = 1/n$ where n is the number of steps considered. Log probabilities are used to obtain better accuracy in the probability estimation, which is not shown here for simplicity but it is available from the source code in bitbucket.

Figure 1 shows a simplified version of how the UMLS concept C0009264 (cold temperature) is considered by the model in a 1-step model. The terms linked to the concept are decomposed into words and counts for the stemmed words (that appear ended with the * character) can be estimated in relation to the concept. In this case, the count for the stem *temperatur** is 2, the count for *low* and *cold* is 1 respectively. Figure 1 shows as well the related concept C0016736 (frostbite). Frequency of the related concept is used to estimate the relation probability.

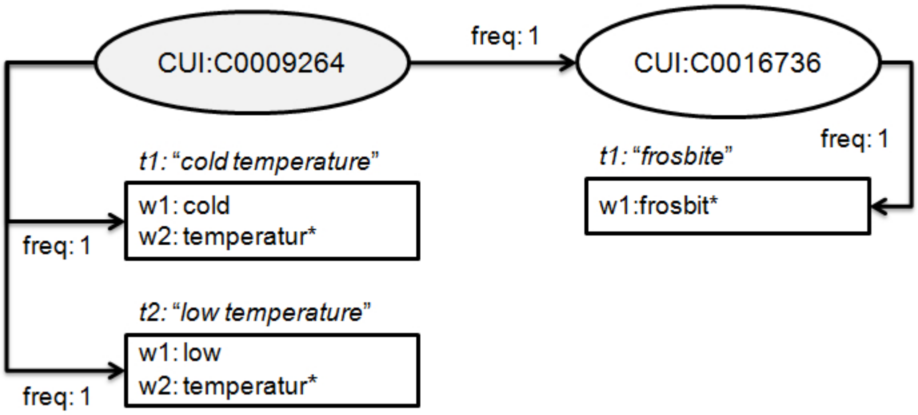

Figure 1: Simplified version of how the UMLS concept C0009264 (cold temperature) is considered by the model in a 1-step model

The final model is smoothed based on Jelinek-Mercer smoothing [7] as shown in the equation 10, where λ has been set to 0.75 based on previous work by Zhai and Lafferty [46]. The background of each word $P(w_j|KB)$ has been estimated over the KB occurrences by applying an add-one smoothing as shown in equation 11. $|N|$ is the count of unique words in the KB, while $|w_i|$ is the count of the i-th word in the KB.

$$P(w_j|c_i) = (1-\lambda)P(w_j|c_i) + \lambda P(w_j|KB) \tag{10}$$

$$P(w_i|KB) = \frac{\alpha + |w_i|}{\alpha \cdot |N| + \sum\limits_{j=1..N} |w_j|} \tag{11}$$

3.2. Using the model in disambiguation

Disambiguation consists of selecting the sense that best fits the context D of an ambiguous word. The context typically consists of the set of words surrounding the ambiguous word. In this work, the context is the MEDLINE abstract containing the ambiguous word. Disambiguation is performed similarly to Naïve Bayes classification, using the following equation, which assumes the independence of words.

$$P(c_j|D) = \frac{P(D|c_j)P(c_j)}{P(D)} \propto P(D|c_j) = \prod_{w_i \in D} P(w_i|c_j) \tag{12}$$

A candidate concept for an ambiguous word is selected according to maximum a posteriori (MAP) of the above expression given the context D of an ambiguous word w and the candidate concepts $C_w \subseteq \mathcal{C}$.

$$c^*(w) = \underset{c \in C_w}{\arg\max} P(D|c) \tag{13}$$

3.3. Model adjustment

In order to establish the β parameters of the traversals, we apply an expectation-maximization (EM) method over the set of contexts, D, where ambiguous words occur. This set of documents can be taken from either the KB (e.g., concept descriptions or definitions) or from an external corpus (e.g., MEDLINE abstracts). In any case, the algorithm does not know a priori the right concept associated to each context, so the method is fully unsupervised.

In the implementation presented in this work, during the expectation step, the concept with the highest probability is assigned to each context D using equation 13, introduced in the previous section. During the maximization step, we use the concept assignment to estimate the β weights. A regularization parameter based on an estimated θ prior (probability of selecting a word from a given step) is used to avoid overfitting.

