A Corpus of Non-Native Written English Annotated for Metaphor

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor
Educational Testing Service
660 Rosedale Road
Princeton, NJ, USA
bbeigmanklebanov,cleong,mflor@ets.org

Abstract

We present a corpus of 240 argumentative essays written by non-native speakers of English annotated for metaphor. The corpus is made publicly available. We provide benchmark performance of state-of-the-art systems on this new corpus, and explore the relationship between writing proficiency and metaphor use.

1 Introduction

With the ubiquity of metaphor across genres of written and spoken communication, the ability of NLP systems to deal with metaphor effectively is an actively researched topic (Veale et al., 2016). Most current work in the supervised machine learning paradigm uses data from the British National Corpus (**BNC**). Beigman Klebanov et al. (2015) reported an evaluation of a metaphor detection system on students' writing; however, their corpus was not released for public use. Our contributions are as follows: (1) We release metaphor annotations of 240 argumentative essays written by non-native speakers of English. This is the first publicly available metaphor annotated data in this genre we are aware of. (2) We evaluate state-of-art (**SoA**) feature sets on the new data. (3) We show that use of argumentation-relevant metaphor is a significant predictor of a holistic score of essay quality, above and beyond essay length.

2 Related Work

Research in automated assessment of students' writing, both native and non-native, is increasingly moving beyond traditional models that emphasize English conventions, sophistication of vocabulary, and organization (Attali and Burstein, 2006). Assessing aspects of content is a rapidly growing research topic, including evaluation of arguments, of the writer's use of information from source materials, of the coherence of the essay, among others (Ghosh et al., 2016; Persing and Ng, 2015; Stab and Gurevych, 2014; Song et al., 2014; Somasundaran et al., 2014; Gurevich and Deane, 2007). Use of metaphor is another aspect of language use that goes beyond grammar and mechanics; recent research suggests that use of metaphor differs with proficiency (Beigman Klebanov et al., 2013), including in non-native writing (Littlemore et al., 2013). On top of serving as a new dataset for metaphor detection experiments, our corpus supports investigation of the relationship between metaphor use and English proficiency.

Most of the recent work on supervised metaphor identification in running text has been done on the VU Amsterdam Metaphor Corpus (**VUA**) (Steen et al., 2010), a large-scale resource containing excerpts from the BNC in four genres (news, academic, fiction, and conversation) annotated for metaphor at the word level (Beigman Klebanov et al., 2016; Haagsma and Bjerva, 2016; Rai et al., 2016; Do Dinh and Gurevych, 2016; Dunn, 2014). Recently, researchers also reported experiments on a corpus of proverbs (Özbal et al., 2016), a corpus of posts to an online breast cancer support group (Jang et al., 2016, 2015), and on argumentative essays (Beigman Klebanov et al., 2015); in these studies, feature sets originally developed for the VUA corpus served as baselines. We follow the same methodology.

3 Metaphor Annotation

3.1 Data

The data was sampled from the publicly available ETS Corpus of Non-Native Written English.[1] The data for annotation was sampled using 8×3×2 design, namely, 5 essays were sampled for each of the eight prompt questions, for three native languages of the writer (Japanese, Italian, Arabic),

[1] https://catalog.ldc.upenn.edu/LDC2014T06

Proceedings of NAACL-HLT 2018, pages 86–91
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

and for two proficiency levels – medium and high. We decided not to include data from low English proficiency writers, as the writing is often barely coherent and the authors' meaning is sometimes difficult to determine. For the experiments reported below, the data was partitioned into 75% training and 25% testing. Data partitions and feature values will be released for public use.[2]

3.2 Annotation

The annotation protocol used in this study was taken from Beigman Klebanov et al. (2013). The protocol was developed for analyzing argumentative writing, and emphasized the identification of argumentation-relevant metaphors. Argumentation-relevant metaphors are, briefly, those that help the author advance her argument. For example, if you are arguing against some action because it would *drain* resources, *drain* is a metaphor that helps you advance your argument, because it presents the expenditure in a very negative way, suggesting that resources would disappear very quickly and without control. Beigman Klebanov et al. (2013) reported inter-annotator agreement of κ = .56-.58 on binary classification of all content words in an essay into metaphor or non-metaphor.

