MT-Speech at SemEval-2022 Task 10: Incorporating Data Augmentation and Auxiliary Task with Cross-Lingual Pretrained Language Model for Structured Sentiment Analysis

Cong Chen, Jiansong Chen, Cao Liu, Fan Yang, Guanglu Wan, Jinxiong Xia,
Meituan, Beijing, China
{chencong29, chenjiansong, liucao}@meituan.com

Abstract

Structured Sentiment Analysis (SSA) is an important component of sentiment analysis, which is a critical task in NLP. Traditional SSA methods cannot capture the cross-lingual interactions between different language and there is insufficient annotated data, especially in the cross-lingual settings. In this paper, we use the pre-trained language models with two auxiliary tasks and adopt data augmentation to address the above problems. Specifically, we employ XLM-RoBERTa to capture the cross-lingual knowledge interactions and enhance the generalization in multilingual/cross-lingual settings. Furthermore, we leverage two data augmentation techniques and propose two auxiliary tasks to improve the performance on the few-shot and zero-shot settings. Experiments demonstrate that our model ranks first on the cross-lingual sub-task and second on the monolingual sub-task of SemEval-2022 task 10.

1 Introduction

Structured Sentiment Analysis (SSA) is an important task in sentiment analysis (Barnes et al., 2021; Liu, 2012; Mitchell et al., 2013). The goal of SSA is to extract all opinion tuples from given texts. The opinion tuple (h, t, e, p) consists of a holder (h) which expresses a polarity (p) towards a target (t) by a textual sentiment expression (e). Benefiting a variety of business applications, such as human-machine dialogue and recommendation systems, SSA has attracted much more attention from both academia and industry (Pang et al., 2008; Mitchell et al., 2013; Xu et al., 2020; Ovrelid et al., 2020; Li et al., 2019).

The mainstream method for SSA is to adopt a pipeline approach by separately performing the subtasks including holder extraction and target extraction. However, such methods can not capture dependencies of multiple sub-tasks. To address this problem, Barnes et al. (2021) leverages graph-based dependency parsing to capture the dependencies among opinion tuples, where sentiment holders, targets and expressions are the nodes, and the relations of them as the arcs. This model has obtained state-of-the-art performance on SSA.

However, the aforementioned methods still suffer from some important issues. Firstly, the knowledge of the pre-trained language models (PLMs) has not been fully exploited. In fact, the cross-lingual PLMs contain rich knowledge of the interactions among different languages. Secondly, the above data-driven models rely on a large amount of annotation data, but there is insufficient or even no annotated data in the real scene. For example, in SemEval-2022 shared task 10 (Barnes et al., 2022), the **MultiB**$_{EU}$ (Barnes et al., 2018) dataset has only 1215 sentences and the **MultiB**$_{CA}$ (Barnes et al., 2018) dataset have only 1341 sentences, and there is no training data for the target language in the cross-lingual setting, which heavily hinders the performance on SSA.

To address the above problems, we propose a unified and end-to-end model for SSA, which performs data augmentation and adopts auxiliary tasks with cross-lingual PLMs. Specifically, we employ XLM-RoBERTa (Conneau and Lample, 2019; Conneau et al., 2019) as the backbone encoder to make use of its multilingual/cross-lingual knowledge. To alleviate the problem of insufficiency or lack of annotated data, we adopt two data augmentation methods: the one is to add in-domain annotated data of the same task under the training stage, and the other is to employ Masked Language Model (MLM) (Devlin et al., 2018) for generating similar texts. Furthermore, in addition to predicting each tuple in the dependency parsing graph simultaneously, we add two auxiliary tasks: 1) sequence labeling to predict the span of the holder / target / expression, and 2) sentiment polarity classification. Note that both of them do not need additional

