

A Pilot Study of Domain Adaptation Effect for Neural Abstractive Summarization

Xinyu Hua and Lu Wang

College of Computer and Information Science

Northeastern University

Boston, MA 02115

hua.x@husky.neu.edu luwang@ccs.neu.edu

Abstract

We study the problem of domain adaptation for neural abstractive summarization. We make initial efforts in investigating what information can be transferred to a new domain. Experimental results on news stories and opinion articles indicate that neural summarization model benefits from pre-training based on extractive summaries. We also find that the combination of in-domain and out-of-domain setup yields better summaries when in-domain data is insufficient. Further analysis shows that, the model is capable to select salient content even trained on out-of-domain data, but requires in-domain data to capture the style for a target domain.

1 Introduction

Recent text summarization research moves towards producing abstractive summaries, which better emulates human summarization process and produces more concise summaries (Nenkova et al., 2011). Built on the success of sequence-to-sequence learning with encoder-decoder neural networks (Bahdanau et al., 2014), there has been growing interest in utilizing this framework for generating abstractive summaries (Rush et al., 2015; Wang and Ling, 2016; Takase et al., 2016; Nallapati et al., 2016; See et al., 2017). The end-to-end learning framework circumvents efforts in feature engineering and template construction as done in previous work (Ganesan et al., 2010; Wang and Cardie, 2013; Gerani et al., 2014; Pighin et al., 2014), by directly learning to detect summary-worthy content as well as generate fluent sentences.

Nevertheless, training such systems requires large amounts of labeled data, which creates a big hurdle for new domains where training data is scant and expensive to acquire. Consequently, we raise the following research questions:

Input (News):The Department of Defense has identified 441 American service members who have died since the start of the Iraq war. It confirmed the death of the following American yesterday: DAVIS, Raphael S., 24, specialist, Army National Guard; Tutwiler, Miss.; 223rd Engineer Battalion.

Abstract: Name of American newly confirmed dead in Iraq ; 441 American service members have died since start of war.

Input (Opinion): WHEN the 1999 United States Ryder Cup team trailed the Europeans, 10-6, going into Sunday’s 12 singles matches at the Country Club outside Boston, Ben Crenshaw, the United States captain, issued a declaration of confidence in his golfers. “I’m a big believer in faith,” Crenshaw said firmly in his Texas twang. “I have a good feeling about this.” The next day, Crenshaw’s cavalry won the firsts even singles matches. With a sudden 13-10 lead, the turnaround put unexpected pressure on the Europeans, . . .

Abstract: Dave Anderson Sports of The Times column discusses US team’s poor performance against Europe in Ryder Cup.

Figure 1: A snippet of sample news story and opinion article from The New York Times Annotated Corpus (Sandhaus, 2008).

- *domain adaptation*: whether we can leverage available out-of-domain abstracts or extractive summaries to help train a neural summarization system for a new domain?

- *transferable component*: what information is transferable and what are the limitations?

In this paper, we attempt to shed some light on the above questions by investigating neural summarization on two types of documents with major difference: news stories and opinion articles from The New York Times Annotated Corpus (Sandhaus, 2008). Sample articles and human written abstracts are shown in Figure 1. We select a reasonably simple task on generating short news summary for multi-paragraph documents.

Contributions. We first investigate the effect of parameter initialization via pre-training on extractive summaries. A large-scale dataset consisting of 1 million article-extract pairs is collected from The New York Times for use. Experimental results show that this step improves summarization performance measured by ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002).

We then treat news stories as source domain and opinion articles as target domain, and make initial tries for understanding the feasibility of domain adaptation. Importantly, by testing on opinion article summarization, the model leveraging data from both source and target domains yields better performance than in-domain trained model when in-domain training data is rare.

Furthermore, we interpret the learned model to understand what information is transferred to a new domain. In general, a model trained on out-of-domain data can learn to detect summary-worthy content, but may not match the generation style in the target domain. Concretely, we observe that the model trained on news domain pays similar amount of attention to summary-worthy content (i.e., words reused by human abstracts) when tested on news and opinion articles. On the other hand, human writers tend to employ new words unseen from the input when constructing opinion abstracts. End-to-end evaluation results imply that the model trained on out-of-domain data fails to capture this aspect.

The above observations suggest that the neural summarization model learns to 1) identify salient content, and 2) generate summaries with a style as in the training data. The first element might be transferable to a new domain, while not so much for the second.