The weight associated to this prior (α) is set to 0.3. More complex regularization methods could be applied [43], but their evaluation is beyond the scope of this paper. A modified version of the log-likelihood including the prior of each β parameter is used. After each iteration t, the log probability of the model is estimated, and the EM method stops when the log probability is not smaller than the previous iteration's log probability. The estimation of parameters at each iteration is defined in equation 14. δ is set to 1 when the ambiguous word w has been disambiguated with concept c in context d.

$$\beta_i^{t+1} = \frac{\sum_d \sum_w \beta_i^t P_i(w|c)\delta(c,d) + \alpha P(\theta_i^t(w))}{\sum_d \sum_w \sum_j \beta_j^t P_j(w|c)\delta(c,d) + \alpha P(\theta_j^t(w))} \tag{14}$$

3.4. Model refinement

As previously mentioned, the initial word-concept model is estimated from the KB data only. We propose exploiting the information available in an external corpora relying on two heuristics. The first heuristic is one sense per discourse [18], namely: all the occurrences of an ambiguous word in a document refer to the same sense. The second heuristic is one sense per collocation [45]. The idea is to identify the terms that tend to happen with each possible sense of the ambiguous word.

We propose refining the estimates iteratively using statistics from the target corpus, which is done in two steps. In the first step, the corpus is annotated with KB concepts, resolving the ambiguities with the word-concept model of the current iteration. In the second step, a new word-concept model is obtained based on the KB statistics and the annotated corpus statistics. The word-concept count in equation 3 also considers counts from the KB and the annotated corpus. The counts from the corpus indicate which terms tend to be used in the same context as the concept. This is different to the approach used in [24, 25], where terms were removed from the resource. The concept-concept count in equation 9 also considers counts from the KB and the corpus. In contrast to previous work, this allows adding information by assigning a weight to the relations.

In the current implementation, the concept-concept counts are based on co-occurrences of concepts at sentence level, even though higher precision information extraction methods can be considered (e.g. syntactic dependencies and relation extraction [31]). The two steps are repeated until the Log Likelihood derived from equation 1 does not increase.

Figure 2 shows a simplified version of the state of the UMLS concept C0009264 (cold temperature) in a 1-step model after a refinement. In contrast to the example in Figure 1, the term *low temperature* appears more frequently in the corpus in comparison to *cold temperature.* In this case, the count for the word *temperatur** is 20, the count for *low* is 18 and the count for *cold* is 2. Similarly, frequencies are updated for the other concepts. Figure 2 shows as well the related concept C0016736 (frostbite) and the newly related concept C0016736 (cold exposure). Frequencies to the related concepts are updated according to the refinement method. In this example, just one occurrence with concept C0016736 was found in the corpus, which is added with the mention from the KB, so its (e.g., MEDLINE abstracts) new frequency is 2. No relation to concept C0016736 (cold exposure) appeared in the KB but this concept was found to appear 15 times with the concept C0009264, thus the frequency is 15.

3.5. Using the proposed model in document ranking

In addition to disambiguation, we propose using the model in document ranking. The ranking of the documents D for a given concept c can be derived from cross-entropy (CE) [25] between the word-concept $P(w_i|c)$ and word-document $P(w_i|D)$ models as follows:

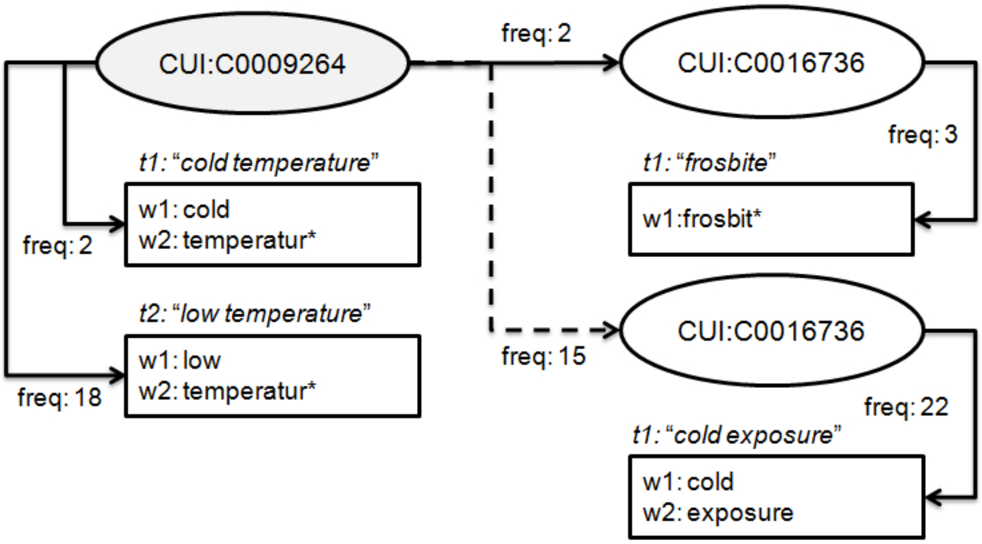

Figure 2: Simplified version of how the UMLS concept C0009264 (cold temperature) is considered by the model in a 1-step model after an example refinement

$$CE(c,D) = \sum_{w_i \in D} P(w_i|c) \cdot logP(w_i|D) \tag{15}$$

Word-document probability can be estimated as shown in equation 16, which combines the maximum likelihood estimation (MLE) with a background probability G using Jelinek-Mercer smoothing. As before, λ is initially set to 0.75.

$$P(w_i|D) = (1-\lambda)\frac{count(w_i)}{\sum_{w \in D} count(w)} + \lambda P(w_i|G) \tag{16}$$

3.6. Experimental setup

3.6.1. Biomedical knowledge base

The biomedical KB used in our experiments is the UMLS. The UMLS is a compendium of a large number of biomedical terminologies and ontologies, and is the largest biomedical terminological resource.

We used the 2012 UMLS version AA with the default installation. We estimated the model using two UMLS Metathesaurus tables available in Rich Release Format (RRF). The *MRCONSO* table was used to estimate the word-concept probabilities in equation 9. The terms linked to the concepts are lowercased, tokenized into individual words, stemmed with the Porter stemmer [34] and filtered using a standard stop word list[3]. The *MRREL* table was used to calculate the traversal probabilities in equation 9. Since synonym information is already obtained from *MRCONSO* and it is not clear how to interpret a synonym relation in *MRREL*, this information from *MRREL* is ignored. More details about these tables can be found from the UMLS web site[4].

3.6.2. Biomedical corpora

We used two data sets, one for model refinement and disambiguation evaluation, and another one for evaluating the retrieval performance. Both sets are derived from MEDLINE.

[3] Stopword list from the SMART system: ftp://ftp.cs.cornell.edu/pub/smart

[4] UMLS site: http://www.nlm.nih.gov/research/umls

The first set is the WSD corpus called MSH (MeSH) WSD [26][5]. This corpus has been generated using MeSH indexing of MEDLINE to determine the correct UMLS concept assigned to an ambiguous word. Using MeSH as reference allows us to automatically build a a large disambiguation corpora which is typically a time intensive processAlthough, the corpus is limited to MeSH headings that can be mapped to UMLS concepts, it contains a more comprehensive set of possible ambiguities than other biomedical WSD corpora (e.g. [44]). MSH WSD contains 203 ambiguous terms, with an average of 2.3 senses per term and a maximum of 100 examples per sense. The context of the ambiguous word is composed of the words in the citation in which the ambiguous word appears. As in the processing of the MRCONSO file, the text is lowercased, tokenized, processed with the Porter stemmer and filtered using the same stopword list.

The ranking set is based on a corpus developed for the evaluation and comparison of algorithms for MeSH indexing [6]. Citations belong to a subset from the 2013 MEDLINE and have been split into 2/3 (94,942 citations) for training and 1/3 (48,911 citations) for testing purposes. MEDLINE is indexed manually using terms from the MeSH controlled vocabulary, thus this indexing was used to build the retrieval data set. As in the disambiguation task, the text is lowercased, tokenized, processed with the Porter stemmer and filtered using the same stopword list.

[5] MSH WSD: http://wsd.nlm.nih.gov/collaboration.shtml

[6] http://ii.nlm.nih.gov/DataSets/index.shtml#2013_MTI_ML

For retrieval evaluation, we have reused the ambiguous terms from the MSH data set as queries, since they can be mapped to MeSH terms. Then, we selected the ones with at least 100 citations in the training set, determined by the MeSH indexing. This totaled 82 terms used as queries.

