Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Review Reading Comprehension and Aspect Based Sentiment Analysis in E-Commerce Reviews

Massimo Ruffolo, Ermelinda Oro, Francesco Visalli, Mariella Pupo, Francesco Luppino, Fausto Pupo

RT-ICAR-CS-19-03 **Agosto 2019**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 8-9C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.icar.cnr.it*
– Sezione di Palermo, Via Ugo La Malfa, 153, 90146 Palermo, URL: *www.icar.cnr.it*

REVIEW READING COMPREHENSION AND ASPECT BASED SENTIMENT ANALYSIS IN E-COMMERCE REVIEWS *

A PREPRINT

Massimo Ruffolo
High Performance Computing and Networking Institute - National Research Counsil
Via Pietro Bucci 8/9C Rende (CS), 87036, Italy
massimo.ruffolo@icar.cnr.it

Ermelind Oro
High Performance Computing and Networking Institute - National Research Counsil
Via Pietro Bucci 8/9C Rende (CS), 87036, Italy
linda.oro@icar.cnr.it

Francesco Visalli
High Performance Computing and Networking Institute - National Research Counsil
Via Pietro Bucci 8/9C Rende (CS), 87036, Italy
francesco.visalli@icar.cnr.it

Mariella Pupo
Altilia srl
TechNest Unical Piazza Vermicelli Rende (CS), 87036, Italy
mariella.pupo@altiliagroup.com

Francesco Luppino
Altilia srl
TechNest Unical Piazza Vermicelli Rende (CS), 87036, Italy
francesco.luppino@altiliagroup.com

Fausto Pupo
Altilia srl
TechNest Unical Piazza Vermicelli Rende (CS), 87036, Italy
fausto.pupo@altiliagroup.com

August 6, 2019

ABSTRACT

E-commerce has become one of the most used purchasing method. Because of this, it is important for customers to quickly understand features of the items they want to buy, and to know opinions of other customers that have already purchased such items. This problem can be recasted as an aspect based sentiment analysis task on product reviews. In order to address the problem, we leveraged BERT, a model to create strong contextual word embeddings. We put ourselves in a real-case scenario of laptop reviews, extracted from the web. We cleaned the data and built the datasets needed for the training of BERT models. In particular, to annotate the fine tuning datasets we used a custom approach based on weak supervision. Weak supervision techniques are becoming more and more

*This work has been supported by POR-CALABRIA

interesting as they allow to alleviate the cost of human annotation that is a major issue in supervised learning.

Keywords Aspect Based Sentiment Analysis · BERT · Weak Supervision · Machine Learning · Deep Learning

1 Introduction

Due to the easy way of finding products and seeing them delivered directly to home and thanks also to the convenience of prices which are often lower than those of physical stores, online commerce (or e-commerce) has become one of the most used purchasing method. Given the huge amount of products that can be found in online stores, it is important for customers to quickly understand features of the items they want to buy, and to know opinions of other customers that have already purchased such items.

The problem of extracting relevant features along with buyers opinions from product reviews available online can be formulated as an aspect based sentiment analysis (ABSA) task. Sentiment analysis is a natural language processing (NLP) problem that mainly aims at detecting the overall polarity (e.g. positive, negative or neutral) of a text. ABSA is a more fine-grained approach to sentiment analysis. Two main subtasks of ABSA are aspect extraction (AE) and aspect sentiment calssification (ASC). AE takes care of finding aspects (e.g. the GPU of a laptop) in pieces of text (the reviews), whereas ASC aims to identify the polarity of such aspects (e.g. negative for GPU).

There are several approaches to AE. Frequency-based methods rely on the concept that only a limited set of words (usually single nouns or compound nouns) appear frequently in a review. These frequent words are considered aspects. It is a simple method, very used in early AE systems. The two main drawbacks are that not all the frequent words are really aspects and that aspects are not only frequent words. The most familiar method using frequency-based approach is [1].

In syntax-based methods (e.g. [2]), aspects are found by way of the syntactical relations they are in. A very simple rule is the association between an aspect and a sentiment word (e.g. good battery), called adjectival modifier relation. These methods make it possible to identify less frequent aspects, on the other hand many syntactical relations are needed.

