CluBERT: A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages

Tommaso Pasini, Federico Scozzafava, and **Bianca Scarlini**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{pasini,scozzafava,scarlini}@di.uniroma1.it

Abstract

Knowing the Most Frequent Sense (MFS) of a word has been proved to help Word Sense Disambiguation (WSD) models significantly. However, the scarcity of sense-annotated data makes it difficult to induce a reliable and high-coverage distribution of the meanings in a language vocabulary. To address this issue, in this paper we present CluBERT, an automatic and multilingual approach for inducing the distributions of word senses from a corpus of raw sentences. Our experiments show that CluBERT learns distributions over English senses that are of higher quality than those extracted by alternative approaches. When used to induce the MFS of a lemma, CluBERT attains state-of-the-art results on the English Word Sense Disambiguation tasks and helps to improve the disambiguation performance of two off-the-shelf WSD models. Moreover, our distributions also prove to be effective in other languages, beating all their alternatives for computing the MFS on the multilingual WSD tasks. We release our sense distributions in five different languages at https://github.com/SapienzaNLP/clubert.

1 Introduction

Word Sense Disambiguation (WSD) is the task of associating a word in context with a meaning from a given inventory of senses (Navigli, 2009). It resides at the core of Natural Language Processing and has been proved to be beneficial to different downstream tasks, e.g., Information Extraction (Delli Bovi et al., 2015) and Machine Translation (Pu et al., 2018). Current approaches to WSD can mainly be divided into supervised and knowledge-based methods. While the former leverage manually-annotated data to train statistical models, the latter exploit the knowledge enclosed within a semantic network to identify the most appropriate meaning of a word in context. Both kinds of approach, however, suffer from the *knowledge acquisition bottleneck* problem (Gale et al., 1992; Pasini, 2020). In fact, since words and senses follow a Zipfian distribution (McCarthy et al., 2004a), information on rare words and meanings is scarce in both semantically-annotated data and knowledge bases. This undermines the ability of supervised and knowledge-based approaches to deal with words unseen at training time, or that have only a few connections within a semantic network. To overcome this limitation, the Most Frequent Sense (MFS) backoff strategy, i.e., tagging a word with its meaning that has been manually annotated as the most frequent one, is employed by both approaches. Nevertheless, while the MFS proved to be a strong baseline in the general-domain setting of WSD, it does not scale over specific domains (Pasini and Navigli, 2020) and its applicability is limited to languages where annotated data are available, i.e., English. Furthermore, the way words and meanings are used changes over time, hence making old annotations unreliable. This is the case with WordNet (Miller et al., 1990), i.e., the most used electronic English dictionary in WSD. WordNet provides information about sense frequency that is either manually-annotated or derived from SemCor (Miller et al., 1993), i.e., a corpus where words are manually tagged with WordNet meanings. However, neither WordNet nor SemCor have been updated in the past 10 years, thus making their information about sense frequency outdated. For example, the WordNet most frequent sense for the noun *pipe* is its *smoking device* meaning, although, nowadays, one would expect the *metal pipe* sense to appear more often in general.

To overcome some of the aforementioned limitations, different approaches to automatically extracting the distribution of senses have been proposed (Pasini and Navigli, 2018; Hauer et al., 2019). However, these fail to match the WordNet MFS

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4008–4018

performance and are either dependent on bilingual corpora (Hauer et al., 2019), or limited to nouns only (Pasini and Navigli, 2018).

In this paper, we present CluBERT, a multilingual cluster-based approach that automatically induces the distribution of word senses from a corpus of raw sentences without relying on manually-annotated data. By exploiting the assumption that similar meanings appear in similar contexts (Reif et al., 2019) and the representational power of BERT (Devlin et al., 2019), CluBERT can learn distributions that are of better quality – according to both intrinsic and extrinsic evaluation – than those extracted either by its competitors, or from manually-curated resources. Furthermore, our approach outperforms its alternatives in all multilingual and most domain-specific WSD test sets. Finally, when used as backoff strategy of a WSD architecture, our automatically-induced distributions are shown to lead the underlying model to higher results than when using the standard manually-curated distributions of WordNet, hence placing themselves as a better and more flexible alternative.

2 Related Work

Word Sense Disambiguation (WSD) is a long-standing problem in Natural Language Processing which was first formulated to address the ambiguity of words in the context of Machine Translation (Weaver, 1949). Nowadays, WSD models can be mainly divided in two groups: knowledge-based and supervised. Knowledge-based methods (Agirre et al., 2014; Moro et al., 2014; Tripodi and Pelillo, 2015) rely on the information enclosed within a semantic network such as WordNet (Miller et al., 1990), a manually-curated resource organised in a graph structure where nodes are concepts and edges are semantic relations between them, or BabelNet (Navigli and Ponzetto, 2010, 2012), a large multilingual knowledge base where synsets are lexicalised in more than 250 languages. Since knowledge-based approaches do not rely on semantically-annotated corpora, they can easily scale over different languages as long as their underlying semantic network supports them (Scarlini et al., 2020; Maru et al., 2019; Scozzafava et al., 2020). Nevertheless, these approaches struggle to remain competitive on English when compared to supervised methods.

