Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation

Roberto Navigli and **Simone Paolo Ponzetto**
Dipartimento di Informatica
Sapienza Università di Roma
{navigli,ponzetto}@di.uniroma1.it

Abstract

We present a multilingual joint approach to Word Sense Disambiguation (WSD). Our method exploits BabelNet, a very large multilingual knowledge base, to perform graph-based WSD across different languages, and brings together empirical evidence from these languages using ensemble methods. The results show that, thanks to complementing wide-coverage multilingual lexical knowledge with robust graph-based algorithms and combination methods, we are able to achieve the state of the art in both monolingual and multilingual WSD settings.

1 Introduction

Nowadays the textual information needed by a user accessing websites for content such as news reports, commentaries and encyclopedic knowledge is provided in an increasingly wide range of languages. For example, even though English is still the majority language of the Web, the Chinese and Spanish languages are moving fast to capture their "juicy share", and more languages are about to join them in the near future. This language explosion clearly forces researchers to focus on the challenging problem of being able to analyze and understand text written in any language. However, it also opens up novel perspectives for multilingual Natural Language Processing (NLP) such as, for instance, the development of approaches aimed at "joining forces" and taking advantage of the lexico-semantic knowledge provided in the different languages to improve text understanding. These two aspects are strongly intertwined: on the one hand, enabling language-independent text understanding would allow for the harvesting of more knowledge in arbitrary languages, while, on the other hand, bringing together the lexical and semantic information available in different languages would improve the quality of text understanding in arbitrary languages.

However, these two goals have hitherto never been achieved, as is attested to by the fact that research in a core language understanding task such as Word Sense Disambiguation (Navigli, 2009, WSD) has always been focused mostly on English. Historically, English became established as the language used and understood by the scientific community and, consequently, most resources were developed for it, including large-scale computational lexicons like WordNet (Fellbaum, 1998) and sense-tagged corpora like SemCor (Miller et al., 1993). As a result WSD in other languages was hindered by a lack of resources, which in turn led to poor results or low involvement on the part of the research community (Magnini et al., 2004; Màrquez et al., 2004; Orhan et al., 2007; Okumura et al., 2010). Nonetheless, already in the 1990s it had been remarked that WSD could be improved by means of multilingual information: a recurring idea proposed by several researchers was that plausible translations of a word in context would restrict its possible senses to a manageable subset of meanings (Dagan et al., 1991; Gale et al., 1992; Resnik and Yarowsky, 1999). While the lack of resources at that time hampered the development of effective multilingual approaches to WSD, recently this idea has been revamped with the organization of SemEval tasks dealing with cross-lingual WSD (Lefever and Hoste, 2010) and cross-lingual lexical substitution (Mihalcea et al., 2010). At the same time, new re-

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1399–1410, Jeju Island, Korea, 12–14 July 2012.

search on the topic has been done, including the use of statistical translations of sentences into many languages as features for supervised models (Banea and Mihalcea, 2011; Lefever et al., 2011), and the projection of monolingual knowledge onto another language (Khapra et al., 2011).

Yet the above two goals, i.e., disambiguating in an arbitrary language and using lexical and semantic knowledge from many languages in a joint way to improve the WSD task, have not hitherto been attained. In this paper, we address both objectives and propose a graph-based approach to multilingual joint Word Sense Disambiguation. Our proposal brings together the lexical knowledge from different languages by exploiting empirical evidence for disambiguation from each of them, and then combining this information in a synergistic way: each language provides a piece of sense evidence for the meaning of a target word in context, and subsequent integration of these various pieces enables them to (soft) constrain each other. The results show that this way we are able to improve over previous, high-performing graph-based methods in both a monolingual and multilingual setting, thus showing for the first time the beneficial effects of exploiting multilingual knowledge in a joint fashion.

2 Related Work

Parallel corpora have been used in the literature for the automatic creation of a sense-tagged dataset for supervised WSD in different languages (Gale et al., 1992; Chan and Ng, 2005; Zhong and Ng, 2009). Other approaches include the use of a coherence index for identifying the tendency to lexicalize senses differently across languages (Ide, 2000) and the clustering of source words which translate into the same target word, then used to perform WSD using a similarity measure (Diab, 2003). A historical approach (Brown et al., 1991) uses bilingual corpora to perform unsupervised word alignment and determine the most appropriate translation for a target word from a set of contextual features.

All the above approaches to multilingual or cross-lingual WSD rely on bilingual corpora, including those which exploit existing multilingual WordNet-like resources (Ide et al., 2002), or use automatically induced multilingual co-occurrence graphs (Silberer and Ponzetto, 2010). However, this requirement is often very hard to satisfy, especially if we need wide coverage. To overcome this limitation, in this work we make use of BabelNet (Navigli and Ponzetto, 2010), a very large multilingual lexical knowledge base. This resource – complementary in nature to other recent efforts presented by de Melo and Weikum (2010), Nastase et al. (2010) and Meyer and Gurevych (2012), *inter alia* – provides a truly multilingual semantic network by combining Wikipedia's multilinguality with the output of a state-of-the-art machine translation system to achieve high coverage for all languages. The key insight here is that Word Sense Disambiguation and Machine Translation (MT) are highly intertwined tasks, as previously shown by Carpuat and Wu (2007) and Chan et al. (2007), who successfully used sense information to boost state-of-the-art statistical MT. In this work we focus instead on the benefits of using multilingual information for WSD by exploiting the structure of a multilingual semantic network.

