STANCY: Stance Classification Based on Consistency Cues

Kashyap Popat[1], **Subhabrata Mukherjee**[2], **Andrew Yates**[1], **Gerhard Weikum**[1]
[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Microsoft Research, Redmond, WA, USA
[1]{kpopat,ayates,weikum}@mpi-inf.mpg.de
[2]subhabrata.mukherjee@microsoft.com

Abstract

Controversial claims are abundant in online media and discussion forums. A better understanding of such claims requires analyzing them from different perspectives. Stance classification is a necessary step for inferring these perspectives in terms of supporting or opposing the claim. In this work, we present a neural network model for stance classification leveraging BERT representations and augmenting them with a novel consistency constraint. Experiments on the *Perspectrum* dataset, consisting of claims and users' perspectives from various debate websites, demonstrate the effectiveness of our approach over state-of-the-art baselines.

1 Introduction

There is an abundance of contentious claims on the Web including controversial statements from politicians, biased news reports, rumors, etc. People express their perspectives about these controversial claims through various channels like editorials, blog posts, social media, and discussion forums. To achieve a deeper understanding of these claims, we need to understand users' perspectives and stance towards the claims. Recent research (FNC-1, 2016; Baly et al., 2018; Chen et al., 2019) has shown stance classification to be a critical step for information credibility and automated fact-checking.

Prior Work and Limitations: Prior approaches for stance classification proposed in Somasundaran and Wiebe (2010); Anand et al. (2011); Walker et al. (2012); Hasan and Ng (2013, 2014); Sridhar et al. (2015); Sun et al. (2018) rely on various linguistic features, e.g., n-grams, dependency parse tree, opinion lexicons, and sentiment to determine the stance of perspectives regarding controversial topics. Ferreira and Vlachos (2016) further incorporate natural language claims and propose a logistic regression model using the lexical and semantic features of claims and perspectives. SemEval tasks (Mohammad et al., 2016; Kochkina et al., 2017) and other approaches (Chen and Ku, 2016; Lukasik et al., 2016; Sobhani et al., 2017) have focused on determining stance only in Tweets.

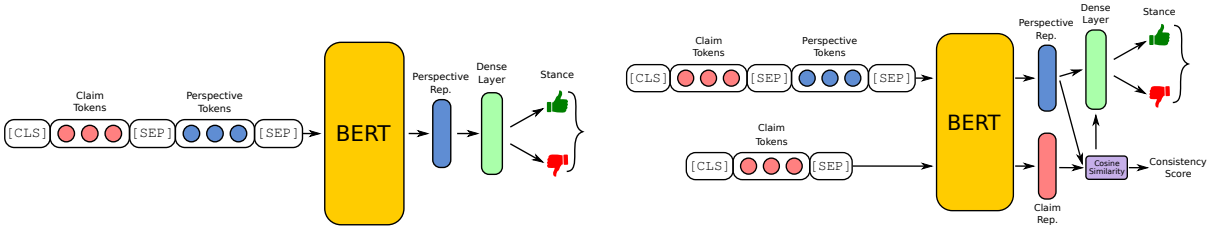
Bar-Haim et al. (2017) propose classifiers based on hand-crafted lexicons to identify important phrases in perspectives and their consistency with the claim to predict the stance However, their model critically relies on manual lexicons and assumes that the important phrases in claims are already identified.

Neural-network-based approaches for stance classification learn the claim and perspective representations separately and later combine them with conditional LSTM encoding (Augenstein et al., 2016), attention mechanisms (Du et al., 2017) or memory networks (Mohtarami et al., 2018). Some neural network models also incorporate lexical features (Riedel et al., 2017; Hanselowski et al., 2018). None of these approaches leverage knowledge acquired from massive external corpora.

Approach and Contributions: To overcome the limitations of prior works, we present STANCY, a neural network model for stance classification. Given an input pair of a claim and a user's perspective, our model predicts whether the perspective is *supporting* or *opposing* the claim. For example, the claim *"You have nothing to worry about surveillance, if you have done nothing wrong"* is supported by the user perspective *"Information gathered through surveillance could be used to fight terrorism"* and opposed by another user perspective *"With surveillance, the user privacy will go away!"*.

