

# Query Expansion with Wikipedia

MARIA-HENDRIKE PEETZ and MARTIN LOPATKA

University of Amsterdam

To address the task of automated query expansion with compound queries we make use of the highly structured Wikipedia corpus. We consider every article linked to by a specific disambiguation page as a unique *sense*. Using language modeling analysis over the articles referred to in order to determine which *sense* is most likely intended by considering the co-occurrence of the different query terms in articles associated with candidate *senses*. Expansion terms are selected from the Wikipedia disambiguation sentence and added to the user defined query. An evaluation of this method is performed with the CLEF english collection and the TREC defined relevance assessments metrics. We observe a notable improvement in mean average precision in both of our experimental conditions, which differ only in the weighting of original query terms and expansion terms.

General Terms: Information Retrieval, Language, Algorithms

Additional Key Words and Phrases: Word Sense Disambiguation, Wikipedia, Language Modeling, Query Expansion

## 1. INTRODUCTION

Many everyday words and phrases have multiple senses. Be it as simple as **bank** or a more elaborate topic representing phrase like **Lord of the Rings**. Trying to retrieve documents about a certain topic, a user might have another sense in mind than the most often occurring sense; instead of the book trilogy the user is searching for a certain card game or the film and uses additional terms. There are two approaches to find appropriate documents, either using a user revision or automatic query expansion. For the latter, topic disambiguation is of crucial value for user experience in this context. Wikipedia, a user created encyclopedia, includes structurally distinct disambiguation pages that can be used for automated query expansion. We use this resource in a language modeling paradigm to estimate the most likely *sense* of a compound query.

### 1.1 Previous Work

Word sense disambiguation (WSD) has been a subject of research for several years. Approaches range from knowledge based algorithms [Lesk 1986] and [Navigli and Velardi 2005] to data-driven methods using supervised and unsupervised learning [Ng and Lee. 1996; Yarowsky 1995; Bauer 2007]. Further work has already been done using Wikipedia for WSD [Mihalcea 2007; Mihalcea and Csomai 2007] but not applying it to a global query expansion intended for queries to another corpus, similar to the use of wordnet. Unlike it has been done with wordnet [Gonzalo et al. 1998] or using latent semantic indexing [Deerwester et al. 1990]. Typical local query expansion techniques used are relevance feedback and pseudo-relevance feedback [Lavrenko and Croft 2001], based on the work by [Rocchio 1971]. Before expanding a query we have to define its hardness, denoting the difficulty to retrieve relevant documents given the query. There are different approaches to find different query

---

We would sincerely thank the creators of the Wikipedia XML corpus. Their work saved us a lot of work and trouble, and made the implementation of our idea within the short amount of time provided possible.

meanings, e.g.: [clu ] clusters documents given a query as many distinct clusters given a certain query indicate its difficulty [Carmel et al. 2006]. A good survey on query expansion can be found in [Xu and Croft 1996].

## 1.2 Outline

Using a language modeling approach based on the common sense based category system of Wikipedia seems feasible for query expansion. In our approach we will define a query as hard, as soon as a subphrase has a Wikipedia disambiguation page (see Section 3.4). We will apply a global query expansion by finding out the correct meaning based on a language model (see Section 3.3) and then using important terms for that meaning as expansion terms (see Section 3.6). We will further evaluate the performance of this approach using the CLEF British newspaper corpus (see Section 4) and propose some future extensions (see Section 5.1).

## 2. WHY WIKIPEDIA?

### 2.1 Wikipedia

Wikipedia is a free, online encyclopedia. It is free in two ways; its contents are freely available and everyone is free to contribute to the corpus so long as they adhere to the Wikipedia standard format. Everyone is able to create and modify existing articles, but also suggest the deletion of inappropriate or wrong content. Since its foundation in 2001, Wikipedia has evolved to the biggest encyclopedia in the world, being available in more than 200 languages with the English version alone featuring 2,373,410 articles, created by the 75000 registered users [Wikipedia 2008c].

