Text Summarization of Product Titles

Joan Xiao
joan.xiao@gmail.com
Figure Eight, Inc.
San Francisco, CA

Robert Munro*
robert.munro@gmail.com
Lilt, Inc.
San Francisco, CA

ABSTRACT

In this work, we investigate the problem of summarizing titles of e-commerce products. With the increase in popularity of voice shopping due to smart phones and (especially) in-home speech devices, it is necessary to shorten long text-based titles to more succinct titles that are appropriate for speech. We present two extractive summarization approaches using bi-directional long short-term memory encoder-decoder network with attention mechanism. The first approach treats the problem as a multi-class named entity recognition problem while the second approach treats it as a binary class named entity recognition problem. As a comparison, we also evaluate two abstractive summarization approaches using the same neural network architecture. We compare the results with automated (ROUGE) and human evaluation. Our experiment demonstrates the effectiveness of both extractive summarization approaches.

KEYWORDS

extractive summarization, abstract summarization, neural networks, voice shopping, named entity recognition

ACM Reference Format:
Joan Xiao and Robert Munro. 2019. Text Summarization of Product Titles. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR 2019 eCom).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Online marketplaces often have millions of products, and the product titles are typically intentionally made quite long for the purpose of being found by search engines. A typical 20-word title can be easily skimmed when it is text, but it provides a bad experience when it needs to be read out loud. With voice shopping estimated to hit $40+ billion across U.S. and U.K. by 2022 [1], short versions or summaries of product titles are desired to improve user experience with voice shopping.

We worked with one of the largest online e-commerce platforms which is also one of the largest producers of in-home devices. They firmly believe that voice-based search is an important future interface for online commerce and they are expanding into speech-based shopping. With them, we identified that a desired short title should contain only the essential words that are present in the original product title, with no additional words. The essential words fall into the following categories:

- BRAND: brand name of the product
- FUNCTION: what the product does
- VARIATION: variation (color, flavor, etc.)
- SIZE: size information
- COUNT: count information

A product title may or may not have all 5 attributes above - often times VARIATION, SIZE, or COUNT may not be present. Some examples of the original product titles and desired short titles are shown in Figure 1.

Summarization techniques are classified into two categories: extractive and abstractive. Extractive summarization identifies and extracts key segments of the text, then assembles them to compose a summary. Abstractive summarization generates a summary from scratch without being constrained to reusing phrases from the original text.

In this work we apply two extractive summarization and two abstractive summarization approaches to summarize a dataset of e-commerce product titles, and compare results using both ROUGE-1 and ROUGE-2 scores and human judgments. The evaluation results show that extractive summarization models consistently perform much better than abstractive summarization models.

We conclude that extractive summarization is effective for title summarization at scale. For titles up to 36 words in length, the summarization is as good as human summarization.

2 BACKGROUND & RELATED WORK

2.1 Extractive Summarization

Most work on automatic summarization has been focusing on extractive summarization. [18] proposed a simple approach to extractive summarization by selecting top sentences ranked by the number of top high frequency words that are contained in the sentences. [12] enhanced this mechanism by utilizing additional information such as cue words, title, heading words and sentence location.

Various approaches based on graphs [13], topic modeling [33] and supervised learning have been proposed since then. Supervised learning methods typically model this as a classification problem on whether a sentence in the original document should be included in the summary or not. Hidden Markov Models [10] and Conditional Random Fields [29] are among the most common supervised learning techniques used for summarization.

Recently deep neural networks [7, 21–23, 35] have become popular for extractive summarization. To date, the majority of these approaches focus on summarizing multiple documents, or a single document with multiple sentences.

