Wikipedia as Sense Inventory to Improve Diversity in Web Search Results

Celina Santamaría, Julio Gonzalo and **Javier Artiles**
nlp.uned.es
UNED, c/Juan del Rosal, 16, 28040 Madrid, Spain
celina.santamaria@gmail.com julio@lsi.uned.es javart@bec.uned.es

Abstract

Is it possible to use sense inventories to improve Web search results diversity for one word queries? To answer this question, we focus on two broad-coverage lexical resources of a different nature: WordNet, as a de-facto standard used in Word Sense Disambiguation experiments; and Wikipedia, as a large coverage, updated encyclopaedic resource which may have a better coverage of relevant senses in Web pages.

Our results indicate that (i) Wikipedia has a much better coverage of search results, (ii) the distribution of senses in search results can be estimated using the internal graph structure of the Wikipedia and the relative number of visits received by each sense in Wikipedia, and (iii) associating Web pages to Wikipedia senses with simple and efficient algorithms, we can produce modified rankings that cover 70% more Wikipedia senses than the original search engine rankings.

1 Motivation

The application of Word Sense Disambiguation (WSD) to Information Retrieval (IR) has been subject of a significant research effort in the recent past. The essential idea is that, by indexing and matching word senses (or even meanings) , the retrieval process could better handle polysemy and synonymy problems (Sanderson, 2000). In practice, however, there are two main difficulties: (i) for long queries, IR models implicitly perform disambiguation, and thus there is little room for improvement. This is the case with most standard IR benchmarks, such as TREC (trec.nist.gov) or CLEF (www.clef-campaign.org) ad-hoc collections; (ii) for very short queries, disambiguation may not be possible or even desirable. This is often the case with one word and even two word queries in Web search engines.

In Web search, there are at least three ways of coping with ambiguity:

- Promoting diversity in the search results (Clarke et al., 2008): given the query "*oasis*", the search engine may try to include representatives for different senses of the word (such as the Oasis band, the Organization for the Advancement of Structured Information Standards, the online fashion store, etc.) among the top results. Search engines are supposed to handle diversity as one of the multiple factors that influence the ranking.

- Presenting the results as a set of (labelled) clusters rather than as a ranked list (Carpineto et al., 2009).

- Complementing search results with search suggestions (e.g. "*oasis band*", "*oasis fashion store*") that serve to refine the query in the intended way (Anick, 2003).

All of them rely on the ability of the search engine to cluster search results, detecting topic similarities. In all of them, disambiguation is implicit, a side effect of the process but not its explicit target. Clustering may detect that documents about the Oasis band and the Oasis fashion store deal with unrelated topics, but it may as well detect a group of documents discussing why one of the Oasis band members is leaving the band, and another group of documents about Oasis band lyrics; both are different aspects of the broad topic Oasis band. A perfect hierarchical clustering should distinguish between the different Oasis senses at a first level, and then discover different topics within each of the senses.

Is it possible to use sense inventories to improve search results for one word queries? To answer

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1357–1366,
Uppsala, Sweden, 11-16 July 2010.

this question, we will focus on two broad-coverage lexical resources of a different nature: WordNet (Miller et al., 1990), as a de-facto standard used in Word Sense Disambiguation experiments and many other Natural Language Processing research fields; and Wikipedia (www.wikipedia.org), as a large coverage and updated encyclopedic resource which may have a better coverage of relevant senses in Web pages.

Our hypothesis is that, under appropriate conditions, any of the above mechanisms (clustering, search suggestions, diversity) might benefit from an explicit disambiguation (classification of pages in the top search results) using a wide-coverage sense inventory. Our research is focused on four relevant aspects of the problem:

1. Coverage: Are Wikipedia/Wordnet senses representative of search results? Otherwise, trying to make a disambiguation in terms of a fixed sense inventory would be meaningless.

2. If the answer to (1) is positive, the reverse question is also interesting: can we estimate search results diversity using our sense inventories?

3. Sense frequencies: knowing sense frequencies in (search results) Web pages is crucial to have a usable sense inventory. Is it possible to estimate Web sense frequencies from currently available information?

4. Classification: The association of Web pages to word senses must be done with some unsupervised algorithm, because it is not possible to hand-tag training material for every possible query word. Can this classification be done accurately? Can it be effective to promote diversity in search results?

