UFRGSent at SemEval-2022 Task 10: Structured Sentiment Analysis using a Question Answering Model

Lucas Rafael Costella Pessutto
Institute of Informatics
UFRGS – Brazil
lrcpessutto@inf.ufrgs.br

Viviane P. Moreira
Institute of Informatics
UFRGS – Brazil
viviane@inf.ufrgs.br

Abstract

This paper describes the system submitted by our team (UFRGSent) to SemEval-2022 Task 10: Structured Sentiment Analysis. We propose a multilingual approach that relies on a Question Answering model to find tuples consisting of holder, target, and opinion expression. The approach starts from general questions and uses the extracted tuple elements to find the remaining components. Finally, we employ an aspect sentiment classification model to classify the polarity of the entire tuple. Despite our method being in a mid-rank position in the SemEval competition, we show that the question-answering approach can achieve good coverage retrieving sentiment tuples, allowing room for improvements in the technique.

1 Introduction

Opinions abound on the Internet nowadays. They are a valuable source of information since people often rely on them for making purchases. Companies can also benefit from this vast amount of opinions, as they do not need to conduct opinion polls or focus groups to measure the acceptance of a particular product (Liu, 2011). The large volume of opinions available becomes hard for humans to process. This leads to the study of ways of automating the processing of opinions, in order to summarize them.

Sentiment Analysis is the field of study which aims at processing the information conveyed by unstructured texts, providing structured information that facilitates the understanding of the opinions, attitudes, or emotions towards a particular entity (Liu, 2011). Sentiment Analysis can be performed at different levels of granularity (entire review, sentence, or aspect). Aspect-Based Sentiment Analysis (ABSA) aims to identify and rate the features (or aspects) of the entity being evaluated. Typically, ABSA involves the following phases: (i) identify and extract entities in reviews; (ii) identify and extract the aspects of an entity; (iii) cluster similar aspects; and (iv) determine the polarity of the sentiment over the entities and the aspects. Most of the research in Sentiment Analysis focuses on solving only one of these phases at a time.

Task 10 in SemEval 2022 – Structured Sentiment Analysis (Barnes et al., 2022) proposes a new approach to tackle the Sentiment Analysis problem, where the elements that constitute an opinion are identified together, in a structured way through a graph. Thus, Structured Sentiment Analysis can be seen as an information extraction task since we want to find the text spans where opinions about a particular feature are expressed (Barnes et al., 2021).

In this paper, we describe UFRGSent, a multilingual approach that relies on a question answering system to find the elements of an opinion present in review texts and a fine-tuned model to classify the polarity of the sentiment tuple. Our average results ranked 20^{th} out of 31 participating systems. Nevertheless, we believe there is room for improvement in our technique.

2 Background and Related Work

An opinion can be defined as a tuple $O = (h, t, e, p)$, where h represents the opinion holder (person who emits the opinion), t is the aspect target of the entity being reviewed, e is the opinion that is being expressed, and p is the sentiment related to the aspect expressed on the review (Liu, 2012).

While many works treat each opinion component separately, some approaches extract them all together, taking advantage of the components being interconnected. Graph neural networks (Barnes et al., 2021; Qian et al., 2021), transition-based neural models (Zhang et al., 2019), and multi-task learning (Chen and Qian, 2020) can be used to accomplish this task. There are also works applying co-extraction to find correlated opinion com-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1360 - 1365
July 14-15, 2022 ©2022 Association for Computational Linguistics

ponents, such as aspect words and corresponding polarities (Luo et al., 2019; He et al., 2019), aspect and opinion terms (Wu et al., 2020), or aspects, opinions, and polarities (Wang et al., 2017; Chen et al., 2021).

Question Answering is a challenging and well-studied problem in the Natural Language Processing field, gaining attention in the last years due to the use of pre-trained Language Models. One of the most popular subtasks of question answering is the Machine Reading Comprehension task. This task consists of, from a text piece (also known as context) and a question, finding the answer to the question in context (Zeng et al., 2020). Chen et al. (2021) proposed using a machine reading comprehension system to solve the problem of aspect sentiment triplet extraction. They use three-turn questions to extract aspects and opinions (in the first two turns) and sentiments (in the last turn), achieving state-of-the-art performance on standard Aspect-Based Sentiment Analysis (ABSA) datasets.

