Zero-shot WordNet Construction using Cross-lingual Embeddings

Anonymous ACL submission

Abstract

Low-resource languages often lack structured text representations (taxonomies, ontologies and lexical databases). In this paper we propose a method for constructing WordNets from Princeton WordNet without translation data or parallel corpora. The proposed method uses cross-lingual word embeddings and outperforms translation-based techniques in F1-score. We also publish automatedly constructed general truncated WordNets and collocation WordNets for 44 languages (including non-European ones).

1 Introduction

There are numerous structured information representations containing texts as titles, descriptions or definitions: e.g. ontologies, taxonomies, and lexical databases. Among such databases we can highlight WordNet(Miller, 1995). WordNet is a lexical database covering various types of relations between words: both semantical and lexical. Semantic concepts called synsets are connected in accordance to the semantic and lexical relations between them. The database has found a very broad usage for many natural language processing and machine learning applications (Kutuzov et al., 2018; Mao et al., 2018). There have been many attempts by researchers to automatically convert WordNet from English into other languages. Most attempts were focused on using machine translation engines, extensive bilingual dictionaries or parallel corpora (Khodak et al., 2017; Neale, 2018) which are often lacking for low-resource languages.

In this paper we propose a method for constructing WordNets using cross-lingual embeddings. Unlike previous attempts our method does not require translation engines or parallel corpora. There have been already works using word embeddings for extending existing WordNets (Sand et al., 2017; Al tarouti and Kalita, 2016) in monolingual settings. However, these methods could not be used for creating a WordNet for another language from scratch.

Word embeddings proved to be a powerful tool for dense text representations after papers by Bengio (Bengio et al., 2003) and Mikolov (Mikolov et al., 2013a). However, first word vector representation models were monolingual only. Soon researchers proposed cross-lingual word embedding models (Mikolov et al., 2013b). There followed several improvements to the original model. In 2016 Arxetxe et al. found that Procrustes refinement gets better results than the original linear transformation method by Mikolov. Also most earlier methods suffered from the "hubness problem" where some words (especially low frequency ones) appear in the top neighbour lists of many other words.

Alexis Conneau et al. in 2017 offered a method called cross-domain similarity local scaling (CSLS) to overcome this problem. They reached 81.7% accuracy for English-Spanish and 83.7% for Spanish-English pairs for top-1500 source queries in a completely unsupervised mode. For English-Russian and Russian-English their results are not as high and they achieved accuracy of 51.7% and 63.7% respectively. Their FastText embeddings were trained on Wikipedia datasets for each respective language. They have published aligned embeddings for 30 languages [1]. Joulin et al. in 2018 found that convex relaxation of the CSLS loss improves the quality of bilingual word alignment. They have also published aligned FastText vectors with vocabularies of more than 2 million words and phrases [2] (Fig. 1).

[1] https://github.com/facebookresearch/MUSE
[2] https://fasttext.cc/docs/en/aligned-vectors.html

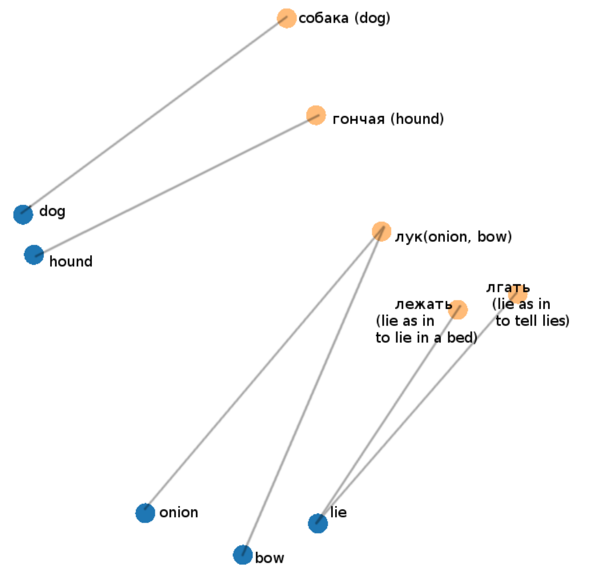

Figure 1: PCA visualization of aligned word embeddings for Russian and English

2 Cross-lingual embeddings

MUSE is based on the work by Conneau et al. (Conneau et al., 2017). It consists of two algorithms. The first one which is used only in unsupervised scenarios is a pair of adversarial neural networks. The first neural network is trained to predict from which distribution $\{X, Y\}$ embeddings come. The second neural networks is trained to modify embeddings X multiplying it by matrix W to prevent the first neural network from making accurate discriminations. Thus, at the end of the training we get a matrix WX which is aligned with matrix Y.