Wikipedia has strict style guidelines, which are actively enforced by the Wikipedia community in order to maintain a consistent quality of content and structure. There are guidelines for every kind of entry, ranging from lead paragraph [Wikipedia 2008b] to disambiguation page [Wikipedia 2008a]. Still, the encyclopedia is created by the masses and contains mistakes — for corpus analysis the incompetent usage of tags and differences in writing style is a big problem.

The basic entity of an entry in Wikipedia is an article. An article can have different forms, from being a template that has to be imported to fix a specific style for all elements of a category system to disambiguation pages. Disambiguation pages for a phrase like **Seven senses of ambiguity** have a very short description of the different topics, in this example we may be referring either to the novel or the literary criticism. These descriptions contain mostly one link to an article covering this material. If there is no article available, the disambiguation page offers a short and concise explanation of the topic [Wikipedia 2008a]. Every article in Wikipedia should consist of a summarizing introductory (lead) paragraph and a more detailed body. Furthermore, it might include templates. Also, content is supposed to be backed-up by citations, articles without sufficient literature are edited or flagged as incomplete.

Still, disambiguation pages and sense descriptions might be opinionated, missing a higher neutral instance. With its large community participating in its creation we assume this opinion being that of the masses [Ciffolilli 2003].

## 2.2 Wordnet vs. Wikipedia

Wordnet [Fellbaum 1998] is an electronic database of word relations written by professional lexicographers. It has been a valuable resource to computational linguists for years, with all open-class words being parts of a synonym set which are interconnected by semantic relations. Each word's sense is described with a short sentence, each open-class word is again linked to its senses. As Wikipedia, it is self-contained within its link structure.

Wordnet has at least two drawbacks compared to the usage of Wikipedia. All synonym sets are created by professionals. This means, the descriptions and their linking is correct. However, the average user does not always think correctly. With Wikipedia being a resource created by the masses, we hope to take advantage of intuitive semantics and not a scientific approach. Further, *topics* are not always covered within wordnet. The phrase **Seven senses of ambiguity** is not explained in Wordnet as it is more concentrated on the basic elements of language (e.g.: **bank**) as compared to phrases. Additionally, wikipedia offers with the articles connected to a certain sense a lot more topic related terms than a short description of a synonym in wordnet<sup>1</sup>.

## 3. MODELING

### 3.1 Data

We used the English XML-corpus created at the University of Amsterdam in August 2007 described in [WikiXML].

We take advantage of the highly structured XML format to facilitate our modeling procedure. Every wikipedia page has two major divisions, the **head** and the **body**. The **head** stores all metadata like the title and the document ID, whereas the **body** contains the relevant content. In the following we will only focus on the for us important features, a complete survey on this corpus can be found in [WikiXML].

**3.1.1 Articles.** In our terminology everything is an article, even templates and disambiguation pages. An article may use different templates and tables, which means a lot of metadata within the body. Additionally, each article might link to an arbitrary number of other articles and can emphasize text. For our purposes we focus on those articles linked by disambiguation pages and which provide specific information exclusive to *one* sense of the title term. These articles are supposed to have a structure which can roughly be divided into two parts, *introduction* and *content*. Examining an average Wikipedia article, we can see that the introduction is separated from the content by a table of contents which is automatically created. Within the corpus this is marked by a `<div id="wx_to"/>` tag. Articles without such a marker are most incomplete and called *stubs* [Wikipedia 2008b].

**3.1.2 Disambiguation pages.** Within the disambiguation body we can find list items, denoted by standard list item HTML tag. From the data we have examined, we can safely assume that each list item is a unique explanation of the title. Further, each disambiguation page uses a disambiguation template with the number

<sup>1</sup>Of course, advanced lesk algorithms use related words descriptions as well to create overlap, still wikipedia has way more words.

9950598. A disambiguation item does not necessarily need to be linked to another more precise article on that topic and may just consist of an explaining sentence (see Figure 1).