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1329 - 1335

Methods	**NoReC**$_{Fine}$	**MultiB**$_{CA}$	**MultiB**$_{EU}$	**OpeNer**$_{EN}$	**OpeNer**$_{ES}$	**MPQA**	**DS**$_{Unis}$	Average
Top1	0.529(2)	0.728(1)	0.739(1)	0.760(2)	0.722(4)	0.447(1)	0.494(1)	0.631(1)
Top2(Ours)	0.524(3)	0.728(1)	0.739(1)	0.763(1)	0.742(1)	0.416(2)	0.485(2)	0.628(2)
Top3	0.533(1)	0.709(3)	0.715(3)	0.756(3)	0.732(3)	0.402(3)	0.463(3)	0.616(3)
Top4	0.504(4)	0.681(6)	0.723(2)	0.747(4)	0.735(2)	0.375(5)	0.410(9)	0.596(4)
Top5	0.483(8)	0.711(2)	0.681(6)	0.727(5)	0.686(7)	0.379(4)	0.373(13)	0.577(5)

Table 1: Comparisons on monolingual evaluation leader board.

Methods	**OpeNer**$_{ES}$	**MultiB**$_{CA}$	**MultiB**$_{EU}$	Average
Top1(Ours)	0.644(1)	0.643(1)	0.632(1)	0.640(1)
Top2	0.618(3)	0.562(7)	0.584(2)	0.588(2)
Top3	0.628(2)	0.607(3)	0.527(4)	0.587(3)
Top4	0.604(5)	0.596(4)	0.512(7)	0.571(4)
Top5	0.589(6)	0.593(5)	0.516(6)	0.566(5)

Table 2: Comparisons on cross-lingual evaluation leader board.

annotations.

We conduct experiments on subtask 1 and subtask 2 of SemEval-2022 shared task on SSA. Experimental results demonstrate that our method outperforms strong baselines. We rank first on the cross-lingual sub-task and rank second on the monolingual subtask in SemEval-2022 task 10[1].

To summarize, our contributions are as follows.

- We leverage cross-lingual pre-trained language models to capture the interactive information knowledge among different languages.
- We combine existing in-domain training data and produce new training data by MLM to alleviate the problem of insufficiency or lack of annotated data.
- We propose two auxiliary tasks that do not require additional annotations to further improve the performance.
- Experimental results demonstrate the effectiveness of our proposed model, and we rank first on the subtask 2 and rank second on the subtask 1 on SemEval-2022 task 10.

2 Method

We incorporate the dependency graph parsing approach (Barnes et al., 2021) into our model. The general architecture is a pre-trained language model (e.g BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), etc..) followed by a three-layers BiLSTMs (Schuster and Paliwal, 1997; Cross and Huang, 2016) and the bilinear (Kiperwasser and Goldberg, 2016; Dozat and Manning, 2016) attention as the decoding component. Hence, we can take advantage of the knowledge of large-scale PLMs (Vaswani et al., 2017; Radford et al., 2019; Raffel et al., 2020; Yang et al., 2019; Wolf et al., 2019) and deep semantic dependency parsing (Dozat and Manning, 2016; Oepen et al., 2020; Kurtz et al., 2020).

2.1 Encoder

We consider several state-of-the-art models as the candidates of our model's backbone, such as Multilingual BERT (mBERT) (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2019), and info-XLM(Chi et al., 2021). Particularly, we choose the XLM-RoBERTa backbone as the baseline. Because subtask 1 is a multilingual problem and subtask 2 is a cross-lingual zero-shot problem. They both benefit from the Translation Language Modeling (TLM) objective in XLM-RoBERTa. The TLM and Masked Language Modeling (MLM) objectives in the XLM-family models perform better than mBERT, which is simply trained on multilingual corpus with the MLM objective. Additionally, XLM-RoBERTa is trained on more data, which makes it more robust. Another reason we choose XLM-RoBERTa is that it is a large open-source model for downstream applications. We did not employ info-XLM as it is trained with the sentence-level classification objective, which is not suited for this task.

[1] https://competitions.codalab.org/competitions/33556

Methods	**MPQA**	**DS**$_{Unis}$	**OpeNer**$_{EN}$	**OpeNer**$_{ES}$	**MultiB**$_{CA}$	**MultiB**$_{EU}$	**NoReC**$_{Fine}$	Average
w2v + BiLSTM	0.103	0.166	0.525	0.526	0.524	0.539	0.320	0.386
mBERT	0.231	0.280	0.571	0.611	0.526	0.517	0.373	0.446
mBERT +BiLSTM	0.266	0.285	0.621	0.614	0.619	0.589	0.386	0.483
XLM-R +BiLSTM	0.332	0.357	0.705	0.654	0.669	0.650	0.481	0.550

Table 3: Monolingual task performances with different encoders. All models use the same bilinear attention decoder. All results are evaluated on the official released development set.

