Word Sense Disambiguation Based on Iterative Activation Spreading with Contextual Embeddings for Sense Matching

Arkadiusz Janz and Maciej Piasecki
Wrocław University of Science and Technology
{arkadiusz.janz|maciej.piasecki}@pwr.edu.pl

Abstract

Many knowledge-based solutions were proposed to solve Word Sense Disambiguation (WSD) problem with limited annotated resources. Such WSD algorithms are able to cover very large sense repositories, but still being outperformed by supervised ones on benchmark data. In this paper, we start with analysis identifying key properties and issues in application of spreading activation algorithms in knowledge-based WSD, e.g. influence of the network local structures, interaction with context information and sense frequency. Taking our observations as a point of departure, we introduce a novel solution with new context-to-sense matching using BERT embeddings, iterative parallel spreading activation function and selective sense alignment using contextual BERT embeddings. The proposed solution obtains performance beyond the state-of-the-art for the contemporary knowledge-based WSD approaches for both English and Polish data.

1 Introduction

Contextual neural embeddings have strongly influenced Word Sense Disambiguation (henceforth WSD), and resulted in extraordinary improvement on benchmark WSD datasets. However, the vast majority of such approaches follow the supervised learning scheme. Thus, they suffer from the lack of annotated data, especially sparse for WSD, and their coverage, i.e. practical applicability, is limited only to a subset of word senses. Moreover, they express bias towards most frequent senses.

Many knowledge-based solutions (i.e. weakly supervised) were proposed to solve the problem of limited sense annotated corpora. They are able to cover very large sense repositories, but still being outperformed by supervised ones on benchmark data. Knowledge-based WSD algorithms were initially based on spreading activation scheme, most on Personalised PageRank algorithm (PPR) (Agirre and Soroa, 2009). PageRank (Brin and Page, 1998) was originally proposed for modelling the Web, a highly connected network with many hubs and loops.As we show in Figure 1, PPR scores are often strangely biased by some local wordnet structures. Knowledge-based WSD approaches interact with the contextual information in a rather shallow way and also are biased by sense frequencies. That is why, we wanted to develop a version spreading activation for WSD which better reflects wordnet structures and more deeply explores context representation by using contextual text embeddings. Our goal was to develop a novel knowledge-based WSD algorithm which combines context-to-sense matching informed by BERT embeddings (Devlin et al., 2019) with a new iterative parallel spreading activation to process the wordnet.

The main contributions of our paper are:

1. a novel iterative parallel spreading activation algorithm for knowledge-based WSD,
2. enhancing spreading activation with context-to-sense matching using BERT embeddings,
3. and promotion of activations that are more central or salient for the given context.

The proposed solution expressed performance beyond the state-of-the-art of the knowledge-based WSD approaches in the all-words tasks for English. In addition, we performed also tests on WSD test data for Polish, a language that is significantly different from English, equiped with a very large wordnet – *plWordNet* (Dziob et al., 2019). Our solution showed superior performance in comparison to the previous approaches on the Polish data.

2 Related Work

Lesk-like (Lesk, 1986) methods use information about wordnet graph structures to a very little extent, e.g. (Banerjee and Pedersen, 2003; Navigli and Ponzetto, 2012), while local subgraphs are

the primary tool for sense description, distinguishing senses in a wordnet, cf (Maziarz et al., 2013). The idea of better exploration of wordnet graphs for WSD appeared in several works, e.g. in (Mihalcea et al., 2004). (Agirre and Soroa, 2009) proposed Personalised PageRank (PPR) algorithm which uses the Princeton WordNet graph (Fellbaum, 1998) with the initial activation limited to nodes (synsets) correspond to the words from a textual context. The initial activation depends on contextual word frequencies. PPR became the core part of the UKB WSD system (Agirre et al., 2014) with WordNet enhanced by several semantic resources, including sense links derived from Princeton WordNet Gloss Corpus (Wor, 2021). UKB refers to sense frequency twice: in initial activation values (normalised together with the word frequency in the context) and finally as a kind of weights to the synset scores. UKB is freely available, but is sensitive to proper setting and selection of knowledge resources. (Agirre et al., 2018) showed that UKB if properly used is still a state-of-the-art knowledge-based WSD system. UKB achieves the best results in a mode called "W2W" (word-to-word), in which the WSD is restarted for each text word separately. This results in several times slower processing, than the standard mode in which all words in the context are disambiguated in one go.

Babelfy (Moro et al., 2014) utilised spreading activation in a indirect way. It was entirely based on BabelNet (Navigli and Ponzetto, 2012) – a complex semantic resource originating from the automated merging WordNet and Wikipedia[1]. Due to the BabelNet content, Babelfy was able to disambiguate words and perform Entity Linking at the same time. Semantic signatures introduce some generalisation, and the extraction of a "dense subgraph" results in a kind of topic related clustering.

