An analysis of language models for metaphor recognition

Arthur Neidlein, Philipp Wiesenbach and **Katja Markert**
Institute of Computational Linguistics
Heidelberg University
{neidlein|wiesenbach|markert}@cl.uni-heidelberg.de

Abstract

We conduct a linguistic analysis of recent metaphor recognition systems, all of which are based on language models. We show that their performance, although reaching high F-scores, has considerable gaps from a linguistic perspective. First, they perform substantially worse on unconventional metaphors than on conventional ones. Second, they struggle with handling rarer word types. These two findings together suggest that a large part of the systems' success is due to optimising the disambiguation of conventionalised, metaphoric word senses for specific words instead of modelling general properties of metaphors. As a positive result, the systems show increasing capabilities to recognise metaphoric readings of unseen words *if* synonyms or morphological variations of these words have been seen before, leading to enhanced generalisation beyond word sense disambiguation.

1 Introduction

Metaphor is a type of figurative language where meaning transfer occurs via similarity between two conceptual domains. In Examples 1 to 3, the metaphors *attacked*, *bashed* and *ceasefire* stem from a transfer from the domain WAR (or FIGHT) to the domain ARGUMENT (Lakoff and Johnson, 1980).[1]

(1) He *attacked* my argument.

(2) He *bashed* my argument.

(3) We declared a *ceasefire* during dinner.

Metaphoric instances can stem from such regular metaphoric patterns equating two domains habitually or conventionally.[2] However, they can also be novel/unconventional such as the famous Emily Dickinson metaphor *Hope is the thing with feathers*, which does not correspond to a well-known metaphorical pattern. Even within a metaphorical pattern such as `Argument is War`, there are degrees of conventionality with Example 1 being more conventional than Examples 2 and 3. To a human, unconventional metaphors tend to be more noticeable.

Metaphor detection has been studied extensively in NLP in recent years (see (Veale et al., 2016; Shutova et al., 2017) for overviews). State-of-the-art approaches in metaphor detection build strongly on language models and word embeddings, with more than half of the participants in the 2020 Shared Task on Metaphor Detection (Leong et al., 2020) using a variant of BERT language models (Devlin et al., 2019). Evaluations on the standard metaphor recognition test sets report scores that creep up steadily, using such methods. We investigate whether these models really are able to learn general properties of metaphor. To do so and to go beyond word sense disambiguation, they should be able to (i) recognise conventional *and* unconventional metaphors (ii) be able to perform well on rarer word types that often

[1] In our examples, metaphoric words are marked in italics.

[2] Lakoff and Johnson (1980) call these patterns *conceptual metaphors*. We will mainly use the term *metaphoric patterns* to distinguish those clearly from specific metaphoric instances, which we simply call metaphors.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3722–3736
Barcelona, Spain (Online), December 8-13, 2020

follow the same metaphoric patterns as frequently seen ones and (iii) be able to generalise across synonyms and morphological variations of word types (such as making the inference from Example 1 to 2).

Our contributions are as follows:

- We conduct a systematic comparison of different sequential metaphor recognition systems on the two most frequently used datasets. Although the two datasets contain different token sets of the *same* underlying corpus (Steen, 2010), we show that one is substantially easier to do well on than the other. We therefore call on future research to stop comparing their results across these two different datasets as this leads to unfair system comparisons.

- We show that the systems behave counter-intuitively by having lower performance on unconventional metaphors than on conventional ones. However, unconventional metaphors are the ones that are particularly relevant as conventional ones can potentially be interpreted with standard word sense disambiguation techniques.

- We show that metaphor recognition systems are strongly dependent on the frequency of word types in the training data.

- As a positive result, we show that the systems have increasing generalisation capabilities in that they perform better on unknown word types if synonyms or morphological variations have been seen in the training data.

2 Models and Datasets

2.1 Models

We report on the following models, all except the baseline being based on a sequence of progressively stronger language models.

Lex-BL is a baseline suggested by Gao et al. (2018) that assigns metaphoric if the word has been annotated as metaphoric more often than literal in the training set, and literal otherwise (including for word types unseen in training).

Wu (Wu et al., 2018) is a system based on skip-gram word2vec (Mikolov et al., 2013), POS tags and word clusters with a CNN and BiLSTM plus ensemble learning, and is the winner of the 2018 Metaphor Detection Shared Task (Leong et al., 2018). As code or system output is not available, we report only the results in their paper and leave it out of fine-grained analysis.

Gao (Gao et al., 2018) uses concatenated GLOVE (Pennington et al., 2014) and ELMO embeddings (Peters et al., 2018) and a BiLSTM.

Mao (Mao et al., 2019) build on Gao et al. (2018) but explicitly model two linguistically-motivated factors that might indicate metaphoricity: firstly, the potential clash between contextual and literal meaning of the word to be labeled, and secondly, the possible conflict between the literal meaning of the word to be labeled and its context.

Dankers (Dankers et al., 2019) enhance a fine-tuned BERT model (Dankers-BERT) with a multitask setup that learns metaphor and emotion labels jointly (Dankers). As code or system output is not available, we report only the results in their paper and leave it out of fine-grained analysis.

Stowe (Stowe et al., 2019) use the ELMO model of Gao et al. (2018) but show that additional, linguistically motivated training data enhances performance. As code or system output is not available, we report only the results in their paper and leave it out of fine-grained analysis.

BERT is a fine-tuned BERT model we implemented. Parameter details are in the Supplement.

ILLI (Gong et al., 2020) is one of the 3 best-performing systems on the 2020 Metaphor Detection Shared Task (Leong et al., 2020). Its most basic form is a simple fine-tuned RoBERTa (Liu et al., 2019) language model (ILLI-ROB). Its most sophisticated version (ILLI-F-ENS) adds a wide variety of linguistic features and an ensemble based on 3 different runs on different train/dev splits. The system code is available but at too short notice for us to conduct a fine-grained analysis of this system yet.

DM is the 2020 Shared Task winner (Su et al., 2020). It uses RoBERTa enriched with POS features and two transformers, one focusing on the whole sentence context and one on a more local context. DM-ENS builds an ensemble across nine different runs of DM. Their system output is available.[3] Our analysis is based on the DM outputs in their `submit` folder, more specifically answer9 for DM as well as ensemble3 for DM-ENS (both for the VUA-ALL-POS task). The results vary only marginally from the reported best results in their paper.

2.2 Datasets

The VUA Metaphor Corpus[4] (Steen, 2010) consists of 115 texts of four different genres: academic, conversation, fiction and news. Each word, including function words, is annotated as *metaphoric* or *literal*, using guidelines based on literal meanings being the more basic or concrete meanings of a word (Group, 2007). Metaphoric readings can still be highly frequent. Example 4 from the corpus contains three conventional, frequent metaphoric readings, including the non-spatial meaning of *in*.

(4) But Nicholas's grand *design collapsed in* 1918

The corpus was used in the 2018 and 2020 VUA Metaphor Detection Shared Task (Leong et al., 2018; Leong et al., 2020).[5] The shared tasks include the VUA-ALL-POS task where all *content* words in a sentence (adjectives, verbs without *have, do, be*, nouns, adjectives) have to be labeled. Although the original corpus also labels function words for metaphoricity, the VUA-ALL-POS task does not evaluate systems on function words. Therefore in Example 4, four tokens (*Nicholas,grand,design,collapsed*) would have to be labeled as metaphoric or literal.[6]

Four other papers that did not participate in the shared task (Gao et al., 2018; Mao et al., 2019; Dankers et al., 2019; Stowe et al., 2019) also use the VUA corpus but use quite different subsets of the corpus than the shared task VUA-ALL-POS does. In the VUA-ALL-POS task all sentences in the VUA texts are used whereas Gao et al. (2018), Mao et al. (2019), Dankers et al. (2019) and Stowe et al. (2019) use a much smaller subset of sentences, for reasons unknown. In addition, these four papers also evaluate on function words in this smaller subset, which makes a substantial difference.

We handle these two setups in two separate tasks: firstly, the original VUA-ALL-POS Shared Task data[7] and secondly, VUA-SEQ which uses the data in (Gao et al., 2018)[8], subsequently used by (Mao et al., 2019; Dankers et al., 2019; Stowe et al., 2019). Statistics on these datasets are given in Table 1. Using only content words means that VUA-ALL-POS evaluates on fewer tokens although it has more sentences and that it contains fewer metaphors per sentence.

3 Results

We use the standard VUA-ALL-POS and VUA-SEQ training/test splits. Evaluation measures are precision, recall and F1 for the metaphoric class as well as accuracy on all target tokens. Table 2 shows overall results. For Lex-BL, Gao, Mao and BERT we had working code and ran that on both datasets as well as reporting original results from the Gao and Mao papers. For DM and DM-ENS we report and analyse output from their Github repository, for others we report only original results from their papers.

Dataset comparison. Results on VUA-ALL-POS are overall considerably lower than on VUA-SEQ for *equivalent* models. For example, our BERT model achieves F1 of 77.5 on VUA-SEQ but only 69.7 on VUA-ALL-POS. Similarly our rerun of Mao et al. (2019) achieves 74.3 on VUA-SEQ (identical to their reported results) but only 65.5 on VUA-ALL-POS. This is because VUA-SEQ also evaluates on function word metaphors that are easier to classify. Therefore, comparisons in various papers that do

[3] `https://github.com/YU-NLPLab/DeepMet`

[4] `http://ota.ahds.ac.uk/headers/2541.xml`

[5] The 2020 Task has also used newly annotated essay data for metaphor recognition in sequences. As the VUA data is up-to-now the by far most frequently used data, we assess the state-of-the art on this dataset.

[6] The label ALL-POS is slightly confusing given that function words are excluded but mainly serves to distinguish from yet another version of the shared task which looks at verbs only.

[7] `https://github.com/EducationalTestingService/metaphor/tree/master/VUA-shared-task`

[8] `https://github.com/gao-g/metaphor-in-context`

VUA Data	#tokens	% M	#S	#M/S
-SEQ$_{all}$	205,425	11.6	10,567	3.4
-SEQ$_{trn}$	116,622	11.2	6,323	3.3
-SEQ$_{dev}$	38,628	11.6	1,550	4.0
-SEQ$_{tst}$	50,175	12.4	2,694	3.4
-ALL-POS$_{all}$	94,807	15.8	16,202	2.4
-ALL-POS$_{trn}$	72,611	15.2	12,122	2.3
-ALL-POS$_{test}$	22,196	17.9	4,080	2.5

Table 1: Statistics for the VUA Datasets, including the number of target tokens to be classified, the percentage of metaphors among all target tokens, the number of sentences and the average number of metaphors in sentences with at least one metaphoric word.

Results on VUA-SEQ

System	P	R	F1	Acc
Lex-BL	68.3	43.6	53.2	90.5
Gao	71.6	73.6	72.6	93.1
Gao Rerun	76.0	69.2	72.4	93.4
Mao	73.0	75.7	74.3	93.8
Mao Rerun	76.2	72.6	74.3	93.8
Dankers-BERT	–	–	76.3	–
Dankers	–	–	76.9	–
Stowe	–	–	73.8	–
BERT	**78.0**	**76.9**	**77.5**	**94.4**

Results on VUA-ALL-POS

System	P	R	F1	Acc
Lex-BL	65.6	35.7	46.2	85.1
Wu	60.8	70.0	65.1	-
Gao Rerun	68.4	59.7	63.8	87.8
Mao Rerun	71.7	60.2	65.5	88.6
BERT	75.4	64.7	69.7	89.9
ILLI-ROB	75.6	68.6	72.0	–
ILLI-F-ENS	74.6	71.5	73.0	
DM	72.8	72.6	72.7	90.2
DM-ENS	**76.5**	**76.8**	**76.6**	**91.6**

Table 2: Results on VUA-SEQ test (50,175 tokens, including function words) and VUA-ALL-POS test (22,196 tokens)

not distinguish the two setups are inherently unfair and this widespread practice should not be continued. Thus, for example, the Gao model is not better than the 2018 Shared Task winner Wu as claimed in Gao et al. (2018), when compared on the same dataset and the same kind of part-of-speech; and the ILLINIMET paper (Gong et al., 2020) disadvantages itself by comparing their results on VUA-ALL-POS negatively to Mao et al. (2019), which they clearly beat when looking at the right dataset comparison.

Language model improvements. The gains achieved by exploiting state-of-the-art language models are usually higher than the ones achieved by additional linguistic modeling or insights. For example, on VUA-SEQ, the gain in moving from ELMO (Gao et al., 2018) to a fine-tuned BERT model with an otherwise similar setup was 4.9 F-measure points, whereas the gain from ELMO (Gao et al., 2018) to the inclusion of more complex linguistic modelling (Mao et al., 2019) is only 1.7 F-measure points. The gain when moving from Dankers-BERT to a multi-task model on top of BERT is only 0.6 F-measure points (Dankers et al., 2019). On VUA-ALL-POS, we again see a steady improvement with better language models, from ELMO in Gao/Mao (F1 65.5) to BERT (F1 69.7) to RoBERTa in ILLI-ROB (F1 72.0). Here again more sophisticated features (from ILLI-ROB's F1 72.0 to ILLI-F-Ens F1 73.0) yield less of an improvement than better language models, although part-of-speech features play a positive role in both ILLI-F-ENS and in DM. Especially important is reducing performance variation by extensive ensemble modeling (from DM's 72.7 F1 to DM-ENS 76.6 F1).

Does this mean that standard language models indeed learn metaphor properties and generalise within metaphorical patterns? We will now conduct further linguistic analysis to adress this question.

4 Analysis

We will now investigate (i) how well the current systems handle conventional vs. novel metaphors, (ii) if they can handle frequent and less frequent word types and (iii) what the influence of morphology as well as semantic similarity is on their capability to handle metaphoric usage of word types not seen in training. In our opinion, frequency and conventionality analysis are crucial to test whether the system mainly recognises frequently seen, word-specific meanings that might also be specified in dictionary entries (see the metaphors in Example 4 in Section 2.2) or whether it is able to generalise the *concept* of metaphor to word types not seen in training or newly occurring metaphoric transfers.

We conduct the analysis on our reruns of Gao and Mao as well as BERT on VUA-SEQ and extend the analysis with the DM models on VUA-ALL-POS. This includes the state-of-the art systems on both datasets as well as 3 different language models (and extensions).

4.1 Novel vs. Conventional Metaphors.

Metaphors can be conventional (Example 4) or novel (see *goose-step* in Example 5 from the VUA corpus).

(5) Ron Todd [...] warned that party leaders could not expect everybody to *'goose-step'* in the same direction [...]

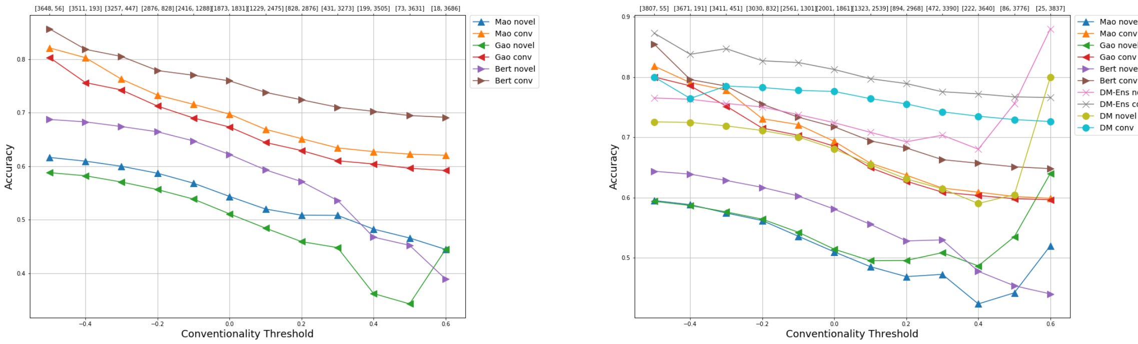
Conventional metaphors are frequent word usages that often have their own dictionary entries whereas novel readings are rare and cannot be found in standard lexical resources. Other aspects also contribute to a metaphor's conventionality, such as whether they do follow a metaphoric pattern. Recognising novel metaphors is important: Shutova (2015) argues "that NLP applications do not necessarily need to address highly conventional and lexicalized metaphors that can be interpreted using standard word sense disambiguation techniques".

Do Dinh et al. (2018) have extended the VUA corpus with reliable novelty scores for *content word metaphors*. Their annotation guidelines define conventionality and novelty based on frequency of use (*often used in everyday language* vs. *not usually used in everyday language*). The scores range from -1 indicating conventional metaphors to 1 for the most novel metaphors. For example, the metaphor in Example 5 has the score 0.765.

Whereas Do Dinh et al. (2018) and Simpson et al. (2019) tackle novelty scoring given gold standard metaphoric/literal information, we investigate how the novelty of a metaphor affects automatic methods for finding metaphors in the first place. Figure 1a and 1b show performance on conventionalised vs. novel metaphors for all systems on metaphoric content words with novelty scores. The x-axis shows the conventionality threshold t and the y-axis shows accuracy/recall. The graphs depict results for conventional metaphors with a novelty score below t and for novel metaphors with a novelty score above t. For example, on VUA-SEQ, our BERT model achieves an accuracy just below 0.6 on the 828 metaphoric content words with a novelty score higher than 0.2. In contrast, it achieves an accuracy of over 70% on the 2876 metaphoric content words with a novelty score lower than 0.2, indicating a substantially better performance on conventional metaphors.

Within each model, the curve for conventionalised metaphors is consistently above the curve for novel metaphors as long as the buckets have a reasonable size.[9] Conventionalised metaphors are therefore recognized much more easily than novel ones. This is interesting as, from a human perspective, novel metaphors are easier to "notice", and suggests that the algorithms might mainly learn different word senses instead of general properties of metaphor, such as the fact that many metaphoric readings show a contrast to their dictionary sense(s) or a contrast to the surrounding context.

[9] DM and DM-ENS perform well for novel metaphors in VUA-ALL-POS from the threshold of 0.6 onwards, a bucket which contains only 25 novel metaphors — classification changes for 3-4 novel items look like very large differences in the graph. It would be interesting to see whether this performance on novel metaphors would hold for (non-existent) larger datasets annotated for novel metaphors.



(a) VUA-SEQ (3704 tokens)

(b) VUA-ALL-POS (3862 tokens)

Figure 1: Accuracy of all models on *metaphoric content words with novelty scores*, measured separately for novel and conventionalised metaphors for different thresholds. Novelty thresholds are shown on the x-axes. The number of instances is shown at the top of the graph with the number of items above the corresponding novelty threshold first and the number of items below the threshold second.

4.2 Frequent vs Infrequent Word Types.

To further investigate how far the algorithms generalise across different word types and their specific meanings, we show the performance of the systems on word types grouped by frequency in the training set in Tables 3 and 4.

fr. in Train	#tok Test	#types Test	Lex-BL	Gao	Mao	Bert
0	4847	2807	– (86.4)	45.1 (86.4)	48.3 (87.3)	53.5 (87.3)
1-10	7631	3349	49.4 (83.1)	63.0 (86.3)	65.1 (86.4)	70.3 (88.6)
11-50	6895	833	55.6 (86.5)	73.9 (90.8)	74.3 (90.7)	79.5 (93.5)
51-100	2805	86	66.7 (90.5)	79.5 (93.6)	80.7 (93.7)	85.7 (95.3)
101 - ∞	27,997	125	60.04 (94.2)	83.2 (97.2)	85.3 (97.5)	86.0 (97.6)
all	50,175	7200	53.2 (90.5)	72.4 (93.4)	74.3 (93.8)	77.5 (94.4)

Table 3: F-measure (accuracy in parenthesis) on different frequency buckets in VUA-SEQ. The frequency buckets are given in the first column, the number of tokens in the test set that belong to each bucket in the second column and the number of types in the test set belonging to each bucket in the third column. For example, there are 2807 word types in the test set that have never been seen in the training set. 125 word types in the test set have been seen over 100 times in training, making up 27,997 tokens of all test tokens.

Overall, F-measure and accuracy increases with the number of times the word type has been seen in training for all models. For example, the best-performing model, BERT, on VUA-SEQ (Table 3) achieves an F-measure of 53.5 on words whose type has not been seen in training, but already 70.3 on words whose type has been seen 1-10 times in training. The one exception is a drop in performance on F-measure for all models on the highest frequency bucket in VUA-ALL-POS (Table 4). Investigation showed that this bucket included only 46 word types, including also word types such as *Yes, Mm, er, also* which were rarely used metaphorically.[10] Thus, the percentage of metaphors in this bucket is much smaller than in the remainder of the corpus, making F-measure more volatile. This is not true for the high frequency bucket in VUA-SEQ which includes many prepositions which are frequently annotated as metaphors (see the non-spatial meaning of *in* in Example 4).

Mao et al. (2019) explicitly encode clashes between literal word meaning and contextual meaning as

[10] Given that the evaluation on VUA-ALL-POS is normally restricted to content words this seems to be a small number of noise words included in the test tokens.

fr. in Train	#tok Test	#types Test	Lex-BL	Gao	Mao	Bert	DM	DM-Ens
0	4050	2583	– (85.6)	45.8 (84.3)	46.1 (86.1)	53.3 (87.2)	61.1 (88.7)	64.9 (89.9)
1-10	7094	3329	49.5 (82.7)	59.0 (85.0)	61.7 (86.2)	67.4 (87.9)	71.8 (88.4)	76.4 (90.0)
11-50	6449	1069	51.1 (86.0)	72.0 (90.6)	73.1 (90.9)	75.3 (91.8)	76.9 (91.8)	81.8 (93.2)
51-100	2180	109	64.6 (85.0)	77.9 (90.1)	77.4 (89.7)	81.6 (92.0)	81.7 (91.8)	84.8 (92.9)
101 - ∞	2423	46	27.6 (88.9)	67.2 (92.5)	68.6 (92.8)	71.1 (93.2)	69.4 (92.4)	75.7 (93.7)
all	22,196	7036	46.2 (85.1)	63.8 (87.8)	65.6 (88.6)	69.7 (89.9)	72.7 (90.2)	76.6 (91.6)

Table 4: F-measure (accuracy in parenthesis) on different frequency buckets in VUA-ALL-POS (all test tokens). Columns correspond to the columns in Table 3.

a metaphoricity indicator on top of Gao's Elmo model, leading to some performance improvements also for word types not seen in training when compared to Gao et al. (2018) (improving from an F1 of 45.1 to 48.3 on unseen word types on VUA-SEQ, Table 3). These improvements are dwarfed by just moving to a stronger language model such as BERT (F1 53.5 on unseen types in VUA-SEQ) but it is possible that the improvements would also carry over when the additional linguistic modelling would be stacked on top of BERT.

It might not seem surprising that all algorithms perform better on word types more often seen in training, but we believe that this type of analysis should be given regularly to check the model's dependence on word-specific labeled data and its ability to generalise.

4.3 The impact of morphology and lexical relations

All models are still able to recognise some metaphors for word types not seen in training (henceforth, *unseen types*). We now investigate when the models are able to generalise to such unseen types.

First, we look at whether performance on unseen word types whose morphological variants have been seen in training is higher than on other unseen word types. This would be plausible as morphological variants will often be close in embedding space and also often undergo the same metaphoric pattern shifts. For example, the AFFECTION IS WARMTH metaphoric pattern is instantiated by *warm greeting*, *warmer greeting* as well as *the warmth of his greeting*. We also hypothesized that inflectional variations probably behave more similarly than derivational variations. We therefore distinguished between exact word type seen, word type not seen but an inflectional variation seen and neither word type nor inflectional variation seen but derivational variant seen. Potential derivational variations were extracted via WordNet (Miller et al., 1990). Tables 5 and 6 show that unseen types where morphological variations had been seen are indeed easier than other unseen types for all systems. For example, on VUA-SEQ F-measure for BERT gradually gets worse from seen types (F1 of 80.3) to 63.8 for word types that have only an inflectional variant seen in training to 54.9 for word types that have only a derivational variant seen in training to 47.4 for word types that have neither itself, nor an inflectional or derivational variant seen in training (Table 5). For all systems but DM-ENS, performance on word types where inflectional variations have been seen is higher than if only derivational variations have been seen. For Gao, Mao and BERT, performance on types where no variation has been seen in training might actually not be better than just assigning *literal* as Lex-BL does for the unseen cases — we see a drop in accuracy compared to Lex-BL for these models (last column in Tables 5 and 6) as well as low precision for metaphor recognition (precision not shown in the tables).

	type seen	infl.var. seen	deriv.var seen	no var seen
Lex-BL	56.9 (90.9)	– (76.0)	– (80.7)	– (89.3)
Gao	75.5 (94.2)	57.0 (80.1)	41.9 (79.0)	39.5 (88.5)
Mao	77.2 (94.5)	56.7 (79.5)	43.1 (80)	43.9 (89.7)
Bert	80.3 (95.2)	63.8 (81.9)	54.9 (82.4)	47.4 (89.0)
#tok test	45,328	879	290	3678
#types test	4393	603	196	2013

Table 5: F-Measure (accuracy) on VUA-SEQ with regard to morphological variants seen in training

	type seen	infl. var. seen	deriv. var seen	no var seen
Lex-BL	51.0 (85.0)	– (78.5)	– (84.4)	– (88.0)
Gao	67.1 (88.6)	55.1 (80.6)	41.5 (82.6)	41.0 (85.6)
Mao	68.7 (89.2)	53.7 (81.6)	42.1 (84.1)	42.0 (87.8)
Bert	72.6 (90.5)	60.5 (83.3)	55.2 (85.9)	48.7 (88.6)
DM	74.7 (90.6)	65.3 (84.7)	64.3 (89.1)	58.1 (89.9)
DM-ENS	78.7 (92.0)	68.8 (86.1)	71.7 (90.6)	60.9 (91.0)
#tok test	18, 146	918	276	2856
#types test	4553	622	187	1777

Table 6: F-measure (accuracy) on VUA-ALL-POS with regard to morphological variants seen in training

In a second study, we look at the performance on unseen word types when synonyms have been seen in training. Synonyms also are close in embedding spaces and also often share metaphorical patterns (see *attack* and *bash* in Example 1 and 2 in the Introduction). Therefore, language models might fare better when a synonym has been seen. We extract synonyms of a word from WordNet. Table 7 shows unseen word types where synonyms have been seen before are indeed easier than unseen word types where synonyms have not been seen. This holds for all systems without exception. For example, on VUA-ALL-POS, performance of the Shared Task Winner DM-ENS has an F1 of 78.8 for seen word types, falling to 68.3 for unseen word types where a synonym has been seen and to 56.8 for unseen word types where no synonym has been seen.

	VUA-SEQ			VUA-ALL-POS		
	type seen	syn seen	no syn seen	type seen	syn seen	no syn seen
Lex-BL	56.9 (90.9)	– (76.1)	– (92.4)	51.0 (85.0)	– (77.7)	– (91.4)
Gao	75.5 (94.2)	52.7 (79.3)	30.5 (90.6)	67.1 (88.6)	52.9 (79.3)	33.1 (87.9)
Mao	77.2 (94.5)	53.6 (79.4)	37.8 (91.9)	68.7 (89.2)	53.5 (82.0)	33.0 (89.2)
Bert	80.3 (95.2)	58.4 (79.9)	44.2 (91.7)	72.6 (90.5)	56.1 (81.2)	47.9 (91.6)
DM	–	–	–	74.7 (90.6)	65.1 (84.2)	53.1 (92.0)
DM-ENS	–	–	–	78.7 (92.0)	68.3 (85.6)	56.8 (93.0)
#tok test	45,328	1792	2055	18,146	1713	2337
#types test	4393	1238	1574	4553	1199	1387

Table 7: F-measure (accuracy in parenthesis) with regard to synonyms seen in training.

In conclusion, current models seem to be able to generalise to a certain degree to unseen word types as long as they are synonyms or morphological variations of seen ones. We give two examples of metaphors in the test set of VUA-ALL-POS (i) that all or most systems identified correctly, (ii) the type of which has not been seen in training and (iii) for which morphological variations or synonyms have been seen. The test example comes first and a similar metaphor from the training set second after an arrow.

(6) . . . to . . . *punctuate* aspects of Holly's life ← *stressed* different facets of Kahlo's public persona

(7) the *richness* of their exquisitely-sculpted decoration ← the colours were *rich*

Of course, due to the black box nature of the language models, the extensive pretraining they undergo before fine-tuning and other interferences such as context similarity, we cannot claim that these were the actual examples that the models generalised from. However, the quantitative data shows that some generalisations do take place.

4.4 The interaction of word frequency and unconventionality

There is a moderate inverse correlation between metaphoric novelty and word frequency (Do Dinh et al., 2018). High frequency words tend to have many conventionalised metaphoric senses; however, low frequency is not necessarily an indication of novel metaphor usage as low frequency words can also follow the metaphoric patterns of their high frequency synonyms (such as *tussle* being used for non-physical arguments just like *attack*). We therefore investigate the interaction between novelty and word frequency, in particular whether for low frequency word types performance still depends on novelty/conventionality.

Figure 2 shows a heatmap that displays the interaction between frequency count in training on the x-axis and conventionality scores on the y-axis for the 3704 VUA-SEQ content word metaphors with novelty scores in the test set. Similar to the analysis in Do Dinh et al. (2018), we see that metaphors using high frequency words normally do not have high novelty scores (right-most column): of 262 metaphoric test tokens whose type has been seen more than 100 times in training, 222 have a novelty score equal or below zero.

We enhance this analysis by showing the accuracy/recall of the BERT model on these subgroups in the fields of the heatmap. We see that even for low frequency words (two left-most columns), conventionality still matters and unconventional metaphors tend to be harder to recognise than conventional ones. For example, even for unseen word types (left-most column), performance on conventional metaphors with novelty scores below 0 is 66%, and then gradually decreases to 53%, 52%, 50% and 38% for less conventional metaphors. Therefore word type frequency does not account for all variation in classifier performance.

The picture is not always completely clear for all models and across both datasets. Especially, small bucket sizes for some fields in the heatmap do not allow firm conclusions. However, the general message holds. Further heatmap examples for other systems on VUA-ALL-POS can be found in the supplementary material.

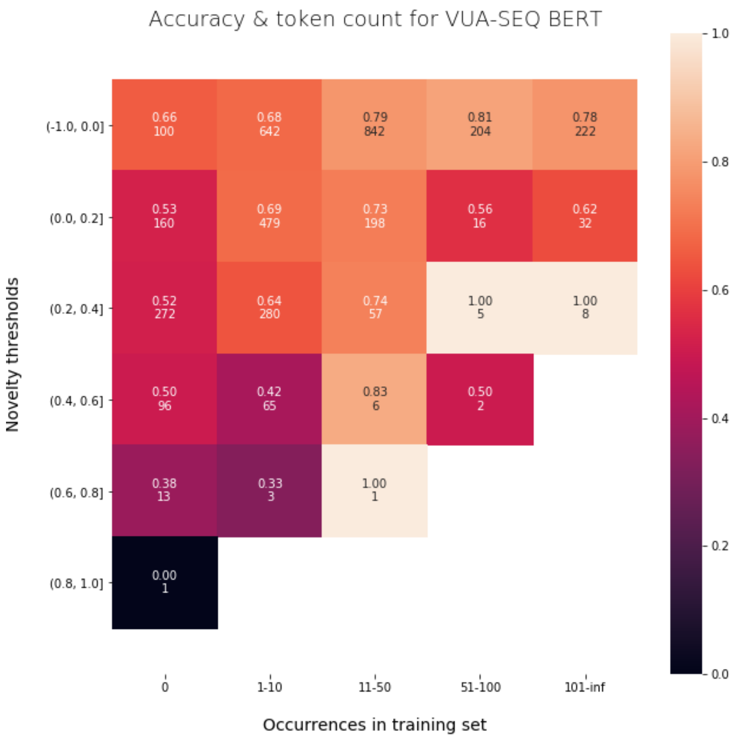


Figure 2: Heatmap showing the interaction between frequency and conventionality as well as classifier performance for the 3704 metaphors with a conventionality score in the VUA-SEQ test set. On the x-axis, we find how often a word type was seen in training. On the y-axis, we have buckets of conventionality scores. In the fields we see the number of test tokens in the bucket as well as accuracy/recall of the BERT model on this bucket.

5 Related Work

Datasets. In some datasets, each word is labeled for metaphoricity (VUA Metaphor corpus, (Steen, 2010)) whereas in others only one target word in a bigram or a sentence is labeled (Mohammad et al., 2016; Shutova et al., 2016; Birke and Sarkar, 2006; Tsvetkov et al., 2014; Turney et al., 2011, among others). We concentrate on datasets where each word is labeled as (i) these are highly appropriate for the sequence labeling tasks that language models excel at and (ii) the 2018 and 2020 Metaphor Shared Tasks (Leong et al., 2018; Leong et al., 2020) use such corpora. We have shown that it matters substantially which dataset partition and setup within the VUA corpus you use and encourage future work to not compare systems working on the two different setups anymore.

Most datasets include only a binary metaphor/literal annotation per word, making it hard to assess system capabilities for the recognition of various metaphor types, such as conventional vs. novel metaphors, deliberately used vs. unintentional metaphors (Steen, 2008) or different domain mappings. Some exceptions exist, such as the conventionality annotation in (Do Dinh et al., 2018; Dunn, 2014), an annotation akin to deliberateness in (Klebanov and Flor, 2013) and annotated domain mappings in (Shutova and Teufel, 2010). However, most of these were small scale and/or are not publically available, the exception being the conventionality ratings by Do Dinh et al. (2018), which we use in this paper.

Metaphor recognition. Data-driven approaches to metaphor recognition (Turney et al., 2011; Tsvetkov et al., 2014; Shutova et al., 2016; Shutova et al., 2017; Bulat et al., 2017; Rei et al., 2017; Köper and im Walde, 2017; Wu et al., 2018; Gao et al., 2018; Gutierrez et al., 2016; Mao et al., 2018; Mao et al., 2019; Dankers et al., 2019; Stowe et al., 2019; Su et al., 2020; Gong et al., 2020, among others) use a variety of information sources such as abstractness/concreteness features, semantic class information, part-of-speech tags, property norms and outside lexical databases as well as multimodal and multilingual information. The recent state of the art models we discuss (Gao et al., 2018; Wu et al., 2018; Mao et al., 2019; Dankers et al., 2019; Stowe et al., 2019; Gong et al., 2020; Su et al., 2020) use sequence labeling and build on embeddings and/or language models. Leong et al. (2020) state that more than half of participants in the 2020 Shared Task use a variation of BERT. We investigate their properties and performance levels in more detail than previously done, including analysis for conventionality, frequency and generalisation via morphology and semantic similarity.

Novel vs. conventionalized metaphors. We investigated how conventionality impacts metaphor recognition. Recent work (Dunn, 2014; Do Dinh et al., 2018; Parde and Nielsen, 2018; Simpson et al., 2019) has assigned novelty scores to (given) metaphors. However, they have either not investigated the influence of novelty on metaphor detection per se or not worked in a sequence labeling, full-text paradigm. We have shown that assigning metaphor novelty scores assuming that metaphors have already been reliably detected is currently somewhat unrealistic as a metaphor's novelty has a strong influence on being detected in the first place by current models.

6 Conclusion and Future Work

We compared several recent models for metaphor recognition, which all build on language models and/or embeddings. We cast doubt on the capability of these models to actually learn general properties of metaphor as the models do not perform well on non-conventionalised metaphors and rarer word types. They excel on frequently seen word types with conventionalised metaphoric meanings, which is more akin to word sense disambiguation. However, they do show generalisation capabilities beyond word sense disambiguation for words unseen in training *if* morphological variations or synonyms have been seen in training.

The latter finding suggests that metaphoric patterns that hold across morphological variations and synonyms might to a degree be learnt by current systems. To follow this up, in the future, we will extend current metaphor corpora by annotating them for metaphorical patterns (or what Lakoff and Johnson (1980) call conceptual metaphors) as well as existence of annotated metaphorical meanings in WordNet.

References

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 Linguistic (EACL)*.

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

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. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1424.

Jonathan Dunn. 2014. Measuring metaphoricity. In *Proceedings of the 52nd Annual Meeting of the ACL (Volume 2: Short Papers)*, pages 745–751.

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

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*, pages 146–153.

Pragglejazz Group. 2007. Mip: a method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

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 ACL (Volume 1: Long Papers)*, pages 183–193.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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.

George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The journal of Philosophy*, 77(8):453–486.

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

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*, pages 18–29.

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. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the ACL*. Association for Computational Linguistics (ACL).

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

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.

Natalie Parde and Rodney D Nielsen. 2018. Exploring the terrain of metaphor novelty: A regression-based approach for automatically scoring metaphors. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

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

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *LREC*, volume 2.

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 NAACL: 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. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with gaussian process preference learning. In *Proceedings of the 57th Conference of the ACL*, pages 5716–5728.

Gerard Steen. 2008. The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol*, 23(4):213–241.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Kevin Stowe, Sarah Moeller, Laura Michaelis, and Martha Palmer. 2019. Linguistic analysis improves neural metaphor detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 362–371.

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*, pages 30–39.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 248–258.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on EMNLP*, 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.

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

7 Supplement to: An analysis of language models for metaphor recognition

7.1 BERT Parameter Details

We use BERT with a dropout and a linear classification layer stacked on top of BERT.

Since BERT has a fixed vocabulary of 30,522 words, we use the standard BERT-tokenizer. Start of sentence and end of sentence markup tokens are prepended and appended to each sentence. We use the uncased base version of BERT (12 Layers, hidden size of 768, 12 heads) since it offers more compact meaning representations than the large version (1024) and is more stable when fine-tuning on small datasets according to Devlin et al. (2019). All hyperparameters are set to standard values. Dropout probability is 0.1. The learning rate is 5e-5 for the Adam optimizer (Kingma and Ba, 2014) after a linear warmup on the first 10% of the training steps. Weight decay is 0.01, where bias vectors and normalization layers are excluded from the decay. The training batch size is set to 16 instances. When facing GPU memory issues this value is decreased to 12. Following Devlin et al. (2019) we use only 3 training epochs. If words in the input sentence get tokenized to multiple sub-tokens, we only use the first sub-token for sequence labeling evaluation.

7.2 Results by POS Tag and genre

It is standard to give the results for the 4 genres in the corpus as well as on different POS. Results for VUA-ALL-POS can be found in the 2018 and 2020 Shared Task reports The corresponding table for VUA-SEQ is given below. We do not observe any differences in tendencies to what has been previously reported: adjectives and nouns are harder than verbs and adverbs; conversational texts are the most difficult genre.

	# Test Tokens	BERT	Gao et al (2018)	Mao et al (2019)
VERB	9,872	74.1	68.4	69.8
NOUN	8,588	68.6	60.0	64.1
ADJ	3,965	64.2	61.9	60.9
ADV	3,393	76.2	60.3	63.6
ADP	5,300	90.9	88.0	89.5
PART	1,463	61.6	56.6	61.2
News	12,324	76.6	72.0	74.3
Conv.	13,270	69.6	64.4	66.1
Fict.	10,966	73.0	66.1	69.6
Acad.	13,615	83.4	78.9	80.3

Table 8: F-measures on VUA-SEQ measured for PoS-tags and Genres.

7.3 Heatmap Examples for VUA-ALL-POS

All heatmaps below show the interaction between frequency and novelty for the 3862 metaphors with a novelty score in VUA-ALL-POS. On the x-axis we find how often a word type was seen in training, on the y-axis we have buckets of novelty scores. In the fields we see the number of test tokens in the bucket as well as accuracy/recall on this bucket. We show heatmaps for the three best-performing models.

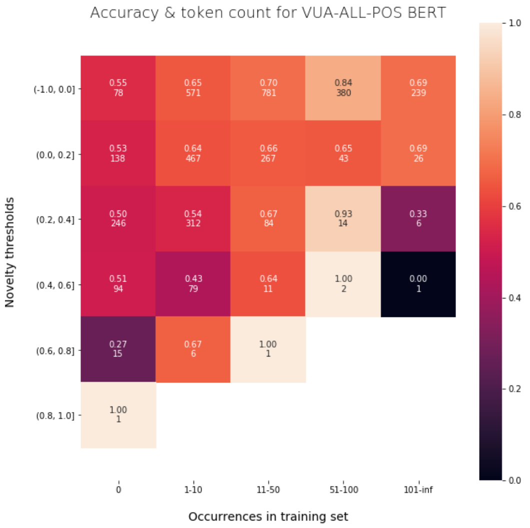


Figure 3: BERT on VUA-ALL-POS

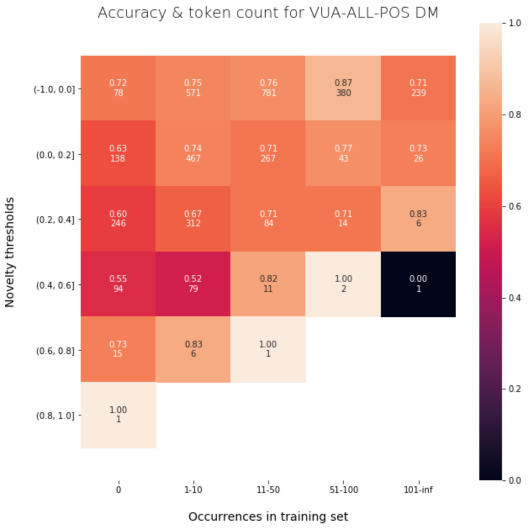


Figure 4: DM on VUA-ALL-POS

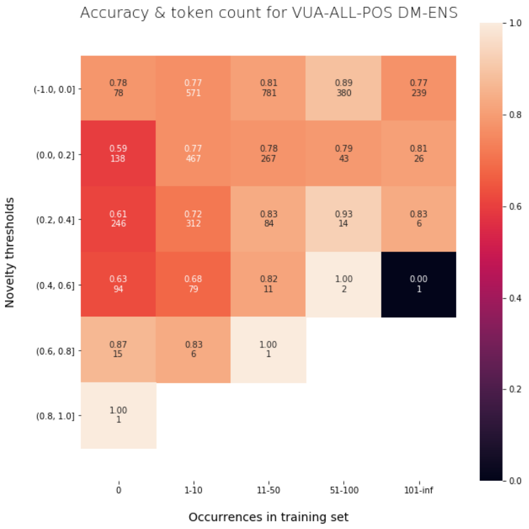


Figure 5: DM-ENS on VUA-ALL-POS