Using Wiktionary as a resource for WSD : the case of French verbs

Vincent Segonne, Marie Candito, Benoît Crabbé

Using Wiktionary as a Resource for WSD : The Case of French Verbs

Vincent Segonne[1], Marie Candito[1], Benoît Crabbé[1]

[1]Université Paris Diderot
[1]Laboratoire de Linguistique Formelle
vincent.segonne@linguist.univ-paris-diderot.fr
marie.candito@linguist.univ-paris-diderot.fr
benoit.crabbe@linguist.univ-paris-diderot.fr

Abstract

As opposed to word sense induction, word sense disambiguation (WSD), whether supervised or semi-supervised, has the advantage of using interpretable senses, but requires annotated data, which are quite rare for most languages except English (Miller et al., 1993). In this paper, we investigate which strategy to adopt to achieve WSD for languages lacking data that was annotated specifically for the task, focusing on the particular case of verb disambiguation in French. We first study the usability of Eurosense (Bovi et al. 2017), a multilingual corpus extracted from Europarl (Kohen, 2005) and automatically annotated with BabelNet (Navigli and Ponzetto, 2010) senses. Such a resource opened up the way to supervised and semi-supervised WSD for resourceless languages like French. While this perspective looked promising, our evaluation showed the annotated senses' quality was not sufficient for supervised WSD on French verbs. Instead, we propose to use Wiktionary, a collaboratively edited, multilingual online dictionary, as a new resource for WSD. Wiktionary provides both sense inventory and manually sense tagged examples which can be used to train supervised and semi-supervised WSD systems. Yet, because senses' distribution differ in lexicographic examples as found in Wiktionary with respect to natural text, we then focus on studying the impact on WSD of the training data size and senses' distribution. Using state-of-the art semi-supervised systems, we report experiments of wiktionary-based WSD for French verbs, evaluated on FrenchSemEval (FSE), a new dataset of French verbs manually annotated with wiktionary senses.

1 Introduction

Word Sense Disambiguation (WSD) is a NLP task aiming at identifying the sense of a word occurrence from its context, given a predefined sense inventory. Although the task emerged almost 70 years ago with the first work on Automatic Machine Translation (Weaver, 1955), it remains unresolved. The recent breakthrough in neural net models allowed a better representation of the context and thus improved the quality of supervised disambiguation systems (Melamud et al., 2016; Yuan et al., 2016; Peters et al., 2018). Nevertheless, although WSD has the advantage of providing interpretable senses (as opposed to the unsupervised task of word sense induction), it also has the drawback of heavily relying on the availability and quality of sense-annotated data, even in the semi-supervised setting.

Now, such data is available in English, essentially with SemCor (Miller et al., 1993), a corpus manually sense-annotated with Wordnet (Miller, 1995) senses. But for most languages, sense disambiguated data are very rare or simply don't exist. This is mainly due to the fact that manual semantic annotation is very costly in time and resources (Navigli, 2009). Nevertheless, Bovi et al. (2017) recently presented Eurosense, a multilingual automatically sense-disambiguated corpus extracted from Europarl (Koehn, 2005) and annotated with BabelNet (Navigli and Ponzetto, 2012) senses.

In this article, we focus on supervised WSD for French verbs and investigate a way to perform the task when no manually sense-annotated training data specifically designed for the task are available.We

focus on verbs because they are known to be central to understanding tasks, but also known to lead to lower WSD performance (Raganato et al., 2017). In section 2 we report a study on the suitability of using Eurosense as training data for our task. Because the results of our evaluation were inconclusive, we decided to explore Wiktionary, a free collaboratively edited multilingual online dictionary which provides a sense inventory and manually sense tagged examples, as resource for WSD. We give a general description of Wiktionary in section 3. In section 4, we present FrenchSemEval, a new manually sense annotated dataset for French verbs, to serve as evaluation data for WSD experiments using Wiktionary as sense inventory and training examples. Because senses' distribution differ in the lexicographic examples found in Wiktionary with respect to natural text, we provide in section 5 a descriptive statistical comparison of the wiktionary example corpus and SemCor. The WSD first experiments are reported in section 6 and we provide an analysis of the results in section 7. We finally conclude and give insights of future work in section 8.

