Leveraging Three Types of Embeddings from Masked Language Models in Idiom Token Classification

Ryosuke Takahashi* **Ryohei Sasano** **Koichi Takeda**
Graduate School of Informatics, Nagoya University
ryosuke.takahashi.cs@gmail.com,
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

Many linguistic expressions have idiomatic and literal interpretations, and the automatic distinction of these two interpretations has been studied for decades. Recent research has shown that contextualized word embeddings derived from masked language models (MLMs) can give promising results for idiom token classification. This indicates that contextualized word embedding alone contains information about whether the word is being used in a literal sense or not. However, we believe that more types of information can be derived from MLMs and that leveraging such information can improve idiom token classification. In this paper, we leverage three types of embeddings from MLMs; uncontextualized token embeddings and masked token embeddings in addition to the standard contextualized word embeddings and show that the newly added embeddings significantly improve idiom token classification for both English and Japanese datasets.

1 Introduction

Potentially idiomatic phrases are often used both in the idiomatic and literal sense. For example, "blew whistle" in (1) is used in the literal sense, whereas that in (2) is used in the idiomatic sense, that is, the meaning of the phrase has shifted and in this case it means *accuse*. Deciding whether each occurrence of a potentially idiomatic phrase is a literal or idiomatic usage is an essential process for text understanding. We call this processing *idiom token classification* following Salton et al. (2016).

(1) The referee blew the whistle to end the match.

(2) I blew the whistle on government corruption.

Recently, contextualized word embeddings have been shown to be useful for word sense disambiguation (Hadiwinoto et al., 2019). Furthermore, Shwartz and Dagan (2019) showed that the contextualized embeddings including BERT (Devlin et al., 2019) are useful for recognizing meaning shift of words in idioms. However, they only used contextualized embeddings, even though comparing them with the standard embeddings of the target word can be beneficial for precise detection of meaning shifts. Thus, in this paper, we propose a method to improved a BERT-based idiom token classifier by leveraging uncontextualized word embeddings.

Specifically, we use the token embedding of BERT, which is the uncontextualized embedding that is input to BERT and the same vector as is used for the prediction in the task of masked language model. Our assumption can be explained using (1) and (2) as follows: since "whistle" in (2) is used as a part of an idiomatic phrase, its contextualized embedding differs more from the uncontextualized embedding of "whistle" than in the case of (1).

Furthermore, we also leverage the masked token embedding of the target word in BERT, which is generated when the target phrase constituents are masked. This embedding can be considered to represent the meaning inferred from its context, and we assume that if the target phrase is used in the literal sense, as in (1), the output embedding will not significantly differ from the original embedding and thus the differences between the BERT embeddings without masking and those with masking are expected to be small.

2 Task and Baseline

2.1 Datasets and Settings

We focus on the idiom token classification of phrases consisting of verb-noun pairs in English and Japanese. As the English dataset, we use the VNC-Tokens dataset[1] (Cook et al., 2008). This dataset consists of 2,984 sentences containing 53 different potentially idiomatic verb-noun pairs in

*Ryosuke Takahashi is currently at SB Technology Corp.

[1] https://people.eng.unimelb.edu.au/paulcook/English_VNC_Cook.zip

Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pages 234 - 239

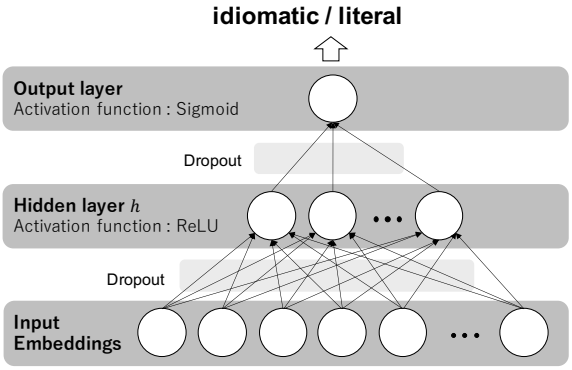

Figure 1: The *Embed-Encode-Predict* model.

English, where each sentence is labeled with "I" (idiomatic), "L" (literal), or "Q" (unknown). We use 28 out of the 53 idioms that have similar numbers of idiomatic and literal occurrences and only those sentences labeled as "I" or "L" following Salton et al. (2016).

As the Japanese dataset, we use the OpenMWE Corpus[2] (Hashimoto and Kawahara, 2008). This dataset consists of 102,846 sentences containing 146 different potentially idiomatic verb-noun pairs in Japanese, where each sentence is labeled with "I" (idiomatic) or "L" (literal). We use 90 out of the 146 idioms for which more than 50 examples for both idiomatic and literal usages are available following Hashimoto and Kawahara (2008).