(Scozzafava et al., 2020) applied PPR on WordNet structure, but significantly expanded it with SyntagNet (Maru et al., 2019) – a large, manually constructed resource of semantic sense collocations. The main limitation of WordNet-based spreading activation is the lack of topical relations between senses. Adding SemCor-derived sense links and connections from wordnet glosses partially resolves this issue. On the other hand the structure of the network becomes more complex. Some solutions solve approach the problem by joining *topic modelling* with knowledge-based techniques. For instance, (Wang et al., 2020) collected a corpus related to words from the WSD test data sets and obtained Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003). The graph was expanded with eXtended WordNet. Finally PPR was applied to the graph, where the initialisation was informed by the similarity of a node to the context. This complex approach requires construction of semantic resources focused on the datasets to be disambiguated.

(Chaplot and Salakhutdinov) adapted LDA topic modelling to represent a text document as derived from synsets (modelled by synset probabilities) and synsets as corresponding to word probability distributions. Prior synset distribution was constrained by wordnet-based synset similarity. This approach depends on the knowledge base to a minimal extent, but it is not clear how it can be expanded to more complex and richer knowledge bases.

The idea of restricting disambiguation context to sense semantically related to the disambiguated words is also central for (O et al., 2018). The authors proposed a distributional representation of senses based on generating pseudo-documents from BabelNet. For each sense, other sense nodes in short distance are retrieved and paths linking senses of the same lemma are searched for and used as pseudo-sentences of pseudo-documents for lemmas, next transformed by Doc2vec (Le and Mikolov, 2014) into a distributional space. Local disambiguation graphs are built sequentially from related sense nodes and next processed by PPR.

Local disambiguation graphs of related senses indirectly address topical homogeneity of word senses in a broader context. (Tripodi and Pelillo, 2017) perceived the WSD problem as a constraint satisfaction problem and model it on the basis of evolutionary game theory. Influence of words on senses by other words is weighted by distributional information. Semantic similarity information is used to calculate the amount of compatibility among the selected senses. Sense frequency from SemCor is indirectly used during disambiguation.

Concerning Polish language, the early work was built upon PageRank-based solutions such as UKB and its variations focused on wordnet expansions (Piasecki et al., 2016; Janz and Piasecki, 2019a,b). However, the frequency distribution of senses was unrepresentative as the large-scale sense-annotated corpora were not available.

[1] https://en.wikipedia.org/

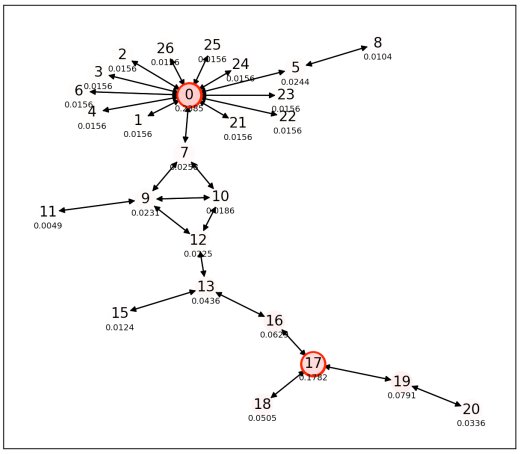

Figure 1: PPR scores computed for artificial data. The graph has been initialized with two seed nodes (v_0, v_{17}). The local structures have a great impact on score distribution.

3 PageRank in Knowledge-based WSD

A wordnet is a mixture of synset and sense relations: some of them directed, other not. Many directed relations exist in one way, but other are in symmetric pairs. Density of a wordnet graph is very diversified – it is not a densely connected graph in general, and many regions may be very sparse. For WSD, a wordnet is typically transformed into a graph with all relations being directed and symmetric and defined on the synset level, e.g. in UKB.

A wordnet has generally a tree-like shape. Its sparseness and shape change after adding links from external resources, e.g. sense links from the gloss corpus. Nevertheless, the final expanded graph for WSD still is not as recursively connected as the Web for which PageRank was designed.

In the case of WSD we assume that from sense nodes activated for a given context activation should evenly spread along the links to senses that are likely to co-occur with them. Activation amount passed to the next nodes should depend mainly on their semantic distance correlated with the number and types of links to be traversed. However, PageRank, modelling of a random walker, seems to work according to a different philosophy. In Fig.1 we have visualised activation scores (below the nodes) obtained with PPR in a simple graph resembling some local wordnet subgraph. The two initial nodes end with the very different final scores. Moreover, we can observe increased activations of nodes located in the same distance from both seeds. This is in contradiction to our above assumptions and results from the recursive character of PPR, e.g. the v_0 high degree and its star-shape local subgraph influence the scoring. In wordnet-based networks such specific local subgraphs, e.g. hub nodes, result from hierarchical categorisation or network editing practices, and do not express node importance due to being 'cited'.