2 Eurosense

In this section we present our investigation on the suitability of using Eurosense as training data for supervised WSD on French verbs. We first present Eurosense and then describe our manual evaluation of the resource regarding French verbs.

Eurosense is a multilingual Europarl-based corpus that was automatically sense-annotated using the BabelNet (Navigli and Ponzetto, 2012) multilingual sense inventory. This sense inventory was originally built by merging the English Wordnet and the English Wikipedia. Subsequent releases integrate senses (mapped or added) from other resources, as listed in BabelNet statistics page[1]. Two versions of Eurosense are available: the "high coverage" corpus and the "high precision" corpus. Both result from jointly disambiguating the parallel Europarl corpus using the Babelfy (Moro et al., 2014) WSD system. A further refinement step was then performed to obtain the high precision version. The refinement aims at enforcing the intra-sentence coherence of annotated senses, in terms of similarity of their corresponding Nasari distributional vectors (Camacho-Collados et al., 2016). The resource was evaluated both intrinsically and extrinsincally.

The intrinsic evaluation was carried out through a manual evaluation of the annotation on four languages (English, French, German and Spanish). Fifty sentences present in all four languages were randomly sampled and manually evaluated both before and after the refinement step. The results showed a good inter-annotator agreement the judges agreed 85% of the time, and the average Kappa score (Cohen, 1968) was 67.7 %) and an improvement of the precision of the annotation after refinement at the expense of a lower coverage. For English, the high-precision Eurosense annotations cover 75% of the content words and have a 81.5% precision[2]. As For French results are lower though: coverage is 71.8% and precision is 63.5%.

Closer to our WSD objective, Bovi et al. (2017) report an extrinsic evaluation of Eurosense that uses it as additional training data for all-words WSD, evaluated on two standard datasets for English, the SemEval 2013 task 12 (Navigli et al., 2013) and the SemEval 2015 task 13 (Moro and Navigli, 2015). The authors compared results of the *It Make Sense* (IMS) system (Zhong and Ng, 2010) when trained on SemCor alone versus SemCor augmented with examples sampled from the high precision Eurosense corpus (up to 500 additional training examples per sense). They report a slight improvement in the latter case, the Fscore rising from 65.3 to 66.4 on SemEval-2013, and from 69.3 to 69.5 on SemEval-2015.

These results give a contrasted picture of the usability of Eurosense as training data for WSD for French: the extrinsic evaluation concerns English, and a setting that uses Eurosense as additional data (picking examples from Eurosense annotated with senses present in SemCor only). For French, the situation is necessarily worse, given that (i) the intrinsic evaluation of Eurosense is lower for French, (ii) we focus on verbs, whose disambiguation is known to be more difficult (Raganato et al., 2017), and (iii),

[1] `https://babelnet.org/stats`

[2] Note the precision is computed as an average between the two annotators, and not with respect to an adjudicated version.

most importantly, Eurosense would be used as sole training data and not as additional data. This poses extra constraints on the quality of the annotations, hence in order to further investigate the usability of Eurosense for our purpose, we decided to first perform another evaluation of the Eurosense annotations, focused on French verbs.

Evaluation of Eurosense on French verbs We randomly selected 50 sentences from the French version of Eurosense's high coverage corpus, and for all the non-auxiliary verbs (160 occurrences), we manually judged whether the Eurosense automatic sense tag was correct : we split the 160 occurrences in three sets, each set being independently annotated by two judges and adjudicated by a third one. The judges were asked to answer "correct" if the Eurosense tag seemed correct, even if some other tag in the BabelNet inventory[3] was possible or even more precise. The agreement between the two judges was 0.72, and the kappa was 0.67, a level described as good in the literature. Note this is a binary task only, which is different from asking the judges to annotate the correct tag as Bovi et al. (2017) did for all parts-of-speech. Yet our agreement score is even lower, shedding light on an overall greater difficulty of judging the annotated sense of verbs. Indeed, we were then able to measure that the proportion of Eurosense automatic annotations that we judged correct after our adjudication is 44% only. Moreover, during this annotation task we could notice that because BabelNet aggregates several data sources (including Wordnet, Verbnet, FrameNet, Wiktionary among others), the BabelNet sense inventory exhibits a high number of senses per verb. To better quantify this observation, we sampled 150 sentences, and measured that the average number of BabelNet senses per verb type occurring in these sentences is 15,5. More importantly, we could notice that the frontiers of the various senses sometimes appeared difficult to judge, making it difficult to grasp the exact perimeter of a sense.These mixed results led us to investigate other sources of sense-annotated data for French.

