International Journal of Geographical Information Science
Vol. 22, No. 3, March 2008, 301–313

Taylor & Francis
Taylor & Francis Group

Research Article

A conceptual density-based approach for the disambiguation of toponyms

DAVIDE BUSCALDI* and PAULO ROSSO
Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain

(*In final form 13 August 2007*)

Nowadays, a huge quantity of information is stored in digital format. A great portion of this information is constituted by textual and unstructured documents, where geographical references are usually given by means of place names. A common problem with textual information retrieval is represented by polysemous words, that is, words can have more than one sense. This problem is present also in the geographical domain: place names may refer to different locations in the world. In this paper we investigate the use of our word sense disambiguation technique in the geographical domain, with the aim of resolving ambiguous place names. Our technique is based on WordNet conceptual density. Due to the lack of a reference corpus tagged with WordNet senses, we carried out the experiments over a set of 1,210 place names extracted from the SemCor corpus that we named GeoSemCor and made publicly available. We compared our method with the most-frequent baseline and the enhanced-Lesk method, which previously has not been tested in large contexts. The results show that a better precision can be achieved by using a small context (phrase level), whereas a greater coverage can be obtained by using large contexts (document level). The proposed method should be tested with other corpora, due to the fact that our experiments evidenced the excessive bias towards the most-frequent sense of the GeoSemCor.

Keywords: Word sense disambiguation; Toponym resolution; Conceptual density; Spatial indexing

1. Introduction

A great portion of the information currently available in digital format is constituted by textual and unstructured documents. The continuous growth of this kind of information and the increasing number of users that can access it constitute a challenge to the developers of information retrieval (IR) systems. One of the most challenging problems is the *ambiguity* of human language. When searching for specific keywords, it is desirable to eliminate occurrences in documents where the word or words are used in an inappropriate sense (Ide and Véronis 1998). Ambiguity can be of various types: proper names may identify different classes of named entities (for instance, 'London' may identify the writer 'Jack London' or a city in the UK), or may be used as a name for different instances of the same class; e.g. 'London' is also a city in Canada. The task of assigning the most appropriate sense to a word within its context is named *word sense disambiguation* (WSD). Notably,

*Corresponding author. Email: dbuscaldi@dsic.upv.es

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810701626251

this is still an open problem in the field of natural language processing (NLP). Many approaches have been developed and evaluated (at *Senseval*[1] and *Semeval*[2] competitions), but no single dominant method has emerged.

Usually, WSD approaches are categorized into *corpus-based* and *knowledge-based*. The former use annotated data to train a model, that is used later in order to carry out the disambiguation process. The latter are based on the use of external resources such as ontologies, thesauri, or dictionaries. On the one hand, corpus-based methods give better results, but they are limited by the lack of annotated corpora; on the other hand, knowledge-based methods do not need training data, but often there are limitations on the cases in which they can be used, resulting in lower coverage and precision (Snyder and Palmer 2004). Previous work demonstrates that WSD is useful for IR only in the case of improving precision (Sanderson 1996, Gonzalo *et al.* 1998, Rosso *et al.* 2004), or if it is used in a restricted domain (Paliouras et al. 1998, Steffen *et al.* 2004).

Our previous experiences at GeoCLEF[3] (Buscaldi *et al.* 2006b,c) drew our attention to the problem of the ambiguity of place names (*toponyms*). In this paper we study the application of a knowledge-based WSD method in the geographical domain, specifically to the disambiguation of toponyms. The method we propose is based on the one (Rosso *et al.* 2003) we developed for the disambiguation of nouns, which implemented a variation of the *Conceptual Density* formula by Agirre and Rigau (1996). We used WordNet (Miller 1995) as an external knowledge resource.

Toponym disambiguation is a relatively new field. From an NLP perspective, it is merely the application of WSD to place names. Its most direct application should be the improvement of the searches both in the Web and in large news collections, due to the fact that it is very common to find geographical information in web pages or news stories (e.g. '*Elections in Italy*', '*Plane crash in Teheran*'). A growing interest in the field of geographical information retrieval (GIR) is testified by the recent creation of the GeoCLEF exercise and the increment of the attendance at the GIR workshops[4] held at the last SIGIR events. The lack of a reference corpus has long been an obstacle to the evaluation of algorithms for toponym resolution (Leidner 2004). Recently, some corpora have been compiled (Garbin and Mani 2005, Leidner 2006), but the lack of a mapping between WordNet and the locations IDs used in these corpora prevented us from evaluating our method with these resources. We overcome this problem by selecting the geographical entities in the SemCor[5] corpus that was originally developed for the WSD task.