3 Multilingual Joint WSD

We present our methodology for multilingual WSD: we first introduce BabelNet, the resource used in our work (Section 3.1) and then present our algorithm for multilingual joint WSD (Section 3.2), including its main components, namely graph-based WSD, ensemble methods and translation weighting (sections 3.3, 3.4 and 3.5).

3.1 BabelNet

BabelNet (Navigli and Ponzetto, 2010) follows the structure of a traditional lexical knowledge base and, accordingly, consists of a labeled directed graph whose nodes represent concepts and named entities, and whose edges express semantic relations between them. Concepts and relations are harvested from the largest available semantic lexicon of English, i.e., WordNet, and a wide-coverage collaboratively-edited encyclopedia, i.e., Wikipedia[1], thus making BabelNet a multilingual 'encyclopedic dictionary' which combines lexicographic information with encyclopedic knowledge on the basis of an unsupervised mapping framework. In addition to a core

[1] http://www.wikipedia.org. In the following, we refer to Wikipedia pages and senses using SMALL CAPS.

semantic network, BabelNet provides a multilingual lexical dimension. Each of its nodes, called *Babel synsets*, contains a set of lexicalizations of the concept for different languages, e.g., { $\mathsf{bank}_n^{\text{EN}}$, $\mathsf{Bank}_n^{\text{DE}}$, $\mathsf{banca}_n^{\text{IT}}$, ..., $\mathsf{banco}_n^{\text{ES}}$ }[2]. Multilingual lexicalizations for all concepts are collected from Wikipedia's inter-language links (e.g., the English Wikipedia page BANK links to the Italian BANCA), as well as by acquiring missing translations by means of a statistical machine translation system applied to sense-tagged data from SemCor and Wikipedia itself – for instance, most occurrences of bank_n^1 in SemCor[3] are translated into German and Italian as Ufer and riva, respectively. As a result of combining human-edited translations from Wikipedia and automatically generated ones from sense-labeled data, BabelNet is able to achieve wide coverage for all its languages (Catalan, English, French, German, Italian and Spanish): accordingly, we chose it to perform graph-based WSD in a multilingual setting since it is specifically focused on lexical knowledge. In addition, BabelNet is available for any language required to perform standard SemEval cross-lingual disambiguation tasks (e.g., Spanish, in order to perform cross-lingual lexical substitution). Since previous work in knowledge-based WSD shows the benefits of using rich lexical resources (Navigli and Lapata, 2010; Ponzetto and Navigli, 2010), BabelNet is a suitable choice for performing graph-based multilingual WSD.

3.2 Exploiting multilingual information in a knowledge-based WSD framework

We present a multilingual approach to WSD which exploits three main factors:

i) the fact that translations of a target word provide complementary information on the range of its candidate senses in context;

ii) the wide-coverage, multilingual lexical knowledge stored in BabelNet;

iii) the support for disambiguation from different languages in a synergistic, unified way.

[2]BabelNet senses are referred to with w_p^l, namely the sense of a word w in a language l with part of speech p.

[3]We denote WordNet senses with w_p^i, namely the i-th sense of a word w with part of speech p.

Algorithm 1 Multilingual joint WSD

Input: a word sequence $\sigma = (w_1, \ldots, w_n)$
a target word $w \in \sigma$
BabelNet BN
an ensemble method M
Output: a distribution of scores for the senses of w

(▷ indicates a comment)

1: $S \leftarrow \mathit{Synsets}_{BN}(w)$
2: $T \leftarrow \{w\}$
3: **for each** $s \in S$
4: $\quad T \leftarrow T \cup getTranslations(s)$
5: $ctx \leftarrow \sigma - \{w\}$
6: ▷ $LScore := \{lScore_{i,j}\}_{i=1,\ldots,|T|,\ j=1,\ldots,|S|}$
7: **for each** $t_i \in T$
8: $\quad \sigma' \leftarrow \{t_i\} \cup ctx$
9: $\quad \triangleright\ G_i := (V_i, E_i)$
10: $\quad G_i \leftarrow createGraph(\sigma', BN)$
11: $\quad$ **for each** $s_j \in S \cap V_i$
12: $\quad\quad lScore_{i,j} \leftarrow score(G_i, s_j)$
13: ▷ $Score := (score_1, \ldots, score_{|S|})$
14: $Score \leftarrow M(LScore)$
15: **return** $Score$

We call this approach *multilingual joint WSD*, since disambiguation is performed by exploiting different languages *together at the same time*. To this end, we first perform graph-based WSD using the target word in context as input, and then combine sense evidence from its translations using an ensemble method. The key idea of our joint approach is that sense evidence from different translations provides complementary views for the senses of a target word in context. Therefore, combining such evidence should produce more accurate sense predictions. We view WSD as a sense ranking problem. Given a word sequence $\sigma = (w_1, \ldots, w_n)$, we disambiguate a target word $w \in \sigma$ by scoring each of its senses and selecting the highest-ranking one:

$$\hat{s} = \underset{s \in \mathit{Synsets}_{BN}(w)}{\arg\max} \; score(s)\ , \qquad (1)$$

where $\mathit{Synsets}_{BN}(w)$ is the set of Babel synsets containing the different senses for w.[4] We score these

