EUROSENSE: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato and Roberto Navigli
Department of Computer Science
Sapienza University of Rome
{dellibovi,collados,raganato,navigli}@di.uniroma1.it

Abstract

Parallel corpora are widely used in a variety of Natural Language Processing tasks, from Machine Translation to cross-lingual Word Sense Disambiguation, where parallel sentences can be exploited to automatically generate high-quality sense annotations on a large scale. In this paper we present EUROSENSE, a multilingual sense-annotated resource based on the joint disambiguation of the Europarl parallel corpus, with almost 123 million sense annotations for over 155 thousand distinct concepts and entities from a language-independent unified sense inventory. We evaluate the quality of our sense annotations intrinsically and extrinsically, showing their effectiveness as training data for Word Sense Disambiguation.

1 Introduction

One of the long-standing challenges in Natural Language Processing (NLP) lies in automatically identifying the meaning of words in context. Various lines of research have been geared towards achieving this goal, most notably Word Sense Disambiguation (Navigli, 2009, WSD) and Entity Linking (Rao et al., 2013, EL). In both tasks, supervised approaches (Zhong and Ng, 2010; Melamud et al., 2016; Iacobacci et al., 2016; Kågebäck and Salomonsson, 2016) tend to obtain the best performances over standard benchmarks but, from a practical standpoint, they lose ground to knowledge-based approaches (Agirre et al., 2014; Moro et al., 2014b; Weissenborn et al., 2015), which scale better in terms of scope and number of languages. In fact, the development of supervised disambiguation systems depends crucially on the availability of reliable sense-annotated corpora, which are indispensable in order to provide solid training and testing grounds (Pilehvar and Navigli, 2014). However, hand-labeled sense annotations are notoriously difficult to obtain on a large scale, and manually curated corpora (Miller et al., 1993; Passonneau et al., 2012) have a limited size. Given that scaling the manual annotation process becomes practically unfeasible when both lexicographic and encyclopedic knowledge is addressed (Schubert, 2006), recent years have witnessed efforts to produce larger sense-annotated corpora automatically (Moro et al., 2014a; Taghipour and Ng, 2015a; Scozzafava et al., 2015; Raganato et al., 2016). Even though these automatic approaches produce noisier corpora, it has been shown that training on them leads to better supervised and semi-supervised models (Taghipour and Ng, 2015b; Raganato et al., 2016; Yuan et al., 2016; Raganato et al., 2017), as well as to effective embedded representations for senses (Iacobacci et al., 2015; Flekova and Gurevych, 2016).

A convenient way of generating sense annotations is to exploit parallel corpora and word alignments (Taghipour and Ng, 2015a): indeed, parallel corpora exist in many flavours (Tiedemann, 2012) and are widely used across the NLP community for a variety of different tasks. In this paper we focus on Europarl (Koehn, 2005)\(^1\), one of the most popular multilingual corpora, originally designed to provide aligned parallel text for Machine Translation (MT) systems. Extracted from the proceedings of the European Parliament, the latest release of the Europarl corpus comprises parallel text for 21 European languages, with more than 743 million tokens overall.

Apart from its prominent role in MT as a training set, the Europarl corpus has been used

\(^1\)http://opus.lingfil.uu.se/Europarl.php
for cross-lingual WSD (Lefever and Hoste, 2010, 2013), including, more recently, preposition sense disambiguation (Gonen and Goldberg, 2016), and widely exploited to develop cross-lingual word embeddings (Hermann and Blunsom, 2014; Gouws et al., 2015; Coulmance et al., 2015; Vyas and Carpuat, 2016; Vulić and Korhonen, 2016; Artetxe et al., 2016) as well as multi-sense embeddings (Ettinger et al., 2016; Šuster et al., 2016).

In this paper, our aim is to augment Europarl with sense-level information for multiple languages, thereby constructing a large-scale sense-annotated multilingual corpus which has the potential to boost both WSD and MT research.

We follow an approach that has already proved effective in a definitional setting (Camacho-Collados et al., 2016a): unlike previous cross-lingual approaches, we do not rely on word alignments against a pivot language, but instead leverage all languages at the same time in a joint disambiguation procedure that is subsequently refined using distributional similarity. We draw on the wide-coverage multilingual encyclopedic dictionary of BabelNet (Navigli and Ponzetto, 2012), which enables us to seamlessly cover lexicographic and encyclopedic knowledge in multiple languages within a unified sense inventory.

As a result of our disambiguation pipeline we obtain and make available to the community EUROSENSE, a multilingual sense-annotated corpus with almost 123 million sense annotations of more than 155 thousand distinct concepts and named entities drawn from the multilingual sense inventory of BabelNet, and covering all the 21 languages of the Europarl corpus. As such EUROSENSE constitutes, to our knowledge, the largest corpus of its kind.