Our model for stance classification leverages representations from the BERT (Bidirectional

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6414–6419, Hong Kong, China, November 3–7, 2019.



(a) $\text{BERT}_{\text{BASE}}$: Fine-tuning BERT for stance classification.

(b) $\text{BERT}_{\text{CONS}}$: Enhancing BERT using the joint loss ($loss_{ce}$ for stance classification and $loss_{cos}$ for consistency).

Figure 1: BERT-based methods for determining the stance of the perspective with respect to the claim.

Encoder Representations from Transformers) neural network model (Devlin et al., 2019). BERT is trained on huge text corpora and serves as background knowledge. We fine-tune BERT for our task which also allows us to jointly model claims and perspectives. Furthermore, we enhance our model by augmenting it with a novel consistency constraint to capture agreement between the claim and perspective.

Key contributions of this paper are:

- **Model**: A neural network model for stance classification leveraging BERT representations learned over massive external corpora and a novel consistency constraint to jointly model claims and perspectives.
- **Interpretability**: A simple approach to interpret the contribution of perspective tokens in deciding their stance towards the claim.
- **Experiments**: Experiments on a recent dataset, *Perspectrum*, highlighting the effectiveness of our approach with error analysis.

2 BERT-based Approaches

In this section, first we describe the base model, $\text{BERT}_{\text{BASE}}$, that is adapted for the stance classification (Chen et al., 2019). Thereafter, we present our consistency-aware model, $\text{BERT}_{\text{CONS}}$.

2.1 Adapting BERT for Stance Classification

The goal of the stance classification task is to determine the stance of the user *Perspective* (P) with respect to the *Claim* (C). Since this task involves a pair of sentences (C and P), we follow the approach for sentence pair classification task as proposed in Devlin et al. (2019); Chen et al. (2019).

In order to obtain the representation $X^{P|C}$ of P with respect to C, this sentence pair is fused into a single input sequence by using a special classification token (`[CLS]`) and a separator token (`[SEP]`): `[CLS]` C_{toks} `[SEP]` P_{toks} `[SEP]`. The input sequences are tokenized using WordPiece tokenization. The final hidden state representation corresponding to the `[CLS]` token is used as $X^{P|C} \in R^H$. The classification probability is given by passing this representation through the softmax layer:

$$\hat{y} = softmax(X^{P|C}W^T) \quad (1)$$

where softmax layer weights $W \in R^{H \times K}$ and K is the number of stance (classification) labels. All the parameters of BERT and W are fine-tuned jointly by minimizing the cross-entropy loss ($loss_{ce}$). The architecture of this model, $\text{BERT}_{\text{BASE}}$, is shown in Figure 1a.

2.2 Consistency-aware Stance Classification

In this setting, we want to incorporate the consistency between the claim (C) and perspective (P) representations. We hypothesize that the latent representations of claim and perspective should be *dissimilar* if the perspective *opposes* the claim, whereas their representations should be *similar* if the claim is *supported* by the perspective. We capture this with the following components.

Claim Representation: To capture the latent representation of the claim, we use only the claim text as the input sequence to BERT, i.e., `[CLS]` C `[SEP]`. The final hidden state of the first input token (`[CLS]`) is used as the claim's representation $X^C \in R^H$.

Perspective Representation: Latent representation of the perspective (with respect to the claim) is captured by fusing the two sequences as described in Section 2.1. We pack the claim and perspective pair as a single input sequence and use the final hidden state of the first input token as the perspective representation $X^{P|C} \in R^H$.

Split	Supporting Pairs	Opposing Pairs	Total Pairs
train	3603	3404	7007
dev	1051	1045	2096
test	1471	1302	2773
Total	6125	5751	11876

Table 1: Perspectrum data statistics.

Capturing Consistency: To incorporate the consistency between claim and perspective representations, we use the *cosine embedding loss*:

$$loss_{cos} = \begin{cases} 1 - \cos(X^C, X^{P|C}) & y_{sim} = 1 \\ max(0, \cos(X^C, X^{P|C})) & y_{sim} = -1 \end{cases}$$

where $\cos(.)$ is the cosine similarity function. y_{sim} is equal to 1 if the perspective is *supporting* the claim (*similar* representations), and -1 if the claim is *opposed* by the perspective (*dissimilar* representations).