In order to provide an initial answer to these questions, we have built a corpus consisting of 40 nouns and 100 Google search results per noun, manually annotated with the most appropriate Wordnet and Wikipedia senses. Section 2 describes how this corpus has been created, and in Section 3 we discuss WordNet and Wikipedia coverage of search results according to our testbed. As this initial results clearly discard Wordnet as a sense inventory for the task, the rest of the paper mainly focuses on Wikipedia. In Section 4 we estimate search results diversity from our testbed, finding that the use of Wikipedia could substantially improve diversity in the top results. In Section 5 we use the Wikipedia internal link structure and the number of visits per page to estimate relative frequencies for Wikipedia senses, obtaining an estimation which is highly correlated with actual data in our testbed. Finally, in Section 6 we discuss a few strategies to classify Web pages into word senses, and apply the best classifier to enhance diversity in search results. The paper concludes with a discussion of related work (Section 7) and an overall discussion of our results in Section 8.

2 Test Set

2.1 Set of Words

The most crucial step in building our test set is choosing the set of words to be considered. We are looking for words which are susceptible to form a one-word query for a Web search engine, and therefore we should focus on nouns which are used to denote one or more named entities. At the same time we want to have some degree of comparability with previous research on Word Sense Disambiguation, which points to noun sets used in Senseval/SemEval evaluation campaigns[1]. Our budget for corpus annotation was enough for two persons-month, which limited us to handle 40 nouns (usually enough to establish statistically significant differences between WSD algorithms, although obviously limited to reach solid figures about the general behaviour of words in the Web).

With these arguments in mind, we decided to choose: (i) 15 nouns from the Senseval-3 lexical sample dataset, which have been previously employed by (Mihalcea, 2007) in a related experiment (see Section 7); (ii) 25 additional words which satisfy two conditions: they are all ambiguous, and they are all names for music bands in one of their senses (not necessarily the most salient). The Senseval set is: {*argument, arm, atmosphere, bank, degree, difference, disc, image, paper, party, performance, plan, shelter, sort, source*}. The bands set is {*amazon, apple, camel, cell, columbia, cream, foreigner, fox, genesis, jaguar, oasis, pioneer, police, puma, rainbow, shell, skin, sun, tesla, thunder, total, traffic, trapeze, triumph, yes*}.

For each noun, we looked up all its possible senses in WordNet 3.0 and in Wikipedia (using

[1] http://senseval.org

Table 1: Coverage of Search Results: Wikipedia vs. WordNet

	Wikipedia		WordNet	
	# senses available/used	# documents assigned to some sense	# senses available/used	# documents assigned to some sense
Senseval set	242/100	877 (59%)	92/52	696 (46%)
Bands set	640/174	1358 (54%)	78/39	599 (24%)
Total	882/274	2235 (56%)	170/91	1295 (32%)

Wikipedia disambiguation pages). Wikipedia has an average of 22 senses per noun (25.2 in the Bands set and 16.1 in the Senseval set), and Wordnet a much smaller figure, 4.5 (3.12 for the Bands set and 6.13 for the Senseval set). For a conventional dictionary, a higher ambiguity might indicate an excess of granularity; for an encyclopaedic resource such as Wikipedia, however, it is just an indication of larger coverage. Wikipedia entries for *camel* which are not in WordNet, for instance, include the Apache Camel routing and mediation engine, the British rock band, the brand of cigarettes, the river in Cornwall, and the World World War I fighter biplane.

2.2 Set of Documents

We retrieved the 150 first ranked documents for each noun, by submitting the nouns as queries to a Web search engine (Google). Then, for each document, we stored both the snippet (small description of the contents of retrieved document) and the whole HTML document. This collection of documents contain an implicit new inventory of senses, based on Web search, as documents retrieved by a noun query are associated with some sense of the noun. Given that every document in the top Web search results is supposed to be highly relevant for the query word, we assume a "one sense per document" scenario, although we allow annotators to assign more than one sense per document. In general this assumption turned out to be correct except in a few exceptional cases (such as Wikipedia disambiguation pages): only nine documents received more than one WordNet sense, and 44 (1.1% of all annotated pages) received more than one Wikipedia sense.

2.3 Manual Annotation

We implemented an annotation interface which stored all documents and a short description for every Wordnet and Wikipedia sense. The annotators had to decide, for every document, whether there was one or more appropriate senses in each of the dictionaries. They were instructed to provide annotations for 100 documents per name; if an URL in the list was corrupt or not available, it had to be discarded. We provided 150 documents per name to ensure that the figure of 100 usable documents per name could be reached without problems.

