Using Modality Exclusivity and Embodiment Norms for Metaphor Detection

Anonymous ACL submission

Abstract

This paper reports a linguistic-enriched method for detecting token-level metaphors at the second shared task on Metaphor Detection. We participate in all four phases of competition with both datasets, i.e. Verbs and AllPOS of VUA and TOFEL. We use four categories of features, i.e. linguistic, collocational, word embeddings and cosine similarity for training and testing based on statistical classifiers. Our system obtains an F-score of 0.652 for the VUA Verbs track, which is 5% higher than the strong baselines. The experimental results across models and datasets have consistently proven the salient contribution of using modality exclusivity and modality shift information for predicting metaphoricity. It is note worthy that this competitive performance is achieved without the implementation of sophisticated deep learning models.

1 Introduction

Metaphors, usually known as conceptual metaphors, are one kind of figurative languages that use conceptual mapping to represent one thing (target domain) as another (source domain) (Lakoff and Johnson, 1980). Metaphors are prevalent in daily life, which play a significant role for people to interpret/understand complex concepts. On the other hand, as a popular linguistic device, metaphors encode versatile ontological information, which usually involve e.g. domain transfer (Ahrens et al., 2003; Ahrens, 2010), sentiment reverse (Steen et al., 2010) or modality shift (Winter, 2019) etc. Therefore, detecting the metaphors in texts are essential for capturing the authentic meaning of the texts, which can benefit many natural language processing applications such as machine translation, dialog systems and sentiment analysis (Tsvetkov et al., 2014). In this shared task, we aim to detect token-level metaphors from plain texts by focusing on content words (Verbs, Nouns, Adjectives and Adverbs) of two corpora: VUA[1] and TOFEL[2]. To better understand the intrinsic properties of metaphors and to provide an in-depth analysis to this phenomenon, we propose a linguistic-enriched model to deal with this task with the use of modality exclusivity and embodiment norms (see details in Section 3).

2 Related Work

Many approaches have been proposed for automatic detection of metaphors, using features of lexical information (Klebanov et al., 2014; Wilks et al., 2013), semantic classes (Klebanov et al., 2016), concreteness (Klebanov et al., 2015), word associations (Xiao et al., 2016), constructions and frames (Hong, 2016) etc. based on models of traditional classifiers (Rai et al., 2016), neural networks (Do Dinh and Gurevych, 2016) or sequential models (Bizzoni and Ghanimifard, 2018).

Despite of the above endeavors, metaphor detection remains a challenging task. The semantic and ontological differences between metaphorical and non-metaphorical expressions are often subtle and probably subjective, which may vary from person to person. To tackle such problems, some people resort to specific domain knowledge (Tsvetkov et al., 2014); some employ lexicons (Mohler et al., 2013; Dodge et al., 2015); some adopt supervised methods (Klebanov et al., 2014, 2015, 2016); and some use attention-based deep learning models to capture both local and contextual information (Wu et al., 2018). These methods show different strengths on detecting metaphors, yet with respective disadvantages, such as having generalization problems, or are lack of explanations to the results. In addition,

[1] http://www.vismet.org/metcor/documentation/home.html

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

the reported performances of metaphor detection so far (around 0.6 F-score) are still not promising, which needs further endeavours in all aspects.

In this work, we adopt supervised machine learning based on four categories of features which include linguistic norms, ngram-word and -pos collocations, word embeddings and cosine similarity between the target nodes and its neighboring words, as well as the strong baselines provided by the organizer of the shared task (Leong et al., 2018; Klebanov et al., 2014, 2015, 2016). Moreover, we use several statistical models and ensemble learning strategies during training and testing so as to testify the cross-model consistency of the improvement by the various features. Detailed methods are given in the following sections.

3 Feature Sets

This work uses four categories of features (16 subsets in all) to represent the nodes and contextual information at hierarchical levels, which include the lexical and syntactic-to-semantic information, sensory modality scales, embodiment ratings (of verbs only), as well as word vectors of the nodes and cosine similarity of node-neighbor pairs, as detailed below.

- **Linguistic Norms:** Two linguistic norms are used to construct four linguistic-enriched feature sets in jsonlines format[3]:
  - **ME** (modality exclusivity): 42 dimension of nodes information
  - **DM** (dominant modality): 2×5 dimension of node-neighbor pairs information
  - **EB** (embodiment): 2 dimension of nodes information
  - **EB-diff** (embodiment differences): 2×5 dimension of node-neighbor pairs information

  The ME and DM feature sets are constructed by using the Lancaster Sensorimotor norms collected by Lynott et al. (2019). The data include measures of sensorimotor strength for 39,707 English words across six perceptual modalities: touch, hearing, smell, taste, vision and interoception, and five action effectors: mouth/throat, hand/arm, foot/leg, head (excluding mouth/throat), torso[4]. As sensorimotor information plays a fundamental role in cognition, these norms provide a valuable knowledge representation to the conceptual categories of the nodes and neighboring words which serve as salient features for inferring metaphors.

