A language-independent LESK based approach to Word Sense Disambiguation

Tommaso Petrolito

Filologia Letteratura e Linguistica, University of Pisa, Italy

tommasouni@gmail.com

Abstract

This paper describes a language-independent LESK based approach to Word Sense Disambiguation (WSD), involving also Vector Space Models applied to the Distributional Semantics Hypotesis. In particular this approach tries to solve some issues that come up with less-resourced languages. The approach also addresses the inadequacy of the Most Frequent Sense (MFS) heuristics to fit specific domain corpora.

1 Introduction

This language independent approach to WSD, even if in a very early stage of development, tries to solve two main problems.

1. Variable quality of glosses and examples (the solution would be to use glosses and examples for the aligned synsets in several languages, we will explain how).

2. Weakness of Most Frequent Sense heuristics for domain corpora (or even general corpora that, for some reasons, are not so similar to the corpus on which the frequencies were calculated), but also lack of synset annotated corpora for several non-English languages (the solution would involve Space Vector Models, we will explain how).

We use Wordnet (WN) resources (Miller et al., 1990; Miller, 1995; Fellbaum, 1998) (synsets glosses and examples) from a specific standpoint: preferring to avoid the usage of monolingual resources, even though the specific task does not involve cross-lingual WSD on aligned parallel corpora.

It has to be pointed out that the approach is being considered 'unsupervised': it does not rely on semantic annotation, although lemmatization and PoS-tagging are taken into account.

In most cases the quality of lexical resources is very variable, even though some languages have good resources, as of course English and, for instance, Italian with both MultiWordnet (Pianta et al., 2002) and ItalWordnet (Roventini et al., 2000).

An example is given with the *dog/câine* (first synset) glosses and examples[1] in Table 1. It is evident that the English synset has a richer gloss.

Assuming a WSD approach involving overlap counts, the English words *Canis*, *wolf*, *breeds* will be counted in; as for the Romanian words *Animal*, *mamifer*, *carnivor* (all IS-A relations), their English lemmas would be reached in any case in an Expanded Gloss implementation.

Anyway, *pază* and *vânătoare* ('guarding' and 'hunting') would be useful for the same task.

In the counterexample given in Table 2, the Romanian gloss is evidently richer than the English one, in particular using a WSD overlapping algorithm that is able to count on *asistenţă*, *socială*, *întreţinerea*, *bătrânilor* ('assistance', 'social', 'maintenance', 'elders') and so on.

In general, it can be noticed how variable the quality is for different corpora and for different synsets.

Anyway, usually English WN provides the best and richest set of examples for a given synset.

This variability in quality is observable also concerning the coverage of different WNs[2] (Bond and Foster, 2013).

[1] For a quick series of examples of this kind, just have a look on multilingual aligned synsets on the MultiWordnet Interface (Ranieri et al., 2004).

See http://multiwordnet.fbk.eu/online/multiwordnet.php

[2] See http://compling.hss.ntu.edu.sg/omw/ and http://globalwordnet.org/wordnets-in-the-world/ for an overview.

Synset	Lang	Gloss
dog,domestic_dog, Canis_familiaris/1	EN	a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds;
câine/1	RO	Animal mamifer carnivor domesticit, folosit pentru pază, vânătoare etc..

Table 1: Synset gloss comparison (EN:dog,domestic_dog,Canis_familiaris/1 – RO:câine/1)

Synset	Lang	Gloss
home,nursing_home, rest_home	EN	an institution where people are cared for;
azil	RO	Instituţie de asistenţă socială pentru întreţinerea bătrânilor, infirmilor, copiilor orfani etc.

Table 2: Synset gloss comparison (EN:home,nursing_home,rest_home – RO:azil)

Many LESK-inspired algorithms have been presented; see for instance Kilgarriff and Rosenzweig (2000a), Kilgarriff and Rosenzweig (2000b), Banerjee and Pedersen (2002) and Basile et al. (2014). Our approach is an adaptation that takes into account the issues about glosses and examples quality.

In particular we try to gain advantage from the usage of better resources available in other languages.

A first baseline attempt was tested here, due to time constraints: relying on English glosses and examples with non-English target corpora. For future work, a more complex adaptation will be attempted, trying to take advantage of glosses and examples in several languages at once.

This kind of approach leads to two main issues.

Either trying to use many glosses and examples from WNs in several languages or trying to use glosses and examples just from the Princeton Wordnet (PWN) working on a non-English corpus, the issue arises of how to compare the contexts of the target words with the glosses and examples of their candidate synsets.