Machine learning methods explored both the supervised and unsupervised approach (e.g. [3] for supervised learning and [4] for unsupervised). Since they need handly written features, supervised learning methods didn't use to be common, it was preferred to use other methods based on features (e.g. frequency-based). The most common unsupervised apporach used is Latent Dirichlet Allocation (LDA) [5], that is a topic modeling technique that generates topics based on word frequency from a set of documents. LDA is useful for finding topics within documents. The problem here is that the generated topic are unlabeled, so there is not a direct connection between topics and aspects.

Regarding ASC task, the first approaches were based on a sentiment dictionary. For example, [6] propagate the known sentiment of a small set of words through the WordNet synonym/antonym graph. Each adjective, present both in the dictionary and in the sentence, votes for a sentiment. The sentence polarity is assigned using majority voting and then is propagated to all the aspects of the sentence. When the number of positive and negative aspects is the same, the majority voting is done on the single aspects (based on the nearness of the adjectives to the aspects). In this case, there could be multiple polarities within the same sentence.

As described for AE, supervised machine learning methods were based on handly written features. The most used for ASC were lexicon informations (e.g. [7]). Unsupervised approach have also been used. In [8], the aspects are used for searching sentiment sentences in the neighborhood of the current sentence (using parsing syntact dependencies). The polarity is determined using an unsupervised technique called relaxation labeling. Anyhow, these unsupervised methods require lot of data that were not always available.

Many approaches have been proposed to combine AE and ASC. Due to the fact that they learn most important features by themselves, deep learning architectures have become the state of the art for many NLP tasks (including AE and ASC). These models are often hungry for data, as they are rich of parameters. One of the critical problem to face off trying to do ABSA in a supervised way is the lack of annotated data. These datasets are annotated by subject matter experts. Alleviating the cost of human annotation is a major issue in supervised learning. One of the methods to deal with annotation problem is called weak supervision. In weak supervision labels are assigned directly by humans (both expert and non-expert of the domain), that have no knowledge about specific machine learning algorithms, by using simple techniques that allow to automate data labeling.

In this paper we leverage most recent deep learning models for NLP, in combination with weak supervision techniques, to face the ABSA problem on a real world use-case scenarios. As working example we make use of reviews about laptops extracted from the web. We have chosen laptops because these products have a complex feature set and are one of the most purchased online. Furthermore, some annotated datasets related to ABSA on laptops are already available [9] in literature. This way, we have the opportunity to compare our weak-supervised ABSA method with existing literature. For the purpose of describing our ABSA method we use a weak-supervision approach that enable to label first relevant features for AE and second sentiment expressions for ASC. In order to verify the effectiveness of our weak-supervision approach we perform tests by using Bidirectional Encoder Representations from Transformers

(BERT) [10] architecture and follow the training procedures described in BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis (BERT for RRC and ABSA) [11] that performs the state of the art on SemEval-2014 Task 4 [9]. It proposes the ABSA task on some dataset of product reviews, one of which is composed by laptop reviews.
From the experiments carried out we can state that BERT is a very powerful tool for performing ABSA tasks. We confirm the state of the art reached in [11], fine-tuning a BERT model post-trained on our custom reviews datasets. Moreover, the results of the experiments on our fine-tuning datasets, annotated by our weak supervision mechanism are resulted interesting and need future insights.
In Section 2 are presented a background of the architecture along with the approach used in this work, comprising reviews extraction from the web, our weak supervision mechanism, data construction and models description. In Section 3 it is discussed the experimental phase and they are showed the results obtained from the experiments on the post-training and fine tuning phases, along with a discussion of such results.

2 The Approach

The proposed approach is composed by 5 phases: *web data extraction, data cleaning, model post-training, weak supervised data annotation, model fine-tuning*. In the following we describe each phase just listed along with the background about the adopted model. We use the laptop use case as working example.