3 Wiktionary as data for WSD

Wiktionary is a collaboratively edited, open-source multilingual online dictionary, hosted by the Wikimedia Foundation. It provides an interesting open-source resource and several studies already showed its usefulness for various NLP tasks (e.g. lemmatization (Liebeck and Conrad, 2015)), especially in the lexical semantic field, for extracting or improving thesauri (Navarro et al., 2009; Henrich et al., 2011; Miller and Gurevych, 2014). In this section we briefly present Wiktionary's most interesting features along with our motivations to investigate the use of this resource for WSD on French verbs.

Wiktionary's main advantages is that it is entirely open-source, multilingual and has a good coverage for a substantial number of languages (according to wiktionary statistics[4], 22 languages have more than $50,000$ wiktionary entries each). Each entry consists of a definition and one or several examples, either attested or created, each example being a potential sense-annotated example for the lemma at hand. Definitions and examples point to other wiktionary pages, which can be useful, although not as useful as if links to wiktionary senses (not pages) would be provided. The structured nature of wiktionary makes it possible to extract wordnets rather easily (as was done for English, German and French by Sérasset (2012), in the RDF format). On the qualitative level, our interest for wiktionary rose after studying random verbal entries for French: we could observe that in general the granularity level is "natural" and that the sense distinctions are easy to grasp. On the quantitative level, we report in table 1 several statistics for the French wiktionary[5], in which it can be seen that the resource is large (we will see in the next section that the coverage in corpus is good indeed).

[3]The sense inventory had been previously extracted via the HTTP API see `https://babelnet.org/guide`

[4]`https://en.wiktionary.org/wiki/Wiktionary:Statistics`

[5]In all the following, all the statistics and work on the French wiktionary corresponds to the 04-20-2018 dump available via Dbnary (Sérasset, 2012).

POS	Nb of entries	Nb of senses	Mean nb of senses per entry	Nb of examples
Noun	81099	112428	1.39	1511517
Verb	27271	41207	1.51	55206
Adj	25865	33732	1.30	46212
Adv	5904	6012	1.29	5904

Table 1: Statistics from French Wiktionary of the 04-20-2018 dump available via the tool of Sérasset (2012)

These advantages come at the cost of Wiktionary's main potential drawback, namely its crowd-sourced nature. Firstly, this means that it is constantly evolving, since any user can edit pages at any time (unless pages that users with more editing rights might have protected). Indeed, new pages are created every day while already existing pages are deleted, modified, merged (note though that every change occurring in the resource is kept in track). Secondly, this means that the resource is not curated by skilled lexicographers only, and the "guidelines" are themselves collaboratively built.

Despite this potential disadvantage, several features of Wiktionary seemed particularly suitable for the task of WSD and this, combined with the fact that sense-annotated data for French verbs are quasi-inexistent[6], makes it a serious candidate for a new resource of WSD. To investigate this opportunity for our objective of French verb WSD, we present FrenchSemEval, a new dataset manually annotated for WSD of French verbs which we used to carry out several evaluations, we describe the new resource in the next section.