The second method is supervised and the aim is to find a linear mapping W between embedding spaces X and Y which can be solved using Orthogonal Procrustes problem:

$$W^* = argmin_W ||WX - Y||_F = UV^T$$

where UV^T is derived using singular value decomposition SVD$(YX^T) = U\Sigma V^T$ This method is used iteratively with the default number of iterations in MUSE equal to 5. As Søgaard, Ruder and Vulić state Procrustes refinement relies on frequent word pairs to serve as reliable anchors.

Conneau et al. also apply cross-domain similarity local scaling to lessen the extent of hubness problem which cross-lingual embeddings are prone to (Dinu et al., 2015). It uses cosine similarity between a source embedding vector x and k target nearest embeddings $\mathcal{N}$ (the default k in MUSE is 10) to generate a dictionary.

$$sim(x, y) = \frac{1}{k}\sum_{i=1}^{K} \cos(x, \mathcal{N}_{Xi});$$
$$\mathcal{N}_X \in Y = \{y_1, ..., y_n\}$$

$$CSLS(x, y) = 2\cos(x, y) - sim_{source}(x, y) - sim_{target}(y, x)$$

Vecmap (Artetxe et al., 2018) is close in its idea to the Procrustes refinement, it computes SVD-factorization SVD$(YX^T) = U\Sigma V^T$ and replaces X and Y with new matrices $X' = U$ and $Y' = V$. The authors also propose normalization and whitening (sphering) transformation. After applying whitening new matrices are equal to: $X' = (X^TX)^{-\frac{1}{2}}$ and $Y' = (Y^TY)^{-\frac{1}{2}}$

Jawanpuria et al. (Jawanpuria et al., 2018) propose a method which is, likewise, based on SVD-factorization but in smooth Riemannian manifolds instead of Euclidean space.

Joulin et al. in 2018 introduced a reformulation of CSLS that generalizes to convex functions (Relaxed CSLS loss). Due to the orthogonality constraint on W and FastText vectors being ℓ_2-normalized $\cos(Wx, y) = x^TW^Ty$ and $||y - W_{Xi}||_2^2 = 2 - 2x^TW^Ty$. The problem can be reformulated to find the k elements of Y which have the largest dot product with W_{Xi}. Thus, RCSLS can be written down as:

$$\min_{W \in \mathcal{O}_d} \frac{1}{n}\sum_{i=1}^{n} -2x_i^TW^Ty_i$$
$$+\frac{1}{k}\sum_{y_j \in \mathcal{N}_Y(W_{Xi})} x_i^TW^Ty_j$$
$$+\frac{1}{k}\sum_{W_{Xj} \in \mathcal{N}_X(y_i)} x_j^TW^Ty_i$$

Thus, RCSLS can be solved using manifold optimization tools (Boumal et al., 2014).

3 Experiments

We reformulated the problem of synset finding as a binary classification problem. The task is to predict for the given (synset, lemma) pair if they are related or not. As training/validation data we used English Princeton WordNet (Miller, 1995) provided by the NLTK package (Bird, 2006). It contains 117'659 synsets. As positive examples

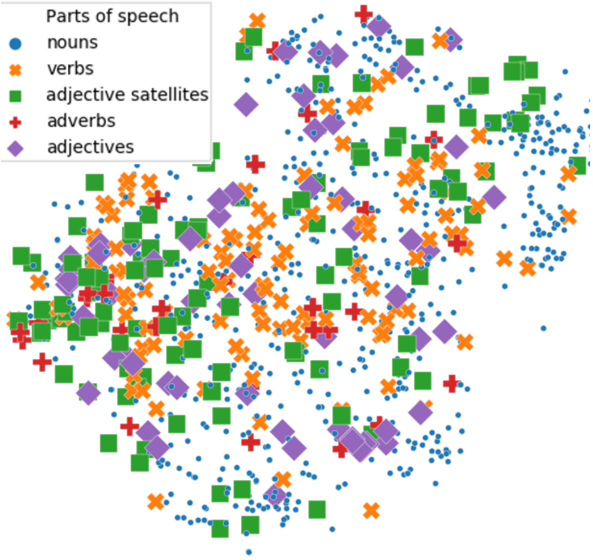

Figure 2: TSNE visualisation of WordNet synset SIF-embeddings

we use (lemma, synset) pairs. As negative examples lemmas from other synsets with the same root are used (chicken.n.01 vs. chicken.n.02). We also added some random words because of scarcity of negative examples. There were also attempts at augmenting training data with the information from the Open Multilingual WordNet (Bond and Foster, 2013). For the final model we used only Finnish Open WordNet because it is 100% full and allows to avoid implicit bias towards Indo-European languages used for testing.