[4]Babel synsets unambiguously identify different senses of the target word, e.g., { $\mathsf{bank}_n^{\text{EN}}$, $\mathsf{Bank}_n^{\text{DE}}$, $\mathsf{banco}_n^{\text{ES}}$ ..., $\mathsf{banca}_n^{\text{IT}}$ } corresponds to the 'financial institute' sense of $\mathsf{bank}_n^{\text{EN}}$ (i.e., bank_n^2 in WordNet).

synsets using Algorithm 1, which we illustrate in the following by means of the example sentence '*bank bonuses are paid in stock*', where we focus on $\mathsf{bank}_n^{\text{EN}}$ as the target word and { $\mathsf{bonus}_n^{\text{EN}}$, $\mathsf{pay}_v^{\text{EN}}$, $\mathsf{stock}_n^{\text{EN}}$ } as its context. The following steps are performed:

Initialization. We start by gathering the data required for disambiguation (lines 1–5). First, we collect in line 1 the set S of Babel synsets corresponding to the different senses of the target word w – namely, the synsets containing the 'financial institution', 'money container', 'building' senses of $\mathsf{bank}_n^{\text{EN}}$, among others. Next, we obtain the multilingual lexicalizations of the target word: to this end, we first include in T the word w itself (line 2), and then iterate through each synset $s \in S$ to collect the translations of each of its senses in the languages of interest (lines 3–4). For instance, given the English word $\mathsf{bank}_n^{\text{EN}}$, we collect its sense-specific German, Italian and Spanish translations and obtain a set of multilingual terms $T = \{ \mathsf{bank}_n^{\text{EN}}, \ldots, \mathsf{Bank}_n^{\text{DE}}, \mathsf{Sparbüchse}_n^{\text{DE}}, \mathsf{Bankgebäude}_n^{\text{DE}}, \ldots, \mathsf{banca}_n^{\text{IT}}, \mathsf{salvadanaio}_n^{\text{IT}}, \ldots, \mathsf{banco}_n^{\text{ES}}, \mathsf{hucha}_n^{\text{ES}} \}$. Finally, we create a disambiguation context ctx by taking the word sequence σ and removing w from it (line 5, as a result, e.g., $ctx = \{ \mathsf{bonus}_n^{\text{EN}}, \mathsf{pay}_v^{\text{EN}}, \mathsf{stock}_n^{\text{EN}} \}$).

Collecting sense distributions. In the next phase (lines 6–12), we collect a scoring distribution over the different synsets S of w for each term $t_i \in T$. Each distribution quantifies the empirical support for the different senses of the target word, obtained using t_i and the context ctx: we store this information in a $|T| \times |S|$ matrix $LScore$, where each cell $lScore_{i,j}$ quantifies the support for synset $s_j \in S$, computed using the term in $t_i \in T$. We calculate the scores as follows:

- We select at each step an element t_i from T (line 7), for instance $\mathsf{banco}_n^{\text{ES}}$.

- Next, we create a multilingual context σ' by combining t_i with the words in ctx (line 8, e.g., we set $\sigma' = \{ \mathsf{banco}_n^{\text{ES}}, \mathsf{bonus}_n^{\text{EN}}, \mathsf{pay}_v^{\text{EN}}, \mathsf{stock}_n^{\text{EN}} \}$.

- We use σ' to build a graph $G_i = (V_i, E_i)$ by computing the paths in BabelNet which connect the synsets of t_i with those of the other words in σ' (line 10, see Section 3.3 for details on the $createGraph$ function). Note that by selecting at each step a different element from T we create a new graph where different sets of Babel synsets get activated by the context words in ctx. In our example, Figures 1(a)–(c) show the graphs obtained by setting at different steps t_i to $\mathsf{bank}_n^{\text{EN}}$, $\mathsf{banco}_n^{\text{ES}}$ and $\mathsf{Bank}_n^{\text{DE}}$, respectively (we show excerpts by using only $\mathsf{stock}_n^{\text{EN}}$ as context word for ease of readability).

- Finally, we compute the support from term t_i for each synset $s_j \in S$ of the target word by applying a graph connectivity measure to G_i and store the result in $lScore_{i,j}$ (lines 11–12). For instance, using degree as graph measure, we can compute the following scores from the graph in Figure 1(b):

	bank_n^2	bank_n^8	bank_n^9
$\mathsf{banco}_n^{\text{ES}}$	2	0	1

By repeating the process for each term in T (lines 7–12) we compute all values in the matrix $LScore$. For instance, given $T = \{\mathsf{bank}_n^{\text{EN}}, \mathsf{banco}_n^{\text{ES}}, \mathsf{Bank}_n^{\text{DE}}\}$, we create the set of graphs in Figures 1(a)–(c), and compute from each of them the following scores (again, using degree as scoring measure):

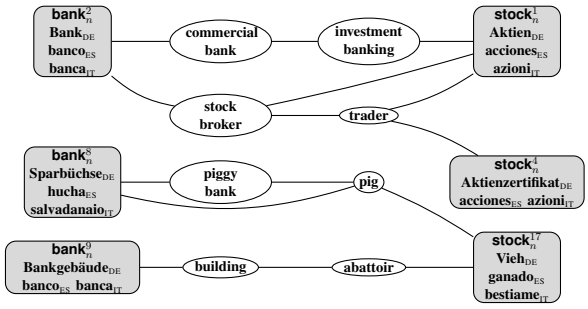
$LScore =$

	bank_n^2	bank_n^8	bank_n^9
$\mathsf{bank}_n^{\text{EN}}$	2	2	1
$\mathsf{banco}_n^{\text{ES}}$	2	0	1
$\mathsf{Bank}_n^{\text{DE}}$	2	0	0

