A Unified Multilingual Semantic Representation of Concepts

José Camacho-Collados, Mohammad Taher Pilehvar and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
`{collados,pilehvar,navigli}@di.uniroma1.it`

Abstract

Semantic representation lies at the core of several applications in Natural Language Processing. However, most existing semantic representation techniques cannot be used effectively for the representation of individual word senses. We put forward a novel multilingual concept representation, called MUFFIN, which not only enables accurate representation of word senses in different languages, but also provides multiple advantages over existing approaches. MUFFIN represents a given concept in a unified semantic space irrespective of the language of interest, enabling cross-lingual comparison of different concepts. We evaluate our approach in two different evaluation benchmarks, semantic similarity and Word Sense Disambiguation, reporting state-of-the-art performance on several standard datasets.

1 Introduction

Semantic representation, i.e., the task of representing a linguistic item (such as a word or a word sense) in a mathematical or machine-interpretable form, is a fundamental problem in Natural Language Processing (NLP). The Vector Space Model (VSM) is a prominent approach for semantic representation, with widespread popularity in numerous NLP applications. The prevailing methods for the computation of a vector space representation are based on distributional semantics (Harris, 1954). However, these approaches, whether in their conventional co-occurrence based form (Salton et al., 1975; Turney and Pantel, 2010; Landauer and Dooley, 2002), or in their newer predictive branch (Collobert and Weston, 2008; Mikolov et al., 2013; Baroni et al., 2014), suffer from a major drawback: they are unable to model individual word senses or concepts, as they conflate different meanings of a word into a single vectorial representation. This hinders the functionality of this group of vector space models in tasks such as Word Sense Disambiguation (WSD) that require the representation of individual word senses. There have been several efforts to adapt and apply distributional approaches to the representation of word senses (Pantel and Lin, 2002; Brody and Lapata, 2009; Reisinger and Mooney, 2010; Huang et al., 2012). However, none of these techniques provides representations that are already linked to a standard sense inventory, and consequently such mapping has to be carried out either manually, or with the help of sense-annotated data. Chen et al. (2014) addressed this issue and obtained vectors for individual word senses by leveraging WordNet glosses. NASARI (Camacho-Collados et al., 2015) is another approach that obtains accurate sense-specific representations by combining the complementary knowledge from WordNet and Wikipedia. Graph-based approaches have also been successfully utilized to model individual words (Hughes and Ramage, 2007; Agirre et al., 2009; Yeh et al., 2009), or concepts (Pilehvar et al., 2013; Pilehvar and Navigli, 2014), drawing on the structural properties of semantic networks. The applicability of all these techniques, however, is usually either constrained to a single language (usually English), or to a specific task.

We put forward MUFFIN (Multilingual, Unified and Flexible INterpretation), a novel method that exploits both structural knowledge derived from semantic networks and distributional statistics from text corpora, to produce effective representations of individual word senses or concepts. Our approach provides multiple advantages in comparison to the previous VSM techniques:

1. *Multilingual*: it enables sense representation in dozens of languages;
2. *Unified*: it represents a linguistic item, irrespective of its language, in a unified seman-
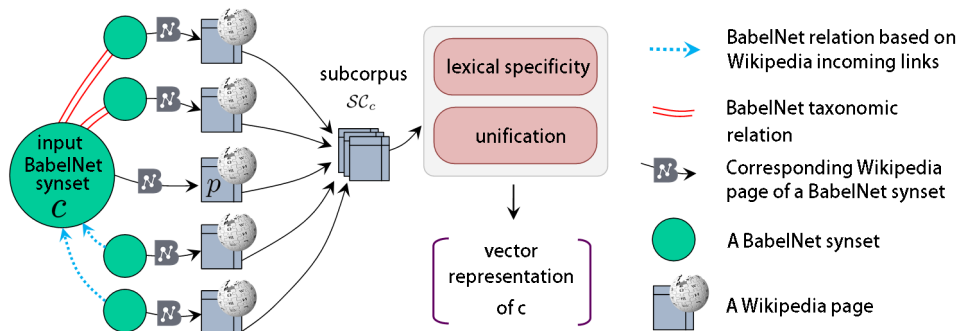
741

Figure 1: Our procedure for constructing a multilingual vector representation for a concept c.

tic space having concepts as its dimensions, permitting direct comparison of different representations across languages, and hence enabling cross-lingual applications;

3. *Flexible*: it can be readily applied to different NLP tasks with minimal adaptation.

We evaluate our semantic representation on two different tasks in lexical semantics: semantic similarity and Word Sense Disambiguation. To assess the multilingual capability of our approach, we also perform experiments on languages other than English on both tasks, and across languages for semantic similarity. We report state-of-the-art performance on multiple datasets and settings in both frameworks, which confirms the reliability and flexibility of our representations.

