Semantic Relatedness for Named Entity Disambiguation Using a Small Wikipedia*

Izaskun Fernandez[1], Iñaki Alegria[2], and Nerea Ezeiza[2]

[1] Tekniker-IK4
ifernandez@tekniker.es
http://www.tekniker.es/
[2] IXA Group
{i.alegria,n.ezeiza}@ehu.es
http://ixa.si.ehu.es/Ixa

Abstract. Resolving Named Entity Disambiguation task with a small knowledge base makes the task more challenging. Concretely, we present an evaluation of the state-of-the-art methods in this task for Basque NE disambiguation based on the Basque Wikipedia. We have used MFS, VSM, ESA and UKB for linking any ambiguous surface NE form occurrence in a text with its corresponding Wikipedia entry in the Basque Wikipedia version. We have analysed their performance with different corpora and as it was expected, most of them perform worse than when using big Wikipedias such as the English version, but we think these results are more realistic for less-resourced languages. We propose a new normalization factor for ESA to minimise the effect of the knowledge base size.

Keywords: Named Entity Disambiguation, Semantic Relatedness, Wikipedia.

1 Introduction

Named Entity Disambiguation (NED) is the task of exploring which real person, place, event... is referred to by a certain surface form. For example, NED must decide whether, in a given context, the surface form *Amstrong* refers to either the cyclist *Lance Armstrong* or the astronaut *Neil Armstrong*.

In many respects, named entities disambiguation task is similar to word sense disambiguation (WSD). WSD task, which has supported large-scale evaluations (SENSEVAL editions[1]), aims to assign dictionary meanings to all the instances of a predetermined set of polysemous words in a corpus. However, these evaluations do not include proper name disambiguation and omit named entity meanings from the targeted semantic labels and the development and test contexts.

Recently, a couple of evaluation campaigns with a specific task dealing with NED have been defined within the TAC-KBP track (Knowledge Base Population), organized in 2009 [9] and 2010 [7] respectively. The entity linking task consists in determining,

* This research was partially supported by MICINN (TIN2010-20218) and the Basque Government (EMAITEK2011:Medicina Digital).

[1] http://www.senseval.org/

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 276–283, 2011.

for a given NE surface form and a particular context, which of the entries in a knowledge base (KB) the NE refers to, provided there is one. In the TAC-KBP track, the English Wikipedia was the KB.

But those results have not been tested yet for smaller Wikipedia versions, such as those written in less-resourced languages. In this work we analyse how characteristics of a smaller Wikipedia (the Basque Wikipedia) change compared to a larger one (the English Wikipedia) and how they can affect the performance of named entity disambiguation systems. The distinctive features of a smaller Wikipedia are the following:

- *Less ambiguity*: since there are fewer articles, the list of possible candidates for an ambiguous entity in the KB will be shorter.
- *More NE without an entry in the KB*: (related to the previous feature) it is possible that more named entities will not be represented in the Wikipedia. This characteristic depends on the KB, but also on the target corpus.
- *Shorter articles*: the percentage of short articles is higher, so the amount of information for disambiguation is smaller.
- *Less links*: (related to the previous feature) we can presume that less links will be available to extract new information.

Although it seems that the first feature leads to an easier task, the rest of them make the task more challenging when working with a small KB.

Our aim is to test some of the most popular NED methods described in the literature for the task of linking named entities in texts to Wikipedia articles, using a small Wikipedia (the Basque Wikipedia, `eu.wikipedia.org`) for this purpose. Concretely we test bag-of-words in a vector space model, the ESA algorithm and a graph-based model (UKB) in this context.

The remainder of the paper is organized as follows. Section 2 describes the state of the art in NED task. Section 3 describes the disambiguation strategy we have applied, specifying the resources and the methods used for it. Section 4 describes the evaluation methodology and in the Section 5 we present the results. Finally, we show some conclusions and future work.

2 Related Work

In the last few years growing interest has been shown in the task of linking Wikipedia and NED. In [3] the authors use several characteristics from an English Wikipedia dump, such as the text of the entries, categories, redirection pages and hyperlinks, which are used to train a supervised model (based on SVM) to link an entity occurrence to a Wikipedia page. They evaluate the system for person name disambiguation, reporting accuracies of 55.4% to 84.8%.

In [4], Cucerzan formalized the disambiguation paradigm, which is based on Wikipedia and includes similar information as the one described in the previous paper in a vector-space model. More specifically, the vectorial representation of a document is compared with the vectorial representation of the Wikipedia entities, which are represented as an extended vector with two main components, corresponding to context and category information. The reported accuracy is 88% to 91%.

Gabrilovich and Markovitch [5] presented Wikipedia-based ESA (Explicit Semantic Analysis) to compute semantic relatedness of documents. ESA works by first building an inverted index from words to all Wikipedia articles that contain them. Then, it estimates a relatedness score for any two documents by using the inverted index to build a vector over Wikipedia articles for each document and by computing the cosine similarity between the two vectors.