In the following section, we will give an overview of the previous efforts in the field of toponym disambiguation. In Section 3, we will provide a brief description of the WordNet ontology. In Section 4, we will describe our WSD method. In Section 5, we will resume the experiments carried out and the systems we compared our method to, together with a description of the corpus we built. Finally, we will give a discussion of the obtained results.

[1] http://www.cs.unt.edu/~rada/senseval/
[2] http://nlp.cs.swarthmore.edu/semeval/
[3] http://ir.shef.ac.uk/geoclef/
[4] http://www.geo.unizh.ch/~rsp/gir06/
[5] http://www.cs.unt.edu/~rada/downloads.html#semcor

2. Previous work on toponym resolution

Toponym resolution can be defined as the task of assigning an ambiguous place name with reference to the actual location that it represents in a given context. For instance, the word '*Cambridge*' is ambiguous. It could be used to represent one of the following locations (according to WordNet):

(i) Cambridge—(a city in eastern England on the River Cam; site of Cambridge University);
(ii) Cambridge—(a city in Massachusetts just north of Boston; site of Harvard University and the Massachusetts Institute of Technology).

As in the generic WSD task, the clues that can be used to disambiguate the word are found in the context; for instance, the presence of '*Boston*' in the context may be a hint that the correct sense of Cambridge is the second one.

Existing methods for the disambiguation of toponyms may be subdivided into three categories:

(i) *map-based*: methods that use an explicit representation of places on a map;
(ii) *knowledge-based*: they exploit external knowledge sources such as gazetteers, Wikipedia or ontologies;
(iii) *data-driven* or *supervised*: based on standard machine learning techniques.

Among the first ones, Smith and Crane (2001) proposed a method for toponym resolution based on the geographical coordinates of places: the locations in the context are arranged in a map, weighted by the number of times they appear. Then, a centroid of this map is calculated and compared with the actual locations related to the ambiguous toponym. The location closest to the 'context map' centroid is selected as the right one. They reported precisions of between 74% and 93% (depending on test configuration), where precision is calculated as the number of correctly disambiguated toponyms divided by the number of toponyms in the test collection. The GIPSY subsystem by Woodruff and Plaunt (1994) is also based on spatial coordinates, although in this case they are used to build polygons. Woodruff and Plaunt (1994) reported issues with noise and runtime problems.

The methods of Olligschlaeger and Hauptmann (1999) and Rauch *et al.* (2003) are based on evidences collected from a variety of sources, especially gazetteers. The information collected in order to disambiguate the place names may vary from population data (references to populous places are more frequent than those to the less populated ones) to the presence of postal addresses. Olligschlaeger and Hauptmann (1999) reported a precision of 75% for their rule-based method. Overell *et al.* (2006) presented a method based on Wikipedia[6], which takes advantage of some of its features, such as the article templates, categories and referents (links to other articles in Wikipedia).

A Naïve Bayes classifier is used by Smith and Mann (2003) to classify place names with respect to the US states or foreign countries. They reported precisions between 21.8% and 87.4%, depending on the test collection used. Garbin and Mani (2005) used a rule-based classifier, obtaining precisions between 65.3% and 88.4%, also depending on the test corpus. The weakness of supervised methods highlights the need for a large quantity of training data in order to obtain a high precision.

[6] http://en.wikipedia.org

Moreover, the inability to classify unseen toponyms is also a major problem that affects this class of methods.

3. The WordNet ontology

WordNet is a complex lexical database of English, developed at the University of Princeton under the direction of G. Miller (Miller 1995). Its last version (3.0) contains 155,327 words grouped into 117,597 *synsets*. A synset (*set* of *syn*onyms) is a group of words that are considered semantically equivalent. An example of synset for a geographical location is the following: (London, Greater London, British capital, capital of the United Kingdom). Each synset is associated to a unique id and a *gloss*, i.e. the definition of the concept (in the case of London: the capital and largest city of England; located on the Thames in southeastern England; financial, industrial and cultural center). Moreover, the most important feature of WordNet is that it also provides a set of semantic relationships which connect different synsets. In figure 1, we show a portion of WordNet surrounding the London synset.

In the example some important semantic relationships are visible; e.g. the *hypernymy* (or *is-a*) relationship. This relationship connects two concepts where one is more general than the other, such as '*clock*' and '*cuckoo clock*'. The inverse relationship (from a more specific concept to a more general one) is called *hyponymy* (i.e. '*cuckoo clock*' is a hyponym of '*clock*'). The *meronymy*, or *part-of*, relationship connects concepts that are a part of the other and vice versa (in the latter case it is named *holonymy*). In the example of figure 1, '*England*' is holonym of '*London*'. Finally, the *instance* relationship connects abstract concepts to real world instances, such as '*clock*' and '*Big Ben*'. Most relationships connect words of the same lexical category, also known as part-of-speech (POS) category, such as those named here, which connect only noun concepts.