2 Methodology

Figure 1 illustrates our procedure for constructing the vector representation of a given concept. We use BabelNet[1] (version 2.5) as our main sense repository. BabelNet (Navigli and Ponzetto, 2012a) is a multilingual encyclopedic dictionary which merges WordNet with other lexical resources, such as Wikipedia and Wiktionary, thanks to its use of an automatic mapping algorithm. BabelNet extends the WordNet synset model to take into account multilinguality: a BabelNet synset contains the words that, in the various languages, express the given concept.

Our approach for modeling a BabelNet synset consists of two main steps. First, for the given synset we gather contextual information from Wikipedia by exploiting knowledge from the BabelNet semantic network (Section 2.1). Then, by analyzing the corresponding contextual information and comparing and contrasting it with the

whole Wikipedia corpus, we obtain a vectorial representation of the given synset (Section 2.2).

2.1 A Wikipedia sub-corpus for each concept

Let c be a concept, which in our setting is a BabelNet synset, and let $\mathcal{W}_c$ be the set containing the Wikipedia page p corresponding to the concept c and all the Wikipedia pages having an outgoing link to p. We further enrich $\mathcal{W}_c$ with the corresponding Wikipedia pages of the hypernyms and hyponyms of c in the BabelNet network. $\mathcal{W}_c$ is the set of Wikipedia pages whose contents are exploited to build a representation for the concept c. We refer to the bag of content words in all the Wikipedia pages in $\mathcal{W}_c$ as the sub-corpus $\mathcal{SC}_c$ for the concept c.

2.2 Vector construction: lexical specificity

Lexical specificity (Lafon, 1980) is a statistical measure based on the hypergeometric distribution. Due to its efficiency in extracting a set of highly relevant words from a sub-corpus, the measure has recently gained popularity in different NLP applications, such as textual data analysis (Lebart et al., 1998), term extraction (Drouin, 2003), and domain-based term disambiguation (Camacho-Collados et al., 2014; Billami et al., 2014). We leverage lexical specificity to compute the weights in our vectors. In our earlier work (Camacho-Collados et al., 2015), we conducted different experiments which demonstrated the improvement that lexical specificity can provide over the popular term frequency-inverse document frequency weighting scheme (Jones, 1972, *tf-idf*). Lexical specificity computes the vector weights for an item, i.e., a word or a set of words, by comparing and contrasting its contextual information with a reference corpus. In our setting, we take the whole Wikipedia as our reference corpus $\mathcal{RC}$ (we use the October 2012 Wikipedia dump).

---

[1] http://www.babelnet.org

Let T and t be the respective total number of tokens in $\mathcal{RC}$ and $\mathcal{SC}_c$, while F and f denote the frequency of a given item in $\mathcal{RC}$ and $\mathcal{SC}_c$, respectively. Our goal is to compute a weight denoting the association of an item to the concept c. For notational brevity, we use the following expression to refer to positive lexical specificity:

$$specificity(T, t, F, f) = -\log_{10} P(X \geq f) \quad (1)$$

where X represents a random variable following a hypergeometric distribution of parameters F, t and T. As we are only interested in a set of items that are representative of the concept being modeled, we follow Billami et al. (2014) and only consider in our final vector the items which are relevant to $\mathcal{SC}_c$ with a confidence higher than 99% according to the hypergeometric distribution ($P(X \geq f) \leq 0.01$).

On the basis of lexical specificity we put forward two types of representations: *lexical* and *unified*. The lexical vector representation lex_c of a concept c has lemmas as its individual dimensions. To this end, we apply lexical specificity to every lemma in $\mathcal{SC}_c$ in order to estimate the relevance of each lemma to our concept c. We use the lexical representation for the task of WSD (see Section 3.2). We describe the unified representation in the next subsection.

2.3 Unified representation

Unlike the lexical version, our *unified* representation has concepts as individual dimensions. Algorithm 1 shows the construction process of a concept's unified vector. The algorithm first clusters together those words that have a sense sharing the same hypernym (h in the algorithm) according to the BabelNet taxonomy (lines 2-4). Next, the specificity is computed for the set of all the hyponyms of h, even those that do not appear in the sub-corpus $\mathcal{SC}_c$ (lines 6-14). Here, F and f denote the aggregated frequencies of all the hyponyms of h in the whole Wikipedia (i.e., reference corpus $\mathcal{RC}$) and the sub-corpus $\mathcal{SC}_c$, respectively.