4 FrenchSemEval : An evaluation corpus for French verb disambiguation

Since the first Senseval evaluation serie in 1998 (Kilgarrif, 1998), a various number of evaluation frameworks were proposed to evaluate different WSD tasks, but only a few include French test datasets (Lefever and Hoste, 2010; Navigli et al., 2013) and unfortunately these only focus on nouns[7]. In this section we present FrenchSemEval[8] a new French dataset in which verb occurrences were manually annotated with Wiktionary senses. Our objective was to evaluate whether Wiktionary's sense inventory is operational for humans to sense-annotate a corpus, and if so, to use it as evaluation data for WSD experiments. We describe the annotation process along with several statistics about the resulting dataset and the quality of the annotations.

4.1 Data selection

To build FrenchSemEval, we chose to focus on moderately frequent and moderately ambiguous verbs. Rare verbs are often monosemous, and very frequent verbs tend to be very polysemous and extremely difficult to disambiguate (we thus left these for future work). FrenchSemEval was built using the following steps: we first selected a vocabulary of verbs based on their frequency in corpus. We selected verbs appearing between 50 and 1000 times in the French Wikipedia (dumped on 2016-12-12 hereafter fr-Wikipedia). Secondly, from this pre-selected list of verbs we extracted those having a number of senses comprised between two and ten in Wiktionary's sense inventory. For these verbs, we chose to extract 50 occurrences primarily from corpora comprising other annotations (the French TreeBank (FTB) (Abeillé

[6]Verbs are annotated with frames in the French FrameNet data (Djemaa et al., 2016), but in such data, only some notional domains were considered, and verb occurrences not pertaining to such domains were not disambiguated.

[7]Except for SensEval1 but only the English dataset was given to public domain.

[8]The dataset is available here `http://www.llf.cnrs.fr/dataset/fse/`

Number of sentences	3121
Number of annotated verb tokens	3199
Number of annotated verb types	66
Mean number of annotations per verb type	48.47
Mean number of senses per verb type	3.83

Table 2: Statistics for the FrenchSemEval corpus (FSE).

and Barrier, 2004) and the Sequoia (Candito and Seddah, 2012) treebank[9]), supplementing the corpus when necessary by occurrences sampled from fr-Wikipedia, in order to reach 50 occurrences per verb.

4.2 Annotation process

The annotation has been performed by three students[10] for nearly a month. We used WebAnno (Yimam et al., 2014; de Castilho et al., 2016) an open-source adaptable annotation tool. Sentences had already been pre-processed into CoNLL format (Nivre et al., 2007) with the *Mind The Gap* (MTG) parser (Coavoux and Crabbé, 2017) and were plugged in WebAnno. We were thus able to provide files (one file per verb) containing sentences in which occurrences of the specific verb were marked for annotation. The annotators were asked to annotate only the marked occurrences. We integrated in WebAnno the sense inventory from Wiktionary, including definitions and examples of senses, and added two extra tags: "OTHER_POS" and "MISSING_SENSE". The former was to use when an occurrence was wrongly tagged as verb, and the latter was to use when the sense of an occurrence didn't exist in the sense inventory. As Wiktionary is constantly evolving through time, we used the 04-20-2018 dump available via Dbnary (Sérasset, 2012). The annotation was performed in double annotation and adjudication.

4.3 Resulting resource

Table 2 reports various statistics about the resulting dataset. It contains 3199 occurrences for 66 different verbs, which means nearly 50 annotated instances per verb (about 100 OTHER_POS occurrences were discarded). The annotators agreed more than 70% of the time and obtained a Kappa score of 0.68 which is good according to the literature. We believe that these metrics indicate an annotation quality which may not be extremely high but still sufficient to validate the coherence of the Wiktionary sense inventory, definitions and examples, despite its non-expert crowd-sourced nature.

5 Descriptive statistical study of the datasets

The best-suited data for training a supervised WSD system is a corpus with sense tags for all content words. Training on such a corpus benefits from basic frequency information found in the corpus. This is particularly striking for WSD, as the "most frequent sense" baseline is known to be very high. In the case of French, as for the majority of languages, we lack such a corpus, and turn to the Wiktionary examples to serve as training examples for a significant portion of the lexicon. Yet, because senses' distribution differ in the lexicographic examples found in Wiktionary with respect to natural text, we first provide some statistics for a running text sense-annotated corpus such as SemCor (for English) versus a lexicographic training set such as Wiktionary examples (for French).