4. Results

The generated model and its refinement based on the UMLS as KB and a corpus derived from MEDLINE for the refinement have been evaluated in the ranking and disambiguation tasks. We determine statistical significance with a randomization version of the two sample t-test [17], which avoids making assumptions on the distribution of the data and allows for a better estimation of significance between the difference of the methods performance.

As mentioned before, we have limited the model to 2-step paths due to computation time and memory requirements. After estimating the beta values for the probabilities in each step k using the EM method, we obtained the beta values: β_0=0.6654, β_1=0.0678, β_2=0.2668. That is, the 0-step model (i.e., considering only words in $lex(c)$) holds the highest weight, followed by the 2-step model. The estimation of the model, which includes the traversal of the KB, took around 2 hours on an Intel Xeon @ 2.40GHz with 5GB of RAM.

The target corpus was processed with MetaMap [5] to map spans of text to UMLS Metathesaurus concepts. No disambiguation provided by MetaMap has been used (default option) so all possible concepts identified by MetaMap are available for model refinement. This is because the disambiguation has to be done by the proposed model that will provide different disambiguation results at each iteration. Candidate senses for ambiguous mappings are based on the result of the EM algorithm. Once the EM algorithm has converged, the most likely concept for each ambiguous word is selected. Once the MetaMap annotations are disambiguated, it is possible to identify which words tend to be used to denote a concept and which concepts are related to each other in the corpus used for refinement. Then, the statistics on term to concept and concept to concept relations are calculated and the EM algorithm is run with these models. The counts in equations 3 and 9 are updated adding these frequencies.

After the refinement of the word-concept model with the target corpus, we obtained a new set of β values: β_0=0.8315, β_1=0.0711, β_2=0.0975. In this case, much more weight is given to the 0-step model. This is because the refinement produces profiles that are considerably larger since more words are linked to the concepts derived from the new relations obtained from co-occurrences found in the corpus. The refinement process took approximately 1 day on an Intel Xeon @ 2.40GHz with 5GB of RAM.

Tables 1 and 2 show the probabilities of words for concepts linked to the ambiguous word *cold*. These concepts are C0009264 (cold temperature), C0024117 (chronic obstructive airway disease) and C0009443 (common cold). Table 1 shows the top words ranked by decreasing probability estimated from the KB for each concept. The top words typically come from synonyms of the concepts followed by words from related concepts.

In Table , 2 probabilities are higher for words linked to common uses of each of the senses of *cold*. For concept C0009443 (*common cold*), words associated with the preferred term, *common cold*, have a higher probability of occurring with the concept than terms *acute coryza* and *acute nasopharyngitis* due to their lower occurrence in the corpus. For concept *C0009264*, we find that even though the top words are the same, the remaining words change, accommodating the words from concepts that tend to co-occur with *C0009264*. The removed words like *frosbite* come from related concepts in the KB. These words do not seem to appear in the context of this concept in the corpus. Accuracy for the ambiguous word *cold* increases from 0.82 with the initial model to almost 0.9 with the refined model.

CUI:C0009264		CUI:C0024117		CUI:C0009443	
Word	Probability	Word	Probability	Word	Probability
cold	0.364455	chronic	0.154269	cold	0.162294
temperatur*	0.296702	obstruct*	0.153241	common	0.135191
low	0.035278	diseas*	0.132414	acut*	0.101557
frostbit*	0.004125	pulmonari*	0.069153	coryza	0.064442
refriger*	0.004123	copd	0.056498	nasopharyng*	0.056429
cryoscienc*	0.004120	lung	0.051505	rhiniti*	0.038304
shiver	0.004111	airwai*	0.025152	infect*	0.036675
hypothermia	0.003935	di*	0.019244	respiratori*	0.014978
freez*	0.002973	no	0.010173	diseas*	0.012207
cryosurgeri*	0.002852	pulm*	0.009851	viral	0.012051

Table 1: Probabilities for words (stemmed using Porter stemmer (stemmed form ended with *)) related to UMLS concepts *C0009264* (cold temperature), *C0024117* (chronic obstructive airway disease) and *C0009443* (common cold) ($P(w_i|c_j)$) related to the term *cold* after tuning the model.