Our binding of a set of sibling words into a single cluster represented by their common hypernym provides two advantages. Firstly, it transforms the representations to a unified semantic space. This space has concepts as its dimensions, enabling their comparability across languages. Secondly, the clustering can be viewed as an implicit disambiguation process, whereby a set of potentially

---

Algorithm 1 Unified Vector Construction

Input: a concept c
Output: the unified vector u_c where $u_c(h)$ is the dimension corresponding to concept h
1: $H \leftarrow \emptyset$
2: **for each** lemma $l \in \mathcal{SC}_c$
3:     **for each** hypernym h of l in BabelNet
4:        $H \leftarrow H \cup \{h\}$
5: vector $u_c \leftarrow$ null vector
6: **for each** $h \in H$
7:     **if** $\exists\, l_1, l_2 \in \mathcal{SC}_c$: l_1, l_2 hyponyms of h and $l_1 \neq l_2$ **then**
8:        $F \leftarrow 0$
9:        $f \leftarrow 0$
10:        **for each** hyponym $hypo$ of h
11:           **for each** lexicalization lex of $hypo$
12:              $F \leftarrow F + freq(lex, \mathcal{RC})$
13:              $f \leftarrow f + freq(lex, \mathcal{SC}_c)$
14:        $u_c(h) \leftarrow specificity(T, t, F, f))$
15: **return** vector u_c

---

ambiguous words are disambiguated into their intended sense on the basis of the contextual clues of the neighbouring content words, resulting in more accurate representations of meaning.

Example. Table 1 lists the top-weighted concepts, represented by their relevant lexicalizations, in the unified vectors generated for the bird and machine senses of the noun *crane* and for three different languages.[2] A comparison of concepts across the two senses indicates the effectiveness of our representation in identifying relevant concepts in different languages, while guaranteeing a clear distinction between the two meanings.

3 Applications

Thanks to their VSM nature and the sense-level functionality, our concept representations are highly flexible, allowing us to adapt and apply them to different NLP tasks with minimal adaptation. In this section we explain how we use our representations in the tasks of semantic similarity (Section 3.1) and WSD (Section 3.2).

Associating concepts with words. Given that our representations are for individual word senses, a preliminary step for both tasks would be to associate the set of concepts, i.e., BabelNet synsets, $\mathcal{C}_w = \{c_1, ..., c_n\}$ with a given word w. In the case when w exists in the BabelNet dictionary, we obtain the set of associated senses of the word as defined in the BabelNet sense inventory.

In order to enhance the coverage in the case of

---

[2]We use the sense notation of Navigli (2009): $word_n^p$ is the nth sense of the *word* with part of speech p.

743

Crane (bird)			Crane (machine)		
English	French	German	English	French	German
shore_bird$_n^1$	‡famille_des_oiseaux$_n^1$	‡vogel-familie$_n^1$	*lifting device$_n^1$	*dispositif de levage$_n^1$	*hebevorrichtung$_n^1$
bird$_n^1$	*limicole$_n^1$	*charadrii$_n^1$	‡construction$_n^4$	navire$_n^1$	radfahrzeug$_n^1$
*wading_bird$_n^1$	oiseau_aquatique$_n^2$	†vogel_gattung$_n^1$	platform$_n^1$	limicole$_n^1$	†lenkfahrzeug$_n^1$
oscine_bird$_n^1$	tollé$_n^2$	wirbeltiere$_n^2$	warship$_n^1$	◇vaisseau$_n^2$	regler$_n^3$
†bird_genus$_n^1$	gallinacé$_n^1$	fleisch$_n^1$	electric circuit$_n^1$	spationef$_n^1$	reisebus$_n^1$
‡bird_family$_n^1$	◇classe$_n^1$	tier um$_n^1$	◇vessel$_n^2$	‡construction$_n^2$	charadrii$_n^1$
◇taxonomic_group$_n^1$	occurence$_n^1$	reiher$_n^1$	boat$_n^1$	†véhicule$_n^3$	güterwagen$_n^2$

Table 1: Top-weighted concepts, i.e., BabelNet synsets, for the bird and machine senses of the noun *crane*. We represent each synset by one of its word senses. Word senses marked with the same symbol across languages correspond to the same BabelNet synset.

words that are not defined in the BabelNet dictionary, we also exploit the so-called Wikipedia piped links. A piped link is a hyperlink appearing in the body of a Wikipedia article, providing a link to another Wikipedia article. For example, the piped link [[dockside_crane|Crane_(machine)]] is a hyperlink that appears as *dockside_crane* in the text, but takes the user to the Wikipedia page titled *Crane_(machine)*. These links provide Wikipedia editors with the ability to represent a Wikipedia article through a suitable lexicalization that preserves the grammatical structure, contextual coherency, and flow of the sentence. This property provides an effective means of obtaining a set of concepts for the words not covered by BabelNet. For the case of our example, the BabelNet out-of-vocabulary word $w = dockside\_crane$ will have in its set of associated concepts $\mathcal{C}_w$ the BabelNet synset corresponding to the Wikipedia page titled *Crane_(machine)*.

3.1 Semantic Similarity

Once we have the set $\mathcal{C}_w$ of concepts associated with each word w, we first retrieve the set of corresponding unified vector representations. We then follow Camacho-Collados et al. (2015) and use square-rooted Weighted Overlap (Pilehvar et al., 2013, WO) as our vector comparison method, a metric that has been shown to suit specificity-based vectors more than the conventional cosine. WO compares two vectors on the basis of their overlapping dimensions, which are harmonically weighted by their relative ranking:

$$WO(v_1, v_2) = \frac{\sum_{q \in O} \left(rank(q, v_1) + rank(q, v_2)\right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}} \quad (2)$$

where O is the set of overlapping dimensions (i.e. concepts) between the two vectors and $rank(q, v_i)$ is the rank of dimension q in the vector v_i.