Supervised approaches, instead, take advantage of sense-annotated data and frame the WSD task as a classification problem, where each word has its own set of labels, i.e., its possible meanings according to a given sense inventory. Ranging from word-based approaches, where a single SVM classifier is specialised in disambiguating only one word in a sentence (Zhong and Ng, 2010; Iacobacci et al., 2016; Yuan et al., 2016), to more general neural architectures that classify all the words together (Raganato et al., 2017a; Vial et al., 2019; Hadiwinoto et al., 2019; Bevilacqua and Navigli, 2020), supervised methods have proved to outperform their knowledge-based counterparts whenever annotated data are available (Scarlini et al., 2019).

Despite the progress and the increment in the overall performance, both kinds of approach still rely, most of the time, on the Most Frequent Sense heuristic whenever a word does not appear tagged in the training set, or the confidence score of its disambiguation is lower than a threshold. The MFS baseline, in fact, has proved to be very competitive (McCarthy et al., 2004a), yet, it is limited to words and senses comprised in a manually-annotated corpus such as SemCor (Miller et al., 1993). To cope with this limitation, several works have been proposed over the years to automatically learn the Most Frequent Sense of a word. A seminal work in this direction was that of McCarthy et al. (2004b), where a thesaurus and the distributional similarity between words were used to find the predominant meaning of a given lemma. More recent works, instead, have focused on inducing the full distribution over the senses of a given word. Bennett et al. (2016) exploited topic modelling techniques, whereas Pasini and Navigli (2018) presented two multilingual approaches that provided full distributions over nominal senses, not only for English, but also for words in other languages.

The work we propose in this paper stands out from previous approaches, exploiting for the first time, to the best of our knowledge, BERT contextualized embeddings together with a knowledge-based WSD model to compute the distribution of word meanings. Our approach is not tied to any specific language and can potentially be applied to all languages supported by both BERT (104) and BabelNet (more than 280).

3 CluBERT

In this Section, we present CluBERT, a multilingual approach for computing the distribution of

CLUSTER 1
The working of **glass** requires lower temperatures.
Vitrinite has a shiny appearance resembling **glass**.
Most of the roof and walls are made out of **glass**.
CLUSTER 2
He asked for a **glass** of water.
It is traditionally served in a **glass**.
He gave him a poison **glass** to drink from.

Table 1: Excerpt of the sentences of two clusters of the noun $glass_n$.

CLUSTER 1		CLUSTER 2	
material$_n$	✓	water$_n$	✓
metal$_n$	✓	wine$_n$	✓
plastic$_n$	✓	drink$_v$	✓
heat$_n$	✗	yellow$_a$	✗
crystal$_n$	✗	thick$_a$	✗

Table 2: Excerpt of the most frequent words (top part) and excluded words (bottom part) for two different clusters of the noun $glass_n$.

word senses from a corpus of raw sentences. Our approach takes as input a corpus $\mathcal{C}$ and a target lexeme l[1] and exploits BERT[2], i.e., a pretrained language model, and BabelNet, i.e., a multilingual knowledge base. We also define the set of possible meanings M_l for the lexeme l as the set of all the synsets[3], i.e., sets of synonyms, in BabelNet which have l among their lexicalizations. CluBERT extracts the sense distribution for l by applying the following three steps:

1. **Sentence Clustering**, which clusters together the sentences of $\mathcal{C}$ in which l appears based on the similarity of their contexts[4].
2. **Cluster Disambiguation**, which assigns to each cluster a distribution over the possible meanings of l in BabelNet by exploiting the context provided by the cluster itself.
3. **Distribution Extraction**, which, given the distributions computed in the previous step, finally derives the general distribution of the senses of l across the corpus C.

3.1 Sentence Clustering

The first step relies on the assumption that different senses of l tend to appear in different contexts and vice versa. Therefore, since BERT has been shown to capture the subtle distinctions between different meanings of the same word (Reif et al., 2019), we employ it to compute the representations of l across different sentences. We thus cluster BERT embeddings in order to group together the occurrences of l that appear in similar contexts and are hence likely to express the same meaning. More in detail, we iterate over all the sentences in $\mathcal{S}_l \subset \mathcal{C}$, i.e., those sentences in $\mathcal{C}$ where l appears, and project them in a latent space by means of BERT. We thereby represent the sentence $\sigma \in \mathcal{S}_l$ as $v^l_\sigma = BERT(\sigma, l)$, i.e., the representation of l in the sentence σ computed by BERT.

