# Learning Outside the Box: Discourse-level Features Improve Metaphor Identification

**Jesse Mu**[1,3], **Helen Yannakoudakis**[2], **Ekaterina Shutova**[4]
[1]Computer Science Department, Stanford University, USA
[2]The ALTA Institute, [3]Dept. of CS & Technology, University of Cambridge, UK
[4]ILLC, University of Amsterdam, The Netherlands
muj@stanford.edu, hy260@cl.cam.ac.uk, e.shutova@uva.nl

## Abstract

Most current approaches to metaphor identification use restricted linguistic contexts, e.g. by considering only a verb's arguments or the sentence containing a phrase. Inspired by pragmatic accounts of metaphor, we argue that broader discourse features are crucial for better metaphor identification. We train simple gradient boosting classifiers on representations of an utterance and its surrounding discourse learned with a variety of document embedding methods, obtaining near state-of-the-art results on the 2018 VU Amsterdam metaphor identification task without the complex metaphor-specific features or deep neural architectures employed by other systems. A qualitative analysis further confirms the need for broader context in metaphor processing.

## 1 Introduction

From *bottled up anger* to *the world is your oyster*, metaphor is a defining component of language, adding poetry and humor to communication (Glucksberg and McGlone, 2001) and serving as a tool for reasoning about relations between concepts (Lakoff and Johnson, 1980). Designing metaphor processing systems has thus seen significant interest in the NLP community, with applications from information retrieval (Korkontzelos et al., 2013) to machine translation (Saygin, 2001).

An important first step in any metaphor processing pipeline is metaphor *identification*. To date, most approaches to its identification operate in restricted contexts, for instance, by only considering isolated verb–argument pairs (e.g. *deflate economy*) (Rei et al., 2017) or the sentence containing an utterance (Gao et al., 2018). However, wider context is crucial for understanding metaphor: for instance, the phrase *drowning students* can be interpreted as literal (in the context of *water*) or metaphorical (in the context of *homework*). Often the context required extends beyond the immediate sentence; in Table 1, coreferences (*them*) must be resolved to understand the arguments of a verb, and a *game* is metaphorical in a political context. Indeed, a rich linguistic tradition (Grice, 1975; Searle, 1979; Sperber and Wilson, 1986) explains metaphor as arising from violations of expectations in a conversational context.

| |
|---|
| "You can't steal their ideas." "No, idiot—not so I can **steal** them." |
| Britain still can't decide when to **play** the mandarinate game of of silence [. . . ] interests and concern of the Chinese government. |

Table 1: Metaphorical examples from the VUA dataset in context. Verb is bolded, arguments underlined. Immediate sentence in black, with further context in gray.

Following these theories, in this paper we argue that metaphor processing models should expand beyond restricted contexts to use representations of wider discourse. We support this claim with two contributions: (1) we develop metaphor identification models which take as input an utterance, its immediate lexico–syntactic context, and broader discourse representations, and demonstrate that incorporating discourse features improves performance; (2) we perform a qualitative analysis and show that broader context is often required to correctly interpret metaphors. To the best of our knowledge, this is the first work to investigate the effects of broader discourse on metaphor identification.[1]

## 2 Related work

Metaphor identification is typically framed as a binary classification task, either with (1) word tu-

[1]Code and data available at https://github.com/jayelm/broader-metaphor.



*Proceedings of NAACL-HLT 2019*, pages 596–601
Minneapolis, Minnesota, June 2 - June 7, 2019. 

ples such as SVO triples (car *drinks* gasoline) or (2) whole sentences as input, where the goal is to predict the metaphoricity of a token in the sentence. Recent work has used a variety of features extracted from these two types of contexts, including selectional preferences (Shutova, 2013; Beigman Klebanov et al., 2016), concreteness/imageability (Turney et al., 2011; Tsvetkov et al., 2014), multi-modal (Tekiroglu et al., 2015; Shutova et al., 2016) and neural features (Do Dinh and Gurevych, 2016; Rei et al., 2017).

At the recent VU Amsterdam (VUA) metaphor identification shared task (Leong et al., 2018), neural approaches dominated, with most teams using LSTMs trained on word embeddings and additional linguistic features, such as semantic classes and part of speech tags (Wu et al., 2018; Stemle and Onysko, 2018; Mykowiecka et al., 2018; Swarnkar and Singh, 2018). Most recently, Gao et al. (2018) revisited this task, reporting state-of-the-art results with BiLSTMs and contextualized word embeddings (Peters et al., 2018). To the best of our knowledge, none of the existing approaches have utilized information from wider discourse context in metaphor identification, nor investigated its effects.