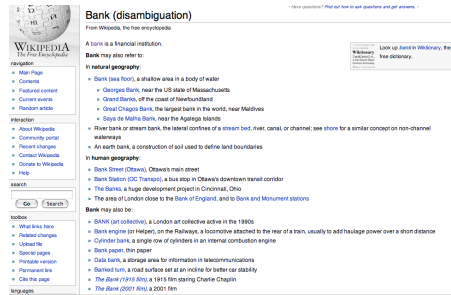


Fig. 1. An example disambiguation page

3.1.3 *Links*. Internal links are used to browse between wikipedia articles. A link contains three important pieces of information, the *document ID*, the *title* of the article, and *describing text*. The document ID can be used to access the document within the corpus as this can be the filename. Looking at two examples in Figure 2 the to-be-shown text is between the link tags (<a>). When the rendering machine has only these links, it uses the title of the link and it therefore becomes a part of the displayed text. Furthermore, there are dead links, or not yet written articles.

```
<a href="/wiki/Herma" title="Herma" wx:linktype="known"
  wx:pagename="Herma" wx:page_id="86289" id="wx7">Herma</a>

<a href="/wiki/Herma" title="Herma" wx:linktype="known"
  wx:pagename="Herma" wx:page_id="86289" id="wx7"/>
```

Fig. 2. Two examples of a link

## 3.2 Levels of word sense

Given the title of a disambiguation page, the different *word senses* are all displayed as disambiguation items. We now describe three different levels of information a word sense might have.

3.2.1 *Disambiguation item*. The most compact representation of a meaning is the disambiguation sentence, see [Wikipedia 2008a], explaining the context of a meaning. In its guidelines, Wikipedia tries to enforce the *contextual* description as compared to a *dictionary* description. Assuming this context is correct, this is the best source of information pertaining to a topic we can have. Unfortunately, due to the lack of words in a short sentence, its probability model might be sensitive to outliers.

**3.2.2 Introduction.** According to the Wikipedia guidelines the introduction must establish a context and be a concise summary of the topic. It is thus more elaborate [Wikipedia 2008b] but still compressed source of information about the word sense. Nevertheless there might be words used that do not have anything to do with the sense. Sometimes an article is linked to a sense, which covers the sense in one section of the article, but this sense is not mainly the topic (for the term **Canton** there is an entry of **flag terminology**). Thus, the introductory paragraph is not a summary of this topic, which decreases its reliability.

**3.2.3 Body.** The body is a detailed description of the topic. It does not necessarily have all information in a compressed way, it is rather elaborate and has some terms which are semantically not very close to the original meaning. Still, we assume that the use of words is representative for the specific topic, especially if we use a general background probability for the English language.

### 3.3 Building a model

Given a query  $q$  consisting of query terms  $t$ , we would like to maximize the the probability that a query is generated by the sense model  $\theta_s$ , with  $s \in S$ , the set of all senses:

$$\arg \max_{s \in S} p(q | \theta_s) \quad (1)$$

Assuming independence between query terms we can derive:

$$p(q | \theta_s) = \prod_{t \in q} p(t | \theta_s) \quad (2)$$

All formulae should be  $p(q | \theta_s, C)$ , with  $C$  being the Wikipedia corpus, but for the sake of readability we omit this.

Given a meaning  $s$ , we have a disambiguation model  $\theta_s^d$ , a introduction model  $\theta_s^i$  and a body model  $\theta_s^b$ . We can now, using 4-step interpolation smoothing [Zhai and Lafferty 2001], calculate  $p(t | \theta_s)$  the following:

$$p(t | \theta_s) = \lambda_1 \cdot p(t | \theta_s^d) + \lambda_2 \cdot p(t | \theta_s^i) + \lambda_3 \cdot p(t | \theta_s^b) + \lambda_4 \cdot p(t), \quad (3)$$

with  $\sum_i \lambda_i = 1$  and  $p(t)$  being the background probability (see Section 3.5).

Using a simple language modeling approach, we assume time-invariance and independence of the words within each section, thus we can say

$$p(t | \theta_s^d) = \frac{f(t, d)}{\sum_{j \in d} f(j, d)} \quad (4)$$

$$p(t | \theta_s^i) = \frac{f(t, i)}{\sum_{j \in i} f(j, i)} \quad (5)$$

$$p(t | \theta_s^b) = \frac{f(t, b)}{\sum_{j \in b} f(j, b)} \quad (6)$$

with  $f(j, d)$  denoting the frequency of a term  $j$  in the disambiguation sentence,  $f(j, i)$  denoting the frequency of a term  $j$  in the introduction paragraph and  $f(j, b)$  denoting the frequency of a term  $j$  in the article as a whole, the body.

