# Speaker-change Aware CRF for Dialogue Act Classification

Guokan Shang[1,2], Antoine J.-P. Tixier[1], Michalis Vazirgiannis[1,3], Jean-Pierre Lorré[2]

[1]École Polytechnique, [2]LINAGORA, [3]AUEB

December 2020

LaTeX of the slides: https://www.overleaf.com/read/phgjpdjvmdpc

# Introduction

Dialogue Act (DA) classification aims at assigning to each utterance in a conversation a DA label to represent its **communicative intention**.

- Useful annotations to many spoken language understanding tasks.

| Change | Speaker | Utterance | DA |
|---|---|---|---|
| - | B | Of course I use, | sd |
| True | A | <laughter>. | x |
| True | B | credit cards. | + |
| False | B | I have a couple of credit cards | sd |
| True | A | **Yeah.** | b |
| True | B | and, uh, use them. | + |
| True | A | Uh-huh, | b |
| False | A | do you use them a lot? | **qy** |
| True | B | Oh, we try not to. | **ng** |

Table: Fragment from SwDA conversation sw3332. **Statement**-non-opinion (sd), Non-verbal (x), Interruption (+), Acknowledge/Backchannel (b), Yes-No-**Question** (qy), Negative non-no **answers** (ng).

$\Rightarrow$ There are dependencies both at the **utterance level** and at the **label level**.

# Related work

## Multi-class classification

Consecutive DA labels are considered to be independent, predicted in isolation.

- **naive Bayes** (Grau et al. 2004), **Maxent** (Venkataraman et al. 2005; Ang, Liu, and E. Shriberg 2005), or **SVM** (Liu 2006).
- **Deep learning models** (Ries 1999; Khanpour, Guntakandla, and Nielsen 2016; Shen and H.-y. Lee 2016; Kalchbrenner and Blunsom 2013; J. Y. Lee and Dernoncourt 2016; Ortega and Vu 2017; Bothe et al. 2018)

## Sequence labeling

DA labels for all the utterances in the conversation are classified together.

- **HMMs** (Stolcke et al. 2000; Surendran and Levow 2006; Tavafi et al. 2013) and **CRFs** (Lendvai and Geertzen 2007; Zimmermann 2009; Kim, Cavedon, and Baldwin 2010)
- Neural sequence labeling architectures: **BiLSTM-Softmax** (W. Li and Wu 2016; Tran, Zukerman, and Haffari 2017; Liu et al. 2017) and **BiLSTM-CRF** (Kumar et al. 2018; Chen et al. 2018; Raheja and Tetreault 2019; R. Li et al. 2019).

BiLSTM-CRF is able to capture the dependencies among consecutive **utterances** (with BiLSTM) and among consecutive DA **labels** (with CRF).

# Motivation

The state-of-the-art works do not take into account the additional **speaker** input sequence.

- This is a major **limitation**.
- This extra input could greatly improve DA prediction.

## Turn management (Sacks, Schegloff, and Jefferson 1974)

- Dialogue participants follow an underlying turn-taking system to occupy or release (not arbitrarily) the speaker role (Petukhova and Bunt 2009).
- $\Rightarrow$ DA transition should be conditioned both on the utterance transition and the **speaker-change** (not speaker-identifier).
- $\Rightarrow$ "A *Question* is usually followed by an *Answer*" (only partially true), + **[*if the speaker changed*]**.

**To address the limitation, we propose a simple modification of the CRF layer that takes speaker-change into account.**
We evaluate our modified CRF layer within the BiLSTM-CRF architecture.

# BiLSTM-CRF

## Notation

$X = \{\mathbf{x}^t\}_{t=1}^T$: the input utterance sequence, of length $T$.

$Y = \{y^t\}_{t=1}^T$: the target label sequence, where $y^t \in \mathcal{Y}$, the DA label set of size $K$.

We use $y^t$ to denote the label and its integer index interchangeably.

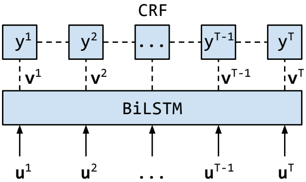


Figure: BiLSTM-CRF. $\{\mathbf{u}^t\}_{t=1}^T$ are utterance embeddings.

