Using Context Information for Knowledge-Based Word Sense Disambiguation

Kiril Simov(✉), Petya Osenova, and Alexander Popov

Institute of Information and Communication Technology, BAS,
Akad. G. Bonchev. 25A, 1113 Sofia, Bulgaria
{kivs,petya,alex.popov}@bultreebank.org

Abstract. One of the most successful approaches to Word Sense Disambiguation (**WSD**) in the last decade has been the knowledge-based approach, which exploits lexical knowledge sources such as Wordnets, ontologies, etc. The knowledge encoded in them is typically used as a sense inventory and as a relations bank. However, this type of information is rather sparse in terms of senses and the relations among them. In this paper we present a strategy for the enrichment of WSD knowledge bases with data-driven relations from a gold standard corpus (annotated with word senses, syntactic analyses, etc.). We focus on English as use case, but our approach is scalable to other languages. The results show that the addition of new knowledge improves the accuracy of WSD task.

Keywords: Knowledge-based word sense disambiguation · Inference of semantic relations · Context representation

1 Introduction

The recent success of *knowledge-based Word Sense Disambiguation* (**KWSD**) approaches depends on the quality of the *knowledge graph* (**KG**) — whether the knowledge represented in terms of nodes and relations (arcs) between them is sufficient for the algorithm to pick the correct senses of the ambiguous words. Several extensions of the KG constructed on the basis of WordNet have been proposed and already implemented. The solutions to Word Sense Disambiguation (**WSD**) related tasks usually employ lexical databases, such as WordNets and ontologies. However, lexical databases suffer from sparseness with respect to the availability and density of relations. One approach towards remedying this problem is the BabelNet [1], which relates several lexical resources — WordNet[1] [2], DBpedia[2], Wiktionary[3], etc. Although such a setting takes into consideration the role of lexical and world knowledge, it does not incorporate contextual knowledge learned from actual texts (such as collocational patterns, for example). This happens because the knowledge sources for WSD systems usually capture only a

[1] https://wordnet.princeton.edu/.

[2] http://wiki.dbpedia.org/.

[3] https://en.wiktionary.org/wiki/Wiktionary:Main_Page.

C. Dichev and G. Agre (Eds.): AIMSA 2016, LNAI 9883, pp. 130–139, 2016.
DOI: 10.1007/978-3-319-44748-3_13

fraction of the relations between entities in the world. Many important relations are not present in the ontological resources but could be learned from texts.

Here we present approaches towards the enrichment of WSD knowledge bases with context information, represented as relations over semantically annotated corpora. These context relations are taken from gold standard corpora. We focus on English (*SemCor* [3]) as use case, but our approach is scalable to other languages as well. Such an approach is justified by the fact that the lexical databases are sparse with respect to the available knowledge, its density and appropriateness. Also, the predominance of paradigmatic knowledge (synonymy, hypernymy, etc.) is balanced by the addition of syntagmatic relations (valency) — see [4]. From the perspective of knowledge representation lexical databases contain terminological knowledge (T-Box in terms of KL-One) and semantically annotated corpora contain world knowledge (A-Box in terms of KL-One). The current paper demonstrates that adding such context information improves the accuracy of Knowledge-based WSD.

The structure of the paper is as follows: the next section discusses the related work on the topic. Section 3 presents the manually annotated with senses resource — *SemCor*. Section 4 introduces the knowledge-based tool for WSD. Section 5 describes the creation of a new knowledge graph on the basis of gloss logical form encoded in eXtended WordNet (XWN). Section 6 demonstrates two approaches to encode sentences as context relations. Section 7 reports on the performed experiments. Section 8 concludes the paper.

2 Related Work

Knowledge-based systems for WSD have proven to be a good alternative to supervised systems, which require large amounts of manually annotated data. Knowledge-based systems require only a knowledge base and no additional corpus-dependent information. An especially popular knowledge-based disambiguation approach has been the use of popular graph-based algorithms known under the name of “Random Walk on Graph” [5]. Most approaches exploit variants of the PageRank algorithm [6]. Agirre and Soroa [7] apply a variant of the algorithm to WSD by translating WordNet into a graph in which the synsets are represented as nodes and the relations between them are represented as arcs. The resulting graph is called a *knowledge graph* in this paper. Calculating the PageRank vector **Pr** is accomplished through solving the equation:

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)\mathbf{v} \tag{1}$$

where M is an $N \times N$ transition probability matrix (N being the number of nodes in the graph), c is the damping factor and $\mathbf{v}$ is an $N \times 1$ vector. In the traditional, static version of PageRank the values of $\mathbf{v}$ are all equal $(1/N)$, which means that in the case of a random jump each vertex is equally likely to be selected. Modifying the values of $\mathbf{v}$ effectively changes these probabilities and thus makes certain nodes more important. The version of PageRank for which the values in $\mathbf{v}$ are not uniform is called *Personalized PageRank* (PPR). The words in the text that are to be disambiguated are inserted as nodes in the KG and are connected to their potential senses via directed arcs. These newly introduced nodes serve to inject initial probability mass (via the vector **v**) and thus to make their associated sense nodes especially relevant in the *knowledge graph.* Applying the PPR algorithm iteratively over the resulting graph determines the most appropriate sense for each ambiguous word. Montroyo et al. [8] present a combination of knowledge-based and supervised systems for WSD, which demonstrate that the two approaches can boost one another, due to the fundamentally different types of knowledge they utilise (paradigmatic vs. syntagmatic). They explore a knowledge-based system that uses heuristics for WSD depending on the position of word potential senses in the WordNet knowledge graph (**WN**). In terms of supervised machine learning based on an annotated corpus, it explores a Maximum Entropy model that takes into account multiple features from the context of the to-be-disambiguated word. This earlier line of research demonstrates that combining paradigmatic and syntagmatic information is a fruitful strategy, but it does so by doing the combination in a postprocessing step, i.e. by merging the output of two separate systems; also, it still relies on manually annotated data for the supervised disambiguation.

The success of KWSD approaches apparently depends on the quality of the knowledge graph – whether the knowledge represented in terms of nodes and relations between them is sufficient for the algorithm to pick the correct senses of ambiguous words. An approach similar to ours is described in [9], which explores the extraction of syntactically supported semantic relations from manually annotated corpora: *SemCor*. *SemCor* was processed with the MiniPar parser and the subject-verb and object-verb relations were extracted. The new relations were represented on several levels: as word-to-class and class-to-class relations. The extracted selectional relations were then added to WordNet. The main difference with our approach is that the set of relations used in our work is larger, including whole sentences in which different n-ary relations are encoded such as subject-verb-object-indirect-object relations, adjective-noun-noun relations, etc.). In our case we have added much more relations.

3 The Sense Annotated Resources: *SemCor* and eXtended WordNet

As it was stated above, our goal is to experiment with different kinds of semantic relations. The relations missing in WordNet are the syntagmatic ones. As sources of such types of relations we consider semantically annotated resources extended with syntactic information. In this case we are able to extract syntagmatic relations between semantic classes of syntactically related words. For English we use a parsebank created over the texts in *SemCor* and XWN which is annotated also with syntax and logical forms. Since *SemCor* has been exploited for extracting relations, it was divided into test and training parts in ratio of one-to-three.

The sense annotations in *SemCor* were also performed manually on the base of WordNet. It comprises texts from Brown corpus[4] which is a balanced corpus. In this respect *SemCor* contains really diverse types of texts. We use *SemCor* in two ways: first, for testing the WSD for English; and second, as a source for extracting of new semantic relations. To achieve this, we parsed *SemCor* with a dependency parser included in the IXA pipeline[5]. Then we divided the corpus into a proportion one-to-three: first part comprises 49 documents (from br-a01 to br-f44) and it was used as a test set in the experiments reported below in the paper. The rest of the documents formed the training set from which the new relations were extracted. The new semantic relations were extracted on the basis of the syntactic relations in the dependency parses of each sentence in the training part of *SemCor*.

4 Knowledge-Based Tool for the WSD

The experiments that serve to illustrate the outlined approaches were carried out with the UKB[6] tool, which provides graph-based methods for WSD and measuring lexical similarity. The tool uses a set of random walk on graph algorithms, described in [7]. The tool builds a knowledge graph over a set of relations that can be induced from different types of resources, such as WordNet or DBPedia; then it selects a context window of open class words and runs the algorithm over the graph. We have used the UKB default settings, i.e. a context window of 20 words that are to be disambiguated together, and 30 iterations of the PPR algorithm. The UKB tool requires two resource files to process the input file. One of the resources is a dictionary file with all lemmas that can be possibly linked to a sense identifier. In our case WordNet-derived relations were used for our knowledge base; consequently, the sense identifiers are WordNet IDs. For instance, a line from the dictionary extracted from WordNet looks like this:

```
predicate 06316813-n:0 06316626-n:0 01017222-v:0
          01017001-v:0 00931232-v:0
```

First comes the lemma associated with the relevant word senses, after the lemma the sense identifiers are listed. Each `ID` consists of eight digits followed by a hyphen and a label referring to the POS category of the word. Finally, a number following a colon indicates the frequency of the word sense, calculated on the basis of a tagged corpus. When a lemma from the dictionary has occurred in the analysis of the input text, the tool assigns all the associated word senses to the word form in the context and attempts to disambiguate its meaning among them.