Finally, the similarity between two words w_1 and w_2 is calculated as the similarity of their closest senses, a prevailing approach in the literature (Resnik, 1995; Budanitsky and Hirst, 2006):

$$sim(w_1, w_2) = \max_{v_1 \in \mathcal{C}_{w_1}, v_2 \in \mathcal{C}_{w_2}} \sqrt{WO(v_1, v_2)} \quad (3)$$

where w_1 and w_2 can belong to different languages. This cross-lingual similarity measurement is possible thanks to the unified language-independent space of concepts of our semantic representations.

3.2 Multilingual Word Sense Disambiguation

In order to be able to apply our approach to WSD, we use the lexical vector lex_c for each concept c. The reason for our choice of lexical vectors in this setting is that they enable a direct comparison of a candidate sense's representation with the context, which is also in the same lexical form. Algorithm 2 summarizes the general framework of our approach. Given a target word w to disambiguate, our approach proceeds by the following steps:

1. Retrieve $\mathcal{C}_w$, the set of associated concepts with the target word w (line 1);

2. Obtain the lexical vector lex_c for each concept $c \in \mathcal{C}_w$ (cf. Section 2);

3. Calculate, for each candidate concept c, a confidence score ($score_c$) based on the harmonic sum of the ranks of the overlapping words between its lexical vector lex_c and the context of the target word (line 5 in Algorithm 2).

744

Algorithm 2 MUFFIN for WSD

Input: a target word w and a document d (context of w)
Output: $\hat{c}$, the intended sense of w
1: **for each** concept $c \in \mathcal{C}_w$
2: $\quad score_c \leftarrow 0$
3: $\quad$ **for each** lemma $l \in d$
4: $\quad\quad$ **if** $l \in lex_c$ **then**
5: $\quad\quad\quad score_c \leftarrow score_c + \big(rank(l, lex_c)\big)^{-1}$
6: $\hat{c} \leftarrow \underset{c \in C_w}{\arg\max}\, score_c$
7: **return** $\hat{c}$

Thanks to the use of BabelNet, our approach is applicable to arbitrary languages. For the task of WSD, we focus on two major sense inventories integrated in BabelNet: Wikipedia and WordNet.

Wikipedia sense inventory. In this case, we obtain the set of candidate senses for a target word by following the procedure described in the beginning of this Section (i.e., associating concepts with words). However, we do not consider those BabelNet synsets that are not associated with Wikipedia pages.

WordNet sense inventory. Similarly, when restricted to the WordNet inventory, we discard those BabelNet synsets that do not contain a WordNet synset. In this setting, we also leverage relations from WordNet's semantic network and its disambiguated glosses[3] in order to obtain a richer set of Wikipedia articles in the sub-corpus construction. The enrichment of the semantic network with the disambiguated glosses has been shown to be beneficial in various graph-based disambiguation tasks (Navigli and Velardi, 2005; Agirre and Soroa, 2009; Pilehvar et al., 2013).

4 Experiments

We assess the reliability of MUFFIN in two standard evaluation benchmarks: semantic similarity (Section 4.1) and Word Sense Disambiguation (Section 4.2).

4.1 Semantic Similarity

As our semantic similarity experiment we opted for word similarity, which is one of the most popular evaluation frameworks in lexical semantics. Given a pair of words, the task in word similarity is to automatically judge their semantic similarity and, ideally, this judgement should be close to that given by humans.

---

[3] http://wordnet.princeton.edu/glosstag.shtml

4.1.1 Datasets

Monolingual. We picked the RG-65 dataset (Rubenstein and Goodenough, 1965) as our monolingual word similarity dataset. The dataset comprises 65 English word pairs which have been manually annotated by several annotators according to their similarity on a scale of 0 to 4. We also perform evaluations on the French (Joubarne and Inkpen, 2011) and German (Gurevych, 2005) adaptations of this dataset.

Cross-lingual. Hassan and Mihalcea (2009) developed two sets of cross-lingual datasets based on the English MC-30 (Miller and Charles, 1991) and WordSim-353 (Finkelstein et al., 2002) datasets, for four different languages: English, German, Romanian, and Arabic. However, the construction procedure they adopted, consisting of translating the pairs to other languages while preserving the original similarity scores, has led to inconsistencies in the datasets. For instance, the Spanish dataset contains the identical pair *mediodia-mediodia* with a similarity score of 3.42 (in the scale [0,4]). Additionally, the datasets contain several orthographic errors, such as *despliege* and *grua* (instead of *despliegue* and *grúa*) and incorrect translations (e.g., the English noun *implement* translated into the Spanish verb *implementar*).