*Research conducted during employment at Figure Eight, Inc.
[1] https://www.prnewswire.com/news-releases/voice-shopping-set-to-jump-to-40-billion-by-2022-rising-from-2-billion-today-300605596.html

SIGIR 2019 eCom, July 2019, Paris, France

Original Title	Elite Platinum Mst-900d Maxi-matic 8.5 Quart Digital Programmable Slow Cooker With Timer Stainless Steel
Short Title	Elite Platinum Slow Cooker Stainless Steel 8.5 Quart
Original Title	The Jetset - 12" Wooden Salad Bowl - Natural Rubber Wood - Product From Thailand
Short Title	The Jetset Wooden Salad Bowl 12"
Original Title	Brekky Bbq Grill Mat - Set Of 3 - Teflon Nonstick Grilling Accessory - Perfect For Charcoal Electric And Gas Grill
Short Title	Brekky Bbq Grill Mat Set Of 3

Figure 1: Examples of original product titles and desired short titles

In our work we focus on extractive summarization on product titles which are single "sentences", although the sentences here are fragments of sentences. Since we identified that a desired short title should contain only the words that fall into the 5 categories (BRAND, FUNCTION, VARIATION, SIZE and COUNT), the problem is reduced to identifying the words in these categories, which can be treated as a Named Entity Recognition problem. Once the essential words are identified, a short title can be composed by assembling these words together.

2.2 Named Entity Recognition

Named Entity Recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, quantities, etc. NER systems have been created using linguistic grammar-based techniques as well as statistical models such as machine learning.

Traditional machine learning approaches have been dominated by applying Hidden Markov Models [6], Decision Trees [28], Support Vector Machines [3], and Conditional Random Fields [20] to hand-crafted features. [9] pioneered a neural network model that requires little feature engineering and instead learns important features from word embeddings [31] trained on large quantities of unlabeled text. Since then, CNN, LSTM, and bidirectional LSTM models using feature extractors for word and characters ([1, 8, 15, 16, 19, 25, 34]) have been reported to achieve start-of-the-art results on CoNLL-2003 NER task [26].

In our work we experiment with two NER based approaches for extractive summarization.

2.3 Abstract Summarization

The task of abstractive sentence summarization was formalized around the DUC-2003 and DUC-2004 competitions [24]. Inspired by the success of attention model in neural machine translation, [5] proposed a sequence-to-sequence encoder-decoder LSTM [14] with attention mechanism for this problem, showing state-of-the-art performance on the DUC tasks. Since then, more work using deep neural networks has been done on focusing on handling out-of-vocabulary words [22] and discouraging repetition [27].

As a comparison with the extractive approaches, we experiment with two abstractive summarization models on the same dataset.

3 OUR APPROACHES

We first manually extracted named entities corresponding to the classes of BRAND, FUNCTION, VARIATION, SIZE, and COUNT, then constructed ground truth labels separately for each model. Once a model is trained, it makes prediction on titles from the test set. In the case of extractive summarization models, shorter titles are composed from the predicted named entities.

Figure 2 illustrates how the labels for each model are generated from the annotations of named entities of a product title. Figure 3 describes how a short title is generated from each model's prediction using the same example.

3.1 Extractive Summarization (Multi-class NER)

We treat the summarization problem as a multi-class sequence labeling problem, where each class corresponds to the category of a word in the product title, i.e., whether a word is a BRAND, FUNCTION, VARIATION, SIZE, COUNT, or none of these. Once we have the predicted classes of all words in the title, we create a short (summary) title by concatenating all words that are classified as having a non-trivial entity class.

In this study, we obtained the ground-truth labels for NER using the data annotation platform Figure Eight. Crowd workers were asked to extract named entities (BRAND, FUNCTION, VARIATION, SIZE, COUNT) from the product titles. We then construct a label for each title using a BIO tag scheme. The product titles and these labels (Figure 2) are then fed into a neural network. For each predicted sequence of a title, we construct a short title using the named entities extracted from the prediction, in the fixed order of BRAND, FUNCTION, VARIATION, SIZE, COUNT (Figure 3).