2.2 Data Augmentation

We provide two data augmentation methods to further boost the performance of our model. First is the in-domain data enhancement (DA1) to better make use of the data in different languages. The second is the MLM data augmentation (DA2).

2.2.1 In-domain Data Enhancement

We combined different dataset that belong to the same domain in the training phase, to help improve generalization. Noticed that the four datasets **MultiB**$_{EU}$, **MultiB**$_{CA}$, **Opener**$_{ES}$, and **Opener**$_{EN}$ (Agerri et al., 2013) are all from the hotel review corpus. We observe these datasets share some common features even though they are of different languages. These languages share the same or similar words for the same objects or concepts. For example, the "hotel" word in Catalan and Spanish are also "hotel", and in Basque it is a similar word "hotela". Besides, the people who use these languages share the same sentiment polarity tendency on the hotel review domain. Combining the four language datasets together as a whole training set will improve the overall performance. We additionally add the Portuguese hotel review dataset (BOTE-rehol)(Barros and Bona, 2021) and the English laptops review dataset (RES14) (Pontiki et al., 2014; Xu et al., 2020) for extra training, which needs to be converted to the same format as this task.

2.2.2 Data Augmentation by Masked Word Generation

The Masked Language Model corrupts the input texts by randomly replacing the tokens with the [MASK] tokens, and predicts the original token at the [MASK] positions. For each sample with valid opinion tuples, by randomly masking a small portion of the tokens in the text, we obtain a new sample whose meaning is similar as the original with the same labels. Note that in this task we do not mask the sentiment expression words as the PLMs may generate words of different polarities which are inconsistent with the original labels.

2.3 Auxiliary Tasks

SSA consists of structure prediction and sentiment polarity classification, and to handle these two tasks in an end-to-end manner is non-trivial. We propose two auxiliary tasks to provide more training signals to the model to better handle structure prediction and polarity classification. For structure prediction, we add a sequence labeling task to explicitly predict the type (target, holder, or expression) of each token. For polarity classification, we add more sentiment polarity classification data as extra tasks. Specifically, we use the average pooling of the model's BiLSTM hidden-states as sentence-level representations. The representation is fed to a multilayer perceptron(MLP) for sentence-level sentiment polarity classification. The total loss is the weighted sum of the main loss and the auxiliary losses:

$$\mathcal{L} = \mathcal{L}^p + (\mathcal{L}^s + \mathcal{L}^c)/2 \quad (1)$$

where $\mathcal{L}^p$ is the primary loss of the SSA task. $\mathcal{L}^s$ and $\mathcal{L}^c$ are the losses for sequence-labeling and classification task, respectively.

3 Experiments

3.1 Experimental data

We use the officially released development set as the test set, and randomly split the original training set into the training and development sets. We keep the size of the split development sets the same as the official released development set.

We transform the BOTE-rehol and RES14 datasets into the graph format and leave the opinion tuples' holders empty since the two datasets do not contain holder labels.

For each dataset, we convert the sentiment graph labels to sequence labeling labels, which will be added as an auxiliary task during training. Additionally, for the **MultiB**$_{CA}$ and **Opener**$_{ES}$ datasets, we make use of the Catalonia Independence Corpus (CIC) (Zotova et al., 2020) as the extra training

Methods	**MultiB**$_{CA}$	**MultiB**$_{EU}$	**OpeNer**$_{ES}$	**OpeNer**$_{EN}$
baseline	0.685	0.650	0.654	0.712
w / DA1	0.727	0.670	0.711	0.729

Table 4: Monolingual performances of data augment (DA) on official released development set.