The problem needs to be addressed, for instance, if the contexts of the target words are in a given language (not English) that is not compatible with an overlapping approach involving words from glosses and examples in many different languages or even just with the English ones.

Second, the widely used Most Frequent Sense (MFS) heuristics (Gale et al., 1992; Miller et al., 1994; Kohomban and Lee, 2005), easily implementable in English by choosing the first synset for the given lemma, cannot be used when working with other languages, as the synset ordering does not mirror sense frequency statistics.

Even working on English, MFS' usefulness varies accordingly to the similarity of the target corpus to SemCor (Mihalcea, 1998), concerning the topic(s) and the granularity of meanings.

Also this issue needs a proper solution and some help can come from Vector Space Models (VSMs) applied to the Distributional Hypothesis (DH) (Harris, 1951; Turney et al., 2010) of Semantics implementing Distributional Semantic Spaces (DSS).

2 Methodology

This approach is organized in two disambiguation steps.

The first (focused on quality) is based mainly on a kind of LESK adapted in the language-independent perspective discussed above and involves WN glosses and examples.

The second (focused on quantity) is based on VSMs and follows the assumption, coherent with the Distributional Hypothesis, that the neighbours of the target word in the Semantic Space are semantically related (in paradigmatic relations) with the target word.

Both these two disambiguation steps will be discussed in this section.

2.1 Language independent LESK algorithm

Our idea consists in counting the overlaps in couples of candidate-synset-bag-of-lemmas and context-bag-of-lemmas. Then, the candidate synset for which the count is higher is chosen.

Let us assume that we use an Italian sentence, but we want to rely on English synsets glosses and examples (we will explain later why we would want to do that).

Let us take the Italian sentence:

> *Il cane abbaia spesso quando fa la guardia ai suoi giocattoli o al suo cibo*
> "The dog often barks when guarding its toys or food"

Given *cane* ("dog") as our target lemma and *n*, 'noun' as part of speech, the algorithm has to fulfill the following steps:

1. Find all the Italian synsets associated to the given lemma and part of speech that are aligned to the English WordNet.

2. For each candidate synset, build a 'bag of lemmas' by retrieving all content words found in the English gloss and example(s) and lemmatizing them.

3. For each sentence (in this case the current sentence containing *cane*), build a context bag of lemmas by taking English glosses and examples of the English synsets aligned to the Italian synsets of the words in the sentence (lemma and part of speech annotations are assumed to be there).

   To avoid a computational nightmare (and maybe also to avoid noise), only unambiguous lemmas and lemmas with a number of synsets less than an upper bound, previously defined, will be taken into account as sources of synset-glosses and synset-examples.

The synset for which the overlapping between the two bags is bigger is the chosen one.

With this approach we want to show that, theoretically, one can benefit from the semantic information available in different languages to help solve the ambiguity, even though the task doesn't start off as multilingual.

This means that theoretically we can disambiguate an Italian text using information from a WN in any language.

Now, let us suppose to use at once pairs of English bags (as explained above) and other pairs of bags of lemmas, built in the same way, but taken from WNs of other languages.

So we will have for each synset of *cane* a bag with lemmas from each language (separately).

Similarly, for the words in the sentence there will be a bag of lemmas for each language.

Let us take one 'monolingual' group at the time.

Each bag-of-lemmas pair (one from the candidate synset, one from the sentence words) will have an overlapping score. We can take into account all the scores, for example by summing them then choose the synset that has the higher total score.

Why should all this improve the results?

Let us suppose to include Romanian WN in these group of wordnets and try to disambiguate *cane* in the same Italian sentence seen above:

> *Il cane abbaia spesso quando fa la guardia ai suoi giocattoli o al suo cibo*
> "The dog often barks when guarding its toys or food"

We point out that `dog.n.01` and `cane.n.01` (respectively the English and Italian first synsets for the Italian lemma *cane*) have glosses and examples with no mention to 'hunting' or 'guarding', while the gloss of the Romanian synset (`câine.n.01`) refers to both.

The context word *guardia* ('guard') would be exploited much better by using the Romanian WN than by using the Italian one, even though the language of the text is Italian.

The same thing could happen with English (or any other language) texts about dogs in which 'guarding' and 'hunting' words are not exploited by a monolingual LESK approach.

This case is an evidence of how a multilingual approach, involving comparisons between the bags for the candidate synsets and for the context in several languages, could enhance overlapping counts and lead to a better synset selection.