Some of post-processing PPR scoring may decrease such negative bias. However, we decided to completely change the way in which spreading activation is performed. In the following section we propose a non-recursive spreading scheme in which every source node independently broadcast activation that is gradually transmitted along the network paths. It is the path characteristic that matters for activation strength. The final activation of the nodes is the result of overlapping of independent waves crossing the network.

4 Fast Spreading Activation and Contextual Matching

Learning from the PPR analysis, but also literature, two aspects seem especially important for knowledge-based WSD. First, spreading activation should transmit support from contextually related senses to the senses of a disambiguated word. Second, not all context words are equally informative for WSD – a good measure for contextual informativeness is needed. Thus, we proposed a redesigned WSD process based on three main components:

1. Use of contextual embeddings (a neural language model) to express similarity of senses and the context.

2. Iterative, parallel spreading of sense support across the network.

3. Identification of contextually salient senses as markers of context semantic dimensions.

A knowledge-base is a graph $G = (V, E)$, where the vertices V are senses, and the edges E – semantic links encoded in an adjacency matrix $\boldsymbol{A}^{N\times N}$: $\boldsymbol{A}_{s_n,s_{n'}} = 1$ if $(s_n, s_{s'}) \in E$, otherwise 0.

A typical WSD knowledge graph is quite sparse. Several fast graph traversal algorithms (Yang et al., 2015) were proposed for sparse adjacency matrices. A simple sparse–matrix dense–vector or even sparse–matrix sparse–vector multiplications can be interpreted as a single traversal step over graph. This property was used in parallel versions of well known graph algorithms e.g. Breadth First Search (BFS) and quick graph traversal algorithms using

GPUs (Gilbert et al., 2006). More specifically, to design a parallel BFS with multiple independent searches one can use sparse–matrix sparse–matrix multiplication (SpMSpM) where the second matrix represents an initial seed of starting nodes. In this work we adapt SpMSpM to design a fast spreading activation algorithm for WSD.

4.1 Parallel Spreading Activation with Contextual Sense Matching

We define spreading activation as a sequence of SpMSpM steps. The process starts from a set of initial seed nodes $T : (t_1, t_2, ..., t_M)$ with sense specific activation weights $\mathbf{w} = (w_1, w_2, ..., w_M)$. The decay factor d dampens the impact of initial nodes on their neighborhood in propagation procedure. In SpMSpM graph traversal framework the seed nodes are encoded as sparse matrix $P^{N\times M}$ using one-hot encoding where $\boldsymbol{P}_{n,m} = 1, n \in [1, N]$, $m \in [1, M]$ if $s_n \in T$ and $s_n = t_m$. SpMSpM allows us to quickly compute consecutive steps of graph traversal process starting independently from different seed nodes.

$$\boldsymbol{P}' = \boldsymbol{AP} \tag{1}$$

As we will show in the next section, with a single multiplication we can generate the output $Q^{N\times M}$ and select M columns from the adjacency matrix A in the first step. Each column $Q_{\star,m}$ represents in fact a set of visited nodes reached from the initial node t_m. Thus, we obtain independent outputs for every single seed node separately.

Parallel Spreading Activation defined in this way is an iterative process. We can easily reuse the outputs of the first multiplication step to generate the K new traversals starting from them.

$$\begin{aligned} \boldsymbol{P}^{(0)}_{n,m} &= \begin{cases} 1, & \text{if } s_n \in T \wedge s_n = t_m \\ 0, & \text{otherwise} \end{cases} \\ \boldsymbol{P}^{(k)} &= \boldsymbol{AP}^{(k-1)},\ k = 0, 1, \ldots, K \end{aligned} \tag{2}$$

To enforce that the consecutive P matrices encode only 0 and 1 where 1 represent visited nodes in k-th traversal step, we apply a simple clipping step using sign function, if non-negative values are greater than one. The clipping should counteract the results of multiple matrix multiplications. Since the multiplication steps accumulate visited nodes in the P matrices we also add subtraction step to ensure that the k-th output matrix contains only newly visited nodes. The final traversal matrix Q for the k-th step with only newly visited nodes is:

$$\begin{aligned} \widetilde{\boldsymbol{P}}^{(k)} &= \text{sign}(\boldsymbol{P}^{(k)}) \\ \boldsymbol{Q}^{(0)} &= \widetilde{\boldsymbol{P}}^{(0)} \\ \boldsymbol{Q}^{(1)} &= \max\{0, \widetilde{\boldsymbol{P}}^{(1)} - \widetilde{\boldsymbol{P}}^{(0)}\} \\ \boldsymbol{Q}^{(k)} &= \max\{0, \widetilde{\boldsymbol{P}}^{(k)} - \widetilde{\boldsymbol{P}}^{(k-1)} - \widetilde{\boldsymbol{P}}^{(k-2)}\} \end{aligned} \tag{3}$$