All 240 essays in our corpus were annotated by two annotators: an annotator with a BA in English and Spanish and experience as an English-as-a-second-language (TESOL) instructor who was hired for this project (annotator A) and the lead author of this paper (annotator B). The annotation procedure was as follows. First, 3 out of 30 essays for each prompt were chosen for training and calibration; the two annotators performed an annotation on the 3 essays, and discussed disagreements. Then, each annotator independently annotated the remaining 27 essays. Inter-annotator agreement was calculated for each of the 27 essays; all essays with $\kappa < 0.5$ were selected, and annotator A was asked to review his annotations of these essays again. Once the essays were returned from annotator A's review, agreement was measured again. If the overall agreement for the set of 27 was below $\kappa = 0.55$, essays that had $\kappa < 0.5$ were selected, and annotator B reviewed her own annotations of those essays. Once these annotations were returned, the final κ for the set of 27 essays was calculated. Average inter-annotator agreement for the first annotation pass (before reviews of their own work by A and B) was κ = 0.56. After reviews by one or both the annotators, the average agreement was $\kappa = 0.62$. For the experiments, we use the union of the two annotations: everything marked as metaphor by at least one annotator is labeled as a metaphor, consistently with the practice in prior work (Beigman Klebanov et al., 2013).

To illustrate the annotation, consider an excerpt from a response to the prompt "It is better to have broad knowledge of many academic subjects than to specialize in one specific subject"; metaphors are italicized:

> I ultimately agree with the fact that it is better to be specialized on a specific subject than to *spread* energy on different subjects. However I say ultimately, because being and staying *focused* on one subject means always to *discard* other subjects. I found the *focus* necessary and very important at a certain *late stage* of the personal working career or academic career. The reason behind this you *build up* some "*spikes* of knowledge" on a *broad* knowledge *platform*. These *sharp spikes* of knowledge will allow you to promote yourself and to *pull* with you the society *forward*.

This excerpt is rich in metaphor, painting knowledge as a tall, spiky, yet sturdy structure one builds on a broad solid platform, to serve as a grip when pulling (others) up; a metaphor of academic subjects as objects that can be neatly isolated from one another, examined in detail, and accepted to removed from possession; a life-as-a-journey element that breaks events in life into "stages". All these are working (not necessarily most elegantly) to support the notion that specialization is feasible at an appropriate time in one's life, and it would make you stand out (in the skyline, so-to-speak).

It is worth pointing out some differences between this annotation and an annotation that would have resulted from the application of the MIPVU protocol used in the VUA corpus.[3] The MIPVU protocol requires an annotator to establish the contextual meaning of a lexical item and then consider whether there exists another meaning (attested in a contemporary dictionary) that is more "basic",

[2] https://github.com/EducationalTestingService/metaphor

[3] MIPVU is an extension of MIP (Pragglejaz, 2007) – Metaphor Identification Procedure.

which is defined as (i) more concrete; (ii) related to bodily action; (iii) more precise (as opposed to vague). Additional words in the excerpt shown above might have been marked as metaphors by the MIPVU protocol – such as *found* and *behind*, since their contextual senses are less concrete than the "see where something is by searching for it"[4] and "at the back of something", respectively – but these do not seem to contribute as directly to the content of the author's argument. It is debatable whether *discard* would be considered a metaphor by MIPVU, its one dictionary sense being "to get rid of something that you no longer want or need," which might or might not be considered the contextual sense depending on whether "something" in the definition is interpreted as a concrete object with shape and size or possibly an abstract entity with ill-defined boundaries. The protocol used here does not require recourse to dictionary definitions, leaving the senses to the annotator's intuition. On average, 3% (0.03) of all words in an essay are marked as metaphor according to this protocol; the standard deviation is 0.02; min = 0; max = 0.1, in the training partition.