2.1 Background: Model Learning by BERT

BERT is mainly a model to create strong contextual word embeddings but it is also a generic architecture for many NLP tasks. BERT models are pre-trained on large corpus in order to learn word embeddings and then are fine-tuned for specific NLP problems. In its pre-training phase, BERT introduces two new tasks: the masked language model (MLM) and the next sentence prediction (NSP), that allow an embedding to learn both from its left and right context. MLM is inspired by the Cloze task [12], it randomly masks some of the tokens of the input and let the model learn to predict the original vocabulary word based only on its context. NSP let the model learn contextual representation beyond the word-level. On the top of the pre-trained model it can be easily built a new model for specific NLP tasks. The new model just needs to be fine-tuned.
BERT is based on the Transformer architecture [13]. It is released in two main variant: BERT base and BERT large. The differences between the two concern the number of Transformer blocks, the hidden size, the number of attention heads and the number of parameters, respectively 12, 768, 12, 110M for the former and 24, 1024, 16, 340M for the latter. The releases differ also in the type of vocabulary used (i.e. cased or uncased) during the pre-training phase (Section 2.3).
BERT is pre-trained on Wikipedia and BooksCorpus dataset [14], it knows nothing about laptops and needs to learn domain specific word embeddings. BERT for RRC and ABSA introduces a new phase after the pre-training and before the fine tuning. In this step, called post-training, it is injected into BERT model the domain knowledge. Moreover, the original work proposes a new task called Review Reading Comprehension (RRC), it is similar to Machine Reading Comprehension [15, 16] but based on reviews of products. Therefore, during the post-training phase, BERT is trained on unannotated reviews dataset by doing MLM and NSP in order to inject the domain knowledge and, at the same time, it is trained on a famous labeled dataset of question answering (Stanford Question Answering Dataset - SQuAD - v1.1 [15]) to inject the task knowledge. Both the knowledge contribute to the achievement of the state of the art for ABSA (for more details we refer to the original paper [11]). The model resulted from the post training phase is fine finally tuned in a supervised way.

2.2 Web Data Extraction and Data Cleaning

In order to inject domain specific knowledge in BERT, we extracted data from the web online retailer web site using web scraping methods.

2.3 Post-Training of the Model

In order to inform BERT about laptops world, as previously mentioned, a post-training phase is required. In this step a pre-trained BERT model (a model that already knows contextual word embeddings out of domain but that is not fine-tuned yet) performs MLM and NSP on a large unlabeled laptop dataset. The MLM task is the same described in BERT original paper and mentioned in 2.1, whereas the NSP task concerns to predict if two sentences belong to the same review or not.
A post-training example has the form as shown in Figure 1 and it is ready to be given in input to a BERT model.

$t1, t2, ..., t_n$ and $t_{n+1}, t_{n+2}, ..., t_m$ are tokens belonging to the first and second sentence respectively. `[CLS]`, `[MASK]`, and `[SEP]` are special tokens used in BERT. The `[CLS]` token is used for sentence classification task during fine tuning phase. The `[MASK]` tokens are the masked ones that have to be predicted to perform the MLM task. The `[SEP]` tokens are used to separate two sentences and to indicate the end of the input.
Tokens of the sentences are output by WordPiece algorithm [17], that is based on a 30522 tokens vocabulary. When a BERT model sees a symbol out of vocabulary, this is splitted in tokens following the WordPiece alghoritm (everytime the alghoritm can do it, otherwise the `[UNK]` token is assigned to the symbol) and so also its embedding. Therefore, it is important that the vocabulary contains the domain words. BERT reservs 994 tokens for this purpose.

`[CLS]` t_1 t_2 `[MASK]` ... t_n `[SEP]` t_{n+1} t_{n+2} `[MASK]` ... `[MASK]` t_m `[SEP]`

Figure 1: Example of BERT input representation for the domain knowledge of post-training phase.

2.4 Dataset Annotation by Weak Supervision

See paper M. Ruffolo, E. Oro, et al. "A Weak Supervision Data Annotation Method for Large Scale Natural Language Understanding Tasks".

