# Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task

**Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov**
Educational Testing Service
xchen002,cleong,mflor,bbeigmanklebanov@ets.org

## Abstract

This paper describes the ETS entry to the 2020 Metaphor Detection shared task. Our contribution consists of a sequence of experiments using BERT, starting with a baseline, strengthening it by spell-correcting the TOEFL corpus, followed by a multi-task learning setting, where one of the tasks is the token-level metaphor classification as per the shared task, while the other is meant to provide additional training that we hypothesized to be relevant to the main task. In one case, out-of-domain data manually annotated for metaphor is used for the auxiliary task; in the other case, in-domain data automatically annotated for idioms is used for the auxiliary task. Both multi-task experiments yield promising results.

## 1 Introduction

We use metaphors in our everyday life as a means of relating our experiences to other subjects and contexts (Lakoff and Johnson, 2008); it is commonly used to help us understand the world in a structured way, and oftentimes in an unconscious manner while we speak and write. It sheds light on the unknown using the known, explains the complex using the simple, and helps us to emphasize the relevant aspects of meaning resulting in effective communication.

There is a large body of work in the literature that discusses how metaphor has been used in the context of political communication, marketing, mental health, teaching, assessment of English proficiency, among others (Beigman Klebanov et al., 2018; Gutierrez et al., 2017; Littlemore et al., 2013; Thibodeau and Boroditsky, 2011; Kaviani and Hamedi, 2011; Kathpalia and Carmel, 2011; Landau et al., 2009; Beigman Klebanov et al., 2008; Zaltman and Zaltman, 2008; Littlemore and Low, 2006; Cameron, 2003; Lakoff, 2010; Billow et al., 1997; Bosman, 1987); see chapter 7 in Veale et al. (2016) for a recent review.

In the NLP universe, there's been substantial recent interest in automated detection of metaphor (Dankers et al., 2019; Mikhalkova et al., 2019; Mao et al., 2019; Igamberdiev and Shin, 2018; Marhula et al., 2019; Markert, 2019; Saund et al., 2019).

This paper describes the ETS entry to the 2020 Metaphor Detection shared task held as a part of the 2nd Workshop on Processing Figurative Language, at ACL 2020[1]. The shared tasks consists of four tracks: all content parts of speech – nouns, verbs, adjectives, and adverbs (AllPOS) and a verbs-only track (Verbs) for two corpora – (a) a corpus of well-edited BNC articles from a variety of genres annotated using the MIP-VU protocol, and (b) a corpus of medium to high quality timed, non-native essays written for the Test of English as a Foreign Language annotated under a different protocol. We participated in all the four tracks.

Our contribution consists of a sequence of experiments using BERT, starting with a baseline, then strengthening it by spell-correcting the TOEFL corpus (section 4). We then devised a multi-task learning setting, where one of the tasks is the token level metaphor classification as per the shared task, while the other is meant to provide additional training that we hypothesized to be relevant to the main task (section 5).

The first multitask learning is the utilization out-of-domain data annotated for metaphor, albeit under a different annotation protocol, by using data from the other competition corpus. Thus, we use metaphor prediction on the VUA corpus as an auxiliary task for the main TOEFL task, and vice versa. We show that this setup resulted in an improved performance on the TOEFL test data but not on VUA data. A sanity-check experiment where the two training datasets were simply merged together yielded performance that was inferior to the baseline for all tracks (section 5.1).

[1] https://sites.google.com/view/figlang2020/home



*Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243

https://doi.org/10.18653/v1/P17

The second auxiliary task is utilization of a large in-domain corpus that we automatically tagged for occurrence of a different type of figurative language phenomenon – idioms. We hypothesize that the affinity between the two ways of figuration might help the system become more sensitive to metaphor by learning to attend to idioms. Our results provide support to the hypothesis, as this setting yielded our best result on the VUA dataset (section 5.2). We provide a discussion of our findings in section 7.