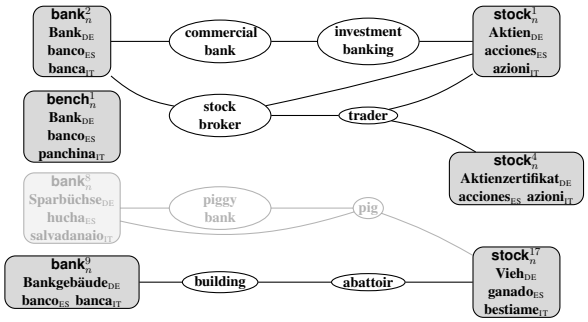
Combining sense distributions. In the last step (line 14) we aggregate the scores associated with each term of T using an ensemble method M (see Section 3.4 for details). For instance, M could simply consist of summing the scores associated with each sense over all distributions and thus return a score of 6, 2, and 2 for bank_n^2, bank_n^8 and bank_n^9, respectively. As a result of the execution of Algorithm 1, the combined scoring distribution is returned (line 15). This sense distribution in turn can be used to select the best sense using Equation 1.

The main hunch behind our approach is that using information from different languages improves disambiguation performance, as in the example of Figure 1 where more accurate disambiguation is performed by combining scores computed from translations in different languages, as opposed to using

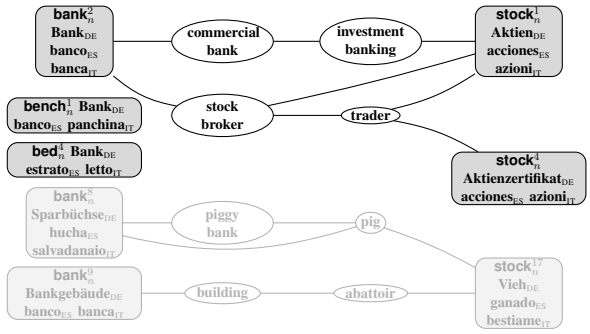
(a) Disambiguation graph using the target word $\text{bank}_n^{\text{EN}}$.



(b) Disambiguation graph using $\text{banco}_n^{\text{ES}}$ as translation.



(c) Disambiguation graph using $\text{Bank}_n^{\text{DE}}$ as translation.



(d) List of corresponding WordNet senses and their glosses

bank_n^2	financial institution that accepts deposits and channels the money into lending activities
bank_n^8	a container (usually with a slot in the top) for keeping money at home
bank_n^9	a building in which the business of banking transacted
stock_n^1	the capital raised by a corporation through the issue of shares entitling holders to an ownership interest
stock_n^4	a certificate documenting the shareholder's ownership in the corporation
stock_n^{17}	any animals kept for use or profit

Figure 1: Multilingual graph construction for the input sentence '*bank bonuses are paid in stock*'. We show excerpts using only $\text{stock}_n^{\text{EN}}$ as context word for ease of readability.

monolingual sense evidence only. Figure 1(a) shows the graph created to disambiguate the English target word $\text{bank}_n^{\text{EN}}$ in our example sentence. In the graph, some of the possible senses of this word are activated, including the correct one (bank_n^2) but also related, yet incorrect ones such as bank_n^8 and bank_n^9. Figure 1(b) and 1(c) show instead the graphs obtained from replacing the target word with its Spanish and German translations, respectively. In these graphs, different subsets of the senses of $\text{bank}_n^{\text{EN}}$ are activated, together with others pertaining to the translations only (e.g., the meaning of $\text{banco}_n^{\text{ES}}$ corresponding to the English bench_n^1). However, the sense that is consistently activated across all graphs is the correct one – i.e., $\text{bank}_n^{\text{EN}}$ as financial institution – which is in fact the sense selected by our multilingual approach by means of combining the scoring distributions from all these graphs.

3.3 Graph-based WSD

We use graph-based algorithms to exploit multilingual knowledge from BabelNet for WSD. These are a natural choice for our approach, since BabelNet is a semantic network, and such algorithms have been shown to achieve high performance across domains (Agirre et al., 2009; Navigli et al., 2011), as well as to compete with supervised methods on a variety of lexical disambiguation tasks (Ponzetto and Navigli, 2010). To this end, we use the method of Navigli and Lapata (2010) and construct a directed graph $G = (V, E)$ for an input word sequence $\sigma = (w_1, \ldots, w_n)$[5] using the lexical and semantic relations found in BabelNet. The result of this procedure is a subgraph of BabelNet containing (1) the senses of the words in context, (2) all edges and intermediate senses found in BabelNet along all paths that connect them. Given G, a target word $w \in \sigma$ and its set of senses in BabelNet $S \subseteq V$, we compute a score distribution $(score_1, \ldots, score_{|S|})$ over S, where $score_j$ refers to the confidence score for the j-th sense of w, e.g. bank_n^2, based on some connectivity measure applied to G. In this paper, we specifically focus on two such measures.

[5] In our experiments we always take σ to be a single sentence, thus disambiguating on a sentence-by-sentence basis.