Once all the sentences in $\mathcal{S}_l$ are associated with a vector, we group contextually-similar sentences together by leveraging the k-means algorithm (Lloyd, 2006). K-means, in fact, creates internally-cohesive clusters that partition $\mathcal{S}_l$ into k disjoint groups. For example, in Table 1 we show an excerpt of two clusters extracted for the lexeme $glass_n$[5]. As one can see, the sentences in each set identify the semantics of the target word, with the upper cluster grouping all sentences related to the *material* meaning of $glass_n$ and the bottom one all those related to its *container* sense. We note that no induction of senses is performed at any stage of our approach.

At the end of this step, the target lexeme l is associated with the set of its clusters $\mathcal{U}_l$.

3.2 Cluster Disambiguation

The second step computes, for each cluster c of the lexeme l, a distribution over the possible senses of l that is specific to c. To this end, by exploiting the lexical context of c, we build its weighted Bag-of-Words representation and use it to compute the cluster-level distribution over the senses in M_l.

BoW construction We are now interested in finding which of the senses of l best suits the context provided by the sentences in c. To this end, we extract the Bag of Words of c BoW_c by considering all the content words in c. BoW_c, in fact, conflates

[1] A lemma with a specific Part-Of-Speech tag.

[2] Across all the experiments we used the multilingual model of BERT, i.e., bert-base-multilingual-cased.

[3] We use sense and synset interchangeably.

[4] As representation for a sentence containing l we use the contextualized representation of l.

[5] We use the lemma$_{POS}$ notation.

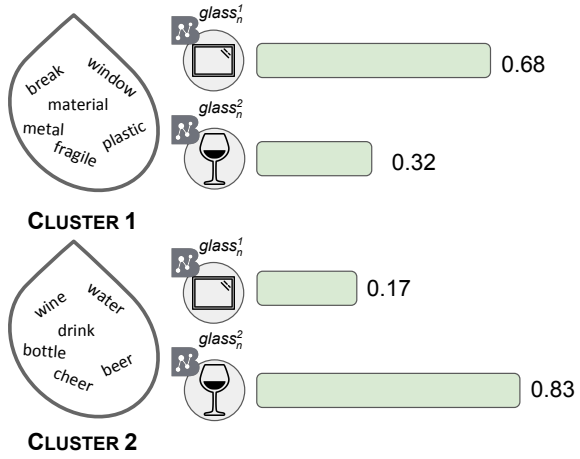

Figure 1: Cluster-level sense distributions for the two clusters of $glass_n$ over its possible meanings in the reference knowledge base.

the contextual information of all the sentences in c in a list of unique words ranked by their frequency within the cluster. We refine BoW_c by retaining only its top n most frequent words, hence filtering out those that are less informative for determining the most suitable meaning of l in c and the stopwords. To showcase the outcome of this step, in Table 2 we report the three most frequent words in the BoW for two clusters of $glass_n$ (top part) along with two excluded words (bottom part). As one can see, the topmost words provide a precise characterization of the semantics of the clusters.

Cluster-Level Sense Distribution We now proceed by computing the probability of l expressing a given sense $s \in M_l$ within a cluster c. To this end, we rank the synsets of l according to their relevance in the BabelNet semantic network with respect to a given set of nodes $M_{BoW_c} = \bigcup_{l' \in BoW_c} M_{l'}$, i.e., the set of all the possible meanings of the words in BoW_c. Thus, we follow Agirre et al. (2014) and employ the PageRank algorithm in its personalised version (Haveliwala et al., 2002, PPR), which computes the probability of reaching a node in the graph when starting from a fixed set of nodes. Formally, we calculate the score of each synset in BabelNet as follows:

$$v^{(t+1)} = (1 - \alpha)v^{(0)} + \alpha A v^{(t)}$$

where A is the row-normalised adjacency matrix of the knowledge base, $v^{(0)}$ is the restart probability distribution, which is zero in every component except for those corresponding to the nodes in M_{BoW_c}, and α is the well-known damping factor which we set to 0.85. We further exploit the contexts in BoW_c by weighting each synset $s \in M_{BoW_c}$ by the sum of the frequencies of its lexicalizations that appear in BoW_c. Finally, after n iterations of the PPR algorithm, we extract the scores for each $s \in M_l$ from v^n and normalise them to build the cluster-level sense distribution d_l^c for the lemma l in the cluster c. As shown in Figure 1, the two clusters of $glass_n$ are now associated with two different distributions over $glass_n$' meanings in BabelNet, i.e., the *container* sense and the *material* sense.

3.3 Distribution Extraction

In this last step, we compute the overall sense distribution of l with respect to the input corpus $\mathcal{C}$. To this end, we leverage the cluster-level distributions and the clusters' sizes to compute the overall distribution over the senses of l as follows:

$$d_l = \frac{\sum_{c \in \mathcal{U}_l} |c| \, d_l^c}{\sum_{c \in \mathcal{U}_l} |c|}$$

where d_l^c is the vector representing the distribution over l's synsets in the cluster c and $\mathcal{U}_l$ is the set of clusters of l. For example, considering the clusters depicted in Figure 1 and their distributions[6], we associate the lexeme $glass_n$ with the distribution $d_{glass_n} = \{glass_n^1 : 0.34, glass_n^2 : 0.66\}$ where $glass_n^1$ is the sense 1 of $glass_n$ in BabelNet.