3 UFRGSent

3.1 Task Description

The Structured Sentiment Analysis task consists in identifying a sentiment graph from a review text r. Such graph can be seen as tuple $t = (h, t, e, p)$, composed by the the holder (h), the target (t), the opinion expression (e), and the sentiment polarity (p). Components h, t, and p are text spans over r, while $s \in \{\text{Positive}, \text{Negative}, \text{Neutral}\}$. A review r can contain none or multiple sentiment tuples.

3.2 Solution Overview

An overview of UFRGSent can be seen in Figure 1. A two-phase process was employed to identify sentiment tuples in a review text r. First, we use a pre-trained question answering model fine-tuned with opinionated texts in order to identify the spans in r that correspond to holder, targets, and opinion expressions. Next, we use another pre-trained aspect sentiment classification model fine-tuned on the training datasets, in order to predict the polarity of each extracted sentiment tuple.

The extraction of sentiment tuples from the review text was made using a question answering model. We iteratively submit three kinds of questions to the model, in order to extract candidates to sentiment tuples.

Tier 1 Questions: the following questions were the first questions posed to the model. As shown, each of the questions allows us to obtain one component of the sentiment tuple.

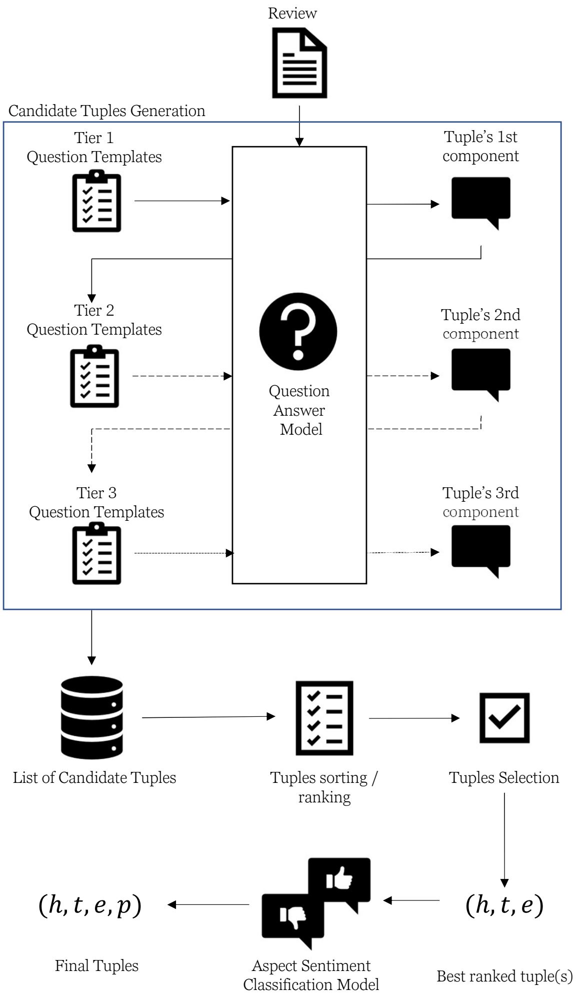

Figure 1: Overview of UFRGSent

Who is the holder?
Answer: $(h, \_\_, \_\_)$
What is the aspect expression?
Answer: $(\_\_, t, \_\_)$
What is the opinion word?
Answer: $(\_\_, \_\_, e)$

We extract the n most likely answers predicted by the model for each question. The model can also predict that the question has no answer on r. When this happens, we interpret it as the absence of sentiment in the sentence and produce the null tuple $(\_\_, \_\_, \_\_)$ as a candidate.

Tier 2 Questions: after the first extraction iteration, we obtain a list of candidate tuples containing just one component. This step aims to extract the second tuple component based on the existing one. These are the templates of questions used in this step, and the candidate tuples that were generated in this phase.

`Who has an opinion about <target>?`
Answer: $(h, t, \_\_)$
`What is the feeling about <target>?`
Answer: $(\_\_, t, e)$
`What is <holder> opining about?`
Answer: $(h, t, \_\_)$
`What is the opinion expressed by <holder>?`
Answer: $(h, \_\_, e)$
`What is <opinion> about?`
Answer: $(\_\_, t, e)$
`Who thinks <opinion>?`
Answer: $(h, \_\_, e)$

We do not use the null tuple in this iteration. If the QA Model returns that a question has no answer in this phase, we create a tuple without the remaining components, which will not be used in the next phase.