3.2 Extractive Summarization (Binary NER)

In this approach, we treat the summarization problem as a binary NER problem, where a word in a title belongs to the positive class if the word is included in the summary, in contrast with the previous multi-class NER model. We re-use the ground-truth labels from multi-class NER task above by transforming each entity class to the positive class ("1") and non-entity class to the negative class ("0"). The product titles and these labels are then fed into a neural network (Figure 2).

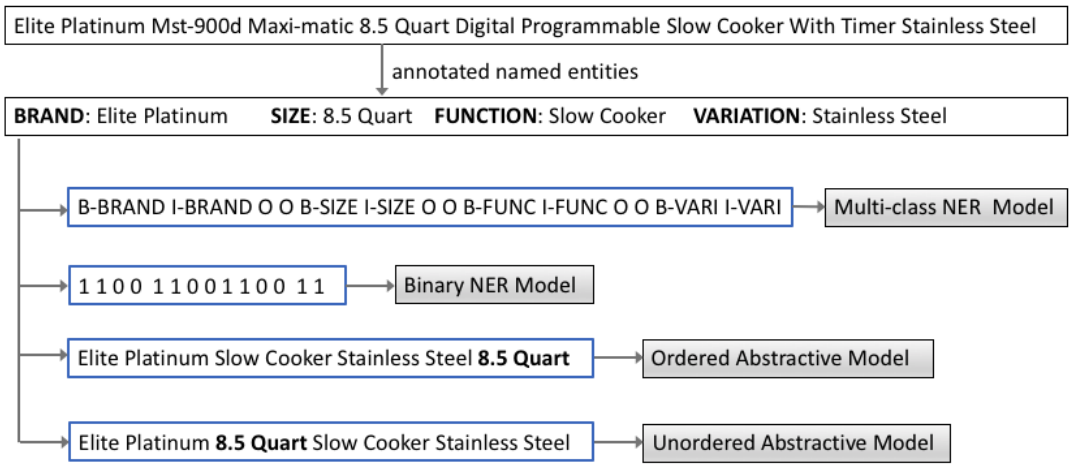

Figure 2: How labels are generated from annotations for each model. Bold words in the labels for the abstractive models indicate the difference in the order of words of the entity SIZE.

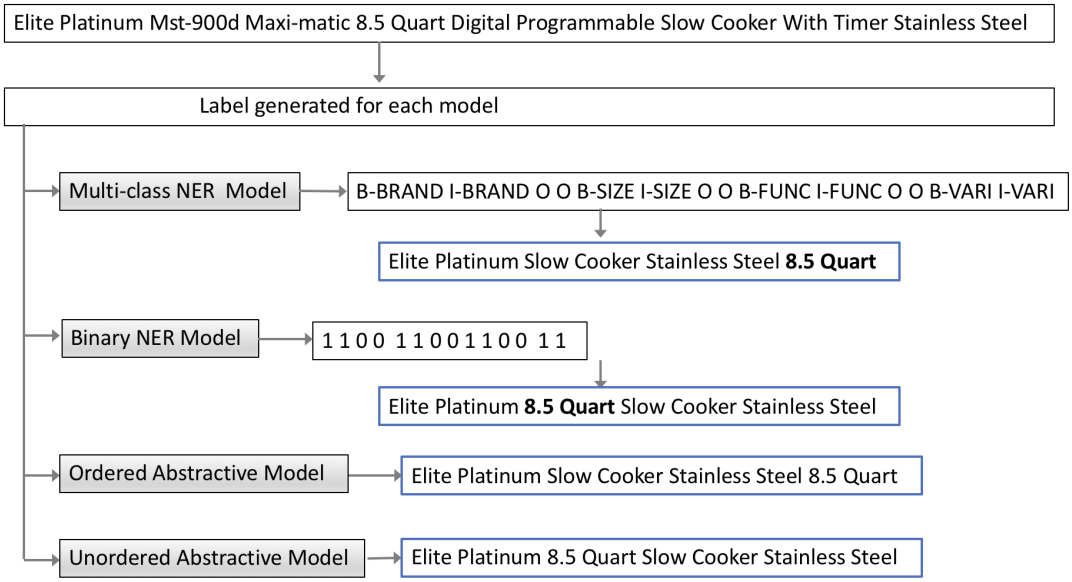

Figure 3: How shorter title is generated from each model's prediction. Bold words in the short titles generated from the extraction models indicate the difference in the order of the entity SIZE.