Degree Centrality (Degree): The first measure ranks the senses of a given word in the graph based on the number of their incident edges, namely:

$$score_j = |\{\{s_j, v\} \in E : v \in V\}| \ .$$

This standard connectivity measure weights a sense as more appropriate if it has a higher degree. We chose context-based Degree since, albeit simple, it had previously been shown to yield a highly competitive performance on various WSD tasks (Navigli and Lapata, 2010; Ponzetto and Navigli, 2010).

Inverse path length sum (PLength): We then developed a graph connectivity measure which scores each sense by summing over the inverse length of all paths which connect it to other senses in the graph:

$$score_j = \sum_{p \in paths(s_j)} \frac{1}{e^{length(p)-1}} \ ,$$

where $paths(s_j)$ is the set of simple paths connecting s_j to the senses of other context words, $length(p)$ is the number of edges in the path p and each path is scored with the exponential inverse decay of the path length. This measure overcomes the locality of Degree by aggregating over all paths between a sense of the target word and those of the context words, thus being able to capture the richness of the BabelNet subgraph and the semantic density of the underlying knowledge base.

3.4 Ensemble methods for multilingual WSD

At the core of our algorithm lies the combination of the scores generated using the different translations of the target word w. For this purpose, we apply so-called *ensemble* methods, which have been shown to improve the performance of both supervised (Florian et al., 2002) and unsupervised WSD systems (Brody et al., 2006). Given $|T|$ lexicalizations and $|S|$ senses for w, the input to the combination component consists of a $|T| \times |S|$ matrix $LScore$, where each cell $lScore_{i,j}$ quantifies the empirical support for sense s_j from a term $t_i \in T$ (see Section 3.2 for an example). The ensemble method computes from this translation-sense matrix a combined scoring, expressing the *joint* confidence across terms in different languages over the set of senses S. In this work, we use the 'Probability Mixture' (PMixture) method proposed by Brody et al. (2006), which they show to be the best performing for WSD. This method takes the scores associated with each term, normalizes and combines them by summing across distributions. Formally, it computes the score for the j-th sense of w as follows:

$$score_j = \sum_{i=1}^{|T|} p(s_{i,j}), \ p(s_{i,j}) = \frac{lScore_{i,j}}{\sum_{s=1}^{|S|} lScore_{i,s}} \ .$$

For instance, using the (normalized) sense distributions from our example, the ensemble distribution will be the following:

	bank_n^2	bank_n^8	bank_n^9
$\mathsf{bank}_n^{\text{EN}}$	0.40	0.40	0.20
$\mathsf{banco}_n^{\text{ES}}$	0.67	0.00	0.33
$\mathsf{Bank}_n^{\text{DE}}$	1.00	0.00	0.00
PMixture	2.07	0.40	0.53

3.5 Weighting multilingual sense distribution

Computing a sense distribution for each translation using the same graph connectivity measure assumes that all translations are equal. However, a *leitmotif* of multilingual WSD research is that translations restrict the set of candidate senses of the target word in the source language. In our example of Figure 1, for instance, $\mathsf{Bank}_n^{\text{DE}}$ provides structural support only for the financial sense of English bank, since this is the only sense it covers. Within our framework this can potentially lead to skewed sense distributions when only some senses of the target word have a translation. In such cases, in fact, scores tend to be concentrated mostly on the senses covered by the translations, with the result that sense evidence for uncovered English senses is disregarded. In order to cope with this issue, we weight the elements of each sense distribution $lScore_i$ for the i-th translation $t_i \in T$ by a factor of $1+\log_2 cov(t_i, w)$, where $cov(t_i, w)$ is the number of Babel synsets where t_i co-occurs with the target word w – i.e., the number of senses of w that it covers (we use the log function to dampen the effect of high coverage values). This is to say, in order to level off the effects of unbalanced sense coverage we assume that, all things being equal, the more senses a translation covers, the stronger the disambiguation evidence it provides in context for specific senses. As a result, the contributions of each translation are weighted differently

and we are thus able to dampen the effects of a highly skewed distribution like, for instance, that of $\mathsf{Bank}_n^{\mathrm{DE}}$:

	bank_n^2	bank_n^8	bank_n^9
$\mathsf{bank}_n^{\mathrm{EN}}$	1.72	1.72	0.86
$\mathsf{banco}_n^{\mathrm{ES}}$	1.34	0.00	0.66
$\mathsf{Bank}_n^{\mathrm{DE}}$	1.00	0.00	0.00
Weighted PMixture	4.04	1.70	1.52

4 Experiments

We evaluate our approach in two different settings, namely a monolingual all-words WSD task in Section 4.1, as well as two different cross-lingual disambiguation gold standards in Section 4.2.