We repeat these steps for each lemma of interest to derive the distribution over its senses in BabelNet.

4 Experimental Setup

We now present a battery of experiments to assess the quality of our induced sense distributions on both intrinsic and extrinsic evaluation tasks. First, we set the parameters of the model, namely, the sense inventory, the corpus, the number of words to retain in each Bag of Words, and the number of clusters to create for each lemma. Then, we evaluate our automatically-induced distributions intrinsically, by computing their distance in comparison to a manually-annotated distribution, and extrinsically, on the standard English and multilingual Word Sense Disambiguation tasks.

System Parameters As sense inventory, we use all the synsets in BabelNet that also contain a sense from WordNet. Concerning the corpus, we use

[6] We consider $|\text{CLUSTER}_1| = 50$ and $|\text{CLUSTER}_2| = 100$.

Wikipedia[7] since it is freely available and covers more than 300 languages and most of the semantic domains. As regards the number of clusters for a given lemma l, we set the parameter k of the k-means algorithm to the number of l's meanings in BabelNet. Finally, we tune the number of words n to retain within each cluster's Bag of Words by manually evaluating the quality of the disambiguation step (see Section 3.2) when varying n between 5 and 20 with a 5 step and set $n = 5$.

We compute the distributions for all the lemmas in English, Italian, Spanish, French and German which have at least one corresponding synset within the sense inventory.

Comparison Systems We compare CluBERT with the most recent and best-performing automatic and manual approaches for sense-distribution learning and MFS detection. As regards the automatic methods for inducing sense distributions, we consider the two knowledge-based and multilingual approaches proposed by Pasini and Navigli (2018), i.e., EnDi and DaD, and the topic modelling-based approach proposed by Bennett et al. (2016), i.e., LexSemTM. We also compare against three other approaches specialised in identifying the MFS of a word, namely, COMP2SENSE (Hauer et al., 2019), which exploits the distance between a word and a sense in a knowledge base, and WCT-VEC (Hauer et al., 2019) and UMFS-WE (Bhingardive et al., 2015), which, instead, leverage the distance between words and sense embeddings.

As for the manually-annotated competitors, we compare against the sense distributions and the MFS of WordNet (Miller et al., 1990). These are both determined by the frequency of the senses in SemCor (Miller et al., 1993), when possible, and by manual annotations of the synsets' ranks, otherwise.

Concerning the multilingual evaluation, instead, we compare CluBERT with EnDi, DaD and the BabelNet MFS, which computes the MFS for a given lemma by taking its highest ranked sense according to BabelNet.

5 Intrinsic Evaluation

In this Section we estimate the quality of our automatically-induced sense distributions by comparing them to gold standard ones. We use the dataset proposed by Bennett et al. (2016) which, contains 50 distinct lemmas annotated with a gold distribution over their senses. Hence, we compare the distributions for the target lemmas induced by CluBERT and its competitors with the manually-annotated ones.

5.1 Evaluation Measures

In order to compare two distributions, we use two measures: the Jensen-Shannon divergence (JSD) and the Weighted Overlap (WO) (Pilehvar et al., 2013). With both metrics, we average all the pairwise similarity between the gold distributions and the ones induced by the systems under comparison.

Jensen-Shannon Divergence The JSD computes a real value expressing the similarity between the two input distributions, which is 0 when they are identical, and higher than 0 otherwise. Formally, given two input distributions d and d', the Jensen-Shannon divergence is defined as follows:

$$JSD(d, d') = \frac{\mathcal{D}(d, M) + \mathcal{D}(d', M)}{2}$$

$$\mathcal{D}(d, d') = \sum_s d(s) log \left(\frac{d(s)}{d'(s)} \right)$$

where $M = \frac{d+d'}{2}$ and $\mathcal{D}$ is the Kullback-Leibler divergence function in which $d(s)$ is the value of the component corresponding to the synset s in the distribution d.

Weighted Overlap The WO measure computes the similarity between two input distributions by harmonically averaging the ranks of the distributions' components when sorted according to their probabilities. Its output value is 1 when the two inputs are identical, and 0 otherwise. Formally, let d and d' be two input distributions, their Weighted Overlap is computed as follows:

$$WO(d, d') = \sum_{i=1}^{|O|} \frac{(r_i + r'_i)^{-1}}{(2i)^{-1}}$$

where O is the set of common components between the input distributions and r_i and r'_i are the ranks of the i-th component in d and d', respectively.

5.2 Results

We now report the results of CluBERT and its competitors in terms of JSD and WO in comparison to the gold distributions provided by Bennett et al. (2016). As one can see from Table 3, CluBERT

[7] We used the June 2019 dump.