  The EB and EB-diff feature sets are constructed by using the embodiment norms collected by Sidhu et al. (2014). Research examining semantic richness effects has shown that multiple dimensions of meaning are activated in the process of word recognition (Yap et al., 2011). This data applies the semantic richness approach to verb stimuli in order to investigate how verb meanings are represented. The norms collected ratings on that dimension for 687 English verbs. The relative embodiment ratings revealed that bodily experience was judged to be more important to the meanings of some verbs (e.g., dance, breathe) than to others (e.g., evaporate, expect), which suggest that relative embodiment is an important aspect of verb meaning which can be very useful indicators of meaning mismatch of verbs.

- **Collocations:** Three sets of collocational features are constructed to represent the lexical, syntactic, grammatical information of the nodes and their neighbors: **Trigram**, **FL**(Fivegram Lemma), **FPOS** (Fivegram POS tags). The two corpora are lemmatized using the nltk spacy WordNetLemmatizer[5] and POS tagged using the nltk averaged_perceptron_tagger[6] before constructing such features.

- **Word Embeddings:** For comparisons, we utilise distributional vector representation of word meaning to the nodes based on the distributional hypothesis (Firth, 1957). Three pre-trained word2Vec models are used: **GoogleNews.300d**, **GloVe.300d** and **Internal-W2V.300d** (pre-trained using the VUA and TOFEL corpora). GoogleNews[7] is pre-trained with Google News corpus with

[3]The feature sets can be accessed through the link: https://github.com/ClaraWan629/Feature-Sets-for-MD

[4]https://osf.io/7emr6/

[5]https://www.nltk.org/_modules/nltk/stem/wordnet.html

[6]https://www.kaggle.com/nltkdata/averaged-perceptron-tagger

[7]https://github.com/mmihaltz/word2vec-GoogleNews-vectors

a million 300-dimension word vector model. GloVe[8] is an unsupervised learning algorithm for obtaining vector representations for words. We use the 300d vectors pre-trained on Wikipedia 2014+Gigaword 5.

- **Cosine Similarity:** We also investigate the cosine similarity (CS) measures for computing word sense distances between the nodes and their neighboring lexical words, based on the hypothesis that words of distant meaning are more likely to be metaphors. Three different sets of CS features are constructed in this work by using the above three different word embedding models: **CS-Google**, **CS-GloVe**, **CS-Internal**.

These features constitute a rather comprehensive representation of the mismatch of the nodes and their neighbors in terms of senses, domains, modalities, agentivity and concreteness etc, which are highly indicative of metaphorical uses and are hence hypothesized as more distinctive features than the strong baselines in Leong et al. (2018).

In addition, we replicate the three strong baselines provided by the organizer for comparison purposes:

- **B1**: lemmatized unigrams (UL)
- **B2**: lemmatized unigrams, generalized WordNet semantic classes, and difference in concreteness ratings between verbs/adjectives and nouns (UL + WordNet + CCDB)
- **B3**: baseline 2 and unigrams, pos tag, topic, concreteness ratings between nodes and up and down words respectively (UL + WordNet + CCDB + U + P + T + CUp + CDown)

4 Classifiers and Experimental Setup

Three traditional classifiers are used for predicting the metaphoricity of the tokens, including Logistic Regression, Linear SVC and Random Forest Classifier. The Machine Learning experiments are run through utilities in the SciKit-Learn Laboratory (SKLL)[9].

For parameter tuning, we use grid search to find optimal parameters for the learners. In addition, we set up the following optimized parameters for the three classifiers:

- Logistic Regression (LR): *'class_weight':'balanced', 'max_iter':5000, 'tol':1*
- Linear SVC (LSVC): *'class_weight':'balanced', 'max_iter':50000, 'C':10*
- Random Forest Classifier (RFC): *'min_samples_split':8, 'max_features':'log2', 'oob_score':'True', 'random_state':10, 'class_weight':'balanced'*

5 Results and Discussions

5.1 Evaluation Results

In order to evaluate the discriminativeness of the various features for metaphor detection and their fitness to the three classifiers, we focus on the VUA Verbs phase and randomly select a development set (4380 tokens) from the training set in proportion to the Train/Test ratio of the four phases. Experiments are run using the three classifiers and the setup in Section 4.

The evaluation results on the individual features in terms of F1-score are summarized in Table 1 below:

Individual	Features	LR	LSVC	RFC
Baseline	B1^{T2}	**.632**	.621	.618
Linguistic	ME^{T1}	**.637**	.636	.632
	DM	.616	.620	.623
	EB	.547	.548	.544
	EB-diff	.322	.321	.302
Collocation	Trigram^{T4}	**.626**	.625	.612
	FL^{T5}	**.624**	.623	.621
	FPOS	.378	.369	.335
Word2Vec	GoogleNews	.605	.607	.603
	GloVe^{T3}	**.630**	.627	.633
	Internal	.569	.555	.568
CS	GoogleNews	.448	.451	.445
	GloVe	.403	.404	.410
	Internal	.436	.421	.402

Table 1: Evaluation Results on Individual Features. T1-5 are the top five features in terms of F1 score.