4 State-of-art feature sets on new data

The task to be performed on the annotated data is to classify all content words (**allPOS**: verbs, nouns, adjectives, and adverbs) or just the verbs (**verbs**) in an essay into those that are being used metaphorically or those that are not. The verbs *have*, *be*, and *do* are excluded from both allPOS and verbs data. Table 1 summarizes the data.

Data	**Training**			**Testing**	
	#T	#I	%M	#T	#I
all-POS	180	26,737	7%	60	9,017
verbs	180	7,016	14%	60	2,301

Table 1: Summary of the data. #T = # of texts; #I = # of instances; %M = proportion of metaphors.

We evaluated the performance of two feature sets. The set **v-16** comes from Beigman Klebanov et al. (2016), addressing metaphoricity of verbs. The feature set **all-15** comes from Beigman Klebanov et al. (2015) and addresses all content words (all-POS). We also ran the v-16 feature set on all-POS data – lemma unigram features are calculated for all words, WordNet lexicographic categories are used only for verbs,[5] and the difference in concreteness feature is calculated for verbs (verb vs its direct object) and for adjectives (adjective vs its head noun). We found that this feature set is competitive with all-15; we therefore include it in the benchmark, as **all-16**. Table 2 summarizes the three feature sets.

Feature Set	**Features**
all-15	unigrams (all POS), part of speech (all POS), topics (all POS), concreteness (all POS), difference in concreteness (v, adj)
v-16	lemma unigrams (v), WordNet lexicographic categories (v), difference in concreteness (v)
all-16	lemma unigrams (all POS), WordNet lexicographic categories (v), difference in concreteness (v, adj)

Table 2: Details of the VUA SoA feature sets

The metaphor detection systems use SoA feature sets with Logistic Regression as the classifier. The systems are evaluated for precision, recall, and F-score for the class "metaphor". The evaluations were performed with scikit-learn (Pedregosa et al., 2011) using the SKLL toolkit[6] that makes it easy to run batch scikit-learn experiments. Table 3 shows the results. Since the data is imbalanced, we applied re-weighting using grid-search optimization, parametrized as in Beigman Klebanov et al. (2015); we also report results with no re-weighting.

Sys	Optimized Re-weighting			No Re-weighting		
	P	**R**	**F**	**P**	**R**	**F**
all-15	.52	.52	.52	.68	.40	.50
all-16	.49	.58	.53	.69	.39	.50
v-16	.50	.64	.56	.69	.39	.50

Table 3: Performance of Logistic Regression with SoA feature sets on essay data.

We note that the performance of the SoA feature sets on the new data is somewhat below the published results for the VUA data. In particular, the v-16 system posted an F-score of 0.60 when trained and tested on verbs-only VUA data (Beigman Klebanov et al., 2016). Improvement of metaphor detection performance is clearly an important avenue for future work.

[4] Sense definitions are quoted from the MacMillan dictionary, used in MIPVU.

[5] We also expanded the WordNet feature set to use the lexicographic categories for verbs and for nouns, but the addition of nominal categories degraded performance; these results are not reported.

[6] http://github.com/ EducationalTestingService/skll

5 Metaphor use and writing proficiency

The main motivation of the study and the annotation campaign is the potential for creating features based on metaphor use for assessing the English language skills of developing writers, under the assumption that argumentation-relevant use of metaphor is a fairly advanced skill that requires solid command of vocabulary and a certain amount of cultural knowledge, among other things. In this section, we consider the relationship between holistic scores of essay quality and use of metaphor, in argumentative essays. Specifically we ask the following research questions: (1) Is there a relationship between essay score and metaphor use? (2) Is this relationship the same for the two definitions of metaphor – argumentation-relevant metaphor and the more traditional MIPVU definition? (3) Does the relationship depend on the specifics of the task set to the writer?