The values of $\boldsymbol{P}$ matrices are manipulated in a way that allows us to prevent backward traversals since $\boldsymbol{AP}$ multiplication does not prevent it. Only two-step subtractions are necessary to completely exclude them from the traversal procedure. The matrix $\boldsymbol{Q}^{(k)}_{n,m} = \max\{0, \widetilde{\boldsymbol{P}}^{(k)}_{n,m} - \widetilde{\boldsymbol{P}}^{(k-1)}_{n,m} - \widetilde{\boldsymbol{P}}^{(k-2)}_{n,m}\}$ contains 1 if a node s_n has been discovered starting from node t_m, otherwise 0.

By repeating this process we obtain a sequence of $(\boldsymbol{Q}^{(1)}, \boldsymbol{Q}^{(2)}, ..., \boldsymbol{Q}^{(K)})$ matrices where K is the total number of traversal steps. The final matrix $\boldsymbol{R}^{N\times M}$ represents accumulated activations computed M times for all nodes in the graph starting from each initial seed node $t_m \in T$ independently.

$$\boldsymbol{R} = \Big(\sum_{k=0}^{K} d^k \boldsymbol{Q}^{(k)}\Big) \tag{4}$$

Contextual Sense Matching The activation scores resulting from the parallel spreading represent support coming from different input nodes. We could immediately combine different activations coming to a node into one scoring value, but signals coming from different input senses may be of different informativeness for the context and a word to be disambiguated. To effectively disambiguate words in the context we need to incorporate only the most relevant signals. To do this, we introduce below a context-sensitive weighting $\mathbf{w}$ for activations coming from different seed nodes.

Two strategies can be applied to compute a contextually sensitive scoring from raw activations. The first one mixes all coming in activations with a dot product of the $\boldsymbol{R}$ rows and weight factors. The second is focused only on the most informative activations by applying maximum function.

$$\boldsymbol{z}_{s_n} = \boldsymbol{R}_{n,\star} \cdot \mathbf{w} \tag{5}$$

$$\widetilde{\boldsymbol{z}}_{s_n} = \max\{\boldsymbol{R}_{n,\star} \odot \mathbf{w}\} \tag{6}$$

$$\tag{7}$$

To implement *Word-to-Word*-like mode of WSD (W2W), known from UKB, we can use a binary masking matrix $\boldsymbol{U}^{N\times M}$ excluding all senses of the same lemma from weight factors $\mathbf{w}$ and traversal-based scoring function z_{s_n}. The output of masking $\boldsymbol{R}'$ can be obtained by applying Hadamard product of the $\boldsymbol{R}$ matrix entries with $\boldsymbol{U}$:

$$\begin{aligned}\boldsymbol{R}' &= \boldsymbol{R} \odot \boldsymbol{U}\\ \boldsymbol{z}'_{s_n} &= \boldsymbol{R}'_{n,\star} \cdot \mathbf{w}\\ \widetilde{\boldsymbol{z}}'_{s_n} &= \max\{\boldsymbol{R}'_{n,\star} \odot \mathbf{w}\}\end{aligned}$$

On-path logit tracking Knowledge-bases are noisy, just like the text data. The lexico-semantic structure of wordnet and its extensions is non-uniform which implies one can find some areas that might be semantically incoherent. On the other hand, lexico-semantic structure usually extends beyond statistical disambiguation context. Additional supervision might be disastrous to model's generalisation ability and decrease its performance on unseen senses. However, by measuring semantic coherence of traversal paths one can reduce underlying noise and *filter* out unnecessary signals reaching target nodes representing disambiguated senses. For this reason, we propose an *on-path-logit-tracking* mechanism such that it uses sense embeddings (see section 4.3) of the intermediate nodes on traversal paths by utilising traversal matrices $(\boldsymbol{Q}^{(1)}, \boldsymbol{Q}^{(2)}, ..., \boldsymbol{Q}^{(K)})$ context-dependent scores. For a given seed node t_m, we analyse the nodes visited on its paths encoded by a sequence $(\boldsymbol{Q}^{(1)}_{*,m}, \boldsymbol{Q}^{(2)}_{*,m}, \dots \boldsymbol{Q}^{(K)}_{*,m})$ of columns in traversal matrices $\boldsymbol{Q}$. Let $(H_m^{(1)}, H_m^{(2)}, \dots, H_m^{(K)})$ represent the sets of visited nodes discovered in each traversal step, starting from the node t_m. We compute a score representing a degree of contextual matching between the seed node t_m and the disambiguated word w_c, by measuring the contextual match of the embeddings of nodes visited during the traversal and contextual embedding of disambiguated word.