Method	JSD (↓)	WO (↑)
CluBERT	**0.085**	**0.958**
EnDi	0.099	0.937
DaD	0.204	0.902
LexSemTM	0.116	0.932
WordNet	0.255	0.837

Table 3: Similarity scores on the Bennett et al. (2016) gold standard in terms of JSD (the lower the better) and Weighted Overlap (the higher the better). Statistically-significant differences between CluBERT and EnDi are underlined.

is the approach that better resembles the human-annotated distributions, in terms of both JSD and WO, achieving 0.085 and 0.958, respectively, and outperforming the previous state of the art on this dataset, i.e., EnDi. Interestingly enough, WordNet is the worst approach across the board scoring more than 0.1 worse than CluBERT on both evaluation measures. We attribute these modest results to the fact that WordNet draws its distribution from annotations that are not up to date. Furthermore, we note that CluBERT results are statistically-significant ($p < 0.1$) when compared to the best competitor systems, i.e., EnDi, on both evaluation measures.

5.3 Error Analysis

By manually inspecting the induced distributions that were most different to the gold ones, we note that the vast majority of CluBERT errors are due to the lack of senses for named entities in our inventory. Indeed, many nouns that are commonly associated with objects or abstract meanings are also used for named entities, e.g., the lexeme *flora*$_n$, which is commonly used to indicate either *the living organism* meaning, or *the plant life of a region* meaning, it is often used in compound nouns used to refer to named entities, such as *F.C. Flora*[8], *William Flora*[9], etc. These occurrences are therefore considered by CluBERT, which, despite being able to cluster them correctly, fails to disambiguate the group containing named entities owing to the fact that the correct meaning is not available within the sense inventory. As a result, most of the clusters where *flora*$_n$ appears as named entity are disambiguated with the *living organism* meaning, thereby contributing to wrongly steering the sense distribution towards this meaning.

Since most of the errors are of this kind, better handling of named entities or the use of a larger sense inventory could further improve the performance of CluBERT.

6 Extrinsic Evaluation

In this Section we evaluate CluBERT's distributions on the English, domain-specific and multilingual all-words WSD tasks. To this end, we leverage the sense distributions to extract a lemma's Most Frequent Sense (MFS), which is then used to annotate each occurrence of the lemma in the test sets. In addition, we also integrate CluBERT MFS into two off-the-shelf WSD models and measure its impact.

Evaluation Datasets We consider all the standard English all-words WSD test sets contained in the framework presented by Raganato et al. (2017b), i.e., Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), SemEval-2015 (Moro and Navigli, 2015) and ALL, i.e., the concatenation of all the previous datasets. As regards the domain-specific evaluation we consider the 6 and 3 domains in SemEval-2013 and SemEval-2015, respectively, and test on each of them separately. As for the multilingual evaluation, instead, we test on the Italian, Spanish, French and German datasets of SemEval-2013 and the Italian and Spanish test sets of SemEval-2015.

We note that both datasets make use of old versions of BabelNet (version 1.1.1 and 2.5, respectively). For this reason, previous works used an in-house mapping between BabelNet versions to make them up to date. However, in this process, several gold instances were lost making the datasets smaller than the original ones. To be fair with other approaches, we compare CluBERT against them on the same datasets on which they tested. Moreover, to encourage future comparisons, we also report CluBERT's performance on the newer versions of both gold standards made available by the Sapienza NLP group at https://github.com/SapienzaNLP/mwsd-datasets, which comprise more instances than the older datasets and feature the latest version of BabelNet (4.0.1)[10]. As

[8] https://en.wikipedia.org/wiki/FC_Flora

[9] https://en.wikipedia.org/wiki/William_Flora

[10] We used the WordNet split as we can only provide senses within the WordNet part of BabelNet.

Method	Senseval2	Senseval3	SemEval-2007	SemEval-2013	SemEval-2015	All
CluBERT	**68.3**	64.6	**55.4**	**69.7**	**68.0**	**66.8**
UMFS-WE	54.8	52.0	38.2	55.2	54.5	53.1
WCT-VEC	56.4	53.8	40.6	54.9	54.0	54.1
COMP2SENSE	51.5	47.0	37.5	54.2	55.0	50.7
WordNet MFS	67.0	**66.0**	55.0	63.0	**68.0**	65.0

Table 4: MFS performance in terms of F1 on all the instances of the test sets in Raganato et al. (2017b). Statistically-significant differences on the ALL dataset between CluBERT and WordNet MFS are underlined.

Method	Precision	Recall	F1
CluBERT	**70.9**	**70.2**	**70.6**
EnDi	66.0	66.0	66.0
DaD	61.0	61.0	61.0
LexSemTM	51.0	48.0	49.0
WordNet MFS	68.0	68.0	68.0

Table 5: MFS performance in terms of Precision, Recall and F1 on the nominal instances of the ALL test set from Raganato et al. (2017b).

a term of comparison, we also report the results of the BabelNet MFS on these datasets. In what follows, we refer to the older versions of the multilingual tasks of SemEval-2013 and SemEval-2015 by juxtaposing the "*" symbol (SemEval-2013* and SemEval-2015*).