2 The Neural Summarization Model

In this work, we choose the attentional sequence-to-sequence model with pointer-generator mechanism (See et al., 2017) for study. Briefly, the model learns to generate a sequence of tokens $\{y_i\}$ based on the following conditional probability: $p(y_i = w | y_1, \dots, y_{i-1}, x) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t$

Here $P_{vocab}(w)$ denotes the probability to generate a new word from vocabulary, p_{gen} is a learned parameter that chooses between generating and copying, depending on the hidden states and attention distribution. This model enhances

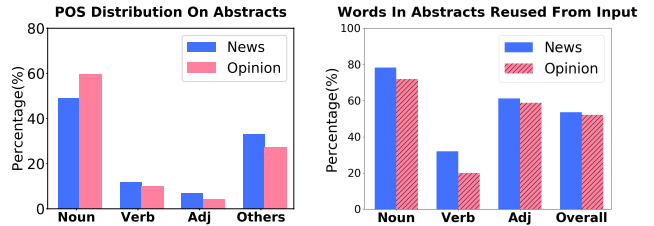


Figure 2: [Left] Part-of-speech (POS) distribution for words in abstracts. [Right] Percentage of words in abstracts that are reused from input, per POS and all words. OPINION abstracts generally reuse less words.

the original attention model (Bahdanau et al., 2014) by incorporating pointer-network (Vinyals et al., 2015), which allows the decoder to copy accurate information from input. Due to space limitation, we refer the readers to original paper (See et al., 2017) for model details.

For experiments, we employ bidirectional recurrent neural network (RNN) as encoder and unidirectional RNN as decoder, both implemented by Long Short Term Memory (LSTM) with 256 hidden units. Input and output data are lowercased as described in (See et al., 2017).

3 Datasets and Experimental Setup

Primary Data. Our primary data source is The New York Times Annotated Corpus (Sandhaus, 2008) (henceforth called NYT-annotated). Compared with other commonly used dataset for abstractive summarization, NYT-annotated has more variation in its abstracts, such as paraphrase and generalization. It also comes with other human labels we could use to characterize the type of articles. The whole dataset consists of 1.8 million articles, of which 650,000 are annotated with human constructed abstracts. Articles longer than 15 tokens and abstracts longer than 10 tokens are extracted for use in our study (as in Figure 1).

The resulting dataset are further separated into two types based on their taxonomy tags¹: NEWS stories and OPINION articles. We believe these two types of documents are different enough in terms of topics, summary style, and lexical level language use, that they could be treated as different domains for our study. We collected 100,824

¹The corpus comes with taxonomic classifiers tags. Articles with tag “News” are treated as news stories; for the rest, the ones with “Opinion”, “Editorial”, or “Features” are treated as opinion articles.

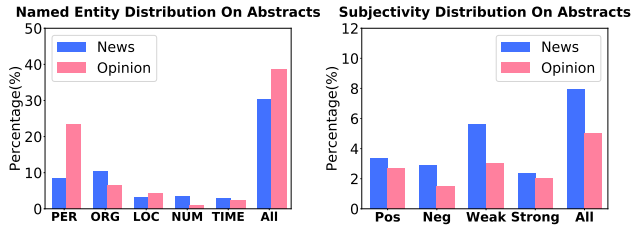


Figure 3: Named Entities distribution (left) and subjective words distribution (right) in abstracts. More PERSON, less ORGANIZATION, and less subjective words are observed in OPINION.

articles for NEWS which is treated as source domain, and 51,214 for OPINION as target domain. The average length for documents of NEWS is 680.8 tokens, and 785.6 tokens for OPINION. The average lengths for abstracts are 23.14 and 19.13 for NEWS and OPINION.

We also make use of the section tag, such as *Business*, *Sports*, *Arts*, to calculate the topic distribution for these two domains. About 57% of the documents of NEWS are about *Sports*, whereas more than 78% documents of OPINION are about *Arts*. We also observe different levels of subjectivity based on the percentage of strong subjective words taken from MPQA lexicon (Wilson et al., 2005). On average 4.1% of the tokens in OPINION articles are strong subjective, compared to 2.9% for NEWS stories. This shows the topics and word usage are essentially different between these two domains.