$$\begin{aligned}\mathcal{G}\left(H_m^{(k)}\right) &= \max_{v\in H_m^{(k)}} \mathcal{G}'(v)\\ \mathcal{G}'(v) &= \max\left\{e(v)\cdot e(w_c|C),\, e(v)\cdot e(w_{c'}|C)\right\}\end{aligned}$$

where w_c and $w_{c'}$ represent, respectively, the disambiguated word itself and another content word in the given disambiguation context. This allows us to incorporate out-of-context senses existing in a wordnet into the final score. The procedure is computed for every seed node from the disambiguation context. We use the scores $\mathcal{G}\left(H_m^{(k)}\right)$ as a replacement for plain reduction model from Equation 4 and plug them into Equation 5. Before we will finally describe the disambiguation model in Sec. 4.4, we introduce the contextual embedding models used for generating weight factors and similarities in contextual sense matching and on-path logit tracking procedures.

4.2 Sense Encoder

To encode wordnet senses for contextual sense selection we use pre-trained BERT model in a similar way to (Du et al., 2019). We modified this architecture by dropping additional MLP layers as our approach is not supervised, see Fig. 2. Sense vector space is generated as follows: for each synset $s_n \in V$ its definition and examples are obtained from Princeton WordNet Gloss Corpus and BERT embeddings are generated for all their tokens. As BERT uses its own tokenizer based on WordPiece (Wu et al., 2016) words are segmented into subtokens and from the sequence of subtoken embeddings we generate a synset embedding by averaging only the embeddings of subtokens being a part of the synset's lemmas in its context (definition or example, see Figure 2). If a synset s_n has both a definition and an example, we generate separate contextual embeddings $e_d(s_n)$ and $e_s(s_n)$ and average them into a single synset embedding $e(s_n)$.

4.3 Context Encoder

Context size is a significant factor in WSD. To verify the impact of this factor on WSD performance we decided to test two context generation methods. First, the context generation heuristic from the UKB implementation (Agirre et al., 2018) takes 30 distinct content words around a target word to be disambiguated. It assumes that the location of target words is known in advance and a specific number of content words can be pre-selected to form the disambiguation context. As we rely on BERT embeddings in our method, not a bag of content words as, e.g. in UKB, we take a sequence of sentences the words belong to as a context. To convert the context to its vector space representation we apply BERT model on concatenation of input sentences and we store output token embeddings for further usage during disambiguation process.

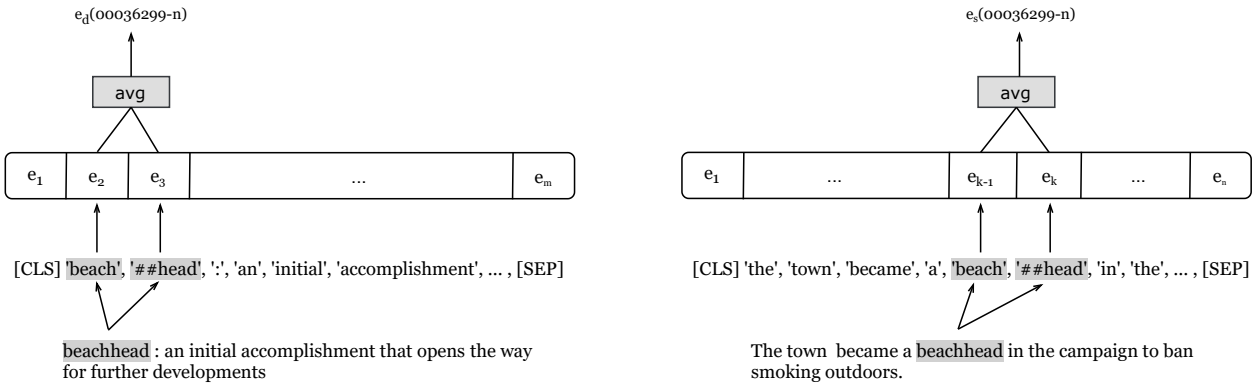

Figure 2: Transformer-based sense encoder used to encode wordnet senses. We modified the architecture proposed by (Du et al., 2019) and adapted it to our setting by dropping MLP layers originally being used to tune the model for supervised setting.