On all the aforementioned datasets we report the results in terms of F1, i.e., the harmonic mean of precision and recall.

Most Frequent Sense Strategy We extract the MFS of a target lemma l from its sense distribution d_l by taking the synset with the highest probability, i.e., $MFS(l) = argmax(d_l)$. Therefore, we use the MFS of a lemma computed according to each system under comparison to tag all of l's occurrences within the test sets.

Domain-Specific WSD Setup To assess the ability of CluBERT to scale over different domains and hence to extract a distribution that is skewed towards the topic of the input corpus, we build 8 distinct domain-specific corpora, one for each domain of SemEval-2013 and SemEval-2015's English datasets. For this purpose, we exploit the 34 domain labels (Camacho-Collados and Navigli, 2017) available in BabelNet together with the mapping between synsets and Wikipedia pages to retrieve those pages that are peculiar to a specific domain, hence building a corpus $\mathcal{C}_{dom}$ specific for the domain dom. We then apply CluBERT, EnDi, DaD and LexSemTM on $\mathcal{C}_{dom}$ and extract their respective MFS specific for each domain[11].

Downstream Task Setup Finally, we test the benefits brought by CluBERT's distributions by including them in a knowledge-based and a supervised approach, namely:

- **UKB**[12] (Agirre et al., 2014): an off-the-shelf state-of-the-art knowledge-based WSD model based on the Personalised PageRank algorithm. When provided, it makes use of the given sense distribution to bias its answers towards the MFS.

- **BiLSTM** (Raganato et al., 2017a): an end-to-end neural sequence model which employs two bidirectional LSTM layers and an attention mechanism trained on multiple tasks, i.e., fine- and coarse-grained WSD and Part-of-Speech tagging. When provided, it makes use of the MFS backoff strategy whenever it comes to disambiguating a lemma unseen during training.

We compare these two models, firstly, when no prior knowledge is supplied, and then, when WordNet (UKB$_{WN}$, BiLSTM$_{WN}$) and CluBERT (UKB$_{CluBERT}$, BiLSTM$_{CluBERT}$) distributions are provided.

6.1 English WSD Results

As one can see from Table 4, CluBERT attains the highest scores across the board, outperforming all the other automatic approaches by more than 10 F1 points. More interestingly, CluBERT surpasses the hitherto unbeaten manual baseline of WordNet by

[11] We do not compare against UMFS-WE, WCT-VEC and COMP2SENSE inasmuch as code and data are not available.

[12] Version 3.2 available at http://ixa2.si.ehu.es/ukb/

Method	SemEval-2013						SemEval-2015		
	Biology	Climate	Finance	Politics	Social Issue	Sport	Math&Computer	Biomedicine	Social Issue
CluBERT	72.9	**70.9**	**69.0**	**79.2**	**70.9**	**61.4**	52.3	**77.3**	**75.2**
DaD	**79.0**	63.0	64.0	67.0	68.0	54.0	59.8	63.9	54.3
EnDi	71.0	53.0	60.0	62.0	63.0	57.0	**63.0**	63.0	55.9
LexSemTM	56.0	47.0	49.0	51.0	52.0	34.0	47.7	63.0	40.7
WordNet MFS	61.0	59.0	52.0	64.0	58.0	56.0	47.2	67.8	62.4

Table 6: MFS performance in terms of F1 on the nominal instances of the different domains in the SemEval-2013 (Navigli et al., 2013) and SemEval-2015 test sets (Moro and Navigli, 2015).

Method	SemEval-2013*				SemEval-2015*	
	IT	ES	DE	FR	IT	ES
CluBERT	**71.7**	**68.7**	**69.1**	**67.1**	**70.4**	**68.8**
DaD	62.9	58.9	65.5	54.3	61.0	58.0
EnDi	46.2	44.6	49.1	54.3	55.0	52.0
BabelNet MFS	52.3	55.6	49.3	55.1	52.0	53.0

Table 7: MFS F1 scores on the nominal instances of the SemEval-2013* and SemEval-2015* multilingual datasets. Statistically significant differences (at χ^2 test) with $p < 0.01$ between CluBERT and the second best performing model are underlined.

a statistically-significant[13] difference (McNemar, 1947) of almost 2 F1 points on the ALL dataset. In order to set a level playing field with EnDi and DaD, which cover nouns only, we also carried out our evaluation on the ALL dataset focusing on its nominal instances. As shown in Table 5, CluBERT attains an F1 score of 70.6, surpassing the best automatic competitor, i.e., DaD, by more than 4 F1 points. More importantly, our induced distributions also outperform the well-known WordNet MFS strategy by 2.6 F1 points in this setting too.