2 Related Work

Extending sense annotations to multiple languages is a demanding endeavor, especially when manual intervention is required. Despite the fact that sense-annotated corpora for a number of languages have been around for more than a decade (Petrolito and Bond, 2014), they either include few samples per word sense, or only cover a restricted set of ambiguous words (Passonneau et al., 2012); as a result, multilingual WSD was until recently almost exclusively tackled using knowledge-based approaches (Agirre et al., 2014; Moro et al., 2014b). Nowadays, however, the rapid development of NLP pipelines for languages other than English has been opening up the possibilities for the automatic generation of multilingual sense-annotated data. Nevertheless, the few approaches that have been proposed so far are either focused on treating each individual language in isolation (Otegi et al., 2016), or limited to short and concise definitional text (Camacho-Collados et al., 2016a).

On the other hand, the use of parallel text to perform WSD (Ng et al., 2003; Lefever et al., 2011; Yao et al., 2012; Bonansinga and Bond, 2016) or even Word Sense Induction (Apidianaki, 2013) has been widely explored in the literature, and has demonstrated its effectiveness in producing high-quality sense-annotated data (Chan and Ng, 2005). This strategy, however, requires word alignments for each language pair to be taken into account, with alignment errors that might propagate and hamper subsequent stages unless human supervision is employed to correct erroneous annotations (Taghipour and Ng, 2015a). Moreover, cross-language disambiguation using parallel text requires a language-independent annotation framework that goes beyond monolingual WordNet-like sense inventories (Lefever et al., 2011) in order for the annotations obtained to be used effectively within end-to-end applications.

With EUROSENSE, instead, the key idea is to exploit at best parallel sentences to provide enriched context for a joint multilingual disambiguation. Using BabelNet, a unified multilingual sense inventory, we obtain language-independent sense annotations for a wide variety of concepts and named entities, which can be seamlessly mapped to individual semantic resources (e.g WordNet, Wikipedia, DBpedia) via BabelNet’s inter-resource mappings.

3 Building EUROSENSE

Following Camacho-Collados et al. (2016a), our fully automatic disambiguation pipeline for constructing EUROSENSE couples a graph-based multilingual joint WSD/EL system, Babelfy (Moro et al., 2014b)3, and a language-independent vector representation of concepts and entities, NASARI (Camacho-Collados et al., 2016b).4 It comprises two stages: multilingual disambigu-
tion (Section 3.1) and refinement based on distributional similarity (Section 3.2).

3.1 Stage 1: Multilingual Disambiguation

As a preprocessing step, we part-of-speech tag and lemmatize the whole corpus using TreeTagger (Schmid, 1995). We perform disambiguation at the sentence level. However, instead of disambiguating each sentence in isolation, language by language, we first identify all available translations of a given sentence and then gather these together into a single multilingual text.

Then, we disambiguate this multilingual text using Babelfy. Given that Babelfy is capable of handling text with multiple languages at the same time, this multilingual extension effectively increases the amount of context for each sentence, and directly helps in dealing with highly ambiguous words in any particular language (as the translations of these words may be less ambiguous in some different language). Moreover, given the multilingual nature of our sense inventory, Babelfy’s high-coherence approach favors naturally sense assignments that are consistent across languages at the sentence level (i.e. those having fewer distinct senses shared by more translations of the same sentence).

As a result, we obtain a full, high-coverage version of EUROSENSE where each disambiguated word or multi-word expression (disambiguated instance) is associated with a coherence score.\(^5\)

3.2 Stage 2: Similarity-based Refinement

In this stage we aim at improving the sense annotations obtained in the previous step (Section 3.1), with a procedure specifically targeted at correcting and extending these sense annotations. In general, graph-based WSD systems, such as Babelfy, have been shown to be heavily biased towards the Most Common Sense (MCS) (Calvo and Gelbukh, 2015). In order to get a handle on this bias and improve our pipeline’s disambiguation accuracy we adopt a refinement based on distributional similarity, which is not affected by the MCS.

To this end, we exploit the 300-dimensional embedded representations of concepts and entities of NASARI to discard or refine disambiguated instances that are less semantically coherent. These NASARI vector representations were constructed by combining structural and distributional knowledge from Wikipedia and WordNet with Word2Vec word embeddings (Mikolov et al., 2013) trained on textual corpora.