1. LSTM (text encoder): utterances $X = \{\mathbf{x}^t\}_{t=1}^T \to$ utterance embeddings $\{\mathbf{u}^t\}_{t=1}^T$.
2. BiLSTM: $\{\mathbf{u}^t\}_{t=1}^T \to$ conversation-level utterance representations $\{\mathbf{v}^t\}_{t=1}^T$.
3. CRF: $\{\mathbf{v}^t\}_{t=1}^T \to$ labels $Y = \{y^t\}_{t=1}^T$

# CRF layer

CRF is a **discriminative** probabilistic graphical framework used to label sequences (Lafferty, McCallum, and Pereira 2001).

$$P(Y|X) = \frac{\exp(\psi(X, Y))}{\sum_{\tilde{Y}} \exp(\psi(X, \tilde{Y}))} \tag{1}$$

where $\psi(X, Y)$ is a feature function that assigns a ***path score*** to the label sequence $Y$, giving the input sequence $X$. $\tilde{Y}$ denotes one of all possible label sequences (paths).

$$\psi(X, Y) = \sum_{t=1}^{T} h(y^t, X) + \sum_{t=1}^{T-1} g(y^t, y^{t+1}) \tag{2}$$

$\psi(X, Y)$ is defined as the sum of ***emission scores*** (or state scores) and ***transition scores*** over all time steps.

$$h(y^t, X) = (\mathbf{W}\mathbf{v}^t + \mathbf{b})[y^t] \tag{3}$$

where the conversation-level utterance representation $\mathbf{v}^t$ is converted into a vector of size $K$.

$$g(y^t, y^{t+1}) = \mathbf{G}[y^t, y^{t+1}] \tag{4}$$

where $\mathbf{G}$ is the label transition matrix of size $K \times K$.

# CRF layer

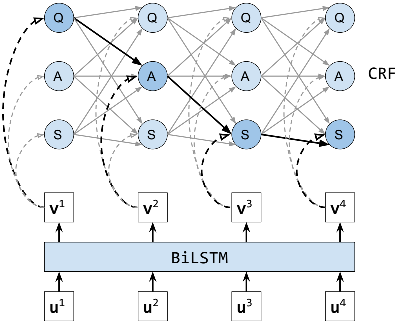


Figure: BiLSTM-CRF for an example.

For a training set of $M$ conversations, the loss can be written as:

$$\mathcal{L} = \sum_{m=1}^{M} -\log P(Y^m|X^m) \tag{5}$$

At test time, the optimal label sequence, i.e., $Y^* = \text{argmax}_{\tilde{Y}} P(\tilde{Y}|X)$ for unseen $X$, is obtained with the Viterbi algorithm (Viterbi 1967), with polynomial complexity $O(TK^2)$.

# Contribution

## Notation

$S = \{s^t\}_{t=1}^{T}$: the sequence of speaker-identifiers.
$Z = \{z^{t,t+1}\}_{t=1}^{T-1}$: **the sequence of speaker-changes**, obtained by comparing neighbors in $S$.
E.g., $z^{2,3} = 0$ means the speaker does not change from time $t = 2$ to $t = 3$.

We extend the original CRF so that it considers as **additional input**, the sequence $Z$.

$$P(Y|X,Z) = \frac{\exp(\psi(X,Y,Z))}{\sum_{\tilde{Y}} \exp(\psi(X,\tilde{Y},Z))} \tag{6}$$

Specifically, transition scores in our modified CRF layer are computed as follows:

$$\begin{aligned} g(y^t, y^{t+1}, z^{t,t+1}) =& (1 - z^{t,t+1}) * \mathbf{G}_0[y^t, y^{t+1}] + \\ & z^{t,t+1} * \mathbf{G}_1[y^t, y^{t+1}] \end{aligned} \tag{7}$$

where $\mathbf{G}_0$ and $\mathbf{G}_1$ are label transition matrices of size $K \times K$, corresponding respectively to the **“speaker unchanged”** and **“speaker changed”** cases.

# Dataset

Switchboard Dialogue Act (SwDA) dataset (Jurafsky, L. Shriberg, and Biasca 1997; Stolcke et al. 2000).

- telephonic conversations recorded between two randomly selected speakers talking about one of various general topics (air pollution, music, football, etc.).
- training, validation and testing partition of 1003, 112, and 19 conversations.
- utterances are annotated with **42** mutually exclusive DA labels
- Inter-annotator agreement is 84%.

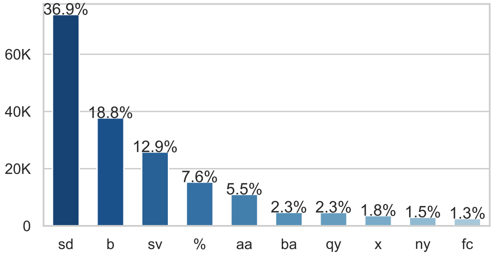


Figure: Counts and frequencies of the 10 most represented DA labels in the SwDA dataset. There are 200444 utterances in total.