Each judge provided annotations for the 4,000 documents in the final data set. In a second round, they met and discussed their independent annotations together, reaching a consensus judgement for every document.

3 Coverage of Web Search Results: Wikipedia vs Wordnet

Table 1 shows how Wikipedia and Wordnet cover the senses in search results. We report each noun subset separately (*Senseval* and *bands* subsets) as well as aggregated figures.

The most relevant fact is that, unsurprisingly, Wikipedia senses cover much more search results (56%) than Wordnet (32%). If we focus on the top ten results, in the bands subset (which should be more representative of plausible web queries) Wikipedia covers 68% of the top ten documents. This is an indication that it can indeed be useful for promoting diversity or help clustering search results: even if 32% of the top ten documents are not covered by Wikipedia, it is still a representative source of senses in the top search results.

We have manually examined all documents in the top ten results that are not covered by Wikipedia: a majority of the missing senses consists of names of (generally not well-known) companies (45%) and products or services (26%); the other frequent type (12%) of non annotated document is disambiguation pages (from Wikipedia and also from other dictionaries).

It is also interesting to examine the degree of overlap between Wikipedia and Wordnet senses. Being two different types of lexical resource, they might have some degree of complementarity. Table 2 shows, however, that this is not the case: most of the (annotated) documents either fit Wikipedia senses (26%) or both Wikipedia and Wordnet (29%), and just 3% fit Wordnet only.

Table 2: Overlap between Wikipedia and Wordnet in Search Results

	# documents annotated with			
	Wikipedia & Wordnet	Wikipedia only	Wordnet only	none
Senseval set	607 (40%)	270 (18%)	89 (6%)	534 (36%)
Bands set	572 (23%)	786 (31%)	27 (1%)	1115 (45%)
Total	1179 (29%)	1056 (26%)	116 (3%)	1649 (41%)

Therefore, Wikipedia seems to extend the coverage of Wordnet rather than providing complementary sense information. If we wanted to extend the coverage of Wikipedia, the best strategy seems to be to consider lists of companies, products and services, rather than complementing Wikipedia with additional sense inventories.

4 Diversity in Google Search Results

Once we know that Wikipedia senses are a representative subset of actual Web senses (covering more than half of the documents retrieved by the search engine), we can test how well search results respect diversity in terms of this subset of senses.

Table 3 displays the number of different senses found at different depths in the search results rank, and the average proportion of total senses that they represent. These results suggest that diversity is not a major priority for ranking results: the top ten results only cover, in average, 3 Wikipedia senses (while the average number of senses listed in Wikipedia is 22). When considering the first 100 documents, this number grows up to 6.85 senses per noun.

Another relevant figure is the frequency of the most frequent sense for each word: in average, 63% of the pages in search results belong to the most frequent sense of the query word. This is roughly comparable with most frequent sense figures in standard annotated corpora such as Semcor (Miller et al., 1993) and the Senseval/Semeval data sets, which suggests that diversity may not play a major role in the current Google ranking algorithm.

Of course this result must be taken with care, because variability between words is high and unpredictable, and we are using only 40 nouns for our experiment. But what we have is a positive indication that Wikipedia could be used to improve diversity or cluster search results: potentially the first top ten results could cover 6.15 different senses in average (see Section 6.5), which would be a substantial growth.

5 Sense Frequency Estimators for Wikipedia

Wikipedia disambiguation pages contain no systematic information about the relative importance of senses for a given word. Such information, however, is crucial in a lexicon, because sense distributions tend to be skewed, and knowing them can help disambiguation algorithms.

We have attempted to use two estimators of expected sense distribution:

- Internal relevance of a word sense, measured as incoming links for the URL of a given sense in Wikipedia.
- External relevance of a word sense, measured as the number of visits for the URL of a given sense (as reported in http://stats.grok.se).

The number of internal incoming links is expected to be relatively stable for Wikipedia articles. As for the number of visits, we performed a comparison of the number of visits received by the bands noun subset in May, June and July 2009, finding a stable-enough scenario with one notorious exception: the number of visits to the noun Tesla raised dramatically in July, because July 10 was the anniversary of the birth of Nicola Tesla, and a special Google logo directed users to the Wikipedia page for the scientist.