Kennedy and Hirst (2012) proposed a more reliable procedure that leverages two existing aligned monolingual word similarity datasets for the construction of a new cross-lingual dataset. To this end, for each two word pairs *a-b* and *a'-b'* in the two datasets, if the difference in the corresponding scores is greater than one, the pairs are discarded. Otherwise, two new pairs *a-b'* and *a'-b* are created with a score equal to the average of the two original pairs' scores. In the case of repeated pairs, we merge them into a single pair with a similarity equal to their average scores. Using this procedure as a basis, Kennedy and Hirst (2012) created an English-French dataset consisting of 100 pairs. We followed the same procedure and built two datasets for English-German (consisting of 125 pairs) and German-French (comprising 96 pairs) language pairs.[4]

4.1.2 Comparison systems

Monolingual. We benchmark our system against four other approaches that exploit

---

[4] The cross-lingual datasets are available at http://lcl.uniroma1.it/sim-datasets/.

English	ρ	r	German	ρ	r	French	ρ	r
MUFFIN	0.83	0.84	MUFFIN	**0.77**	**0.76**	MUFFIN	**0.71**	**0.77**
SOC-PMI	–	0.61	SOC-PMI	–	0.27	SOC-PMI	–	0.19
PMI	–	0.41	PMI	–	0.40	PMI	–	0.34
Retrofitting	0.74	–	Retrofitting	0.60	–	Retrofitting	0.61	–
LSA-Wiki	0.69	0.65	–	–	–	LSA-Wiki	0.52	0.57
Wiki-wup	–	0.59	Wiki-wup	–	0.65			
SSA	0.83	**0.86**	Resnik	–	0.72			
NASARI	0.84	0.82	Lesk_hyper	–	0.69			
ADW	**0.87**	0.81						
Word2Vec	–	0.84						
PMI-SVD	–	0.74						
ESA	–	0.72						

Table 2: Spearman (ρ) and Pearson (r) correlation performance of different systems on the English, German and French RG-65 datasets.

Wikipedia as their main knowledge resource: SSA[5] (Hassan and Mihalcea, 2011), ESA (Gabrilovich and Markovitch, 2007), Wiki-wup (Ponzetto and Strube, 2007), and LSA-Wiki (Granada et al., 2014). We also provide results for systems that use distributional semantics for modeling words, both the conventional co-occurrence based approach, i.e., PMI-SVD (Baroni et al., 2014), PMI and SOC-PMI (Joubarne and Inkpen, 2011), and Retrofitting (Faruqui et al., 2015), and the newer word embeddings, i.e., Word2Vec (Mikolov et al., 2013). For Word2Vec and PMI-SVD, we use the pre-trained models obtained by Baroni et al. (2014).[6] As for WordNet-based approaches, we report results for Resnik (Resnik, 1995) and ADW (Pilehvar et al., 2013), which take advantage of its structural information, and Lesk_hyper (Gurevych, 2005), which leverages definitional information in WordNet for similarity computation. Finally, we also report the performance of our earlier work NASARI (Camacho-Collados et al., 2015), which combines knowledge from WordNet and Wikipedia for the English language in its setting without the Wiktionary synonyms module.

Cross-lingual. We compare the performance of our approach against the best configuration of the CL-MSR-2.0 system (Kennedy and Hirst, 2012), which exploits Pointwise Mutual Information (PMI) on a parallel corpus obtained from

the English and French versions of WordNet. Since two of our cross-lingual datasets are newly-created, we developed three baseline systems to enable a more meaningful comparison. To this end, we first use Google Translate to translate the non-English side of the dataset to the English language. Accordingly, three state-of-the-art graph-based and corpus-based approaches were used to measure the similarity of the resulting English pairs. As English similarity measurement systems, we opted for ADW (Pilehvar et al., 2013), and the best predictive (Mikolov et al., 2013, Word2Vec) and co-occurrence (i.e., PMI-SVD) models obtained by Baroni et al. (2014).[7] In our experiments we refer to these systems as *pivot*, since they use English as a pivot for computing semantic similarity. As a comparison, we also show results for MUFFIN$_{pivot}$, which is the variant of our system applied to the same automatically translated monolingual datasets.

4.1.3 Results

Monolingual. We show in Table 2 the performance of different systems in terms of Spearman and Pearson correlations on the English, German, and French RG-65 datasets. On the German and French datasets, our system outperforms the comparison systems according to both evaluation measures. It achieves considerable Spearman and Pearson correlation leads of 0.1 and 0.2, respectively, on the French dataset in comparison to the best system. Also on the English RG-65 dataset, our system attains competitive performance according to both Spearman and Pearson correla-

---

[5]SSA involves several parameters tuned on datasets that are constructed on the basis of MC-30 and RG-65.

[6]We report the best configuration of the systems on the RG-65 dataset out of their 48 configurations. The corpus used to train the models contained 2.8 billion tokens, including Wikipedia (Baroni et al., 2014).