Characterizing Two Domains. Here we characterize the difference between NEWS and OPINION by analyzing the distribution of word types in abstracts and how often human reuse words from input text to construct the summaries. Overall, 81.3% of the words in NEWS abstracts are reused from input, compared with 75.8% for OPINION. The distribution for words of different part-of-speech is displayed on the left of Figure 2, which shows that there are relatively more Nouns in OPINION. In the same figure, we display the percentage of words in abstract that are reused from input, which suggests that human tends to reuse more nouns and verbs for NEWS abstracts. Furthermore, the distribution of Named Entities words and subjective words in abstracts are depicted in Figure 3.

Model Pre-training Dataset. We further collect lead paragraphs and article descriptions for

1,435,735 articles from The New York Times API². About 71% of these descriptions are the first sentences in the lead paragraphs, and thus can be considered as extractive summaries. About one million lead paragraph and description pairs are retained for pre-training³ (henceforth NYT-extract).

Training Setup. We randomly divide NYT-annotated into training (75%), validation (15%), and test (10%) for both news and opinion. Experiments are conducted with the following setups:

- 1) IN-DOMAIN: Training and testing are done in the same domain, for NEWS and OPINION;
- 2) OUT-OF-DOMAIN: training on source domain NEWS, and testing on target domain OPINION;
- 3) MIX-DOMAIN: training on source domain NEWS and then on target domain OPINION, and testing on OPINION. Training stops when the trend of loss function on validation set starts increasing.

Evaluation Metrics. We use automatic evaluation on recall-oriented ROUGE (Lin, 2004) and precision-oriented BLEU (Papineni et al., 2002). We consider ROUGE-2 which measures bigram recall, and ROUGE-L which takes into account the longest common subsequence. We also evaluate on BLEU which measures precision up to bigrams.

4 Results

Effect of Pre-training with Extracts. We first evaluate whether pre-training can improve summarization performance for IN-DOMAIN setups, where we initialize model parameters by training on NYT-extract for about 20,000 iterations. Otherwise, parameters are randomly initialized. Results are displayed in Table 1. We also consider two baselines, BASELINE1 outputs the first sentence, BASLINE2 selects the first 22 (news) and 15 (opinion) tokens (with similar lengths as human summaries).

As can be seen, the pre-training step improves performance for NEWS, whereas the performance on OPINION remains roughly the same. This might be due to the fact that news abstracts reuse more

²<https://developer.nytimes.com>

³Unsupervised language model (Ramachandran et al., 2016) can also be used for parameter initialization before our pre-training step. Here our goal is to allow the model to learn searching for summary-worthy content, in addition to grammaticality and language fluency.

words from input, which are closer to extractive summaries than opinion abstracts.

	R-2	R-L	BLEU	Avg Len
<i>Test on News</i>				
BASELINE1	23.5	35.4	19.9	28.94
BASELINE2	19.5	30.1	19.5	22.00
IN-DOMAIN	23.3	34.1	21.3	22.08
IN-DOMAIN + pre-train	24.2	34.5	22.4	21.59
<i>Test on Opinion</i>				
BASELINE1	17.9	26.6	11.4	28.18
BASELINE2	12.9	20.5	11.7	15.00
IN-DOMAIN	19.8	31.9	19.9	14.60
IN-DOMAIN + pre-train	19.9	31.8	19.4	14.22

Table 1: Evaluation based on ROUGE-2 (R-2), ROUGE-L (R-L), and BLEU (multiplied by 100) for in-domain training.

Effect of Domain Adaptation. Here we evaluate on domain adaptation, where OPINION is the target domain. From Figure 4, we can see that when In-domain data is insufficient Mix-domain training yields better performance. As more In-domain training data becomes available, it outperforms Mix-domain training. Baseline for selecting the first sentence as summary is also displayed. Sample summaries in Figure 5 also shows that OUT-OF-DOMAIN training tends to generate summary in similar style to the source domain, while MIX-DOMAIN training introduces the style of the target domain. In our dataset, the first sentences of summaries for OPINION are usually in the form of *[PERSON] reviews/criticizes/columns [EVENT]*, but the summaries for NEWS usually start with event descriptions directly. Such style difference is reflected in OUT-OF-DOMAIN and MIX-DOMAIN too.

