Enhancing Word Embeddings with Knowledge Extracted from Lexical Resources

Magdalena Biesialska*, Bardia Rafieian*, Marta R. Costa-jussà
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
{magdalena.biesialska,bardia.rafieian,marta.ruiz}@upc.edu

Abstract

In this work, we present an effective method for semantic specialization of word vector representations. To this end, we use traditional word embeddings and apply specialization methods to better capture semantic relations between words. In our approach, we leverage external knowledge from rich lexical resources such as BabelNet. We also show that our proposed post-specialization method based on an adversarial neural network with the Wasserstein distance allows to gain improvements over state-of-the-art methods on two tasks: word similarity and dialog state tracking.

1 Introduction

Vector representations of words (embeddings) have become the cornerstone of modern Natural Language Processing (NLP), as learning word vectors and utilizing them as features in downstream NLP tasks is the de facto standard. Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) are typically trained in an unsupervised way on large monolingual corpora. Whilst such word representations are able to capture some syntactic as well as semantic information, their ability to map relations (e.g. synonymy, antonymy) between words is limited. To alleviate this deficiency, a set of refinement post-processing methods—called retrofitting or semantic specialization—has been introduced. In the next section, we discuss the intricacies of these methods in more detail.

To summarize, our contributions in this work are as follows:

• We introduce an improved post-specialization method (dubbed WGAN-postspec), which demonstrates improved performance as compared to state-of-the-art DFFN (Vulić et al., 2018) and AuxGAN (Ponti et al., 2018) models.

• We show that the proposed approach achieves performance improvements on an intrinsic task (word similarity) as well as on a downstream task (dialog state tracking).

2 Related Work

Numerous methods have been introduced for incorporating structured linguistic knowledge from external resources to word embeddings. Fundamentally, there exist three categories of semantic specialization approaches: (a) joint methods which incorporate lexical information during the training of distributional word vectors; (b) specialization methods also referred to as retrofitting methods which use post-processing techniques to inject semantic information from external lexical resources into pre-trained word vector representations; and (c) post-specialization methods which use linguistic constraints to learn a general mapping function allowing to specialize the entire distributional vector space.

In general, joint methods perform worse than the other two methods, and are not model-agnostic, as they are tightly coupled to the distributional word vector models (e.g. Word2Vec, GloVe). Therefore, in this work we concentrate on the specialization and post-specialization methods. Approaches which fall in the former category can be considered local specialization methods, where the most prominent examples are: retrofitting (Faruqui et al., 2015) which is a post-processing method to enrich word embeddings with knowledge from semantic

* Equal contribution
lexicons, in this case it brings closer semantically similar words. Counter-fitting (Mrkšić et al., 2016) likewise fine-tunes word representations; however, conversely to the retrofitting technique it counter-fits the embeddings with respect to the given similarity and antonymy constraints. Attract-Repel (Mrkšić et al., 2017b) uses linguistic constraints obtained from external lexical resources to semantically specialize word embeddings. Similarly to counter-fitting it injects synonymy and antonymy constraints into distributional word vector spaces. In contrast to counter-fitting, this method does not ignore how updates of the example word vector pairs affect their relations to other word vectors.

On the other hand, the latter group, post-specialization methods, performs global specialization of distributional spaces. We can distinguish: explicit retrofitting (Glavaš and Vulić, 2018) that was the first attempt to use external constraints (i.e. synonyms and antonyms) as training examples for learning an explicit mapping function for specializing the words not observed in the constraints. Later, a more robust DFFN (Vulić et al., 2018) method was introduced with the same goal — to specialize the full vocabulary by leveraging the already specialized subspace of seen words.

3 Methodology

In this paper, we propose an approach that builds upon previous works (Vulić et al., 2018; Ponti et al., 2018). The process of specializing distributional vectors is a two-step procedure (as shown in Figure 1). First, an initial specialization is performed (see 3.1). In the second step, a global specialization mapping function is learned, allowing to generalize to unseen words (see 3.2).

