SEW-EMBED at SemEval-2017 Task 2: Language-Independent Concept Representations from a Semantically Enriched Wikipedia

Claudio Delli Bovi and **Alessandro Raganato**
Department of Computer Science
Sapienza University of Rome
{dellibovi,raganato}@di.uniroma1.it

Abstract

This paper describes SEW-EMBED, our language-independent approach to multilingual and cross-lingual semantic word similarity as part of the SemEval-2017 Task 2. We leverage the Wikipedia-based concept representations developed by Raganato et al. (2016), and propose an embedded augmentation of their explicit high-dimensional vectors, which we obtain by plugging in an arbitrary word (or sense) embedding representation, and computing a weighted average in the continuous vector space. We evaluate SEW-EMBED with two different off-the-shelf embedding representations, and report their performances across all monolingual and cross-lingual benchmarks available for the task. Despite its simplicity, especially compared with supervised or overly tuned approaches, SEW-EMBED achieves competitive results in the cross-lingual setting (3rd best result in the global ranking of subtask 2, score 0.56).

1 Introduction

Semantic similarity is a well established research area of Natural Language Processing, concerned with measuring the extent to which two linguistic items are similar (Budanitsky and Hirst, 2006). In particular, word similarity is nowadays a widely used evaluation benchmark for word and sense representations (Turney and Pantel, 2010).

While many classical approaches to word similarity have been limited to the English language (Gabrilovich and Markovitch, 2007; Mihalcea, 2007; Pilehvar et al., 2013; Baroni et al., 2014), a growing interest for multilingual and cross-lingual models is emerging (Hassan and Mihalcea, 2011; Camacho Collados et al., 2016) and it is accompanied by the development of multilingual benchmarks (Gurevych, 2005; Granada et al., 2014; Camacho Collados et al., 2015).

In this respect Wikipedia, as one of the most popular semi-structured resources in the field (Hovy et al., 2013), provides a convenient bridge to multilinguality, with several million inter-language links among articles refferring to the same concept or entity. In fact, a number of successful approaches to semantic similarity make explicit use of Wikipedia, from ESA (Gabrilovich and Markovitch, 2007) to NASARI (Camacho Collados et al., 2016). Others, like SENSEMBED (Iacobacci et al., 2015), report state-of-the-art results when trained on an automatically disambiguated version of a Wikipedia dump. Regardless of whether Wikipedia is seen as a multilingual semantic network of concepts and entities or as a sense-annotated corpus, *hyperlinks* (inter-page links) constitute its key structural property: in light of this, Raganato et al. (2016) addressed the sparsity problem of original hyperlinks and developed SEW[1], a semantically enriched Wikipedia where the overall number of linked mentions has been more than tripled by solely exploiting the structure of Wikipedia itself and the wide-coverage sense inventory of BabelNet (Navigli and Ponzetto, 2012)[2].

In addition to building the corpus, the authors used SEW's sense annotations to construct vector representations of concepts and entities from the BabelNet sense inventory, and tested them on multiple semantic similarity tasks. Being defined at the concept level, SEW's representations are inherently multilingual: however, they consist of high-

[1] http://lcl.uniroma1.it/sew
[2] http://babelnet.org

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 252–257,
Vancouver, Canada, August 3 - 4, 2017.