We have measured correlation between the relative frequencies derived from these two indicators and the actual relative frequencies in our testbed. Therefore, for each noun w and for each sense w_i, we consider three values: (i) proportion of documents retrieved for w which are manually assigned to each sense w_i; (ii) inlinks(w_i): relative amount of incoming links to each sense w_i; and (iii) visits(w_i): relative number of visits to the URL for each sense w_i.

We have measured the correlation between these three values using a linear regression correlation coefficient, which gives a correlation value of .54 for the number of visits and of .71 for the number of incoming links. Both estimators seem

Table 3: Diversity in Search Results according to Wikipedia

	average # senses in search results			average coverage of Wikipedia senses		
	Bands set	Senseval set	Total	Bands set	Senseval set	Total
First 10 docs	2.88	3.2	3.00	.21	.21	.21
First 25	4.44	4.8	4.58	.28	.33	.30
First 50	5.56	5.47	5.53	.33	.36	.34
First 75	6.56	6.33	6.48	.37	.43	.39
First 100	6.96	6.67	6.85	.38	.45	.41

to be positively correlated with real relative frequencies in our testbed, with a strong preference for the number of links.

We have experimented with weighted combinations of both indicators, using weights of the form $(k, 1-k), k \in \{0, 0.1, 0.2 \ldots 1\}$, reaching a maximal correlation of .73 for the following weights:

$$\text{freq}(w_i) = 0.9 * \text{inlinks}(w_i) + 0.1 * \text{visits}(w_i) \quad (1)$$

This weighted estimator provides a slight advantage over the use of incoming links only (.73 vs .71). Overall, we have an estimator which has a strong correlation with the distribution of senses in our testbed. In the next section we will test its utility for disambiguation purposes.

6 Association of Wikipedia Senses to Web Pages

We want to test whether the information provided by Wikipedia can be used to classify search results accurately. Note that we do not want to consider approaches that involve a manual creation of training material, because they can't be used in practice.

Given a Web page p returned by the search engine for the query w, and the set of senses $w_1 \ldots w_n$ listed in Wikipedia, the task is to assign the best candidate sense to p. We consider two different techniques:

- A basic Information Retrieval approach, where the documents and the Wikipedia pages are represented using a Vector Space Model (VSM) and compared with a standard cosine measure. This is a basic approach which, if successful, can be used efficiently to classify search results.

- An approach based on a state-of-the-art supervised WSD system, extracting training examples automatically from Wikipedia content.

We also compute two baselines:

- A random assignment of senses (precision is computed as the inverse of the number of senses, for every test case).

- A most frequent sense heuristic which uses our estimation of sense frequencies and assigns the same sense (the most frequent) to all documents.

Both are naive baselines, but it must be noted that the most frequent sense heuristic is usually hard to beat for unsupervised WSD algorithms in most standard data sets.

We now describe each of the two main approaches in detail.

6.1 VSM Approach

For each word sense, we represent its Wikipedia page in a (unigram) vector space model, assigning standard tf*idf weights to the words in the document. idf weights are computed in two different ways:

1. Experiment **VSM** computes inverse document frequencies in the collection of retrieved documents (for the word being considered).

2. Experiment **VSM-GT** uses the statistics provided by the Google Terabyte collection (Brants and Franz, 2006), i.e. it replaces the collection of documents with statistics from a representative snapshot of the Web.

3. Experiment **VSM-mixed** combines statistics from the collection and from the Google Terabyte collection, following (Chen et al., 2009).

The document p is represented in the same vector space as the Wikipedia senses, and it is compared with each of the candidate senses w_i via the cosine similarity metric (we have experimented

with other similarity metrics such as χ^2, but differences are irrelevant). The sense with the highest similarity to p is assigned to the document. In case of ties (which are rare), we pick the first sense in the Wikipedia disambiguation page (which in practice is like a random decision, because senses in disambiguation pages do not seem to be ordered according to any clear criteria).

We have also tested a variant of this approach which uses the estimation of sense frequencies presented above: once the similarities are computed, we consider those cases where two or more senses have a similar score (in particular, all senses with a score greater or equal than 80% of the highest score). In that cases, instead of using the small similarity differences to select a sense, we pick up the one which has the largest frequency according to our estimator. We have applied this strategy to the best performing system, VSM-GT, resulting in experiment **VSM-GT+freq**.