## 3 Data

Following past work, we use the *Verbs* subset of the VUA metaphor corpus (Steen et al., 2010) used in the above shared task. The data consists of 17240 training and 5873 test examples, equally distributed across 4 genres of the British National Corpus: Academic, Conversation, News, and Fiction. All verbs are annotated as metaphorical or literal in these texts. We sample 500 examples randomly from the training set as a development set.

## 4 Models

For each utterance, our models learn generic representations of a *verb lemma*,[2] its syntactic *arguments*, and its broader discourse *context*. We concatenate these features into a single feature vector and feed them into a gradient boosting decision tree classifier (Chen and Guestrin, 2016).[3] By observing performance differences when using the lemma only (L), lemma + arguments (LA), or lemma + arguments + context (LAC), we can investigate the effects of including broader context.

To obtain arguments for verbs, we extract subjects and direct objects with Stanford CoreNLP (Manning et al., 2014). 67.4% of verb usages in the dataset have at least one argument; absent arguments are represented as zero vectors. To obtain the broader *context* of a verb, we take its surrounding paragraph as defined by the BNC; the average number of tokens in a context is 97.3. Figure 1 depicts the feature extraction and classification pipeline of our approach.

To learn representations, we use several widely-used embedding methods:[4]

**GloVe** We use 300-dimensional pre-trained GloVe embeddings (Pennington et al., 2014) trained on the Common Crawl corpus as representations of a lemma and its arguments. To learn a context embedding, we simply average the vectors of the tokens in the context. Out-of-vocabulary words are represented as a mean across all vectors.

**doc2vec** We use pretrained 300-dimensional paragraph vectors learned with the distributed bag-of-words method of Le and Mikolov (2014) (colloquially, doc2vec), trained on Wikipedia (Lau and Baldwin, 2016). Here, paragraph vectors are learned to predict randomly sampled words from the paragraph, ignoring word order. To extract representations for verbs and arguments, we embed one-word "documents" consisting of only the word itself.[5] We use a learning rate $\alpha = 0.01$ and 1000 epochs to infer vectors.

**Skip-thought** We use pretrained skip-thought vectors (Kiros et al., 2015) learned from training an encoder–decoder model to reconstruct the surrounding sentences of an input sentence from the Toronto BooksCorpus (Zhu et al., 2015). From this model, we extract 4800-dimensional representations for verb lemma, arguments, and contexts.

**ELMo** Finally, we use ELMo, a model of deep contextualized word embeddings (Peters et al., 2018). We extract 1024-dimensional representations from the last layer of a stacked BiLSTM

[2]The lemmatized form of the verb has improved generalization in other systems (Beigman Klebanov et al., 2016).

[3]We use the default parameters of the XGBoost package: a maximum tree depth of 3, 100 trees, and $\eta = 0.1$.

[4]These methods differ significantly in dimensionality and training data. Our intent is not to exhaustively compare these methods, but rather claim generally that many embeddings give good performance on this task.

[5]Since some methods provide only document embeddings and not word embeddings, for consistency, in all methods we use the same embedding process even for single-word verbs and arguments.

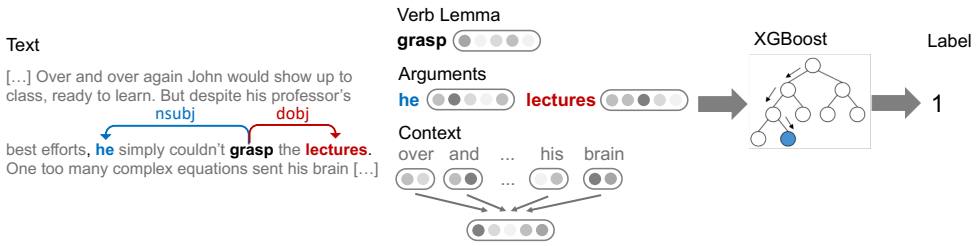


Figure 1: The general feature extraction and classification pipeline of our approach.

trained on Wikipedia and monolingual news data from WMT 2008–2012. To learn embeddings for verbs and arguments, we extract representations for sentences containing only the word itself. To learn context embeddings, we again average the constituent word embeddings.

## 5 Evaluation

For each embedding method, we evaluate the three configurations of features—L, LA, and LAC—on the VUA shared task train/test split, reporting precision, recall and F1 score. Since we are interested in whether incorporating broader context significantly improves identification performance, we compare successive model predictions (LAC vs. LA; LA vs. L) using the *mid-p* variant of McNemar's test for paired binary data (Fagerland et al., 2013).

### 5.1 Comparison Systems

We first compare our models to the baselines of the VUA shared task (Leong et al., 2018): *Baseline 1*, a logistic regression classifier trained only on one-hot encodings of verb lemmas; and *Baseline 2*, the same classifier with additional WordNet class and concreteness features. We also compare to the best systems submitted to the VUA shared task: Wu et al. (2018), an ensemble of 20 CNN-BiLSTMs trained on word2vec embeddings, part-of-speech tags, and word embedding clusters; and Stemle and Onysko (2018), a BiLSTM trained on embeddings from English language learner corpora.