### 3.4 Finding ambiguous words

Given a query  $q$ , we wish to know which terms or combination of terms is ambiguous. Disambiguation titles often do not consist of only one word (e.g.: book or movie titles). Accordingly, we take all  $n$ -grams of the words until  $n$  is the maximal length of a title and compare these  $n$ -grams with the titles in our list. If we have a match we find the highest probable meaning and add the expansion terms (see Section 3.6) to the list of query terms. In this way our query expansion process is transparent to the user.

### 3.5 Background Probability

The background probability  $p(t)$  for a term should use the whole corpus of wikipedia as a resource. Unfortunately, due to time restrictions we only use 5% of the articles as resource to assemble word probabilities. This might cause serious problems with respect to smoothing and cause zero probabilities. To compensate for this shortcoming we use random sampling to select the 5% used.

### 3.6 Expansion terms

We now have a certain sense  $m$  for a word (sequence)  $t$  and are searching for an effective way to expand our query. Query expansion is a delicate process; on one hand it might cause a drifting in topic which is intended by the user, on the other hand it might be too general and fails improve search performance. We considered several methods of expanding a query given a sense  $m$ , but due to time restrictions we implemented a combination that seemed the sensible.

Words within a sentence are usually only linked when they are considered as important [Mihalcea and Csomai 2007] or generally, a good means of disambiguation. We could therefore use the words or titles of the links as query extensions. Unfortunately there are phrases that contain no links at all and therefore consist of a word distribution and with no expansion terms at all. This leads to a different approach which uses the whole sentence. The disadvantage of this approach is clearly the use of irrelevant words (stopwords) and re-occurring words. We could therefore take the set of words within the sentence and apply a stopword list to circumvent these issues.

As a further resource of information, the linked article might be considered. In this case the choosing expansion terms might be manifold; ranging from the usage of link terms over emphasized words (indicated by italics) to section titles. Furthermore, all these terms could be used, using different weights, based on their type but also on the occurrence within the document.

Different senses can have similar word distributions and sometimes different  $p(q | \theta_s)$  are very close together. To avoid strong query drift to a sense that does not have a high probability the usage of an  $\epsilon$ , indicating the difference between the two senses with the highest probability. The smaller  $\epsilon$  is, the less certain we can be about the topic we are looking for. With this approach we may assume that the user is not looking for both senses.

Our final implementation uses the disambiguation sentence and an  $\epsilon$  to decide whether the query is expanded or not.

### 3.7 Implementation

The fundamental code of this approach uses the programming language `python`, version 2.5. We make extensive use of regular expressions to clean up the articles for our needs. Furthermore we have developed an index structure for different senses, allowing us fast and efficient access after the senses are indexed.

**3.7.1 Link to article.** A disambiguation sentence can have different links to related articles. We are only interested in the ID of the article containing more information about the sense described in the disambiguation sentence. Naïvely, we simply take the link with the same title or embedded text as the disambiguation page, assuming that the linked articles' title is the same as the disambiguation page. This assumption is not necessarily valid, i.e. for the disambiguation page of the word `herm`, having linked articles with the titles `Hermaphrodite`, `Herm`, `Hermit` or `Herm, Landes`. This naïve approach would only match one of the articles links and omit the rest. Using the *Jacquard Similarity* [Jaccard 1912] between the articles' title and the disambiguation terms helps at least to also find `Herm, Landes` but still does not find the remaining ones. We then make use of a typical dictionary language feature of stating the article describing the sense first, as also proposed in the guidelines of Wikipedia [Wikipedia 2008a]. This approach works for at least a small subset of data. Additionally, we deleted the little phrase (`disambiguation`) out of every disambiguation page's title.