In this study, we adopt the zero-shot setting because we are interested in detecting meaning shifts of words that are not included in the training data. Specifically, we employ the one-versus-rest scheme with the fully zero-shot setting. That is, we build a classifier for each phrase, which is trained on the phrases that contain neither the verb nor the noun that makes up the target phrase. For example, when building a classifier for *blew whistle*, we exclude phrases whose verb is *blew* or whose noun is *whistle* from the training data. We take one fifth of each training dataset as development data.

2.2 Baseline Systems

As the baseline system, we adopted a minimal *Embed-Encode-Predict* model (Shwartz and Dagan, 2019) that uses only contextualized embeddings of the constituent words of the target phrase as input. The reason for adopting a relatively simple model as a baseline is that the purpose of this study is to confirm the effectiveness of the newly added embeddings. Figure 1 shows the outline of the model, which consists of an input layer, a hidden layer, and an output layer. The output layer predicts whether the input phrase is idiomatic or literal. The size of the hidden layer is half of the input embedding size in all models in the paper. We applied dropout on the input embeddings and hidden layer. The dropout rates are both 50%.

Models	English	Japanese
Majority Baseline	0.672	0.629
Salton et al. (2016)	0.780	-
Hashimoto and Kawahara (2008)	-	0.740
BERT[$\boldsymbol{v}_{\mathrm{V}}$]	0.829	0.816
BERT[$\boldsymbol{v}_{\mathrm{N}}$]	0.836	0.821
BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$]	**0.840**	**0.823**

Table 1: Macro-averaged accuracy for baseline systems.

As the input, we used $[\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}]$, a concatenation of the contextualized embeddings of the verb and noun that comprise the target phrase. We used the pre-trained models `BERT-Base, Uncased`[3] for English and `BERT-Base, WWE`[4] for Japanese. Both models have 12 layers and 768 hidden dimensions per token. Japanese sentences were tokenized by Juman++[5] in advance. We used the development data to determine the number of training epochs and to determine which BERT hidden layer to use as the input embeddings of the *Embed-Encode-Predict* model. We refer to this model as BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$]. In addition, we developed models that only leverages one of the contextualized embeddings $\boldsymbol{v}_{\mathrm{V}}$ and $\boldsymbol{v}_{\mathrm{N}}$ to confirm the importance of each embedding. We refer to them as BERT[$\boldsymbol{v}_{\mathrm{V}}$] and BERT[$\boldsymbol{v}_{\mathrm{N}}$], respectively.

For reference, we also implemented support vector machine (SVM) based models with the features used in previous work. For English, we employed Salton et al. (2016)'s model that leveraged Skip-Thought Vectors (Kiros et al., 2015) as features. For Japanese, we implemented the features used by Hashimoto and Kawahara (2008), consisting of POS, lemma, token n-gram, hypernym, domain, voice, negativity, modality, adjacency, and adnominal information.

Table 1 lists the macro-averaged accuracy for each baseline model with the accuracy of the majority baseline. Each accuracy is the average of 5 runs with different random seeds. For both English

[2] http://openmwe.sourceforge.jp/Idiom/corpus/OpenMWE-Corpus-0.02.tar.bz2

[3] https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

[4] http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/JapaneseBertPretrainedModel/Japanese_L-12_H-768_A-12_E-30_BPE_WWM.zip

[5] https://github.com/ku-nlp/jumanpp

and Japanese dataset, BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$] achieved the highest accuracy, which demonstrates that BERT embeddings are useful for idiom token classification even in a zero-shot setting and supposedly capture the general characteristic of idiomaticity. We measured the statistical significance between BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$] and the other models with an approximate randomization test (Chinchor, 1992) with 99,999 iterations and significance level $\alpha = 0.05$ after Bonferroni correction. We found significant differences against the Majority Baseline and Salton et al. (2016) with respect to English and against Majority Baseline and Hashimoto and Kawahara (2008) with respect to Japanese.

3 Leveraging Additional Embeddings

The relatively high performance of BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$] in a zero-shot setting indicates that the standard BERT embeddings contain information about how much the meaning differs from the standard meaning of the words that comprise the phrase. However, the performance of idiom token classification can be improved by explicitly incorporating the standard meaning of the constituent words and the meaning inferred from its context.

3.1 Additional embeddings

We add two types of embeddings to BERT[$\boldsymbol{v}_{\mathrm{V}}$;$\boldsymbol{v}_{\mathrm{N}}$]: uncontextualized token embeddings and masked token embeddings of the phrase constituents.