6.2 WSD Approach

We have used TiMBL (Daelemans et al., 2001), a state-of-the-art supervised WSD system which uses Memory-Based Learning. The key, in this case, is how to extract learning examples from the Wikipedia automatically. For each word sense, we basically have three sources of examples: (i) occurrences of the word in the Wikipedia page for the word sense; (ii) occurrences of the word in Wikipedia pages pointing to the page for the word sense; (iii) occurrences of the word in external pages linked in the Wikipedia page for the word sense.

After an initial manual inspection, we decided to discard external pages for being too noisy, and we focused on the first two options. We tried three alternatives:

- **TiMBL-core** uses only the examples found in the page for the sense being trained.
- **TiMBL-inlinks** uses the examples found in Wikipedia pages pointing to the sense being trained.
- **TiMBL-all** uses both sources of examples.

In order to classify a page p with respect to the senses for a word w, we first disambiguate all occurrences of w in the page p. Then we choose the sense which appears most frequently in the page according to TiMBL results. In case of ties we pick up the first sense listed in the Wikipedia disambiguation page.

We have also experimented with a variant of the approach that uses our estimation of sense frequencies, similarly to what we did with the VSM approach. In this case, (i) when there is a tie between two or more senses (which is much more likely than in the VSM approach), we pick up the sense with the highest frequency according to our estimator; and (ii) when no sense reaches 30% of the cases in the page to be disambiguated, we also resort to the most frequent sense heuristic (among the candidates for the page). This experiment is called TiMBL-core+freq (we discarded "inlinks" and "all" versions because they were clearly worse than "core").

6.3 Classification Results

Table 4 shows classification results. The accuracy of systems is reported as precision, i.e. the number of pages correctly classified divided by the total number of predictions. This is approximately the same as recall (correctly classified pages divided by total number of pages) for our systems, because the algorithms provide an answer for every page containing text (actual coverage is 94% because some pages only contain text as part of an image file such as photographs and logotypes).

Table 4: Classification Results

Experiment	Precision
random	.19
most frequent sense (estimation)	.46
TiMBL-core	.60
TiMBL-inlinks	.50
TiMBL-all	.58
TiMBL-core+freq	**.67**
VSM	.67
VSM-GT	.68
VSM-mixed	.67
VSM-GT+freq	**.69**

All systems are significantly better than the random and most frequent sense baselines (using $p < 0.05$ for a standard t-test). Overall, both approaches (using TiMBL WSD machinery and using VSM) lead to similar results (.67 vs. .69), which would make VSM preferable because it is a simpler and more efficient approach. Taking a

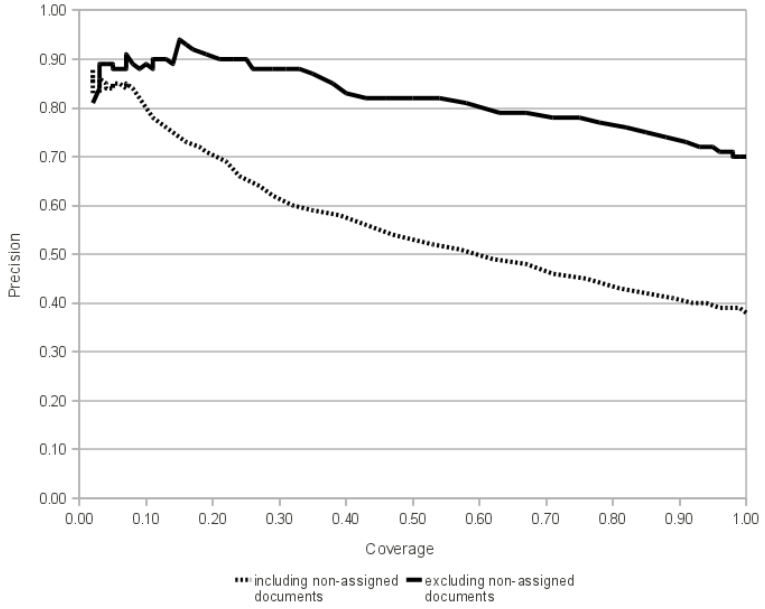

Figure 1: Precision/Coverage curves for VSM-GT+freq classification algorithm

closer look at the results with TiMBL, there are a couple of interesting facts:

- There is a substantial difference between using only examples taken from the Wikipedia Web page for the sense being trained (TiMBL-core, .60) and using examples from the Wikipedia pages pointing to that page (TiMBL-inlinks, .50). Examples taken from related pages (even if the relationship is close as in this case) seem to be too noisy for the task. This result is compatible with findings in (Santamaría et al., 2003) using the Open Directory Project to extract examples automatically.