This demonstrates that CluBERT's distributions are of higher quality than those induced by any of the other automatic and manual competitors.

6.2 Domain-Specific WSD Results

We now focus on testing our distributions on the domain-specific documents available in the SemEval-2013 and SemEval-2015 WSD test sets. As shown in Table 6, CluBERT outperforms all the other competitors on 7 out of the 9 domains by several points, falling behind DaD on the Biology domain and behind EnDi on the Math&Computer one. This is mainly due to the fact that the senses in these two domains are poorly connected in BabelNet, hence making them hard to reach when applying the PPR algorithm (see Section 3.2). DaD, which also exploits the BabelNet graph, seems to be more robust to this event inasmuch as it relies directly on the connections between domains and synsets and not only on those between words and concepts, as CluBERT does. Nevertheless, when the senses of the target domain are well framed within the semantic network, our approach proves to be able to induce a distribution that accurately reflects the way the meanings of a word are spread within the input corpus. In fact, CluBERT achieves the best results on all the other domains, with the highest improvement of 12.2 F1 points over the current state of the art on the Politics domain of SemEval-2013.

WordNet, instead, shows poor performance in this setting, too. In fact, its MFS information is designed to work on a general domain setting and it cannot be customised easily for other scenarios. All these results further corroborate our findings in the intrinsic evaluation, and they highlight the fact that WordNet distributions no longer reflect the way senses are spread across a corpus.

6.3 Multilingual WSD Results

We now investigate the capabilities of CluBERT to scale over different languages by evaluating it on the multilingual Word Sense Disambiguation tasks of SemEval-2013* and SemEval-2015*. As can be seen from Table 7, the differences in results between CluBERT and the other systems under comparison remain consistent with those reported for English. Our approach, in fact, achieves on average a significant improvement of approximately 9 F1 points over the existing state of the art. This demonstrates that CluBERT makes efficient use of its two complementary resources, i.e., BabelNet and BERT, in this way making up for the paucity of data in non-English languages. Conversely, EnDi and DaD suffer from this shortcoming and perform either poorly (EnDi), or not consistently across languages (DaD). As for the performance on the newer versions of the datasets (Table 8), we note

[13] χ^2 test for statistical significance with $p < 0.05$.

	SemEval-2013				SemEval-2015	
Method	IT	ES	DE	FR	IT	ES
CluBERT	66.6	69.5	72.3	62.3	62.8	61.5
BabelNet MFS	53.2	60.3	76.6	60.0	54.2	50.1

Table 8: MFS F1 scores on all instances of the WordNet split of SemEval-2013 and SemEval-2015 multilingual datasets mapped to the latest BabelNet version (4.0.1). Data available at https://github.com/SapienzaNLP/mwsd-datasets.

Method	Precision	Recall	F1
UKB	63.1	63.1	63.1
UKB$_{WN}$	67.1	67.1	67.1
UKB$_{CluBERT}$	67.7	67.7	**67.7**
BiLSTM	68.1	61.6	64.7
BiLSTM$_{WN}$	69.6	69.6	69.6
BiLSTM$_{CluBERT}$	69.9	69.9	**69.9**

Table 9: UKB and BiLSTM Precision, Recall and F1 with and without the MFS backoff strategy on the ALL test set in Raganato et al. (2017b).

that CluBERT outperforms the BabelNet MFS on all languages but German. The drop in performance on SemEval-2015 when compared to the older version of the dataset, is mainly due to the fact that the datasets now also include all the non-nominal instances which were excluded before to be fair with the other competitors. As for future comparisons, we highly encourage the community to consider the results in Table 8 for CluBERT as they are computed on larger and more updated versions of the datasets.

6.4 Downstream Task Results

Finally, we assess CluBERT MFS effectiveness when used as backoff strategy in two off-the-shelf WSD approaches, i.e., UKB and the BiLSTM with attention model presented by Raganato et al. (2017b) (see Section 6). In Table 9 we report the performance of the two models without MFS, with WordNet MFS and with CluBERT MFS on the ALL WSD dataset. As one can see, not only does our MFS provide a large boost of 4.6 and 5.2 F1 points when compared with the base models without backoff strategy, but it also leads the two systems to attain better performance than when using the WordNet MFS. This strengthens our previous findings and crowns CluBERT as the best backoff strategy compared to all its alternatives.

These results open up to new scenarios where the CluBERT MFS might be preferred as backoff strategy for WSD models to the well-established WordNet MFS. In fact, CluBERT attains higher results than WordNet on several WSD datasets, while at the same time assuring greater flexibility. In fact, whereas WordNet MFS is static, CluBERT can be run on different corpora and can therefore adapt the sense distributions to various circumstances and different languages.

7 Conclusions

In this paper we presented CluBERT, an automatic multilingual approach which induces the distribution of word senses in an arbitrary input corpus by exploiting the contextual information coming from BERT and the lexical-semantic knowledge available in BabelNet. CluBERT attains state-of-the-art results on both intrinsic and extrinsic evaluations, also beating the widely-used and manually-curated WordNet MFS.