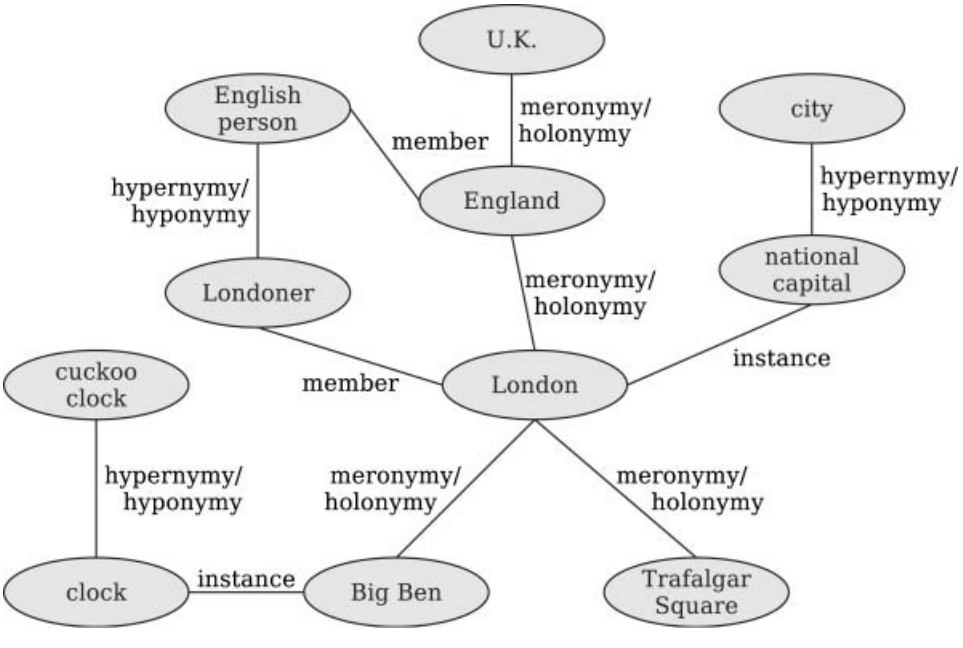

Figure 1. A graph representation of a portion of WordNet surrounding the *London* synset.

WordNet has been widely used in NLP, mainly because of its role as sense inventory. It was also employed to semantically annotate the Brown corpus (Kucera and Franics 1967), obtaining the SemCor (*Sem*antic *Cor*respondance) corpus (Landes *et al.* 1998). In SemCor every word belonging to the noun, verb, adjective and adverb POS categories has been labeled with a WordNet sense. It is often used as a training corpus for supervised word sense disambiguation methods.

4. Conceptual density-based word sense disambiguation

Conceptual density (CD) was introduced by Agirre and Rigau (1996) as a measure of the correlation between the sense of a given word and its context. It is computed on WordNet subhierarchies, determined by the hypernymy relationship. The disambiguation algorithm by means of CD consists of the following steps:

(i) select the next ambiguous word w, with $|w|$ senses;
(ii) select the context $\bar{c}_w$, i.e. a sequence of words, for w;
(iii) build $|w|$ subhierarchies, one for each sense of w;
(iv) for each sense s of w, calculate CD_s;
(v) assign to w the sense which maximizes CD_s.

Our formulation of the Conceptual Density of a WordNet subhierarchy s is (Rosso *et al.* 2003):

$$CD(m, f, n) = m^{\alpha}\left(\frac{m}{n}\right)^{\log f}, \tag{1}$$

where m are the *relevant* synsets in the subhierarchy, n is the total number of synsets in the subhierarchy, and f is the rank of frequency of the word sense related to the subhierarchy (e.g. 1 for the most frequent sense, 2 for the secondone, etc.). The inclusion of the frequency rank means that less frequent senses are selected only when $m/n \geqslant 1$. The relevant synsets are both the synsets of the word to disambiguate and those of the context words. Our formulation allows solving some problems with the original CD due to the higher granularity of newer WordNet versions.

The WSD system based on this formula obtained 81.5% in precision over the nouns in the SemCor (baseline: 75.5%, calculated by assigning to each noun its most frequent sense), and participated at the Senseval-3 competition as the CIAOSENSO system (Bscaldi *et al.* 2004), obtaining 75.3% in precision over nouns in the all-words task (baseline: 70.1%). These results were obtained with a context window of only two nouns, the one preceding and the one following the word to disambiguate.

