Cross-Task Knowledge Transfer for Query-Based Text Summarization

Elozino Egonmwan †* **Vittorio Castelli** ‡ **Md Arafat Sultan** ‡
† University of Lethbridge, Lethbridge, AB, Canada
‡ IBM Research AI
elozino.egonmwan@uleth.ca, vittorio@us.ibm.com, arafat.sultan@ibm.com

Abstract

We demonstrate the viability of knowledge transfer between two related tasks: machine reading comprehension (MRC) and query-based text summarization. Using an MRC model trained on the SQuAD1.1 dataset as a core system component, we first build an *extractive* query-based summarizer. For better precision, this summarizer also compresses the output of the MRC model using a novel sentence compression technique. We further leverage pre-trained machine translation systems to *abstract* our extracted summaries. Our models achieve state-of-the-art results on the publicly available CNN/Daily Mail and Debatepedia datasets, and can serve as simple yet powerful baselines for future systems. We also hope that these results will encourage research on transfer learning from large MRC corpora to query-based summarization.

1 Introduction

Query-based single-document text summarization is the process of selecting the most relevant points in a document for a given query and arranging them into a concise and coherent snippet of text. The query can range from an individual word to a fully formed natural language question. *Extractive* summarizers select verbatim the most relevant span of text in the source, while *abstractive* summarizers further paraphrase the selected content for better clarity and brevity.

By and large, existing approaches train models using summarization data corpora (Nema et al., 2017; Hasselqvist et al., 2017), which are of moderate size. At the same time, large corpora are available for related tasks, specifically machine reading comprehension (MRC) and machine translation (MT). To find out if such corpora have utility for summarizers, we propose methods to directly produce extractive and abstractive query-based summaries from pretrained MRC and MT modules, requiring no further adaptation or transfer learning steps.

In our experiments, this approach outperforms existing methods, suggesting a novel route to query-based summarization: pre-training systems on such related tasks, where an abundance of training data is enabling extremely rapid progress (Wang et al., 2018; Sun et al., 2018; Vaswani et al., 2017), and using summarization-specific corpora for transfer learning.

The main contributions of this work are:

- We show how existing off-the-shelf components for tasks other than query-based summarization are competitive with the state-of-the-art in the field, even without model adaptation or transfer learning – we hope to encourage researchers to more closely examine transfer learning among these tasks.

- Specifically, we show how processing the output of an MRC system (trained on the SQuAD1.1 dataset (Rajpurkar et al., 2016)) with a simple rule-based sentence compression module that operates on the dependency parse (de Marneffe and Manning, 2008) of the answer sentence yields results that are better than those of query-based extractive summarizers trained for the specific dataset.

- We demonstrate how a sequence-to-sequence model (Sutskever et al., 2014) that uses two machine translation engines—from and to English, respectively—applied to the output of the above, yields results that are better than query-based abstractive summarizers trained for the specific dataset.

*Work done at IBM.

Passage: people whether overweight or not are still people. you can not compare a person with a suitcase. suitcases don't live and breathe. this rule is the same with weight. excess weight in a suitcase is not comparable with a fat person . **Query:** is it necessary to charge fat passengers extra when flying? **Reference Summary:** there is no comparison between a person and a suitcase.
Our method (abstractive) : The overweight in the bag can't be compared with the fat guy. **Diversity driven attention model:** beings are definitely by the <unk> to illegal illegal.

Table 1: Example/comparison of our *abstractive* summary on a Debatepedia sample with the output of the diversity driven attention model of Nema et al. (2017). Our generated summary is relevant to the query.

Passage (truncated): [...] offensive italian football expert and author john foot explained how paulo berlusconi 's words were offensive on several levels . " it is an insult , " foot told cnn [...] **Query:** john foot **Reference Summary:** italian football expert and author john foot says paulo berlusconi 's words are offensive on several levels .
Our method (extractive) : offensive italian football expert and author john foot explained how paulo berlusconi 's words were offensive on several levels .

Table 2: Example of our *extractive* summary on an example from the query-based version of CNN/Daily Mail (Hermann et al., 2015).

2 Task Definition

Given a document $D = (S_1, ..., S_n)$ with n sentences comprising of a set of words $D_W = \{d_1, ..., d_w\}$, and a query $Q = (q_1, ..., q_m)$ with m words, one desires to produce an *extractive* (S_E) or *abstractive* (S_A) summary that provides information about the answer to Q, where $S_E \subseteq D_W$ and $S_A = \{w_1, ..., w_s\} \mid \exists w_i \notin D_W$. Tables 1 and 2 show examples of abstractive and extractive summaries, respectively.

3 Method

Our proposed system comprises of three modules for extractive summarization: retrieval of candidate answer phrases using a reading comprehension system, sentence extraction, and sentence compression. Additionally we utilize two MT modules (English to Spanish and back) to paraphrase for abstractive summarization.