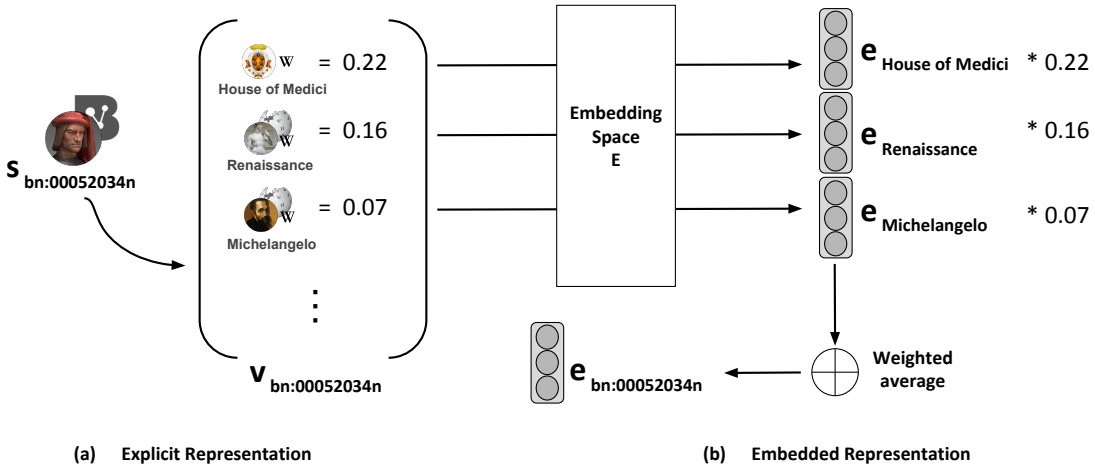

Figure 1: Illustrative example of SEW-EMBED's embedded representation (*b*) for the BabelNet entity *Lorenzo de Medici* (`bn:00052034n`) obtained from the corresponding explicit representation (*a*).

dimensional sparse vectors, not immediately comparable with existing approaches, especially those based on word embeddings, and less flexible to use within downstream applications.

In this paper we propose SEW-EMBED, an embedded augmentation of SEW's original representations in which sparse vectors, defined in the high-dimensional space of Wikipedia pages, are mapped to continuous vector representations via a weighted average of embedded vectors from an arbitrary, pre-specified word (or sense) representation. Regardless of the particular representation used, the resulting vectors are still defined at the concept level, and hence immediately expendable in a multilingual and cross-lingual setting.

We describe and evaluate SEW-EMBED with two off-the-shelf embedded representations: the popular word embeddings of Word2Vec (Mikolov et al., 2013a) and the embedded concept representations of NASARI (Camacho Collados et al., 2016)[3]. We report and discuss the results obtained by both versions on all monolingual and cross-lingual benchmarks available for the task (Camacho Collados et al., 2017), and include a comparison with the original explicit representations of Raganato et al. (2016).

2 Background: Developing a Semantically Enriched Wikipedia

The approach used by Raganato et al. (2016) to develop SEW relies on a cascade of *hyperlink propagation heuristics*, applied to an English Wikipedia dump after some standard pre-processing. In general terms, each propagation heuristic identifies a list of BabelNet synsets to be propagated across a given Wikipedia page p; then, for each synset, occurrences of any of its potential lexicalizations are detected and added as new sense annotations for p. Raganato et al. (2016) distinguishes between intra-page and inter-page heuristics (depending on whether the synsets propagated across p are collected from the same page), but all of them share a common assumption: every occurrence of an ambiguous mention within p refers to the same underlying sense (*one sense per page*) and hence it is annotated with the same synset.[4] After all heuristics have been applied, overlapping mentions and duplicates are removed by enforcing a conservative policy which favors intra-page annotations over inter-page ones, and selects the longest match in case of overlapping annotations of the same type.

The result of this process is SEW, a Wikipedia-based corpus with over 200 million sense annotations of BabelNet synsets for all open-class parts of speech (nouns, verbs, adjectives, and adverbs).

3 SEW-EMBED: Building Vectors from Sense Annotations

In this section we provide the details of SEW-EMBED. We start by briefly describing the original explicit representations based on SEW (Section 3.1) and then our embedded augmentation (Section 3.2). The workflow of our procedure is depicted in Figure 1 with an illustrative example.

[3] http://lcl.uniroma1.it/nasari

[4] Although restrictive, this assumption is surprisingly accurate, as shown also in previous work (Wu and Giles, 2015).

3.1 Explicit Representation

As a starting point, we consider the Wikipedia-based representation (WB-SEW) by Raganato et al. (2016), in which each concept or entity s in the BabelNet sense inventory is represented as a vector v_s where dimensions are Wikipedia pages. For each Wikipedia page p in SEW, the corresponding component of v_s is computed as the estimated frequency of s appearing as sense annotation in p. Frequency is estimated using *lexical specificity* (Lafon, 1980), a statistical measure based on the hypergeometric distribution, particularly suitable for extracting an accurate set of representative terms for a given subcorpus SC of a reference corpus RC. We applied the procedure described by Camacho Collados et al. (2016), with the single page p as SC and the whole SEW as RC. As a result we obtain v_s, a rather sparse vector in which non-zero components correspond to the Wikipedia pages where s appears as a hyperlink; the weight ω_p associated with each component reflects the representativeness of s in the context described by p (Figure 1a).