Joint Loss: The classification probabilities are determined by concatenating $X^{P|C}$ and $\cos(X^C, X^{P|C})$ and passing it through a softmax layer. However, unlike the $\text{BERT}_{\text{BASE}}$ configuration, parameters of the consistency-aware model are learned by optimizing the joint loss function: $loss = loss_{ce} + loss_{cos}$. With this joint loss function, we enforce consistency between latent representations of the claim and perspective. The architecture of this consistency-aware model, $\text{BERT}_{\text{CONS}}$, is shown in Figure 1b.

3 Experimental Setup

For our experiments, we consider the base version of BERT[1] with 12 layers, 768 hidden size, and 12 attention heads. We fine-tune BERT-based models using the Adam optimizer with learning rates $\{1, 3, 5\} \times 10^{-5}$ and training batch sizes $\{24, 28, 32\}$. We choose the best parameters based on the development split of the dataset. For measuring the performance, we use per-class and macro-averaged Precision/Recall/F1.

3.1 Dataset

We evaluate our approach on the *Perspectrum* dataset (Chen et al., 2019). *Perspectrum* contains claims and users' perspectives from various online debate websites like `idebate.com`, `debatewise.org`, and `procon.org`. Each claim has different perspectives along with the stance (*supporting* or *opposing* the claim). We use the same train/dev/test split as provided in the released dataset. Statistics of the dataset is shown in Table 1.

3.2 Baselines

We use the following baselines:

LSTM: A long short-term memory (LSTM) model, in which we pass the claim and perspective word representations (using GloVE-6B word embeddings of size 300) through a bidirectional LSTM. Then we concatenate the final hidden states of the claim and perspective, and pass it through dense layers with ReLU activations.

ESIM: An enhanced sequential inference model (ESIM) for natural language inference proposed in Chen et al. (2017).

MLP: Multi-layer perceptron (MLP) based model using lexical and similarity-based features – presented as a *simple but tough-to-beat baseline* for stance detection in Riedel et al. (2017).

WordAttn: Our implementation of word-by-word attention-based model using long short-term memory networks (Rocktäschel et al., 2016).

LangFeat: A random forest classifier using linguistic lexicons like NRC lexicon (Mohammad and Turney, 2010), hedges (e.g., *possibly*, *might*, etc.), positive/negative sentiment words (Hu and Liu, 2004), MPQA subjective lexicon (Wilson et al., 2005) and bias lexicon (Recasens et al., 2013) along with sentiment scores as features.

$\text{BERT}_{\text{BASE}}$: Approach proposed in Chen et al. (2019) (as described in Section 2.1).

Human: Human performance on this task as reported in Chen et al. (2019).

4 Results and Discussion

Stance classification performance of our model and the baselines on the *test* split of the Perspectrum dataset are presented in Table 2. Our consistency-aware model $\text{BERT}_{\text{CONS}}$ outperforms all the other baselines. It achieves a performance improvement of about 2 points in F1-score over the strong baseline corresponding to the $\text{BERT}_{\text{BASE}}$ model (p-value of $4.985\text{e}{-4}$ as per the McNemar test). This highlights the value addition achieved by incorporating consistency cues. Since the BERT-based models incorporate the knowledge acquired from massive external corpora, our model, $\text{BERT}_{\text{CONS}}$, captures better semantics and outperforms the other baselines.

[1]BERT implementation: `https://git.io/fhbJQ`

Approach	Supporting			Opposing			Overall (Macro)		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
LSTM	63.42	58.80	61.02	56.99	61.67	59.24	60.20	60.24	60.13
ESIM	64.38	61.32	62.81	58.53	61.67	60.06	61.46	61.50	61.44
MLP	64.53	60.98	62.71	58.50	62.14	60.26	61.51	61.56	61.48
WordAttn	64.43	63.43	63.93	59.40	60.45	59.92	62.07	62.03	62.04
LangFeat	63.74	75.05	68.94	64.75	51.77	57.53	64.24	63.41	63.23
$\text{BERT}_{\text{BASE}}$	78.43	80.08	79.25	76.95	**75.12**	76.02	77.69	77.60	77.63
$\text{BERT}_{\text{CONS}}$	**79.05**	**84.64**	**81.75**	**81.14**	74.65	**77.76**	**80.09**	**79.65**	**79.95**
Human	-	-	-	-	-	-	91.3	90.6	90.9

Table 2: Comparison of our approach $\text{BERT}_{\text{CONS}}$ with different baseline models for stance classification.