3.1 Machine Reading Comprehension

MRC requires the identification of a contiguous span of words in a passage that answers a given query (Rajpurkar et al., 2016; Wang et al., 2018; Hu et al., 2017). We use the MRC model by Wang et al. (2016b) trained on the SQuAD1.1 dataset (Rajpurkar et al., 2016) to identify the top *n* (empirically: *n*=5) possibly overlapping candidate answer phrases, or *chunks*, for the given query. The chunks are typically short, 3.2 words on average in the training set. Obviously, chunks from MRC are not meant to be summaries, but in our system they help the summarizer focus on the regions of the input document that appear related to the query.

3.2 Sentence Extraction

Sentence extraction consists of selecting the sentences containing the top n chunks produced by MRC. This is in contrast to methods based on sentence ranking algorithms such as those used in (Boudin et al., 2015; Parveen and Strube, 2015; Nallapati et al., 2017; Cheng and Lapata, 2016). For our experiments, we impose the constraint that the candidate answer chunks for each query be contained in a single sentence. Hence, starting from $n = 5$, we iteratively reduce *n* until the top *n* candidate chunks are all contained in one sentence.

3.3 Sentence Compression

Sentence extraction often produces results that are much longer than those in the reference summaries—the training data (Table 4) suggests that 20 words is a good upper limit for the length of the summaries. We address this problem by introducing a novel sentence compression framework based on pruning the dependency parses of the sentences. Our approach is partially inspired by the work of Wang et al. (2016a), which performs sentence compression based on constituency parses. The intuition is that dependency parses better capture the semantic relations between words than constituents, which actually model syntactic structure.

Input Sentence: it is ridiculous to suggest governments should restrict their own ability to help their economies. **Paraphrase (with MT):** It is **absurd** to suggest that governments **impose limits** on their ability to help their economies.
Input Sentence: this favoritism would only increase that of which the laws are trying to suppress . **Paraphrase (with MT):** These **nepotism** will only increase the laws that you try to suppress.

Table 3: Examples of some of our paraphrased sentences using an MT system. Bolded words are novel.

Given a summary with length ≥ 20, we obtain the dependency parses of its sentences using the IBM Watson NLU toolkit. Next, we remove words in the sentences (starting from the rear) that are not in a dependency relationship with any of the candidate phrases, until the summary length limit is reached.

3.4 Back Translation

Recent research has shown gains in leveraging on the enormous corpora in machine translation (MT) for paraphrasing (Mallinson et al., 2017; Wieting and Gimpel, 2017). Inspired by such research and our fundamental goal of investigating the viability of cross-task knowledge transfer for query-based summarization, we paraphrase our extracts using an off-the-shelf MT system[1]. The final English paraphrase of the input sentence is obtained by translating it into Spanish and back-translating the translation into English. We experimented with English-French-English and English-Italian-English as well as with multi-hops approaches before settling on the English-Spanish pair, based on subjective analysis of the results. Table 3 shows examples of paraphrased sentences using back-translation.

4 Experiments

We test our approach on two publicly available datasets—Debatepedia (Nema et al., 2017) for abstractive summarization, and the version of CNN/Daily Mail that was adapted in (Hermann et al., 2015) for both extractive and abstractive summarization. No training was involved; the test sets were simply passed through the modules discussed in section 3.

[1]The MT engine is used in the IBM Watson Language Translator service.

	CNN.	Deb.
Test	14,725	979
Avg. #words/psg.	776	70
Avg. #words/query	2	11
Avg. #words/summ.	14	10

Table 4: Statistics of the dataset test samples after processing by the Wang et al. (2016b) MRC system's preprocessing module. Note that the preprocessor fails to parse 2-3% of the test samples in each dataset.

4.1 Datasets

We processed the CNN/DM[2] and Debatepedia[3] datasets using the respective official Python scripts to yield the corpora with passages, queries and summaries tailored to the queries (Table 4). CNN/DM is much larger in terms of both the number of samples and the lengths of passages, with short queries consisting of few words, mostly entity names. Debatepedia is a smaller dataset, but the queries are fully-formed natural language questions. Interestingly, although our MRC system was originally designed to answer full-length questions, as our results show later in this section, it identifies key regions of the document remarkably well in both test sets.

4.2 Evaluation

As customary in summarization tasks, we evaluate our system using ROUGE (Lin, 2004)—a family of metrics that compute the textual overlap between the output and the reference summary. The publicly available `ROUGE 2.0 toolkit`[4] was used as the implementation.

4.3 Results

Tables 5 and 6 summarize the performances of our model and other published models on Debatepedia and CNN/Daily Mail, respectively. Our models, both extractive and abstractive, outperform the published results on both test sets.