CUI:C0009264		CUI:C0024117		CUI:C0009443	
Word	Probability	Word	Probability	Word	Probability
cold	0.511172	chronic	0.191317	cold	0.479154
temperatur*	0.172350	obstruct*	0.190142	common	0.274660
low	0.149683	diseas*	0.189146	acut*	0.017842
exposur*	0.000891	pulmonari*	0.138241	coryza	0.015999
studi*	0.000613	lung	0.048191	nasopharyng*	0.009669
gene	0.000605	copd	0.045529	infect*	0.007214
activ*	0.000592	cold	0.023634	rhiniti*	0.006479
stress	0.000568	airwai*	0.004760	respiratori*	0.003581
protein	0.000563	patient*	0.003714	upper	0.003322
rat	0.000475	di*	0.002211	viral	0.003287

Table 2: Probabilities for words (stemmed using Porter stemmer (stemmed form ended with *)) related to concepts *C0009264* (cold temperature), *C0024117* (chronic obstructive airway disease) and *C0009443* (common cold) ($P(w_i|c_j)$) related to the term *cold* after tuning the refined model

4.1. Disambiguation performance

The disambiguation results are compared to state-of-the-art algorithms already evaluated on the MSH WSD dataset (see Table 3), which are briefly described in turn. Machine Readable Dictionary (MRD) and 2-MRD build a concept profile vector assigning weights to words related to concepts [20, 26]. Automatic Extracted Corpus (AEC) uses the UMLS Metathesaurus to build queries used to collect training data for each ambiguous concept and then train a Naïve Bayes classifier. Structural Semantic Integration (SSI) and SSI+Information Content (SSI+IC) [38] use a model from the Metathesaurus that is enriched by co-occurrence information available from the UMLS distribution. PageRank [2] uses a graph based approach to perform the selection (we use the results presented in [16]). MRD+KMeans and AEC+KMeans combine MRD and AEC predictions with k-means [22]. CPIDF builds concept profiles for the whole of MEDLINE based on the same queries as the AEC method [21].

Naïve Bayes (NB) has been used as well as baseline, even though it is consider an upper bound of the results since a supervised method is expected to perform better than an unsupervised method on this task. NB results were obtained in 10-fold cross-validation using the MSH WSD data set. More details are available from [26].

We find that the proposed method outperforms existing unsupervised methods, including the AEC algorithm and the SSI+IC that already combine KB data and co-occurrence information from MEDLINE. This improvement is statistically significant ($p<0.00001$) compared to MRD and AEC, and the refined model significantly outperforms all the methods ($p<0.00001$). Improvements are significant even when Bonferroni corrections are applied to correct for multiple comparisons.

4.2. Document ranking

We have also evaluated the capability of the model to rank documents. The ranking benchmark queries are based on a subset of the MeSH headings available from the MSH WSD set. The queries are built using the words extracted for each one of the relevant concepts from UMLS. MeSH indexing of the citation is used as ground truth, as indicated before. For each retrieval evaluated method, the top 1000 retrieved documents sorted by relevance are

Method	Accuracy
MRD [26]	0.807
2-MRD [26]	0.780
AEC [26]	0.838
SSI [38]	0.743
SSI+IC [38]	0.860
PageRank [16]	0.786
MRD+KMeans [22]	0.874
AEC+KMeans [22]	0.865
CPIDF [21]	0.877
Naïve Bayes [26]	0.930
0-step model	0.829
2-step model	0.863
Refined model	0.891

Table 3: Disambiguation results on the MSH WSD data set. Baseline WSD methods results are shown with the reference to the article reporting the result.

selected for each query. *Trec_eval*[7] was used to perform the evaluation with the standard retrieval measures.

The baseline is based on a Kullback-Leibler retrieval (pr.simple_kl_dir) using Lemur [4]. We also included using pseudo-relevance feedback with the top 10 documents (pr.mixfb_kl_dir). In addition, we have implemented the pseudo-feedback method by Tao and Zhai [42] (kl_feedback) with the basic version of Kullback-Leibler retrieval (kl_divergence).