5.1 Data

We use six sets from three testing programs. **MGrF** and **MGrS** – Mixed Graduate Free and Source-based, respectively – come from a test of English administered to domestic and international applicants to graduate schools in the U.S. **IColF** and **IColS** – International College Free and Source-based – come from a test of English mainly administered to international applicants to U.S. colleges and universities. **DTLF** and **DTLS** – Domestic Teacher Licensure Free and Source-based – come from a test of English administered domestically to those wishing to obtain teaching certification in the U.S. The datasets vary in population (domestic vs international, early stages of higher education vs advanced) and in the tasks – for each test, one of the tasks is the standard defend-your-position-on-an-issue task (F), while the other (S) requires test-takers to use source texts to summarize, criticize, or draw on arguments presented therein. Table 4 summarizes the data.

5.2 Method

We quantify the extent of metaphor use in an essay as the logarithm of metaphor frequency per 1,000 words. Given the tendency of essay length to be a strong predictor of proficiency scores, our evaluation metric is partial correlation with essay score controlling for length.

For metaphor detection, we train all-16 model with no re-weighting.[7] We augment the 240-essay corpus described here with an additional set of 141 essays annotated using the same protocol on proprietary data from the same program as MGrF. Performance for a system trained on this combined set of essays is shown as Arg in Table 4; a system that uses the same features and the same training regime on VUA data is shown as VUA in Table 4.

5.3 Results

Dataset	#	Score	Performance	
	Essays	Scale	Arg	VUA
MGrF$^+$	40,000	1-6	.166	.020
MGrS	40,000	1-6	.060	.006
IColF$^+$	40,000	1-5	.159	.070
IColS	40,000	1-5	.067	.052
DTLF	10,000	1-6	.121	.029
DTLS	10,000	1-6	.092	.019

Table 4: Partial correlation controlling for length between essay score and metaphor use, for a system trained on essays (Arg) vs VUA data. The underlined figures are not statistically significant ($p > 0.01$). Plus signs are explained in the text.

First, we observe that Arg shows statistically significant partial correlations for all datasets; namely, use of argumentation-relevant metaphor provides information about essay score above and beyond essay length.

Second, Arg outperforms VUA. In some cases, the difference could be attributed to data being in-domain; the sets marked with a plus in Table 4 are taken from the same testing programs as the training data for Arg, although the specific prompts are different. However, Arg does better across the board, including data completely unrelated to the annotation campaign. It is likely that the protocol that emphasized specifically the need to pay attention to the role played by the metaphor in the author's argument is at least partially responsible for the higher performance indicators.

Next, we observe better performance on F sets, those with a very general, single-sentence prompt (see example in section 3.2) than on S datasets with extensive prompts that directed test-takers to criticize, summarize, or draw upon arguments presented in specific textual sources. Again, this

[7] Precision-oriented detection models aggregated through a logarithmic or square-root functions are common in the automated essay scoring literature (Gamon et al., 2013; Chen et al., 2017; Attali and Burstein, 2006).

could be due to the short-prompt-based arguments being more in line with the annotated data; however, since there is a similar tendency for the VUA-trained system, it could also be a more general issue of the extent to which the author controls the vocabulary in her essay. With extensive prompts that contain information that needs to be reflected in the essay, a substantial part of the vocabulary is forced by the prompt and not drawn from the author's more creative faculties and knowledge.

6 Conclusion

We present a corpus of argumentative essays written by non-native speakers of English annotated for metaphor. The corpus is made publicly available; this is the first publicly available metaphor annotated data in this genre we are aware of. We provide benchmark performance on this new corpus. We also show that use of argumentation-relevant metaphor provides information about English proficiency above and beyond essay length, especially in the standard defend-a-position-on-an-issue essays where authors have fuller control of their vocabulary.