[7]http://clic.cimec.unitn.it/composes/semantic-vectors.html

Measure	FR-EN	EN-DE	DE-FR
MUFFIN	**0.83**	**0.76**	**0.83**
MUFFIN$_{pivot}$	**0.83**	0.73	0.79
ADW$_{pivot}$	0.80	0.73	0.72
Word2Vec$_{pivot}$	0.75	0.69	0.77
PMI-SVD$_{pivot}$	0.76	0.72	0.65
CL-MSR-2.0	0.30	–	–

Table 3: Pearson correlation performance of different similarity measures on the three cross-lingual RG-65 datasets.

tions. We note that most state-of-the-art systems on the dataset (e.g., ADW) are restricted to the English language only.

Cross-lingual. Pearson correlation results on the three cross-lingual RG-65 datasets are presented in Table 3. Similarly to the monolingual experiments, our system proves highly reliable in the cross-lingual setting, improving the performance of the comparison systems on all three language pairs. Moreover, MUFFIN$_{pivot}$ attains the best results among the *pivot* systems on all datasets, confirming the reliability of our system in the monolingual setting. We note that since the cross-lingual datasets were built by translating the word pairs in the original English RG-65 dataset, the pivot-based comparison systems proved to be highly competitive, outperforming the CL-MSR-2.0 system by a considerable margin.

4.2 Word Sense Disambiguation

4.2.1 Wikipedia

In this setting, we selected the SemEval 2013 all-words WSD task (Navigli et al., 2013) as our evaluation benchmark. The task provides datasets for five different languages: Italian, English, French, Spanish and German. There are on average 1123 words to disambiguate in each language's dataset. As comparison system, we provide results for the best-performing participating system on each language. We also show results for the state-of-the-art WSD system of Moro et al. (2014, Babelfy), which relies on random walks on the BabelNet semantic network and a set of graph heuristic algorithms. Finally, we also report results for the Most Frequent Sense (MFS) baseline provided by the task organizers.

We follow Moro et al. (2014) and back off to the MFS baseline in the case when our system's

judgement does not meet a threshold θ. Similarly to Babelfy, we tuned the value of the threshold θ on the trial dataset provided by the organizers of the task. We tuned θ with step size 0.05 (hence, 21 possible values in [0,1]), obtaining an optimal value of 0.85 in the trial set, a value which we use across all languages.

Table 4 lists the F1 percentage performance of different systems on the five datasets of the SemEval-2013 all-words WSD task. Despite not being tuned to the task, our representations provide competitive results on all datasets, outperforming the sophisticated Babelfy system on the Spanish and German languages. The variant of our system not utilizing the MFS information in the disambiguation process ($\theta = 0$), i.e., MUFFIN⋆, also shows competitive results, outperforming the best system in the SemEval-2013 dataset on all languages. Interestingly, MUFFIN⋆ proves highly effective on the French language, surpassing not only the performance of our system using the MFS information, but also attaining the best overall performance.

4.2.2 WordNet

As regards the WordNet disambiguation task, we take as our benchmark the two recent SemEval English all-words WSD tasks: the SemEval-2013 task on Multilingual WSD (Navigli et al., 2013) and the SemEval-2007 English Lexical Sample, SRL and All-Words task (Pradhan et al., 2007). The all-words datasets of the two tasks contain 1644 instances (SemEval-2013) and 162 noun instances (SemEval-2007), respectively.

As comparison system, we report the performance of the best configuration of the top-performing system in the SemEval-2013 task, i.e., UMCC-DLSI (Gutiérrez et al., 2013). We also show results for the state-of-the-art supervised system (Zhong and Ng, 2010, IMS), as well as for two graph-based approaches that are based on random walks on the WordNet graph (Agirre and Soroa, 2009, UKB w2w) and the BabelNet semantic network (Moro et al., 2014, Babelfy). We follow Babelfy and also exploit the WordNet's sense frequency information from the SemCor sense-annotated corpus (Miller et al., 1993). However, instead of simply backing off to the most frequent sense, we propose a more meaningful exploitation of this information. To this end, we compute the relevance of a specific sense as the average of its normalized sense frequency and its corresponding

System	MFS Back off	Italian	English	French	Spanish	German
MUFFIN	✓	81.9	84.5	71.4	**85.1**	**83.1**
MUFFIN⋆		67.9	73.5	**72.3**	81.1	76.1
Babelfy	✓	**84.3**	**87.4**	71.6	83.8	81.6
Best SemEval 2013 system	✓	58.3	54.8	60.5	58.1	61.0
MFS	-	82.2	80.2	69.1	82.1	83.0

Table 4: F1 percentage performance on the SemEval-2013 Multilingual WSD datasets using Wikipedia as sense inventory.

score ($score_c$ in Algorithm 2) given by our system. The sense with the highest overall relevance value is then picked as the intended sense.