Methods	**MultiB**$_{CA}$	**MultiB**$_{EU}$	**OpeNer**$_{ES}$
OpeNer$_{EN}$	0.574	0.438	0.630
w / DA1	0.600	0.550	0.620
w / DA1-2	0.623	0.567	0.657

Table 5: Cross-lingual performances of data augment (DA) on official released development set.

data of the polarity classification task.

3.2 Implementation details

We adopt the head-first setting (Barnes et al., 2021) which sets the first token within each span as the head of the span with all other tokens within that span as dependents. The root node is represented by the first token within the sentiment expression. For any word that is tokenized into a head-token followed by several sub-tokens, we set the head-token as the head and its following sub-tokens as the dependents. We use *holder, targ, exp-Positive, exp-Negative, exp-Neutral, None* as labels to denote different node types. The relation between each node is expressed by the attention value between the head-tokens of the nodes.

We try different combinations to get the best results for different subtasks. With XLM-RoBERTa-large as the backbone, details combinations are listed in Table 6 for the four hotel review datasets (**MultiB**$_{EU}$, **MultiB**$_{CA}$, **Opener**$_{ES}$ and **Opener**$_{EN}$). For **DS**$_{Unis}$ dataset (Toprak et al., 2010), we chose English RES14 dataset which also has few holders elements as its in-domain dataset. As most of the sentences in the two dataset are expressed without holders elements.

When generating new samples via MLM, for each sentence with at least one valid sentiment tuple, we mask one position i at a time and feed the masked sentence to the XLM-RoBERTa-large model. The PLM generates a word based on the highest probability p_i. We pick the top 5 most confident samples ranked by the PLM's output probability P_i for $i \in n$, where n denotes the possible masked positions. And set a threshold p as 0.85 to filter out any samples with a probability lower than the threshold. Repeated samples are not considered valid. The generated samples are treated as supplementary data to the original dataset.

For domain adaptation, we further pre-trained XLM-RoBERTa-large with all the data from the released datasets via Mask Language Modeling (MLM)(Devlin et al., 2018). We pick the best checkpoint according to the lowest perplexity on the development set.

3.3 Overall Comparisons

Comparison Settings. Firstly, we compare our model with other participant teams on the leader board of the structured sentiment competition. Table 1 and Table 2 record the comparison results of the monolingual and cross-lingual evaluation, respectively.

Comparison Results. (1) Our methods rank second and first on the monolingual and cross-lingual evaluation, respectively, which demonstrates the effectiveness of our proposed model. (2) Our model remarkably outperforms the top2 team in the cross-lingual subtask, which indicates our model has better generalization on the zero-shot cross-lingual settings.

3.4 Effectiveness of Cross-lingual Pre-trained Language Model

Comparison Settings. To prove the effectiveness of XLM-RoBERTa[2], a cross-lingual pre-trained language model, we compare it with the following baselines: 1) w2v + BiLSTM, BiLSTMs with word2vec (Mikolov et al., 2013) word embeddings; 2) mBERT, the Multilingual BERT (Devlin et al., 2018); 3) mBERT + BiLSTM; 4) XLM-RoBERTa + BiLSTM.

Comparison Results. (1) Table 3 demonstrates that XLM-RoBERTa + BiLSTM obtains the best performance among all of the benchmarks, and the average score outperforms the strongest baseline (mBERT + BiLSTM) by 6.7%. It proves that our model has great generalization ability. (2) BiLSTM can improve the performance by 3.7%, which indicates the BiLSTM layer can capture sequence information, which is beneficial to sequence encoding (Cross and Huang, 2016).

[2]We leverage the large version of XLM-RoBERTa to improve performances.

Methods	**MultiB**$_{CA}$	**MultiB**$_{EU}$	**OpeNer**$_{ES}$
Data combination	**OpeNer**$_{ES}$ **MultiB**$_{EU}$	**OpeNer**$_{EN}$ **MultiB**$_{CA}$ **OpeNer**$_{ES}$ bote-rehol	**OpeNer**$_{EN}$ **MultiB**$_{CA}$

Table 6: In domain data combination for cross-lingual evaluation. No target language data is participated in training and development.