The second method uses a sliding window of three sentences and disambiguating content words in the middle one. In this case, we use BERT to obtain contextual embeddings of all content words. As both context and sense encoders work in the same vector space the fit between sense candidates and context words corresponds to vector similarity. With the context embedding model we generate contextual word embeddings $e(w_c|C)$ for each content word $w_c \in C$ being a part of input context C using the same sub-token embedding averaging model, see Fig. 2).

4.4 Sense Selection

Spreading activation scores are an input to the word sense selection. For the seed senses T extracted from the context C we initialise weight factors $\mathbf{w}$ for weighting the outputs of spreading activation function. Sense-specific activations in $\mathbf{w}$, cf Sec. 4.1, are based on cosine similarity $\text{sim}(e(w_c|C), e(t_m))$ between the embeddings of the input seed senses $t_m \in T$ (Sec. 4.2) and the contextual embedding of the disambiguated word w_c (Sec. 4.3), where $t_m \notin S(w_c)$.

Word-to-word masking The nodes $t_m \in S(w_c)$ representing the senses of the lemma w_c are excluded. Technically, the obtained $\mathbf{w}$ factors are directly plugged into scoring function $\widetilde{z}'_{s_n} = \max\{\boldsymbol{R}'_{n,\star} \odot \mathbf{w}\}$ where we take the maximum and the masking matrix U simulates W2W behaviour (see Sec. 4.1). The $\widetilde{z}'_{s_n}$ values are used as the first factor in our disambiguation function.

Disambiguation The disambiguation model merges two aforementioned factors. The first factor z_s is based on spreading activation with contextual sense matching and on-path logit tracking. The second factor g_s is computed simply as a dot product between a candidate sense embedding and the contextual embedding of a disambiguated word. For a set of sense candidates $s \in S(w_c)$ of the disambiguated word w_c, the final score is:

$$score(s) = \frac{z'_s + g_s}{2}$$

$$\hat{s} = \underset{s \in S(w_c)}{\text{argmax}}\{score(s)\} \quad (8)$$

This model does not use word or sense frequencies yet. However, we can easily include them by changing the weight factors $\mathbf{w}$ or by multiplying the output scoring o_{s_n} with the frequency factor of a specific sense. We have chosen the latter.

5 Experiments

In this section we report the setup and the results of our experimental part.

5.1 Setting

We focused on comparison with several other knowledge-based solutions (see Sec.2) as well as analysis of the impact of sense frequency and an underlying knowledge graph on the performance of the proposed method. We tested two knowledge graphs: a graph based on Princeton WordNet expanded with eXtended WordNet (WN), and next it further expanded with syntagmatic links (SGN) as in (Maru et al., 2019; Scozzafava et al., 2020). We also evaluated the proposed model with two different context generation heuristics (see Sec. 4.3).

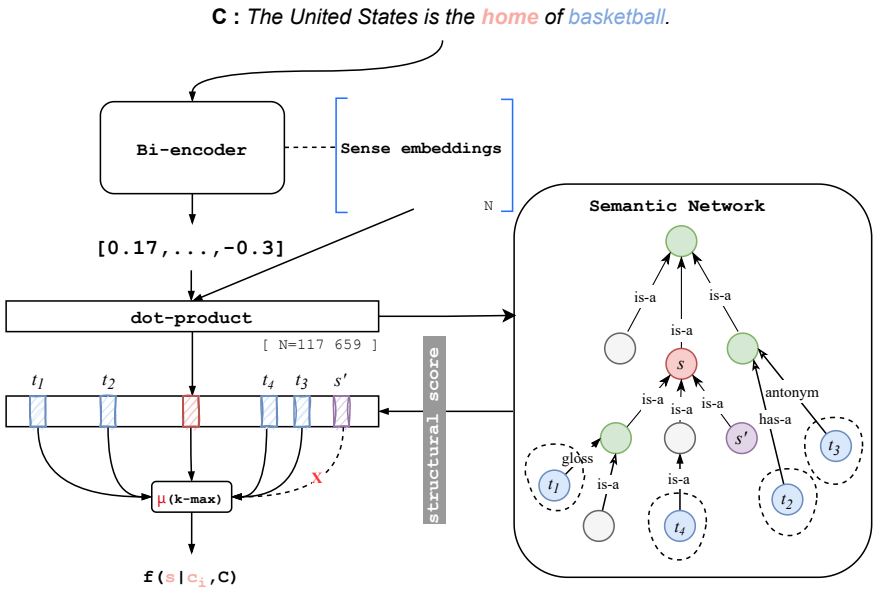

Figure 3: A general view on the proposed WSD model. The spreading activation model is provided with *sense-to-context* similarity scores computed with contextual embeddings. The scores from spreading activation model are combined with candidate sense score, excluding the incoming activations from the nodes representing disambiguated word (UKB's *word-to-word* heuristic). To avoid biasing towards frequent senses, we use *k-max* selection ($k = 3$) of incoming activation scores.