Additionally, we put forward a hybrid system that combines our system with IMS, hence benefiting from the judgements made by two systems that utilize complementary information. Our system makes judgements based on global contexts, whereas IMS exploits the local context of the target word. To this end, we compute the relevance of a specific sense as the average of the normalized scores given by IMS and our system ($score_c$ in Algorithm 2). We refer to this hybrid system as MUFFIN+IMS.

Table 5 reports the F1 percentage performance of different systems on the datasets of SemEval-2013 and SemEval-2007 English all-words WSD tasks. We also report the results for the MFS baseline, which always picks the most frequent sense of a word. Similarly to the disambiguation task on the Wikipedia sense inventory, MUFFIN proves to be quite competitive on the WordNet disambiguation task, while surpassing the performance of all the comparison systems on the SemEval-2013 dataset. On the SemEval-2007 dataset, IMS achieves the best performance, thanks to its usage of large amounts of manually and semi-automatically tagged data. Finally, our hybrid system, MUFFIN+IMS, provides the best overall performance on the two datasets, showing that our combination of the two WSD systems that utilize different types of knowledge was beneficial.

5 Related work

We briefly review the recent literature on the two NLP tasks to which we applied our representations, i.e., Word Sense Disambiguation and semantic similarity.

WSD. There are two main categories of WSD techniques: knowledge-based and supervised

System	SemEval-2013	SemEval-2007
MUFFIN	**66.0**	66.0
UKB	61.3	56.0
UMCC-DLSI	64.7	–
IMS	65.3	**67.3**
Babelfy	65.9	62.7
MFS	63.2	65.8
MUFFIN+IMS	66.9	68.5

Table 5: F1 percentage performance on the SemEval-2013 and SemEval-2007 (noun instances) English All-words WSD datatets using WordNet as sense inventory.

(Navigli, 2009). Supervised systems such as IMS (Zhong and Ng, 2010) analyze sense-annotated data and model the context in which the various senses of a word usually appear. Despite their accuracy for the words that are provided with suitable amounts of sense-annotated data, their applicability is limited to those words and languages for which such data is available, practically limiting them to a small subset of words mainly in the English language. Knowledge-based approaches (Sinha and Mihalcea, 2007; Navigli and Lapata, 2007; Agirre and Soroa, 2009) significantly improve the coverage of supervised systems. However, similarly to their supervised counterparts, knowledge-based techniques are usually limited to the English language.

Recent years have seen a growing interest in cross-lingual and multilingual WSD (Lefever and Hoste, 2010; Lefever and Hoste, 2013; Navigli et al., 2013). Multilinguality is usually offered by methods that exploit the structural information of large-scale multilingual lexical resources such as Wikipedia (Gutiérrez et al., 2013; Manion and Sainudiin, 2013; Hovy et al., 2013). Babelfy (Moro et al., 2014) is an approach with state-of-the-art performance that relies on random walks

on BabelNet multilingual semantic network (Navigli and Ponzetto, 2012a) and densest subgraph heuristics. However, the approach is limited to the WSD and Entity Linking tasks. In contrast, our approach is global as it can be used in different NLP tasks, including WSD.

Semantic similarity. Semantic similarity of word pairs is usually computed either on the basis of the structural properties of lexical databases and thesauri, or by comparing vectorial representations of words learned from massive text corpora. Structural approaches usually measure the similarity on the basis of the distance information on semantic networks, such as WordNet (Budanitsky and Hirst, 2006), or thesauri, such as Roget's (Morris and Hirst, 1991; Jarmasz and Szpakowicz, 2003). The semantic network of WordNet has also been used in more sophisticated techniques such as those based on random graph walks (Ramage et al., 2009; Pilehvar et al., 2013), or coupled with the complementary knowledge from Wikipedia (Camacho-Collados et al., 2015). However, these techniques are either limited in the languages to which they can be applied, or in their applicability to tasks other than semantic similarity (Navigli and Ponzetto, 2012b).

Corpus-based techniques are more flexible, enabling the training of models on corpora other than English. However, these approaches, either in their conventional co-occurrence based form (Gabrilovich and Markovitch, 2007; Landauer and Dumais, 1997; Turney and Pantel, 2010; Bullinaria and Levy, 2012), or the more recent predictive models (Mikolov et al., 2013; Collobert and Weston, 2008; Pennington et al., 2014), are restricted in two ways: (1) they cannot be used to compare word senses; and (2) they cannot be directly applied to cross-lingual semantic similarity. Though the first problem has been solved by multi-prototype models (Huang et al., 2012), or by the sense-specific representations obtained as a result of exploiting WordNet glosses (Chen et al., 2014), the second problem remains unaddressed. In contrast, our approach models word senses and concepts effectively, while providing a unified representation for different languages that enables cross-lingual semantic similarity.