Table 1: Training data for chicken.n.01 (the flesh of a chicken used for food)

Word	Synset	Target
chicken	chicken.n.01	1
poulet	chicken.n.01	1
yellow	chicken.n.01	0
chickenhearted	chicken.n.01	0
visible	chicken.n.01	0

Synset embeddings were calculated using averaged synset lemma embedding and the definition embedding. We used averaging weights 0.2 for the lemma and 0.8 for the definition. SIF (smooth inverse frequency) and TF-IDF (term frequency–inverse document frequency) averaging schemes were used for definition embeddings. SIF (Arora et al., 2017) embeddings use pre-trained word vectors. For each sentence s this model first creates a vectorized averaged representation V_s.

$$V_s = \frac{1}{|s|}\sum_{w \in s} \frac{a}{a + p(w)} V_w$$

where V_w is the word unigram probability and a is a scalar (set to 1e-3 by default). After that all sentence embeddings are grouped into a matrix where u is its first singular vector. The final sentence embedding is computed using this singular vector u.

$$V_s = V_s - uu^T V_s$$

For each lemma we used its embedding from the corresponding cross-lingual pre-trained model ((Conneau et al., 2017) or (Joulin et al., 2018)) for the language.

Each synset vector is also augmented with information about its part-of-speech and the synset number.

Predicting synset relations is not a trivial task even in a monolingual setting. E.g. we failed to get any meaningful cluster representation for synsets using TSNE (van der Maaten and Hinton, 2008) (Fig. 2). Moreover, there is not much training data and models are prone to overfitting. Thus, we introduced an ensemble of 4 LGBM-models (Ke et al., 2017) and 4 dense 3-layered fully-connected neural networks with dropout as regularization (Srivastava et al., 2014). In our case fine-tuning parameters using half of data not only did not bring any benefit to the final score but even decreased it significantly (about 3 F1-score points).

We also attempted to fine tune input data using PCA (principal component analysis) and UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) but it did not provide any gains, and brought about worse results.

For testing there were used two manually annotated datasets provided in the paper (Khodak et al., 2017) for Russian and French languages respectively. Each dataset consists of 600 target language words from three parts of speech (nouns, verbs and adjectives). Each word has some true senses (synsets) and false ones (Table 2). The original test procedure did not penalize models for synsets and words that they do not contain. In our case the RCSLS-model has full coverage of the dataset.

4 Multiword Expressions

Multiword expressions (MWE) are notoriously difficult to process and to model. A typical phrasing scheme used by Mikolov in Word2Vec (2013a) is a very simple and rather efficient way to take phrase context into consideration and not just con-

Table 2: Test data by (Khodak et al., 2017)

Word	Target	Synset	Definition
адрес	1	address.n.01	(computer science) the code that identifies where a piece of information is stored
	0	address.n.03	the act of delivering a formal spoken communication to an audience
aise	1	comfortable.a.01	providing or experiencing physical well-being or relief
	0	comfortable.s.03	more than adequate; Example 1: the home team had a comfortable lead

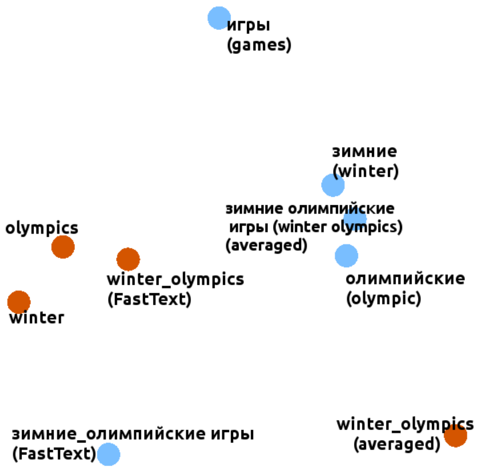

Figure 3: TSNE visualisation of averaged and FastText induced embeddings for MWE

sider it as an average of constituents 3.

$$score(w_i, w_j) = \frac{count(w_i w_j) - \delta}{count(w_i) * count(w_j)}$$

Other metrics such as PMI (Bouma, 2009) or a special model for identifying MWE using a corpus like PARSEME (Savary et al., 2018) might have been preferable. Yet in this work we were constrained by the multi-word expressions scheme used in pretrained embeddings.