3.1 Initial Specialization

In this step a subspace of distributional vectors for words that occur in the external constraints is specialized. To this end, fine-tuning of seen words can be performed using any specialization method. In this work, we utilize Attract-Repel model (Mrkšić et al., 2017b) as it offers state-of-the-art performance. This method allows to make use of both synonymy (attract) and antonymy (repel) constraints. More formally, given a set \( \mathcal{A} \) of attract word pairs and a set of \( \mathcal{R} \) of repel word pairs, let \( \mathcal{V}_s \) be the vocabulary of words seen in the constraints. Hence, each word pair \( (v_l, v_r) \) is represented by a corresponding vector pair \( (x_l, x_r) \). The model optimization method operates over mini-batches: a mini-batch \( \mathcal{B}_{\mathcal{A}} \) of synonymy pairs (of size \( k_1 \)) and a mini-batch \( \mathcal{B}_{\mathcal{R}} \) of antonymy pairs (of size \( k_2 \)). The pairs of negative examples \( T_A(\mathcal{B}_{\mathcal{A}}) = \left\{ (t^1_l, t^1_r), \ldots, (t_{k_1}^1, t_{k_1}^1) \right\} \) and \( T_R(\mathcal{B}_{\mathcal{R}}) = \left\{ (t^2_l, t^2_r), \ldots, (t_{k_2}^2, t_{k_2}^2) \right\} \) are drawn from \( 2(k_1 + k_2) \) word vectors in \( \mathcal{B}_{\mathcal{A}} \cup \mathcal{B}_{\mathcal{R}} \).

The negative examples serve the purpose of pushing synonym pairs closer and pushing antonym pairs further away with respect to their corresponding negative examples. For synonyms:

\[
A(\mathcal{B}_{\mathcal{A}}) = \sum_{i=1}^{k_1} \left[ \tau(\delta_{\text{att}} + x^l \cdot t^i_l - x^r \cdot t^i_r) + \right. \\
+ \left. \tau \left( \delta_{\text{att}} + x^l \cdot t^i_r - x^r \cdot t^i_l \right) \right] 
\]  

where \( \tau \) is the rectifier function, and \( \delta_{\text{att}} \) is the similarity margin determining the distance between synonymy vectors and how much closer they should be comparing to their negative examples. Similarly,
While post-specialization we call our model well as its improved variant with gradient penalty $\lambda$ where (WGAN-GP) (Gulrajani et al., 2017). For brevity, 3.2 Proposed Post-Specialization Model is the original vector for the word $x$. One of the benefits of using WGANs over post-specialization brella term for the ial network (WGAN) (Arjovsky et al., 2017) as distance incorporated in the generative adversarial network (GAN) (Goodfellow et al., 2014). More specifically, the mapping function is learned through a combination of a standard layer classification network, tries to distinguish the corresponding initially specialized word vectors $x'_i$. Therefore, a global mapping function is trained using word vector pairs, such that $(x_i, x'_i) = \{x_i \in X_s, x'_i \in X'_s\}$. On the other hand, the discriminator $D(\theta_D)$, which is a multi-layer classification network, tries to distinguish the generated samples from the initially specialized vectors sampled from $X'_s$. In this process, the differences between predictions and initially specialized vectors are used to improve the generator, resulting in more realistically looking outputs.

