Exposé

Training recursive compositional models with hierarchical linguistic information for semantic tasks in NLP

by Arne Binder

Motivation. Compositional Distributional Semantics Models (CDSMs)(Clark, Coecke, and Sadrzadeh 2008; Grefenstette and Sadrzadeh 2011) are Vector Space Models (VSMs) (Salton, Wong, and C.-S. Yang 1975) that produce vector representations for sequences of tokens by composing word embeddings in a meaningful manner.

CDSMs based on gated Recurrent Neural Networks (RNNs)(Hochreiter and Schmidhuber 1997) produce promising results for internal representations of short to medium length textual input on several semantic tasks like language modeling (Sundermeyer, Schluter, and Ney 2012), parsing (Dyer et al. 2016), image caption generation (Vinyals et al. 2014) or machine translation (Wu et al. 2016), since RNNs are capable of contextualized token processing. But they still fail to handle long range dependencies as they suffer from vanishing gradients and the memory capacity of their inner states is restricted. Furthermore, RNNs are computational expensive because they are inherently sequential and therefore not parallelizable.

Summation or averaging composition models such as fastText (Joulin et al. 2017) demonstrate that avoiding any explicit structure like word order information may also perform quite well at least for short texts. These bag-of-words models are very advantageous in means of training time and memory consumption, therefore enabling for huge amounts of training data, but fail to capture more complex semantic interactions like negation, especially for larger documents.

However, it is up to debate which composition functions perform well for multi-sentence documents while remaining as precise as for short input.

Idea. In this thesis, we analyze recursive neural CDSM trained with hierarchical structured linguistic information. Recursive Neural Networks (RecNNs) (Goller and Kuchler 1996; Socher et al. 2011) generalize RNNs by allowing arbitrary trees as input structure instead of just linear sequences. The consequences of this generalization are twofold: On

the one hand, distances of tokens that are potentially related[1] may decrease with respect to the input graph. This may lead to easier contextualization and, consequently, more precise interpretations of individual tokens. Furthermore, by using functions like averaging or summation to compose child embeddings at the tree nodes, computation cost should decrease. On the other hand, RecNNs introduce another degree of structural complexity and require pre-calculated input structures.

We investigate if it is possible to combine the best of both worlds, the speed and size of bag-of-words models and the context awareness of sequence models, by using tree structured embedding composition. By doing so, it may be possible to enable semantic embedding calculation for a large range of text sizes. In that sense we analyze the performance of tree structured models and compare it with bag-of-words and sequence models with regard to different document sizes.

Implementation. To achieve these goals, we implement the following composition models: 1) a tree structured RecNN model as our candidate model, 2) a bag-of-words model, and 3) a RNN based sequence model. We use linguistic dependency parse information and paragraph structure to construct the tree hierarchy for the candidate model. Since the impact of added structural information is hard to evaluate, we lend from the two competitor models to construct the tree model. Considering one tree node, we use a bag-of-words approach to combine all child embeddings to a single one (reduction) and execute a RNN step to incorporate the current word embedding (mapping). If structure that equals one of the edge cases (i.e. the degenerated trees: sequence or depth one tree) is fed to this tree model, it strongly resembles one of the competitor models. Furthermore, the separation in map and reduce functions allows to experiment if reduction should precede the mapping or the other way around. Depending on its implementation, the former can reduce computational costs as the RNN step may be executed only once per inner tree node. The latter should allow more precise contextualization, but results in as many RNN step executions as for the sequence model.

Recently, the attention mechanism (Bahdanau, Cho, and Bengio 2014; Xu et al. 2015) was successfully applied in neural Natural Language Processing (NLP) tasks (Zhuang and Chang 2017; Vaswani et al. 2017) and gained attention due to its simplicity. We will test whether the tree model will benefit from using attention as reduction function leading to a hierarchical attention model similar to Z. Yang et al. (2016), but in a dynamic fashion.