The second resource file required for running the tool is the set of relations that is used to construct the knowledge graph over which algorithms are run.

[4] http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM.

[5] http://ixa.si.ehu.es/Ixa.

[6] http://ixa2.si.ehu.es/ukb/.

As an initial knowledge graph we are using the resource files for version 3.0, distributed together with the tool, have been used in our experiments. The distribution of UKB comes with a file containing the standard lexical relations defined in WordNet, such as hypernymy, meronymy, etc., as well as with a file containing relations derived on the basis of common words found in the synset glosses, which have been manually disambiguated. The format of the relations in the KG is as follows:

```
u:SynSetId01 v:SynSetId02 s:Source d:w
```

where `SynSetId01` is the identifier of the first synset in the relation, `SynSetId02` is the identifier of the second synset, `Source` is the source of the relation, and `w` is the weight of the relation in the graph. In the experiments reported in the paper, the weight of all relations is set to 0.

This tool is used for performing all the experiments reported in the next section. The goal in this paper is to investigate the impact of the different sets of relations over the knowledge graph.

5 New Knowledge Graph from Logical Form

Here we present an approach towards the enrichment of WSD knowledge bases with relations from gold standard corpora. In our previous work we focused on Bulgarian (*BTB* [10]) and English (*SemCor* [3]) corpora as use cases and as sources of new semantic relations. The extraction of new semantic relations from gold corpora is a mechanism for balancing the predominance of paradigmatic knowledge (synonymy, hypernymy, etc.) by the addition of syntagmatic relations.

The new relations are extracted from eXtended WordNet (XWN) by using the logical form of glosses in WordNet. This corpus was already used for the extraction of semantic relations from the co-occurrences of the synset concept and the concepts assigned during the annotation to the words in the gloss. For example, the synset {disyllable, dissyllable} — 06290539-n, is defined by “a word having two syllables.” After the analysis, the following synsets are selected: 06286395-n — *word*, 06304671-n — *syllable*, 02203362-v — *have*. Each of these synsets is related to the synset which the gloss belongs to[7]:

```
u:06290539-n v:06286395-n
u:06290539-n v:06304671-n
```

The logical form for this gloss in XWN is the following

```
disyllable:NN(x1) ->
   word:NN(x1) have:VB(e1, x1, x2)
   two:JJ(x2) syllable:NN(x2)
```

[7] In the knowledge graph constructed in this way and distributed with the UKB system, the relation between the noun synset and the verb synset for *have* is not presented.

In our opinion, each predicate that originates from a verbal, adjectival, adverbial, or prepositional lemma expresses an event. In the example, `have:VB(e1, x1, x2)` denote the event of “holding” of object denoted by `x2` by the object denoted by `x1`. Both of these objects are participants of the event of “holding” `e1`. From this we extract the following relations:

```
u:02203362-v v:06286395-n
u:02203362-v v:06304671-n
u:06286395-n v:06304671-n
```

In the case of {ice-cream cone} defined by “ice cream in a crisp conical wafer” the following logical form is presented:

```
ice-cream_cone:NN(x1) ->
   ice_cream:NN(x1) in:IN(x1, x2)
   crisp:JJ(x2) conical:JJ(x2)
   wafer:NN(x2)
```

From it we have extracted relations between “ice cream” and “wafer” on the basis of the predicate `in:IN(x1, x2)`, also between “crisp” and “wafer”, “conical” and “wafer”, and between “crisp” and “conical” in the appropriate senses. This set of relations forms a knowledge graph which we denote as **WNGL** — knowledge graph constructed on the basis of the logical form of the glosses in WordNet.

6 Semantically Annotated Sentences as Context for WSD

We consider each sentence in a semantically annotated corpora as representation of a context in which the WSD is already performed by an expert. These contexts are similar to the context created by the UKB system during the WSD task. In our view this explains the good results reported below. In this section we present two approaches to represent such contexts as knowledge graphs for WSD.

The extraction of contexts from manually annotated corpora with word senses can be performed in different ways. The connection between the nodes that are semantically annotated can be determined at least in two ways:

- As a sequence of nodes corresponding to the order of the words in the sentence;
- As a syntactic structure of nodes corresponding to a parse of the sentence.

The second approach is demonstrated on the basis of SemCor. The training part of SemCor was parsed with the dependency parser from the IXA pipeline. The set of relations presented here is based on the dependency tree for each sentence. Each node in the dependency tree corresponds to a new node for the relevant word. Then it is related to the head node. An additional relation points to the node corresponding to the WordNet Synset. Since SemCor consists of text fragments, the sentences that are from the same fragment are connected via relations between the roots of the dependency trees. The root of the second tree is related to the root of the first sentence, the root of the third sentence to the root of the second sentence, etc.

For example, the sentence "Evidence that other sources of financing are unavailable must be provided." is analyzed with a dependency parser of IXA pipeline. From this analysis we construct a set of relations corresponding to the syntactic tree. Figure 1 depicts the top

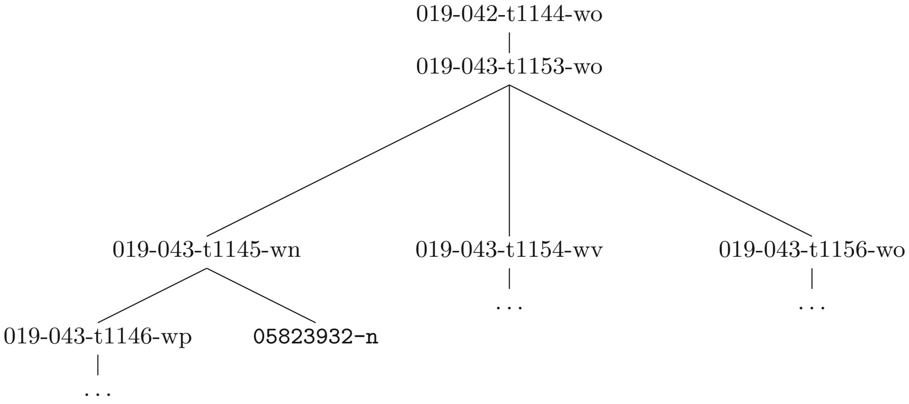

Fig. 1. The top fragment of the dependency tree

where 019-043-t****-** represents the nodes in the dependency tree, the first three digits represent: the number of the file from which the sentence is selected, the number of the sentence and the number of tokens in the sentence. Nodes like 05823932-n ("Evidence") are from the knowledge graph of WordNet V3.0. There are nodes of the syntactic tree that are not mapped to a synset, because not each word in the sentence is mapped to a synset. The relation between 019-043-t1153-wo and 019-042-t1144-wo is the relation between the root of the sentence and the root of the previous sentence. The set is called **GraphRelSC**. Note that the parsing of the sentences is done automatically. Thus, there might be errors.

The first approach mentioned above on the representation of context is illustrated on the basis of glosses in XWN. For each ⟨gloss⟩ element in XWN we consider the element ⟨wsd⟩, containing the words of the gloss with assigned synset id from WordNet V2.0:

```
<wsd>
<wf>a</wf>
<wf wn20="ENG20-05501538-n" wnsn="1">kind</wf>
<wf>of</wf>
<wf wn20="ENG20-02650459-n">artificial heart</wf>
<wf>that</wf>
<wf wn20="ENG20-02139918-v">has</wf>
```

```
<wf wn20="ENG20-02526983-v">been</wf>
<wf wn20="ENG20-01123102-v">used</wf>
<wf>with</wf>
<wf>some</wf>
<wf wn20="ENG20-06869923-n">success</wf>
</wsd>
```

We have performed the following operations:

- The synset id for synset of the gloss is added as a first element;
- The WordNet V2.0 ids are converted to WordNet V3.0 ids as they are used in UKB knowledge graphs;
- For each word annotated with WordNet id we created a node which connects to the node of the corresponding WordNet synset and to the node of the preceding word annotated with WordNet synset id in the gloss.

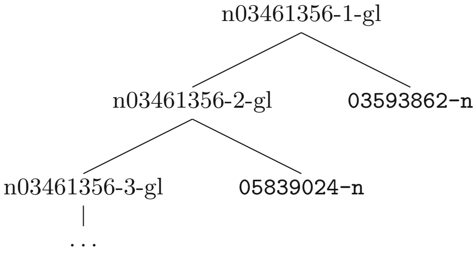

Fig. 2. The beginning of graph for the sequence of words in the gloss and their relation to synsets nodes.

Having performed this procedure on the above gloss example we created a set of relations depicted in Fig. 2. The nodes with labels `n03461356-*-gl` are nodes corresponding to the word in the gloss and the other nodes are corresponding to the synsets in WordNet V3.0. We call the resulting set of relations **WN30glCon**.