# Results

| Model | | BiLSTM input | CRF extra input | Accuracy (% ± SD) |
|---|---|---|---|---|
| **a**) | Our CRF | $\mathbf{u}^t$ | SC | **78.70** ± .37 |
| a1) | | $\mathbf{u}^t$ + SI | SC | 78.32 ± .28 |
| a2) | | $\mathbf{u}^t$ + SC | SC | 78.65 ± .47 |
| **b**) | Vanilla CRF | $\mathbf{u}^t$ | - | 77.69 ± .38 |
| b1) | | $\mathbf{u}^t$ + SI | - | 77.86 ± .61 |
| b2) | | $\mathbf{u}^t$ + SC | - | 78.33 ± .71 |
| c) | Softmax | $\mathbf{u}^t$ | - | 77.80 ± .48 |
| c1) | | $\mathbf{u}^t$ + SI | - | 77.73 ± .44 |
| c2) | | $\mathbf{u}^t$ + SC | - | 78.33 ± .49 |
| a) + b) ensembling | | $\mathbf{u}^t$ | SC | **78.89** ± .20 |
| a) + b) joint training | | $\mathbf{u}^t$ | SC | 78.27 ± .47 |

Table: Results, averaged over 10 runs and 42 DA labels. SI: speaker-identifier, SC: speaker-change, $\mathbf{u}^t$: utterance embedding, ±: standard deviation.

# Analysis

**Our CRF vs. Vanilla CRF**

- $\Rightarrow$ our model a) outperforms the base model b) by 1%, over 42 labels.
- $\Rightarrow$ The boost is greater than the gains of 0.26% (Liu et al. 2017) and 0.09% (Bothe et al. 2018) reported by previous attempts at leveraging speaker information.

**Confusion matrices**

- 10 most frequent labels (91%) $\Rightarrow$ outperforms on a majority of them, but not on `sd`.
- 10 best predicted labels (20%) and 10 worst predicted labels (40%) $\Rightarrow$ Our model is most useful for the difficult and rare DAs requiring speaker-change awareness.

**Different ways of incorporating speaker information**

- concatenate the one-hot encoded SI vector (of size 2) and the binary speaker-change vector (of size 1) with $\mathbf{u}^t$ the utterance embedding.

**BiLSTM-CRF VS. BiLSTM-Softmax**

- $\Rightarrow$ competitive, this finding is not surprising and consistent with the results reported in recent works on other tasks (Reimers and Gurevych 2017; Yang, Liang, and Zhang 2018; Cui and Zhang 2019).

**Ensembling vs. joint training**

- Ensembling: combines the predictions of the two trained models by averaging their emission and transition scores respectively.
- Joint training: $\mathbf{G}_{basis}[y^t, y^{t+1}] + (1 - z^{t,t+1}) * \mathbf{G}_0[y^t, y^{t+1}] + z^{t,t+1} * \mathbf{G}_1[y^t, y^{t+1}]$

# Confusion matrices for the 10 most frequent labels

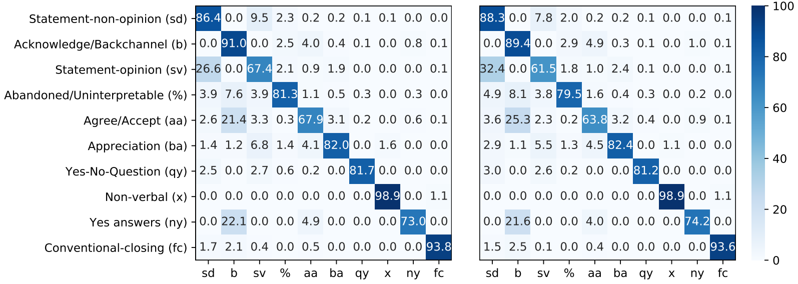


Figure: Normalized confusion matrices, averaged over 10 runs, for the 10 most frequent DA labels (90.9% of all annotations). Left: our model, right: base model. Rows (columns) correspond to true (predicted) classes.

| | | P | R | F1 |
|---|---|---|---|---|
| Our | sd | 80.49 | 86.36 | 83.32 |
| | sv | 71.54 | 67.41 | 69.42 |
| Vanilla | sd | 77.83 | 88.32 | 82.74 |
| | sv | 73.24 | 61.48 | 66.84 |

Table: Precison, Recall, and F1 score (%) of our model vs. base model on the `sd` and `sv` labels.