- Our estimation of sense frequencies turns out to be very helpful for cases where our TiMBL-based algorithm cannot provide an answer: precision rises from .60 (TiMBL-core) to .67 (TiMBL-core+freq). The difference is statistically significant ($p < 0.05$) according to the t-test.

As for the experiments with VSM, the variations tested do not provide substantial improvements to the baseline (which is .67). Using idf frequencies obtained from the Google Terabyte corpus (instead of frequencies obtained from the set of retrieved documents) provides only a small improvement (VSM-GT, .68), and adding the estimation of sense frequencies gives another small improvement (.69). Comparing the baseline VSM with the optimal setting (VSM-GT+freq), the difference is small (.67 vs .69) but relatively robust ($p = 0.066$ according to the t-test).

Remarkably, the use of frequency estimations is very helpful for the WSD approach but not for the SVM one, and they both end up with similar performance figures; this might indicate that using frequency estimations is only helpful up to certain precision ceiling.

6.4 Precision/Coverage Trade-off

All the above experiments are done at maximal coverage, i.e., all systems assign a sense for every document in the test collection (at least for every document with textual content). But it is possible to enhance search results diversity without annotating every document (in fact, not every document can be assigned to a Wikipedia sense, as we have discussed in Section 3). Thus, it is useful to investigate which is the precision/coverage trade-off in our dataset. We have experimented with the best performing system (VSM-GT+freq), introducing a similarity threshold: assignment of a document to a sense is only done if the similarity of the document to the Wikipedia page for the sense exceeds the similarity threshold.

We have computed precision and coverage for every threshold in the range $[0.00 - 0.90]$ (beyond 0.90 coverage was null) and represented the results in Figure 1 (solid line). The graph shows that we

can classify around 20% of the documents with a precision above .90, and around 60% of the documents with a precision of .80.

Note that we are reporting disambiguation results using a conventional WSD test set, i.e., one in which every test case (every document) has been manually assigned to some Wikipedia sense. But in our Web Search scenario, 44% of the documents were not assigned to any Wikipedia sense: in practice, our classification algorithm would have to cope with all this noise as well. Figure 1 (dotted line) shows how the precision/coverage curve is affected when the algorithm attempts to disambiguate all documents retrieved by Google, whether they can in fact be assigned to a Wikipedia sense or not. At a coverage of 20%, precision drops approximately from .90 to .70, and at a coverage of 60% it drops from .80 to .50. We now address the question of whether this performance is good enough to improve search results diversity in practice.

6.5 Using Classification to Promote Diversity

We now want to estimate how the reported classification accuracy may perform in practice to enhance diversity in search results. In order to provide an initial answer to this question, we have re-ranked the documents for the 40 nouns in our testbed, using our best classifier (VSM-GT+freq) and making a list of the top-ten documents with the primary criterion of maximising the number of senses represented in the set, and the secondary criterion of maximising the similarity scores of the documents to their assigned senses. The algorithm proceeds as follows: we fill each position in the rank (starting at rank 1), with the document which has the highest similarity to some of the senses which are not yet represented in the rank; once all senses are represented, we start choosing a second representative for each sense, following the same criterion. The process goes on until the first ten documents are selected.

We have also produced a number of alternative rankings for comparison purposes:

- **clustering (centroids)**: this method applies Hierarchical Agglomerative Clustering – which proved to be the most competitive clustering algorithm in a similar task (Artiles et al., 2009) – to the set of search results, forcing the algorithm to create ten clusters. The centroid of each cluster is then selected as one of the top ten documents in the new rank.

- **clustering (top ranked)**: Applies the same clustering algorithm, but this time the top ranked document (in the original Google rank) of each cluster is selected.

- **random**: Randomly selects ten documents from the set of retrieved results.

- **upper bound**: This is the maximal diversity that can be obtained in our testbed. Note that coverage is not 100%, because some words have more than ten meanings in Wikipedia and we are only considering the top ten documents.