When we considered adapting this algorithm to the disambiguation of toponyms, we realized that the hypernymy relationship was not suitable. For instance, *Cambridge*(*1*) and *Cambridge*(*2*) are both instances of the '*city*' concept and therefore, they share the same hypernym. The result is that the subhierachies are composed only by the synsets of the two senses of '*Cambridge*', and they are left undisambiguated because their density is the same (which in both cases is 1).

Our idea is to consider the *holonymy* relationship instead of hypernymy. With this relationship it is possible to create subhierarchies that allow discerning different locations (having the same name) in a more effective way. For instance, the last three holonyms for '*Cambridge*' are:

(i) Cambridge→England→UK
(ii) Cambridge→Massachusetts→New England→USA

The best choice for context words is represented by other place names, because holonymy is always defined through them and because they constitute the actual 'geographical' context of the toponym we are disambiguating. In figure 2 we show an example of a holonym tree obtained for the disambiguation of '*Georgia*' with the contexts of '*Atlanta*', '*Savannah*' and '*Texas*', from the following fragment of text extracted from the br-a01 file of SemCor: "Hartsfield has been mayor of **Atlanta**, with exception of one brief interlude since 1937. His political career goes back to his election to city council in 1923. The mayor's present term of office expires on Jan. 1. He will be succeeded by Ivan Allen Jr., who became a candidate in the Sept. 13 primary after Mayor Hartsfield announced that he would not run for reelection. **Georgia** Republicans are getting strong encouragement to enter a candidate in the 1962 governor's race, a top official said on Wednesday. Robert Snodgrass, state GOP chairman, said a meeting held Tuesday night in Blue Ridge brought enthusiastic responses from the audience. State Party Chairman James W. Dorsey added that enthusiasm was picking up for a state rally to be held on Sept. 8 in **Savannah** at which newly elected **Texas** Sen. John Tower will be the featured speaker."

According to WordNet *Georgia* may refer to 'a state in southeastern United States' or a 'republic in Asia Minor on the Black Sea separated from Russia by the Caucasus mountains'.

As one would expect, the holonyms of the context words populate exclusively the subhierarchy related to the first sense (the area filled with a diagonal hatching in figure 2); this is reflected in the CD formula, which returns a CD value 4.29 for the first sense ($m=8$, $n=11$, $f=1$) and 0.33 for the second one ($m=1$, $n=5$, $f=2$). In this work, we considered as relevant also those synsets which belong to the paths of the context words that fall into a subhierarchy of the toponym to disambiguate.

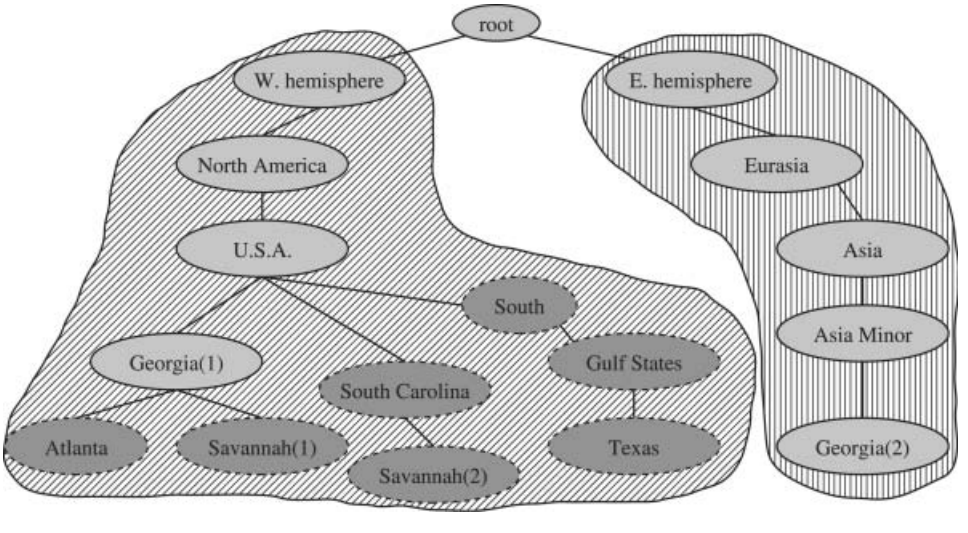

Figure 2. Example of holonym hierarchy for the disambiguation of *Georgia*, with context: {*Atlanta*, *Savannah*, *Texas*} from the file br-a01 of SemCor. Nodes are synsets, dark grey nodes are synsets of context words.