# Visualization of transition matrices 1/2

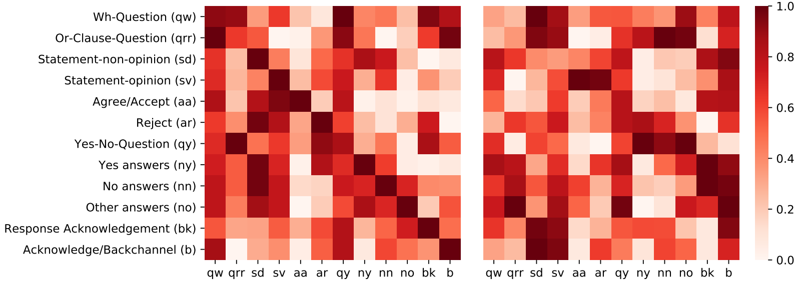


Figure: Normalized transition matrices (averaged over 10 runs). Left: $\mathbf{G}_0$ (speaker unchanged) and Right: $\mathbf{G}_1$ (speaker changed) of **our CRF layer**. The darker, the greater the score.

# Visualization of transition matrices 2/2

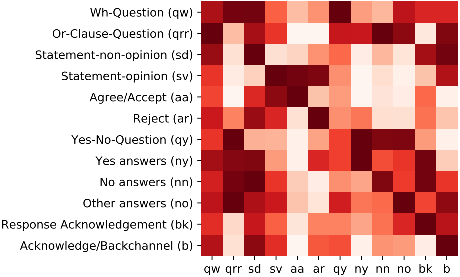


Figure: Normalized transition matrix (averaged over 10 runs). **G** of **vanilla CRF layer**. The darker, the greater the score.

# Conclusion

- A modified CRF layer that takes as extra input the sequence of speaker-changes was proposed. Code is publicly available: `https://bitbucket.org/guokan_shang/da-classification`.
- Experiments showed that our CRF layer outperforms vanilla CRF ⇒ taking speaker information into consideration was beneficial.
- Visualizations confirmed that our improved CRF was able to learn complex speaker-change aware DA transition patterns in an end-to-end way.

## Future work

Future research should be devoted to address the limitation of the Markov property of CRF layer, by developing a model that is capable of capturing longer-range dependencies within and among the three sequences: that of speakers, utterances, and DA labels.

## Acknowledgments

This work was supported by the LinTO meeting assistant project (Lorré et al. 2019). `https://linto.ai`

# References I


- Viterbi, Andrew (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: **IEEE transactions on Information Theory** 13.2, pp. 260–269.
- Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation". In: **Language** 50.4, pp. 696–735. ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/412243.
- Jurafsky, Dan, Liz Shriberg, and Debra Biasca (1997). "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual". In: **Institute of Cognitive Science Technical Report**. URL: https://web.stanford.edu/~jurafsky/ws97/manual.august1.html.
- Ries, Klaus (1999). "HMM and neural network based speech act detection". In: **1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)**. Vol. 1. IEEE, pp. 497–500.
- Stolcke, Andreas et al. (2000). "Dialogue act modeling for automatic tagging and recognition of conversational speech". In: **Computational Linguistics** 26.3, pp. 339–374. URL: https://www.aclweb.org/anthology/J00-3003.
- Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: **Proceedings of the Eighteenth International Conference on Machine Learning**. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1558607781.
- Grau, Sergio et al. (2004). "Dialogue act classification using a Bayesian approach". In: **9th Conference Speech and Computer**.
- Ang, Jeremy, Yang Liu, and Elizabeth Shriberg (2005). "Automatic dialog act segmentation and classification in multiparty meetings". In: **Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.** Vol. 1. IEEE, pp. I–1061.
- Venkataraman, Anand et al. (2005). "Does active learning help automatic dialog act tagging in meeting data?" In: **Ninth European Conference on Speech Communication and Technology**.
- Liu, Yang (2006). "Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus". In: **Ninth International Conference on Spoken Language Processing**.
- Surendran, Dinoj and Gina-Anne Levow (2006). "Dialog act tagging with support vector machines and hidden Markov models". In: **Ninth International Conference on Spoken Language Processing**.
- Lendvai, Piroska and Jeroen Geertzen (2007). "Token-based chunking of turn-internal dialogue act sequences". In: **Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue**, pp. 174–181.
- Petukhova, Volha and Harry Bunt (2009). "Who's next? Speaker-selection mechanisms in multiparty dialogue". In: **Workshop on the Semantics and Pragmatics of Dialogue**.
- Zimmermann, Matthias (2009). "Joint segmentation and classification of dialog acts using conditional random fields". In: **Tenth Annual Conference of the International Speech Communication Association**.
- Kim, Su Nam, Lawrence Cavedon, and Timothy Baldwin (Oct. 2010). "Classifying Dialogue Acts in One-on-One Live Chats". In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. Cambridge, MA: Association for Computational Linguistics, pp. 862–871. URL: https://www.aclweb.org/anthology/D10-1084.