For each predicted sequence of a title, we construct a short title by including the words predicted in positive class, in the same order as they appear in the original title (Figure 3).

3.3 Abstractive Summarization (Ordered)

For the abstractive summarization task, the ground-truth labels are constructed from the annotated named entities in the order of BRAND, FUNCTION, VARIATION, SIZE, and COUNT, same as in the multi-class NER approach (Figure 2).

3.4 Abstractive Summarization (Unordered)

Since the ground-truth labels for the abstractive summarization approach above are generated in a specific order, the words in the short title may not occur in the same order as they do in the source. We are curious to know whether the re-ordering of the words affects the result of the summarization. Therefore, we made one change from the ordered abstractive summarization approach, using the same annotated named entities but keeping the words in the same order as they originally appear in the source (Figure 2).

Model	Test Set		1000 Random Titles	
	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2
NER_GOLD	-	-	75.32	50.13
Multi-class NER	**84.71**	65.98	75.00	50.43
Binary NER	84.09	**67.87**	**75.06**	**58.07**
Ordered Abstractive	78.83	47.85	67.47	41.66
Unordered Abstractive	80.70	64.91	72.01	53.92

Table 1: ROUGE-1 and ROUGE-2 on test set and 1000 random titles. Bold indicates the model with the highest ROUGE-1 or ROUGE-2 score on each dataset.

NER_Gold	Multi-class NER	Binary NER	Ordered Abstractive	Unordered Abstractive	Human Summarization
7.02 ± 1.72	6.77 ± 1.75	6.78 ± 1.80	6.39 ± 1.79	6.47 ± 1.70	7.70 ± 1.76

Table 2: Human evaluation on accuracy.

Method	Succinctness	Combined (accuracy and succinctness)
NER_Gold	9.54 ± 0.83	8.28 ± 0.96
Multi-class NER	9.53 ± 0.85	8.15 ± 0.97
Binary NER	9.53 ± 0.77	8.16 ± 1.02
Human Summarization	8.76 ± 1.35	8.23 ± 1.09

Table 3: Human evaluation on succinctness, and combined evaluation on accuracy and succinctness.

Method	% of Titles with Factual Errors
Ordered Abstractive	29.1
Unordered Abstractive	26.8
Human Summarization	0.19

Table 4: Human evaluation on non-factualness.

4 EXPERIMENTAL SETUP

4.1 Dataset

Our dataset consists of 56,200 product titles in English, randomly selected from the following categories:

- Baby Products
- Beauty
- Drugstore
- Fresh Perishable
- Fresh Produce
- Grocery
- Home
- Kitchen
- Office Products
- Pantry

The dataset is randomly split into a training set of size 37,300, a validation set of size 9,300, and a test set of size 9,600.

4.2 Evaluation

We evaluated the four approaches with the standard ROUGE metric [17], reporting the F1 scores on each model's test set for ROUGE-1 and ROUGE-2 against their corresponding ground truth labels.

In addition, we selected 1000 random product titles from the test set and asked the crowd workers to manually summarize them. The crowd workers were instructed to summarize in a similar manner to how the short titles of the NER model are generated: identify keywords corresponding to BRAND, FUNCTION, VARIATION, SIZE and COUNT, and then create a short title using these keywords in the order they appear in this list.