Table 5: Enhancement of Search Results Diversity

rank@10	# senses	coverage
Original rank	2.80	49%
Wikipedia	**4.75**	**77%**
clustering (centroids)	2.50	42%
clustering (top ranked)	2.80	46%
random	2.45	43%
upper bound	6.15	97%

All experiments have been applied on the full set of documents in the testbed, including those which could not be annotated with any Wikipedia sense. Coverage is computed as the ratio of senses that appear in the top ten results compared to the number of senses that appear in all search results.

Results are presented in Table 5. Note that diversity in the top ten documents increases from an average of 2.80 Wikipedia senses represented in the original search engine rank, to 4.75 in the modified rank (being 6.15 the upper bound), with the coverage of senses going from 49% to 77%. With a simple VSM algorithm, the coverage of Wikipedia senses in the top ten results is 70% larger than in the original ranking.

Using Wikipedia to enhance diversity seems to work much better than clustering: both strategies to select a representative from each cluster are unable to improve the diversity of the original ranking. Note, however, that our evaluation has a bias towards using Wikipedia, because only Wikipedia senses are considered to estimate diversity.

Of course our results do not imply that the Wikipedia modified rank is better than the original

Google rank: there are many other factors that influence the final ranking provided by a search engine. What our results indicate is that, with simple and efficient algorithms, Wikipedia can be used as a reference to improve search results diversity for one-word queries.

7 Related Work

Web search results clustering and diversity in search results are topics that receive an increasing attention from the research community. Diversity is used both to represent sub-themes in a broad topic, or to consider alternative interpretations for ambiguous queries (Agrawal et al., 2009), which is our interest here. Standard IR test collections do not usually consider ambiguous queries, and are thus inappropriate to test systems that promote diversity (Sanderson, 2008); it is only recently that appropriate test collections are being built, such as (Paramita et al., 2009) for image search and (Artiles et al., 2009) for person name search. We see our testbed as complementary to these ones, and expect that it can contribute to foster research on search results diversity.

To our knowledge, Wikipedia has not explicitly been used before to promote diversity in search results; but in (Gollapudi and Sharma, 2009), it is used as a gold standard to evaluate diversification algorithms: given a query with a Wikipedia disambiguation page, an algorithm is evaluated as promoting diversity when different documents in the search results are semantically similar to different Wikipedia pages (describing the alternative senses of the query). Although semantic similarity is measured automatically in this work, our results confirm that this evaluation strategy is sound, because Wikipedia senses are indeed representative of search results.

(Clough et al., 2009) analyses query diversity in a Microsoft Live Search, using click entropy and query reformulation as diversity indicators. It was found that at least 9.5% - 16.2% of queries could benefit from diversification, although no correlation was found between the number of senses of a word in Wikipedia and the indicators used to discover diverse queries. This result does not discard, however, that queries where applying diversity is useful cannot benefit from Wikipedia as a sense inventory.

In the context of clustering, (Carmel et al., 2009) successfully employ Wikipedia to enhance automatic cluster labeling, finding that Wikipedia labels agree with manual labels associated by humans to a cluster, much more than with significant terms that are extracted directly from the text. In a similar line, both (Gabrilovich and Markovitch, 2007) and (Syed et al., 2008) provide evidence suggesting that categories of Wikipedia articles can successfully describe common concepts in documents.

In the field of Natural Language Processing, there has been successful attempts to connect Wikipedia entries to Wordnet senses: (Ruiz-Casado et al., 2005) reports an algorithm that provides an accuracy of 84%. (Mihalcea, 2007) uses internal Wikipedia hyperlinks to derive sense-tagged examples. But instead of using Wikipedia directly as sense inventory, Mihalcea then manually maps Wikipedia senses into Wordnet senses (claiming that, at the time of writing the paper, Wikipedia did not consistently report ambiguity in disambiguation pages) and shows that a WSD system based on acquired sense-tagged examples reaches an accuracy well beyond an (informed) most frequent sense heuristic.

8 Conclusions

We have investigated whether generic lexical resources can be used to promote diversity in Web search results for one-word, ambiguous queries. We have compared WordNet and Wikipedia and arrived to a number of conclusions: (i) unsurprisingly, Wikipedia has a much better coverage of senses in search results, and is therefore more appropriate for the task; (ii) the distribution of senses in search results can be estimated using the internal graph structure of the Wikipedia and the relative number of visits received by each sense in Wikipedia, and (iii) associating Web pages to Wikipedia senses with simple and efficient algorithms, we can produce modified rankings that cover 70% more Wikipedia senses than the original search engine rankings.