3.2 Embedded Representation

In order to compute the embedded augmentation of an explicit vector v_s, obtained as in Section 3.1 for a given concept or entity s, we follow Camacho Collados et al. (2016) and exploit the compositionality of word embeddings (Mikolov et al., 2013b). According to this property, the representation of an arbitrary compositional phrase can be expressed as the combination (typically the average) of its constituents' representations. We build on this property and plug a pre-trained embedding representation into the explicit representation of Raganato et al. (2016). In particular, we consider each dimension p (i.e. Wikipedia page, cf. Section 3.1) of v_s and map it to the embedding space E provided by the pre-trained representation to obtain an embedded vector e_p. Such mapping depends on the specific embedding representation:

- In case of a *word* embedding representation we consider the Wikipedia page title as lexicalization of p and then retrieve the associated pre-trained embedding. If the title is a multi-word expression and no embedding is available for the whole expression, we exploit compositionality again and average the embedding vectors of its individual tokens;
- In case of a *sense* or *concept* embedding representation we instead exploit BabelNet's inter-resource links, and map p to the target sense inventory for which the corresponding embedding vector can be retrieved.

The embedded representation e_s of s (Figure 1b) is then computed as the weighted average over all the embedded vectors e_p associated with the dimensions of v_s:

$$e_s = \frac{\sum_{p \in v_s} \omega_p \, e_p}{\sum_{p \in v_s} \omega_p} \quad (1)$$

where ω_p is the lexical specificity weight of dimension p. In contrast to a simple average, here we exploit the ranking of each dimension p (represented by ω_p) and hence give more importance to the higher weighted dimensions of v_s.

3.3 Word Similarity

In order to calculate similarity at the word level, we follow other sense-based approaches (Pilehvar et al., 2013; Camacho Collados et al., 2016) and adopt a strategy that selects, for a given word pair w_1 and w_2, the *closest* pair of candidate senses:

$$Sim(w_1, w_2) = \max_{s_1 \in S_{w_1},\, s_2 \in S_{w_2}} \sigma(\vec{s_1}, \vec{s_2}) \quad (2)$$

where S_w is the set of candidate senses of w in the BabelNet sense inventory, and $\vec{s}$ is the vector representation associated with $s \in S_w$. As similarity measure σ we use standard *cosine similarity* for SEW-EMBED (Section 3.2), and *weighted overlap* (Pilehvar et al., 2013) for the explicit representations based on SEW (Section 3.1).

Finally, we rely on a back-off strategy that set $Sim(w_1, w_2) = 0.5$ (i.e. the middle point in our similarity scale) when no candidate sense is found for either w_1 or w_2.

4 Experiments

In this section we report and discuss the performance of SEW-EMBED on the monolingual and cross-lingual benchmark of the Semeval 2017 Task 2 (Camacho Collados et al., 2017). We consider two versions of SEW-EMBED: one based on the pre-trained word embeddings of Word2Vec (Mikolov et al., 2013a, **SEW-EMBED**$_{w2v}$)[5], and another one based on the

[5] We utilized the pre-trained models available at https://code.google.com/archive/p/word2vec. These models were trained on a Google News corpus of about 100 billion words.

	EN			FA			DE			IT			ES		
	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean
SEW-EMBED$_{w2v}$	0.56	0.58	0.57	0.38	0.40	0.39	0.45	0.45	0.45	0.57	0.57	0.57	0.61	0.62	0.62
SEW-EMBED$_{Nasari}$	0.57	0.61	0.59	0.30	0.40	0.34	0.38	0.45	0.42	0.56	0.62	0.59	0.59	0.64	0.62
SEW	0.61	0.67	0.64	**0.51**	**0.56**	**0.53**	**0.51**	**0.53**	**0.52**	**0.63**	**0.70**	**0.66**	**0.60**	**0.66**	**0.63**
NASARI	**0.68**	**0.68**	**0.68**	0.41	0.40	0.41	**0.51**	0.51	0.51	0.60	0.59	0.60	**0.60**	0.60	0.60

Table 1: Results on the multilingual word similarity benchmarks (subtask 1) of Semeval 2017 task 2, in terms of Pearson correlation (r), Spearman correlation (ρ), and the harmonic mean of r and ρ.