Opposing Class	Supporting Class
unauthorized, falsely, even though, unlike, cannot, not everyone, could strike, could further weaken, jeopardize, impacts, may not provide, ...	enabling, ensuring, prevail, positive discrimination, gains, help reduce, would improve, right, would allow, encourage, more effective, ...

Table 3: Top phrases for determining stance.

4.1 Interpreting Token-level Contribution

Due to the massive structure of BERT with a complex attention mechanism, it is difficult to interpret the significance of different lexical units in the text. Therefore, we propose a simple technique to interpret the contribution of each token in the text in determining the stance.

Given the claim (C) and perspective (P) pair, we tokenize P into *phrases*. We record the change in stance classification probabilities by adding one perspective phrase at a time to the input:

$$\Delta_i = |BERT_{CONS}(C, P_i) - BERT_{CONS}(C, P_{i-1})|$$

where P_i is the prefix of P up to the i^{th} phrase. This helps us in understanding the contribution of each perspective phrase towards determining the stance – the larger the change in the classification probabilities, the larger the contribution. For this analysis, we consider unigrams and chunks from a shallow parser as phrases. The top contributing phrases for the *supporting* and *opposing* classes are shown in Table 3.

4.2 Error Analysis

In this section, we analyze why the task of stance classification is challenging and why the performance of the best model configuration is far from human performance as observed by the performance gap in Table 2.

Negations: One of the major challenges in solving this task is understanding negations and their scope. For example, given the claim *"College education is worth it"*, the perspective *"Many college graduates are employed in jobs that **do not** require college degrees"* is opposing the claim. However, our model is not able to capture that the negation phrase *'do not require'* opposes the claim. On the other hand, the presence of negation in the perspective does not necessarily imply that it is opposing the claim. Contrast this with the claim *"Chess must be at the Olympics"* and perspective *"Chess is currently **not** an Olympic sport, but it should be"* – where the negation is merely a part of the statement and the stance is given by the discourse segment following *'but'*.

Commonsense: Determining the stance may require commonsense knowledge. For example, the claim *"Chess must be at the Olympics"* is opposed by the perspective *"Olympic sports are supposed to be **physical**"*. To understand this, the model should have the background knowledge that chess is not a physical sport.

Semantics: Understanding the stance also involves a deeper understanding of semantics. For example, given the claim *"Make all museums free of charge"* is opposed by the perspective *"State funding should be used **elsewhere**"*. Here, the word *'elsewhere'* is the key cue which determines the stance. However, the presence of the word *'elsewhere'* does not necessarily imply that the perspective is opposing the claim. For instance, the perspective *"We could spend the money **elsewhere**"* is supporting the claim *"The EU should significantly reduce the amount it spends on agricultural production subsidies"*. Hence, the polarity of the word *'elsewhere'* is determined by the context and semantics of the statement.

5 Conclusion

In this work, we propose a consistency-aware neural network model for stance classification. Our model leverages representations from the BERT model trained over massive external corpora and a novel consistency constraint to jointly model claims and perspectives. Our experiments on a recent benchmark highlight the advantages of our approach. We also study the gap in human performance and the performance of the best model for stance classification.

Acknowledgments

This research was partly supported by the ERC Synergy Grant "imPACT" (No. 610150).

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9, Portland, Oregon. Association for Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557. Association for Computational Linguistics.

Wei-Fan Chen and Lun-Wei Ku. 2016. UTCNN: a deep learning model of stance classification on social media text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1635–1645, Osaka, Japan. The COLING 2016 Organizing Committee.

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*.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

FNC-1. 2016. Fake news challenge stage 1 (fnc-1): Stance detection.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *CoRR*, abs/1704.07221.

Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, Berlin, Germany. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, New Orleans, Louisiana. Association for Computational Linguistics.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *ICLR*.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 116–125, Beijing, China. Association for Computational Linguistics.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596, Montréal, Canada. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.