We further classify the words in gold-standard summaries based on if they are seen in abstracts during training and then whether they are taken from the input text. We examine whether they are generated correctly. Full training set of opinion is used for in-domain and mix-domain training. Table 2 shows that among in-domain models, the model trained for news are superior at generating tokens mentioned in the input, compared to the model trained for opinion (33.7% v.s. 22.0%). Nonetheless, model trained for opinion is better at generating new words not in the input (8.2% vs. 2.6%). This is consistent with our observation that in opinion domain human editors favors

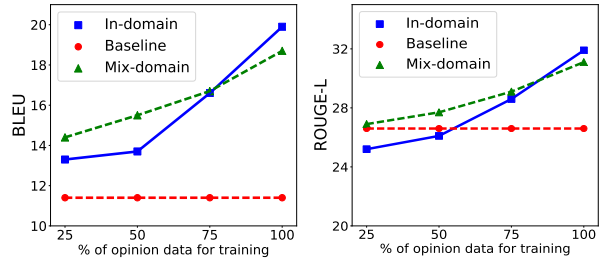


Figure 4: BLEU (left) and ROUGE-L (right) performance on In-domain and Mix-domain setup over different amount of training data. As the training data increases, In-domain outperforms Mix-domain training.

	Seen in Training (%)						Unseen (%)
	In Input			Not In Input			
	Gen	Mis	Total	Gen	Mis	Total	
<i>Test on News</i>							
IN-DOMAIN	33.7	40.9	74.6	2.6	19.3	21.9	3.5
<i>Test on Opinion</i>							
IN-DOMAIN	22.0	43.3	65.3	8.2	22.1	30.3	4.5
OUT-OF-DOMAIN	19.9	45.3	65.2	1.1	29.2	30.3	
MIX-DOMAIN	18.6	46.6	65.2	6.3	23.9	30.2	

Table 2: Comparison of generated (Gen) and missed (Mis) tokens for different training setups. We divide token in goldstandard summaries by 1) if it is seen in abstracts during training, and 2) if it is in the input text.

new words different from the input.

Further Analysis. Here we study what information is transferable cross domains by investigating the attention weights assigned to the input text.

What can be transferred. We start with input words with highest attention weights when generating the summaries. Among these, we show the percentage over different word categories as in Table 3. For named entities, model trained on out-of-domain data pays more attention to PERSON and less attention to ORGANIZATION, while the in-domain trained model does reverse. This is consistent with the fact that opinion abstracts contains more PERSON and less ORGANIZATION than news abstracts (see Figure 3). This suggests that the identification of summary-worthy named entities might be transferable from NEWS to OPINION. Similar effect is also observed for nouns and verbs, though less significant.

Attention change for domain adaptation. We also examine the percentage of attention paid to summary-worthy words. For every output token we pick the input token with highest attention weight, and count the ones reused by hu-

Human: stephen holden reviews carnegie hall concert celebrating music of judy garland. singers include her daughter, lorna luft.
Out-of-Domain: article discusses possibility of carnegie hall in carnegie hall golf tournament.
Mix-Domain: stephen holden reviews performance by jazz singer celebration by rainbow and garland at carnegie, part of tribute hall.
Human: janet maslin reviews john grisham book the king of torts .
Out-of-Domain: interview with john grisham of legal thriller is itself proof for john grisham 376 pages.
Mix-Domain: janet maslin reviews book the king of torts by john grisham .
Human: anthony tommasini reviews 23d annual benefit concert of richard tucker music foundation , featuring members of metropolitan opera orchestra led by leonard slatkin .
Out-of-Domain: final choral society and richard tucker music foundation , on sunday night in [UNK] fisher hall , will even longer than substantive 22d gala last year .
Mix-Domain: anthony tommasini reviews 23d annual benefit concert of benefit of richard tucker music.

Figure 5: Sample summaries based on OUT-OF-DOMAIN and MIX-DOMAIN training on opinion articles.

man. For IN-DOMAIN test on NEWS, on average 29.57% of the output tokens have highest attention on summary-worthy words. For OUT-OF-DOMAIN test on OPINION, the number is 15.93%; for MIX-DOMAIN, it is 26.08%. This shows the ability to focus on salient words is largely kept for MIX-DOMAIN training. Additionally, as can be seen in Table 3, model trained on MIX-DOMAIN puts more attention weights on PERSON (and all named entities) and nouns, but less attention on verbs and subjective words, compared with the model trained OUT-OF-DOMAIN. This again aligns with our observation for the domain difference based on abstracts as in Figures 2 and 3.