Graph-based models have been also used to face this problem [6]. In this work, the authors propose a method, which uses a graph model using multiple features extracted from Wikipedia, to estimate Semantic Relatedness over the Wikipedia-based graph. They exploit the obtained relatedness values to resolve the NED problem, obtaining 91.46% and 89.83% accuracy in two different evaluation sets.

A wide range of methods and combinations have been developed for the entity linking task in TAC2009 and TAC2010 ([9], [7]). This task is similar to the one we explain in this paper: given a name and its context, the system must decide whether this name corresponds to an entry in a database from Wikipedia and, if so, which one.

3 Experimental Settings

The NED process aims to create a mapping between the surface form of an entity and its unique meaning in the KB if it exists. So a dictionary indicating all possible entries in the KB for a surface form is needed. In this work we use a Basque Wikipedia dump dated March 2006 as KB so as to build the mapping dictionary and extract features for the different disambiguation methods. The algorithms, given a surface form of a named entity and its occurrence context (a paragraph), must decide whether the name corresponds to an entry in the Wikipedia and, if so, to which one. For the evaluation, we have used a news corpus. All these resources are detailed in the following sections.

3.1 Resources

Named Entity Ambiguity Dictionary. To build the mapping dictionary based on the Basque Wikipedia, we try to derive all the possible ambiguous forms for each Wikipedia entry. Once all the surface forms are generated, we represent them in a dictionary where each surface form is defined by a set of Basque Wikipedia entries. We apply the following strategy to generate the dictionary:

- *The title itself* is considered a possible surface form for the entry.
- When a title has more than one word, we *generate all the possible combinations replacing each word (except the last one) with its initial*, and we add them as surface forms. For instance, to deal with the entry *Juan Jose Ibarretxe*, we add the surface forms *Juan J. Ibarretxe*, *J. Jose Ibarretxe* and *J. J. Ibarretxe*.
- *Each word in the title* is considered a surface form. For instance, three new entries are created in the previous example: *Juan*, *Jose* and *Ibarretxe*.

While exploiting the Wikipedia characteristics we also enrich the set of surface forms. Since if an entry in Wikipedia refers to another entry, it should be linked to its main entry, those anchors can also be considered surface forms for the target entry. As redirect entries represent just another form to mention the linked entry, they are also added to the set of surface forms for the target entry.

Finally, we also use the content of disambiguation pages. Unlike in redirect pages, disambiguation titles are considered surface forms for the entries listed in them, because, as the name implies, they are entries that refer to different Wikipedia entries, so they are ambiguous names.

News Corpus. Since there is no standard Basque corpus defined for the NED evaluation task, we have generated a repository for that purpose using pieces of news of the 2002 year edition of the *Euskaldunon Egunkaria* newspaper. To be precise, we have used an annotated version of this news corpus, processed in the context of HERMES project (http://nlp.uned.es/hermes/). The corpus has 40,648 articles and 135,505 NEs.

It is very common in news texts to use the entire or long form of an entity (i.e. *Aimar Olaizola*) in its first occurrence and shorter forms (i.e. *Aimar* or *Olaizola*) in the later occurrences within the same item of news. These short surface forms have higher ambiguity, since they are not as specific as the longer ones, so they can refer to a larger set of NEs.

We do not take advantage of this feature during the disambiguation task, because the aim of our work is to evaluate different methods for short contexts which might not have more than one instance of the same entity. However, it is very useful to generate a corpus that does not need any hand revision for the evaluation. We have created a test-corpus (Corpus A) which only includes texts that meet the following conditions: they have instances of short entities and their longer unambiguous forms in the same item of news; and the unambiguous forms have their corresponding Wikipedia entry.

This way, for every ambiguous NE, we know which Basque Wikipedia entry it must be linked to as a result of the disambiguation. Looking at the previous example, if *Aimar Olaizola* appears in Wikipedia and it is unambiguous at the piece of news we are analysing, the paragraphs where *Aimar* or *Olaizola* surface forms occur will be selected for this test corpus.

Another test-corpus was built in a regular way, collecting news paragraphs with at least one entity, no matter if there was a longer NE containing it along the news or if it was represented in the Wikipedia. Since there was no automatic way to know which the corresponding Wikipedia entry was, if any, for each NE in this example set, the corpus (henceforth Corpus B) was manually disambiguated, linking each NE occurrence to its corresponding Wikipedia entry, when possible. Even if in this work no training process has been applied, Corpus B was divided into two groups in order to use one for the tuning process (Corpus B-dev) and the second one for evaluation (Corpus B-eval).