Methods	**MPQA**	**DS**$_{Unis}$	**OpeNer**$_{EN}$	**OpeNer**$_{ES}$	**MultiB**$_{CA}$	**MultiB**$_{EU}$	**NoReC**$_{Fine}$
baseline	0.296	0.337	0.648	0.641	0.662	0.647	0.400
w / Auxiliary-task	0.305	0.346	0.674	0.660	0.687	0.657	0.411

Table 7: Performances on the official released development set with auxiliary tasks. We use RoBERTa-base (Liu et al., 2019) for **MPQA** (Wiebe et al., 2005), **DS**$_{Unis}$ and **OpeNer**$_{EN}$, bert-base-spanish-wwm-cased (Cañete et al., 2020) for **OpeNer**$_{ES}$, RoBERTa-base-ca (Armengol-Estapé et al., 2021) for **MultiB**$_{CA}$, berteus-base-cased (Agerri et al., 2020) for **MultiB**$_{EU}$, and norwegian-RoBERTa-base `https://huggingface.co/patrickvonplaten/norwegian-roberta-base` for **NoReC**$_{Fine}$ (Øvrelid et al., 2020).

3.5 Effectiveness of Data Augmentation

Comparison Settings. In order to demonstrate the effectiveness of data augmentation, we utilize existing training data for data augmentation (DA1) including **MultiB**$_{EU}$ **MultiB**$_{CA}$, **OpeNer**$_{ES}$ and **OpeNer**$_{EN}$. Furthermore, we leverage MLM to generate new training data for data augmentation (DA2). We record the performance in Table 4 and Table 5, where "w/" means "with", and "DA1-2" means "DA1 combined with DA2".

Comparison Results. We can conclude the following from Table 4 and Table 5: both DA1 and DA2 contribute to performance improvement, with performance increases on almost every benchmark. Specifically, the performance has remarkably improved in the cross-lingual settings, and data augmentation is more helpful on the few-shot and zero-shot settings.

3.6 The Effectiveness of Auxiliary Tasks

As shown in Table 7, we leverage the auxiliary tasks including sequence labeling and sentiment polarity classification to improve the performances. We can observe that the auxiliary tasks improve performances on all of the datasets, which demonstrate the effectiveness of the two auxiliary tasks.

4 Conclusion

This paper studies the task of structured sentiment analysis. In order to deal with the problems of poor interactions of different languages and lack of annotation data, we adopt the cross-lingual pre-trained language model and adopt data augmentation and auxiliary tasks. Specifically, we employ XLM to capture the interactive information in the pre-training stage. Furthermore, we leverage two data augmentation strategies and two auxiliary tasks to improve the performance for lack of training data. Experiments demonstrate the effectiveness of our models. Our models rank first on the cross-lingual sub-task and rank second on the monolingual sub-task of SemEval-2022 task 10.

References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. In *Sociedad Española para el Procesamiento del Lenguaje Natural*, volume 51, pages 215–218.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*

(LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

Jeremy Barnes, Oberländer Laura Ana Maria Kutuzov, Andrey and, Enrica Troiano, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, Erik Velldal, and Stephan Oepen. 2022. SemEval-2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle. Association for Computational Linguistics.

José Meléndez Barros and Glauber De Bona. 2021. A deep learning approach for aspect sentiment triplet extraction in portuguese. In *Brazilian Conference on Intelligent Systems*, pages 343–358. Springer.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and M. Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *NAACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

James Cross and Liang Huang. 2016. Incremental parsing with minimal features using bi-directional lstm. *ArXiv*, abs/1606.06406.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

E. Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Robin Kurtz, Stephan Oepen, and Marco Kuhlmann. 2020. End-to-end negation resolution as graph parsing. In *IWPT*.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. *ArXiv*, abs/1811.05082.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *EMNLP*.

Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Timothy J. O'Gorman, Nianwen Xue, and Daniel Zeman. 2020. Mrp 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. In *CONLL*.

Lilja Ovrelid, Petter Maehlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for norwegian. In *LREC*.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *COLING 2014*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Elena Zotova, Rodrigo Agerri, Manuel Nunez, and German Rigau. 2020. Multilingual stance detection: The catalonia independence corpus. *arXiv preprint arXiv:2004.00050*.