[9]The FTB contains 18500 sentences, from articles from Le Monde newspaper, and Sequoia contains 3099 sentences from Europarl, the European Medicine Agency, a regional newspaper (L'Est Républicain) and fr-Wikipedia.

[10]None of them had previous experience in annotation.

Language	Corpus (# annotations)	AMBIG_trainSI		AMBIG_fullSI	
		type	token	type	token
English	SemCor (88334)	1.97	7.91	3.24	10.94
	SenseEval2 (517)	4.90	6.7	7.58	10.28
	SemEval 2007 (296)	5.15	6.89	7.78	10.17
	SenseEval 2015 (251)	5.69	6.25	8.48	9.16
French	Wiktionary (55206)	1.66	5.49	1.74	5.68
	FSE (3199)	6.02	6.74	6.15	6.91

Table 3: Ambiguity rates for verbs, in the English usual training set (SemCor) and usual evaluation sets, and in the French training set (Wiktionary) and evaluation set (FSE). **AMBIG_trainSI** corresponds to using for the number of senses the sense inventory in the corresponding training corpus, whereas **AMBIG_fullSI** corresponds to using the full sense inventory.

5.1 Comparison of the sense distribution in training examples

We study here the distribution of the annotated senses in training data. When looking at the number of training examples per sense, we obtain an average of 9.6 and a mean absolute deviation of 11.9 for SemCor, whereas the average is only 2.0 for FR-Wiktionary, and the mean absolute deviation is 0.9. It is clear that using wiktionary examples will lack the genre-dependent but nonetheless very informative information of sense frequency in corpus.

5.2 Evaluation of the task difficulty: comparison of ambiguity rates

We now turn to comparing the difficulty of the WSD task, when tested on English SenseEval datasets versus on FrenchSemEval. Note that performance of WSD systems cannot be used for that purpose, given that it is not comparable across languages and datasets. For a corpus consisting in a sequence of tokens $t_1 \dots t_N$, we rather compute the average ambiguity rate that a WSD system has to face, in two settings:

- **token_AMBIG_fullSI**: the ambiguity rate per token, using the full sense inventory:

$$\frac{1}{N}\sum_{i=1}^{N} \text{n\_senses}(t_i)$$

- **token_AMBIG_trainSI**: the ambiguity rate per token, using the sense inventory found in the training corpus

$$\frac{1}{N}\sum_{i=1}^{N} \text{attested\_n\_senses}(t_i)$$

For further information, although not directly measuring corpus WSD difficulty, we also provide the ambiguity rate per verb type, both using the full inventory or that attested in the training set (shown in the "type" columns in Table 3).

We report these metrics in Table 3. When studying the difference between the "fullSI" versus "trainSI" modes, namely when using the full sense inventory versus that found in the training set, we have a different trend for the English corpora (containing natural text) and the French ones: for SemCor and the English evaluation sets, there is a drop of ambiguity in trainSI mode. This illustrates the usual difficulty to cover rare senses in a corpus of natural text. Note though that for the French corpora, based on the wiktionary inventory, there is almost no difference between the two modes of computation, illustrating that almost all senses have examples in wiktionary.

When comparing, for each language, the figures for the training corpora (SemCor and Wiktionary examples) and for the evaluation datasets, it can be noted that the average ambiguity per token is similar

for training and evaluation datasets, but the average ambiguity per type is much smaller for the training corpora (3.24 for SemCor, and 1.74 for Wiktionary). This is because the lexicon covered in the training corpora is much larger, and contains many more monosemic verbs. .

As far as training corpora are concerned, it can be seen that the overall average ambiguity is higher for SemCor than for Wiktionary (e.g. in fullSI mode, 10.94 per token ambiguity for SemCor, versus 5.68 for Wiktionary). It shows that the sense inventory for Wiktionary is slightly less ambiguous than Wordnet's (both for the senses found in SemCor, and overall).