We have provided an example showing that this approach can be applied also by building many sub-bags in distinct languages (and this was the full original idea): for each synset existing in English, Italian and Romanian (for example) a list containing the three monolingual bags can be built and the synset-scores can take into account the overlapping in all the languages (summing the overlapping scores together), taking advantage from eventual better quality (or even just few lucky occurring keywords) in the glosses and examples in other languages.

2.1.1 Candidate synsets scoring

For the future, a more complex and representative scoring measure will be defined, maybe taking into account the good example provided by Basile et al. (2014) based on different weights for lemmas.

In the current version, due to time constraints, each synset gains a very simple score equal to the number of lemmas shared by the candidate-synset-bag and the context-synsets-bag (that is the union of the single synset-bags occurring in the sentence).

Only one specific customization is added to this naive scoring approach: unambiguous lemmas in the context have double weight (so their overlapping will be counted twice).

2.1.2 Results

Here we show a baseline experiment exploiting only English glosses and examples on an Italian target corpus.

If we set a configuration that takes context lemmas from words linked to a certain number of synsets (up to 6), this algorithm tags correctly the 36.17% of words in the Italian MultiSemCor (Pianta and Bentivogli, 2003; Bentivogli and Pianta, 2005; Bentivogli et al., 2005).

If we use it to remove the wrong synsets it works much better: removing, for each target word, synsets with score lower than max_score/2, 65% of words still have right synsets in the remaining set of synsets.

2.2 Paradigmatic relations algorithm

As for the second issue, concerning the Most Frequent heuristics, VSMs could provide a big help.

In particular, while the first disambiguation step focuses on the specificity of meanings observed in the specific contexts, a help from distributional quantities would focus on the frequencies of co-occurrences, thus providing a frequency based heuristics.

So, while the LESK based approach is context-dependent (so it will select different synsets for different usages of the same lemma in different contexts), the highest frequency heuristics would just help by pushing for the only one synset (always the same) that is the most frequent for the given lemma (independently whether observed in different contexts) in the corpus on which the frequencies have been measured.

A way to reproduce that kind of heuristics, even for languages with lack in synsets-annotated corpora[3](Petrolito and Bond, 2014) (even well resourced languages as Italian cannot provide such resources for corpora other than SemCor), could be implemented as a WSD algorithm involving a Distributional Semantic Space.

An example is provided by (McCarthy et al., 2004).

McCarthy et al. (2004) use a thesaurus, acquired from automatically parsed text, based on the method of Lin (1998), in order to find the predominant sense of a target word.

This thesaurus provides, through distributional similarity scores, the nearest neighbours to each target word. Then they use the WordNet similarity package (Patwardhan and Pedersen, 2003) to obtain semantic similarity measures to weight the contribution that each neighbour gives to the various senses of the target word.

Here we do something similar, but we specifically exploit paradigmatic relations.

1. In the DSS, neighbour words with high cosine similarity share the same contexts and are therefore supposed to be in paradigmatic relation.

2. Also through WordNet we can infer words in paradigmatic relation with the target word, such as hypernyms, hyponyms, cohyponyms, synonyms and antonyms.

Also this method consists of measuring the overlapping between bags of lemmas, as for the algorithm described previously.

The first bag contains the *N* (chosen arbitrarily) neighbours of the target lemma in the Semantic Space.

Both the target lemma and the neighbours are combinations lemma-PoS (the lemma and PoS information is assumed to be there).

The second bag contains lemmas taken through an exploration of the paradigmatic relations in the WN ontology for the candidate synset.

So for the candidate synset the following will be taken: lemmas, antonyms of lemmas, hyper-

[3] See `http://globalwordnet.org/wordnet-annotated-corpora/`

nyms lemmas, hyponyms lemmas, cohyponyms (hyponyms of the hypernyms) lemmas.

Actually, also in this bag (as for the one of the neighbours) instead of simple lemmas, we have combinations of lemma-PoS (in this case the information is obviously provided because we are taking lemmas from WN synsets).

Then the synset with maximum score among the overlapping values between the Semantic Space 'neighbourhood' and the paradigmatic-relations-bag is selected.

2.2.1 Candidate synsets scoring

Also in this case the score is a simple count of the intersection between the two bags.

2.2.2 Results

Also this algorithm on its own is not achieving good performances, only a 34.5% of correct annotations on the Italian SemCor.