Uncontextualized token embeddings We use the token embedding of BERT, which is the uncontextualized embedding that is input to BERT and the same vector as is used for the prediction in the task of masked language model in BERT. This embedding can be considered to represent the standard meaning of the word and thus if the target phrase is used in the literal sense, the BERT embeddings, which are contextualized, should be similar to the token embeddings. We refer to the uncontextualized token embeddings of a verb and a noun as $\boldsymbol{v}_{\mathrm{V}\_t}$ and $\boldsymbol{v}_{\mathrm{N}\_t}$, respectively.

Masked token embeddings We use the hidden layer of BERT when the target token is replaced with a special token `[MASK]`. This embedding can be considered to represent the meaning inferred from its context. If the target phrase is used in the literal sense, the differences between the BERT embeddings without masking and those with masking are expected to be small. We refer to the masked token embeddings of a verb and a noun as $\boldsymbol{v}_{\mathrm{V}\_m}$ and $\boldsymbol{v}_{\mathrm{N}\_m}$, respectively.

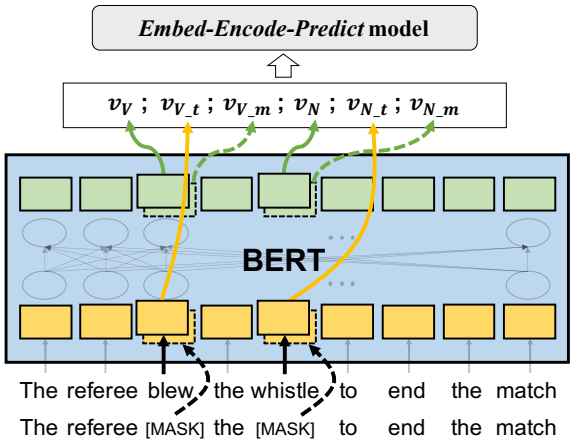

Figure 2: Overview of the proposed model.

Embeddings	English	Japanese
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$	0.840	0.823
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_t}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_t}$	0.859	0.842
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_m}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_m}$	0.852	0.829
$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_t}; \boldsymbol{v}_{\mathrm{V}\_m}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_t}; \boldsymbol{v}_{\mathrm{N}\_m}$	**0.865**	**0.847**

Table 2: Macro-averaged accuracy for different combinations of input embeddings.

Figure 2 shows the overview of the proposed model. When a sentence containing the target phrase is given, a *masked sentence*, in which the verb and noun that comprise the phrase are masked, is generated and input to the BERT in addition to the original sentence. Then, $\boldsymbol{v}_{\mathrm{V}}$, $\boldsymbol{v}_{\mathrm{V}\_t}$, $\boldsymbol{v}_{\mathrm{V}\_m}$, $\boldsymbol{v}_{\mathrm{N}}$, $\boldsymbol{v}_{\mathrm{N}\_t}$, and $\boldsymbol{v}_{\mathrm{N}\_m}$ are extracted and their concatenation is input to the *Embed-Encode-Predict* model.

3.2 Experiments and analysis

We performed the idiom token classification experiments with the additional embeddings. Table 2 lists the macro-averaged accuracy for different combinations of input embeddings. We can confirm that leveraging uncontextualized token embeddings and masked token embeddings in addition to the standard BERT embeddings is beneficial for idiom token classification. The statistical significance test shows that the difference between the accuracy of BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_t}; \boldsymbol{v}_{\mathrm{V}\_m}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_t}; \boldsymbol{v}_{\mathrm{N}\_m}$] and that of BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{N}}$] are significant for both English and Japanese datasets. The accuracy of BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_t}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_t}$] was slightly better than that of BERT[$\boldsymbol{v}_{\mathrm{V}}; \boldsymbol{v}_{\mathrm{V}\_m}; \boldsymbol{v}_{\mathrm{N}}; \boldsymbol{v}_{\mathrm{N}\_m}$]. We can say that the difference between the standard BERT embeddings and the uncontextualized token embed-

Usage	English		Japanese	
	$\boldsymbol{v}$ vs. $\boldsymbol{v}_t$	$\boldsymbol{v}$ vs. $\boldsymbol{v}_m$	$\boldsymbol{v}$ vs. $\boldsymbol{v}_t$	$\boldsymbol{v}$ vs. $\boldsymbol{v}_m$
Literal	**0.157**	**0.593**	**0.197**	**0.545**
Idiomatic	0.122	0.517	0.166	0.428

Table 3: Means of the cosine similarities of standard BERT embeddings ($\boldsymbol{v}$) against uncontextualized token embeddings ($\boldsymbol{v}_t$) and masked token embeddings ($\boldsymbol{v}_m$) for literal and idiomatic cases, respectively.

dings should be a good indicator of idiomaticity.