Table 1. Evaluation corpora

	# Examples	# Ambiguous Ex.	NIL	Ambiguity
Corpus A	**6,500**	**4,376**	**0**	**67.32%**
Corpus B-dev	**532**	**295**	**70**	**55.45%**
Corpus B-eval	**500**	**300**	**63**	**60%**

Table 1 summarizes the main features of the mentioned corpora, including the number of NE to disambiguate; the number of the ambiguous examples in the set, considering non-ambiguous those surface forms that are only defined by one Wikipedia entry in the NE ambiguity dictionary; the number of examples that have no corresponding NE disambiguation form represented in Wikipedia, in which case the system should answer NIL; and finally the ambiguity rate.

3.2 Methods

The methods we want to test with a small Wikipedia KB are some of the most popular in the literature: bag-of-words in a vector space model (VSM), ESA and UKB, a graph-based model. As it is usual in WSD, the baseline is calculated using the most frequent entity among the candidates (MFS). To compute the most frequent entity, we use the number of in-links of each Wikipedia entry. The entry with the highest number of in-links is considered the most frequent one.

VSM. Vector Space Model [2] for the resolution of NE ambiguity represents all the Wikipedia entries using bag-of-words vectors, being each word position measured with its *tfidf* value in the corresponding Wikipedia entry. Based on this information, when a new NE occurrence has to be disambiguated, its context is represented in the same vector space with bag-of-words modelling. This vector is compared to the ones corresponding to the Wikipedia disambiguation pages for the given ambiguous NE form defined at the dictionary, computing the cosine of each vector pair. The Wikipedia entry with the highest cosine value will be the one proposed as the disambiguation form.

ESA. Explicit Semantic Analysis (ESA) [5] is a vector space comparison algorithm based on Wikipedia articles. For a candidate text, each dimension in its ESA vector corresponds to a Wikipedia article, with the score being the similarity of the text with the article text, subject to *tfidf* weighting. The relatedness of two texts is computed as the cosine similarity of their ESA vectors.

Since in NED task the aim is to compare an input text with a set of Wikipedia entries, it is not necessary to construct the entire ESA vector. Estimating the similarity measures for that set of Wikipedia entries is enough, as it is formalized by Sorg and Cimiano [10]. So what we need is just to compute the similarity values for a given ESA vector dimension.

For computing the association strength between an A_j Wikipedia article and a T input text, ESA applies the following metric:

$$ESASimilarity(T, A_j) = \sum_{w_i \epsilon T} v_i * k_j$$

where k_j is the tfidf value of w_i in A_j, and v_i the tfidf value of w_i in T

Balanced ESA. The association strength defined for ESA tends to promote short articles over long ones when articles in the KB are very different in terms of length. This effect disappear when articles are similar in extent or at least of a minimum length, as it happens when this algorithm is tested in a bigger KB such as the English Wikipedia. In order to reduce this adverse effect, we have introduced a normalization factor to the original ESA strength association measure, which takes into account the number of words shared between the Wikipedia article and the input text. So the new association strength estimation is defined as follows:

$$bESAsimilarity(T, A_j) = ESASimilarity(T, A_j) * \frac{Count_T}{|T|}$$

where $Count_T$ is the number of words shared by a T input text and an A_j article

UKB. UKB [1] is a Personalized PageRank method that aims to obtain a relatedness score between a pair of texts by performing random walks over a graph to compute a stationary distribution for each text. In order to apply it in the context of NED based on the Basque Wikipedia, it is necessary to build two resources: the Basque Wikipedia as a graph, and a dictionary.

In order to construct the graph structure of the Basque Wikipedia 2006 dump, we simply treat the articles as vertices, and the links between articles as edges as in [12]. The graph has 63,106 vertices and 458,026 edges.

The aim of this task is to disambiguate surface forms linking them with a particular Wikipedia article, provided there is one. Therefore, the mapping dictionary should have available correspondences between NE ambiguous forms, surface forms, and the set of Basque Wikipedia articles that could be their disambiguation forms. This resource has been described in 3.1.

4 Evaluation and Results

The goal of the NED tool is to give, for a certain NE, the Wikipedia entry with the highest score among the candidates. When no candidate is found in the dictionary, no answer is possible and the system must return NIL. When there is a scoreless tie, the tool can have one of the following behaviours:

1. Silent mode. The system does not take any decision on a tie.
2. Tie-break mode. The system makes a decision by applying a random answer or MFS.

In silent mode, we evaluate the performance of the algorithms in terms of *F-Score*. As there is neither special treatment nor a defined thershold for NIL prediction, the system only returns NIL when we force an answer (in tie-break mode) and there is no candidate. Otherwise, it decides at random or using MFS. Thus, in tie-break mode, we treat NIL as a possible answer and, we compute *accuracy* accordingly.