6 Experiments on supervised WSD

To investigate the suitability of using Wiktionary for supervised WSD on French verbs, we evaluated state-of-the-art supervised WSD systems on FrenchSemEval, using the examples of Wiktionary's senses as training data. As for the representation of the instances we used two different models that we describe below. We then applied a supervised disambiguation method to evaluate the performance of the models. We first describe the models we used to obtain vector representations of the instances and the disambiguation algorithm we used for evaluation. Then we propose several experiments based on FSE and finally we evaluate the models using Wiktionary as input for disambiguation.

6.1 Models for context representations

AWE We implemented a simple model that we use as baseline. We first train a word2vec (Mikolov et al., 2013) model on fr-Wikipedia[11] to obtain non contextual word vectors. We then represent the context of an occurrence by averaging the vectors of the words found in its context window, which we defined as the 5 words on the left and 5 words on the right of the target word. This is a common model often referred in the literature as averaged-word-embeddings (AWE).

C2V Context2vec (Melamud et al., 2016) is a recurrent neural model that learns a function mapping the context around a target word to a vectorial representation. The context2vec model represents the context using a bi-directional recurrent neural network (Hochreiter and Schmidhuber, 1997) that allow us to take the context of the sentence into account, thus contrasting with AWE. All codes and implementations are available publicly so we only adapted it and trained the model on the whole French Wikipedia. We then applied the learnt model to obtain vectorial representations of our target verb occurrences.

6.2 Supervised disambiguation algorithm

We replicated the supervised WSD method used in (Yuan et al., 2016): a sense representation is computed from annotated data by averaging the context vector representation of its instances, in our case the Wiktionary examples or instances from FSE. Then each test instance is sense tagged with the sense whose representation is the closest, based on cosine similarity.

6.3 Protocol

Wiktionary experiment We did a first experiment simply using the examples of the senses in the Wiktionary sense inventory as training data and then we performed disambiguation on FSE.

In domain experiments In order to better identify the potential error sources, we also performed experiments with "in-domain" training instances, namely instances directly taken from FSE. To evaluate the impact of the number of training examples per sense, a property that is quite different for a lexicographic training set as opposed to a corpus-based training set, we performed experiments on different sets, using N_{max} a varying maximum number of examples per sense. More precisely, for each verb we

[11] We used the fr-Wikipedia dump of 10-20-2017

selected respectively 1, 2, 5 and 10 maximum training examples per sense from the dataset and evaluated the disambiguation on the remaining examples.

6.4 Results and Analysis

The results of our experiments on Wiktionary are presented in table 4. Although both automatic systems perform better than the Most Frequent Sense baseline (MFS), using only the literary Wiktionary examples to build a classifier for newspaper and Wikipedia test instances remains a rather adversarial setup[12]. We thus investigated two potential ways to leverage the hardness of this initial setup: domain adaptation and the amount of training examples.

Models	score
MFS	0.30
AWE	0.40
C2V	**0.43**

Table 4: WSD accuracies when training on Wiktionary examples, and testing on FSE.

To study the effect of the amount of training instances, since Wiktionary is limited in terms of number of examples per sense, we switched to using FSE both for training and testing. We used a variable number of maximum training examples per sense from $N_{max} = 1$ to $N_{max} = 10$ and we used the remaining examples as test set[13]. The results of these experiments are summarized in Table 5 and illustrated in Figure 1.

Let us observe first the impact the amount of training data. The mean number of examples in Wiktionary for the verbs occurring in FSE is $N_{avg} = 3.1$ and the results show that all classifiers dramatically improve when the available training examples per verb grows up to $N_{max} = 10$. This means that if we were able to expand with absolute certainty the small amount of examples in Wiktionary, we could get a much higher disambiguation performance.

Second, using the same setup we can compare the behaviour of the classifiers when predicting out of domain (Table 4) with in domain predictions (Table 5). Recall that Wiktionary examples are often long literary sentences whereas the test instances are sampled from newspaper or Wikipedia. Again as Wiktionary has $N_{avg} \approx 3$ training examples per sense we can see that the domain adaptation effect is worth roughly 20 points in accuracy.