In general, for the GAN model we can define the loss $L_G$ of the generator as:

$$L_G = - \sum_{i=1}^{n} \log P(\text{spec} = 1|G(x_i; \theta_G); \theta_D) - \sum_{i=1}^{m} \log P(\text{spec} = 0|x'_i; \theta_D)$$  \hfill (5)

While the loss of the discriminator $L_D$ is given as:

$$L_D = - \sum_{i=1}^{n} \log P(\text{spec} = 0|G(x_i; \theta_G); \theta_D) - \sum_{i=1}^{m} \log P(\text{spec} = 1|x'_i; \theta_D)$$  \hfill (6)

In principle, the losses with Wasserstein distance can be formulated as follows:

$$L_G = -\frac{1}{n} \sum_{i=1}^{n} D(G(x_i; \theta_G); \theta_D)$$  \hfill (7)

and

$$L_D = \frac{1}{m} \sum_{i=1}^{m} D(x'_i; \theta_D) - \frac{1}{n} \sum_{i=1}^{n} D(G(x_i; \theta_G); \theta_D)$$  \hfill (8)
An alternative scenario with a gradient penalty (WGAN-GP) requires adding gradient penalty $\lambda$ coefficient in the Eq. (8).

4 Experiments

Pre-trained Word Embeddings. In order to evaluate our proposed approach as well as to compare our results with respect to current state-of-the-art post-specialization approaches, we use popular and readily available 300-dimensional pre-trained word vectors. Word2Vec (Mikolov et al., 2013) embeddings for English were trained using skip-gram with negative sampling on the cleaned and tokenized Polyglot Wikipedia (Al-Rfou’ et al., 2013) by Levy and Goldberg (2014), while German and Italian embeddings were trained using CBOW with negative sampling on WacKy corpora (Dinu et al., 2015; Artetxe et al., 2017, 2018). Moreover, GloVe vectors for English were trained on Common Crawl (Pennington et al., 2014).

Linguistic Constraints. To perform semantic specialization of word vector spaces, we exploit linguistic constraints used in previous works (Zhang et al., 2014; Ono et al., 2015; Vulić et al., 2018) (referred to as external) as well as introduce a new set of constraints collected by us (referred to as babelnet) for three languages: English, German and Italian. We use constraints in two different settings: disjoint and overlap. In the first setting, we remove all linguistic constraints that contain any of the words available in SimLex (Hill et al., 2015), SimVerb (Gerz et al., 2016) and WordSim (Leviant and Reichart, 2015) evaluation datasets. In the overlap setting, we let the SimLex, SimVerb and WordSim words remain in the constraints. To summarize, we present the number of word pairs for English, German and Italian constraints in Table 1.

Let us discuss in more detail how the lists of constraints were constructed. In this work, we use two sets of linguistic constraints: external and babelnet. The first set of constraints was retrieved from WordNet (Fellbaum, 1998) and Rogets Thesaurus (Kipfer, 2009), resulting in 1,023,082 synonymy and 380,873 antonymy word pairs. The second set of constraints, which is a part of our contribution, comprises synonyms and antonyms obtained using NASARI lexical embeddings (Camacho-Collados et al., 2016) and BabelNet (Navigli and Ponzetto, 2012). As NASARI provides lexical information for BabelNet words in five languages (EN, ES, FR, DE and IT), we collected each word with its related BabelNetID (a sense database identifier) to extract the list of its synonyms and antonyms using BabelNet API.

Furthermore, to improve the list of Italian words, we also followed the approach proposed by Scheible and Schulte im Walde (2014) that presents a new collection of semantically related word pairs in German, which was compiled through human evaluation. Relying on GermaNet and the respective JAVA API, the list of the word pairs was generated with a sampling technique. Finally, we used these word pairs in our experiments as external resources for the German language.