Another baseline is based on the supervised learning algorithm, Support Vector Machine (SVM)[8]. The model has been trained on a subset of 95 thousand citations, and documents in the evaluation set have been ranked

[7] http://trec.nist.gov/trec_eval

[8] http://ii.nlm.nih.gov/MTI_ML

according to the distance to the hyperplane. This method is an upper bound baseline, since it is not expected that any unsupervised method would improve it.

Results show that the proposed method significantly performs better than standard IR methods ($p < 0.001$, which is significant even when Bonferroni corrections are applied to correct for multiple comparisons), and that the refinement method outperforms pseudo-relevance feedback approaches.

Method	MAP	P@10	Rretr
pr.simple_kl_dir	0.2944	0.6146	7339
pr.mixfb_kl_dir	0.2985	0.6366	7435
kl_divergence	0.2955	0.5866	7399
kl_feedback [42]	0.3055	0.5902	7776
SVM	0.3544	0.6537	8317
2-step model	0.3025	0.6341	7372
Refined model	0.3176	0.6463	7799

Table 4: Document ranking results in terms of mean average precision (MAP), precision @ 10 (P@10) and the number of Relevant Retrieved (Rretr).

4.3. Discussion

We have proposed an estimation of a word-concept model that improves performance in disambiguation and document ranking by capturing statistical data from a large KB. In addition, we showed that the effectiveness of the proposed model can be further improved by combining corpora co-occurrence statistics. As shown in Table 3, the 2-step model performs better than any unsupervised method built solely on KB information. The refined model, that integrates information as well MEDLINE statistics, performs better than any of the compared KB methods. Regardless of the large number of potentially false positive relations extracted by co-occurrences, the model refinement improves the performance of the initial model only based on the KB. The improvement of the resulting model is global, since the refinement is done on the whole of the KB, and not by a single concept as in [25].

In the document ranking results, we showed significant improvement in ranking over other methods. This may in part be due to the disambiguation performance. The model integrates words from the synonyms and related concepts, which effectively improved baseline performance. Despite the disambiguation performance, the retrieval differences are not equally significant, which indicates that other factors beyond ambiguity are relevant for retrieval. Similar impact of WSD but with a different model and different data sets was observed in [47].

One of the current limitations of the method is the cost of traversing the KB to estimate the probabilities and the cost of the refinement, which is quite expensive with the current implementation. However, all this is only needed to be done once per concept. Once this is done, both disambiguation and document ranking are performed very quickly.

Additionally, larger k-step models will not only require more time ,but more memory as well, since the chance of relating all vocabulary words and concepts is higher. Notice that the number of words and concepts is over 1 million. On the other hand, it is unclear if there will be any positive effect in performance when larger k-step models are considered.

The method estimates word-concept probabilities. Higher order n-grams or terms could be considered as well in the model, which would use more precise features than single words (unigrams). A term-concept model could be estimated in the same way as presented in the Methods section but instead of words, higher order n-grams or terms should be used. Probabilities from models based on different features (i.e., unigrams and n-grams) could be combined to improve the performance of individual models. On the other hand, while terms are easily identified in the KB, the identification of these terms in text might not be perfect thus adding noise when using a term-concept model.

5. Conclusion and Future Work

Results show that the proposed method improves both word sense disambiguation and document ranking with respect to state-of-the-art methods. The current work considers only two traversal steps, further research is required to replace larger traversal steps efficiently.

The current estimation and refinement of the model does not rely on any training data and performance could be further enhanced if some training data is made available. Another possible application of the presented statistical method is text categorization, which could profit from the combination of knowledge based information and information derived from the training data. Another issue to be explored is to identify Gene Ontology [6] concepts, which is difficult to perform with traditional named entity resolution approaches. The Gene Ontology Annotation database [13] could be used to train the model.

We plan to extend this preliminary work to more general domains than the biomedical one, by using Wikipedia or more structured data sets like DBpedia. The proposed model has been evaluated in disambiguation and document ranking but we are interested in further evaluating it in other text mining tasks such as knowledge acquisition [35, 25], identification of context words for language generation [36], similarity between concepts and semantic distance [32].

The refinement of the model used in this work relies on co-occurrences, which potentially provides a large number of false positives. We would like to integrate additional relation extraction methods, but the difficulty is obtaining training data for all possible relation types. Although, methods based on open information extraction [8] could be considered.