Models	N_{max} = 1	N_{max} = 2	N_{max} = 3	N_{max} = 5	N_{max} = 10
MFS	0.32	0.38	0.45	0.52	0.70
AWE	0.44	0.53	0.58	0.64	0.70
C2V	0.5	0.57	0.62	0.68	**0.74**
Mean number of training ex. per sense	1	1.81	2.61	3.86	6.30
Mean size training data per verb	3.83	6.95	9.81	14.8	24.15
Mean size test data per verb	44.63	41.51	38.65	33.66	24.31

Table 5: Training on FSE examples, with varying maximum number of examples per sense (N_{max}). Top: WSD accuracies. Bottom: training / test sets statistics.

Third, we can observe that the 3 classifiers (MFS, AWE,C2V) do behave consistently in the different configurations. To understand why MFS is a rather weak predictor in our setup, we have to recall that

[12] As a comparison, results for supervised WSD for English verbs are around 0.55 in the benchmark of Raganato et al. (2017).

[13] We say that we use N_{max} as a maximum number of training examples because some senses may have only $K < N_{max}$ annotated instances in the whole data set.

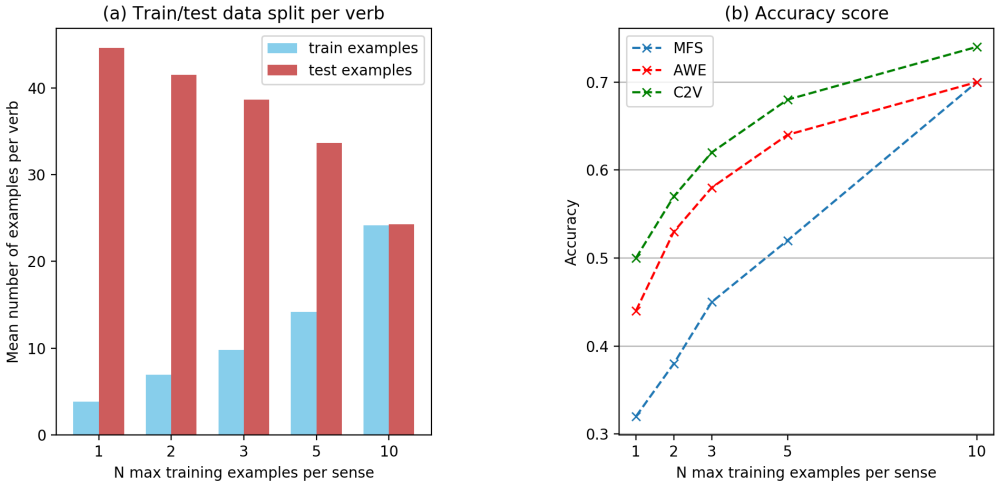

Figure 1: Illustration of results reported in table 5.

contrary to Semcor and Senseval, FSE is built following a lexicographic perspective: the sentences are sampled in a non natural way (e.g. monosemic words are excluded). We observe (table 5) that in these conditions the MFS is a weaker (although still strong) baseline and that standard sequential neural models are able to outperform it, especially when there are few training examples. Among the neural models C2V performs consistently better than AWE but we believe that these models or some extensions of these models to be designed in the future might well still show significant improvements.

Lastly, we can observe that among all models, C2V performs best on every test setup. It supports (Melamud et al., 2016)'s results, especially regarding the fact that Context2vec succeeds better in capturing context information than the common averaging of word vector representations.

7 Conclusions and future work

Word Sense Disambiguation is a task rarely seen for languages other than English. One obvious reason to explain that is the lack of costly sense annotated resources for those languages. In this paper we provide some elements seeking to set up a methodology to perform word sense disambiguation for other languages than English, such as French, without requiring the cost of annotating sense disambiguated corpora.