Initial Specialization and Post-Specialization. Although, initially specialized vector spaces show gains over the non-specialized word embeddings, linguistic constraints represent only a fraction of their total vocabulary. Therefore, semantic specialization is a two-step process. Firstly, we perform initial specialization of the pre-trained word vectors by means of Attract-Repel (see 2) algorithm. The values of hyperparameter are set according to the default values: $\lambda_{reg} = 10^{-9}$, $\delta_{sim} = 0.6$, $\delta_{ant} = 0.0$ and $k_1 = k_2 = 50$. Afterward, to perform a specialization of the entire vocabulary, a global specialization mapping function is learned. In our WGAN-postspec proposed approach, the post-specialization model uses a GAN with improved loss functions by means of the Wasserstein distance and gradient penalty. Importantly, the optimization process differs depending on the algorithm implemented in our model. In the case of a vanilla GAN (AuxGAN), standard stochastic gradient descent is used. While in the WGAN model we employ RMSProp (Tieleman and Hinton, 2012). Finally, in the case of the WGAN-GP, Adam (Kingma and Ba, 2015) optimizer is applied.
Table 1: Number of synonym and antonym word pairs for English, German and Italian in two settings: babelnet, external + babelnet.

|                | Synonyms | Antonyms |
|----------------|----------|----------|
|                | overlap  | disjoint | overlap  | disjoint | overlap  | disjoint |
| English        | 3,522,434| 3,521,366| 3,515,111| 1,358,358| 1,087,814| 807,399  |
| German         | 4,545,045| 4,396,350| 3,515,111| 1,360,040| 1,089,338| 808,605  |
| Italian        | 1,024    | 843      | 1,11      | 99       | 99       | 98       |

Table 2: Spearman’s $\rho$ correlation scores on SimLex-999 (SL), SimVerb-3500 (SV) and WordSim-353 (WS). Evaluation was performed using constraints in three settings: (a) external, (b) babelnet, (c) external + babelnet.

5 Results

5.1 Word Similarity

We report our experimental results with respect to a common intrinsic word similarity task, using standard benchmarks: SimLex-999 and WordSim-353 for English, German and Italian, as well as SimVerb-3500 for English. Each dataset contains human similarity ratings, and we evaluate the similarity measure using the Spearman’s $\rho$ rank correlation coefficient. In Table 2, we present results for English benchmarks, whereas results for German and Italian are reported in Table 3.

Word embeddings are evaluated in two scenarios: disjoint where words observed in the benchmark datasets are removed from the linguistic constraints; and overlap where all words provided in the linguistic constraints are utilized. We use the overlap setting in a downstream task (see 5.2).

In the tasks we report scores for Original (non-specialized) word vectors, initial specialization method Attract-Repel (Mrkšić et al., 2017b), and three post-specialization methods: DFFN (Vulić et al., 2018), AuxGAN (Ponti et al., 2018) and our proposed model WGAN-postspec (in two scenarios: WGAN and WGAN-GP).

The results suggest that the post-specialization methods bring improvements in the specialization of the distributional word vector space. Overall, the highest correlation scores are reported for the models with adversarial losses. We also observe that the proposed WGAN-postspec achieves fairly consistent correlation gains with GloVe vectors on the SimLex dataset. Interestingly, while exploiting additional constraints (i.e. external + babelnet) generally boosts correlation scores for German and Italian, the results are not conclusive in the case of English, and thus they require further investigation.
Table 3: Spearman’s ρ correlation scores on SimLex-999 (SL) and WordSim-353 (WS). Evaluation was performed using constraints in three settings: (a) external, (b) babelnet, (c) external + babelnet.

|        | German Word2Vec | Italian Word2Vec |
|--------|----------------|-----------------|
|        | overlap disjoint | overlap disjoint |
| ORIGINAL | 0.358 0.538 0.538 0.538 0.356 0.563 0.356 0.563 |
| ATTRACT-REPEL | a 0.360 0.537 0.538 0.538 0.376 0.568 0.366 0.559 |
|          | b 0.358 0.538 0.538 0.538 0.366 0.568 0.366 0.559 |
|          | c 0.360 0.538 0.538 0.538 0.378 0.566 0.370 0.564 |
| DFFN    | a 0.366 0.422 0.370 0.452 0.381 0.512 0.365 0.519 |
|          | b 0.354 0.538 0.348 0.538 0.364 0.559 0.361 0.560 |
|          | c 0.359 0.541 0.358 0.533 0.376 0.561 0.369 0.559 |
| AUXGAN  | a 0.331 0.532 0.325 0.535 0.362 0.561 0.348 0.560 |
|          | b 0.369 0.552 0.373 0.561 0.361 0.559 0.364 0.563 |
|          | c 0.369 0.564 0.365 0.556 0.365 0.566 0.368 0.563 |
| WGAN    | a 0.331 0.528 0.327 0.531 0.361 0.558 0.344 0.558 |
|          | b 0.364 0.558 0.367 0.559 0.359 0.553 0.367 0.559 |
|          | c 0.371 0.559 0.364 0.560 0.367 0.567 0.370 0.562 |