References

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring With e-rater®V.2. *Journal of Technology, Learning, and Assessment* 4(3). https://www.learntechlib.org/p/103244/.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*. Association for Computational Linguistics, Atlanta, Georgia, pages 11–20. http://www.aclweb.org/anthology/W13-0902.

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*. Association for Computational Linguistics, Denver, Colorado, pages 11–20. http://www.aclweb.org/anthology/W15-1402.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 101–106. http://anthology.aclweb.org/P16-2017.

Jing Chen, Mo Zhang, and Isaac I. Bejar. 2017. An investigation of the e-rater automated scoring engine's grammar, usage, mechanics, and style microfeatures and their aggregation model. *ETS Research Report Series* https://doi.org/10.1002/ets2.12131.

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*. Association for Computational Linguistics, San Diego, California, pages 28–33. http://www.aclweb.org/anthology/W16-1104.

Jonathan Dunn. 2014. Multi-dimensional abstractness in cross-domain mappings. In *Proceedings of the Second Workshop on Metaphor in NLP*. Association for Computational Linguistics, Baltimore, MD, pages 27–32. http://www.aclweb.org/anthology/W14-2304.

Michael Gamon, Martin Chodorow, Claudia Leacock, and Joel Tetreault. 2013. Grammatical Error Detection in Automatic Essay Scoring and Feedback. In Mark Shermis and Jill Burstein, editors, *Handbook for Automated Essay Evaluation*, New York: Taylor and Francis.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 549–554. http://anthology.aclweb.org/P16-2089.

Olga Gurevich and Paul Deane. 2007. Document similarity measures to distinguish native vs. non-native essay writers. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Association for Computational Linguistics, Rochester, New York, pages 49–52. http://www.aclweb.org/anthology/N/N07/N07-2013.

Hessel Haagsma and Johannes Bjerva. 2016. Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*. Association for Computational Linguistics, San Diego, California, pages 10–17. http://www.aclweb.org/anthology/W16-1102.

Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rose. 2016. Metaphor detection with topic transition, emotion and cognition in context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 216–225. http://www.aclweb.org/anthology/P16-1021.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rose. 2015. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, pages 384–392. http://aclweb.org/anthology/W15-4650.

Jeannette Littlemore, Tina Krennmayr, James Turner, and Sarah Turner. 2013. An investigation into metaphor use at different levels of second language writing. *Applied Linguistics* https://doi.org/10.1093/applin/amt004.

Gözde Özbal, Carlo Strapparava, Serra Sinem Tekiroglu, and Daniele Pighin. 2016. Learning to identify metaphors from a corpus of proverbs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2060–2065. https://aclweb.org/anthology/D16-1220.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 543–552. http://www.aclweb.org/anthology/P15-1053.

Group Pragglejaz. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol* 22(1):1–39. https://www.tandfonline.com/doi/abs/10.1080/10926480709336752.

Sunny Rai, Shampa Chakraverty, and Devendra K. Tayal. 2016. Supervised metaphor detection using conditional random fields. In *Proceedings of the Fourth Workshop on Metaphor in NLP*. Association for Computational Linguistics, San Diego, California, pages 18–27. http://www.aclweb.org/anthology/W16-1103.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING*. pages 950–961. http://aclweb.org/anthology/C/C14/C14-1090.pdf.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 69–78. http://www.aclweb.org/anthology/W14-2110.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In Junichi Tsujii and Jan Hajic, editors, *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. Dublin City University and Association for Computational Linguistics, pages 1501–1510. http://www.aclweb.org/anthology/C14-1142.

Gerard Steen, Aletta Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification*. Amsterdam: John Benjamins.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies* 9(1):1–160. https://www.morganclaypool.com/doi/abs/10.2200/S00694ED1V01Y201601HLT031.