Tier 3 Questions: finally, we use the candidate tuples obtained in the previous step, with two components of the tuple, to obtain the remaining expression. These are the templates of questions for this step.

```
How <holder> feels about <target>?
Who thinks <target> is <opinion>?
What <holder> expressed <opinion>
about?
```

There are two possible outcomes for this phase – a complete tuple (h, t, e) or a tuple with two components, for the cases in which the question produces no answer.

Candidate Ranking and Selection: At the end of the Tier 3 questions, UFRGSent produces a list of candidate tuples, containing all answers generated by the iterative procedure previously described. The next step is ranking these tuples according to some criteria. Finally, based on the final ranking, we select the top-k answers to find the subset of tuples that best represent the structured sentiment on that review. If the null tuple were selected as the best answer in this phase, we conclude that the review has no sentiment.

Aspect Sentiment Classification: After determining the first three components of the sentiment tuple, we use an aspect sentiment classification model. This model receives the review and the aspect-phrase as inputs and outputs the polarity of the aspect in the review.

4 Experimental Setup

4.1 Question Answering Model

We fine-tuned BERT multilingual (Devlin et al., 2019) for the Question Answering task using the training data provided. To convert the original data into a question-answer dataset, we employed the following technique: Each sentence becomes a context. For each sentence, we generate the questions following the templates presented in Section 3.2. If a sentence does not have any sentiment annotation, we only generate Tier-1 questions with the null answer. Otherwise, we generate the three tiers of questions for the sentences containing structured sentiment annotations. For the sentences whose tuples do not contain some component (*i.e.,* part of the sentiment graph is not in the sentence), we do not generate questions with the missing component. For example, we only include the question `Who has an opinion about <target>?` for the sentences that have the aspect annotations.

The QA model was fine-tuned for two epochs, using the script provided by HuggingFace Transformers (Wolf et al., 2020)[1]. We tested the final model over dev datasets provided in the task, obtaining an F-1 score of 74.68.

4.2 Aspect Sentiment Classification Model

For the polarity prediction task, we used LCF-BERT (Zeng et al., 2019), which is provided by PyABSA[2]. We used an existing trained model created over 14 datasets in two languages (English and Chinese) as our base model. The base model was fine-tuned for ten epochs on the training dataset. The tests over the dev datasets yielded an F1 score of 75.13% and an Accuracy of 88.54%. The model only accepts as input a review and an aspect. In the case of sentences which not contain the target component, we feed the model with the opinion expression. The holder information is not used in this task.

4.3 Ranking and Tuple Selection

Our team employed two simple heuristics to sort the candidate tuples and select the final tuples. To sort the candidate tuples, we ranked them according to the number of times that the tuple was generated by the question answering procedure. Since we make multiple questions, it is common for an answer to be generated many times.

Once the candidate tuples are sorted, we selected the most frequent answers as our final tuples. If the most frequent was the null answer, we assume

[1] https://github.com/huggingface/transformers/tree/main/examples/pytorch/question-answering

[2] https://github.com/yangheng95/PyABSA

the review has no sentiment. On the other hand, if our technique generated more than one answer, we prune the answer set by removing the occurrence of the null tuple and removing overlapping tuples, *i.e.,* the tuples that have conflicting spans over the review. For example, in the sentence "I love the food." the spans "food" and "the food" overlap. In that case, we only keep the first tuple that appears in the ranking.

4.4 Datasets

We evaluated UFRGSent using seven datasets, namely NoReC (Øvrelid et al., 2020) that contains professional reviews in Norwegian; MultiBooked_eu and MultiBooked_ca (Barnes et al., 2018) with hotel reviews in Basque and Catalan, respectively; OpeNER_en and OpeNER_es (Agerri et al., 2013) that contain hotel reviews in English and Spanish, respectively; MPQA (Wiebe et al., 2005) a dataset of news wires in English; and Darmstadt Service Reviews (Toprak et al., 2010) with English reviews from online universities. Table 1 shows statistics of the datasets.