## 2 Datasets

### 2.1 VUA corpus

We use the VU Amsterdam Metaphor Corpus (VUA) (Steen et al., 2010) as provided by the shared task organizers. The dataset consists of 117 fragments sampled across four genres from the British National Corpus: Academic, News, Conversation, and Fiction. Each genre is represented by approximately the same number of tokens, although the number of texts differs greatly, where the news archive has the largest number of texts. The data is annotated using the MIP-VU procedure with a strong inter-annotator reliability of $\kappa > 0.8$. It is based on the MIP procedure (Praglejaz, 2007), extending it to handle metaphoricity through reference (such as marking *did* as a metaphor in *As the weather broke up, so did their friendship*) and allow for explicit coding of difficult cases where a group of annotators could not arrive at a consensus. Note that we only considered words marked as metaphors decided as such by the shared task organizers. The VUA dataset and annotations is the same as the one used in the first shared task on metaphor detection (Leong et al., 2018).

### 2.2 TOEFL corpus

This data labeled for metaphor was sampled from the publicly available ETS Corpus of Non-Native Written English[2] (Blanchard et al., 2013) and was first introduced by (Beigman Klebanov et al., 2018). The annotated data comprises essay responses to eight persuaisve/argumentative prompts, for three native languages of the writer (Japanese, Italian, Arabic), and for two proficiency levels – medium and high. The data was annotated using the protocol in Beigman Klebanov and Flor (2013), that emphasized argumentation-relevant metaphors. Average inter-annotator agreement was $\kappa$ = 0.56 – 0.62, for multiple passes of the annotation (see (Beigman Klebanov et al., 2018) for more details). For the experiments, we used the metaphor annotations marked as such by the organizers. We used 180 essays for training and 60 essays for testing, as provided by the shared task organizers. Tables 1 and 2 show some descriptive characteristics of the data: the number of texts, sentences, tokens, and class distribution information for Verbs and AllPOS tracks for the two corpora – VUA and TOEFL.

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

## 3 Baseline system

We build our baseline system based on BERT (Devlin et al., 2018). BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) is a transformer (Vaswani et al., 2017) model that is pre-trained on a large quantity of texts, and obtained state-of-the-art performance on many NLP benchmarks (Wang et al., 2018; Zellers et al., 2018; Rajpurkar et al., 2016). Since its introduction, there have been many improvements over the original BERT model, such as RoBERTa (Liu et al., 2019), ERNIE (Sun et al., 2019), XLNet (Yang et al., 2019); we use the most basic model (**bert-base-uncased**).

We fine-tune the BERT model as a standard token classification task, that is, after obtaining the contextualized embeddings of a sentence, we apply a linear layer followed by softmax on each token to predict whether it is metaphorical or not. Fig 1 shows the architecture of the baseline model. We tune the hyperparameters based on cross-validation on training data. The fold partitions for the VUA corpus are the same as the ones used for experiments in Beigman Klebanov et al. (2016). For the TOEFL corpus, we obtained the folds information from the shared task organizers directly. We select batch size in $\{16, 32, 64\}$, number of training epochs in $\{2, 3, 4, 5\}$, and use a fixed learning rate of $3 \times 10^{-5}$. We also apply the learning rate scheduler known as *slanted triangular* (Howard and Ruder, 2018). Due to the imbalanced class distribution in our data (see Table 2), the positive class is up-weighted by a factor of 3. The same setting applies to experiments described in all the following sections.

| Datasets | VUA | | TOEFL | |
|---|---|---|---|---|
| | Train | Test | Train | Test |
| #texts | 90 | 27 | 180 | 60 |
| #sents | 12,123 | 4,081 | 2,741 | 968 |

Table 1: Number of texts and sentences for both VUA and TOEFL datasets.

| Datasets | VUA | | | | TOEFL | | | |
|---|---|---|---|---|---|---|---|---|
| | Verbs | | All POS | | Verbs | | All POS | |
| | Train | Test | Train | Test | Train | Test | Train | Test |
| #tokens | 17,240 | 5,873 | 72,611 | 22,196 | 7,016 | 2,301 | 26,737 | 9,014 |
| %M | 29% | – | 18% | – | 13% | – | 7% | – |

Table 2: Number of tokens and percentage of metaphors breakdown for both VUA and TOEFL datasets, grouped by Verbs and AllPOS.

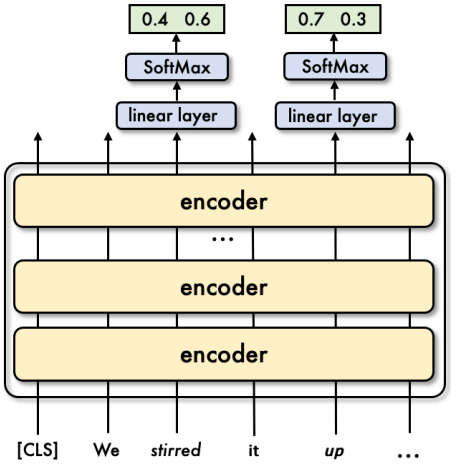


Figure 1: Baseline system architecture. The output is a pair of probabilities – the first for class 0 (non-metaphor) and the second for class 1 (metaphor).

## 4 Experiment 1: Spell correction system

Proper automatic detection of lexically-anchored phenomena in text often depends on availability of correct spelling in the text. The contribution of spelling correction to other tasks has been documented previously, especially for English texts produced by non-native learners of English (Rozovskaya and Roth, 2016; Granger and Wynne, 1999). Essays written by TOEFL test-takers are known to contain a considerable amount of spelling errors (Flor et al., 2015). To alleviate this, we used a state-of-the-art automatic spelling corrections system (Flor et al., 2019) to correct spelling in the TOEFL dataset. Specifically, for the training partition of the TOEFL dataset, the system corrected 1553 errors in 180 essays, and 510 errors in 60 essays of the test partition.

## 5 Multi-task system

As we fine-tune BERT on relatively small datasets, we attempted to enrich the learning with partially relevant additional materials through a multi-task setting - adding auxiliary tasks and train the metaphor detection task with them. The auxiliary tasks are described in sections 5.1 and 5.2.

The model we use for multi-task learning is as follows: Instead of directly making predictions based on the output embeddings of BERT, the embeddings are first projected to a lower-dimensional representation by a linear layer; each task then has its own classifier on top of that linear layer. The architecture of the model is shown in Fig. 2.

During training, the data in batches of the metaphor task (our main task) and the auxiliary tasks are mixed and trained on in an interleaved manner; the specifics will be described for each of the auxiliary tasks separately. In order for the main task to dominate the learning, we also scale the gradient of the auxiliary tasks by a factor of 0.1. The hyperparameters are selected in the same way as described in section 3.

### 5.1 Experiment 2: Learning from out-of-domain data

Since both the VUA and the TOEFL corpora are annotated for metaphors, using one to help the other during learning could potentially provide additional relevant training data. However, since the data is from different types of texts and dif-

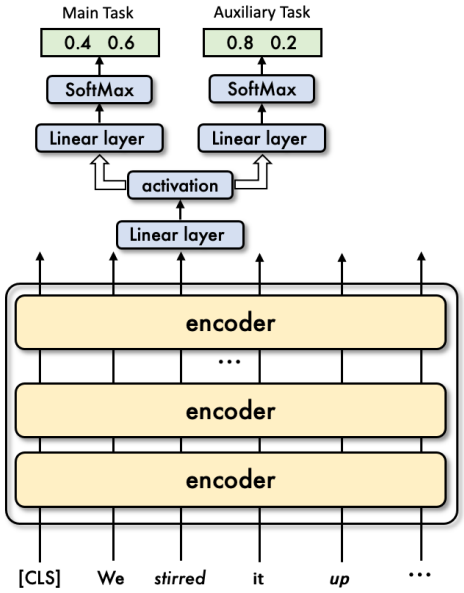


Figure 2: Multi-task system architecture. The output is a pair of probabilities – the first for class 0 (non-metaphor) and the second for class 1 (metaphor).

ferent genres (well-edited BNC text in academic, news, conversation and fiction genres vs relatively short English language learner essays), and since the guidelines under which the two datasets were annotated are different, it is possible that each corpus is only partially or indirectly relevant to the other. We experimented with both a straightforward merging of the training sets of the two datasets (as a preview – this did not produce good results) and with a multi-task setting where the other corpus is used for the auxiliary task.

We use the same batch size and learning rate for the main task and the auxiliary task. The batches from the two tasks are interleaved uniformly; as there are roughly four times more sentences in the VUA corpus than in the TOEFL corpus, there are five batches of the VUA task following every batch of the TOEFL task. We do not sub-sample the VUA corpus or over-sample the TOEFL corpus.

### 5.2 Experiment 3: Learning from another type of figurative language

Differently from Experiment 2, where we utilized an out-of-domain dataset annotated for the same phenomenon (albeit under somewhat different annotation protocol), in Experiment 3 we are attempting to make use of a different but related phenomenon – a different type of figurative language, namely, idioms. The metaphorical underpinnings of many idiomatic expressions have been noted in psycholinguistic literature (Gibbs and O'Brien, 1990; Nunberg et al., 1994; Glucksberg, 2001).

The main idea here is that one or more of the words participating in idiomatic expressions are often used metaphorically. Thus, in CUTTING EDGE both the words are used metaphorically; in PAY *attention*, *false* STEP, *helping* HAND, GOLDEN *opportunity*, and *social* LADDER, the capitalized word is a metaphor while the other is not. There are also idioms where none of the words are used metaphorically such as *matter of fact*, *other than*, and *once in a while*. Still, it appears likely that the preponderance of metaphors within idiomatic expressions would be higher than in non-idiomatic language. It is also possible that learning to detect idioms – a different but related type of figurative language – could help with metaphor detection, as these might tend to be used in similar contexts. Experiment 3 is an attempt to explore these observations by setting idiom detection as an auxiliary task for the main metaphor detection task.

Although there exists considerable prior research on automatic detection of idioms (for a brief review see Flor and Beigman Klebanov (2018)), idiom detection systems are typically constrained to very small sets of idioms or to particular types of expressions (e.g. verb-noun constructions). We opted to use a system that marks candidate expressions but does not verify their idiomaticity in the given context. The advantage of this particular system is that it has very wide coverage. We assume that many of the idioms found in a particular corpus might be well-known idioms that are listed in various dictionaries. Our system (Flor and Beigman Klebanov, 2018) is equipped with a dictionary of about 5000 English idiomatic expressions (culled from Wiktionary), and performs a flexible search for idioms and their syntactic and lexical variants in running text. In fact, it performs a simultaneous flexible pattern matching. The idiom detection system looks only for expressions that have more than one word, and excludes common greeting phrases (e.g. 'have a nice day'), phrasal verbs and verb+preposition constructions (unless they are part of a larger idiom). The system marks expressions that potentially might be instances of idioms, but it does not perform idiom/non-idiom classification. For the present experiment we used this system with rather conservative settings that yielded precision

of 0.571 in our previous evaluations on a subset of the TOEFL data; see the leftmost column in Figure 1-A in Flor and Beigman Klebanov (2018) for the details of the configuration. Based on the prior evaluation, for this system configuration, most of the errors were cases where the expression is identified correctly but it is used literally rather than idiomatically.

We ran the idiom-candidate marking system on TOEFL-11[3] essays and on the BNC corpus (excluding texts of the shared task). In total, the system detected 3,581 different idiom types in the BNC, with 179,967 instances of (candidate) idioms; in the TOEFL-11 data, we found 504 different idiom types, with 3,908 instances. There is somewhat more idiom usage per sentence in the BNC than in the TOEFL data: The system identified an idiom in 3% of all BNC sentences and in 2.2% for of TOEFL-11 sentences. Table 3 shows the 20 most frequently found (candidate, or unverified) idioms in the BNC and TOEFL data; the lists contain a mix of idioms that contain and do not contain metaphors.

The idiom detection auxiliary task is also formalized as a token classification task: Given a sentence, predict for each token whether it is part of an idiom. Given the size of the BNC corpus, we only sample a small subset of it for training: 10,000 sentences with idioms and 10,000 sentences without idioms. For the TOEFL-11 data, we keep all sentences with idioms and sample the same number (3,908) of sentences without idioms.

## 6 Results

Tables 4 and 5 show performance of the various systems on AllPOS and Verbs-only tasks, respectively, for both VUA and TOEFL data. Since it is clear that spelling correction is useful for improving performance on TOEFL data, we used the spell corrected version of the data for all the systems from experiments 2 and 3 on TOEFL data. Our best-performing systems reported here are also benchmarked against other participating systems in the shared task summary report (Leong et al., 2020). We obtained a ranking of 2nd and 4th in the VUA and TOEFL tasks, respectively.

Since VUA data contains well-edited BNC text, we did not run spelling correction on VUA data. For the Verbs tracks, we experimented with both (a) training on AllPOS data and evaluating on the Verbs-only subset of the test data, and (b) training and testing on Verbs only subsets. Version (a) yielded better results, which are reported here.

## 7 Discussion

First, we observe that comparative results across the different systems are highly consistent for AllPOS and Verbs-only settings; we therefore focus on AllPOS in the discussion.

Combining the training data from TOEFL and VUA sets does not result in better performance on either test set (see $D_{all}$ in Tables 4, 5). This could be due to both out-of-domain nature of the two corpora with respect to each other, to the difference in the guidelines under which the two corpora were annotated, and/or to the difference in the distribution of metaphors vs non-metaphors in the two corpora (see Table 2 for class distribution information).

However, when set up as a multi-task system with a shared representation, using data from VUA as part of the training process results in better performance on TOEFL test data, with a 2.6 points F1 score gain for AllPOS (0.666 vs 0.692 in Table 4). Thus, it appears that using the VUA data as part of the training process through the shared representation but without the TOEFL training process sustaining a loss for mis-classifying instances from VUA (as was the case when the training sets were merged), the system has apparently successfully acquired useful information that helped boost performance on TOEFL test data.

It is interesting to note that the multi-task version of the setting for using out-of-domain data did not result in improvements on VUA test data (F1 score of 0.717 vs 0.715). The drop in recall and increase in precision observed for the $D_{mt}$ model on VUA data is consistent with the direction of the results where the two training datasets were simply merged into a bigger training dataset ($D_{all}$). It appears that under the guidelines in which TOEFL data was annotated where argumentation-relevant metaphors are detected intuitively, without recourse to a standard dictionary, the annotation outcomes are more conservative: Some instances that the system trained on VUA data considered metaphorical were not considered so in a system that was exposed to TOEFL data

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

| BNC | TOEFL 11 |
|---|---|
| find oneself | long time |
| other than | need-to-know |
| long time | pay attention |
| great deal | matter of fact |
| once again | other than |
| ups and downs | day-to-day |
| once more | find oneself |
| much less | long run |
| come through | stay at home |
| old woman | play games |
| bear in mind | great deal |
| cup of tea | jack of all trades |
| day-to-day | side effect |
| ask the question | much less |
| let alone | change one's mind |
| need-to-know | ask the question |
| common law | again and again |
| close one's eyes | well and good |
| blue-eyed | tell the truth |
| change one's mind | once again |

Table 3: Top 20 most frequently observed (unverified) idioms in the BNC and TOEFL 11 corpora. Note: Hyphens are treated as between-word delimiters and are optionally matched. Thus, both "stay-at-home" and "day to day" will be matched, even though these are not the canonical forms of the idioms on the list.

during training. For example, the three underlined words in the following sentence were classified as metaphors by the version that was trained on VUA data only, but were classified as non-metaphors after augmentation with the TOEFL data: "A less direct measure which is applicable only to the most senior management is to observe the fall or rise of the share price when a particular executive leaves or joins a company." Of these, *senior* and *leaves* are metaphors according to VUA ground truth, while *observe* is not. Overall, the drop in recall was not sufficiently offset by the increase in precision (although there is a small improvement in F1 score for the Verbs only data – from 0.756 to 0.762, see Table 5). Still, our results suggest that if one is interested in a precision-focused system, using TOEFL data in a multi-task setting when training and testing on VUA could be beneficial, as $D_{mt}$ achieved the best precision on the VUA dataset among all the compared systems.

We next turn to the experiments with an auxiliary idiom detection task. We observe that on VUA data this resulted in a 2-point increase in

| All POS | | | | | | |
|---|---|---|---|---|---|---|
| System | VUA | | | TOEFL | | |
| | P | R | F | P | R | F |
| BL | .721 | .713 | .718 | .701 | .563 | .624 |
| Sp | – | – | – | .656 | .676 | **.666** |
| $D_{all}$ | .728 | .676 | .701 | .576 | .637 | .605 |
| $D_{mt}$ | .741 | .692 | .715 | .669 | .717 | **.692** |
| $I_{mt}$ | .721 | .749 | **.734** | .718 | .616 | .663 |

Table 4: AllPOS performance. BL = baseline BERT system; Sp = baseline BERT system trained and tested on spell-corrected TOEFL data; $D_{all}$ = baseline BERT system trained on combined TOEFL and VUA data; $D_{mt}$ = a multi-task system using out-of-domain metaphor annotated data; $I_{mt}$ = multi-task system using idiom detection as an auxiliary task.

| Verbs | | | | | | |
|---|---|---|---|---|---|---|
| System | VUA | | | TOEFL | | |
| | P | R | F | P | R | F |
| BL | .725 | .790 | .756 | .624 | .694 | .657 |
| Sp | – | – | – | .674 | .694 | **.684** |
| $D_{all}$ | .747 | .733 | .740 | .614 | .664 | .638 |
| $D_{mt}$ | .754 | .772 | .762 | .747 | .661 | **.702** |
| $I_{mt}$ | .732 | .823 | **.775** | .705 | .631 | .667 |

Table 5: Verbs performance. BL = baseline BERT system; Sp = baseline BERT system trained and tested on spell-corrected TOEFL data; $D_{all}$ = baseline BERT system trained on combined TOEFL and VUA data; $D_{mt}$ = a multi-task system using out-of-domain metaphor annotated data; $I_{mt}$ = multi-task system using idiom detection as an auxiliary task.

F1 score – with no penalty in precision, the system gained about 3.5 points in recall (0.713 vs 0.749 on AllPOS; 0.790 vs 0.823 on Verbs). This confirms the usefulness of attending to a related type of figurative language through an auxiliary task – even though the identification of idioms was done using an automated procedure and therefore is quite noisy.

To examine the impact of the idiom auxiliary task, we used one of the cross-validation folds as development set. Looking at instances tagged as non-metaphor by the baseline model and as metaphor by the current model, there are two observations. First, of the 236 VUA sentences with newly tagged metaphors, only 9 sentences contained an idiom, according to our idiom detection system. Thus, it does not appear to be the case that it is specifically metaphors within known idioms that the system has now learned to find; this

is a tentative conclusion, however, as it is also possible that these sentences did contain idioms that were either not on the list of the 5,000 the system is searching for, or are on the list and are present in the text but are not detected by the system. Secondly, it appears that the system has learned some sentence-level characteristics of sentences that contain figurative language, in that quite often multiple words in the same sentence got tagged as metaphors: the 236 sentences contained 323 newly tagged metaphors. The most extreme case is that of 4 new words in the same sentence being tagged as metaphors (italicized): "This desire that can not find its *name* (though it would *dare speak*, if it could) is *pleasurable*."

Using idioms for an auxiliary task did not help with TOEFL data. We also tried using the BNC idiom data instead of the TOEFL 11 idiom data; this resulted in comparable performance, still without improvement over the spell-checked single-task version. Since results on VUA suggest that idioms could provide useful information for metaphor detection, we intend to further pursue this line of work by attending more closely to the different types of idiomatic expressions that might be more or less useful for metaphor detection, and by improving the idiom detection mechanism.

## 8 Conclusion

This paper describes the ETS entry to the 2020 Metaphor Detection shared task held as a part of the 2nd Workshop on Processing Figurative Language, at ACL 2020. We participated in all four tracks – Verbs and AllPOS for each of VUA and TOEFL datasets. Our contribution consists of a sequence of experiments using BERT, starting with a baseline, then strengthening it by spell-correcting the TOEFL corpus, followed by a multi-task learning setting, where one of the tasks is the token-level metaphor classification as per the shared task, while the other is meant to provide additional training that we hypothesized to be relevant to the main task.

The first multitask learning is the utilization of out-of-domain data annotated for metaphor, albeit under a different annotation protocol, by using data from the other competition corpus; this manipulation helped improve F1 scores for metaphor class on TOEFL test data, but not on VUA data. The second auxiliary task is utilization of a large in-domain corpus that we automatically tagged for occurrence of a different type of figurative language phenomenon – idioms. This manipulation resulted in an improved performance on VUA data, but not on TOEFL data. Given the promising results with idiom auxiliary task, we intend to continue work in this direction by improving automatic detection of idioms and by a finer-grained analysis of the contribution of various types of idioms to improve metaphor detection.

## References


Beata Beigman Klebanov, Daniel Diermeier, and Eyal Beigman. 2008. Lexical cohesion analysis of political speech. *Political Analysis*, 16(4):447–463.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2018. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.

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)*, volume 2, pages 101–106.

Richard M Billow, Jeffrey Rossman, Nona Lewis, Deberah Goldman, and Charles Raps. 1997. Observing expressive and deviant language in schizophrenia. *Metaphor and Symbol*, 12(3):205–216.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Jan Bosman. 1987. Persuasive effects of political metaphors. *Metaphor and Symbol*, 2(2):97–113.

Lynne Cameron. 2003. *Metaphor in educational discourse*. A&C Black.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Michael Flor and Beata Beigman Klebanov. 2018. Catching idiomatic expressions in EFL essays. In *Proceedings of the Workshop on Figurative Language Processing*, pages 34–44, New Orleans, LA.

Michael Flor, Michael Fried, and Alla Rozovskaya. 2019. A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, page 76–86, Florence, Italy.

Michael Flor, Yoko Futagi, Melissa Lopez, and Matthew Mulholland. 2015. Patterns of misspellings in L2 and L1 English: a view from the ETS spelling corpus. *Bergen Language and Linguistics Studies*, 6.

R. W. Gibbs and J. E. O'Brien. 1990. Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. *Cognition*, 36:35–68.

Sam Glucksberg. 2001. *Understanding Figurative Language: from metaphors to idioms*. Oxford University Press, New York, NY.

Sylviane Granger and Martin Wynne. 1999. Optimising measures of lexical variation in EFL learner corpora. In *Corpora Galore*, pages 249–257, Amsterdam. Rodopi.

E Dario Gutierrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Timour Igamberdiev and Hyopil Shin. 2018. Metaphor identification with paragraph and word vectorization: An attention-based neural approach. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Sujata S Kathpalia and Heah Lee Hah Carmel. 2011. Metaphorical competence in ESL student writing. *RELC Journal*, 42(3):273–290.

Hossein Kaviani and Robabeh Hamedi. 2011. A quantitative/qualitative study on metaphors used by persian depressed patients. *Archives of Psychiatry and Psychotherapy*, 4(5-13):110.

George Lakoff. 2010. *Moral politics: How liberals and conservatives think*. University of Chicago Press.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Mark J Landau, Daniel Sullivan, and Jeff Greenberg. 2009. Evidence that self-relevant motives and metaphoric framing interact to influence political and social attitudes. *Psychological Science*, 20(11):1421–1427.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

Chee Wee 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.

Jeannette Littlemore, Tina Krennmayr, James Turner, and Sarah Turner. 2013. An investigation into metaphor use at different levels of second language writing. *Applied linguistics*, 35(2):117–144.

Jeannette Littlemore and Graham Low. 2006. Metaphoric competence, second language learning, and communicative language ability. *Applied linguistics*, 27(2):268–294.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898.

Joanna Marhula, Justyna Polak, Maria Janicka, and Aleksander Wawer. 2019. Recognizing metaphor: how do non-experts and machines deal with a metaphor identification task? In *BOOK OF ABSTRACTS*, page 70.

Katja Markert. 2019. Literature list ps/hs ss2019 figurative language resolution.

Elena Mikhalkova, Nadezhda Ganzherli, Vladislav Maraev, Anna Glazkova, and Dmitriy Grigoriev. 2019. A comparison of algorithms for detection of "figurativeness" in metaphor, irony and puns. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 186–192. Springer.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Pragglejaz. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Alla Rozovskaya and Dan Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2205–2215, Berlin, Germany. Association for Computational Linguistics.

Carolyn Saund, Marion Roth, Mathieu Chollet, and Stacy Marsella. 2019. Multiple metaphors in metaphoric gesturing. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 524–530. IEEE.

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*, volume 14. John Benjamins Publishing.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2):e16782.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Gerald Zaltman and Lindsay H Zaltman. 2008. *Marketing metaphoria: What deep metaphors reveal about the minds of consumers*. Harvard Business Press.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.