We expect that the testbed created for this research will complement the - currently short - set of benchmarking test sets to explore search results diversity and query ambiguity. Our testbed is publicly available for research purposes at http://nlp.uned.es.

Our results endorse further investigation on the use of Wikipedia to organize search results. Some limitations of our research, however, must be

noted: (i) the nature of our testbed (with every search result manually annotated in terms of two sense inventories) makes it too small to extract solid conclusions on Web searches (ii) our work does not involve any study of diversity from the point of view of Web users (i.e. when a Web query addresses many different use needs in practice); research in (Clough et al., 2009) suggests that word ambiguity in Wikipedia might not be related with diversity of search needs; (iii) we have tested our classifiers with a simple re-ordering of search results to test how much diversity can be improved, but a search results ranking depends on many other factors, some of them more crucial than diversity; it remains to be tested how can we use document/Wikipedia associations to improve search results clustering (for instance, providing seeds for the clustering process) and to provide search suggestions.

Acknowledgments

This work has been partially funded by the Spanish Government (project INES/Text-Mess) and the Xunta de Galicia.

References

R. Agrawal, S. Gollapudi, A. Halverson, and S. Leong. 2009. Diversifying Search Results. In *Proc. of WSDM'09*. ACM.

P. Anick. 2003. Using Terminological Feedback for Web Search Refinement : a Log-based Study. In *Proc. ACM SIGIR 2003*, pages 88–95. ACM New York, NY, USA.

J. Artiles, J. Gonzalo, and S. Sekine. 2009. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. 2009.*

T. Brants and A. Franz. 2006. Web 1T 5-gram, version 1. Philadelphia: Linguistic Data Consortium.

D. Carmel, H. Roitman, and N. Zwerdling. 2009. Enhancing Cluster Labeling using Wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 139–146. ACM.

C. Carpineto, S. Osinski, G. Romano, and Dawid Weiss. 2009. A Survey of Web Clustering Engines. *ACM Computing Surveys*, 41(3).

Y. Chen, S. Yat Mei Lee, and C. Huang. 2009. PolyUHK: A Robust Information Extraction System for Web Personal Names. In *Proc. WWW'09 (WePS-2 Workshop)*. ACM.

C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proc. SIGIR'08*, pages 659–666. ACM.

P. Clough, M. Sanderson, M. Abouammoh, S. Navarro, and M. Paramita. 2009. Multiple Approaches to Analysing Query Diversity. In *Proc. of SIGIR 2009*. ACM.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2001. TiMBL: Tilburg Memory Based Learner, version 4.0, Reference Guide. Technical report, University of Antwerp.

E. Gabrilovich and S. Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India*.

S. Gollapudi and A. Sharma. 2009. An Axiomatic Approach for Result Diversification. In *Proc. WWW 2009*, pages 381–390. ACM New York, NY, USA.

R. Mihalcea. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of NAACL HLT*, volume 2007.

G. Miller, C. R. Beckwith, D. Fellbaum, Gross, and K. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicograph*, 3(4).

G.A Miller, C. Leacock, R. Tengi, and Bunker R. T. 1993. A Semantic Concordance. In *Proceedings of the ARPA WorkShop on Human Language Technology. San Francisco, Morgan Kaufman.*

M. Paramita, M. Sanderson, and P. Clough. 2009. Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto task 2009. *CLEF working notes*, 2009.

M. Ruiz-Casado, E. Alfonseca, and P. Castells. 2005. Automatic Assignment of Wikipedia Encyclopaedic Entries to Wordnet Synsets. *Advances in Web Intelligence*, 3528:380–386.

M. Sanderson. 2000. Retrieving with Good Sense. *Information Retrieval*, 2(1):49–69.

M. Sanderson. 2008. Ambiguous Queries: Test Collections Need More Sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 499–506. ACM New York, NY, USA.

C. Santamaría, J. Gonzalo, and F. Verdejo. 2003. Automatic Association of Web Directories to Word Senses. *Computational Linguistics*, 29(3):485–502.

Z. S. Syed, T. Finin, and Joshi. A. 2008. Wikipedia as an Ontology for Describing Documents. In *Proc. ICWSM'08*.