The current refinement implementation does not try identifying new synonyms of existing concepts but only tries to quantify how often they are being used with a given concept. Furthermore, it does not try identifying new concepts missing in the KB. It could be worth exploring information extraction methods to identify new synonyms [10] of existing concepts and new concepts.

6. Acknowledgements

We thank anonymous reviewers for their very useful comments and suggestions. Special thanks to Bridget McInnes and Maika Vicente Navarro for proofreading the manuscript. The work was supported by the CICYT project TIN2011-24147 from the Spanish Ministry of Economy and Competitiveness (MINECO).

References

[1] Agirre, E., Soroa, A., 2009. Personalizing pagerank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 33–41.

[2] Agirre, E., Soroa, A., Stevenson, M., 2010. Graph-based word sense disambiguation of biomedical documents. Bioinformatics 26 (22), pp. 2889–2896.
URL http://www.ncbi.nlm.nih.gov/pubmed/20934991

[3] Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M., Wächter, T., 2009. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. BMC Bioinformatics 10 (1), 28.

[4] Allan, J., Callan, J., Collins-Thompson, K., Croft, B., Feng, F., Fisher, D., Lafferty, J., Larkey, L., Truong, T. N., Ogilvie, P., et al., 2003.

The lemur toolkit for language modeling and information retrieval. The Lemur Project.[WWW document] http://lemurproject. org (accessed 25 January 2012).

[5] Aronson, A., Lang, F., 2010. An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17 (3), pp. 229–236.

[6] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al., 2000. Gene ontology: tool for the unification of biology. Nature genetics 25 (1), pp. 25–29.

[7] Bahl, L. R., Jelinek, F., Mercer, R., 1983. A maximum likelihood approach to continuous speech recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on (2), pp. 179–190.

[8] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., Etzioni, O., 2007. Open information extraction for the web. In: IJCAI. Vol. 7. pp. 2670–2676.

[9] Berlanga, R., Nebot, V., Pérez, M., 2014. Tailored semantic annotation for semantic search. Web Semantics: Science, Services and Agents on the World Wide Web, Available online 18 July 2014, doi:10.1016/j.websem.2014.07.007.

[10] Blair, D. R., Wang, K., Nestorov, S., Evans, J. A., Rzhetsky, A., 2014. Quantifying the impact and extent of undocumented biomedical synonymy. PLoS computational biology 10 (9), e1003799.

[11] Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. the Journal of Machine Learning Research 3, pp. 993–1022.

[12] Bodenreider, O., 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Research 32 (suppl 1), D267–D270.

[13] Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R., 2004. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucleic acids research 32 (suppl 1), D262–D266.

[14] Cao, G., Nie, J.-Y., Bai, J., 2005. Integrating word relationships into language models. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. pp. 298–305.

[15] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., Blei, D. M., 2009. Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. pp. 288–296.

[16] Cheng, W., Preiss, J., Stevenson, M., 2012. Scaling up wsd with automatically generated examples. In: Proceedings of the 2012 Workshop

on Biomedical Natural Language Processing. Association for Computational Linguistics, pp. 231–239.

[17] Cohen, P. R., 1995. Empirical Methods for Artificial Intelligence. MIT Press, Cambridge, MA, USA.

[18] Gale, W., Church, K., Yarowsky, D., 1992. One sense per discourse. In: Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, pp. 233–237.

[19] Humphrey, S., Rogers, W., Kilicoglu, H., Demner-Fushman, D., Rindflesch, T., 2006. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. Journal of the American Society for Information Science and Technology (Print) 57 (1), pp. 96–113.

[20] Jimeno-Yepes, A., Aronson, A., 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. BMC Bioinformatics 11:565.

[21] Jimeno-Yepes, A., Aronson, A. R., 2012. Integration of umls and medline in unsupervised word sense disambiguation. In: 2012 AAAI Fall Symposium Series.

[22] Jimeno Yepes, A., Aronson, A. R., 2012. Knowledge-based and knowledge-lean methods combined in unsupervised word sense disam-

biguation. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. ACM, pp. 733–736.

[23] Jimeno-Yepes, A., Berlanga-Llavori, R., Rebholz-Schuchmann, D., 2010. Applications of ontologies and text mining in the biomedical domain. Ontology Theory, Management, and Design: Advanced Tools and Models, pp. 261–283.