When considering input corpora that come from specific domains, CluBERT showed an unmatched nimbleness in shaping the distributions accordingly, hence outperforming its manual and automatic competitors on most domains. Similarly, our approach demonstrated its ability to scale well on different languages, attaining state-of-the-art results on the multilingual WSD tasks. Finally, when injecting CluBERT MFS into off-the-shelf WSD models, we showed that it brings greater benefits than the WordNet MFS. We release the sense distributions in five different languages at https://github.com/SapienzaNLP/clubert.

As future work, we plan to refine our approach by exploiting other strategies for weighting the words in the clusters and to leverage them for automatically building multilingual sense-tagged corpora.

Acknowledgments

The authors wish to greatly thank Claudio Delli Bovi for his comments, suggestions and late-night discussions on the manuscript.

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Andrew Bennett, Timothy Baldwin, Jey Han Lau, Diana McCarthy, and Francis Bond. 2016. LexSemTm: A Semantic Dataset Based on All-words Unsupervised Sense Distribution Learning. In *Proc. of ACL*, pages 1513 – 1524.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In *Proc. of ACL*.

Sudha Bhingardive, Dhirendra Singh, V Rudramurthy, Hanumant Redkar, and Pushpak Bhattacharyya. 2015. Unsupervised Most Frequent Sense Detection using Word Embeddings. In *Proc. of NAACL*, pages 1238–1243.

Jose Camacho-Collados and Roberto Navigli. 2017. BabelDomains: Large-Scale Domain Labeling of Lexical Resources. In *Proc. of ACL*, volume 2, pages 223–228.

Claudio Delli Bovi, Luis Espinosa Anke, and Roberto Navigli. 2015. Knowledge Base Unification via Sense Embeddings and Disambiguation. In *Proc. of EMNLP*, pages 726–736.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: overview. In *Proc. of Senseval-2*, pages 1–5.

William A. Gale, Kenneth Church, and David Yarowsky. 1992. A method for disambiguating word senses in a corpus. *Computers and the Humanities*, 26:415–439.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. In *Proc. of EMNLP-IJCNLP*, pages 5296–5305.

Bradley Hauer, Yixing Luan, and Grzegorz Kondrak. 2019. You Shall Know the Most Frequent Sense by the Company it Keeps. In *Proc. of ICSC*, pages 208–215.

T. Haveliwala, A. Gionis D. Klein, and P. Indyk. 2002. Evaluating Strategies for Similarity Search on the Web. In *Proc. of WWW*, pages 432–442.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Sisambiguation: An Evaluation Etudy. In *Proc. of ACL*, volume 1, pages 897–907.

S. Lloyd. 2006. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theor.*, 28(2):129–137.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In *Proc of EMNLP-IJCNLP*, pages 3525–3531.

Diana McCarthy, Rob Koeling, and Julie Weeds. 2004a. Ranking WordNet senses automatically. *Recall*, 40:60.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004b. Finding predominant senses in untagged text. In *Proc. of ACL*, pages 280–287.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A Semantic Concordance. In *Proc. of DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, N.J.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proc. of SemEval-2015*, pages 288–297.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244.

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

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proc. of SemEval-2013*, volume 2, pages 222–231.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proc. of ACL*, pages 216–225.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Tommaso Pasini. 2020. The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation. In *Proc. of IJCAI*. International Joint Conferences on Artificial Intelligence Organization.

Tommaso Pasini and Roberto Navigli. 2018. Two Knowledge-based Methods for High-Performance Sense Distribution Learning. In *Proc. of AAAI*, pages 5374–5381.

Tommaso Pasini and Roberto Navigli. 2020. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103215.

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

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

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating Weakly Supervised Word Sense Disambiguation Into Neural Machine Translation. *TACL*, 6:635–649.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proc. of EMNLP*, pages 1156–1167.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proc. of EACL*, pages 99–110.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Proc. of NeurIPS*, pages 8592–8600.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just OneSeC for Producing Multilingual Sense-Annotated Data. In *Proc. of ACL*, volume 1, page 699709.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proc. of AAAI*.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with Syntagmatic Information-for Multilingual Word Sense Disambiguation. In *Proc. of ACL*.

Benjamin Snyder and Martha Palmer. 2004. The English all-Words Task. In *Proc. of Senseval-3*, pages 41–43.

Rocco Tripodi and Marcello Pelillo. 2015. WSD-games: A Game-Theoretic Algorithm for Unsupervised Word Sense Disambiguation. In *Proc. of SemEval*, pages 329–334.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proc. of Global Wordnet Conference*.

Warren Weaver. 1949. Translation. In *Machine Translation of Languages: Fourteen Essays*, pages 15–23.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-Supervised Word Sense Disambiguation with Neural Models. *Proceedings of COLING*, pages 1374–1385.

Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proc. of ACL*, pages 78–83.