5. Experiments

The holonym-based CD disambiguator described in the previous section was tested over a collection of 1,210 toponyms. Its results were compared with the *most frequent* (MF) baseline, obtained by assigning to each toponym its most frequent sense, and with another WordNet-based method which uses its glosses, and those of its context words, to disambiguate it. We were not able to compare our method with any map-based method, principally because WordNet does not provide coordinates of the geographical entities. Some efforts for the integration of WordNet with geographical gazetteers have been undertaken (Buscaldi *et al.* 2006a), but a ready-to-use mapping still does not exist. Neither did we carry out a comparison with a corpus-based method because of the small amount of data contained in the collection.

5.1 *The GeoSemCor corpus*

For the evaluation of our algorithm we decided to use the SemCor corpus, limited to its geographical names, since the other available resources are not labeled with WordNet senses. We identified the place names with the help of WordNet itself: if a synset (corresponding to the combination of the word—the *lemma* tag—with its sense label—*wnsn*) had the synset *location* among its hypernyms, then we labeled the respective word with a *geo* tag (for instance, <wf geo=true cmd=done pos=NN lemma=dallas wnsn=1 lexsn=1:15:00::>Dallas</wf>). The resulting *GeoSemCor* collection contains 1,210 toponyms and may be downloaded from the following link: http://www.dsic.upv.es/~dbuscaldi/resources/geosemcor2.0.tar.gz. Sense labels are those of WordNet 2.0 (SemCor 2.0). The format is based on the SGML of SemCor. Details of GeoSemCor are shown in table 1. Please note that the polysemy count is based on the number of senses in WordNet and not on the number of places that a name can represent. For instance, ''London'' in WordNet has two senses, but only the first of them corresponds to the city, because the second one is the surname of the US writer ''Jack London''. However, in GeoSemCor only the instances related to toponyms have been labeled with the *geo* tag.

In order to give the reader an impression of a processed sentence, we show a section of text from the br-m02 file of GeoSemCor:

<s snum=74>
<wf cmd=done pos=RB lemma=here wnsn=1 lexsn=4:02:00::>Here</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=people wnsn=1 lexsn=1:14:00::>peoples</wf>
<wf cmd=done pos=VB lemma=speak wnsn=3 lexsn=2:32:02::>spoke</wf>
<wf cmd=ignore pos=DT>the</wf>

Table 1. GeoSemCor statistics.

total toponyms	1,210
polysemous toponyms	709
avg. polysemy	2.151
labeled with MF sense	1,140*
labeled with the second sense	53
labeled with a sense >2	17

*: corresponding to the 94.2%.

<wf cmd=done pos=NN lemma=tongue wnsn=2 lexsn=1:10:00::>tongue</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf geo=true cmd=done pos=NN lemma=iceland wnsn=1 lexsn=1:15:00::>Iceland</wf>
<wf cmd=ignore pos=IN>because</wf>
<wf cmd=ignore pos=IN>that</wf>
<wf cmd=done pos=NN lemma=island wnsn=1 lexsn=1:17:00::>island</wf>
<wf cmd=done pos=VBD ot=notag>had</wf>
<wf cmd=done pos=VB ot=idiom>gotten_the_jump_on</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=Hawaiian wnsn=1 lexsn=1:1:00::>Hawaiian</wf>
<wf cmd=done pos=NN lemma=American wnsn=1 lexsn=1:18:00::>American</wf>
[…]
</s>

The *cmd* attribute indicates whether the tagged word is a stop-word or not. The *wnsn* and *lexsn* attributes indicate the senses of the tagged word. The attribute *lemma* indicates the base form of the tagged word. Finally, *geo=true* tells us that the word represents a geographical location. The 's' tag indicates the sentence boundaries.

5.2 *The enhanced Lesk algorithm*

Banerjee and Pedersen (2002) presented a WordNet-enhanced version of the well-known dictionary-based algorithm proposed by Lesk (1986). The original Lesk algorithm was based on the comparison of the gloss of the word to disambiguate with the context words and their glosses. This enhancement takes into account the glosses of concepts related to the word to be disambiguated by means of various WordNet relationships. Then the similarity between a sense of the word and the context is calculated by means of *overlaps*. The word is assigned the sense which obtains the best overlap match with the glosses of the context words and their related synsets. In WordNet there can be seven relationships for each word, this means that for every pair of words up to 49 relationships have to be considered. The similarity measure based on Lesk has been demonstrated as one of the best measures for the semantic relatedness of two concepts by Patwardhan *et al.* (2003).