Dataset	Lang	# sent	#h	#t	#e
NoReC	NO	11,437	1,128	8,923	11,115
MultiBooked_eu	EU	1,521	296	1,775	2,328
MultiBooked_ca	CA	1,678	235	2,336	2,756
OpeNER_es	ES	2,057	255	3,980	4,388
OpeNER_en	EN	2,494	413	3,850	4,150
MPQA	EN	10,048	2,279	2,452	2,814
Darmstadt_unis	EN	2,803	86	1,119	1,119

Table 1: Statistics of Datasets (obtained from `https://github.com/jerbarnes/semeval22_structured_sentiment`).

4.5 Evaluation Metric

The evaluation metric used to assess the quality of the participating systems was the Sentiment Graph F_1 (Barnes et al., 2021). This metric evaluates an entire tuple (h, t, e, p). A true positive is an exact match between the predicted and golden graphs, weighting the overlaps of the spans for each tuple component, averaged across the three spans. *Precision* is calculated by weighting the number of correctly predicted tokens divided by the total predicted tokens, while *Recall* is the ratio between the number of correctly predicted tokens and the number of golden tokens.

5 Results

Table 2 shows the official results of UFRGSent. We varied two parameters of our method during the runs – the types of questions used in the question-answering model to extract tuple candidates and the number of answers retrieved for each question. The run that achieved the best average result uses the target and opinion questions and one answer to generate the candidate tuples. The same result was obtained when we generated five or ten answers. This run produced the best result in four out of seven datasets – MultiBooked_ca, NoReC, OpeNER_es, and OpeNER_en.

Considering the Darmstadt_unis dataset, the best result was when we just considered the aspect questions and generated one answer. On the other hand, the best results for MPQA and MultiBooked_eu datasets were obtained using the three types of questions and one, five, or ten answers.

We noticed a significant loss in performance when generating three answers. We conclude that this happened because the first answer generated is the expected response most of the time. Producing additional answers tends to introduce noise in the candidate tuples. However, increasing the number of answers makes the correct answer be generated more times, yielding the right tuple choice.

In comparison with other participants, our average score was the 20^{th} result out of 31 participating systems. The dataset in which we achieved the best rank was MultiBooked_ca (15^{th}), while our worst results were Darmstadt_unis and MPQA datasets (20^{th} out of 31). Although our team did not submit results for the cross-lingual task, we emphasize that our solution is entirely multilingual – there is only one model, which extracted sentiment tuples for all datasets simultaneously. Therefore, our solution can be extended to any other language among the 104 languages present in multilingual BERT.

In order to assess the quality of our candidate tuple extraction, we measured the coverage of the question-answering results (*i.e.,* the percentage of gold tuples present in the set of candidate tuples). This measurement was done on the development dataset, for which we know beforehand the gold sentiment graphs. The results of the experiment can be seen in Table 3.

We set the hyper-parameter k to one, extracting only one answer per question. The evaluation was made in two ways – the exact match between gold tuple and candidates and the overlap, in which a pair of tuples containing an overlap between their tokens is considered a correct match. The experiment was repeated, varying the type of questions

Configuration		Dataset							Avg. Score
Question Types	# Ans	Darmstadt_unis	MPQA	MultiBooked_ca	MultiBooked_eu	NoReC	OpeNER_en	OpeNER_es	
H - T - E	1	0.230	**0.232**	0.505	**0.467**	0.251	0.431	0.399	0.359
	3	0.061	0.071	0.217	0.226	0.088	0.151	0.133	0.135
	5	0.230	**0.232**	0.505	**0.467**	0.251	0.431	0.399	0.359
	10	0.005	**0.232**	0.505	**0.467**	0.251	0.431	0.399	0.327
T - E	1	0.242 (20)	0.217 (20)	**0.521 (15)**	0.463 (17)	**0.270 (19)**	**0.452 (18)**	**0.427 (19)**	**0.370 (20)**
	3	0.082	0.042	0.284	0.286	0.135	0.232	0.204	0.18
	5⋆	0.242	0.217	**0.521**	0.463	**0.270**	**0.452**	**0.427**	**0.370**
	10⋆	0.242	0.217	**0.521**	0.463	**0.270**	**0.452**	**0.427**	**0.370**
T	1	**0.283**	0.231	0.456	0.374	0.241	0.399	0.338	0.332
	3	0.007	0.009	0.050	0.038	0.019	0.041	0.035	0.029
	5⋆	0.283	0.231	0.456	0.374	0.241	0.399	0.338	0.029
	10⋆	0.283	0.231	0.456	0.374	0.241	0.399	0.338	0.029
E	1	0.244	0.206	0.486	0.459	0.252	0.417	0.383	0.35
	3	0.029	0.011	0.141	0.148	0.100	0.063	0.086	0.083
	5⋆	0.244	0.206	0.486	0.459	0.252	0.417	0.383	0.35
	10⋆	0.244	0.206	0.486	0.459	0.252	0.417	0.383	0.35