**3.7.2 Word distributions.** Even though emphasizing words increases their importance we ignored all mark-up; emphasizing mark-up was completely removed as well as section and link mark-up. In the latter case we remember the fact that something is a link for further extensions of the program. Templates were completely removed, but as templates contain important information (e.g. categories of animals), we tried to extract as much content as possible from the mark-up prior to its deletion.

Using the information about our data, explained in Section 3.1, we used regular expressions to extract the different text excerpts. The elaboration of all regular expressions would be out of the scope of this paper, but let us discuss one typical example: the extraction of the introduction section. We first had to find the body of the article: `<body>(.*?)</body>`, using non-greediness and thus respecting the tag ends. From that we could first try to match the table of contents, which is supposed to end the introductory paragraph:

```
<div id="wx_article">(.*?)<div id="wx_toc"\>/>
```

If the article is too small and does not contain enough sections, a table of contents is not included and we have to search for the end of the first section:

```
<div id="wx_article">(.*?)</wx:section>
```

For articles that are not valid XML, these regular expressions may yield inconsistent results. We therefore, at the end, delete everything that seems to be XML, before tokenizing. The remaining words were tokenized using the tokenizer distributed with the tree tagger [Schmid et al. 2007].

**3.7.3 Index structure.** Our index structure has a list storing all titles of the senses and the link to a list of senses. As this data structure might become very large, we decided to use a filename as link, which yields a file containing a list of

meanings associated with the title of the disambiguation page. This file is only read in if the title within our list matches. This is due to the practical limitation of system memory. Each sense is an instance of the class `Meaning`, storing word distributions for the disambiguation sentence as well as for the introduction section and for the body of the associated article. Further, it stores the articles id and a list of expansion terms. Each sense  $s$  can, given a query  $q$ , return  $p(q | \theta_s)$  with  $\theta_s$  being based on the word distribution using the model described in section 3.3.

## 4. EXPERIMENTS

### 4.1 Setting

The experimental paradigm we have used to evaluate our query expansion process was designed as follows. We used the LEMUR information retrieval platform [CMU and UMASS ] using a stopwords list distributed by SNOWBALL. We did not use stemming. As a test corpus, we used the CLEF newspaper collection, containing 169477 documents with 89924497 tokens and 312375 types. The created index has a size of 727 MB and took approximately two hours to build on an average home computer. Analogously to our disambiguation index, we applied the Tree-Tagger tokenizer [Schmid et al. 2007].

### 4.2 Results

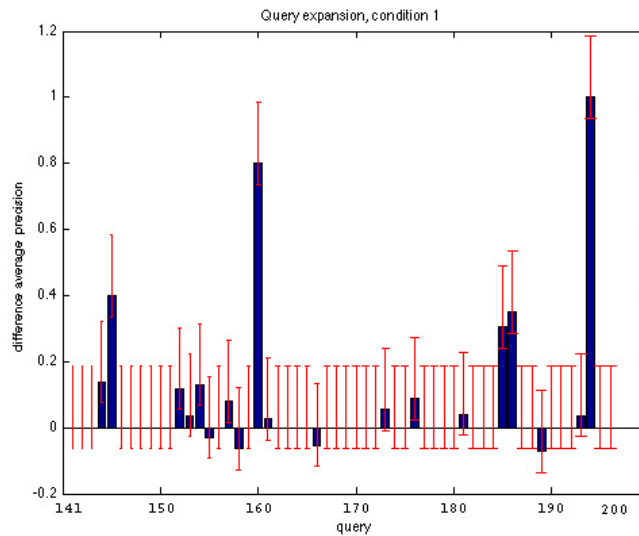


Fig. 3. Unweighted Expansion Terms

The baseline condition to which we compare our approach uses only the titles of the queries since using the title and description would yield queries that do not resemble those submitted by real users [Jansen et al. 1998]. All queries contain between two and five terms. The baseline performance achieved was 2.14% mean