6 Conclusions

This paper presented MUFFIN, a new multilingual, unified and flexible representation of individual word senses. Thanks to its effective combination of distributional statistics and structured knowledge, the approach can compute efficient representations of arbitrary word senses, with high coverage and irrespective of their language. We evaluated our representations on two different NLP tasks, i.e., semantic similarity and Word Sense Disambiguation, reporting state-of-the-art performance on several datasets. Experimental results demonstrated the reliability of our unified representation approach, while at the same time also highlighting its main advantages: multilinguality, owing to its effective application within and across multiple languages; and flexibility, owing to its robust performance on two different tasks.

Acknowledgments

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of EACL*, pages 33–41.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL*, pages 19–27.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

Mokhtar-Boumeyden Billami, José Camacho-Collados, Evelyne Jacquey, and Laurence Kister. 2014. Semantic annotation and terminology validation in full scientific articles in social sciences and humanities (annotation sémantique et validation terminologique en texte intégral en shs) [in french]. In *Proceedings of TALN 2014*, pages 363–376.

Samuel Brody and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *Proceedings of EACL*, pages 103–111.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-

occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

José Camacho-Collados, Mokhtar Billami, Evelyne Jacquey, and Laurence Kister. 2014. Approche statistique pour le filtrage terminologique des occurrences de candidats termes en texte intégral. In *JADT*, pages 121–133.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pages 1025–1035.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, pages 1606–1615.

Lev Finkelstein, Gabrilovich Evgeniy, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, and Ruppin Eytan. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–1611.

Roger Granada, Cassia Trojahn, and Renata Vieira. 2014. Comparing semantic relatedness between word pairs in Portuguese using Wikipedia. In *Computational Processing of the Portuguese Language*, pages 170–175.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of IJCNLP*, pages 767–778.

Yoan Gutiérrez, Yenier Castañeda, Andy González, Rainel Estrada, D. Dennys Piug, I. Jose Abreu, Roger Pérez, Antonio Fernández Orquín, Andrés Montoyo, Rafael Muñoz, and Franc Camara. 2013. UMCC_DLSI: Reinforcing a ranking algorithm with sense frequencies and multidimensional semantic resources to solve multilingual word sense disambiguation. In *Proceedings of SemEval 2013*, pages 241–249.

Zellig Harris. 1954. Distributional structure. *Word*, 10:146–162.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of EMNLP*, pages 1192–1201.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of AAAI*, pages 884,889.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882.

Thad Hughes and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of EMNLP-CoNLL*, pages 581–589.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of RANLP*, pages 212–219.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Colette Joubarne and Diana Inkpen. 2011. Comparison of semantic similarity for different languages using the Google n-gram corpus and second-order co-occurrence measures. In *Advances in Artificial Intelligence*, pages 216–221.

Alistair Kennedy and Graeme Hirst. 2012. Measuring semantic relatedness across languages. In *Proceedings of xLiTe: Cross-Lingual Technologies Workshop at the Neural Information Processing Systems Conference*.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1:127–165.

Tom Landauer and Scott Dooley. 2002. Latent semantic analysis: theory, method and application. In *Proceedings of CSCL*, pages 742–743.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211.

Ludovic Lebart, A Salem, and Lisette Berry. 1998. *Exploring textual data*. Kluwer Academic Publishers.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of SemEval 2010*, pages 82–87, Uppsala, Sweden.

Els Lefever and Veronique Hoste. 2013. SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. In *Proceedings of SemEval 2013*, pages 158–166, Atlanta, USA.

Steve L. Manion and Raazesh Sainudiin. 2013. Daebak!: Peripheral diversity for multilingual Word Sense Disambiguation. In *Proceedings of SemEval 2013*, pages 250–254.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

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, Plainsboro, N.J.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1).

Roberto Navigli and Mirella Lapata. 2007. Graph connectivity measures for unsupervised Word Sense Disambiguation. In *Proceedings of IJCAI*, pages 1683–1688.

Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli and Simone Paolo Ponzetto. 2012b. BabelRelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of AAAI*, pages 108–114.

Roberto Navigli and Paola Velardi. 2005. Structural Semantic Interconnections: a knowledge-based approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1088.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proceedings of SemEval 2013*, pages 222–231.

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

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of KDD*, pages 613–619.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of ACL*, pages 468–478.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351.

Simone Paolo Ponzetto and Michael Strube. 2007. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research (JAIR)*, 30:181–212.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of SemEval*, pages 87–92.

Daniel Ramage, Anna N. Rafferty, and Christopher D. Manning. 2009. Random walks for text semantic similarity. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 23–31.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of ACL*, pages 109–117.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Gerard Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Ravi Sinha and Rada Mihalcea. 2007. Unsupervised graph-based Word Sense Disambiguation using measures of word semantic similarity. In *Proceedings of ICSC*, pages 363–369.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, and Aitor Soroa. 2009. WikiWalk: random walks on Wikipedia for semantic relatedness. In *Proceedings of the Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49.

Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A wide-coverage Word Sense Disambiguation system for free text. In *Proceedings of the ACL System Demonstrations*, pages 78–83.