4.1 Monolingual WSD

Experimental setting. We first evaluate the performance of multilingual joint WSD on a standard monolingual dataset, namely the SemEval-2010 domain WSD task 17 (Agirre et al., 2010), since it provides the latest dataset for fine-grained WSD in English. We opt for an English all-words task for two main reasons: first, it is a well-established and widely-participated task in the WSD community – thus ensuring a comparison of our method with a wide range of state-of-the-art approaches, including other graph-based techniques (e.g., Personalized PageRank), as well as weakly-supervised and supervised approaches (see Agirre et al. (2010) for details on the participating systems); second, we want to assess whether a multilingual approach benefits lexical disambiguation in *all* settings, namely *even* in a standard monolingual one. We use in our experiments the dataset's nouns-only subset (1032 instances), since BabelNet currently contains multilingual lexicalizations for nouns only (and thus no multilingual strategy can be applied to other parts of speech). We perform graph-based WSD with BabelNet in two different configurations, namely a monolingual and multilingual setting. The multilingual system performs WSD by means of the full joint multilingual approach described in Algorithm 1. The monolingual approach, instead, simply uses the English input sentence for disambiguation – that is, we skip lines 3–4 of Algorithm 1. Knowledge-based systems typically suffer from a low recall – i.e., they cannot provide an answer if no information can be found with senses of the context words. To overcome this issue, in both settings we use a type-based fallback strategy which assigns to the target word the sense which has been most frequently assigned by the system to other instances of the word in the dataset.

	Algorithm	P	R	F_1
Monolingual graph	Degree	50.6	45.2	47.7
	PLength	51.0	47.3	49.1
Multilingual ensemble	Degree†	53.9	48.6	51.1
	PLength†	**54.3**	50.2	**52.2**
	SemCor MFS	51.9	**51.2**	51.5
	Random	25.3	25.3	25.3

Table 1: Performance on SemEval-2010 all-words domain WSD (nouns only subset). Best results for each measure are bolded. † indicates statistically significant differences with respect to the monolingual setting.

Results and discussion. We report our results in terms of precision (P), recall (R) and F_1 measure in Table 1, where we compare the monolingual variant (rows 1–2 of the table) with our multilingual approach (rows 3–4). Following standard practice, (1) we benchmark our method against two baselines, namely a random sense assignment and the most frequent sense (MFS) from SemCor; (2) we test for statistical significance by computing a 95% confidence interval on the recall score (i.e., the main evaluation measure for the WSD task) using bootstrap resampling (Noreen, 1989).

The results show that our multilingual approach improves over the monolingual one by a substantial (i.e., statistically significant) margin. Combining multilingual information from different languages yields a higher precision (+3.3 for both graph algorithms) and recall (+3.4 and +2.9 for Degree and PLength, respectively). Manual inspection of the output reveals that these increases in precision are due to translations in different languages constraining each other – e.g., an implausible English sense is 'ruled out' from the sense distributions of the other languages (cf. the example in Figure 1). The increases in recall, instead, indicate that using translations triggers responses in those cases where no sense of the English target word can be connected to the senses of the context words – i.e., some trans-

	Algorithm	P	R	F_1
Monolingual graph	Degree	52.0	51.3	51.6
	PLength	55.0	54.2	54.6
Multilingual ensemble	Degree†	61.6	59.5	60.5
	PLength†	**62.5**	**60.4**	**61.4**
	CFILT	61.4	59.4	60.4
	IIITH	56.4	55.3	55.8

Table 2: Performance on SemEval-2010 all-words domain WSD (nouns only subset) using the most frequent sense assigned by the system as back-off strategy when no sense assignment is attempted.

lations activate senses in the knowledge base which are closer to the senses of the context words. The result is an overall increase in F_1 measure of 3.4 and 3.1 points for Degree and PLength, respectively, which makes it possible for us to beat the MFS baseline (notably a difficult competitor for WSD systems). Among the different graph algorithms, PLength consistently outperforms Degree: however, the differences are not statistically significant.

In order to better understand the impact of our approach we follow previous work (e.g., Navigli and Lapata (2010)) and explore a weakly-supervised setting where the system attempts no sense assignment if the highest score among those assigned to the senses of a target word is below a certain threshold. If this is the case, in order to provide an answer for all items, we output the most frequent sense assigned by the system to other instances of the target word, and fall back to SemCor's MFS if no assignment has been attempted. We estimate the optimal value for the threshold by maximizing F_1 on a development set obtained by combining the Senseval-2 (Palmer et al., 2001) and Senseval-3 (Snyder and Palmer, 2004) English all-words datasets. The results for this setting are shown in Table 2, where we also compare with the top-performing systems from the SemEval competition, namely CFILT (Kulkarni et al., 2010) and IIITH (Reddy et al., 2010).

By complementing our multilingual method with the MFS heuristic we achieve a performance comparable with the state of the art on this task. Again, the multilingual ensemble approach consistently outperforms the monolingual one and enables us to achieve the best overall results for this dataset: without multilingual information, in fact, we achieve only average performance above the MFS level, whereas by effectively combining sense evidence from multilingual translations we are able to boost the F_1 measure by a 6-8 point margin, and thus outperform the top-ranking SemEval systems. While differences with CFILT are not statistically significant, we still take this to be good news, since our system is general purpose in nature and, accordingly, does not use any domain information such as manually-labeled examples for the most frequent domain words (CFILT) or a domain-specific sense ranking (IIITH).

4.2 Cross-lingual lexical disambiguation

Using a multilingual lexical resource makes it possible to perform WSD in any of its languages. Accordingly, we complement our evaluation on English texts with a second set of experiments where we quantify the impact of our approach on a lexical disambiguation task in a multilingual setting. To this end, we use the SemEval-2010 cross-lingual lexical substitution (Mihalcea et al., 2010, CL-LS, henceforth) and WSD (Lefever et al., 2011, CL-WSD) tasks and evaluate our methodology on performing disambiguation across different languages. Both cross-lingual WSD tasks cast disambiguation as a word translation problem: given an English polysemous noun in context as input, the system disambiguates it by providing a translation into another language (translations are deemed correct if they preserve the meaning of the source word in the target language). Their main difference, instead, lies in the range of translations which are assumed to be valid: that is, while CL-LS assumes no predefined sense inventory (i.e., any translation can be potentially correct), CL-WSD makes use of a sense inventory built on the basis of the Europarl corpus (Koehn, 2005).