5.3 *Measures*

There are four measures that are commonly used for the evaluation of WSD methods. *Precision*, or *Accuracy*, is calculated as the number of correctly disambiguated words divided by the number of disambiguated words. *Recall* is the number of correctly disambiguated words divided by the total number of words in the collection. *Coverage* is the number of disambiguated words, either correctly or wrongly, divided by the total number of words. Finally, the *F-measure* is a combination of precision and recall, calculated as their harmonic mean:

$$\frac{2 * precision * recall}{precision + recall} \tag{2}$$

Precision, recall and coverage are usually given as percentages. The *F*-measure is generally represented by a value between 0 and 1.

6. Results

The experiments were carried out considering three kinds of contexts:

(i) *sentence* context: the context words are all the toponyms within the same sentence;
(ii) *paragraph* context: all toponyms in the same paragraph of the word to be disambiguated;
(iii) *document* context: all toponyms contained in the document are used as context.

Most WSD methods use a context window of a fixed size (e.g. two words, four words, etc.). In our case we realized that in some documents there are many geographical terms which could be considered as context words; on the other hand, it is difficult to find more than two or three geographical terms in a sentence, and setting a larger context size would be useless. Therefore, we did not use a fixed context size. The average sizes obtained by taking into account the above context types are displayed in table 2.

It can be observed that there is a small difference between the uses of sentence and paragraph, whereas the context size when using the entire document is three times more than the one obtained by taking into account of the paragraph. Tables 3–5

Table 2. Average context size depending on context type.

context type	avg. context size
sentence	2.09
paragraph	2.92
document	9.73

Table 3. Results obtained using sentence as context.

system	precision	recall	coverage	F-measure
CD-1	94.7%	56.7%	59.9%	0.709
CD-0	92.2%	78.9%	85.6%	0.850
Enh. Lesk	96.2%	53.2%	55.3%	0.685

Table 4. Results obtained using paragraph as context.

system	precision	recall	coverage	F-measure
CD-1	94.0%	63.9%	68.0%	0.761
CD-0	91.7%	76.4%	83.4%	0.833
Enh. Lesk	95.9%	53.9%	56.2%	0.689

Table 5. Results obtained using document as context.

system	precision	recall	coverage	F-measure
CD-1	92.2%	74.2%	80.4%	0.822
CD-0	89.9%	77.5%	86.2%	0.832
Enh. Lesk	99.2%	45.6%	45.9%	0.625

summarize the results obtained by our systems and the enhanced Lesk algorithm for each context type. In the tables, CD-1 indicates the CD disambiguator, CD-0 a variant we introduced to improve coverage by assigning a density 0 to all the sub-hierarchies composed of a single synset (in equation (1) these sub-hierarchies would obtain 1 as weight); *Enh.Lesk* refers to the method by Banerjee and Pedersen (2002).

The obtained results show that CD-based methods are very precise when the smallest context is used, but there are many cases in which the context is empty and, therefore, it is impossible to calculate the CD. On the other hand, as one would expect, when the largest context is used with the increase of coverage and recall, precision drops below the most frequent baseline. However, we observed that 100% coverage cannot be achieved by CD due to some issues with the structure of WordNet. In fact, there are some 'critical' situations where CD cannot be computed, even when a context is present. This occurs when the same place name can refer to a place and another one it contains: for instance, '*New York*' is used to refer both to the city and the state it is contained in (i.e. its holonym). The result is that two senses fall within the same subhierarchy, thus not allowing assignment of a unique sense to '*New York*'.

Nevertheless, even with this problem, the CD-based methods obtain a greater coverage than the enhanced Lesk method. We suppose that this may be due to the fact that the glosses of the hypernyms or hyponyms, rarely used because of the context, are composed exclusively of geographical names (for instance, with respect to the gloss of *city*, the direct hypernym of 'Cambridge' is 'a large and densely populated urban area; moreover, the gloss of *city* may include several independent administrative districts; ''Ancient Troy was a great city'''—this means that an overlap will be found only if 'Troy' is in the context). It was quite surprising that the best precision (and the worst coverage) for the enhanced Lesk was obtained with the largest context; Banerjee and Pedersen (2002) did not test their algorithm with large contexts, on the basis of the observations of Choueka and Lusignan (1985), who found that human beings make disambiguation decisions based on very short windows of context, usually no more than two words on the left and two on the right. We analyzed the results and we observed that the greater the context, the higher the probability to obtain the same overlap for different senses, with the consequence that the coverage drops. By knowing the number of monosemous locations in GeoSemCor (501) we are able to calculate the minimum coverage that a system can obtain (41.4%), close to the value obtained with the enhanced Lesk and document context (45.9%). This explains also the correlation of high precision with low coverage, due to the monosemous names.

