Token-Level Metaphor Detection using Neural Networks

Erik-Lân Do Dinh[†] **& Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de

Abstract

Automatic metaphor detection usually relies on various features, incorporating e.g. selectional preference violations or concreteness ratings to detect metaphors in text. These features rely on background corpora, hand-coded rules or additional, manually created resources, all specific to the language the system is being used on. We present a novel approach to metaphor detection using a neural network in combination with word embeddings, a method that has already proven to yield promising results for other natural language processing tasks. We show that foregoing manual feature engineering by solely relying on word embeddings trained on large corpora produces comparable results to other systems, while removing the need for additional resources.

1 Introduction

According to Lakoff and Johnson (1980), metaphors are cognitive mappings of concepts from a source to a target domain. While in some works identifying those mappings (*conceptual metaphors*) themselves is the subject of analysis, we concern ourselves with detecting their manifestations in text (*linguistic metaphors*).

Various features have been designed to model either representation of metaphor, prominently e.g. violations of (generalized) selectional preferences in grammatical relations (Wilks, 1978; Shutova, 2013), concreteness ratings to model the difference between source and target concepts (Tsvetkov et al., 2014; Turney and Assaf, 2011), supersenses and hypernym relations (Tsvetkov et al., 2013), or topic models (Heintz et al., 2013; Beigman Klebanov et al., 2014). Some of these features can be obtained in an unsupervised way, but many require additional resources such as concreteness databases or word taxonomies. While this is a good approach for resource-rich languages, this poses problems for languages where such resources are not readily available. Approaches to alleviate this issue often make use of bilingual dictionaries or machine translation (Tsvetkov et al., 2013; Dunn et al., 2014), in itself introducing the need for a new resource resp. introducing a possible new source for misclassification.

In this paper, we present a novel approach for metaphor detection using neural networks on the token-level in running text, relying solely on word embeddings used in context. In recent years, neural networks have been used to solve natural language processing tasks with great effect, but so far have not been applied to metaphor detection. While our approach still has to be tested on data in other languages, it already shows promising results on English data, all the more considering it is not using an elaborate feature set, deriving the representation only from distributed and local context.

We start in Section 2 by discussing previous work on metaphor detection which compares to our work in at least one aspect: granularity of classification, language/resource independence, or the usage of word embeddings. Section 3 details the architecture of our neural network. In Section 4, we describe the used resources and training data and present the results of our method. Concluding this paper, in Section 5 we also give an outlook for future work.

Proceedings of The Fourth Workshop on Metaphor in NLP, pages 28–33,
San Diego, CA, 17 June 2016. ©2016 Association for Computational Linguistics

2 Related Work

While some approaches aim at inferring conceptual metaphors, here we mainly discuss works which detect linguistic metaphors, although differing in classification granularity (i.e. detection on token, construction, or sentence level).

Beigman Klebanov et al. (2014) used logistic regression to assess (binary) metaphoricity on the token level, considering only content words. They built upon their work which used unigram, POS, topic model, and concreteness features from a concreteness ratings list, by implementing a reweighing scheme to correct for the imbalanced class distribution in metaphor data (Beigman Klebanov et al., 2015). Re-weighting significantly improved results on their test data, and allows for task-dependent tuning to focus on precision or recall. They also experiment with differences between concreteness ratings in certain grammatical constructions, interpreting the rather small performance increases as an indicator of concreteness possibly not being a defining factor of metaphoricity.

An ensemble approach to detect sentences containing metaphors has been implemented by Dunn et al. (2014). Their system extracts candidate token pairs from parsed text using manually defined syntactic patterns, before applying classifiers—one of them being an overlap classifier for source and target concepts. Employing concreteness ratings it selects the most concrete words related to a given concept, and WordNet for extracting a set of basic semantic categories to which the concepts are mapped. They further use machine translation for concreteness ratings and non-English WordNet-like taxonomies to extend the system to different languages.

Tsvetkov et al. (2013) employed random forests for metaphor detection on adjective-noun constructions. Database-provided features such as concreteness and imageability values are complemented by low-dimensional word embeddings and supersenses. By means of bilingual dictionaries they test their system on datasets in different languages, allowing for the continued use of English resources for feature extraction (e.g. concreteness and imageability databases, WordNet for supersense extraction, etc.).

As an exception to the other discussed systems, Mohler et al. (2014) present a complex system for conceptual metaphor detection, which requires linguistic metaphors identified by a separate system as the input data. For clustering words into conceptual classes, they compare the usage of dependency based vector representations, traditional LSA, and dense word embeddings. Their concept-mapping approach yields significant improvements in accuracy for three languages when using word embeddings; however, for English, LSA produced the best results.