# References II


- Kalchbrenner, Nal and Phil Blunsom (Aug. 2013). "Recurrent Convolutional Neural Networks for Discourse Compositionality". In: **Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality**. Sofia, Bulgaria: Association for Computational Linguistics, pp. 119–126. URL: https://www.aclweb.org/anthology/W13-3214.
- Tavafi, Maryam et al. (Aug. 2013). "Dialogue Act Recognition in Synchronous and Asynchronous Conversations". In: **Proceedings of the SIGDIAL 2013 Conference**. Metz, France: Association for Computational Linguistics, pp. 117–121. URL: https://www.aclweb.org/anthology/W13-4017.
- Khanpour, Hamed, Nishitha Guntakandla, and Rodney Nielsen (Dec. 2016). "Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network". In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 2012–2021. URL: https://www.aclweb.org/anthology/C16-1189.
- Lee, Ji Young and Franck Dernoncourt (2016). "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks". In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. San Diego, California: Association for Computational Linguistics, pp. 515–520. DOI: 10.18653/v1/N16-1062. URL: http://aclweb.org/anthology/N16-1062.
- Li, Wei and Yunfang Wu (Dec. 2016). "Multi-level Gated Recurrent Neural Network for dialog act classification". In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1970–1979. URL: https://www.aclweb.org/anthology/C16-1185.
- Shen, Sheng-syun and Hung-yi Lee (2016). "Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection". In: **arXiv preprint arXiv:1604.00077**.
- Liu, Yang et al. (Sept. 2017). "Using Context Information for Dialog Act Classification in DNN Framework". In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2170–2178. DOI: 10.18653/v1/D17-1231. URL: https://www.aclweb.org/anthology/D17-1231.
- Ortega, Daniel and Ngoc Thang Vu (Aug. 2017). "Neural-based Context Representation Learning for Dialog Act Classification". In: **Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue**. Saarbrücken, Germany: Association for Computational Linguistics, pp. 247–252. DOI: 10.18653/v1/W17-5530. URL: https://www.aclweb.org/anthology/W17-5530.
- Reimers, Nils and Iryna Gurevych (2017). "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks". In: **arXiv preprint arXiv:1707.06799**.
- Tran, Quan Hung, Ingrid Zukerman, and Gholamreza Haffari (Apr. 2017). "A Hierarchical Neural Model for Learning Sequences of Dialogue Acts". In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**. Valencia, Spain: Association for Computational Linguistics, pp. 428–437. URL: https://www.aclweb.org/anthology/E17-1041.
- Bothe, Chandrakant et al. (May 2018). "A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks". In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L18-1307.

# References III


- Chen, Zheqian et al. (2018). "Dialogue Act Recognition via CRF-Attentive Structured Network". In: **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, pp. 225–234. ISBN: 9781450356572. DOI: 10.1145/3209978.3209997. URL: https://doi.org/10.1145/3209978.3209997.
- Kumar, Harshit et al. (2018). "Dialogue Act Sequence Labeling Using Hierarchical Encoder With CRF". In: **AAAI Conference on Artificial Intelligence**. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16706.
- Yang, Jie, Shuailong Liang, and Yue Zhang (2018). "Design challenges and misconceptions in neural sequence labeling". In: **arXiv preprint arXiv:1806.04470**.
- Cui, Leyang and Yue Zhang (2019). "Hierarchically-Refined Label Attention Network for Sequence Labeling". In: **arXiv preprint arXiv:1908.08676**.
- Li, Ruizhe et al. (Nov. 2019). "A Dual-Attention Hierarchical Recurrent Neural Network for Dialogue Act Classification". In: **Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)**. Hong Kong, China: Association for Computational Linguistics, pp. 383–392. DOI: 10.18653/v1/K19-1036. URL: https://www.aclweb.org/anthology/K19-1036.
- Lorré, Jean-Pierre et al. (2019). "LinTO: Assistant vocal open-source respectueux des données personnelles pour les réunions d'entreprise". In: **APIA**, p. 63.
- Raheja, Vipul and Joel Tetreault (June 2019). "Dialogue Act Classification with Context-Aware Self-Attention". 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3727–3733. DOI: 10.18653/v1/N19-1373. URL: https://www.aclweb.org/anthology/N19-1373.