[1] regarding the interpretation process

Data and Evaluation. As argued in Binder (2018) with regard to the Distributional Hypothesis (Harris 1954) semantic relatedness (Resnik 1999; Budanitsky and Hirst 2006) prediction is one fundamental task to evaluate semantic vector space. Although there are well curated relatedness-labeled datasets at paraphrase and sentence level (Pavlick et al. 2015; Dolan and Brockett 2005; Marelli et al. 2014; Cer et al. 2017), there is a lack of super-sentence relatedness corpora. As we are especially interested in scaling beyond sentence boundaries we seek to circumvent this shortcoming by exploiting interlinking information in Wikipedia articles. We heuristically take an article that is mentioned in the *See Also* section of another one as semantically related to that article. We use this link prediction task to train and evaluate our embedding models. Taking the English portion of Wikipedia results in a dataset of ∼1 million documents[2] that occur in at least one of these links. For computational reasons we restrict the documents to the article abstracts. To bypass as much preprocessing hurdles as possible we make use of the DBpedia NIF (Dojchinovski, Hernandez, and Ackermann 2018) dataset[3]. It consists of cleaned, plain Wikipedia article text, but enhanced with structural information extracted from Wikipedia HTML data such as annotations for sections, paragraphs and titles or anchors for intra-Wikipedia links. We will use this structural data in combination with dependency parse information to dynamically construct the tree model.

Furthermore, we evaluate the resulting embedding models on suitable NLP tasks such as the BioASQ Task A challenge[4]. This real world task requires to predict the Medical Subject Headings (MeSH)[5] assigned to a PubMed abstract. As we focus on the impact of structure to composition, we restrict the BioASQ dataset to the subset of *structured abstracts*[6]. These PubMed abstracts are separated into labeled paragraphs and represent approximately one third of the total BioASQ dataset[7].

Finally, we create term frequency-inverse document frequency (TF-IDF) representations as baseline embeddings and compare the performance on the individual tasks.

[2] of a total of ∼5 million English articles

[3] http://wiki.dbpedia.org/dbpedia-nif-dataset

[4] http://participants-area.bioasq.org/general_information/Task6a/

[5] https://www.nlm.nih.gov/mesh/

[6] https://www.nlm.nih.gov/bsd/policy/structured_abstracts.html

[7] The BioASQ 2018 dataset consists of ∼13.5 million documents

List of Abbreviations

CDSM Compositional Distributional Semantics Model

NLP Natural Language Processing

RecNN Recursive Neural Network

RNN Recurrent Neural Network

TF-IDF term frequency-inverse document frequency

VSM Vector Space Model

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: arXiv: 1409.0473 [cs, stat]. URL: http://arxiv.org/abs/1409.0473 (visited on 01/28/2018).

Binder, Arne (2018). *Comparison of Two Semantic Aware Composition Models for Word Embeddings and Its Relation to Dependency Type Information.*

Budanitsky, Alexander and Graeme Hirst (2006). "Evaluating Wordnet-Based Measures of Lexical Semantic Relatedness". In: *Computational Linguistics* 32.1, pp. 13–47. URL: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.13 (visited on 08/24/2017).

Cer, Daniel et al. (2017). "SemEval-2017 Task 1: Semantic Textual Similarity-Multilingual and Cross-Lingual Focused Evaluation". In: *arXiv preprint arXiv:1708.00055*. URL: https://arxiv.org/abs/1708.00055 (visited on 09/26/2017).

Clark, Stephen, Bob Coecke, and Mehrnoosh Sadrzadeh (2008). "A Compositional Distributional Model of Meaning". In: *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*. Oxford, pp. 133–140.

Dojchinovski, Milan, Julio Hernandez, and Markus Ackermann (2018). "DBpedia NIF: Open, Large-Scale and Multilingual Knowledge Extraction Corpus". In: https://2018.eswc-conferences.org/wp-content/uploads/2018/02/ESWC2018_paper_136.pdf, p. 15.