### 5.2 Results

Results for our models are presented in Table 2. Interestingly, most of the simple lemma models (L) already perform at Baseline 2 level, obtaining F1 scores in the range 60–62. This is likely due to the generalization made possible by dense representations of lemmas (vs. one-hot encodings) and the more powerful statistical classifier used. As expected, the addition of argument information consistently enhances performance.

| Model | | P | R | F1 |
|---|---|---|---|---|
| Baseline 1 (lemma) | | 51.0 | 65.4 | 57.3 |
| Baseline 2 (+WN, concrete) | | 52.7 | 69.8 | 60.0 |
| Stemle and Onysko (2018) | | 54.7 | **77.9** | 64.2 |
| Wu et al. (2018) | | **60.0** | 76.3 | **67.2** |
| GloVe | L (lemma) | 51.6 | 74.1 | 60.8 |
| | LA (+ args) | 54.0 | 74.4 | 62.6$^{***}$ |
| | LAC (+ ctx) | 56.7 | 76.8 | 65.2$^{***}$ |
| doc2vec | L | 48.8 | 72.1 | 58.2 |
| | LA | 50.5 | 71.4 | 59.1$^{**}$ |
| | LAC | 52.7 | 72.2 | 60.9$^{***}$ |
| skip-thought | L | 53.5 | 76.1 | 62.8 |
| | LA | 57.0 | 74.0 | 64.3$^{***}$ |
| | LAC | 59.5 | 75.4 | 66.5$^{***}$ |
| ELMo | L | 51.3 | 74.9 | 60.9 |
| | LA | 56.0 | 73.5 | 63.6$^{***}$ |
| | LAC | 58.9 | 77.1 | 66.8$^{***}$ |

$^{**},^{***}$ Significant improvement over previous model ($p < 0.01, 0.001$).

Table 2: Metaphor identification results.

Crucially, the addition of broader discourse context improves performance for all embedding methods. In general, we observe consistent, statistically significant increases of 2-3 F1 points for incorporating discourse. Overall, all LAC models except doc2vec exhibit high performance, and would have achieved second place in the VUA shared task. These results show a clear trend: the incorporation of discourse information leads to improvement of metaphor identification performance across models.

Table 3 displays the performance breakdown by genre in the VUA test set for our best performing model (ELMo LAC) and selected comparison systems. Echoing Leong et al. (2018), we observe that the *Conversation* and *Fiction* genres are consistently more difficult than the *Academic* and *News* genres across all models. This is partially because in this dataset, metaphors in these genres are rarer, occuring 35% of the time in *Academic* and 43% in *News*, but only 15% in *Conversation* and 24% in *Fiction*. In addition, for our model specifically, Conversation genre contexts are much

| Genre | Model | P | R | F1 |
|---|---|---|---|---|
| Academic | Baseline 2 | 70.7 | 83.6 | **76.6** |
| | Wu et al. (2018) | **74.6** | 76.3 | 75.5 |
| | ELMo LAC | 65.4 | **86.8** | 74.6 |
| Conversation | Baseline 2 | 30.1 | 82.1 | 44.1 |
| | Wu et al. (2018) | 40.3 | 65.6 | **50.3** |
| | ELMo LAC | **42.6** | 56.0 | 48.4 |
| Fiction | Baseline 2 | 40.7 | 66.7 | 50.6 |
| | Wu et al. (2018) | **54.5** | **78.4** | **57.6** |
| | ELMo LAC | 48.2 | 63.0 | 54.6 |
| News | Baseline 2 | 67.7 | 68.9 | 68.3 |
| | Wu et al. (2018) | **69.4** | 74.4 | 71.8 |
| | ELMo LAC | 65.2 | **80.0** | **71.8** |

Table 3: Performance breakdown by genre for ELMo LAC model and comparison systems.

| | Args | Sentence | Paragraph |
|---|---|---|---|
| Overall | 40 | 49 | 11 |
| Model errors | | | |
| ELMo L | 37 | 50 | 13 |
| ELMo LA | 36 | 49 | 15 |
| ELMo LAC | 39 | 53 | 8 |

Table 4: Types of context required to interpret metaphors in the development set, both overall (first row) and for model errors. Each row is a separate (but overlapping) sample from the development set.

shorter on average (23.8 vs. 97.3).

Our best performing model (ELMo LAC) is within 0.4 F1 score of the first-place model in the VUA shared task (Wu et al., 2018). The GloVe LAC model would also have obtained second place at 65.2 F1, yet is considerably simpler than the systems used in the shared task, which employed ensembles of deep neural architectures and hand-engineered, metaphor-specific features.