	DE-ES			DE-FA			DE-IT			EN-DE			EN-ES		
	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean
SEW-EMBED$_{w2v}$	0.52	0.54	0.53	0.42	0.44	0.43	0.52	0.52	0.52	0.50	0.53	0.51	0.59	0.60	0.59
SEW-EMBED$_{Nasari}$	0.47	0.55	0.51	0.35	0.45	0.39	0.47	0.55	0.51	0.46	0.55	0.50	0.59	0.63	0.61
SEW	**0.57**	**0.61**	**0.59**	**0.53**	**0.58**	**0.56**	**0.59**	**0.64**	**0.61**	0.58	**0.62**	**0.60**	0.61	**0.63**	0.61
NASARI	0.55	0.55	0.55	0.46	0.45	0.46	0.56	0.56	0.56	**0.60**	0.59	**0.60**	**0.64**	**0.63**	**0.63**

	EN-FA			EN-IT			ES-FA			ES-IT			IT-FA		
	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean	r	ρ	Mean
SEW-EMBED$_{w2v}$	0.46	0.49	0.48	0.58	0.60	0.59	0.50	0.53	0.52	0.59	0.60	0.60	0.48	0.50	0.49
SEW-EMBED$_{Nasari}$	0.41	0.52	0.46	0.59	0.65	0.62	0.44	0.54	0.48	0.58	0.64	0.61	0.42	0.52	0.47
SEW	**0.58**	**0.63**	**0.61**	0.64	**0.71**	**0.68**	**0.59**	**0.65**	**0.62**	**0.63**	**0.70**	**0.66**	**0.59**	**0.65**	**0.62**
NASARI	0.52	0.49	0.51	**0.65**	0.65	0.65	0.49	0.47	0.48	0.60	0.59	0.60	0.50	0.48	0.49

Table 2: Results on the cross-lingual word similarity benchmarks (subtask 2) of Semeval 2017 task 2, in terms of Pearson correlation (r), Spearman correlation (ρ), and the harmonic mean of r and ρ.

embedded concept vectors of NASARI (Camacho Collados et al., 2016, **SEW-EMBED**$_{Nasari}$). In all test sets, the figures of **SEW-EMBED**$_{w2v}$ correspond to the results of SEW-EMBED reported in the task description paper (Camacho Collados et al., 2017). We additionally include the results obtained by the original explicit representations based on SEW (cf. Section 3.1) and by the NASARI baseline, and use them as comparison systems across Sections 4.1 and 4.2.[6]

4.1 Subtask 1: Multilingual Word Similarity

Table 1 shows the overall performance on multilingual word similarity for each monolingual dataset. Both **SEW-EMBED**$_{w2v}$ and **SEW-EMBED**$_{Nasari}$ achieve comparable results: their correlation figures are in the same ballpark as the NASARI baseline for Italian, Farsi, and Spanish; instead, they lag behind in English and German. Most surprisingly, however, the explicit representations based on SEW show an impressive performance, and reach the best result overall in 4 out of 5 benchmarks: this might suggest that many word pairs across the test sets are actually being associated with concepts or entities that are well connected in the semantically enriched Wikipedia, and hence the corresponding sparse vectors are representative enough to provide meaningful comparisons. In general, the performance decrease on German and Farsi for all comparison systems is connected to the lack of coverage: both SEW and SEW-EMBED use the back-off strategy (cf. Section 3.3) 70 times for Farsi (14%) and 54 times (10.8%) for German.

4.2 Subtask 2: Cross-lingual Word Similarity

Table 2 reports the overall performance on cross-lingual word similarity for each language pair. Consistently with the multilingual evaluation (Section 4.1), both **SEW-EMBED**$_{w2v}$ and **SEW-EMBED**$_{Nasari}$ achieve comparable results in the majority of benchmarks. All approaches based on SEW seem to perform globally better in a cross-lingual setting: on average, the harmonic mean of r and ρ is 2.2 points below the NASARI baseline (compared to 3.2 points in the evaluation of Section 4.1). This suggests the potential of Wikipedia as a bridge to multilinguality: in fact, even though SEW was constructed automatically on the English Wikipedia, knowledge transfers rather well via inter-language links and has a considerable impact on the cross-lingual performance.

[6] For an extensive comparison including all participating systems in the task, the reader is referred to the task description paper (Camacho Collados et al., 2017).

Again, the best figures are consistently achieved by the explicit representations based on SEW: the improvement in terms of harmonic mean of r and ρ is especially notable in benchmarks that include a less-resourced language such as Farsi (+11.75% on average compared to the NASARI baseline). This improvement does not occur with SEW-EMBED, since in that case sparse vectors are eventually mapped to an embedding space trained specifically on an English corpus.