3 Data Set

All the experiments have been done on the Italian MultiSemCor (Pianta and Bentivogli, 2003; Bentivogli and Pianta, 2005; Bentivogli et al., 2005) corpus, already sense-tagged.

SemCor is the perfect data set for this task, as it is the first case of corpus annotated with WN synsets and it is available in various languages (English, Italian, Romanian and Japanese). The Italian MultiSemCor contains 14,144 sentences and 261,283 tokens, 119,802 of which are annotated with senses.

The availability in a good number of languages makes MultiSemCor a good resource to try this language-independent approach.

Also the NTU-Multilingual Corpus (NTU-MC) (Tan and Bond, 2011) could be a perfect resource for this kind of experiments.

NTU-MC is a corpus designed to be multilingual from the start. It contains parallel text in eight languages: English, Mandarin Chinese, Japanese, Indonesian, Korean, Arabic, Vietnamese and Thai.

4 Implementation and Evaluation

The two algorithms have been implemented as Python scripts importing the NLTK (Bird, 2006) WN Interface and the Gensim (Řehůřek and Sojka, 2010) `word2vec` (Mikolov et al., 2013) library.

At first, the two algorithms were implemented separately, achieving the results discussed in Subsection 2.1.2 and Subsection 2.2.2, then the two algorithms have been implemented together in sequence.

The first algorithm has been used for a first disambiguation step excluding candidate synsets with scores lower than the 50% of the maximum, then the second algorithm has been applied taking into account only the remaining candidate synsets (provided by the first step of disambiguation) instead of considering all the possible synsets.

When the candidate with higher score in the paradigmatic relations algorithm differs from the one with higher score in the LESK based one, the two scores are normalized in a minimum-maximum 0-1, range and the candidate synset with the highest average is chosen.

The results have improved a lot achieving an encouraging result: 38.67% of the content words have been correctly annotated, with a maximum number of 6 synsets for the context words.

5 Future Work

There is reason to hope that some further attempts based on the approach described in this paper will lead to significant improvements in language-independent WSD.

A first improvement will be exploiting the disambiguated glosses at least for English, as most of the English glosses are disambiguated.

A second improvement will be the extension of the LESK based algorithm with other languages; considered that many glosses are translations of English, we should focus on Merge WNs (Dutch, Polish, etc) in particular.

To do that it will be useful to extend NLTK multilingual support: the `.definition()` and `.examples()` methods of WN synsets would be much more useful for tasks like this by exploiting a `lang` attribute.

A third improvement will be a further development of scores definitions and a complete testing of parameters like: for the first algorithm, the lower bound for the candidate synsets to be saved and passed to the second step of disambiguation and the upper bound for the number of synsets of the context words; for the second algorithm, the number of neighbours or even try to include the approach defined by McCarthy et al. (2004).

References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING*, pages 1591–1600.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Luisa Bentivogli, Emanuele Pianta, and Marcello Ranieri. 2005. Multisemcor: an english italian aligned corpus with a shared inventory of senses. In *Proceedings of the Meaning Workshop*, volume 90.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439.

Zellig S Harris. 1951. Methods in structural linguistics.

Adam Kilgarriff and Joseph Rosenzweig. 2000a. English senseval: Report and results. In *LREC*.

Adam Kilgarriff and Joseph Rosenzweig. 2000b. Framework and results for english senseval. *Computers and the Humanities*, 34(1-2):15–48.

Upali S Kohomban and Wee Sun Lee. 2005. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics.

Rada Mihalcea. 1998. Semcor semantically tagged corpus. *Unpublished manuscript*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244.

G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *In Proceedings of the Human Language Technology Workshop*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Siddharth Patwardhan and Ted Pedersen. 2003. The cpan wordnet:: similarity package.

Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*.

Emanuele Pianta and Luisa Bentivogli. 2003. Translation as annotation. In *Proceedings of the AI* IA 2003 Workshop "Topics and Perspectives of Natural Language Processing in Italy*, pages 40–48. Citeseer.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the first international conference on global WordNet*, volume 152, pages 55–63.

Marcello Ranieri, Emanuele Pianta, and Luisa Bentivogli. 2004. Browsing multilingual information with the multisemcor web interface. In *Proceedings of the LREC 2004 Satellite Workshop on The Amazing Utility of Parallel and Comparable Corpora*, pages 38–41. Citeseer.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. http://is.muni.cz/publication/884893/en.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. Italwordnet: a large semantic database for italian. In *LREC*.

Liling Tan and Francis Bond. 2011. Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). In *PACLIC*, pages 362–371. Citeseer.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.