Table 2: Official Results of UFRGSent in terms of Sentiment Graph F_1. Best results for a given dataset are in bold. ⋆ denotes that the results were obtained in the post-evaluation phase. The numbers between parentheses indicate the position achieved by UFRGSent in the competition.

Dataset	H - T - E		T		E		H	
	Exact	Overlap	Exact	Overlap	Exact	Overlap	Exact	Overlap
Darmstadt_unis	65.7%	76.2%	58.9%	67.7%	56.9%	67.3%	58.9%	63.3%
MPQA	75.1%	87.3%	72.2%	82.9%	71.2%	82.5%	72.7%	83.9%
MultiBooked_ca	41.1%	77.9%	30.2%	61.1%	32.6%	64.9%	10.9%	24.9%
MultiBooked_eu	42.1%	77.4%	29.8%	58.3%	33.6%	58.7%	13.6%	33.2%
NoReC	42.0%	79.7%	33.1%	64.9%	31.9%	66.3%	32.4%	42.2%
OpeNER_en	37.9%	76.5%	28.6%	57.2%	29.7%	58.5%	12.2%	42.8%
OpeNER_es	35.8%	72.8%	27.7%	50.4%	27.4%	57.8%	0.1%	26.0%

Table 3: Coverage of Extraction for Question-Answering System

used to extract the candidates. We first use the three types of questions together, and then we evaluate the coverage achieved by each type of question individually.

The results show that our question-answering technique provides good coverage of sentiment tuples. For darmstadt_unis and mpqa, we have an exact match with over 65% coverage. We also improved the measure on datasets with less coverage in the exact match experiments, considering tuple overlap. All datasets achieved at least 70% coverage in the overlap experiments.

Using only one component to extract candidate tuples reduces coverage for the question answering for all datasets. While the mpqa dataset was less affected by removing the components (loss of 2.9% in coverage using targets, 3.9% using opinion expressions, and 2.4% using holder questions). Other datasets had significant drops in coverage, especially when just holder questions were considered.

6 Conclusion

In this paper, we described our system submitted to SemEval-2022 Task 10. We designed a multilingual approach that relies on a QA system and an ASC model to find the Sentiment Graphs. The key idea is that the joint and incremental extraction of holder, target, and opinion helps to achieve a good coverage for UFRGSent.

In these preliminary experiments, we could not establish the quality of our tuple ranking and selection methods and we leave it for future work. Additionally, we are interested in understanding how well our multilingual model performs against a monolingual version of our technique, and how other state-of-the-art QA models can improve the extraction of sentiment tuples.

Acknowledgements.

This work was partially supported by CNPq/Brazil and by CAPES Finance Code 001.

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.

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)*.

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.

Jeremy Barnes, Andrey Kutuzov, Laura Ana Maria Oberländer, 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)*.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. *arXiv preprint arXiv:2103.07665*.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515.

Bing Liu. 2011. Opinion mining and sentiment analysis. In *Web data mining: exploring hyperlinks, contents, and usage data*, 2 edition, chapter 11. Springer Science & Business Media.

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

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: Dual cross-shared RNN for aspect term-polarity co-extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 591–601.

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.

Yong Qian, Zhongqing Wang, Rong Xiao, Chen Chen, and Haihong Tang. 2021. SGPT: Semantic graphs based pre-training for aspect-based sentiment analysis. *arXiv preprint arXiv:2105.12305*.

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.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms.

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, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. LCF: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16).

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21).

Meishan Zhang, Qiansheng Wang, and Guohong Fu. 2019. End-to-end neural opinion extraction with a transition-based model. *Information Systems*, 80:56–63.