We then asked different crowd workers to compare the short titles produced from the models with the human summarization results on the following metrics:

- **Accuracy**: on the scale of 1-10, how accurately each short title describes the product.
- **Non-factualness**: whether the short title has factual errors. Only the two abstractive models were compared with human summarization.
- **Succinctness**: on the scale of 1-10, how succinct each short title is. A short title is rated as 10 if it does not contain any non-essential words that can be removed without affecting how accurately it describes the product. The abstractive models are excluded from this evaluation due to the non-factualness problem.

For each metric above, 3 crowd workers were assigned to rate the short titles of each product title, and the average of the 3 workers' ratings is used as the aggregated rating.

Finally, in order to have a single metric to evaluate the short titles (excluding the titles generated from the abstractive models), we combined the human evaluation ratings on accuracy and succinctness by taking the average of these two ratings for each title.

4.3 Model Architecture

For simplicity, we used the same bi-directional LSTM encoder/decoder network with attention mechanism for all 4 approaches. Both encoder and decoder are two-layer LSTMs with 512 hidden units. Dropout [30] is used at the decoder and both source and target word embeddings, and beam search of length 5 is used during inference. We trained the models on Amazon SageMaker [2].

5 RESULTS

5.1 Results on Test Set

Table 1 lists the ROUGE-1 and ROUGE-2 F1 scores on each model's test set against their corresponding ground truth labels. On both metrics, the two extractive models perform better than the two abstractive models, and Unordered Abstractive does better than Ordered Abstractive.

5.2 Results Compared with Human Summarization

Table 1 also shows the F1 scores of the ROUGE-1 and ROUGE-2 on the 1000 random titles when evaluated against human summarization. For comparison purpose, we added the short titles generated from the labels used by the NER model, and it is named as "NER_Gold" in the table.

ANOVA and post-hoc tests on ROUGE-1 scores show that there is no significant difference between the two extractive models, the extractive models are significantly better than both abstractive models, and the unordered abstractive model is significantly better than the ordered abstractive model.

On ROUGE-2 scores, the binary NER model is significantly better than the unordered abstractive model, which is better than multi-class NER and NER_Gold, which are better than the ordered abstractive model.

It is interesting to note that the unordered abstractive model achieves higher scores than the ordered abstractive model, and it even achieves higher ROUGE-2 score than the multi-class NER model. This suggests that preserving the order of the words in the target labels has a significant impact on the abstractive model's performance.

For both ROUGE-1 and ROUGE-2 scores, there is no statistically significant difference between multi-class NER and NER_Gold.

5.3 Human Evaluation on Accuracy

Table 2 lists the average and standard deviation of the crowd workers' rating on all 5 versions of short titles, plus the human summarized titles.

ANOVA and post-hoc test on the ratings show results consistent with the ROUGE-1 evaluation performed above: there is no significant difference among the extractive models and among the abstractive models. However, NER_Gold is rated as significantly higher than the two NER models, due to the fact that the NER models fail to identify some named entities in some cases. And not surprisingly, human summarization is rated as being the most accurate among all.

[2]https://aws.amazon.com/sagemaker/

5.4 Human Evaluation on Non-Factualness

The abstractive models are known to struggle with handling out-of-vocabulary words and often make non-factual errors [27]. We were curious about whether the two abstractive models perform differently in terms of non-factualness. Table 4 shows the percentage of the titles are rated as having factual errors. ANOVA Test shows that there is no significant difference between the two abstractive models.

5.5 Human Evaluation on Succinctness

As the abstractive models make factual errors, this evaluation includes only the extractive models and human summarization.

Table 3 shows the average and standard deviation of the human evaluation results on succinctness. There is no statistical difference among the extractive models and NER_Gold, but interestingly human summarization is rated as the least succinct among all. Some examples (Figure 4) indicate that human summarization tends to include words related to product variations which are not captured by the models, and human raters do not think these variations are essential to describe the product.