For this purpose we considered using Eurosense and Wiktionary as training data for Verb Sense Disambiguation. As our first experiments with Eurosense turned out to be inconclusive, we then turned our attention to Wiktionary. We studied how to use it as a resource for Word Sense Disambiguation and we develop FrenchSemEval, a new French WSD evaluation dataset, thanks to which we were able to extract preliminary evaluation results. Our current results showed that the Wiktionary sense inventory has an appropriate granularity for a good quality sense annotation, and that training on Wiktionary examples only leads to encouraging WSD results for verbs. But we could also quantify the gain in performance that could be obtained by adding a moderate number of seed instances. Hence automating the selection and annotation of additional instances might pay off to improve verb sense disambiguation.

References

Abeillé, A. and N. Barrier (2004). Enriching a french treebank. In *Proceedings of LREC 2004, Lisbon, Portugal*.

Bovi, C. D., J. Camacho-Collados, A. Raganato, and R. Navigli (2017). Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 2*, 594–600.

Camacho-Collados, J., M. T. Pilehvar, and R. Navigli (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence 240*, 36–64.

Candito, M. and D. Seddah (2012). Le corpus sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.

Coavoux, M. and B. Crabbé (2017, April). Incremental discontinuous phrase structure parsing with the gap transition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, pp. 1259–1270. Association for Computational Linguistics.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin 70*(4), 213.

de Castilho, R. E., E. Mujdricza-Maydt, S. M. Yimam, S. Hartmann, I. Gurevych, A. Frank, and C. Biemann (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pp. 76–84.

Djemaa, M., M. Candito, P. Muller, and L. Vieu (2016, May). Corpus Annotation within the French FrameNet: a Domain-by-domain Methodology (regular paper). In Calzolari, Choukri, Declerck, Goggi, and Grobelnik (Eds.), *LREC 2016), Portoroz, Slovenia*, pp. 3794–3801.

Henrich, V., E. Hinrichs, and T. Vodolazova (2011). Semi-automatic extension of germanet with sense definitions from wiktionary. In *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, pp. 126–130.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Kilgarrif, A. (1998). Senseval: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the first international conference on language resources and evaluation (LREC 1998), Granada, Spain*, pp. 581–588.

Lefever, E. and V. Hoste (2010). Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 15–20. Association for Computational Linguistics.

Liebeck, M. and S. Conrad (2015, July). Iwnlp: Inverse wiktionary for natural language processing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, pp. 414–418. Association for Computational Linguistics.

Melamud, O., J. Goldberger, and I. Dagan (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 51–61.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

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

Miller, G. A., C. Leacock, R. Tengi, and R. T. Bunker (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pp. 303–308. Association for Computational Linguistics.

Miller, T. and I. Gurevych (2014). Wordnet—wikipedia—wiktionary: Construction of a three-way alignment. In *LREC*, pp. 2094–2100.

Moro, A. and R. Navigli (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 288–297.

Moro, A., A. Raganato, and R. Navigli (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics 2*, 231–244.

Navarro, E., F. Sajous, B. Gaume, L. Prévot, H. ShuKai, K. Tzu-Yi, P. Magistry, and H. Chu-Ren (2009). Wiktionary and nlp: Improving synonymy networks. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27. Association for Computational Linguistics.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR) 41*(2), 10.

Navigli, R., D. Jurgens, and D. Vannella (2013). Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Volume 2, pp. 222–231.

Navigli, R. and S. P. Ponzetto (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence 193*, 217–250.

Nivre, J., J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Raganato, A., J. Camacho-Collados, and R. Navigli (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Volume 1, pp. 99–110.

Sérasset, G. (2012). Dbnary: Wiktionary as a lmf based multilingual rdf network. In *Language Resources and Evaluation Conference, LREC 2012*.

Weaver, W. (1955). Translation. *Machine translation of languages 14*, 15–23.

Yimam, S. M., C. Biemann, R. E. de Castilho, and I. Gurevych (2014). Automatic annotation suggestions and custom annotation layers in webanno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 91–96.

Yuan, D., J. Richardson, R. Doherty, C. Evans, and E. Altendorf (2016). Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.

Zhong, Z. and H. T. Ng (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pp. 78–83. Association for Computational Linguistics.