- In table 6 we show a comparison of the best results for each method and each measure with the most frequent baseline. It can be observed that although knowledge-based methods may achieve better precision, the *F*-measure obtained by the MF heuristic is very high. It is a well-known fact that human annotations, taken as a gold standard, are biased in favor of the first WordNet sense, which corresponds to the most frequent (Fernández-Amoró *et al.* 2001). Moreover, WordNet is not as rich as some specialized geographical resources such as the Getty thesaurus of geographical names (TGN[7]) or the GNS and GNIS gazetteers[8]. For instance, whereas WordNet returns only three results

[7] http://www.getty.edu/research/conducting_research/vocabularies/tgn/
[8] http://earth-info.nga.mil/gns/html/index.html and http://geonames.usgs.gov

Table 6. Comparison of the best results * obtained by the knowledge-based systems with the MF baseline.

system	precision	recall	coverage	F-measure
CD-1	94.7% (s)	74.2% (d)	80.4% (d)	0.822 (d)
CD-0	92.2% (s)	78.9% (s)	86.2% (d)	0.850 (s)
Enh. Lesk	99.2% (d)	53.9% (p)	56.2% (p)	0.689 (p)
Most Frequent	94.2%	94.2%	100.0%	0.942

*: (s) indicates that the result has been obtained by using sentences as context, (p) paragraphs and (d) documents.

for '*Springfield*', a search in the GNIS returns 129 places named 'Springfield'. This poverty of geographical information greatly simplifies the task (some places appear in WordNet with a single sense, although they are actually polysemous, e.g. '*Genoa*'). Therefore, it is not clear how the performance of our method might be affected by scaling up to these resources.

7. Conclusions and future work

Toponym disambiguation is an open problem in GIR. At this moment, an evaluation framework for the testing and the comparison of various methods does not exist. We tested a conceptual density-based method over the GeoSemCor, a resource we extracted from the SemCor corpus and that we are making available. Our method was compared with the most frequent baseline and the enhanced Lesk method. The obtained results expose the limits of both WordNet as a resource for the disambiguation of toponyms and of GeoSemCor as a resource for the testing of disambiguation methods. WordNet is particularly poor in terms of geographical information with respect to gazetteers, whereas GeoSemCor is biased towards the most frequent senses. Another issue raised by our work concerns 'critical' arrangements of the synsets in the WordNet holonym/meronym hierarchy that limit the maximum coverage attainable by CD methods. Finally, the results support the fact that small contexts give higher precisions than larger contexts, also in the case of geographical terms and not only for generic noun disambiguation as we observed in our previous works. The comparison with the enhanced Lesk shows that our CD-based method has a wider coverage and therefore obtains better results in terms of *F*-measure. We plan to test our method with other resources such as gazetteers or geographical thesauri, using for the evaluation of the corpus compiled by Leidner (2006). We would also like to carry out a comparison with a map-based method in order to provide a comprehensive study of the currently available toponym disambiguation methods.

Acknowledgements

This work has been partially supported by the MCyT TIN2006-15265-C06-04 Research Project.

References

AGIRRE, E. and RIGAU, G., 1996, Word sense disambiguation using conceptual density. In *Proceedings of the 16th conference on computational linguistics* (*COLING '96*), pp. 16–22 (Copenhaghen: Association for Computational Linguistics).

Banerjee, S. and Pedersen, T., 2002, An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Computational Linguistics and Intelligent Text Processing, Third International Conference*, A. Gelbukh (Ed.). 2276 of Lecture Notes in Computer Science, pp. 136–145 (Berlin: Springer).

Buscaldi, D., Rosso, P. and Masulli, F., 2004, The upv-unige-CIAOSENSO WSD System. In *Proceedings of the Senseval-3 workshop, ACL 2004*, pp. 77–82 (Barcelona: Association of Computational Linguistics).

Buscaldi, D., Rosso, P. and Peris, P., 2006a, Inferring geographical ontologies from multiple resources for geographical information retrieval. In *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2006*, pp. 52–55 (Seattle: Association for Computational Linguistics).

Buscaldi, D., Rosso, P. and Sanchis, E., 2006b, Using the WordNet ontology in the GeoCLEF geographical information retrieval Task. In *Accessing Multilingual Information Repositories*, C. Peters, F.C. Gey, J. Gonzalo, H. Mller, G.J. Jones, M. Kluck, B. Magnini, M. de Rijke and D. Giampiccolo (Eds). 4022 of Lecture Notes in Computer Science, pp. 939–946 (Berlin: Springer).

Buscaldi, D., Rosso, P. and Sanchis, E., 2006c, WordNet-based index terms expansion for geographical information retrieval. In *Proceedings of the GeoCLEF 2006 Workshop*, (Alicante, Spain).

Choueka, Y. and Lusignan, S., 1985, Disambiguation by short contexts. *Computers and the Humanities*, **19**, pp. 147–157.