Our approach to lexical disambiguation involves two steps: first, given a target word in context, we disambiguate it as usual to the highest-ranked Babel synset; next, given the translations in the selected synset, we return the most suitable lexicalization in the language of interest. Since the selected synset can contain multiple translations in a target language for the input English word, we explore using an unsupervised strategy to select the most reliable translation from multiple candidates. To this end, we return for each test instance only the

	Algorithm	P/R/F_1
	Baseline	23.80
Monolingual graph	Degree	30.52
	PLength	30.64
Multilingual ensemble	Degree	32.21
	PLength	**32.47**
	UBA-T	32.17

Table 3: Performance on SemEval-2010 lexical substitution (best results are bolded).

most frequent translation found in the Babel synset. Given that the two tasks make different assumptions on the sense inventory (no fixed inventory for CL-LS vs. Europarl-based for CL-WSD), the frequency of a translation is calculated as either the number of Babel synsets in which it occurs (CL-LS), or its frequency of alignment with the target word, as obtained by applying GIZA++ (Och and Ney, 2003) to Europarl (CL-WSD). To provide an answer for all instances, we return this most frequent translation even when no sense assignment is attempted – i.e., no sense of the target word is connected to any other sense of the context words – or a tie occurs.

Results and discussion. We report our results for CL-LS and CL-WSD in Tables 3 and 4. We evaluate using the nouns-only subset of the CL-LS dataset and the full CL-WSD dataset, consisting of 300 and 1,000 instances of nouns in context, respectively. The evaluation scheme is based on the SemEval-2007 English lexical substitution task (McCarthy and Navigli, 2009), and consists of an adaptation of the metrics of precision and recall for the translation setting. For each task, we compare our monolingual and multilingual approaches against the best performing SemEval systems for these tasks, namely UBA-T (Basile and Semeraro, 2010) and UVT-v (van Gompel, 2010) for CL-LS and CL-WSD, respectively, as well as a recent supervised proposal that exploits automatically generated multilingual features from parallel text and translated contexts (Lefever et al., 2011, Parasense). For each task we also report its official baseline, namely the first translation from an online-dictionary[6] for CL-LS, and the most frequent word alignment obtained by applying GIZA++ to the Europarl data for CL-WSD.

Our cross-lingual results confirm all trends of the English monolingual evaluation, namely that: a) our joint multilingual approach substantially improves over the simple monolingual graph-based approach; b) it enables us to achieve state-of-the-art performance for these tasks. In the case of both CL-LS and CL-WSD, using a rich multilingual knowledge base like BabelNet makes it possible to achieve a respectable performance already with the simple monolingual approach, thus indicating the viability of a knowledge-rich approach to sense-driven word translation. The use of multilingual ensembles always improves the monolingual setting for all languages, and allows us to achieve the best overall results for both CL-LS and CL-WSD. Similarly to the case of monolingual WSD, manual inspection of the output reveals that translations help us rule out incorrect senses and let the disambiguation algorithm focus on the more coherent set of senses for the input context in a way similar to the one highlighted by the example in Figure 1. As a result of this we are able to improve the performance of both monolingual Degree and PLength, and compete with the state of the art on all disambiguation tasks.

5 Conclusions

In this paper we presented a multilingual joint approach to WSD. Key to our methodology is the effective use of a wide-coverage multilingual knowledge base, BabelNet, which we exploit to perform graph-based WSD across languages and combine complementary sense evidence from translations in different languages using an ensemble method. This is the first proposal to exploit structured multilingual information within a joint, knowledge-rich framework for WSD. The APIs to perform multilingual WSD using BabelNet are freely available for research purposes (Navigli and Ponzetto, 2012b).

Thanks to multilingual joint WSD we achieve state-of-the-art performance on three different gold standards. The good news about these results is that not only can further advances be achieved by using multilingual lexical knowledge, but, more importantly, that combining multilingual sense evidence from different languages at the same time yields consistent improvements over a monolingual ap-

[6] www.spanishdict.com

		French P/R/F_1	German P/R/F_1	Italian P/R/F_1	Spanish P/R/F_1
	Baseline	21.25	13.16	15.18	19.74
	UvT-v	N/A	N/A	N/A	23.39
	Parasense	24.54	16.88	18.03	22.80
Monolingual graph	Degree	22.94	17.15	18.03	22.48
	PLength	23.42	17.72	18.19	22.76
Multilingual ensemble	Degree	24.02	18.07	18.93	23.51
	PLength	**24.61**	**18.26**	**19.05**	**23.65**

Table 4: Results on the SemEval-2010 cross-lingual WSD dataset (best results are bolded).

proach in both monolingual and cross-lingual lexical disambiguation tasks – that is, 'joining forces pays off'. Effectively leveraging multilingual knowledge for WSD helps overcome the shortcomings of the underlying resource (noise, coverage, etc.), thus indicating that further performance boosts can come in the future from even better multilingual lexical resources. Moreover, our methodology is general-purpose and can be adapted to tasks other than WSD: in fact, we have already taken the first steps in this direction by showing the beneficial effects of a joint multilingual approach to computing semantic relatedness (Navigli and Ponzetto, 2012a). In addition, we plan in the very near future to generalize our multilingual joint approach and apply it to high-end tasks such as multilingual textual entailment (Mehdad et al., 2011) and sentiment analysis (Lu et al., 2011) – so as to provide a general framework for knowledge-rich multilingual NLP.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

BabelNet and its API are available for download at `http://lcl.uniroma1.it/babelnet`.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2009. Knowledge-based WSD on specific domains: performing better than generic supervised WSD. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 1501–1506.