Table 4: DST results for English.

5.2 Dialog State Tracking

We also evaluate our proposed approach on a dialog state tracking (DST) downstream task. This task is a standard language understanding task, which allows to differentiate between word similarity and relatedness. To perform the evaluation we follow previous works (Henderson et al., 2014; Williams et al., 2016; Mrkšić et al., 2017b). Concretely, a DST model computes probability based only on pre-trained word embeddings. We use Wizard-of-Oz (WOZ) v.2.0 dataset (Wen et al., 2017; Mrkšić et al., 2017a) composed of 600 training dialogues as well as 200 development and 400 test dialogues.

In our experiments, we report results with a standard joint goal accuracy (JGA) score. The results in Table 4 confirm our findings from the previous word similarity task, as initial semantic specialization and post-specialization (in particular WGAN-postspec) yield improvements over original distributional word vectors. We expect this conclusion to hold in all settings; however, additional experiments for different languages and word embeddings would be beneficial.

6 Conclusion and Future Work

In this work, we presented a method to perform semantic specialization of word vectors. Specifically, we compiled a new set of constraints obtained from BabelNet. Moreover, we improved a state-of-the-art post-specialization method by incorporating adversarial losses with the Wasserstein distance. Our results obtained in an intrinsic and an extrinsic task, suggest that our method yields performance gains over current methods.

In the future, we plan to introduce constraints for asymmetric relations as well as extend our proposed method to leverage them. Moreover, we plan to experiment with adapting our model to a multilingual scenario, to be able to use it in a neural machine translation task. We make the code and resources available at: https://github.com/mbiesialska/wgan-postspec

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work is supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund through the postdoctoral senior grant Ramón y Cajal and by the Agencia Estatal de Investigación through the projects EUR2019-103819 and PCIN-2017-079.

References

Rami Al-Rfou’, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML17, page 214223. JMLR.org.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pages 5012–5019.
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.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. *Proceedings of ICLR*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Christian Fellbaum. 1998. Wordnet: An electronic lexical database mit press.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.

Goran Glavaš and I. Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45, Melbourne, Australia. Association for Computational Linguistics.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 2672–2680, Cambridge, MA, USA. MIT Press.

Ishaan Gulrajani, Furuk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

B.A. Kipfer. 2009. *Roget’s 21st Century Thesaurus (3rd Edition)*. Philip Lief Group.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017a. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017b. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

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

Masatake Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989, Denver, Colorado. Association for Computational Linguistics.
Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 282–293, Brussels, Belgium. Association for Computational Linguistics.

Silke Scheible and Sabine Schulte im Walde. 2014. A database of paradigmatic semantic relation pairs for German nouns, verbs, and adjectives. In Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing, pages 111–119, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Irene Sucameli and Alessandro Lenci. 2017. Parad-it: Eliciting Italian paradigmatic relations with crowdsourcing. In Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2):26–31.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 516–527, New Orleans, Louisiana. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. Dialogue & Discourse, 7:4–33.

Jingwei Zhang, Jeremy Salwen, Michael Glass, and Alfio Gliozzo. 2014. Word semantic representations using Bayesian probabilistic tensor factorization. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1522–1531, Doha, Qatar. Association for Computational Linguistics.