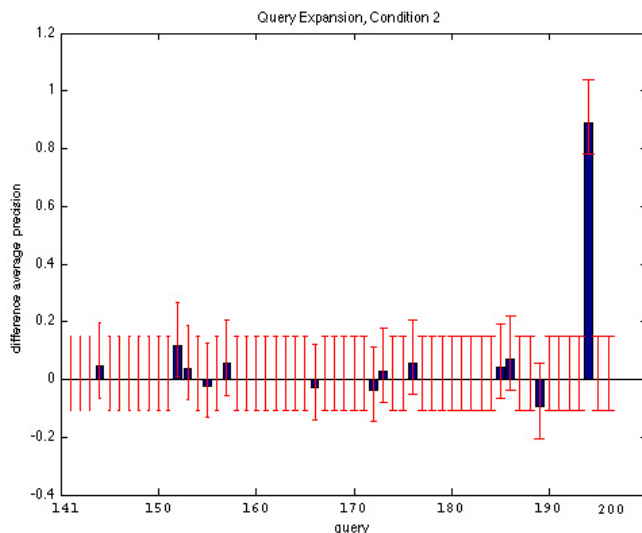


Fig. 4. Weighted Expansion Terms

average precision (MAP) within the top 10 documents retrieved. This result was scored using the TREC relevance assessment metric provided with the CLEF collection.

Choosing appropriate smoothing parameters is always subject to experimentation. Due to time restrictions we decided to use parameters based on common sense rather than doing extensive experimentation. With the disambiguation sentence being the most reliable information source, we set  $\lambda_1$  to 0.5, bearing in mind that due to the usually small size of the disambiguation sentence more weight should be put on this language model. We used incrementally decreasing  $\lambda$  values for larger portions of each article,  $\lambda_2$  being 0.3 and  $\lambda_3$  0.15. As such, the background probability acts only to smooth out zero probabilities and is set to 0.05 for  $\lambda_4$ .

Using only the title terms from the standard queries we generated a list of expansion terms for each query and repeated an identical evaluation process. The resulting MAP was 5.09%, a notable improvement from the baseline condition. Due to the occasional occurrence of query drift we repeated the evaluation process weighting the expansion terms by a factor of 0.2. This led to a decrease in performance with a MAP of 4.6%. A per-query analysis of our results can be seen below. *Condition 1* corresponds to the expanded query with no weighting (refer to Figure 3) and *Condition 2* shows the results of weighting the expansion terms (see Figure 4). *Condition 1* is for more query terms significantly better than *Condition 2*, but both being in more cases significantly better than the baseline.

The individual query analysis allows us to examine specific instances where performance changed most dramatically. Substantial improvement was seen in query 198 in the unweighted condition. The original query was **Honorary Oscar for Italian Directors**. The generated expansion terms were: **unpaid**, **part-time**,

diplomatic, consul, academy, awards, popularly, called, oscars, film, person, responsible, orchestrating, artistic, dramatic and aspects. The addition of the terms `academy`, `awards`, and `film` yield documents that more closely match the relevance assessment criteria provided. An example in which precision declines with the inclusion of expansion terms is seen in query 141. The original query is `Wimbledon Lady Winners`, the expansion terms added were `championships`, `championship`, `grass`, `court`, `tennis`, `tournament`, `grand` and `slam`. Clearly, the additional terms allow for more general candidate results and thus demonstrate the occurrence of query drift.

## 5. CONCLUSION

### 5.1 Future Work

One limitation of this approach is the potential for query drift. In cases where the Wikipedia disambiguation page links to several slightly different senses of a term that are still semantically close, a large number of expansion terms are generated. It appears that reducing the emphasis on expansion terms by means of reduced weighting is ineffective. Future improvement would necessitate the exploration of alternative means of accounting for query drift. One such approach would be the use of a custom stopword list that eliminates high frequency words specific to the disambiguation descriptions.

A practical consideration regarding the use of Wikipedia is the volatile nature of some content. Being an inherently dynamic corpus provides the advantage of being constantly up to date which is useful as users tend to submit queries pertaining to current events, however the use of an XML dump is a suboptimal approach. The re-implementation of our algorithm to function as a crawler that dynamically retrieves expansions terms from the live corpus would take advantage of the dynamic nature of Wikipedia.