5 Related Work

Domain adaptation has been studied for a wide range of natural language processing tasks (Blitzer et al., 2007; Florian et al., 2004; Daume III, 2007; Foster et al., 2010). However, little has been done for investigating summarization systems (Sandu et al., 2010; Wang and Cardie, 2013). To the best of our knowledge, we are the first to study the adaptation of neural summarization models for

<i>Src</i> → <i>Trt</i>	IN-DOMAIN	OUT-OF-DOMAIN	MIX-DOMAIN
	<i>News</i> → <i>News</i>	<i>News</i> → <i>Opin</i>	<i>News</i> + <i>Opin</i> → <i>Opin</i>
PER	7.9%	8.7%	15.1% ↑
ORG	10.9%	6.9%	8.2% ↑
All NEs	26.7%	23.6%	31.6% ↑
Noun	41.2%	36.2%	43.3% ↑
Verb	10.3%	6.7%	5.5% ↓
Positive	5.6%	5.1%	4.5% ↓
Negative	2.5%	2.2%	2.1% ↓

Table 3: Attention distribution on different word categories. We consider input words with highest attention weights when generating the summaries, and characterize them by Named Entity, POS tag, and Subjectivity. The arrows shows the change with regard to OUT-OF-DOMAIN.

new domain. Furthermore, Recent work in neural summarization mainly focuses on specific extensions to improve system performance (Rush et al., 2015; Takase et al., 2016; Gu et al., 2016; Nallapati et al., 2016; Ranzato et al., 2015). It is unclear how to adapt the existing neural summarization systems to a new domain when the training data is limited or not available. This is a question we aim to address in this work.

6 Conclusion

We investigated domain adaptation for abstractive neural summarization. Experimental results showed that pre-training model with extractive summaries helps. By analyzing the attention weight distribution over input tokens, we found the model was capable to select salient information even trained on out-of-domain data. This points to future directions where domain adaptation techniques can be developed to allow a summarization system to learn content selection from out-of-domain data while acquiring language generating behavior with in-domain data.

Acknowledgments

This work was supported in part by National Science Foundation Grant IIS-1566382 and a GPU gift from Nvidia. We thank three anonymous reviewers for their valuable suggestions on various aspects of this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly

- learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- John Blitzer, Mark Dredze, Fernando Pereira, et al. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*. volume 7, pages 440–447.
- Hal Daume III. 2007. **Frustratingly easy domain adaptation**. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 256–263. <http://www.aclweb.org/anthology/P07-1033>.
- R Florian, H Hassan, A Ittycheriah, H Jing, N Kambhatla, X Luo, N Nicolov, and S Roukos. 2004. A statistical model for multilingual entity detection and tracking. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*. Association for Computational Linguistics, Boston, Massachusetts, USA, pages 1–8.
- George Foster, Cyril Goutte, and Roland Kuhn. 2010. **Discriminative instance weighting for domain adaptation in statistical machine translation**. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '10, pages 451–459. <http://dl.acm.org/citation.cfm?id=1870658.1870702>.
- Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, pages 340–348.
- Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bitan Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *EMNLP*. pages 1602–1613.
- Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. **Incorporating copying mechanism in sequence-to-sequence learning**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1631–1640. <http://www.aclweb.org/anthology/P16-1154>.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, volume 8.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rns and beyond. *CoNLL 2016* page 280.
- Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval* 5(2–3):103–233.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.
- Daniele Pighin, Marco Cornolti, Enrique Alfonseca, and Katja Filippova. 2014. **Modelling events through memory-based, open-ie patterns for abstractive summarization**. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 892–901. <http://www.aclweb.org/anthology/P14-1084>.
- Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. **A neural attention model for abstractive sentence summarization**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 379–389. <http://aclweb.org/anthology/D15-1044>.
- Evan Sandhaus. 2008. The new york times annotated corpus, 2008. *Linguistic Data Consortium, PA*.
- Oana Sandu, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2010. Domain adaptation to summarize human conversations. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Association for Computational Linguistics, pages 16–22.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
- Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. **Neural headline generation on abstract meaning representation**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1054–1059. <https://aclweb.org/anthology/D16-1112>.
- Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*. pages 2692–2700.
- Lu Wang and Claire Cardie. 2013. **Domain-independent abstract generation for focused meeting summarization**. In *Proceedings of the 51st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Sofia, Bulgaria, pages 1395–1405. <http://www.aclweb.org/anthology/P13-1137>.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 47–57. <http://www.aclweb.org/anthology/N16-1007>.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 347–354.