Schulder and Hovy (2014) provide an inherently resource and language independent method by using tf.idf for metaphor detection, requiring only large background corpora and a small set of training sentences. Employing a bootstrapping approach with manually selected seed terms, they achieve rather modest results based on tf.idf alone, but emphasize its potential usefulness as a feature in more advanced multi-feature detection systems.

Closely related, work in metonymy identification also has made use of word embeddings. Among other features commonly used for metaphor detection, such as abstractness or selectional restrictions, Zhang and Gelernter (2015) test different representations of words to detect metonymy using an SVM. To that end they employ word embeddings, LSA, and one-hot-encoding—however, their results do not show clear superiority of one representation over the others.

3 Neural Networks for Metaphor Detection

We propose a neural network approach to metaphor detection in conjunction with word embeddings, i.e. dense vector representations of words. For this purpose, we experiment with multilayer perceptrons (MLP), fully connected feedforward neural networks with an input layer, one or more hidden layers, and an output layer (Figure 1). MLPs have been successfully applied to a variety of standard NLP preprocessing tasks, such as part-of-speech (POS) tagging, chunking, or named entity recognition, but also to more advanced tasks like semantic role labeling (Collobert et al., 2011). In this spirit, we treat metaphor detection as a tagging problem. To that end, we extend and modify a framework for named entity recognition (Reimers et al., 2014),

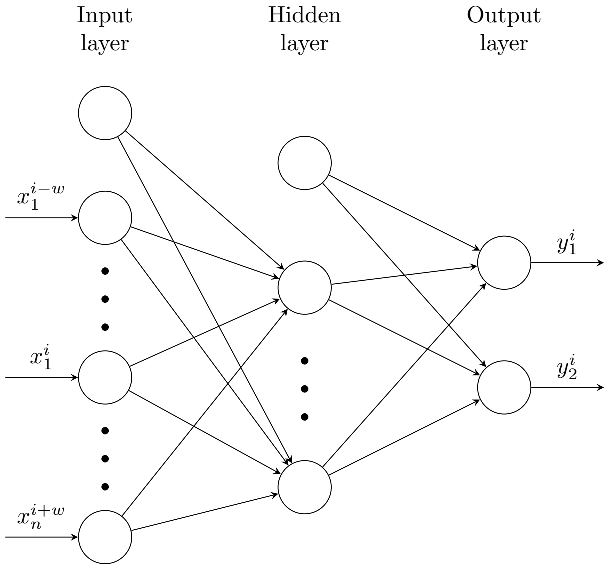

Figure 1: Multilayer perceptron with one hidden layer. Input x for a token at position i is a concatenation of word embedding vectors x^{i-w}, ..., x^{i+w} within a given window w, modeling the local context of the token. The input vector x can be enriched with additional features by concatenating embeddings or numerical values. The output layer uses logistic regression to calculate y_1^i and y_2^i, interpretable as probabilities for metaphorical and literal use of token i.

which is built using the Python deep learning library Theano (Bastien et al., 2012).

The input layer is using a lookup function to extract existing, pre-trained word embeddings for all content words in the data set. More specifically, for each content token, we concatenate the embeddings of the surrounding tokens within a window of $w = 5$ (including non-content tokens, and padded with a randomly created embedding at sentence begin and end), which showed the most promising results in preliminary tests. The rationale of re-using this window approach employed by Collobert et al. (2011) is that, similar to POS tagging on a functional level, context is needed for metaphor detection on a semantic level. We also conduct experiments where we additionally use POS tag embeddings as features, as well as concreteness ratings for comparison purposes. Embeddings/Values for those additional features are appended to the respective word embedding, before concatenating them. We set the hyperbolic tangent as the non-linear activation function in our hidden layer. To prevent overfitting, we employ dropout, which essentially amounts to the creation of several unique neural networks where nodes in the hidden layer are removed at random—those unique networks are averaged over the batches. At test time, the complete network is used. The output layer is implemented as a softmax layer which employs logistic regression to classify a token as being used metaphorically or literally.

We use stochastic gradient descent to train the network, employing negative log-likelihood as the loss function to minimize. In addition to learning the parameters of the hidden and output layer via backpropagation, we also extend that learning to the input (lookup) layer, which allows us to adapt the pretrained embeddings to the training data. We adjust the learning rate using an annealing schedule, decreasing it linearly with respect to epochs, but keeping it above a set minimum.

The network design thus incorporates various hyper-parameters, of which the following had the largest impact: window size, number of nodes in the hidden layer, initial learning rate (which is gradually lowered to ensure convergence), and mini-batch size (which determines over how many instance vectors the weight updates should be averaged). We use grid search to determine the best performing setting of our network on the validation set(s), and tune it according to the best average performance (F1) over different subcorpora of the used data (Table 1).

Genre	content tokens	met. tokens
academic	36,015	15.6%
conversation	19,807	10.6%
fiction	23,163	14.5%
news	24,880	19.6%
overall	103,865	15.4%

Table 1: Subcorpora of the VU Amsterdam Metaphor Corpus, showing contained content tokens (with POS noun, verb, adjective, adverb), and percentage of metaphorically used tokens.