5.6 Combined Human Evaluation on Accuracy and Succinctness

Table 3 also shows the average and standard deviation of the combined human evaluation results. Again, there is no statistically significant difference between the two extractive models, and it is interesting to note that even though NER_Gold is significantly better than the two extractive models, there is no statistically significant difference between human summarization and any of the other 3 versions.

To understand how the ratings vary with the length of product titles, we show in Figure 5 the average combined rating broken down by number of words in the product titles. And Table 5 shows the word count distribution of these product titles. We see that the two NER models perform very close to human summarization unless the product titles are extremely long (with more than 37 words, which accounts for only 0.2% of the titles).

6 CONCLUSION

We applied four different deep learning based approaches to product title summarization on a dataset of 56,200 product titles and used both ROUGE scores and human judgments to evaluate the results on a random 1000 titles from the test set. The evaluation results show that extractive summarization models consistently perform much better than the abstractive summarization models, and overall there is no statistically significant difference between the two extractive models and human summarization.

There are several avenues for future work. First, in this study we used the same neural network architecture for all models, so we did not use the latest and greatest neural network architecture for NER, and this is evident in the gap in accuracy between NER_Gold and NER models when the product titles are longer (Figure 5). We plan to adopt the state-of-the-art architectures such as Elmo [25] and Flair [1] contextual embeddings for the two NER models for future study. In addition, we plan to experiment with self-attention transformer [32] based models such as OpenAI GPT [2], BERT [11] and [4].

Original Title	Dixon Ticonderoga Wood-cased # 4 Extra Hard Pencils Box Of 12 Yellow 13884
Multi-class NER	Dixon Ticonderoga Pencils Yellow Box Of 12
Human Summarization	Dixon Ticonderoga Wood-cased Pencils Yellow Box Of 12

Original Title	Labvon Bluetooth Speaker With Enhanced Bass Hands Free Potable Wireless Speaker For Iphone / Ipad / Ipod / Mp3 Player / Laptop
Multi-class NER	Labvon Bluetooth Speaker
Human Summarization	Labvon Bluetooth Speaker With Hands Free Potable Wireless

Figure 4: Human summarization is rated lower than NER models on succinctness for some product titles.

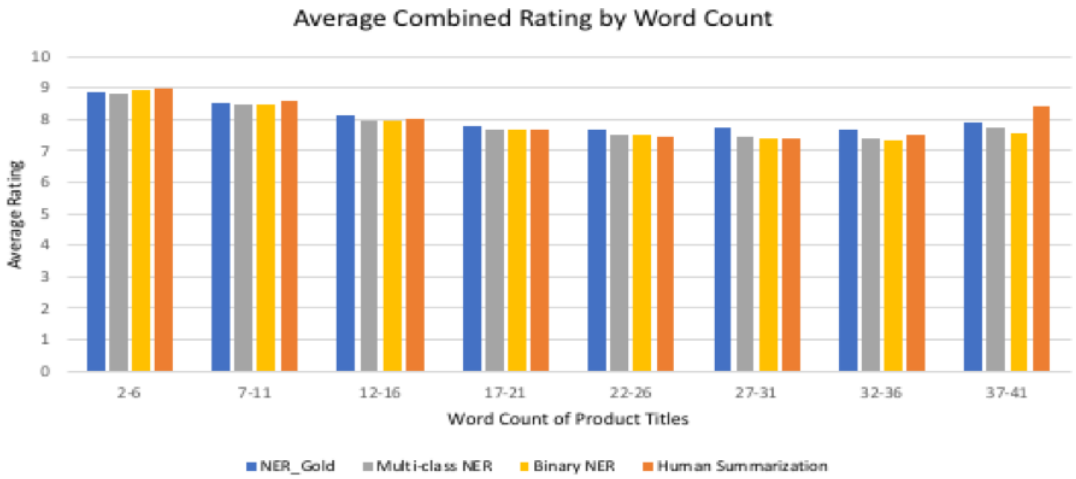

Figure 5: Average combined rating by word count, showing that automated (extractive) summarization is equal to human-summarization for titles up to 36 words in length.