Dolan, William B. and Chris Brockett (2005). "Automatically Constructing a Corpus of Sentential Paraphrases". In: *Proc. of IWP*. URL: https://pdfs.semanticscholar.org/4753/54f10798f110d34792b6d88f31d6d5cb099e.pdf (visited on 06/12/2017).

Dyer, Chris et al. (2016). "Recurrent Neural Network Grammars". In: arXiv: 1602.07776 [cs]. URL: http://arxiv.org/abs/1602.07776 (visited on 06/02/2018).

Goller, C. and A. Kuchler (1996). "Learning Task-Dependent Distributed Representations by Backpropagation through Structure". In: vol. 1. IEEE, pp. 347–352. ISBN: 978-0-7803-3210-2. DOI: 10.1109/ICNN.1996.548916. URL: http://ieeexplore.ieee.org/document/548916/ (visited on 01/22/2018).

Grefenstette, Edward and Mehrnoosh Sadrzadeh (2011). "Experimental Support for a Categorical Compositional Distributional Model of Meaning". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1394–1404.

Harris, Zellig S. (1954). "Distributional Structure". In: WORD 10.2-3, pp. 146–162. ISSN: 0043-7956, 2373-5112. DOI: 10.1080/00437956.1954.11659520. URL: http://www.tandfonline.com/doi/full/10.1080/00437956.1954.11659520 (visited on 09/05/2017).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural computation* 9 8, pp. 1735–80.

Joulin, Armand et al. (2017). "Bag of Tricks for Efficient Text Classification". In: Association for Computational Linguistics, pp. 427–431. DOI: 10.18653/v1/E17-2068. URL: http://aclweb.org/anthology/E17-2068 (visited on 05/16/2018).

Marelli, Marco et al. (2014). "A SICK Cure for the Evaluation of Compositional Distributional Semantic Models." In: *LREC*, pp. 216–223. URL: http://clic.cimec.unitn.it/marco/publications/marelli-etal-sick-lrec2014.pdf (visited on 06/12/2017).

Pavlick, Ellie et al. (2015). "PPDB 2.0: Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, pp. 425–430. URL: http://www.aclweb.org/anthology/P15-2070.

Resnik, Philip (1999). "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language". In: *J. Artif. Intell. Res.(JAIR)* 11, pp. 95–130. URL: https://www.jair.org/media/514/live-514-1722-jair.pdf.

Salton, Gerard, A. Wong, and Chung-Shu Yang (1975). "A Vector Space Model for Automatic Indexing". In: *Commun. ACM* 18, pp. 613–620.

Socher, Richard et al. (2011). "Parsing Natural Scenes and Natural Language with Recursive Neural Networks". In: p. 8.

Sundermeyer, Martin, Ralf Schluter, and Hermann Ney (2012). "LSTM Neural Networks for Language Modeling". In: p. 4.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 08/17/2017).

Vinyals, Oriol et al. (2014). "Show and Tell: A Neural Image Caption Generator". In: arXiv: 1411.4555 [cs]. URL: http://arxiv.org/abs/1411.4555 (visited on 05/18/2018).

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: arXiv: 1609.08144 [cs]. URL: http://arxiv.org/abs/1609.08144 (visited on 02/13/2018).

Xu, Kelvin et al. (2015). "Show, Attend and Tell: Neural Image CaptionGeneration with Visual Attention". In: p. 10.

Yang, Zichao et al. (2016). "Hierarchical Attention Networks for Document Classification". In: Association for Computational Linguistics, pp. 1480–1489. DOI: 10.18653/v1/N16-1174. URL: http://aclweb.org/anthology/N16-1174 (visited on 05/18/2018).

Zhuang, Wenli and Ernie Chang (2017). "Neobility at SemEval-2017 Task 1: An Attention-Based Sentence Similarity Model". In: arXiv: 1703.05465 [cs]. URL: http://arxiv.org/abs/1703.05465 (visited on 06/20/2017).