4 Experiments

For our experiments, we use the pre-trained 300-dimensional word embeddings created with word2vec[1] using the Google News dataset (Mikolov

[1] `https://code.google.com/p/word2vec/`

Genre	B	Token			Token+POS			Token+POS+Conc		
	F1	P	R	F1	P	R	F1	P	R	F1
acprose	.2650	.5791	.5382	.5579	.5775	.5426	.5595	.5841	.5412	**.5618**
convrsn	.2056	.6561	.4863	.5586	.6597	.4941	**.5650**	.6595	.4784	.5545
fiction	.2824	.5597	.4637	.5072	.5703	.4725	**.5168**	.5819	.4527	.5093
news	.3240	.6412	.6293	.6352	.6480	.6293	**.6385**	.6431	.6190	.6308
average	.2735	.6034	.5415	.5707	.6074	.5460	**.5751**	.6110	.5359	.5710
complete	.2703	.5834	.5235	.5518	.5879	.5255	.5550	.5899	.5355	**.5614**

Table 2: Results on the test sets for the tuned neural network. Showing precision (P), recall (R) and F1-measure (F1) regarding metaphorically used tokens, for different feature combinations and VUAMC subcorpora, as well as for the whole corpus (complete). *B* denotes a pseudo-baseline classifying all tokens as metaphorical.

et al., 2013). Training and testing data is taken from the VU Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010), a subset of the BNC Baby in which each token is annotated as being used literally or metaphorically. This is done using various fine-grained tags; however, we only use the most clear cut tag *mrw* (metaphor-related word), labeling everything else as literal. Furthermore, because generally only the detection of metaphoricity of content tokens is of interest, we only incorporate labels for tokens having one of the following POS tags (as supplied with the VUAMC): noun, verb, adjective, adverb. Also auxiliary verbs, having lemmas *have*, *be*, or *do*, were filtered out. An overview over the remaining tokens in the subcorpora can be seen in Table 1. The system is trained on each contained genre (*news*, *conversation*, *fiction*, *academic*) separately; for each subcorpus we use a random subset of 76% of the data as a training set, 12% as development set and 12% as test set. We also extend our system to incorporate 10-dimensional POS embeddings, which were initialized randomly and updated over the course of the network's training phase. For comparison, we finally include concreteness values taken from Brysbaert et al. (2014), and train an additional network with it.

The reported inter-annotator agreement for the VUAMC is 0.84 in terms of Fleiss' Kappa—in comparison, treating the gold annotations in the test set (complete) and our system as annotators, we achieve 0.56, indicating room for improvement. More detailed results for our final network as evaluated on the test sets can be seen in Table 2. We observe that extending the word embeddings with the 10-dimensional POS embeddings only has a small influence on the results. Adding the concreteness values does not significantly change the results compared to the token-only based approach, which could be due to several factors. Firstly, the used concreteness values cover only about 80% of the data, with the remaining 20% being assigned a default neutral value. The one-dimensionality of the concreteness feature is also likely to be part of the problem, demanding for a better representation. Last, there is a chance of the word embeddings implicitly capturing the concreteness which needs to be investigated further.

We added a pseudo-baseline (B) where each token is labeled as metaphorical; this is handily beaten by our system. However, more informative is a comparison with the work done by other researchers. Beigman Klebanov et al. (2015) use the same classification granularity for their experiments, but employ cross-validation on 77% of the VUAMC for their evaluation. Still, comparing the results gives an indication of our system's performance. Their best performing feature set shows similar performance on the *academic* dataset, achieving an F1 score of 0.564 compared to 0.558 for our token-only based approach; the *fiction* subcorpus yields 0.493, respectively 0.507. Larger gaps can be observed on the *news* and especially the *conversation* sets, where they report results of 0.590 and 0.396 respectively, compared to our scores of 0.635 and 0.559. The strong performance on the *news* subcorpus can partly be attested to our choice of word embeddings, which were constructed using news texts and thus best capture usage of words in this genre.

We also trained and tested our network on the complete corpus (*complete*, again using a 76%/12%/12% split), with the results indicating that it generalizes rather well. Looking at the data, we can observe some limitations to our approach. E.g., in the sentence "To throw up an impenetrable Berlin Wall between you and them could be tactless.", "Berlin" and "Wall" are wrongly being tagged as literal, because the used context window is too small to detect their metaphoric usage. Similar problems occur in other cases where insufficient information is available, because of too short sentences, and also at the beginning or end of a sentence, where the vectors are padded with generic embeddings and thus contain less information. Such cases could be treated by using larger parts of text instead of sentences, or by adding topical information gained in unsupervised fashion, as it has been done in related work, e.g. via topic models.