2.5 Fine Tuning of the Model

In fine tuning phase, the post-trained models are specialized for downstream language tasks. When the AE task is addressed in a supervised way, it is usually casted as a classification problem at token level, where each token of the sentence has to be labeled within the set {*Begin, Inside, Outside*}. If a token is labeled with *Begin*, it means that token is an aspect. If it is followed by some tokens labeled with *Inside*, that *Begin*-labeled token is the beginning of an aspect composed by more than one word (the *Inside*-labeled tokens), otherwise it is a single word aspect. The tokens labeled with *Outside* are out of the aspect. The classification is done applying a fully connected layer and a softmax for each position of the sentence. The input of this step are the embedded tokens get from the post-training phase (Figure 2).
The ASC task is a sentence level classification problem. It is a subsequent task of AE, that is, it is needed to assign a polarity to every pair (*sentence*, *aspect*), for each aspect identified in the previous step. The polarity is within the set {*positive, negative, neutral*}. Softmax is applied along the labels dimension on `[CLS]` (see Section 2.3). An example of the input of the classifier is in Figure 3. $t1, ..., t_n$ are tokens belonging to the aspect, while $t_{n+1}, ..., t_m$ belong to the sentence.

3 Experiments

In this section we first describe the datasets and the experimental settings used for the post-training and the fine tuning phases. Then, we present the experiments done and the results obtained with a discussion of these.

3.1 Dataset and Experimental Settings

Within the research project called *"APPIA: The dynamic Altilia Price, Product and market Intelligence Advisor"*, we built two dataset, one for the domain knowledge of the post-training phase, and another for the ASC task, used in the fine tuning phase, which we call APPIA-ASC.
The domain knowledge dataset was built starting from the laptop reviews extracted from the web. We took 127843 reviews from a total of 130210, the rest were reserved for fine-tuning datasets to avoid bias problem. We splitted our reviews into 689012 sentences from which we generated 1055552 post-training examples. Moreover, in the post-training phase, we leverage the original SQuAD v1.1 [15] dataset (that contains 87599 training examples from 442 Wikipedia articles) for the task knowledge.
In order to compare our weak-supervised ABSA method, we trained ASC classifiers on SemEval 2014 Task 4 Subtask

`[CLS]` t_1 t_2 ... t_{n-1} t_n `[SEP]`

Figure 2: Example of BERT input representation for the AE of fine tuning phase.

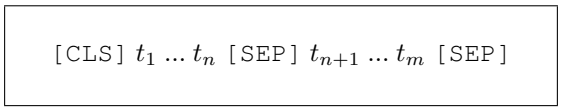

Figure 3: Example of BERT input representation for the ASC of fine tuning phase.

2 [9] (SemEval from now on), a famous ABSA dataset in literature.
APPIA-ASC was built starting from the laptop reviews that weren't used for the domain knowledge dataset. It was annotated by our weak supervision mechanism and has the same cardinality as SemEval. We have also maintained the same percentage of positive, negative and neutral examples.
We used the hyperparameters suggested in the original paper. We fixed the maximum length of post-training examples to 320 and used a batch size of 16 both for task and domain knowledge. We adopted the Adam algorithm [18] as optimizer for the gradient descent with a learning rate of 3e-5. We did 70000 training steps that approximately correspond to an entire pass on the domain knowledge dataset.
We fine-tuned our models for 4 epochs fixing the maximum sequence length to 100. We used a batch size of 32 and the Adam optimizer with a learning rate of 3e-5.

3.2 Post-Training

To let BERT learn word domain embeddings properly, we calculated the most common words (using TF) in our laptop reviews and replaced the unused tokens in the input WordPiece vocabulary with the top 994 words, ordered by term frequency.
Starting from the pre-trained BERT base uncased model (Section 2.1), we did a post-training based on the SQuAD v1.1 dataset for the task knowledge and on our one million examples dataset (the one built from the laptop reviews extracted from the web) for the domain knowledge. We call this model APPIA-PT.
At the same time, in order to verify the behavior of the loss function, we post-trained another model on the same task knowledge dataset but leveraging the domain knowledge dataset used in the original work. Let's call this BERT-PT. Since the number of examples is approximately the same, we used the same hyperparameters for both trainings. The models coverge to the same loss value, oscillating between 1.5 and 2.