7 Experiments

The experiments that illustrate the outlined approaches were carried out with the UKB tool, which provides graph-based methods for WSD and measuring lexical similarity. We have performed experiments with two algorithms implemented within UKB: Static and PPRw2w — see [7]. We have selected these two because the latter has shown the best results during our previous experiments and the first one is the fastest one. As it was mentioned, we exploit the semantically annotated corpus *SemCor*. As baselines we consider the results achieved with the standard knowledge graphs distributed within UKB system: **WN** for the relations in WordNet and **WNG** for the relations extracted from **XWN** on the basis of co-occurrences.

The new knowledge graphs are: **WNGL** — a knowledge graph based on analysis of the logical forms of the glosses in **XWN**; **GraphRelSC** — a knowledge graph comprising sentences from *SemCor*; and **WN30glCon** — a knowledge graph comprising sentences from **XWN**. The results for these knowledge graphs are compared to the baselines and also to some combinations of them. The results show improvement on *SemCor* for both algorithms — Table 1.

Table 1. Comparison of the results for the standard knowledge graphs with the newly constructed knowledge graphs.

KG	*Static*	*PPRw2w*
WN	56.60	56.35
WNG	56.00	57.33
WN + WNG	**59.55**	**62.24**
WNGL	60.46	60.35
WN + WNGL	66.61	67.19
WN + WN30glCon	67.00	66.42
WN + GraphRelSC	67.04	65.97
WN + GraphRelSC + WNGL	68.41	68.51
WN + WN30glCon + GraphRelSC	68.74	68.15
WN + WN30glCon + GraphRelSC + WNGL	**68.77**	68.48
WN + WNG + WN30glCon + GraphRelSC + WNGL	68.39	**68.59**

The results show the following important facts: (1) the combination of relations from different sources might improve the results significantly[8]; (2) the improvement is not monotonic with respect to the number of the relations. Obviously the topology of the graph plays an important role for the Random Walk on Graphs algorithms. Also in current experimental setup only local context in the text is considered. Thus, if two senses share local connectivity in the text, they will be hard for disambiguation even when more relations are added. This problem will be studied in our future work.

8 Conclusion

The experiments with adding various bundles of relations from WordNet and from syntactically and semantically annotated corpora for English have shown several directions to be considered in our future work.

[8] This result for English is far from state-of-the-art, but it is based only on 25 % of *SemCor*. Also, our goal here is only to compare the various knowledge graphs.

First of all, the addition of syntagmatic syntactic-based relations in form of context improves the results of KWSD task, since they balance the paradigmatic lexical relations. Then, the accuracy depends also on the integrity of the domain – in more homogeneous domains the accuracy is more stable and increases, while in more heterogeneous domains the accuracy drops. We consider the accuracy as a measure of quality of the knowledge graph with respect to the KWSD task. The conclusion is that adding important linguistic and world knowledge in form of relations between lexical concepts does not necessarily improve the quality of the knowledge graph.

Another issue is the differing impact of the various relations on the knowledge graph. Since the quantity of the added information is huge, our idea was to reduce it through the selection of the contributing relations without losing the quality of the result. This strategy is not trivial. It requires a lot of sets of experiments as well as new mechanisms for evaluating the graph and optimizing the algorithm.

Acknowledgements. This research has received partial support by the EC's FP7 project: "QTLeap: Quality Translation by Deep Language Engineering Approaches" (610516).

References

1. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012). Elsevier
2. Fellbaum, C. (ed.): WordNet an Electronic Lexical Database. The MIT Press, Cambridge (1998)
3. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proceedings of HLT 1993, pp. 303–308 (1993)
4. Cruse, D.A.: Lexical Semantics. Cambridge University Press, Cambridge (1986)
5. Agirre, E., de López, O., Soroa, A.: Random walks for knowledge-based word sense disambiguation. Comput. Linguist. **40**(1), 57–84 (2014)
6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. **56**(18), 3825–3833 (2012)
7. Agirre, E., Soroa, A.: Personalizing PageRank for word sense disambiguation. In: Proceedings of 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 33–41 (2009)
8. Montoyo, A., Suárez, A., Rigau, G., Palomar, M.: Combining knowledge-and corpus-based word-sense-disambiguation methods. J. Artif. Intell. Res. (JAIR) **23**, 299–330 (2005)
9. Agirre, E., Martinez, D.: Integrating selectional preferences in WordNet. In: Proceedings of 1st International WordNet Conference (2002)
10. Popov, A., Kancheva, S., Manova, S., Radev, I., Simov, K., Osenova, P.: The sense annotation of BulTreeBank. In: Proceedings of TLT13, pp. 127–136 (2014)