Unique problems arise with classification of tokens in the (often colloquial) *conversation* texts. The sentences in these transcripts are sometimes missing words, have non-grammatical structure, or are wrongly split. For example, consider the sentence, "Yeah, I want a whole with that whole." The only content words are "want", "whole", and "whole", of which the latter two are being wrongly tagged as metaphoric by our system. However, even additional context likely would not improve the classification in this case—because of missing words and incomplete sentences, even humans have a hard time grasping the meaning of this sentence in (textual) context, let alone assessing metaphoricity.

When examining errors by POS tag (Table 3), we note that verbs get misclassified twice as often as nouns, which seems intuitive given that verbs generally are more polysemous than nouns. We can observe increased error rates for tokens that are tagged as being ambiguous between nouns and other POS; considering only these, the percentage of misclassified tokens is double that of the whole noun set. Proper nouns are being tagged as literal by the system in all cases, differing from the gold annotations for just 5 out of 813 instances. However, a metaphoric meaning for 4 of those annotations could be disputed—at least, these annotations are inconsistent with other annotations in the corpus.

POS	misclassified	total	percentage
Noun	556	5700	9.75%
Verb	673	3449	19.51%
Adj./Adverb	468	3629	12.90%

Table 3: Number of misclassified tokens in the *complete* test set, detailed by coarse-grained POS tag.

Two subclasses of adverbs stand out as yielding substantially higher error rates than the remaining adjectives and adverbs: adverbially used prepositions (e.g. "out" in "carry out"), and borderline cases between adverb and preposition (e.g. "on" in "what's going on"). Arguably these could be considered non-content words and filtered out as well, which would further increase F1 values of our system.

5 Conclusion

We presented a novel approach for supervised metaphor detection, combining pre-trained word embeddings with a neural network architecture. Although we showed that our approach works well with English data, comparing favorably to a state-of-the-art system that employs elaborate features, we still need to examine the performance of our approach on other corpora and different languages, e.g. for historical German data, for which annotated data is still under construction. However, we deem our approach especially suited for this kind of task, as it already shows promising results without any additional features other than basic word embeddings, which can be created in an unsupervised fashion. In future work, we also want to analyze the impact of using genre-specific embeddings, as well as the influence of the training set size on the neural network and its results. Natural next steps also include experiments with more advanced network structures such as Recurrent Neural Networks, specifically Long-Short Term Memory networks (LSTMs).

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the German Institute for Educational Research (DIPF) as part of the graduate program "Knowledge Discovery in Scientific Literature" (KDSL).

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: new features and speed improvements. In *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*, pages 1–10, Stateline, NV, USA.

Beata Beigman Klebanov, Chee Wee Leong, Michael Heilman, and Michael Flor. 2014. Different Texts, Same Metaphors: Unigrams and Beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, Baltimore, MD, USA. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised Word-Level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, CO, USA. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–11.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Jonathan Dunn, Jon Beltran de Heredia, Maura Burke, Lisa Gandy, Sergey Kanareykin, Oren Kapah, Matthew Taylor, Dell Hines, Ophir Frieder, David Grossman, Newton Howard, Moshe Koppel, Scott Morris, Andrew Ortony, and Shlomo Argamon. 2014. Language-Independent Ensemble Approaches to Metaphor Identification. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 6–12.

Ilana Heintz, Ryan Gabbard, Donald S Black, Marjorie Freedman, Ralph Weischedel, and San Diego. 2013. Automatic Extraction of Linguistic Metaphor with LDA Topic Modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, GA, USA. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago, IL, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, pages 1–9, Stateline, NV, USA.

Michael Mohler, Bryan Rink, David Bracewell, and Marc Tomlinson. 2014. A Novel Distributional Approach to Multilingual Conceptual Metaphor Recognition. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, number 2, pages 1752–1763.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. GermEval-2014: Nested Named Entity Recognition with Neural Networks. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 117–120, Hildesheim, Germany. Universitätsverlag Hildesheim.

Marc Schulder and Eduard Hovy. 2014. Metaphor Detection through Term Relevance. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 18–26, Baltimore, MD, USA. Association for Computational Linguistics.

Ekaterina Shutova. 2013. Metaphor Identification as Interpretation. In *Proceedings of *SEM*, pages 276–285, Atlanta, GA, USA. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU.* John Benjamins, Amsterdam.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-Lingual Metaphor Detection Using Common Semantic Features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, GA, USA. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D Turney and Dan Assaf. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, United Kingdom. Association for Computational Linguistics.

Yorick Wilks. 1978. Making Preferences More Active. *Artificial Intelligence*, 11(3):197–223.

Wei Zhang and Judith Gelernter. 2015. Exploring Metaphorical Senses and Word Representations for Identifying Metonyms. *http://arxiv.org/abs/1508.04515*.