Fernández-Amorós, D., Gonzalo, J. and Verdejo, F., 2001, The role of conceptual relation in word sense disambiguation. In *Proceedings of the 6th International Workshop NLDB'01*, pp. 87–98 (Madrid: GI).

Garbin, E. and Mani, I., 2005, Disambiguating toponyms in news. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (*HLT05*), Vancouver, British Columbia, Canada, pp. 363–370 (Morristown, NJ: Association for Computational Linguistics).

Gonzalo, J., Verdejo, F., Chugur, I. and Cigarrán, J., 1998, Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL '98 workshop on the Usage of WordNet for NLP*, pp. 38–44 (Montreal: Association for Computational Linguistics).

Ide, N. and Véronis, J., 1998, Word sense disambiguation: the state of the art. *Computational Linguistics*, **24**, pp. 1–40.

Kucera, H. and Francis, W.N., 1967, *Computational Analysis of Present-Day American English* (Brown University Press).

Landes, S., Leacock, C. and Tengi, R., 1998, Building semantic concordances. In *WordNet: An Electronic Lexical Database*, C. Fellbaum (Ed.), pp. 199–216 (Cambridge, MA: MIT Press).

Leidner, J.L., 2004, Towards a reference corpus for automatic toponym resolution evaluation. In *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2004* (Sheffield, UK).

Leidner, J.L., 2006, An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, **30**, pp. 400–417.

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 5th annual international conference on Systems documentation*, (SIGDOC '86), pp. 24–26 (Toronto: Association for Computing Machinery).

Miller, G.A., 1995, WordNet: A lexical database for English. *Communications of the ACM*, **38**, pp. 39–41.

Olligschlaeger, A. and Hauptmann, A., 1999, Multimodal information systems and GIS: The informedia digital video library. In *Proceedings of the 1999 ESRI User Conference* (San Diego, CA).

OVERELL, S., MAGALHAES, J. and RÜGER, S., 2006, Place disambiguation with co-occurrence models. In *Proceedings of the GeoCLEF 2006 Workshop*, C. Peters (Ed.) (Alicante, Spain).

PALIOURAS, M., KARKALETSIS, V. and SPYROPOULOS, C., 1998, Machine learning for domain-adaptive word sense disambiguation. In *Proceedings of the Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, LREC 1998*, (Granada, Spain).

PATWARDHAN, S., BANERJEE, S. and PEDERSEN, T., 2003, Using measures of semantic relatedness for word sense disambiguation. In *Computational Linguistics and Intelligent Text Processing, 4th International Conference*, A. Gelbukh (Ed.), 2588 of *Lecture Notes in Computer Science*, pp. 241–257 (Berlin: Springer).

RAUCH, E., BUKATIN, M. and BAKER, K., 2003, A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pp. 50–54 (Edmonton: Association for Computational Linguistics).

ROSSO, P., FERRETTI, E. and VIDAL, V., 2004, Text Categorization and Information Retrieval Using WordNet Senses. In *Proceedings of the 2nd Global WordNet Conference* (*GWC 2004*), pp. 299–304 (Masaryk University, Brno, Czech Republic).

ROSSO, P., MASULLI, F., BUSCALDI, D., PLA, F. and MOLINA, A., 2003, Automatic noun sense disambiguation. In *Computational Linguistics and Intelligent Text Processing, 4th International Conference*, A. Gelbukh (Ed.) 2588 of *Lecture Notes in Computer Science*, pp. 273–276 (Berlin: Springer).

SANDERSON, M., 1996, Word sense disambiguation and information retrieval. PhD Thesis, University of Glasgow, Glasgow, Scotland, UK.

SMITH, D.A. and CRANE, G., 2001, Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, 2163 of *Lecture Notes in Computer Science*, pp. 127–137 (Berlin: Springer).

SMITH, D.A. and MANN, G.S., 2003, Bootstrapping toponym classifiers. In *Proceedings of the HLTNAACL 2003 workshop on Analysis of geographic references*, pp. 45–49 (Morristown, NJ: Association for Computational Linguistics).

SNYDER, B. and PALMER, M., 2004, The English all-words task. In *Proceedings of the Senseval-3 workshop, ACL 2004*, pp. 41–43 (Barcelona: Association for Computational Linguistics).

STEFFEN, D., SACALEANU, B. and BUITELAAR, P., 2004, Domain specific sense disambiguation with unsupervised methods. *LDV Forum*, **19**, pp. 93–101.

WOODRUFF, A. and PLAUNT, C., 1994, GIPSY: Automated geographic indexing of text documents. *Journal of the American Society of Information Science*, **45**, pp. 645–655.