In Table 1, the top five features with the LR classifier are highlighted in bold. Results show that the best individual feature is ME, followed by B1, W2V.GloVe, Trigram and FL. The performances

[8] https://nlp.stanford.edu/projects/glove/

[9] https://skll.readthedocs.io/en/latest/index.html

of the three classifiers are quite close for all features, with LR performing slightly better. To test the combined power of these features for metaphor detection, we also conduct evaluation on fused features, as shown in Table 2 below:

Fused Features	LR	LSVC	RFC
B2	.641	.636	.635
B3	.631	.630	.628
Top3	.653	.649	.650
Top4	.668	.666	.659
Top5	.669	.665	.668
Linguistic+B2	.655	.654	.652
Collocation+B2	.659	.658	.655
Word2Vec+B2	.637	.636	.637
CS+B2	.639	.637	.636
Selected	.672	.670	.671

Table 2: Evaluation Results on Fused Features

Results in Table 2 show that B2 is a stronger baseline than B3, so we use B2 as the comparison basis. Among the four categories of features, the linguistic and collocational features in combination with B2 achieve the greatest improvement by around 1.5% F1-score. The top 3-5 features also improve the performance by 1-2% F1-score. However, the word embeddings and cosine similarity features show no improvement over baseline 2. Finally, we use selected 12 features (excluding the W2V features) by automatic feature selection algorithm and have achieved the best results for evaluation (.672 F1 for LR).

5.2 Results of Test Sets

We use the best feature sets and classifier (LR) in the above evaluation for the final submission. The released results of our system on the test sets of the four phases in terms of F1-score are summarized in Table 3 below:

Phase/Method	B2	Top5	L+B2	Selected
VUA-Verbs	.600	.645	.642	**.652**
VUA-AllPOS	.589	.597	.591	**.603**
TOFEL-Verbs	.555	.588	.581	**.596**
TOFEL-AllPOS	.543	.550	.552	**.560**

Table 3: Released Final Results of Our System

In Table 3, 'L+B2' stands for 'Linguistic feature fused with baseline 2' and the best results are highlighted in bold. In addition to the best methods, we also submit the Top5 features and the 'L+B2' features which all show consistent improvement (1-5% F1) over baseline 2. The evaluation results and the final released results have both proven the effectiveness of using the linguistic features, esp. the Modality Exclusivity representations for metaphor detection.

6 Conclusion

We presented a linguistically enhanced method for token level metaphor detection using 16 features sets of four categories based on simple classifiers. As suggested by the results, the modality exclusivity and embodiment norms provide conceptual and bodily information for representing the nodes and the context, which help improve the performance of metaphor detection over the three strong baselines at a great extent. We consider to apply this series of feature sets to the state-of-the-art deep learning models to further testify the effectiveness of this method for metaphor detection in future.

Acknowledgments

This work is partially supported by the GRF grant (PolyU 156086/18H) and the Post-doctoral project (no. 4-ZZKE) at the Hong Kong Polytechnic University.

References

Kathleen Ahrens. 2010. Mapping principles for conceptual metaphors. *Researching and applying metaphor in the real world*, 26:185.

Kathleen Ahrens, Siaw Fong Chung, and Chu-Ren Huang. 2003. Conceptual metaphors: Ontology-based representation and corpora driven mapping principles. In *Proceedings of the ACL 2003 workshop on Lexicon and figurative language-Volume 14*, pages 36–42. Association for Computational Linguistics.

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.

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

Ellen K Dodge, Jisup Hong, and Elise Stickles. 2015. Metanet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49.

Raymond Firth. 1957. 2. a note on descent groups in polynesia. *Man*, 57:4–8.

Jisup Hong. 2016. Automatic metaphor detection using constructions and frames. *Constructions and frames*, 8(2):295–322.

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.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20.

Beata Beigman Klebanov, Chee Wee Leong, E Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106.

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

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

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

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

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

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

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796.

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

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 First Workshop on Metaphor in NLP*, pages 36–44.

Bodo Winter. 2019. Synaesthetic metaphors are neither synaesthetic nor metaphorical. *Perception metaphors*, pages 105–126.

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.

Ping Xiao, Khalid Alnajjar, Mark Granroth-Wilding, Kat Agres, and Hannu Toivonen. 2016. Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations. In *Proceedings of the 7th International Conference on Computational Creativity (ICCC). Paris, France*.

Melvin J Yap, Sarah E Tan, Penny M Pexman, and Ian S Hargreaves. 2011. Is more always better? effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, 18(4):742–750.