## 6 Qualitative analysis

To better understand the ways in which discourse information plays a role in metaphor processing, we randomly sample 100 examples from our development set and manually categorize them by the amount of context required for their interpretation. For instance, a verb may be interpretable when given just its arguments (direct subject/object), it may require context from the enclosing sentence, or it may require paragraph-level context (or beyond). We also similarly analyze 100 sampled errors made on the development set by the ELMo L, LA, and LAC models, to examine whether error types vary between models.

Our analysis in Table 4 shows that 11% of examples in the development set require paragraph-level context for correct interpretation. Indeed, while such examples are frequently misclassified by the L and LA models (13%, 15%), the error rate is halved when context is included (8%).

Table 5 further presents examples requiring at least paragraph-level context, along with gold label and model predictions. Out of the 31 unique such examples identified in the above analyses, we found 11 (35%) requiring explicit coreference resolution of a pronoun or otherwise underspecified noun (e.g. Table 5 row 1) and 5 (16%) which reference an entity or event implicitly (*ellipsis*; e.g. Table 5 row 2). However, we also observed 4 errors (13%) due to examples with non-verbs and incomplete sentences and 11 examples (35%) where not even paragraph-level context was sufficient for interpretation, mostly in the *Conversation* genre, demonstrating the subjective and borderline nature of many of the annotations.

This analysis shows *a priori* the need for broader context beyond sentence-level for robust metaphor processing. Yet this is not an upper bound on performance gains; the general improvement of the LAC models over LA shows that even when context is not strictly necessary, it can still be a useful signal for identification.

## 7 Conclusion

We presented the first models which leverage representations of discourse for metaphor identification. The performance gains of these models demonstrate that incorporating broader discourse information is a powerful feature for metaphor identification systems, aligning with our qualitative analysis and the theoretical and empirical evidence suggesting metaphor comprehension is heavily influenced by wider context.

Given the simplicity of our representations of context in these models, we are interested in future models which (1) use discourse in more sophisticated ways, e.g. by modeling discourse relations or dialog state tracking (Henderson, 2015), and (2) leverage more sophisticated neural architectures (Gao et al., 2018).

## Acknowledgments

We thank anonymous reviewers for their insightful comments, Noah Goodman, and Ben Leong for assistance with the 2018 VUA shared task data.

| Sentence | Gold label | LA | LAC |
|---|---|---|---|
| A major complication [. . . ] is that the environment can rarely be treated as in a laboratory experiment. **Given** this, determining the nature of the interactions between the variables becomes a matter of major difficulty. | 0 | 1 | 0 |
| For example, on high policy common opinion said that there was nothing for it but to stay in the ERM. He **stayed** in, and the recession worsened. | 1 | 0 | 1 |

Table 5: Examples where context helps, along with gold label (0 – literal; 1 – metaphor) and model predictions (LA, LAC). Verb is bolded, arguments underlined. Additional context (needed for interpretation) in gray.

We thank the Department of Computer Science and Technology and Churchill College, University of Cambridge for travel funding. Jesse Mu is supported by a Churchill Scholarship and an NSF Graduate Research Fellowship. Helen Yannakoudakis was supported by Cambridge Assessment, University of Cambridge. We thank the NVIDIA Corporation for the donation of the Titan GPU used in this research.

## References

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33.

Morten W Fagerland, Stian Lydersen, and Petter Laake. 2013. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13(1):91.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613.

Sam Glucksberg and Matthew S McGlone. 2001. *Understanding figurative language: From metaphor to idioms*. Oxford University Press, Oxford.

Herbert P Grice. 1975. Logic and conversation. In Peter Cole and Jerry L Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. Academic Press, New York.

Matthew Henderson. 2015. Machine learning for dialog state tracking: A review. In *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 39–47.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Agnieszka Mykowiecka, Aleksander Wawer, and Malgorzata Marciniak. 2018. Detecting figurative word occurrences using recurrent neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 124–127.

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*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546.

Ayse Pinar Saygin. 2001. Processing figurative language in a multi-lingual task: Translation, transfer and metaphor. In *Proceedings of the Workshop on Corpus-based and Processing Approaches to Figurative Language*.

John Searle. 1979. Metaphor. In *Expression and Meaning: Studies in the Theory of Speech Acts*, pages 76–116. Cambridge University Press, Cambridge and New York.

Ekaterina Shutova. 2013. Metaphor identification as interpretation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 276–285.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Harvard University Press, Cambridge, MA.

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 Publishing Company, Amsterdam.

Egon Stemle and Alexander Onysko. 2018. Using language learner data for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 133–138.

Krishnkant Swarnkar and Anil Kumar Singh. 2018. Di-LSTM contrast : A deep neural network for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 115–120.

Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39.

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 (Volume 1: Long Papers)*, volume 1, pages 248–258.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 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.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.