Eneko Agirre, Oier López de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. Semeval-2010 task 17: All-words Word Sense Disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 75–80.

Carmen Banea and Rada Mihalcea. 2011. Word Sense Disambiguation with multilingual features. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS 2011)*, pages 25–34.

Pierpaolo Basile and Giovanni Semeraro. 2010. UBA: Using automatic translation and Wikipedia for cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 242–247.

Samuel Brody, Roberto Navigli, and Mirella Lapata. 2006. Ensemble methods for unsupervised WSD. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 97–104.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pages 264–270.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning (EMNLP-CoNLL-07)*, pages 61–72.

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up Word Sense Disambiguation via parallel texts. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, pages 1037–1042.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation improves Statistical Ma-

chine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 33–40.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pages 130–137.

Gerard de Melo and Gerhard Weikum. 2010. MENTA: Inducing multilingual taxonomies from Wikipedia. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM-10)*, pages 1099–1108.

Mona Diab. 2003. *Word Sense Disambiguation within a Multilingual Framework*. Ph.D. thesis, University of Maryland, College Park, Maryland.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Radu Florian, Silviu Cucerzan, Charles Schafer, and David Yarowsky. 2002. Combining classifiers for Word Sense Disambiguation. *Natural Language Engineering*, 8(4):1–14.

William A. Gale, Kenneth Church, and David Yarowsky. 1992. Using bilingual materials to develop Word Sense Disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112.

Nancy Ide, Tomaz Erjavec, and Dan Tufiş. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on WSD: Recent Successes and Future Directions*, pages 54–60.

Nancy Ide. 2000. Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34:223–234.

Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for WSD. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (ACL-11)*, pages 561–569.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*.

Anup Kulkarni, Mitesh Khapra, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. CFILT: Resource conscious approaches for all-words domain specific WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 421–426.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 15–20.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for Word Sense Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (ACL-11)*, pages 317–322.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (ACL-11)*, pages 320–330.

Bernardo Magnini, Danilo Giampiccolo, and Alessandro Vallin. 2004. The Italian lexical sample task at Senseval-3. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 17–20.

Lluis Màrquez, Mariona Taulé, Antonia Martí, Núria Artigas, Mar García, Francis Real, and Dani Ferrés. 2004. Senseval-3: The Spanish lexical sample task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 21–24.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (ACL-11)*, pages 1336–1345.

Christian M. Meyer and Iryna Gurevych. 2012. Ontowiktionary – Constructing an ontology from the collaborative online dictionary Wiktionary. In Maria Teresa Pazienza and Armando Stellato, editors, *Semi-Automatic Ontology Development: Processes and Resources*. IGI Global, Hershey, Penn.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 9–14.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308.

Vivi Nastase, Michael Strube, Benjamin Börschinger, Caecilia Zirn, and Anas Elghafari. 2010. WikiNet: A very large scale multi-lingual concept network. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, (LREC '10)*.

Roberto Navigli and Mirella Lapata. 2010. An experimental study on graph connectivity for unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Anaylsis and Machine Intelligence*, 32(4):678–692.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (ACL-10)*, pages 216–225.

Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelRelate! A joint multilingual approach to computing semantic relatedness. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI-12)*.

Roberto Navigli and Simone Paolo Ponzetto. 2012b. Multilingual WSD with just a few lines of code: The BabelNet API. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, (ACL-12). System Demonstrations*.

Roberto Navigli, Stefano Faralli, Aitor Soroa, Oier Lopez de Lacalle, and Eneko Agirre. 2011. Two birds with one stone: Learning semantic models for Text Categorization and Word Sense Disambiguation. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM-11)*, pages 2317–2320.

Roberto Navigli. 2009. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Eric W. Noreen, editor. 1989. *Computer-intensive methods for testing hypotheses: an introduction*. New York, N.Y.: John Wiley.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. SemEval-2010 task: Japanese WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 69–74.

Zeynep Orhan, Emine Çelik, and Demirgüç Neslihan. 2007. SemEval-2007 task 12: Turkish lexical sample task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 59–63.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 21–24.

Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (ACL-10)*, pages 1522–1531.

Siva Reddy, Abhilash Inumella, Diana McCarthy, and Mark Stevenson. 2010. IIITH: Domain specific Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 387–391.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Journal of Natural Language Engineering*, 5(2):113–133.

Carina Silberer and Simone Paolo Ponzetto. 2010. UHD: Cross-lingual Word Sense Disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 134–137.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41–43.

Maarten van Gompel. 2010. UvT-WSD1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 238–241.

Zhi Zhong and Hwee Tou Ng. 2009. Word Sense Disambiguation for all words without hard labor. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 1616–1622.