Table 3: MWE examples for the aligned English RCSLS embeddings

nearest_city
fiscal_code
architectural_style
population_metro
winners_share
third_team
people_from_new_orleans
parent_agency

Another problem with FastText MWE is that the model is trained using the Wikipedia corpora. This leads to many artifact multi-word expressions corresponding to Wikipedia categories (Fig 3). Still we decided to publish WordNet for this collocations as may of them are still relevant and correspond to multi-part verbs and collocations (e.g. 'take notes', 'take away'). It also highlights the importance of another approach for MWE in word embedding models.

5 WordNet construction

In line with previous works (Vossen, 2013; Tufis et al., 2006) we extend an existing WordNet and match words from the target language to existing Princeton WordNet synsets.

Despite being easy to train, WordNet construction imposes significant computational problems in our case because for every word from the vocabulary we need to compare it with every possible synset. For this reason we used several heuristics. 1) removed strings with punctuation besides the underscore symbol. 2) identified language using (Joulin et al., 2016) for each string (FastText embeddings are noisy and contain a lot of samples from other languages). All computations were vectorized. Ensemble methods also are easy to parallelize using the multiprocessing realization in Python and using bufferized numpy-arrays allows to increase the batch size and avoid copying data (Gorelick and Ozsvald, 2014).

Preselected words using a simple model with a low threshold (0.2). However, we still had to limit the size of our WordNets. That is why we publish automated WordNets for all collocations and most common 10000 lemmas for 44 languages. To increase the quality of the generated models it was decided to sacrifice recall for the sake of precision and we increased the model confidence threshold to 0.6.

6 Results

As can be seen from table 4 cross-lingual embedding methods outperform translation methods in most categories except Russian nouns without being fine-tuned on the test (as the work by Khodak). Actually, f1-score fine-tuning using the validation set even decreased the test set. Moreover, cross-lingual methods are reported to work in a completely unsupervised way. Even in the unsupervised mode they are easier to come by because because they require only a limited bilingual dictionary. High-quality translation engines are still inadequate for low-resource languages (and for many rare languages they reportedly outperform Transformer-based models (Conneau et al., 2018)).

Using information from another language is also helpful. It provides up to 1.5 F1-score point performance boost for some parts of speech. However, the English-only model also outperforms previous works for French and is slightly better than our best-performing model in verb representation. It should also be noted that in the experiments with MUSE-embeddings (not tested for RCSLS embeddings) data from other languages with smaller WordNets (e.g. Polish) from the Extended Open Multilingual WordNet decreased the results by 1.2 - 2 accuracy points for the validation dataset.

The SIF embedding scheme provides an advantage of 3.5 F1-score points in comparison with simple TF-IDF averaging.

MUSE-embeddings perform slightly worse than RCSLS. However, it should be also noted that the vocabulary for MUSE embeddings is only 200'000 words vs 2'000'000 for RCSLS. However, it should not have substantially influenced test results because of the chosen test procedure.

Ensembling gives a major performance boost. However, it may be partially attributed to out lack of investment into fine-tuning of individual models. Individual models are also almost as performant for French as previous multi-stepped procedures that used translation engines and clustering. However, they fail for Russian which can be attributed to overfitting to the original English dataset. Simple averaging between models helps to mitigate it.

7 Conclusion

Cross-lingual embeddings turned out to be an efficient method for cross-lingual WordNet extension. This technique is not limited to WordNet construction, and can also be used for other types of similar structures (e.g. taxonomies and ontologies). We also published truncated and collocation WordNets for 44 languages which can be used in future research.

Our work has shown that it is possible to build a WordNet for a new language without corpora or translation engines for the target language. Cross-lingual embeddings used in this work are fine-tuned with parallel dictionaries. However, an interesting direction of improvement would be to use fully-unsupervised models that do even rely on any parallel data at all.

8 Phraser

Moreover, what we find amusing is that English validation set results are similar to the results for the test set in another language.

Many researchers have used a similar method for detecting new hypernyms-hyponyms relations beyond WordNet for English. The work by (Sanchez and Riedel, 2017) has provided an overview of such methods. After getting rid of noisy hypernym-hyponym samples models may achieve models may achieve up to 81.2 % in accuracy score.

Some works propose to change the embeddings training process to incorporate hierarchical information (Alsuhaibani et al., 2019).