For each sentence, we first identify a subset \(D\) of high-confidence disambiguations\(^7\) from among those given by Babelfy in the previous step. Then, we calculate the centroid of all the NASARI vectors corresponding to the elements of \(D\), and we re-disambiguate the mentions associated with the remaining low-confidence disambiguated instances (i.e. those not in \(D\)), by picking, for each mention \(w\), the concept or entity \(\hat{s}\) whose NASARI vector\(^8\) is closest to the centroid of the sentence:

\[
\hat{s} = \arg\max_{s \in S_w} \cos \left( \frac{\sum_{d \in D} \vec{d}_w^T \vec{s}}{|D|}, \vec{s} \right)
\]

(1)

where \(S_w\) is the set of all candidate senses for mention \(w\) according to BabelNet. Cosine similarity (\(\cos\)) is used as similarity measure. Finally, in order to discard less confident annotations, we consider the cosine value associated with each refined disambiguation as confidence score, and use it to compare each disambiguated instance against an empirically validated threshold of 0.75.

As a result, we obtain the refined high-precision version of EUROSENSE, where each disambiguated instance is associated with both a coherence score and a distributional similarity score.

4 Corpus and Statistics

Table 1 reports general statistics on EUROSENSE regarding both its high-coverage (cf. Section 3.1) and high-precision (cf. Section 3.2) versions. Joint multilingual disambiguation with Babelfy generated more than 215M sense annotations of 247k distinct concepts and entities, while similarity-based refinement retained almost 123M high-confidence instances (56.96% of the total), covering almost 156k distinct concepts and entities. 42.40% of these retained annotations were corrected or validated using distributional similarity. As expected, the distribution over parts of speech is skewed towards nominal senses (64.79% before refinement and 81.79% after refinement).

\(^5\)We rely on the internal preprocessing pipeline of Babelfy for those languages not supported by TreeTagger.

\(^7\)As in Camacho-Collados et al. (2016a), coherence score is given by the normalized number of connections of a given concept within the sentence.

\(^8\)Given a concept or entity \(s\) we indicate with \(\vec{s}\) its corresponding NASARI vector.
Table 1: General statistics on EUROSENSE before (full) and after refinement (refined) for all the 21 languages. Language-specific figures are also reported for the 4 languages of the intrinsic evaluation.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>EN</th>
<th>FR</th>
<th>DE</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Annotations</td>
<td>215 877 109</td>
<td>26 455 574</td>
<td>22 214 996</td>
<td>16 888 108</td>
<td>21 486 532</td>
</tr>
<tr>
<td>Distinct lemmas covered</td>
<td>567 378</td>
<td>60 853</td>
<td>30 474</td>
<td>66 762</td>
<td>43 892</td>
</tr>
<tr>
<td>Distinct senses covered</td>
<td>247 706</td>
<td>138 115</td>
<td>65 301</td>
<td>75 008</td>
<td>74 214</td>
</tr>
<tr>
<td>Average coherence score</td>
<td>0.19</td>
<td>0.19</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td><strong>Refined</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Annotations</td>
<td>122 963 111</td>
<td>15 441 667</td>
<td>12 955 469</td>
<td>9 165 112</td>
<td>12 193 260</td>
</tr>
<tr>
<td>Distinct lemmas covered</td>
<td>453 063</td>
<td>42 947</td>
<td>23 603</td>
<td>50 681</td>
<td>31 980</td>
</tr>
<tr>
<td>Distinct senses covered</td>
<td>155 904</td>
<td>86 881</td>
<td>49 189</td>
<td>52 425</td>
<td>52 859</td>
</tr>
<tr>
<td>Average coherence score</td>
<td>0.29</td>
<td>0.28</td>
<td>0.25</td>
<td>0.28</td>
<td>0.27</td>
</tr>
</tbody>
</table>

followed by verbs (19.26% and 12.22%), adjectives (11.46% and 5.24%) and adverbs (4.48% and 0.73%). We note that the average coherence score increases from 0.19 to 0.29 after refinement, suggesting that distributional similarity tends to favor sense annotations that are also consistent across different languages. Table 1 also includes language-specific statistics on the 4 languages of the intrinsic evaluation, where the average lexical ambiguity ranges from 1.12 senses per lemma (German) to 2.26 (English) and, as expected, decreases consistently after refinement.

Interestingly enough, if we consider all the 21 languages, the total number of distinct lemmas covered is more than twice the total number of distinct senses: this is a direct consequence of having a unified, language-independent sense inventory (BabelNet), a feature that sets EUROSENSE apart from previous multilingual sense-annotated corpora (Otegi et al., 2016). Finally we note from the global figures on the number of covered senses that 109 591 senses (44.2% of the total) are not covered by the English sense annotations: this suggests that EUROSENSE relies heavily on multilinguality in integrating concepts or named entities that are tied to specific social or cultural aspects of a given language (and hence would be underrepresented in an English-specific sense inventory).