Table 2 shows the results obtained by each algorithm. The first column describes the results of the baseline system (MFS), which is used as a reference in tie-break mode. In silent mode, we want to point out that, unexpectedly, ESA obtains the worst results, because of the low recall. We think that the small size of the Wikipedia articles makes it difficult to obtain strong similarity measures when we compare them with the context paragraphs of the target NE instance. However, we intend to take a deeper view on the results to confirm this suspition. Using bESA we obtain better results than with ESA or VSM. Nevertheless, UKB obtains the best results, outperforming significantly the rest

Table 2. Results of the algorithms

	MFS	**VSM**	**ESA**	**bESA**	**UKB**
A – Silent	68.32%	70.15%	66%	71.25%	81.91%
A – Random	68.32%	74.03%	71%	75.94%	82.8%
A – MFS	68.32%	75.53%	72.43%	77.66%	82.8%
B-Dev - Silent	72%	69%	59.4%	66.6%	75.7%
B-Dev - Random	72%	70.3%	60.9%	67.8%	75.9%
B-Dev - MFS	72%	70.4%	61%	68%	75.9%
B-Eval - Silent	70.4%	67%	58%	65%	76%
B-Eval - Random	70.4%	69%	61.2%	68.4%	76.2%
B-Eval - MFS	70.4%	70%	61.6%	68.4%	76.2%

of the methods and it is the only method that achieves better results than MFS in Corpus B.

Corpus B-Dev and Corpus B-Eval have similar size (500 NEs) and the results are comparable between them, but there is a drop of 5-7 points with respect to Corpus A in the case of ESA, bESA and UKB. We think this is due to the low recall and the bad assignment of the NIL choice, but it requires deeper analysis.

In tie-break mode, as it was expected, the results are better when MFS is applied. For Corpus B we observe that the results are 3 points better in average, while for Corpus A the difference is higher, because the NEs in Corpus A have always an answer in Wikipedia. The improvement is not so important for UKB. The main reason is that UKB has significantly higher recall, so there are less scoreless ties to break.

5 Conclusions

We have presented the work we have done in the field of Named Entity Disambiguation (NED) based on the Basque Wikipedia. Being the Basque Wikipedia a small KB, we have tested most of the state-of-the-art algorithms in order to evaluate their performance using small resources instead of big KBs.

Despite being ESA the most popular algorithm for semantic relatedness estimation, we have seen that for a small Wikipedia, with short and few entries, it does not perform so well. To minimise the negative effect caused by the KB size, we have proposed a new normalization factor for ESA, which provides better performance than the original, even getting one of the best performances in terms of accuracy and stability.

The UKB algorithm has not achieved very good results for NED task using the English Wikipedia as KB for graph construction (A. Soroa, pers. comm.). But surprisingly, for Basque, not only does it perform well, but it has also turned out to be the best of the tested algorithms for every evaluation corpora. We are examining the results on Corpus B-Dev in order to clarify the reasons of this.

We think that improving the NIL assignment will be a key work for the future. Finally, we consider interesting as future work to combine the different algorithms to get better results, especially in terms of recall.

References

1. Agirre, E., Soroa, A.: Personalizing PageRank for Word Sense Disambiguation. In: Proceedings of the 12th Conference of the European chapter of the Association for Computational Linguistics, pp. 33–41 (2009)
2. Baeza, R., Ribeiro, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston (1999)
3. Bunescu, R.C., Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation. In: 1th Conference of the European Chapter of the Association for Computational Linguistics, pp. 9–16 (2006)
4. Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. Empirical Methods in Natural Language Processing (2007)
5. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 6–12 (2007)
6. Gentile, A.L., Zhang, Z., Xia, L., Iria, J.: Cultural Knowledge for Named Entity Disambiguation: A Graph-Based Semantic Relatedness Approach. Serdica Journal of Computing 4(2), 217–242 (2010)
7. Ji, H., Grishman, R., Dang, H.T., Griffitt, K., Ellis, J.: Overview of the TAC 2010 Knowledge Base Population Track. In: Proceedings of Text Analysis Conference (2010)
8. Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for wordsense identification. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)
9. McNamee, P., Dang, H.T.: Overview of the TAC 2009 Knowledge Base Population track. In: Proceedings of the Second Text Analysis Conference (2009)
10. Sorg, P., Cimiano, P., Enriching, P.: the crosslingual link structure of Wikipedia-A classification-based approach. In: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (2008)
11. Strube, M., Ponzeto, S.P.: WikiRelate! Computing semantic relatedness using Wikipedia. In: Proceedings of the AAAI 2006, pp. 1419–1424 (2006)
12. Yeh, E., Ramage, D., Manning, C., Agirre, E., Soroa, A.: WikiWalk: Random walks on Wikipedia for Semantic Relatedness. In: Proceedings of ACL Workshop TextGraphs-4: Graph-based Methods for NLP (2009)