4.3 General Discussion

Overall, SEW-EMBED reached the 4th and 3rd positions in the global rankings of subtask 1 and 2 respectively (with scores 0.552 and 0.558, not including the NASARI baseline). Thus, perhaps surprisingly, the embedded augmentation yielded a considerable decrease in terms of global performance in both subtasks, where the original explicit representations of SEW achieved a global score of 0.615 in subtask 1, and a global score of 0.63 in subtask 2 (cf. Sections 4.1-4.2).[7]

Intuitively, multiple factors might have influenced this negative result:

- **Dimensionality Reduction.** Converting an explicit vector (with around 4 million dimensions) into a latent vector of a few hundred dimensions leads inevitably to losing some valuable information, and hence to a decrease in the representational power of the model. Such a phenomenon was also shown by Camacho Collados et al. (2016), where the lexical and unified representations of NASARI tend to outperform the embedded representation on several word similarity and sense clustering benchmarks;

- **Lexical Ambiguity.** While the original concept vectors of SEW are defined in the unambiguous semantic space of Wikipedia pages, we constructed their embedded counterparts via the word-level representations of their lexicalized dimensions (Section 3.2); hence, when moving to the word level, we ended up conflating the different meanings of an ambiguous word or expression;[8]

- **Non-Compositionality.** The compositional properties of word embeddings that we assumed in Section 3.2 falls short in many cases, such as idiomatic expressions or named entity mentions (e.g. *Wall Street*, or *New York*). The explicit vectors of SEW, instead, do not require the compositional assumption and always consider a multi-word expression as a whole.

Even though the embedded representations of SEW do not match up to the accuracy of explicit ones on experimental benchmarks, they are on the other hand more convenient in terms of compactness and flexibility (due to the reduced dimensionality), and also in terms of comparability, as they are defined in the same vector space of Word2Vec-based representations such as the embedded vectors of NASARI (Camacho Collados et al., 2016) or DECONF (Pilehvar and Collier, 2016).

5 Conclusion

In this paper we presented SEW-EMBED, a language-independent concept representation approach which we put forward as a competitor system in the Semeval-2017 Task 2 (Camacho Collados et al., 2017). SEW-EMBED is tied to a Wikipedia-based sense-annotated corpus, SEW (Raganato et al., 2016), obtained automatically by exploiting the hyperlink structure of Wikipedia and the wide-coverage sense inventory of BabelNet. SEW is used to construct sparse vector representations in the space of Wikipedia pages, which are then mapped to an embedded representation by plugging in an arbitrary word (or sense) embedding model and computing a weighted average. We described and evaluated SEW-EMBED on all benchmarks available for the task, together with the explicit sparse vectors originally proposed by Raganato et al. (2016). In spite of the methodological simplicity of the approach (which was designed as an extrinsic test bed for the quality of SEW's annotations), global figures put SEW-EMBED close to, or on par with, state-of-the-art approaches such as NASARI. In particular, we showed that a cross-lingual setting yields the best overall improvement for concept representations based entirely on SEW, suggesting its potential for multilingual and cross-lingual applications.

[7] The global score is computed as the average harmonic mean of Pearson and Spearman correlation on the best four (subtask 1) and six (subtask 2) individual benchmarks (Camacho Collados et al., 2017).

[8] E.g., in SEW-EMBED$_{w2v}$, the distinct explicit dimensions represented in SEW by the Wikipedia pages BANK and BANK (GEOGRAPHY) were both mapped to the Word2Vec embedding of *bank*.

References

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

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

José Camacho Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proc. of SemEval*. pages 15–26.

José Camacho Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets. In *Proc. of ACL*. pages 1–7.

José Camacho Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence* 240:36–64.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proc. 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 the Second International Joint Conference on Natural Language Processing*. pages 767–778.

Samer Hassan and Rada Mihalcea. 2011. Semantic Relatedness Using Salient Semantic Analysis. In *Proc. of AAAI*. pages 884–889.

Eduard 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.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proc. of ACL*. pages 95–105.

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

Rada Mihalcea. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. In *Proc. of NAACL-HLT*. pages 196–203.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR Workshop*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Proc. of NIPS*. pages 3111–3119.

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.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. In *Proc. of EMNLP*. pages 1680–1690.

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.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2016. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In *Proc. of IJCAI*. pages 2894–2900.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37(1):141–188.

Zhaohui Wu and C. Lee Giles. 2015. Sense-aware Semantic Analysis: A Multi-prototype Word Representation Model using Wikipedia. In *Proc. of AAAI*. pages 2188–2194.