A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task

Chee Wee Leong[1], **Beata Beigman Klebanov**[1],
Chris Hamill[1], **Egon Stemle**[2,3*], **Rutuja Ubale**[1†] **and Xianyang Chen**[1]
[1]Educational Testing Service
[2]Eurac Research, Institute for Applied Linguistics
[3]Masaryk University, Faculty of Informatics
{cleong,bbeigmanklebanov,chamill,xchen002}@ets.org
[†]rubale@etscanada.ca
[*]egon.stemle@eurac.edu

Abstract

In this paper, we report on the shared task on metaphor identification on VU Amsterdam Metaphor Corpus and on a subset of the TOEFL Native Language Identification Corpus. The shared task was conducted as apart of the ACL 2020 Workshop on Processing Figurative Language.

1 Introduction

Metaphor use in everyday language is a way to relate our physical and familiar social experiences to a multitude of other subjects and contexts (Lakoff and Johnson, 2008); it is a fundamental way to structure our understanding of the world even without our conscious realization of its presence as we speak and write. It highlights 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.

Metaphor has been studied 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.

We report on the second shared task on automatic metaphor detection, following up on the first shared task held in 2018 (Leong et al., 2018). We present the shared task and provide a brief description of each of the participating systems, a comparative evaluation of the systems, and our observations about trends in designs and performance of the systems that participated in the shared task.

2 Related Work

Over the last decade, automated detection of metaphor has become an popular topic, which manifests itself in both a variety of approaches and in an increasing variety of data to which the methods are applied. In terms of methods, approaches based on feature-engineering in a supervised machine learning paradigm explored features based on concreteness and imageability, semantic classification using WordNet, FrameNet, VerbNet, SUMO ontology, property norms, and distributional semantic models, syntactic dependency patterns, sensorial and vision-based features (Bulat et al., 2017; Köper and im Walde, 2017; Gutierrez et al., 2016; Shutova et al., 2016; Beigman Klebanov et al., 2016; Tekiroglu et al., 2015; Tsvetkov et al., 2014; Beigman Klebanov et al., 2014; Dunn, 2013; Neuman et al., 2013; Mohler et al., 2013; Hovy et al., 2013; Tsvetkov et al., 2013; Turney et al., 2011; Shutova et al., 2010; Gedigian et al., 2006); see Shutova et al. (2017) and Veale et al. (2016) for reviews of supervised as well as semi-supervised and unsupervised approaches. Recently, deep learning methods have been explored for token-level metaphor detection (Mao et al., 2019; Dankers et al., 2019; Gao et al., 2018; Wu et al., 2018; Rei et al., 2017; Gutierrez et al., 2017; Do Dinh and Gurevych, 2016).

In terms of data, researchers used specially constructed or selected sets, such as adjective noun pairs (Gutierrez et al., 2016; Tsvetkov et al., 2014), WordNet synsets and glosses (Mohammad et al., 2016), annotated lexical items (from a range of word classes) in sentences sampled from corpora (Özbal et al., 2016; Jang et al., 2015; Hovy et al., 2013; Birke and Sarkar, 2006), all the way to annotation of all words in running text for metaphoricity (Beigman Klebanov et al., 2018; Steen et al., 2010); Veale et al. (2016) review various annotated datasets.

Proceedings of the Second Workshop on Figurative Language Processing, pages 18–29

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

3 Task Description

The goal of this shared task is to detect, at the word level, all content word metaphors in a given text. We are using two datasets – VUA and TOEFL, to be described shortly. There are two tracks for each dataset, for a total of four tracks: **VUA All POS**, **VUA Verbs**, **TOEFL All POS**, and **TOEFL Verbs**. The **AllPOS** track is concerned with the detection of all content words, i.e., nouns, verbs, adverbs and adjectives that are labeled as metaphorical while the **Verbs** track is concerned only with verbs that are metaphorical. We excluded all forms of *be*, *do*, and *have* for both tracks. For each dataset, each participating individual or team can elect to compete in the All POS track, Verbs track, or both. The competition is organized into two phases: training and testing.

3.1 Datasets

3.1.1 VUA corpus

We use the VU Amsterdam Metaphor Corpus (VUA) (Steen et al., 2010). The dataset consists of 117 fragments sampled across four genres from the British National Corpus: Academic, News, Conversation, and Fiction. The data is annotated using the MIPVU procedure with a strong inter-annotator reliability of $\kappa > 0.8$ (Steen et al., 2010). The VUA dataset and annotations is the same as the one used in the first shared task on metaphor detection (Leong et al., 2018), where the reader is referred for further details.

3.1.2 TOEFL corpus

This data labeled for metaphor was sampled from the publicly available ETS Corpus of Non-Native Written English[1] and was first introduced by (Beigman Klebanov et al., 2018). The annotated data comprises essay responses to eight persuasive/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:

> "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 and Flor (2013)

Average inter-annotator agreement was κ = 0.56-0.62, for multiple passes of the annotation (see (Beigman Klebanov et al., 2018) for more details). We use the data partition from Beigman Klebanov et al. (2018), with 180 essays as training data and 60 essays as testing data.

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 datasets.

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.

To facilitate the use of the datasets and evaluation scripts beyond this shared task in future research, the complete set of task instructions and scripts are published on Github[2]. We also provide a set of features used to construct one of the baseline classification models for prediction of metaphor/non-metaphor classes at the word level, and instructions on how to replicate that baseline.

3.2 Training phase

In this first phase, data is released for training and/or development of metaphor detection models. Participants can elect to perform cross-validation on the training data, or partition the training data further to have a held-out set for preliminary evaluations, and/or set apart a subset of the data for development/tuning of hyper-parameters. However the training data is used, the goal is to have N final systems (or versions of a system) ready for evaluation when the test data is released.

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

[2] https://github.com/EducationalTestingService/metaphor/tree/master/NAACL-FLP-shared-task, https://github.com/EducationalTestingService/metaphor//tree/master/TOEFL-release

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 VUA and TOEFL datasets.

3.3 Testing phase

In this phase, instances for evaluation are released.[3] Each participating system generated predictions for the test instances, for up to N models.[4] Predictions are submitted to CodaLab[5] and evaluated automatically against the gold-standard labels. Submissions were anonymized. The only statistics displayed were the highest score of all systems per day. The total allowable number of system submissions per day was limited to 5 per team per track. The metric used for evaluation is the F1 score (least frequent class/label, which is "metaphor") with Precision and Recall also available via the detailed results link in CodaLab.

The shared task started on January 12, 2020 when the training data was made available to registered participants. On February 14, 2020, the testing data was released. Submissions were accepted until April 17, 2020. Table 3 shows the submission statistics for systems with a system paper. Generally, there were more participants in the VUA tracks than in TOEFL tracks, and in All POS tracks than in Verbs tracks. In total, 13 system papers were submitted describing methods for generating metaphor/non-metaphor predictions.

	#teams	#submissions
VUA-AllPOS	13	210
VUA-Verbs	11	167
TOEFL-AllPOS	9	247
TOEFL-Verbs	9	181

Table 3: Participation statistics for all tracks.

[3]In principle, participants could have access to the test data by independently obtaining the VUA corpus. The shared task was based on a presumption of fair play by participants.

[4]We set N=12.

[5]https://competitions.codalab.org/competitions/22188

4 Systems

We first describe the baseline systems. Next, we briefly describe the general approach taken by every team. Interested readers can refer to the teams' papers for more details.

4.1 Baseline Classifiers

We make available to shared task participants a number of features from prior published work on metaphor detection, including unigram features, features based on WordNet, VerbNet, and those derived from a distributional semantic model, POS-based, concreteness and difference in concreteness, as well as topic models.

We adopted three informed baselines from prior work. As **Baseline 1: UL + WordNet + CCDB**, we use the best system from Beigman Klebanov et al. (2016). The features are: lemmatized unigrams, generalized WordNet semantic classes, and difference in concreteness ratings between verbs/adjectives and nouns (UL + WN + CCDB).[6] **Baseline 2: bot.zen** is one of the top-ranked systems in the first metaphor shared task in 2018 by Stemle and Onysko (2018) that uses a bi-directional recursive neural network architecture with long-term short-term memory (LSTM BiRNN) and implements a flat sequence-to-sequence neural network with one hidden layer using TensorFlow and Keras in Python. The system uses fastText word embeddings from different corpora, including learner corpus and BNC data. Finally, **Baseline 3: BERT** is constructed by fine-tuning the BERT model (Devlin et al., 2018) in a standard token classification task: 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. Chen et al. (2020) gives more details about the architecture of this baseline. For Verbs tracks, we tune the system on All POS data and test on Verbs,

[6]Baseline 1 is "all-16" in Beigman Klebanov et al. (2018)

as this produced better results during preliminary experimentation than training on Verbs only.

4.2 System Descriptions

illiniMet: RoBERTa embedding + Linguistic features + Ensemble Gong et al. (2020) used RoBERTa to obtain a contextualized embedding of a word and concatenate it with features extracted from linguistic resources (e.g. WordNet, VerbNet) as well as other features (e.g. POS, topicality, concreteness) previously used in the first shared task (Leong et al., 2018) before feeding them into a fully-connected Feedforward network to generate predictions. During inference, an ensemble of three independently trained models using different train/development splits is proposed to yield a final prediction based on majority vote. Using just RoBERTa without linguistic features in an ensemble also generates competitive performance.

DeepMet: Global and local text information + Transformer stacks Su et al. (2020) proposed a reading comprehension paradigm for metaphor detection, where the system seeks to understand the metaphoricity role of each word token in a shorter sequence within a given sentence. Features belonging to five different categories are provided as inputs to the network i.e. global text context, local text context, query word, general POS, fine-grained POS. The features are then mapped onto embeddings before going into Transformer stacks and ensemble for inference. An ablation experiment was also performed with the observation that fine-grained POS and global text features are the most helpful for detecting metaphors.

umd_bilstm: Bi-LSTM + Embeddings + Unigram Lemmas + Spell Correction Kuo and Carpuat (2020) explored the effectiveness of additional features by augmenting the basic contextual metaphor detection system developed by Gao et al. (2018) with one-hot unigram lemma features in addition to GloVe and ELMo embeddings. The authors also experimented with a spell-corrected version of TOEFL data and found it further improves the performance of the Bi-LSTM system.

atr2112: Residual Bi-LSTM + Embeddings + CRF + POS + WN Rivera et al. (2020) proposed a deep architecture that takes as inputs ELMo embeddings that represent words and lemmas, along with POS labels and WordNet synsets. The inputs are processed by a residual Bi-LSTM, then by a number of additional layers, with a final CRF sequence labeling step to generate predictions.

Zenith: Character embeddings + Similarity Networks + Bi-LSTM + Transformer Kumar and Sharma (2020) added lexical and orthographic information via character embeddings in addition to GloVe and ELMo embeddings for an enriched input representation. The authors also constructed a similarity metric between the literal and contextual representations of a word as another input component. A Bi-LSTM network and Transformer network are trained independently and combined in an ensemble. Eventually, adding both character-based information and similarity network are the most helpful, as evidenced by results obtained using cross-validation on the training datasets.

rowanhm: Static and contextual embeddings + concreteness + Multi-layer Perceptron Maudslay et al. (2020) created a system that combines the concreteness of a word, its static embedding and its contextual embedding before providing them as inputs into a deep Multi-layer Perceptron network which predicts word metaphoricity. Specifically, the concreteness value of a word is formulated as a linear interpolation between two reference vectors (concrete and abstract) which were randomly initialized and learned from data.

iiegn: LSTM BiRNN + metadata; combine TOEFL and VUA data Stemle and Onysko (2020) used an LSTM BiRNN classifier to study the relationship between the metadata in the TOEFL corpus (proficiency, L1 of the author, and the prompt to which the essay is responding) and classifier performance. The system is an extension of the authors' system for the 2018 shared task (Stemle and Onysko, 2018) that served as one of the baseline in the current shared task (see section 4.1). Analyzing the training data, the authors observed that essays written by more proficient users had significantly more metaphors, and that essays responding to some of the prompts had significantly more metaphors than other prompts; however, using proficiency and prompt metadata explicitly in the classifier did not improve performance. The authors also experimented with combining VUA and TOEFL data.

Duke Data Science: BERT, XNET language models + POS tags as features for a Bi-LSTM classifier Liu et al. (2020) use pre-trained BERT and XLNet language models to create contextualized embeddings, which are combined with

POS tags to generate features for a Bi-LSTM for token-level metaphor classification. For the testing phase, the authros used an ensemble strategy, training four copies of the Bi-LSTM with different initializations and averaging their predictions. To increase the likelihood of prediction of a metaphor label, a token is declared a metaphor if: (1) its predicted probability is higher than the threshold, or (2) if its probability is three orders of magnitude higher than the median predicted probability for that word in the evaluation set.

chasingkangaroos: RNN + BiLSTM + Attention + Ensemble Brooks and Youssef (2020) use an ensemble of RNN models with Bi-LSTMs and bidirectional attention mechanisms. Each word was represented by an 11-gram and appeared at the center of the 11-gram; each word in the 11-gram was represented by a 1,324 dimensional word embedding (concatenation of ELMo and GloVe embeddings). The authors experimented with ensembles of models that implement somewhat different architecture (in terms of attention) and models trained on all POS and on a specific POS.

Go Figure!: BERT + multi-task + spell correction + idioms + domain adaptation Chen et al. (2020) baseline system (also one of the shared task baselines, see section 4.1) uses BERT – after obtaining the contextualized embeddings of a sentence, a linear layer is applied followed by softmax on each token to predict whether it is metaphorical or not. The authors spell-correct the TOEFL data, which improves performance. Chen et al. (2020) present two multi-task settings: In the first, metaphor detection on out-of-domain data is treated as an auxiliary task; in the second, idiom detection on in-domain data is the auxiliary task. Performance on TOEFL is helped by the first multi-task setting; performance on VUA is helped by the second.

UoB team: Bi-LSTM + GloVe embeddings + concreteness Alnafesah et al. (2020) explore ways of using concreteness information in a neural metaphor detection context. GloVe embeddings are used as features to an SVM classifier to learn concreteness values, training it using human labels of concreteness. Then, for metaphor detection, every input word is represented as a 304-dimensional vector – 300 dimensions are GloVe pre-trained embeddings, plus probabilities for the four concreteness classes. These representations of words are given as input to a Bi-LSTM which outputs a sequence of labels. Results suggest that explicit concreteness information helps improve metaphor detection, relative to a baseline that uses GloVe embeddings only.

zhengchang: ALBERT + BiLSTM Li et al. (2020) use a sequence labeling model based on ALBERT-LSTM-Softmax. Embeddings produced by BERT serve as input to BiLSTM, as well as to the final softmax layer. The authors report on experiments with inputs to BERT (single-sentence vs pairs; variants using BERT tokenization), spell-correction of the TOEFL data, and CRF vs softmax at the classification layer.

PolyU-LLT: Sensorimotor and embodiment features + embeddings + n-grams + logistic regression classifier Wan et al. (2020) use sensorimotor and embodiment features. They use the Lancaster Sensorimotor norms (Lynott et al., 2019) that include measures of sensorimotor strength for about 40K English words across six perceptual modalities (e.g., touch, hearing, smell), and five action effectors (mouth/throat, hand/arm, etc), and embodiment norms from Sidhu et al. (2014). The authors also use word, lemma, and POS n-grams; word2vec and GloVe word embeddings, as well as cosine distance measurements using the embeddings. The different features are combined using logistic regression and other classifiers.

5 Results and Discussion

Table 4 present the results for All POS and Verbs tracks for VUA data. Table 5 present the results for All POS and Verbs tracks for TOEFL data.

5.1 Trends in system design

The clearest trend in the 2020 submissions is the use of deep learning architectures based on BERT (Devlin et al., 2018) – more than half of the participating systems used BERT or its variant. The usefulness of BERT for metaphor detection has been shown by Mao et al. (2019), where a BERT-based system posted F1 = 0.717 on VUA AllPOS, hence our use of a BERT-based system as Baseline 3.

Beyond explorations of neural architectures, we also observe usage of new lexical, grammatical, and morphological information, such as fine-grained POS, spell-corrected variants of words (for TOEFL data), sub-word level information (e.g., character embeddings), idioms, sensorimotor and embodiment-related information.

Rank	Team	P	R	F1
	All POS			
1	DeepMet	.756	.783	.769
2	Go Figure!	.721	.748	.734
3	illiniMet	.746	.715	.730
4	rowanhm	.727	.709	.718
5	Baseline 3: BERT	.712	.725	.718
6	zhengchang	.696	.729	.712
7	chasingkangaroos	.702	.704	.703
8	Duke Data Science	.662	.699	.680
9	Zenith	.630	.716	.670
10	umd_bilstm	.733	.601	.660
11	atr2112	.599	.672	.633
12	PolyU-LLT	.556	.660	.603
13	iiegn	.601	.591	.596
14	UoB team	.653	.548	.596
15	Baseline 2: bot.zen	.612	.575	.593
16	Baseline 1: UL + + WN + CCDB	.510	.696	.589
	Verbs			
1	DeepMet	.789	.819	.804
2	Go Figure!	.732	.823	.775
3	illiniMet	.761	.781	.771
4	Baseline 3: BERT	.725	.789	.756
5	zhengchang	.706	.811	.755
6	rowanhm	.734	.779	.755
7	Duke Data Science	.712	.749	.730
8	Zenith	.667	.775	.717
9	umd_bilstm	.597	.806	.686
10	atr2112	.652	.718	.683
11	PolyU-LLT	.608	.703	.652
12	iiegn	.587	.691	.635
13	Baseline 2: bot.zen	.605	.666	.634
14	Baseline 1: UL + + WN + CCDB	.527	.698	.600

Table 4: **VUA Dataset**: Performance and ranking of the best system per team and baselines, for All POS track (top panel) and for Verbs track (bottom panel).

Rank	Team	P	R	F1
	All POS			
1	DeepMet	.695	.735	.715
2	zhengchang	.755	.666	.707
3	illiniMet	.709	.697	.703
4	Go Figure!	.669	.717	.692
5	Duke Data Science	.688	.651	.669
6	Baseline 3: BERT	.701	.563	.624
7	Zenith	.607	.634	.620
8	umd_bilstm	.629	.593	.611
9	iiegn	.596	.579	.587
10	PolyU-LLT	.523	.602	.560
11	Baseline 2: bot.zen	.590	.517	.551
12	Baseline 1: UL + + WN + CCDB	.488	.576	.528
	Verbs			
1	DeepMet	.733	.766	.749
2	zhengchang	.735	.720	.728
3	illiniMet	.731	.707	.719
4	Go Figure!	.747	.661	.702
5	Duke Data Science	.687	.707	.697
6	Baseline 3: BERT	.624	.694	.657
7	Zenith	.669	.638	.653
8	umd_bilstm	.668	.562	.611
9	PolyU-LLT	.584	.609	.596
10	Baseline 2: bot.zen	.566	.595	.580
11	Baseline 1: UL + + WN + CCDB	.504	.641	.564
12	iiegn	.622	.487	.546

Table 5: **TOEFL Dataset**: Performance and ranking of the best system per team and baselines, for All POS track (top panel) and for Verbs track (bottom panel).

5.2 Performance wrt 2018 shared task

Since the same VUA dataset was used in 2020 shared task as in the 2018 shared task, we can directly compare the performance of the best systems to observe the extent of the improvement. The best system in 2018 performed at F1 = 0.651; the best performance in 2020 is more than 10 points better – F1 = 0.769. Indeed, the 2018 best performing system would have earned the rank of 11 in the 2020 All POS track, suggesting that the field has generally moved to more effective models than those proposed for the 2018 competitions.

The best result posted for the 2020 shared task is on par with state-of-art for VUA All POS task: Dankers et al. (2019) reported F1 = 0.769 for a multi-task learning setting utilizing emotion-related information. The best results obtained by participants of the 2020 shared task for TOEFL are state-of-the-art, improving upon Baseline 1, which is the best published result for this dataset (Beigman Klebanov et al., 2018).

5.3 Performance across genres: VUA

Table 6 shows performance by genre for the VUA data All POS track. The patterns are highly consistent across systems, and replicate those observed for the 2018 shared task – Academic and News genres are substantially easier to handle than Fiction and Conversation. The gap between the best and worst performance across genres for the same system remains wide – between 11.4 F1 points and 24.3 F1 points. Somewhat encouragingly, the gap is narrower for the better performing systems – the top 6 systems show the smallest gaps between best and worst genres (11.4-14.0).

5.4 Performance on VUA vs TOEFL data

Table 7 shows performance and ranks of the best systems for teams that participated in both VUA and TOEFL AllPOS tracks, along with baselines. Overall, the relative performance rankings are consistent – F1 scores are correlated at r = .92 and team ranks are correlated at r = 0.95 across the two datasets. All teams posted better performance on the VUA data than on the TOEFL

Team	All VUA	Acad.	Conv.	Fiction	News	Best to Worst
atr2112	.633	.716 (1)	.510 (4)	.558 (3)	.641 (2)	.206
chasingkangaroos	.703	.761 (1)	.599 (4)	.651 (3)	.714 (2)	.162
PolyU-LLT	.603	.719 (1)	.482 (3)	.476 (4)	.634 (2)	.243
DeepMet	**.769**	**.810** (1)	**.681** (4)	**.718** (3)	**.790** (2)	.129
UoB team	.596	.686 (1)	.485 (4)	.511 (3)	.582 (2)	.201
iiegn	.596	.669 (1)	.521 (3)	.500 (4)	.626 (2)	.169
umd_bilstm	.660	.724 (1)	.537 (4)	.606 (3)	.670 (2)	.187
illiniMet	**.730**	**.768** (1)	**.654** (4)	**.688** (3)	**.743** (2)	.114
rowanhm	.718	.760 (1)	.631 (4)	.678 (3)	.730 (2)	.129
Zenith	.670	.730 (1)	.566 (4)	.583 (3)	.697 (2)	.164
Duke Data Science	.680	.742 (1)	.572 (4)	.617 (3)	.697 (2)	.170
Go Figure!	**.734**	**.784** (1)	**.644** (4)	**.692** (3)	**.741** (2)	.140
zhengchang	.712	.752 (1)	.634 (4)	.669 (3)	.723 (2)	.118
Baseline 3: BERT	.718	.767 (1)	.640 (4)	.684 (3)	.719 (2)	.127
Baseline 2: bot.zen	.593	.673 (1)	.487 (4)	.521 (3)	.602 (2)	.186
Baseline 1: UL+ +WN+CCDB	.589	.721 (1)	.472 (3)	.458 (4)	.606 (2)	.263
Av. rank among genres	–	1.00	3.81	3.19	2.00	.169

Table 6: **VUA Dataset**: Performance (F1-score) of the best systems submitted to All-POS track by genre subsets of the test data. In parentheses, we show the rank of the given genre within all genres for the system. The last column shows the overall drop in performance from best genre (ranked 1) to worst (ranked 4). The top three performances for a given genre are boldfaced.

Team	VUA (rank)	TOEFL (rank)	Diff.
Baseline 1: UL+ +WN+CCDB	.59 (12)	.53 (12)	.06
Baseline 2: bot.zen	.59 (11)	.55 (11)	.04
Baseline 3: BERT	.72 (4)	.62 (6)	.09
PolyU-LLT	.60 (9)	.56 (10)	.04
DeepMet	.77 (1)	.72 (1)	.05
iiegn	.60 (10)	.59 (9)	.01
umd_bilstm	.66 (8)	.61 (8)	.05
illiniMet	.73 (3)	.70 (3)	.03
Zenith	.67 (7)	.62 (7)	.05
Duke Data Science	.68 (6)	.67 (5)	.01
Go Figure!	.73 (2)	.69 (4)	.04
zhengchang	.71 (5)	.71 (2)	.01

Table 7: **VUA vs TOEFL**: Performance (F1 scores) and rankings of participants in both VUA and TOEFL All POS competitions. Column 4 shows the difference in F1 performance between VUA and TOEFL data.

data; the difference (see column 4 in Table 7) averaged 4 F1 points, ranging from just half a F1 point (zhengchang) to 5 F1 points (DeepMet, umd_bilstm, Zenith). The BERT baseline posted a relatively large difference of 9 F1 points; this could be because BNC data is more similar to the data on which BERT has been pre-trained than TOEFL data. We note, however, that participating systems that used BERT showed a smaller performance gap between VUA and TOEFL data; in zhengchang the gap is all but eliminated. This suggests that a BERT-based system with parameters optimized for performance on TOEFL data can close this gap.

Considering TOEFL data as an additional genre, along with the four genres represented in VUA, we observe that it is generally harder than Academic and News, and is commensurate with Fiction in terms of performance, for the three systems with best VUA All POS performance (DeepMet: 0.72 both, Go Figure!: 0.69 both, illiniMet: 0.69 for VUA Fiction, .70 for TOEFL); a caveat to this observation is that the difference between VUA and TOEFL is not only in genre but in the metaphor annotation guidelines as well.

5.5 Performance by proficiency: TOEFL

Table 8 shows performance for All POS track on the TOEFL data by the writer's proficiency level – high or medium. We note that the quality of the human annotations does not appear to differ substantially by proficiency: The average inter-annotator agreement for the high proficiency essays was κ = 0.619, while it was κ = 0.613 for the medium proficiency essays. We observe that generally systems tend to perform better on the higher proficiency essays, although two of the 12 systems posted better performance on the medium proficiency data. However, even though the medium proficiency essays might have deficiencies in grammar, spelling, coherence and other properties of the essay that could interfere with metaphor detection, we generally observe rela-

tively small differences in performance by proficiency – up to 3.5 F1 points, with a few exceptions (zhengchang, Go Figure!). Interestingly, automatic correction of spelling errors does not seem to guarantee a smaller gap in performance (see Chen et al. (2020), Go Figure!).

Team	All	High	Med.	Diff.
PolyU-LLT	.560	.567 (1)	.552 (2)	.015
DeepMet	**.715**	**.724** (1)	**.706** (2)	.018
iiegn	.587	.592 (1)	.583 (2)	.009
umd_bilstm	.611	.620 (1)	.601 (2)	.019
illiniMet	**.703**	**.717** (1)	**.690** (2)	.027
Zenith	.620	.637 (1)	.604 (2)	.033
Duke Data Science	.669	.660 (2)	**.677** (1)	.017
Go Figure!	.692	.713 (1)	.671 (2)	.042
zhengchang	**.707**	**.741** (1)	.674 (2)	.067
Baseline 3: BERT	.624	.636 (1)	.612 (2)	.024
Baseline 2: bot.zen	.551	.535 (2)	.567 (1)	.032
Baseline 1: UL+ WordNet+CCDB	.528	.533 (1)	.524 (2)	.009
Av. rank	–	1.16	1.83	.03

Table 8: **TOEFL Dataset**: Performance (F1-score) of the best systems submitted to All-POS track by proficiency level (high, medium) subsets of the test data. In parentheses, we show the rank of the given proficiency level within all levels for the system. The last column shows the overall drop in performance from best proficiency level (ranked 1) to worst (ranked 4). The top three performances for a given genre are boldfaced.

5.6 Part of Speech

Table 9 shows the performance of the systems submitted to the All POS tracks for VUA and TOEFL data broken down by part of speech (Verbs, Nouns, Adjectives, Adverbs). As can be observed both from the All POS vs Verbs tracks (Tables 4 and 5) and from Table 9, performance on Verbs is generally better than on All POS.[7]

For VUA data, all but one systems perform best on Verbs, followed by Adjectives and Nouns, with the worst performance generally observed for Adverbs. These results replicate the findings from the 2018 shared task and follow the proportions of metaphors in the respective parts of speech, led by Verbs (30%), Adjectives (18%), Nouns (13%), Adverbs (8%). The average gap between best and worst POS performance has also stayed similar – 11 F1 points (it was 9% in 2018).

[7]Performance on Verbs track and performance on Verbs as part of All POS track might differ, since for Verbs track, participants could train their system on verbs-only data, whereas we took submissions to All POS track and analyzed by POS for Table 9.

For the TOEFL data, the situation is quite different. Adjectives lead the scoreboard for all but 3 systems, with Adverbs and Verbs coming next, while Nouns proved to be the most challenging category for all participating systems. Furthermore, the gap between best and worst POS performance is large – 17 F1 points on average, ranging between 11 and 22 points. The best performance on Nouns is only F1 = 0.641; it would have ranked 10th out of 12 on Adjectives. The proportions of metaphorically used Verbs (13%), Adjectives (8%), Nouns (4%), and Adverbs (3%) (based on training data) perhaps offer some explanation of the difficulty with nouns, since nominal metaphors seem to be quite rare. Stemle and Onysko (2020) observed that metaphors occur more frequently in responses to some essay prompts that to others among the 8 prompts covered in the TOEFL dataset; moreover, for some prompts, a metaphor is suggested in the prompt itself and occurs frequently in responses (e.g. whether *broad* knowledge is better than specialized knowledge). It is possible that prompt-based patterns interact with POS patterns in ways that affect relative ease or difficulty of POS for metaphor identification.

6 Acknowledgements

As organizers of the shared task, we would like to thank all the teams for their interest and participation. We would also like to thank Ton Veale, Eyal Sagi, Debanjan Ghosh, Xinhao Wang, and Keelan Evanini for their helpful comments on the paper.

Team	All-POS	Verbs	Adjectives	Nouns	Adverbs	Best to Worst
VUA Dataset						
atr2112	.633	.683 (1)	.602 (2)	.595 (3)	.560 (4)	.12
chasingkangaroos	.703	.737 (1)	.678 (2)	.678 (2)	.648 (4)	.09
PolyU-LLT	.603	.625 (1)	.595 (2)	.581 (3)	.552 (4)	.07
DeepMet	.769	.800 (1)	.733 (3)	.749 (2)	.732 (4)	.07
UoB team	.596	.626 (1)	.587 (2)	.569 (3)	.506 (4)	.12
iiegn	.596	.635 (1)	.581 (2)	.558 (3)	.513 (4)	.12
umd_bilstm	.660	.700 (1)	.642 (2)	.630 (3)	.514 (4)	.19
illiniMet	.730	.770 (1)	.693 (3)	.705 (2)	.633 (4)	.14
rowanhm	.718	.753 (1)	.660 (3)	.706 (2)	.644 (4)	.11
Zenith	.670	.715 (1)	.621 (3)	.637 (2)	.612 (4)	.10
Duke Data Science	.680	.724 (1)	.614 (4)	.654 (2)	.625 (3)	.11
Go Figure!	.734	.775 (1)	.683 (3)	.708 (2)	.681 (4)	.09
zhengchang	.712	.755 (1)	.655 (4)	.684 (2)	.659 (3)	.10
Baseline 3: BERT	.718	.756 (1)	.672 (3)	.695 (2)	.672 (3)	.08
Baseline 2: bot.zen	.593	.637 (1)	.564 (2)	.553 (3)	.513 (4)	.12
Baseline 1: UL + WN + CCDB	.589	.616 (1)	.557 (3)	.564 (2)	.542 (4)	.07
Av. rank among POS	–	1.00	2.69	2.38	3.81	.11
TOEFL Dataset						
PolyU-LLT	.560	.587 (2)	.630 (1)	.462 (4)	.517 (3)	.17
DeepMet	.715	.749 (3)	.757 (2)	.610 (4)	.800 (1)	.19
iiegn	.587	.617 (3)	.667 (1)	.465 (4)	.632 (2)	.20
umd_bilstm	.611	.652 (2)	.693 (1)	.478 (4)	.627 (3)	.22
illiniMet	.703	.718 (3)	.770 (2)	.609 (4)	.786 (1)	.18
Zenith	.620	.650 (2)	.703 (1)	.505 (4)	.600 (3)	.20
Duke Data Science	.669	.697 (3)	.725 (2)	.555 (4)	.741 (1)	.19
Go Figure!	.692	.697 (2)	.749 (1)	.641 (4)	.691 (3)	.11
zhengchang	.707	.728 (3)	.759 (1)	.620 (4)	.731 (2)	.14
Baseline 3: BERT	.624	.644 (2)	.689 (1)	.541 (4)	.583 (3)	.15
Baseline 2: bot.zen	.551	.565 (2)	.611 (1)	.485 (4)	.490 (3)	.13
Baseline 1: UL + WN + CCDB	.528	.543 (2)	.618 (1)	.415 (4)	.531 (3)	.20
Av. rank among POS	–	2.42	1.25	4.00	2.33	.17

Table 9: **VUA and TOEFL Datasets by POS**: Performance (F1-score) of the best systems submitted to All-POS track by POS subsets of the test data. In parentheses, we show the rank of the given POS within all POS for the system. The last column shows the overall drop in performance from best POS (ranked 1) to worst (ranked 4).

References

Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. 2020. Augmenting neural metaphor detection with concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

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.

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.

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.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

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

Jennifer Brooks and Abdou Youssef. 2020. Metaphor detection using ensembles of bidirectional recurrent neural networks. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 523–528.

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

Xianyang Chen, Chee Wee Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: Ets team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

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, Hong Kong, China. Association for Computational Linguistics.

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*.

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.

Jonathan Dunn. 2013. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Conference on Empirical Methods in Natural Language Processing*.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48. Association for Computational Linguistics.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

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.

E Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 183–193.

Dirk Hovy, Shashank Shrivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.

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*, pages 384–392.

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.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

Tarun Kumar and Yashvardhan Sharma. 2020. Character aware models with similarity learning for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

Kevin Kuo and Marine Carpuat. 2020. Evaluating a bilstm model for metaphor detection in toefl essays. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

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, 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.

Shuqun Li, Jingjie Zeng, Jinhui Zhang, Tao Peng, Liang Yang, and Hongfei Lin. 2020. Albert-BiLSTM for sequential metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

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.

Jerry Liu, Nathan OH́ara, Alex Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The Lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, pages 1–21.

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, Florence, Italy. Association for Computational Linguistics.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PloS one*, 8(4):e62343.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroglu. 2016. Prometheus: A corpus of proverbs annotated with metaphors. In *LREC*.

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.

Andrés Torres Rivera, Antoni Oliver, Salvador Climent, and Marta Coll-Florit. 2020. Neural metaphor detection with a residual bilstm-crf model. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

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.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics.

David Sidhu, Rachel Kwan, Penny Pexman, and Paul Siakaluk. 2014. Effects of relative embodiment in lexical and semantic processing of verbs. *Acta psychologica*, 149:32–39.

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.

Egon Stemle and Alexander Onysko. 2018. Using language learner data for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, New Orleans,LA.

Egon Stemle and Alexander Onysko. 2020. Testing the role of metadata in metaphor identification. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

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.

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

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.

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.

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. Association for Computational Linguistics.

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

Mingyu Wan, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang. 2020. Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, Seattle, WA.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Thu ngn at naacl-2018 metaphor shared task: Neural metaphor detecting with cnn-lstm model. In *Proceedings of the Workshop on Figurative Language Processing*, New Orleans,LA.

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