Word Count	2-6	7-11	12-16	17-21	22-26	27-31	32-36	37-41
% of Titles	11.9	39.9	20.6	10.7	7.8	6.8	2.1	0.2

Table 5: Word count distribution of product titles.

These models do not use recurrent neural networks therefore do not restrict their prediction performance to short sequences, and all have achieved competitive results on CoNLL 2003 NER task.

Second, for abstractive summarization, even with the high percentage of titles making non-factual errors (Table 4), the ROUGE-1 and ROUGE-2 and human evaluation on accuracy are still considerably high, which suggests that abstractive summarization may achieve good results if the non-factual errors are eliminated. We plan to explore the copy mechanism in pointer and generator approaches ([22, 27]) in future study.

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1638–1649. https://www.aclweb.org/anthology/C18-1139

[2] Radford Alec, Narasimhan Karthik, Salimans Tim, and Ilya Sutskever Openai. 2018. *Improving Language Understanding by Generative Pre-Training*. Technical Report. https://doi.org/10.1093/aob/mcp031

[3] Masayuki Asahara and Yuji Matsumoto. 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 8–15. https://doi.org/10.3115/1073445.1073447

[4] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke S. Zettlemoyer, and Michael Auli. 2019. Cloze-driven Pretraining of Self-attention Networks. *CoRR* abs/1903.07785 (2019).

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2015).

[6] Daniel M. Bikel, Scott Miller, Richard M. Schwartz, and Ralph M. Weischedel. 1997. Nymble: a High-Performance Learning Name-finder. In *ANLP*.

[7] Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. *CoRR* abs/1603.07252 (2016).

[8] Jason P. C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4 (2016), 357–370.

[9] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research* 12 (2011), 2493–2537.

[10] John M. Conroy and Dianne P. O'Leary. 2001. Text Summarization via Hidden Markov Models. In *SIGIR*.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018).

[12] H. P. Edmundson. 1969. New Methods in Automatic Extracting. *J. ACM* 16 (1969), 264–285.

[13] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *J. Artif. Intell. Res.* 22 (2004), 457–479.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (1997), 1735–1780.
[15] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015).
[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *HLT-NAACL*.
[17] Chin-Yew Lin. 2004. ROUGE: A Package For Automatic Evaluation Of Summaries. In *ACL 2004*.
[18] Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2 (1958), 159–165.
[19] Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *CoRR* abs/1603.01354 (2016).
[20] Andrew McCallum and Wei Li. 2003. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *CoNLL*.
[21] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016. SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. In *AAAI*.
[22] Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, ÃĞaglar GülÃğehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *CoNLL*.
[23] Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2017. Classify or Select: Neural Architectures for Extractive Document Summarization. *CoRR* abs/1611.04244 (2017).
[24] Paul Over, Hoa Dang, and Donna K. Harman. 2007. DUC in context. *Inf. Process. Manage.* 43 (2007), 1506–1520.
[25] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*.
[26] Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *CoNLL*.
[27] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*.
[28] Satoshi Sekine. 1998. Description of the Japanese NE System Used for MET-2. In *MUC*.
[29] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document Summarization Using Conditional Random Fields. In *IJCAI*.
[30] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15 (2014), 1929–1958.
[31] Florian Strub, Harm de Vries, Jérémie Mary, Bilal Piot, Aaron C. Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*.
[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NIPS*.
[33] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Multi-Document Summarization using Sentence-based Topic Models. In *ACL/IJCNLP*.
[34] Zhilin Yang, Ruslan R. Salakhutdinov, and William W. Cohen. 2016. Multi-Task Cross-Lingual Sequence Tagging from Scratch. *CoRR* abs/1603.06270 (2016).
[35] Wenpeng Yin and Yulong Pei. 2015. Optimizing Sentence Modeling and Selection for Document Summarization. In *IJCAI*.