3.3 Fine Tuning

In this step, the post-trained models are fine-tuned for the specific NLP problems. In Table 1 the results of the fine tuning phase of the ASC task (starting from the APPIA-PT model) are showed. These are reported as averages of 10 runs (10 different random seeds for random batch generation) and show the performance on the APPIA-ASC and SemEval datasets (including cross-comparison between the training set and the test set). For example the value of 70.24 is the Accuracy obtained using the APPIA-ASC training set and the SemEval test set.
The results on the uncrossed datasets are excellent. We confirm the state of the art for the SemEval-2014 Task 4 Subtask 2, claimed in the original work. Better results are obtained on the uncrossed APPIA-ASC datasets, probably because the data are annotated in such a way that the network can learn it more easily. More interesting results are obtained looking to the cross-comparison between APPIA-ASC and SemEval. It is important to emphasize that APPIA-ASC was automatically annotated, without any human involvement during the labeling process. The discrepancy between the comparisons is due to the fact that the datasets were annotated differently. In fact, while APPIA-ASC was automatically labelled, SemEval was annotated by experienced human annotators[2].
In Table 2 we present the results of the fine tuning phase of the ASC task, starting from the BERT-PT model. It can be seen how the results are almost the same of the classifier obtained by fine tuning APPIA-PT. Therefore, the fine tuning step is totally independent of the post-training datasets. This underlines how the post-training phase was carried out correctly, but most of all it is a sign of the large expressiveness of the BERT model.
Regarding the AE task, more experiments are needed.

4 Conclusions

In this work we addressed the problem of extracting relevant features along with buyers opinions from laptop reviews available online as an ABSA task. In order to do this, we extracted and cleaned laptop reviews from the web with whom we built datasets for BERT models. We presented a custom weak supervision meachanism used for annotating the

[2]https://alt.qcri.org/semeval2014/task4/

Table 1: Results of ASC in Accuracy and Macro-F1(MF1), using APPIA-PT.

		Test Set			
		APPIA-ASC		SemEval	
		Acc.	MF1	Acc.	MF1
Training Set	APPIA-ASC	87.31	86.09	70.24	67.55
	SemEval	70.17	66.66	78.13	75.05

Table 2: Results of ASC in Accuracy and Macro-F1(MF1), using BERT-PT.

		Test Set			
		APPIA-ASC		SemEval	
		Acc.	MF1	Acc.	MF1
Training Set	APPIA-ASC	87.56	86.26	70.08	67.72
	SemEval	69.51	66.35	78.07[1]	75.08[1]

[1] Claimed in the original paper.

datasets for the fine tuning phase.
We confirm the state of the art on SemEval-2014 Task 4 Subtask 2 using APPIA-PT, the model post-trained on our custom dataset of laptop reviews extracted from the web. The results of the experiments suggest that our weak supervision mechanism works well. However, more future experiments are needed to undestand and bridge the distance between the way of labeling of human annotators and automated processes as weak supervision techniques.

References

[1] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In Deborah L. McGuinness and George Ferguson, editors, *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*, pages 755–760. AAAI Press / The MIT Press, 2004.

[2] Yanyan Zhao, Bing Qin, Shen Hu, and Ting Liu. Generalizing syntactic structures for product attribute candidate extraction. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 377–380. The Association for Computational Linguistics, 2010.

[3] Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1035–1045. ACL, 2010.

[4] Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya, and Srujana Merugu. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, pages 498–509. SIAM / Omnipress, 2011.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM, 2004.

[7] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, and Chun Jin. Red opal: product-feature scoring from reviews. In Jeffrey K. MacKie-Mason, David C. Parkes, and Paul Resnick, editors, *Proceedings 8th ACM Conference on Electronic Commerce (EC-2007), San Diego, California, USA, June 11-15, 2007*, pages 182–191. ACM, 2007.

[8] Ana-Maria Popescu, Bao Nguyen, and Oren Etzioni. OPINE: extracting product features and opinions from reviews. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 32–33. The Association for Computational Linguistics, 2005.

[9] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014.*, pages 27–35. The Association for Computer Linguistics, 2014.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[11] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2324–2335. Association for Computational Linguistics, 2019.

[12] Wilson L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017.

[14] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society, 2015.

[15] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016.

[16] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics, 2018.

[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.