5 Experimental Evaluation

We assessed the quality of EUROSENSE’s sense annotations both intrinsically, by means of a manual evaluation on four samples of randomly extracted sentences in different languages (Section 5.1), as well as extrinsically, by augmenting the training set of a state-of-the-art supervised WSD system (Zhong and Ng, 2010) and showing that it leads to consistent performance improvements over two standard WSD benchmarks (Section 5.2).

5.1 Intrinsic Evaluation: Annotation Quality

In order to assess annotation quality directly, we carried out a manual evaluation on four different languages (English, French, German and Spanish) with 2 human judges per language. We sampled 50 random sentences across the subset of sentences in EUROSENSE featuring a translation in all 4 languages, totaling 200 sentences overall.

For each sentence, we evaluated all sense annotations both before and after the refinement stage, along with the sense annotations obtained by a baseline that disambiguates each sentence in isolation with Babelfy. Overall, we manually verified a total of 5818 sense annotations across the three configurations (1518 in English, 1564 in French, 1093 in German and 1643 in Spanish). In every language the two judges agreed in more than 85% of the cases, with an inter-annotator agreement in terms of Cohen’s kappa (Cohen, 1960) above 60% in all evaluations (67.7% on average).

Results, reported in Table 2, show that joint multilingual disambiguation improves consistently over the baseline. The similarity-based refinement boosts precision even further, at the expense of a reduced coverage (whereas both Babelfy and the baseline attempt an answer for every disambiguation target). Over the 4 languages, sense annotations appear to be most reliable for German, which is consistent with its lower lexical ambiguity on the corpus (cf. Section 4).

5.2 Extrinsic Evaluation: Word Sense Disambiguation

We additionally carried out an extrinsic evaluation of EUROSENSE by using its refined sense an-
Table 2: Precision (Prec.) and coverage (Cov.) of EUROSENSE, manually evaluated on a random sample in 4 languages. Precision is averaged between the two judges, and coverage is computed assuming each content word in the sense inventory to be a valid disambiguation target.

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>FR</th>
<th>DE</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babelfy</td>
<td>76.1</td>
<td>100</td>
<td>59.1</td>
<td>100</td>
</tr>
<tr>
<td>EUROSENSE (full)</td>
<td>80.3</td>
<td>100</td>
<td>67.9</td>
<td>100</td>
</tr>
<tr>
<td>EUROSENSE (refined)</td>
<td>81.5</td>
<td>75.0</td>
<td>71.8</td>
<td>63.5</td>
</tr>
</tbody>
</table>

Table 3: F-Score on all-words WSD.

<table>
<thead>
<tr>
<th></th>
<th>SemEval-2013</th>
<th>SemEval-2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS_SemCor</td>
<td>65.3</td>
<td>69.3</td>
</tr>
<tr>
<td>IMS_OMSTI</td>
<td>65.0</td>
<td>69.1</td>
</tr>
<tr>
<td>IMS_EUROSENSE</td>
<td>66.4</td>
<td>69.5</td>
</tr>
<tr>
<td>UKB</td>
<td>59.0</td>
<td>61.2</td>
</tr>
<tr>
<td>UKB_w2w</td>
<td>62.9</td>
<td>63.3</td>
</tr>
<tr>
<td>MCS</td>
<td>63.0</td>
<td>67.8</td>
</tr>
</tbody>
</table>

6 Release

EUROSENSE is available at [http://lcl.uniroma1.it/eurosense](http://lcl.uniroma1.it/eurosense). We release two different versions of the corpus:

- A high-coverage version, obtained after the first stage of the pipeline, i.e. multilingual joint disambiguation with Babelfy. Here, each sense annotation is associated with a coherence score (cf. Section 3.1);

- A high-precision version, obtained after the similarity-based refinement with NASARI. In this version, sense annotations are associated

7 Conclusion

In this paper we presented EUROSENSE, a large multilingual sense-annotated corpus based on Europarl, and constructed automatically via a disambiguation pipeline that exploits the interplay between a joint multilingual disambiguation algorithm and a language-independent vector-based representation of concepts and entities. Crucially, EUROSENSE relies on the wide-coverage unified sense inventory of BabelNet, which enabled the disambiguation process to exploit at best parallel text and enforces cross-language coherence among sense annotations. We evaluated EUROSENSE both intrinsically and extrinsically, showing that it provides reliable sense annotations that improve supervised models for WSD.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE Contract No. 726487.

Claudio Delli Bovi is supported by a Sapienza Research Grant ‘Avvio alla Ricerca 2016’. Jose Camacho-Collados is supported by a Google PhD Fellowship in Natural Language Processing.
References