We assumed that when the target phrase is used in the literal sense, the uncontextualized token embeddings and the masked token embeddings tend to be similar to the standard BERT embeddings. To verify this assumption, we calculated the means of their cosine similarities for the literal and idiomatic cases, respectively. Table 3 lists the means of the cosine similarities. For English dataset, the mean of the cosine similarities between the uncontextualized token embeddings and standard BERT embeddings for the literal cases was 0.157, which was larger than that for the idiomatic cases, 0.122. Similarly, the mean of the cosine similarities between the masked token embeddings and standard BERT embeddings for the literal cases was 0.593, which was larger than that for the idiomatic cases, 0.517. The same trend can be observed for the Japanese dataset. It has been confirmed that all the differences are statistically significant. These results support our assumption.

4 Related Work

Several researchers have tackled the task of idiom token classification. Hashimoto and Kawahara (2008) is one of the earliest works. They created a Japanese annotated data for idiom token classification and proposed an SVM-based model with a set of features that commonly used for WSD. Fazly et al. (2009) proposed statistical measures that quantify the degree of lexical, syntactic, and overall fixedness of a verb noun combination. Sporleder and Li (2009) proposed a model for unsupervised idiom token classification based on the observation that literally used expressions typically exhibit cohesive ties with the surrounding discourse, while idiomatic expressions do not.

Li and Sporleder (2010) explored various features, such as global lexical context, discourse cohesion, syntactic structure, and local lexical features. They reported that global lexical context and discourse cohesion were most effective for idiom token classification. Peng et al. (2014) treated idiom token identification as a problem of outlier detection. They extracted topics from paragraphs containing idioms and from paragraphs containing literals by using Latent Dirichlet Allocation (LDA).

A broad range of neural network-based models have been proposed in recent years. Gharbieh et al. (2016) obtained phrase representations by averaging skip-gram (Mikolov et al., 2013) vectors of words that appear around the target phrase and applied them to idiom token classification. Salton et al. (2016) constructed an SVM-based classifier using the distributed representation of sentences generated by the Skip-Thought model (Kiros et al., 2015). King and Cook (2018) improved the performance of word embedding-based methods by incorporating syntactic and lexical patterns of idiomatic expressions.

More recently, methods using contextualized word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have been proposed. Shwartz and Dagan (2019) showed that the contextualized embeddings of constituent words were useful for recognizing meaning shifts of phrases. Hashempour and Villavicencio (2020) and Kurfalı and Östling (2020) worked on the idiom token classification task using BERT embeddings and reported that the BERT-based model achieved high accuracy in a phrase-specific setting. Garcia et al. (2021) proposed probing measures to examine how accurately idiomaticity in noun compounds is captured in vector space models and concluded that idiomaticity is not yet accurately represented by contextualized word embeddings.

Studies that used multiple types of embeddings in BERT, similar to our method, include the work by Zhang et al. (2020) and Yamada et al. (2021). Zhang et al. used the weighted sum of the input embedding and the mask embedding for spelling error correction whereas Yamada et al. used the weighted sum of the input embedding and the mask embedding for semantic frame induction.

5 Conclusion

We demonstrate that leveraging uncontextualized token embeddings and masked token embeddings in addition to the standard contextualized word embeddings significantly improve idiom token classification in a zero-shot setting. We also show that the results of investigating the similarities of these embeddings for each of the literal and idiomatic cases support our assumption that the uncontextu-

alized token embeddings and the masked token embeddings tend to be similar to the standard BERT embeddings when the target phrase is used in the literal meaning. One of the advantages of the proposed method is that it does not require training a new model because it extracts and uses embeddings with different properties from the same language model. We believe that the three types of embedding introduced in this study can be applied to other natural language tasks.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 21K12012.

References

Nancy Chinchor. 1992. The statistical significance of the MUC-4 results. In *Proceedings of the 4th Message Understanding Conference (MUC)*, pages 30–50.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop on Towards a Shared Task for Multiword Expressions (MWE)*, pages 19–22.

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 (NAACL-HLT)*, pages 4171–4186.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3551–3564.

Waseem Gharbieh, Virendra Bhavsar, and Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions (MWE)*, pages 112–118.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. 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 5297–5306.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon (CogALex)*, pages 72–80.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 992–1001.

Milton King and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 345–350.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3294–3302.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE)*, pages 85–94.

Linlin Li and Caroline Sporleder. 2010. Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 683–691.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027.

Matthew 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)*, pages 2227–2237.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 194–204.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 754–762.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021. Semantic frame induction using masked word embeddings and two-step clustering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 811–816.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 882–890.