As a text encoder we used a pre-trained BERT$_{BASE}$ (Devlin et al., 2019) uncased model with hidden 12 layers, 12 attention heads and the hidden layer size of 768. For the Polish dataset we used PolBERT uncased model which is pre-trained on Polish corpora and has the same BERT$_{BASE}$[2] architecture. For Polish data we present only the results of the model without sense frequency factor since a large sense-annotated corpora do not exist for the Polish language, so we could not compute frequency scores for senses.

5.2 Evaluation Corpora

The performance of our method was measured using English all-words WSD framework (Raganato et al., 2017) built upon Senseval-2 (Palmer et al., 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013) and SemEval-2015 (Moro and Navigli, 2015) datasets. To compute performance metrics we used a standard scorer provided with this framework. We also conducted the evaluation on the Polish annotated corpora prepared for PolEval'2020 competition (Ogrodniczuk and Łukasz Kobyliński, 2020) in a similar way.

[2] https://github.com/kldarek/polbert

5.3 Parameter Tuning

To tune the parameters of the activation spreading algorithm and the sense selection function we decided to utilise available wordnet data and glosses from Princeton WordNet Gloss Corpus.

5.4 Results and Discussion

Sense frequency We compare the results from literature with the performance of our method measuring the impact of SemCor-based sense frequencies. The WSD models without prior information of sense frequencies showed lower performance on almost every dataset. Still, our model performed quite well in comparison with PageRank-based model implemented in UKB and WoSeDon, even if we compare the model without prior sense frequency information with the models using this prior during the disambiguation (see Table 1).

Knowledge graph We can also notice that the knowledge graph itself has a great impact on WSD performance. The methods proposed in the literature usually utilise eXtended WordNet (e.g. UKB) which introduces additional semantic links extracted from Princeton WordNet Gloss Corpus as a basis for disambiguation process. However, the resources like BabelNet or SyntagNet have been showed to increase the performance even more.

Table 1: F1-scores computed for different evaluation datasets in all–words WSD competition. The methods using sense frequencies from SemCor (SF) were marked with ✓symbol. We also mentioned the knowledge bases used in cited methods (KB column). The (Wang et al., 2020) approach has used a knowledge base augmented with additional documents retrieved from external corpora (WN†).

Method	**KB**	**SF**	**Test set**					
			S2	S3	S7	S13	S15	All
(Agirre et al., 2018)	WN	✓	68.8	66.1	53.0	68.6	70.3	67.3
(Moro et al., 2014)	BN	?	67.0	63.5	51.6	66.4	70.3	65.5
(Maru et al., 2019)	SGN	✓	71.2	71.6	59.6	72.4	75.6	69.3
(Janz and Piasecki, 2019a)	WN	✓	69.6	66.5	52.8	68.6	70.2	67.7
(Chaplot and Salakhutdinov)	WN	?	69.0	66.9	55.6	65.3	69.6	66.9
(Tripodi and Pelillo, 2017)	BN	?	61.2	59.1	43.3	70.8	–	–
(Scozzafava et al., 2020)	SGN	✓	71.6	72.0	59.3	72.2	75.8	71.7
(Wang et al., 2020)	WN†	?	72.7	71.5	61.5	**76.4**	79.5	73.5
(Wang et al., 2020)*	WN†	?	71.9	69,9	60,5	75,7	79,0	72,5
The proposed model								
Parallel Spreading Activation	WN	✓	72.9	71.0	61.8	74.9	78.9	73.1
Parallel Spreading Activation	X-WN	✓	**75.3**	**72.2**	**63.9**	76.2	**81.0**	**74.8**

Table 2: F1-scores computed for different models on test in Polish language. We used the test data prepared for PolEval's Task 3: All-words WSD competition (Janz et al., 2020).

Method	**Test set**	
	SPEC	KPWr-100
(Kłeczek, 2020)	58.40	59.40
(Janz et al., 2020)	62.28	64.65
Parallel Spreading Activation	**65.79**	**66.12**

In this work we analysed the performance of our model working with SyntagNet knowledge-graph as it appeared to be very effective for WSD. We noticed that the model has obtained the best results among other knowledge-based solutions. We did not test BabelNet, as it is not open and we could not get access to this resource. When we compare the methods based on WordNet+eXtended WordNet, our model has obtained better results than PageRank solutions which suggests that selective approach is indeed more effective.