Word embeddings methods are preferable for short texts (Maslova and Potapov, 2017).

References

Feras Al tarouti and Jugal Kalita. 2016. Enhancing Automatic Wordnet Construction Using Word Embeddings. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 30–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mohammed Alsuhaibani, Takanori Maehara, and Danushka Bollegala. 2019. Joint Learning of Hierarchical Word Embeddings from a Corpus and a Taxonomy. pages 1–19.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. *ICLR*, pages 1–14.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings.

Y Bengio, R Ducharme, and P Vincent. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Table 4: Results for WordNet-synset prediction

Method	POS	F1 French	F1 Russian
Extended Open Multilingual Wordnet(Bond and Foster, 2013)	Adj.	40.8	41.3
	Noun	43.8	53.1
	Verb	29.4	34.8
	Total	38.0	43.1
Synset Representation + Linear-WSI (Khodak et al., 2017)	Adj.	62.5	64.9
	Noun	66.0	**67.61**
	Verb	55.9	49.7
	Total	61.5	60.7
Ensemble model (SIF + Non-English data + RCSLS)	Adj.	**62.8**	**65.0**
	Noun	**71.8**	65.1
	Verb	60.0	**54.8**
	Total	**64.1**	**61.0**
Ensemble model (SIF + **Only-English** data + RCSLS)	Adj.	62.3	64.6
	Noun	70.9	63.6
	Verb	**60.3**	53.6
	Total	63.9	60.1
Ensemble model (SIF + Non-English data + **MUSE**)	Adj.	61.0	64.8
	Noun	71.3	64.1
	Verb	59.0	54.3
	Total	63.9	60.5
Ensemble model (**TFIDF** + Non-English data + RCSLS)	Adj.	62.3	63.0
	Noun	68.1	59.5
	Verb	53.9	48.0
	Total	60.7	56.5
Single **LGBM**-model (SIF + Non-English data + RCSLS)	Adj.	59.4	61.4
	Noun	69.2	63.4
	Verb	57.5	49.2
	Total	61.3	57.0
Single **NN**-model (SIF + Non-English data + RCSLS)	Adj.	60.5	62.9
	Noun	69.5	62.5
	Verb	56.3	51.1
	Total	61.1	58.5

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1352–1362.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. 2014. Manopt, a Matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. *arxiv.org*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *In Proceedings of the 3rd In- ternational Conference on Learning Representations (ICLR2015), workshop track*.

Micha Gorelick and Ian Ozsvald. 2014. *High Performance Python: Practical Performant Programming for Humans*. " O'Reilly Media, Inc.".

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2018. Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-yan Liu. 2017. LightGBM : A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 3149–3157.

Mikhail Khodak, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. 2017. Automated WordNet Construction Using Word Embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 12–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrey Kutuzov, Mohammad Dorgham, Oleksiy Oliynyk, Chris Biemann, and Alexander Panchenko. 2018. Learning Graph Embeddings from WordNet-based Similarity Measures.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Rui Mao, Guanyi Chen, Ruizhe Li, and Chenghua Lin. 2018. ABDN at SemEval-2018 Task 10: Recognising Discriminative Attributes using Context Embeddings and WordNet. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1017–1021, Stroudsburg, PA, USA. Association for Computational Linguistics.

Natalia Maslova and Vsevolod Potapov. 2017. Neural network doc2vec in automated sentiment analysis for short informal texts. In *Lecture Notes in Computer Science*, volume 10458 LNAI, pages 546–554.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pages 3111–3119.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Steven Neale. 2018. A Survey on Automatically-Constructed WordNets and their Evaluation: Lexical and Word Embedding-based Approaches. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ivan Sanchez and Sebastian Riedel. 2017. How Well Can We Predict Hypernyms from Word Embeddings? A Dataset-Centric Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 401–407.

Heidi Sand, Erik Velldal, and Lilja Øvrelid. 2017. Wordnet extension via word embeddings: Experiments on the Norwegian Wordnet. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 298–302.

Agata Savary, Marie Candito, V Barbu Mititelu, Eduard Bejček, Fabienne Cap, and M van Gompel. 2018. PARSEME multilingual corpus of verbal multiword expressions.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Dan Tufis, Verginica Barbu Mititelu, Luigi Bozianu, and Catalin Mihaila. 2006. Romanian wordnet: New developments and applications. In *Proceedings of the 3rd Conference of the Global WordNet Association*, pages 337–344.

Piek. Vossen. 2013. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic.