Contextual Modulation for Relation-Level Metaphor Identification

Omnia Zayed, John P. McCrae, Paul Buitelaar
Insight SFI Research Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway, Ireland
{omnia.zayed, john.mccrae, paul.buitelaar}@insight-centre.org

Abstract

Identifying metaphors in text is very challenging and requires comprehending the underlying comparison. The automation of this cognitive process has gained wide attention lately. However, the majority of existing approaches concentrate on word-level identification by treating the task as either single-word classification or sequential labelling without explicitly modelling the interaction between the metaphor components. On the other hand, while existing relation-level approaches implicitly model this interaction, they ignore the context where the metaphor occurs. In this work, we address these limitations by introducing a novel architecture for identifying relation-level metaphoric expressions of certain grammatical relations based on contextual modulation. In a methodology inspired by works in visual reasoning, our approach is based on conditioning the neural network computation on the deep contextualised features of the candidate expressions using feature-wise linear modulation. We demonstrate that the proposed architecture achieves state-of-the-art results on benchmark datasets. The proposed methodology is generic and could be applied to other textual classification problems that benefit from contextual interaction.

1 Introduction

Despite its fuzziness, metaphor is a fundamental feature of language that defines the relation between how we understand things and how we express them (Cameron and Low, 1999). A metaphor is a figurative device containing an implied mapping between two conceptual domains. These domains are represented by its two main components, namely the tenor (target domain) and the vehicle (source domain) (End, 1986). According to the conceptual metaphor theory (CMT) of Lakoff and Johnson (1980), which we adopt in this work, a concept such as *"liquids"* (source domain/vehicle) can be borrowed to express another such as *"emotions"* (target domain/tenor) by exploiting single or common properties. Therefore, the conceptual metaphor *"Emotions are Liquids"* can be manifested through the use of linguistic metaphors such as *"pure love"*, *"stir excitement"* and *"contain your anger"*. The interaction between the target and the source concepts of the expression is important to fully comprehend its metaphoricity.

Over the last couple of years, there has been an increasing interest towards metaphor processing and its applications, either as part of natural language processing (NLP) tasks such as machine translation (Koglin and Cunha, 2019), text simplification (Wolska and Clausen, 2017; Clausen and Nastase, 2019) and sentiment analysis (Rentoumi et al., 2012) or in more general discourse analysis use cases such as in analysing political discourse (Charteris-Black, 2011), financial reporting (Ho and Cheng, 2016) and health communication (Semino et al., 2018).

Metaphor processing comprises several tasks including identification, interpretation and cross-domain mappings. Metaphor identification is the most studied among these tasks. It is concerned with detecting the metaphoric words or expressions in the input text and could be done either on the sentence, relation or word levels. The difference between these levels of processing is extensively studied in (Zayed et al., 2020). Identifying metaphors on the word-level could be treated as either *sequence labelling* by deciding the metaphoricity of each word in a sentence given the context or *single-word classification* by deciding the metaphoricity of a targeted word. On the other hand, relation-level identification looks at specific grammatical relations such as the *dobj* or *amod* dependencies and checks the metaphoricity of the verb or the adjective given its association with the noun. In

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 388–406
November 16 - 20, 2020.

relation-level identification, both the source and target domain words (the tenor and vehicle) are classified either as a metaphoric or literal expression, whereas in word-level identification only the source domain words (vehicle) are labelled. These levels of analysis (paradigms) are already established in literature and adopted by previous research in this area as will be explained in Section 2. The majority of existing approaches, as well as the available datasets, pertaining to metaphor processing focus on the metaphorical usage of verbs and adjectives either on the word or relation levels. This is because these syntactic types exhibit metaphoricity more frequently than others according to corpus-based analysis (Cameron, 2003; Shutova and Teufel, 2010).

Although the main focus of both the relation-level and word-level metaphor identification is discerning the metaphoricity of the vehicle (source domain words), the interaction between the metaphor components is less explicit in word-level analysis either when treating the task as sequence labelling or single-word classification. Relation-level analysis could be viewed as a deeper level analysis that captures information that is not captured on the word-level through modelling the influence of the tenor (e.g.noun) on the vehicle (e.g. verb/adjective). There will be reasons that some downstream tasks would prefer to have such information (i.e. explicitly marked relations), among these tasks are metaphor interpretation and cross-domain mappings. Moreover, employing the wider context around the expression is essential to improve the identification process.

This work focuses on relation-level metaphor identification represented by verb-noun and adjective-noun grammar relations. We propose a novel approach for context-based textual classification that utilises affine transformations. In order to integrate the interaction of the metaphor components in the identification process, we utilise affine transformation in a novel way to condition the neural network computation on the contextualised features of the given expression. The idea of affine transformations has been used in NLP-related tasks such as visual question-answering (de Vries et al., 2017), dependency parsing (Dozat and Manning, 2017), semantic role labelling (Cai et al., 2018), coreference resolution (Zhang et al., 2018), visual reasoning (Perez et al., 2018) and lexicon features integration (Margatina et al., 2019).

Inspired by the works on visual reasoning, we use the candidate expression of certain grammatical relations, represented by deep contextualised features, as an auxiliary input to modulate our computational model. Affine transformations can be utilised to process one source of information in the context of another. In our case, we want to integrate: 1) the deep contextualised-features of the candidate expression (represented by ELMo sentence embeddings) with 2) the syntactic/semantic features of a given sentence. Based on this task, affine transformations have a similar role to attention but with more parameters, which allows the model to better exploit context. Therefore, it could be regarded as a form of a more sophisticated attention. Whereas the current "straightforward" attention models are overly simplistic, our model prioritises the contextual information of the candidate to discern its metaphoricity in a given sentence.

Our proposed model consists of an affine transform coefficients generator that captures the meaning of the candidate to be classified, and a neural network that encodes the full text in which the candidate needs to be classified. We demonstrate that our model significantly outperforms the state-of-the-art approaches on existing relation-level benchmark datasets. The unique characteristics of tweets and the availability of Twitter data motivated us to identify metaphors in such content. Therefore, we evaluate our proposed model on a newly introduced dataset of tweets (Zayed et al., 2019) annotated for relation-level metaphors.

2 Related Work

Over the last decades, the focus of computational metaphor identification has shifted from rule-based (Fass, 1991) and knowledge-based approaches (Krishnakumaran and Zhu, 2007; Wilks et al., 2013) to statistical and machine learning approaches including supervised (Gedigian et al., 2006; Turney et al., 2011; Dunn, 2013a,b; Tsvetkov et al., 2013; Hovy et al., 2013; Mohler et al., 2013; Klebanov et al., 2014; Bracewell et al., 2014; Jang et al., 2015; Gargett and Barnden, 2015; Rai et al., 2016; Bulat et al., 2017; Köper and Schulte im Walde, 2017), semi-supervised (Birke and Sarkar, 2006; Shutova et al., 2010; Zayed et al., 2018) and unsupervised methods (Shutova and Sun, 2013; Heintz et al., 2013; Strzalkowski et al., 2013). These approaches employed a variety of features to design their models. With the advances in neu-

ral networks, the focus started to shift towards employing more sophisticated models to identify metaphors. This section focuses on current research that employs neural models for metaphor identification on both word and relation levels.

Word-Level Processing: Do Dinh and Gurevych (2016) were the first to utilise a neural architecture to identify metaphors. They approached the problem as sequence labelling where a traditional fully-connected feed-forward neural network is trained using pre-trained word embeddings. The authors highlighted the limitation of this approach when dealing with short and noisy conversational texts. As part of the NAACL 2018 Metaphor Shared Task (Leong et al., 2018), many researchers proposed neural models that mainly employ LSTMs (Hochreiter and Schmidhuber, 1997) with pre-trained word embeddings to identify metaphors on the word-level. The best performing systems are: THU NGN (Wu et al., 2018), OCOTA (Bizzoni and Ghanimifard, 2018) and bot.zen (Stemle and Onysko, 2018). Gao et al. (2018) were the first to employ the deep contextualised word representation ELMo (Peters et al., 2018), combined with pre-trained GloVe (Pennington et al., 2014) embeddings to train bidirectional LSTM-based models. The authors introduced a sequence labelling model and a single-word classification model for verbs. They showed that incorporating the context-dependent representation of ELMo with context-independent word embeddings improved metaphor identification. Mu et al. (2019) proposed a system that utilises a gradient boosting decision tree classifier. Document embeddings were employed in an attempt to exploit wider context to improve metaphor detection in addition to other word representations including GLoVe, ELMo and skip-thought (Kiros et al., 2015). Mao et al. (2018, 2019) explored the idea of selectional preferences violation (Wilks, 1978) in a neural architecture to identify metaphoric words. Mao's proposed approaches emphasised the importance of the context to identify metaphoricity by employing context-dependent and context-independent word embeddings. Mao et al. (2019) also proposed employing multi-head attention to compare the targeted word representation with its context. An interesting approach was introduced by Dankers et al. (2019) to model the interplay between metaphor identification and emotion regression. The authors introduced multiple multi-task learning techniques that employ hard and soft parameter sharing methods to optimise LSTM-based and BERT-based models.

Relation-Level Processing: Shutova et al. (2016) focused on identifying the metaphoricity of adjective/verb-noun pairs. This work employed multimodal embeddings of visual and linguistic features. Their model employs the cosine similarity of the candidate expression components based on word embeddings to classify metaphors using an optimised similarity threshold. Rei et al. (2017) introduced a supervised similarity network to detect adjective/verb-noun metaphoric expressions. Their system utilises word gating, vector representation mapping and a weighted similarity function. Pre-trained word embeddings and attribute-based embeddings (Bulat et al., 2017) were employed as features. This work explicitly models the interaction between the metaphor components. Gating is used to modify the vector of the verb/adjective based on the noun, however the surrounding context is ignored by feeding only the candidates as input to the neural model which might lead to loosing important contextual information.

Limitations: As discussed, the majority of previous works adopted the word-level paradigm to identify metaphors in text. The main distinction between the relation-level and the word-level paradigms is that the former makes the context more explicit than the latter through providing information about not only where the metaphor is in the sentence but also how its components come together through hinting at the relation between the tenor and the vehicle. Stowe and Palmer (2018) showed that the type of syntactic construction a verb occurs in influences its metaphoricity. On the other hand, existing relation-level approaches (Tsvetkov et al., 2014; Shutova et al., 2016; Bulat et al., 2017; Rei et al., 2017) ignore the context where the expression appears and only classify a given syntactic construction as metaphorical or literal. Studies showed that the context surrounding a targeted expression is important to discern its metaphoricity and fully grasp its meaning (Mao et al., 2018; Mu et al., 2019). Therefore, current relation-level approaches will only be able to capture commonly used conventionalised metaphors. In this work, we address these limitations by introducing a novel approach to textual classification which employs contextual information from both the targeted expression under study and the wider

context surrounding it.

3 Proposed Approach

Feature-wise transformation techniques such as feature-wise linear modulation (FiLM) have been recently employed in many applications showing improved performance. They became popular in image processing applications such as image style transfer (Dumoulin et al., 2017); then they found their way into multi-modal tasks, specifically visual question-answering (de Vries et al., 2017; Perez et al., 2018). They also have been shown to be effective approaches for relational problems as mentioned in Section 1. The idea behind FiLM is to condition the computation carried out by a neural model on the information extracted from an auxiliary input in order to capture the relationship between multiple sources of information (Dumoulin et al., 2018).

Our approach adopts Perez's (2018) formulation of FiLM on visual reasoning for metaphor identification. In visual reasoning, image-related questions are answered by conditioning the image-based neural network (visual pipeline) on the question context via a linguistic pipeline. In metaphor identification, we can consider that the image in our case is the sentence that has a metaphoric candidate and the auxiliary input is the linguistic interaction between the components of the candidate itself. This will allow us to condition the computation of a sequential neural model on the contextual information of the candidate and leverage the feature-wise interactions between the conditioning representation and the conditioned network. To the best of our knowledge, we are the first to propose such contextual modulation for textual classification in general and for metaphor identification specifically.

Our proposed architecture consists of *a contextual modulation* pipeline and *a metaphor identification linguistic* pipeline as shown in Figure 1. The input to the contextual modulator is the deep contextualised representation of the candidate expression under study (which we will refer to as targeted expression[1]) to capture the interaction between its components. The linguistic pipeline employs an LSTM encoder which produces a contextual representation of the provided sentence where the targeted expression appeared. The model is trained end-to-end to identify relation-level metaphoric expressions focusing on verb-noun and adjective-noun grammatical relations. Our model takes as input a sentence (or a tweet) and a targeted expression of a certain syntactic construction and identifies whether the candidate in question is used metaphorically or literally by going through the following steps:

Condition: In this step the targeted expression is used as the auxiliary input to produce a conditioning representation. We first embed each candidate of verb-direct object pairs[2] (v, n) using ELMo sentence embeddings to learn context-dependent aspects of word meanings c_{vn}. We used the 1,024-dimensional ELMo embeddings pre-trained on the One Billion Word benchmark corpus (Chelba et al., 2014). The sentence embeddings of the targeted expression are then prepared by implementing an embeddings layer that loads these pre-trained ELMo embeddings from the TensorFlow Hub[3]. The layer takes in the raw text of the targeted expression and outputs a fixed mean-pooled vector representation of the input as the contextualised representation. This representation is then used as an input to the main component of this step, namely a contextual modulator. The contextual modulator consists of a fully-connected feed-forward neural network (FFNN) that produces the conditioning parameters (i.e. the shifting and scaling coefficients) that will later modulate the linguistic pipeline computations. Given that c_{vn} is the conditioning input then the contextual modulator outputs γ and β, the context-dependent scaling and shifting vectors, as follows:

$$\begin{aligned} \gamma(c_{vn}) &= W_\gamma c_{vn} + b_\gamma, \\ \beta(c_{vn}) &= W_\beta c_{vn} + b_\beta \end{aligned} \tag{1}$$

where W_γ, W_β, b_γ, b_β are learnable parameters.

Embed: Given a labelled dataset of sentences, the model begins by embedding the tokenised sentence S of words $w_1, w_2, ..., w_n$, where n is the number of words in S, into vector representations using GloVe embeddings. We used the uncased 200-dimensional GloVe embeddings pre-trained on ∼2 billion tweets and contains 1.2 million words.

Encode: The next step is to train a neural network with the obtained embeddings. Since context is important for identifying metaphoricity, sentence

[1] Targeted expressions are already annotated in the dataset and initially obtained either manually or automatically using a dependency parser as will be described in Section 4.

[2] We do the same for subject-verb and adjective-noun pairs but, for simplicity, we demonstrate the process with verb-direct object pairs.

[3] https://www.tensorflow.org/hub

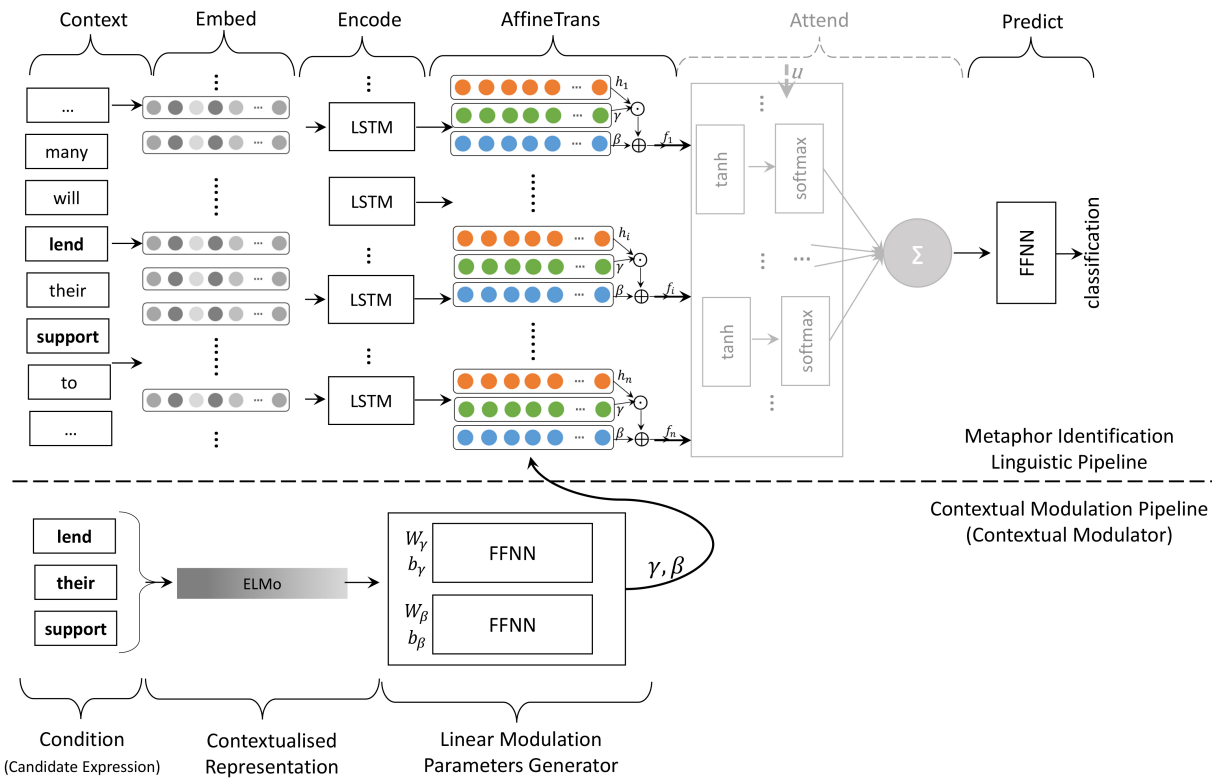

Figure 1: The proposed framework for relation-level metaphor identification showing the contextual modulation in detail. The attention process is greyed out as we experimented with and without it.

encoder is a sensible choice. We use an LSTM sequence model to obtain a contextual representation which summarises the syntactic and semantic features of the whole sentence. The output of the LSTM is a sequence of hidden states $h_1, h_2, ..., h_n$, where h_i is the hidden state at the i^{th} time-step.

Feature-wise Transformation: In this step, an affine transformation layer, hereafter *AffineTrans* layer, applies a feature-wise linear modulation to its inputs, which are: 1) the hidden states from the encoding step; 2) the scaling and shifting parameters from the conditioning step. By feature-wise, we mean that scaling and shifting are applied to each encoded vector for each word in the sentence.

$$f(h_i, c_{vn}) = \gamma(c_{vn}) \odot h_i + \beta(c_{vn}) \quad (2)$$

Attend: Recently, attention mechanisms have become useful to select the most important elements in a given representation while minimising information loss. In this work, we employ an attention layer based on the mechanism presented in (Lin et al., 2017). It takes the output from the *AffineTrans* layer as an input in addition to a randomly initialised weight matrix W, a bias vector b and a learnable context vector u to produce the attended output as follows:

$$e_i = tanh(W f_i + b) \quad (3)$$

$$\alpha_i = softmax(u e_i) \quad (4)$$

$$r = \sum_{i=1}^{n} \alpha_i f_i \quad (5)$$

Our model is trained and evaluated with and without the attention mechanism in order to differentiate between the effect of the feature modulation and the attention on the model performance.

Predict: The last step is to make the final prediction using the output from the previous step (attended output in case of using attention or the *AffineTrans* layer output in case of skipping it). We use a fully-connected feed-forward layer with a sigmoid activation that returns a single (binary) class label to identify whether the targeted expression is metaphoric or not.

4 Datasets

The choice of annotated dataset for training the model and evaluating its performance is determined by the level of metaphor identification. Given

the distinction between the levels of analysis, approaches addressing the task on the word-level are not fairly comparable to relation-level approaches since each task addresses metaphor identification differently. Therefore, the tradition of previous work in this area is to compare approaches addressing the task on the same level against each other on level-specific annotated benchmark datasets (Zayed et al., 2020).

Following prior work in this area and in order to compare the performance of our proposed approach with other relation-level metaphor identification approaches, we utilise available annotated datasets that support this level of processing. The existing datasets are either originally prepared to directly support relation-level processing such as the TSV (Tsvetkov et al., 2014) dataset and the Tweets dataset by Zayed et al. (2019) or adapted from other word-level benchmark datasets to suit relation-level processing such as the adaptation of the benchmark datasets TroFi (Birke and Sarkar, 2006) and VU Amsterdam metaphor corpus (VUAMC) (Steen et al., 2010) by Zayed et al. (2020) and the adaptation of the MOH (Mohammad et al., 2016) dataset by Shutova et al. (2016). Due to space limitation, we include in Appendix A: 1) examples of annotated instances from these datasets showing their format as: sentence, targeted expression and the provided label; 2) the statistics of these datasets including their size and percentage of metaphors.

Relation-Level Datasets: These datasets focus on expressions of certain grammatical relations. Obtaining these relations could be done either automatically by employing a dependency parser or manually by highlighting targeted expressions in a specific corpus. Then, these expressions are manually annotated for metaphoricity given the surrounding context. There exist two benchmark datasets of this kind, namely the **TSV** dataset and Zayed et al. (2019) **Tweets** dataset, hereafter **ZayTw** dataset. The former focuses on discerning the metaphoricity of adjective-noun expressions in sentences collected from the Web and Twitter while the latter focuses on verb-direct object expressions in tweets.

Adapted Word-Level Datasets: Annotated datasets that support word-level metaphor identification are not suitable to support relation-level processing due to the annotation difference (Shutova, 2015; Zayed et al., 2020). To overcome the limited availability of relation-level datasets, there has been a growing effort to enrich and extend benchmark datasets annotated on the word-level to suit relation-level metaphor identification. Although it is non-trivial and requires extra annotation effort, Shutova et al. (2016) and Zayed et al. (2020) introduced adapted versions of the MOH, TroFi and VUAMC datasets to train and evaluate models that identify metaphors on the relation-level. Since the **MOH** dataset was originally created to identify metaphoric verbs on the word-level, its adaptation by Shutova et al. (2016), also referred to as **MOH-X** in several papers, focused on extracting the verb-noun grammar relations using a dependency parser. The dataset is relatively small and contains short and simple sentences that are originally sampled from the example sentences of each verb in WordNet (Fellbaum, 1998). The **TroFi** dataset was designed to identify the metaphoricity of 50 selected verbs on the word-level from the 1987-1989 Wall Street Journal (WSJ) corpus. The **VUAMC** (Steen et al., 2010) is the largest corpus annotated for metaphors and has been employed extensively by models developed to identify metaphors on the word-level. However, models designed to support relation-level metaphor identification can not use it in its current state. Therefore, previous research focusing on relation-level processing (Rei et al., 2017; Bulat et al., 2017; Shutova et al., 2016; Tsvetkov et al., 2014) did not train, evaluate or compare their approaches using it. Recently, a subset of the VUAMC was adapted to suit relation-level analysis by focusing on the training and test splits provided by the NAACL metaphor shared task. This corpus subset as well as the TroFi dataset are adapted by Zayed et al. (2020) to suit identifying metaphoric expressions on the relation-level, focusing on verb-direct object grammar relations (i.e *dobj* dependencies). The Stanford dependency parser was utilised to extract these relations which were then filtered to ensure quality.

5 Experiments

5.1 Experimental Setup

We employ a single-layer LSTM model with 512 hidden units. The Adadelta algorithm (Zeiler, 2012) is used for optimisation during the training phase and the binary cross-entropy is used as a loss function to fine tune the network. The reported results are obtained using batch size of 256 instances for the ZayTw dataset and 128 instances for the

other employed datasets. L_2-regularisation weight of 0.01 is used to constraint the weights of the contextual modulator. In all experiments, we zero-pad the input sentences to the longest sentence length in the dataset. All the hyper-parameters were optimised on a randomly separated development set (validation set) by assessing the accuracy. We present here the best performing design choices based on experimental results but we highlight some other attempted considerations in Appendix B. We implemented our models using Keras (Chollet et al., 2015) with the TensorFlow backend. We are making the source code and best models publicly available[4]. To ensure reproducibility, we include the sizes of the training, validation and test sets in Appendix B as well as the best validation accuracy obtained on each validation set. All the results presented in this paper are obtained after running the experiments five times with different random seeds and taking the average.

In this work, we selected the following state-of-the-art models pertaining to relation-level metaphor identification for comparisons: the cross-lingual model by (Tsvetkov et al., 2014), the multimodal system of linguistic and visual features by (Shutova et al., 2016), the ATTR-EMBED model by Bulat et al. (2017) and the supervised similarity network (SSN) by Rei et al. (2017). We consider the SSN system as our baseline. For fair comparisons, we utilised their same data splits on the five employed benchmark datasets described in Section 4.

5.2 Excluding *AffineTrans*

We implemented a simple LSTM model to study the effect of employing affine transformations on the system performance. The input to this model is the tokenised sentence S which is embedded as a sequence of vector representations using GloVe. These sequences of word embeddings are then encoded using the LSTM layer to compute a contextual representation. Finally, this representation is fed to a feed-forward layer with a sigmoid activation to predict the class label. We used this model with and without the attention mechanism.

5.3 Results

We conduct several experiments to better understand our proposed model. First, we experiment with the simple model introduced in Section 5.2. Then, we train the proposed models on the benchmark datasets discussed in Section 4. We experiment with and without the attention layer to assess its effect on the model performance. Furthermore, we compare our model to the current work that addresses the task on the relation-level, in-line with our peers in this area. Tables 1 and 2 show our model performance in terms of precision, recall, F1-score and accuracy.

Since the source code of Rei's (2017) system is available online[5], we trained and tested their model using the ZayTw dataset as well as the adapted VUAMC and TroFi dataset in an attempt to study the ability of their model to generalise when applied on a corpus of a different text genre with wider metaphoric coverage including less common (conventionalised) metaphors.

6 Discussion

Overall performance. We analysed the model performance by inspecting the classified instances. We noticed that it did a good job identifying conventionalised metaphors as well as uncommon ones. Appendix A shows examples of classified instances by our system from the employed benchmark datasets. Our model achieves significantly better F1-score over the state-of-the-art SSN system (Rei et al., 2017) under the one-tailed paired *t-test* (Yeh, 2000) at p-value$<$0.01 on three of the five employed benchmark datasets. Moreover, our architecture showed improved performance over the state-of-the-art approaches on the TSV and MOH datasets. It is worth mentioning that the size of their test sets is relatively smaller; therefore any change in a single annotated instance drastically affects the results. Moreover, the approach proposed by Tsvetkov et al. (2014) relies on hand-coded lexical features which justifies its high F1-score.

The effect of contextual modulation. When excluding the *AffineTrans* layer and only using the simple LSTM model, we observe a significant performance drop that shows the effectiveness of leveraging linear modulation. This layer adaptively influences the output of the model by conditioning the identification process on the contextual information of the targeted expression itself which significantly improved the system performance, as observed from the results. Moreover, employing the contextualised representation of the targeted expression, through ELMo sentence embeddings,

[4] https://github.com/OmniaZayed/affineTrans_metaphor_identification

[5] https://github.com/marekrei/ssn

	ZayTw (test-set)				TSV (test-set)			
	Prec.	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**
Tsvetkov et al. (2014)	-	-	-	-	-	-	**0.85**	-
Shutova et al. (2016) (multimodal)	-	-	-	-	0.67	**0.96**	0.79	-
Bulat et al. (2017) (ATTR-EMBED)	-	-	-	-	0.85	0.71	0.77	-
Rei et al. (2017) (SSN)	0.543	**1.0**	0.704	0.543	**0.903**	0.738	0.811	0.829
Simple LSTM	0.625	0.758	0.685	0.621	0.690	0.58	0.630	0.66
Simple LSTM (+ Attend)	0.614	0.866	0.718	0.631	0.655	0.55	0.598	0.63
Our AffineTrans	**0.804**	0.769	**0.786***	**0.773**	0.869	0.80	**0.834**	**0.84**
Our AffineTrans (+ Attend)	0.758	**0.812**	0.784*	0.757	0.875	0.77	0.819	0.83

Table 1: Our proposed architecture performance compared to the state-of-the-art approaches on the benchmark datasets ZayTw and TSV. *Statistically significant (p-value$<$0.01) compared to the SSN system (Rei et al., 2017).

	adapted MOH (10-fold)				adapted TroFi (test-set)				adapted VUAMC (test-set)			
	Prec.	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**
Rei et al. (2017) (SSN)	0.736	0.761	0.742	0.748	0.620	0.892	0.732	0.628	0.475	0.532	0.502	0.558
Simple LSTM	0.757	0.773	0.759	0.759	0.70	0.751	0.725	0.674	0.510	0.339	0.407	0.587
Simple LSTM (+ Attend)	0.746	0.782	0.757	0.752	0.759	0.853	0.803*	0.761	0.575	0.423	0.487	0.627
Our AffineTrans	**0.804**	0.748	0.771	**0.780**	**0.852**	**0.909**	**0.879***	**0.858**	**0.712**	0.639	0.673*	**0.741**
Our AffineTrans (+ Attend)	0.753	**0.813**	**0.779**	0.773	0.841	0.870	0.856*	0.832	0.686	**0.679**	**0.683***	0.736

Table 2: Our proposed architecture performance compared to the state-of-the-art approaches on the adapted benchmark datasets MOH, TroFi and VUAMC. *Statistically significant (p-value$<$0.01) compared to the SSN system (Rei et al., 2017). We could not include Shutova et al. (2016) results on the MOH dataset since they used different test settings, thus their results will not be strictly comparable.

was essential to explicitly capture the interaction between the verb/adjective and its accompanying noun. Then, the *AffineTrans* layer was able to modulate the network based on this interaction.

The effect of attention. It is worth noting that the attention mechanism did not help much in our *AffineTrans* model because affine transformation itself could be seen as playing a similar role to attention, as discussed in Section 1. In attention mechanisms important elements are given higher weight based on weight scaling whereas in linear affine transformation scaling is done in addition to shifting which gives prior importance (probability) to particular features. We are planning to perform an in-depth comparison of using affine transformation verses attention in our future work.

Error analysis. An error analysis is performed to determine the model flaws by analysing the predicted classification. We examined the false positives and false negatives obtained by the best performing model, namely *AffineTrans* (without attention). Interestingly, the majority of false negatives are from the political tweets in ZayTw dataset. Table 3 lists some examples of misclassified instances in the TSV and ZayTw datasets. Some instances could be argued as being correctly classified by the model. For instance, *"spend capital"* could be seen as a metaphor in that the noun is an abstract concept referring to actual money. Examples of misclassified instances from the other employed datasets are presented in Appendix A. Interestingly, we noticed that the model was able to spot mistakenly annotated instances. Although the adapted VUMAC subset contains various expressions which should help the model perform better, we noticed annotation inconsistency in some of them. For example, the verb *"choose"* associated with the noun *"science"* is annotated once as metaphor and twice as literal in very similar contexts. This aligns well with the findings of Zayed et al. (2020) who questioned the annotation of around 5% of the instances in this subset mainly due to annotation inconsistency.

Analysis of some misclassified verbs. We noticed that sometimes the model got confused while identifying the metaphoricity of expressions where the verb is related to emotion and cognition such as: *"accept, believe, discuss, explain, experience, need, recognise,* and *want"*. Our model tends to classify them as not metaphors. We include different examples from the ZayTw dataset of the verbs *"experience"* and *"explain"* with different associated nouns along with their gold and predicted classifications in Appendix A. Our model's prediction seems reasonable given that the instances in the training set were labelled as not metaphors. It is

	ZayTw		TSV	
	Tweet	**Prob.**	**Sentence**	**Prob.**
False Negative	hard to **resist the feeling** that remain is further [...]	0.46	You have a **shiny goal** in mind that is distracting you with its awesomeness.	0.49
	@abpi uk: **need #euref final facts**? read why if [...]	0.08	The first hours of a **shaky ceasefire** are not "the best of times".	0.14
	#ivoted with a black pen. do not **trust pencils**. [...]	0.003	The French bourgeoisie has rushed into a **blind alley**.	0.00
False Positive	[...] this guy would **spend so much political capital** trying to erase the [...]	0.96	I could hear the **shrill voices** of his sisters as they dash about their store helping customers.	0.98
	#pencilgate to **justify vitriolic backlash** if #remain wins [...]	0.94	[...] flavoring used in cheese, meat and fish to give it a **smoky flavor** could in fact be toxic.	0.82
	@anubhuti921 @prasannas it **adds technology** to worst of old police state practices, [...]	0.76*	Usually an overly dry nose is a precursor to a **bloody nose**.	0.64

Table 3: Misclassified examples by our *AffineTrans* model (without attention) from ZayTw and TSV test sets. Sentences are truncated due to space limitations. *Our model was able to spot some mistakenly annotated instances.

not clear why the gold label for *"explain this mess"* is not a metaphor while it is metaphor for *"explain implications"*; similarly, the nouns *"insprirations"* and *"emotions"* with the verb *"experience"*.

7 Conclusions

In this paper, we introduced a novel architecture to identify metaphors by utilising feature-wise affine transformation and deep contextual modulation. Our approach employs a contextual modulation pipeline to capture the interaction between the metaphor components. This interaction is then used as an auxiliary input to modulate a metaphor identification linguistic pipeline. We showed that such modulation allowed the model to dynamically highlight the key contextual features to identify the metaphoricity of a given expression. We applied our approach to relation-level metaphor identification to classify expressions of certain syntactic constructions for metaphoricity as they occur in context. We significantly outperform the state-of-the-art approaches for this level of analysis on benchmark datasets. Our experiments also show that our contextual modulation-based model can generalise well to identify the metaphoricity of unseen instances in different text types including the noisy user-generated text of tweets. Our model was able to identify both conventionalised common metaphoric expressions as well as less common ones. To the best of our knowledge, this is the first attempt to computationally identify metaphors in tweets and the first approach to study the employment of feature-wise linear modulation on metaphor identification in general. The proposed methodology is generic and can be applied to a wide variety of text classification approaches including sentiment analysis or term extraction.

Acknowledgments

This work was supported by Science Foundation Ireland under grant number SFI/12/RC/2289_2 (Insight).
We would like to thank the anonymous reviewers of this paper for their helpful comments and feedback. Special thanks for the anonymous meta-reviewer for steering an effective and constructive discussion about this paper which we realised its results through the experienced, extensive and beneficial meta-review. Sincere thanks to Mennatullah Siam for the insightful discussions about the technical part of this paper.

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 329–336, Trento, Italy.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101, New Orleans, LA, USA.

David B. Bracewell, Marc T. Tomlinson, Michael Mohler, and Bryan Rink. 2014. A tiered approach

to the recognition of metaphor. *Computational Linguistics and Intelligent Text Processing*, 8403:403–414.

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*, EACL '17, pages 523–528, Valencia, Spain.

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 2753–2765, Santa Fe, NM, USA. Association for Computational Linguistics.

Lynne Cameron. 2003. *Metaphor in educational discourse*. Advances in Applied Linguistics. Continuum, London, UK.

Lynne Cameron and Graham Low. 1999. *Researching and Applying Metaphor*. Cambridge Applied Linguistics. Cambridge University Press, Cambridge, UK.

Jonathan Charteris-Black. 2011. Metaphor in Political Discourse. In *Politicians and Rhetoric: The Persuasive Power of Metaphor*, pages 28–51. Palgrave Macmillan UK, London.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *The 15th Annual Conference of the International Speech Communication Association*, INTERSPEECH '14, pages 2635–2639, Singapore.

François Chollet et al. 2015. Keras.

Yulia Clausen and Vivi Nastase. 2019. Metaphors in text simplification: To change or not to change, that is the question. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 423–434, Florence, Italy.

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 '19, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the 4th Workshop on Metaphor in NLP*, pages 28–33, San Diego, CA, USA.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. 2018. Feature-wise transformations. *Distill*. Https://distill.pub/2018/feature-wise-transformations.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Jonathan Dunn. 2013a. Evaluating the premises and results of four metaphor identification systems. In *Proceedings of the 14th International conference on Computational Linguistics and Intelligent Text Processing*, volume 7816 of *CICLing '13*, pages 471–486, Samos, Greece. Springer Berlin Heidelberg.

Jonathan Dunn. 2013b. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, GA, USA. Association for Computational Linguistics.

Laurel J End. 1986. Grounds for metaphor comprehension. *Knowledge and language*, pages 327–345.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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*, EMNLP '18, pages 1412–1424, Brussels, Belgium.

Andrew Gargett and John Barnden. 2015. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 21–30, Denver, CO, USA. Association for Computational Linguistics.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, ScaNaLU '06, pages 41–48, New York City, NY, USA.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphors with lda topic modeling. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 58–66, Atlanta, GA, USA.

Janet Ho and Winnie Cheng. 2016. Metaphors in financial analysis reports: How are emotions expressed? *English for Specific Purposes*, 43:37 – 48.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 52–56, Atlanta, GA, USA.

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*, SIGDIAL '15, pages 384–392, Prague, Czech Republic.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, NIPS '15, pages 3294–3302.

Beata Beigman Klebanov, Ben 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, Baltimore, MD, USA. Association for Computational Linguistics.

Arlene Koglin and Rossana Cunha. 2019. Investigating the post-editing effort associated with machine-translated metaphors: a process-driven analysis. *The Journal of Specialised Translation*, 31(01):38–59.

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*, SENSE '18, pages 24–30, Valencia, Spain.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, NY, USA.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago, USA.

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, New Orleans, LA, USA.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

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 Association for Computational Linguistics*, ACL '18, pages 1222–1231, Melbourne, Australia.

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*, ACL '19, pages 3888–3898, Florence, Italy.

Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. 2019. Attention-based conditioning methods for external knowledge integration. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3944–3951, Florence, Italy.

Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, *Sem '16, pages 23–33, Berlin, Germany.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 27–35, Atlanta, GA, USA.

Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Learning outside the box: Discourse-level features improve metaphor identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, Minneapolis, MN, USA.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI '18, New Orleans, LA, USA.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, NAACL-HLT '18, New Orleans, LA, USA.

Sunny Rai, Shampa Chakraverty, and Devendra K. Tayal. 2016. Supervised metaphor detection using conditional random fields. In *Proceedings of the 4th Workshop on Metaphor in NLP*, pages 18–27, San Diego, CA, USA.

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*, EMNLP '17, pages 1537–1546, Copenhagen, Denmark.

Vassiliki Rentoumi, George A. Vouros, Vangelis Karkaletsis, and Amalia Moser. 2012. Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective. *ACM Transactions on Speech and Language Processing*, 9(3):1–31.

Elena Semino, Zsofia Demjen, Andrew Hardie, Sheila Alison Payne, and Paul Edward Rayson. 2018. *Metaphor, Cancer and the End of Life: A Corpus-based Study*. Routledge, London, UK.

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

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, pages 160–170, San Diego, CA, USA.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 978–988, Atlanta, GA, USA.

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*, COLING '10, pages 1002–1010, Beijing, China.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC '10, pages 255–261, Malta.

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*. Converging evidence in language and communication research. John Benjamins Publishing Company.

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

Kevin Stowe and Martha Palmer. 2018. Leveraging syntactic constructions for metaphor identification. In *Proceedings of the Workshop on Figurative Language Processing*, pages 17–26, New Orleans, LA, USA.

Tomek Strzalkowski, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Samira Shaikh, Ting Liu, Boris Yamrom, Kit Cho, Umit Boz, Ignacio Cases, and Kyle Elliot. 2013. Robust extraction of metaphor from novel data. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 67–76, Atlanta, GA, USA.

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*, ACL '14, pages 248–258, Baltimore, MD, USA.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 45–51, Atlanta, GA, USA.

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 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Edinburgh, Scotland, UK.

Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron Courville. 2017. Modulating early visual processing by language. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, pages 6597–6607, Long Beach, California, USA.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 36–44, Atlanta, GA, USA.

Magdalena Wolska and Yulia Clausen. 2017. Simplifying metaphorical language for young readers: A corpus study on news text. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 313–318, Copenhagen, Denmark. Association for Computational Linguistics.

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, New Orleans, LA, USA.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics*, volume 2 of *COLING '00*, pages 947–953, Saarbruecken, Germany.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2018. Phrase-level metaphor identification using distributed representations of word meaning. In *Proceedings of the Workshop on Figurative Language Processing*, pages 81–90, New Orleans, LA, USA.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2019. Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In *Proceedings of the 2nd Conference on Language, Data and Knowledge*, LDK '19, Leipzig, Germany.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Adaptation of word-level benchmark datasets for relation-level metaphor identification. In *Proceedings of the Second Workshop on Figurative Language Processing*, Online.

Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Rui Zhang, Cícero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2 (short papers) of *ACL '18*, pages 102–107, Melbourne, Australia. Association for Computational Linguistics.

A Datasets Statistics and Analysis

A.1 Benchmark Datasets Statistics

Table 4 shows the statistics of the benchmark datasets employed in this work, namely the relation-level datasets TSV[6] and ZayTw in addition to the adapted TroFi[7], VUAMC[8] and MOH[9] datasets. Table 5 shows examples of annotated instances from each dataset.

A.2 Datasets Analysis

Examples of correctly classified instances from the employed datasets: We show examples of correctly classified instances by our best performing model. Table 6 comprises examples from the relational-level datasets TSV and ZayTw. Table 7 lists examples from the adapted MOH and TroFi datasets as well as the adapted VUAMC.

Examples of misclassified instances by our model in the tweets dataset: Examples of misclassified instances from the TSV and ZayTw datasets as well as the adapted MOH, TroFi and VUAMC datasets are given in Table 8. Our model spotted some instances that are mistakenly annotated in the original datasets.

Missclassified Verbs: Table 9 shows examples from the ZayTw dataset of the verbs *"experience"* and *"explain"* with different associated nouns along with their gold and predicted classifications.

B Design Considerations

B.1 Experimental Settings

The word embeddings layer is intialised with the pre-trained GloVe embeddings. We used the uncased 200-dimensional GloVe embeddings pre-trained on ∼2 billion tweets and contains 1.2 million words. We did not update the weights of these embeddings during training. Table 10 shows the sizes of the training, validation and test sets of each employed dataset for as well as the corresponding best obtained validation accuracy by the the *Affine-Trans* model (without attention). All experiments are done on a NVIDIA Quadro M2000M GPU and the average running time for the proposed models is around 1 hour for maximum of 100 epochs.

B.2 Other Trials

Sentence Embedding: We experimented with different representations other than GLoVe in order to embed the input sentence. We tried to employ the contextualised pre-trained embeddings ELMo and BERT either instead of the GloVe embeddings or as additional-features but no further improvements were observed on both validation and test sets over the best performance obtained. Furthermore, we experimented with different pre-trained GloVe embeddings including the uncased 300-dimensional pre-trained vectors on the Common Crawl dataset but we did notice any significant improvements.

Sentence Encoding: The choice of using the simple LSTM to encode the input was based on several experiments on the validation set. We tried bidirectional LSTM but observed no further improvement. This is due to the nature of the relation-level metaphor identification task itself as the tenor (e.g. noun) affects the metaphoricity of the vehicle (e.g. verb or adjective) so a single-direction processing was enough.

[6] https://github.com/ytsvetko/metaphor

[7] http://natlang.cs.sfu.ca/software/trofi.html

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

[9] http://saifmohammad.com/WebPages/metaphor.html

Dataset	Syntactic structure	Text type	Size	% Metaphors	Average Sentence Length
The adapted TroFi Dataset	verb-direct object	50 selected verbs (News)	1,535 sentences	59.15%	48.5
The adapted VUAMC (NAACL Shared Task subset)	verb-direct object	known-corpus (The BNC)	5,820 sentences	38.87%	63.5
The adapted MOH Dataset	verb-direct object; subject-verb	selected examples (WordNet)	647 sentences	48.8%	11
The TSV Dataset	adjective–noun	selected examples (Web/Tweets)	1,964 sentences	50%	43.5
The ZayTw Dataset	verb-direct object	Tweets (general and political topics)	2,531 tweets	54.8%	34.5

Table 4: Statistics of the employed benchmark datasets to train and evaluate our proposed models highlighting the used experimental setting and links to the data sources in the footnotes. The adapted versions are available upon request from their corresponding authors.

Dataset	Sentence	Targeted Expression	Gold Label
TSV	Chicago is a big city, with a lot of everything to offer.	big city	0
	It 's a foggy night and there are a lot of cars on the motorway.	foggy night	0
	Their initial icy glares had turned to restless agitation.	icy glares	1
	And he died with a sweet smile on his lip.	sweet smile	1
ZayTw	insanity. ok to abuse children by locking them in closet, dark room and damage their psyche, but corporal punishment not ok? twisted!	abuse children	0
	nothing to do with your lot mate #ukip ran hate nothing else and your bloody poster upset the majority of the country regardless in or out	upset the majority	0
	nothing breaks my heart more than seeing a person looking into the mirror with anger & disappointment, blaming themselves when someone left.	breaks my heart	1
	how quickly will the warring tories patch up their differences to preserve power? #euref	patch up their differences	1
The adapted TroFi	A Middle Eastern analyst says Lebanese usually drink coffee at such occasions; Palestinians drink tea.	drink coffee	0
	In addition, the eight-warhead missiles carry guidance systems allowing them to strike Soviet targets precisely.	strike Soviet targets	0
	He now says that specialty retailing fills the bill, but he made a number of profitable forays in the meantime.	fills the bill	1
	A survey of U.K. institutional fund managers found most expect London stocks to be flat after the fiscal 1989 budget is announced, as Chancellor of the Exchequer Nigel Lawson strikes a careful balance between cutting taxes and not overstimulating the economy.	strikes a careful balance	1
The adapted VUAMC (NAACL Shared Task)	Among the rich and famous who had come to the salon to have their hair cut, tinted and set, Paula recognised Dusty Springfield, the pop singer, her eyes big and sooty , her lips pearly pink, and was unable to suppress the thrill of excitement which ran through her.	recognised Dusty Springfield	0
	But until they get any money back, the Tysons find themselves in the position of the gambler who gambled all and lost .	get any money	0
	The Labour Party Conference: Policy review throws a spanner in the Whitehall machinery	throws a spanner	1
	Otherwise Congress would have to face the consequences of automatic across-the-board cuts under the Gramm-Rudman-Hollings budget deficit reduction law.	face the consequences	1
MOH-X	commit a random act of kindness.	commit a random act	0
	The smoke clouded above the houses.	smoke clouded	0
	His political ideas color his lectures.	ideas color	1
	flood the market with tennis shoes.	flood the market	1

Table 5: Examples of annotated instances from the employed relation-level datasets showing their format as: sentence, targeted expression and the provided label.

Model	ZayTw		TSV	
Classification	**Expression**	**Prob.**	**Expression**	**Prob.**
Metaphor	poisoning our democracy	0.999	rich history	0.999
	binding the country	0.942	rocky beginning	0.928
	see greater diversity	0.892	foggy brain	0.873
	patch up their differences	0.738	steep discounts	0.723
	seeking information	0.629	smooth operation	0.624
	retain eu protection	0.515	dumb luck	0.512
Not Metaphor	shake your baby	0.420	filthy garments	0.393
	enjoy a better climate	0.375	clear day	0.283
	improve our cultural relations	0.292	slimy slugs	0.188
	placate exiters	0.225	sour cherries	0.102
	betrayed the people	0.001	short walk	0.014
	washing my car	0.000	hot chocolate	0.000

Table 6: Examples of correctly classified instances by our *AffineTrans* model (without attention) from the ZayTw and TSV datasets showing the classification probability.

Model	adapted MOH		adapted TroFi		adapted VUAMC	
Classification	**Expression**	**Prob.**	**Expression**	**Prob.**	**Expression**	**Prob.**
Metaphor	absorbed the knowledge	0.987	grasped the concept	0.985	bury their reservations	0.999
	steamed the young man	0.899	strike fear	0.852	reinforce emotional reticence	0.871
	twist my words	0.770	ate the rule	0.781	possess few doubts	0.797
	color my judgment	0.701	planted a sign	0.700	suppress the thrill	0.647
	poses an interesting question	0.543	examined the legacy	0.599	considers the overall effect	0.568
	wears a smile	0.522	pumping money	0.529	made no attempt	0.517
Not Metaphor	shed a lot of tears	0.484	pumping power	0.427	send the tape	0.482
	abused the policeman	0.361	poured acid	0.314	asking pupils	0.389
	tack the notice	0.274	ride his donkey	0.268	removes her hat	0.276
	stagnate the waters	0.148	fixed the dish	0.144	enjoying the reflected glory	0.188
	paste the sign	0.002	lending the credit	0.069	predict the future	0.088
	heap the platter	0.000	destroy coral reefs	0.000	want anything	0.000

Table 7: Examples of correctly classified instances by our *AffineTrans* model (without attention) from the adapted MOH, TroFi and VUAMC datasets showing the classification probability.

	Dataset	Sentence	Prob.
False Negative	TroFi	Unself-consciously , the littlest cast member with the big voice steps into the audience in one number to open her wide cat-eyes and throat to **melt the heart** of one lucky patron each night.	0.295
		Lillian Vernon Corp., a mail-order company, said it is experiencing delays in **filling orders** at its new national distribution center in Virginia Beach,Va.	0.006
	VUAMC	It is a curiously paradoxical foundation uponupon which to **build a theory** of autonomy.	0.410
		It has turned up in Canberra with Japan to develop Asia Pacific Economic Co-operation (APEC) and a new 12-nation organisation which will **mimic the role** of the Organisation for Economic Co-operation and Development in Europe.	0.000
	MOH	When does the **court of law sit**?	0.499
		The **rooms communicated**.	0.000
	TSV	It was great to see a **warm reception** for it on twitter.	0.488
		An **honest meal** at a reasonable price is a rarity in Milan.	0.000
	ZayTw	#brexit? we **explain likely implications** for business insurances on topic of #eureferendum	0.2863
		@abpi uk: **need #euref final facts**? read why if you care about uk life sciences we're #strongerin.	0.0797
False Positive	TroFi	As the struggle enters its final weekend , any one of the top contenders could **grasp his way** to the top of the greasy pole.	0.998*
		Southeastern poultry producers fear **withering soybean supplies** will force up prices on other commodities.	0.507
	VUAMC	Or after we **followed the duff advice** of a legal journalist in a newspaper?	0.999*
		Aristotle said something very interesting in that extract from the Politics which I quoted earlier; he said that women have a deliberative faculty but that it **lacks full authority**.	0.525
	MOH	All our **planets condensed** out of the same material.	0.999
		He bowed before the King.	0.868
	TSV	Bags two and three will only have **straight edges** along the top and the bottom.	0.846
		Mountain climbers at **high altitudes** quickly acquire a tan from the sun.	0.986
	ZayTw	delayed flight in fueturventura due to french strikes **restricting access** across french airspace =/ hopefully get back in time to #voteleave	0.9589
		in manchester more young people are expected to **seek help** in the coming months and years #cypiapt #mentalhealth	0.7055*

Table 8: Misclassified examples by our *AffineTrans* model (without attention) from the TSV test set as well as the adapted MOH, TroFi and VUAMC test sets. *Our model was able to spot some mistakenly annotated instances in the dataset.

	Expression	tweet	Predicted	Prob.	Gold
experience	the inspiration	relive the show , re - listen to her messages, re - experience the inspiration, refuel your motivation	**0**	0.220	1
	your emotions	do not be afraid to experience your emotions; they are the path to your soul. trust yourself enough to feel what you feel.	0	0.355	0
	this shocking behaviour	a friend voted this morning & experienced this shocking behaviour. voting is everyone 's right. #voteremain	0	0.009	0
explain	likely implications	#brexit? we explain likely implications for business insurances on topic of #eureferendum	**0**	0.2866	1
	this mess	@b_hanbin28 ikr same here :D imagine hansol & shua trynna explain this mess to other members :D	0	0.109	0
	the rise	loss aversion partly explains the rise of trump and ukip	1	0.618	1

Table 9: Examples of classified instances of the verbs *"experience"* and *"explain"* in the ZayTw test set.

Dataset	**Train**	**Validation**	**Test**	**split %**	**Validation Accuracy**	**@epoch**
The adapted TroFi Dataset	1,074	150	312	70-10-20	0.914	40
The adapted VUAMC	3,535	885	1,398	-	0.748	20
The adapted MOH Dataset	582 per fold	-	65 per fold	10-fold cross-validation	-	-
The TSV Dataset	1,566	200	200	-	0.905	68
The ZayTw Dataset	1,661	360	510	70-10-20	0.808	29

Table 10: Experimental information of the five benchmark datasets including the best obtained validation accuracy by the *AffineTrans* model (without attention). We preserved the splits used in literature for the VUAMC and TSV datasets.