[24] Jimeno-Yepes, A., Berlanga-Llavori, R., Rebholz-Schuhmann, D., 2009. Terminological cleansing for improved information retrieval based on ontological terms. In: Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval. ACM, pp. 6–14.

[25] Jimeno-Yepes, A., Berlanga-Llavori, R., Rebholz-Schuhmann, D., 2010. Ontology refinement for improved information retrieval. Information Processing & Management 46 (4), pp. 426–435.

[26] Jimeno-Yepes, A. J., McInnes, B. T., Aronson, A. R., 2011. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics 12 (1), 223.

[27] Leacock, C., Miller, G. A., Chodorow, M., 1998. Using corpus statistics and wordnet relations for sense identification. Computational Linguistics 24 (1), pp. 147–165.

[28] McInnes, B., June 2008. An unsupervised vector approach to biomedical term disambiguation: Integrating UMLS and Medline. In: Proceedings

of the ACL-08: HLT Student Research Workshop. Association for Computational Linguistics, Columbus, Ohio, pp. 49–54.
URL http://www.aclweb.org/anthology/P/P08/P08-3009

[29] McInnes, B., Pedersen, T., Carlis, J., 2007. Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. In: AMIA Annual Symposium Proceedings. Vol. 2007. American Medical Informatics Association, pp. 533–537.

[30] Navigli, R., Velardi, P., 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (7), pp. 1075–1086.

[31] Nebot, V., Berlanga, R., 2014. Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. Knowl. Inf. Syst. 38 (2), pp. 365–389.
URL http://dx.doi.org/10.1007/s10115-012-0590-x

[32] Pesquita, C., Faria, D., Falcao, A. O., Lord, P., Couto, F. M., 2009. Semantic similarity in biomedical ontologies. PLoS computational biology 5 (7), e1000443.

[33] Plaza, L., Jimeno-Yepes, A., Díaz, A., Aronson, A., 2011. Studying the correlation between different word sense disambiguation methods and

summarization effectiveness in biomedical texts. BMC Bioinformatics 12 (1), 355.

[34] Porter, M. F., 1980. An algorithm for suffix stripping. Program: electronic library and information systems 14 (3), pp. 130–137.

[35] Potter, S., 2003. A survey of knowledge acquisition from natural language. TMA of Knowledge Acquisition from Natural Language 2003.

[36] Reiter, E., Dale, R., Feng, Z., 2000. Building natural language generation systems. Vol. 33. MIT Press.

[37] Schuemie, M. J., Kors, J. A., Mons, B., 2005. Word sense disambiguation in the biomedical domain: an overview. Journal of Computational Biology 12 (5), pp. 554–565.

[38] Singh, B., van Mulligen, E. M., Kors, J. A., 2013. A scalable knowledge-based method for biomedical term disambiguation. JAMIA.

[39] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al., 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 25 (11), pp. 1251–1255.

[40] Spasic, I., Ananiadou, S., McNaught, J., Kumar, A., 2005. Text mining and ontologies in biomedicine: making sense of raw text. Briefings in bioinformatics 6 (3), pp. 239–251.

[41] Stevenson, M., Agirre, E., Soroa, A., 2011. Exploiting domain information for word sense disambiguation of medical documents. Journal of the American Medical Informatics Association, pp. 235–240.

[42] Tao, T., Zhai, C., 2006. Regularized estimation of mixture models for robust pseudo-relevance feedback. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 162–169.

[43] Ueda, N., Nakano, R., 1998. Deterministic annealing em algorithm. Neural Networks 11 (2), pp. 271–282.

[44] Weeber, M., Mork, J. G., Aronson, A. R., 2001. Developing a test collection for biomedical word sense disambiguation. In: Proceedings of the AMIA Symposium. American Medical Informatics Association, p. 746.

[45] Yarowsky, D., 1993. One sense per collocation. In: Proceedings of the workshop on Human Language Technology. Association for Computational Linguistics, pp. 266–271.

[46] Zhai, C., Lafferty, J., 2004. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS) 22 (2), pp. 179–214.

[47] Zhong, Z., Ng, H. T., 2012. Word sense disambiguation improves information retrieval. In: Proceedings of the 50th Annual Meeting of the

Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, pp. 273–282.