Implementationally, we noticed that in some disambiguation pages the most common usage of a term or topic is described in an introductory sentence, occurring before the itemize environment. The code should be adjusted accordingly.

Diverging from the conventions of a language modeling approach we ignored document length. Previous work in language modeling shows that longer documents tend to have higher probabilities of general word occurrence and are therefore preferred over shorter ones. The use of a document length dependent prior probability would account for any problems resulting from the omission of explicit document length consideration.

We observed that some query expansion terms should be treated as a single term as they relate to a single subject. This is shown in one of our expansion examples, `grand` and `slam` are treated as two distinct expansion terms but would be more valuable if used as a single term.

Further experimentation regarding the use of different smoothing parameters and  $\epsilon$  would be beneficial in finding optimal performance settings. With the availability of Wikipedia in a multitude of different languages it should be feasible to use this approach for other languages than English. Despite smaller corpora, we expect equally good results as the approach itself is language independent.

In order to assess the suitability of the Wikipedia corpus as a means of generating

appropriate expansion terms a comparative study between the approach described here and a similar approach using Wordnet would be valuable.

## 5.2 Discussion

Query expansion using terms drawn from Wikipedia disambiguation pages shows promise as a method for automatically addressing ambiguity in user performed searches. Due to the user-centered nature of the content contained within this corpus we believe it provides a more natural distribution of terminology and topics than Wordnet. This augments the usefulness of a language modeling approach in deciding which terms are appropriate in expanding only one *sense* of a query term. Query disambiguation by the method described herein may be refined to yield robust disambiguation for queries to any English corpus.

## REFERENCES

- <http://clusty.com/>.
- BAUER, D. 2007. Learning the semantics of wikipedia hyperlinks. B.Sc. Thesis.
- CARMEL, D., YOM-TOV, E., DARLOW, A., AND PELLEG, D. 2006. What makes a query difficult? In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 390–397.
- CIFFOLILLI, A. 2003. Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of wikipedia. *First Monday* 8, 12 (December).
- CMU, T. L. P. AND UMASS. Lemur toolkit for language modeling and information retrieval.
- DEERWESTER, S. C., DUMAIS, S. T., LANDAUER, T. K., FURNAS, G. W., AND HARSHMAN, R. A. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41, 6, 391–407.
- FELLBAUM, C. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- GONZALO, J., VERDEJO, F., CHUGUR, I., AND CIGARRÁN, J. M. 1998. Indexing with wordnet synsets can improve text retrieval. *CoRR cmp-lg/9808002*.
- JACCARD. 1912. The distribution of the flora of the alpine zone. *New Phytologist* 11, 37–50.
- JANSEN, B. J., SPINK, A., BATEMAN, J., AND SARACEVIC, T. 1998. Real life information retrieval: a study of user queries on the web. *SIGIR Forum* 32, 1, 5–17.
- LAVRENKO, V. AND CROFT, W. B. 2001. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 120–127.
- LESK, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*. Toronto.
- MIHALCEA, R. 2007. Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT 2007*.
- MIHALCEA, R. AND CSOMAI, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, New York, NY, USA, 233–242.
- NAVIGLI, R. AND VELARDI, P. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. Vol. 27.
- NG, H. AND LEE, H. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of ACL*. New Mexico.
- ROCCHIO, J. 1971. *SMART Retrieval System*. Prentice Hall, Chapter Relevance Feedback in Information Retrieval, 313–323.
- SCHMID, H., BARONI, M., ZANCHETTA, E., AND STEIN, A. 2007. The enriched treetagger system. In *proceedings of the EVALITA 2007 workshop*.
- WIKIPEDIA. 2008a. Guideline for disambiguation pages.
- WIKIPEDIA. 2008b. Guideline for lead sections.
- WIKIPEDIA. 2008c. Wikipedia statistics.
- WIKIXML. Wikipedia xml corpus.
- XU, J. AND CROFT, W. B. 1996. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 4–11.
- YAROWSKY, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL 1995*. Cambridge.
- ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Research and Development in Information Retrieval*. 334–342.

May 20, 2008