6 Conclusions

We propose the Parallel Spreading Activation with Contextual Sense Matching (PSA) method for knowledge-based, weakly supervised WSD. Its core is a novel spreading activation algorithm that is based on the idea of iterative spreading of support from the context seed senses across the network. The activation comes to the candidate senses from different directions and can be combined into the final score according to a selected scheme. This spreading scheme seems to fit better to the character of the wordnet-based semantic networks. Moreover, it allows for efficient implementation based on the multiplication of sparse matrices. The contextual sense matching function uses contextual embeddings for more accurate and selective information processing to avoid unnecessary mixing of all input signals from disambiguation context and reduce the impact of knowledge-base imperfections. We showed that two kinds of contextual information, namely informativeness of seed senses for the disambiguated word and association of the seed senses with the semantic dimensions of the context can be introduced into our spreading activation model on the basis of contextual embeddings, in our case we used BERT for this purpose. It is worth to notice that our approach uses versatile, general neural language models, and does not require construction of any further WSD-specific text models. We provide the code and the data at https://gitlab.clarin-pl.eu/knowledge-extraction/prototypes/wsd-psa.

Acknowledgments

The work was partially supported by (1) the Polish Ministry of Education and Science, the CLARIN-PL project (agreement no. 2022/WK/09); (2) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure.

References

[online]. 2021. [link].

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84.

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of Open Source NLP software: UKB is inadvertently state-of-the-art in knowledge-based WSD. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 29–33, Melbourne, Australia. Association for Computational Linguistics.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, page 805–810, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(null):993–1022.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Devendra Singh Chaplot and Ruslan Salakhutdinov. Knowledge-based Word Sense Disambiguation using Topic Models. In *32nd AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, USA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Jiaju Du, Fanchao Qi, and Maosong Sun. 2019. Using BERT for word sense disambiguation. *arXiv preprint arXiv:1909.08358*.

Agnieszka Dziob, Maciej Piasecki, and Ewa K. Rudnicka. 2019. plWordNet 4.1 – a linguistically motivated, corpus-based bilingual resource. In *Proceedings of the Tenth Global Wordnet Conference : July 23-27, 2019, Wrocław (Poland)*, pages 353–362.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

John R Gilbert, Steve Reinhardt, and Viral B Shah. 2006. High-performance graph algorithms from parallel sparse matrices. In *International Workshop on Applied Parallel Computing*, pages 260–269. Springer.

Arkadiusz Janz, Joanna Chlebus, Agnieszka Dziob, and Maciej Piasecki. 2020. Results of the poleval 2020 shared task 3: Word sense disambiguation. *Proceedings of the PolEval 2020 Workshop*, pages 65–77.

Arkadiusz Janz and Maciej Piasecki. 2019a. A Weakly supervised word sense disambiguation for Polish using rich lexical resources. *Poznan Studies in Contemporary Linguistics*, 55(2):339 – 365.

Arkadiusz Janz and Maciej Piasecki. 2019b. Word Sense Disambiguation based on Constrained Random Walks in Linked Semantic Networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 516–525, Varna, Bulgaria. INCOMA Ltd.

Dariusz Kłeczek. 2020. Polbert: Attacking polish nlp tasks with transformers. In *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceeding SIGDOC '86 Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM Press.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3525–3531.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Langauge Resources and Evaluation*, 47(3):769–796.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. PageRank on semantic networks, with application to Word Sense Disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

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

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

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.

Dongsuk O, Sunjae Kwon, Kyungsun Kim, and Youngjoong Ko. 2018. Word Sense Disambiguation Based on Word Similarity Calculation Using Word Vector Representation from a Knowledge-based Graph. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2704–2714, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2020. *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24, Toulouse, France. Association for Computational Linguistics.

Maciej Piasecki, Paweł Kędzia, and Marlena Orlińska. 2016. plWordNet in word sense disambiguation task. In *Proceedings of the 8th Global Wordnet Conference, Bucharest, 27-30 January 2016*, pages 280–289. Global Wordnet Association.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

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

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

Rocco Tripodi and Marcello Pelillo. 2017. A game-theoretic approach to word sense disambiguation. *Computational Linguistics*, 43(1):31–70.

Yinglin Wang, Ming Wang, and Hamido Fujita. 2020. Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems*, 190:105030.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

C. Yang, Y. Wang, and J. D. Owens. 2015. Fast Sparse Matrix and Sparse Vector Multiplication Algorithm on the GPU. In *2015 IEEE International Parallel and Distributed Processing Symposium Workshop*, pages 841–847.