The extractive performance on CNN/DM indicates that the combination of a reading compre-

[2]https://github.com/helmertz/querysum-data/

[3]https://github.com/PrekshaNema25/DiverstiyBasedAttentionMechanism

[4]https://rxnlp.com/rouge-2-0/

Abstractive	R-1	R-2	R-L
Diversity (Nema et al., 2017)	41.26	18.75	40.43
RSA (Baumel et al., 2018)	53.09	16.10	46.18
Ours	**64.43**	**18.93**	**46.80**

Table 5: ROUGE (%) performances of our model and competing models on the Debatepedia dataset. Our model outpeforms both baselines on all metrics.

Extractive	R-1	R-2	R-L	R-SU4
QSum (Hasselqvist et al., 2017)	33.81	18.19	29.22	17.49
Ours	**65.45**	**30.07**	**60.40**	**36.62**
Abstractive				
QSum (Hasselqvist et al., 2017)	18.25	5.04	16.17	6.13
Ours	**58.46**	**25.12**	**54.32**	**32.06**

Table 6: ROUGE (%) scores of our models and the competing model on the CNN/Daily Mail dataset. Our proposed approach yields the best system for both extractive and abstractive summarization.

hension system and a syntax-driven compression module can be highly effective in identifying regions in a document that contain key information with respect to a given query. Moreover, the abstractive performances on both test sets show the effectiveness of machine translation as a paraphrasing component for abstractive summarization. In particular, in the CNN/DM test set the improvement over the baseline is greater in the abstractive than in the extractive case, again suggesting that both text selection and MT-based paraphrasing contribute to the gain.

5 Related Work

Text summarization has long been an active area of research and query-based summarization has gained momentum more recently. Classical summarization models usually identify salient parts of a text by encapsulating manually crafted rules into linear functions (Lin and Bilmes, 2011) which are solved using integer linear programming (ILP) (Nayeem and Chali, 2017; Boudin et al., 2015), conditional random fields (CRF) (Shen et al., 2007), or graph algorithms (Parveen and Strube, 2015; Erkan and Radev, 2004). More recently, neural networks, mostly with an encoder-decoder framework (Bahdanau et al., 2014), have been used to learn the underlying features (Jadhav and Rajan, 2018; Nallapati et al., 2016) trained by minimizing the cross-entropy loss (Nallapati et al., 2017) or reinforcement learning (Narayan et al., 2018; Paulus et al., 2017).

Our baseline models for query-based summarization (Nema et al., 2017; Hasselqvist et al., 2017) are both implemented on the encoder-decoder framework with the former incorporating a diversity function in their model aimed at minimizing the problem of repetitive word generation inherent in encoder-decoder models. However our approach is similar to neither, as our goal is not to train a query-based summarizer from scratch but rather to investigate the competitiveness of using pre-trained models for closely related tasks—i.e., MRC and MT—on query-based summarization.

6 Conclusions

We described an approach to extractive and abstractive summarization that relies on components designed for different tasks: MRC, sentence compression, and MT. We have shown that retrieving the top n answer chunks from a passage with an MRC system and trimming the corresponding sentences using their dependency trees yields an extractive summarizer that outperforms published results on a publicly available dataset. We also showed that using MT to produce a paraphrase of the answers yields a high-performance abstractive summarization method.

This work lays the foundations for transfer learning based approaches that use summarization data to adapt MRC models for summarization. We also envision: i) using summarization data to learn how to re-rank top n candidates from back-translation; ii) replacing the pruning system with a trained sequence-to-sequence model with an objective function that incorporates readability; and

iii) computing the AMR parse (Banarescu et al., 2013) of the candidate answers followed by text generation (Song et al., 2018) instead of using MT.

Acknowledgments

We thank the reviewers for their valuable comments and suggestions. We also thank Zhiguo Wang and Preksha Nema for clarification of their 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*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186. Association for Computational Linguistics.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Florian Boudin, Hugo Mougard, and Benoit Favre. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Johan Hasselqvist, Niklas Helmertz, and Mikael Kågebäck. 2017. Query-based abstractive summarization using neural networks. *arXiv preprint arXiv:1712.06100*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2017. Reinforced mnemonic reader for machine reading comprehension. *arXiv preprint arXiv:1705.02798*.

Aishwarya Jadhav and Vaibhav Rajan. 2018. Extractive summarization with swap-net: Sentences and words from alternating pointer networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 142–151.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 881–893.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. *Stanford typed dependencies manual*. Stanford.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1747–1759.

Mir Tafseer Nayeem and Yllias Chali. 2017. Extract with order for coherent multi-document summarization. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 51–56.

Preksha Nema, Mitesh M Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1063–1072.

Daraksha Parveen and Michael Strube. 2015. Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In *IJCAI*, pages 1298–1304.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *IJCAI*, volume 7, pages 2862–2867.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for amr-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626. Association for Computational Linguistics.

Fu Sun, Linyang Li, Xipeng Qiu, and Yang Liu. 2018. U-net: Machine reading comprehension with unanswerable questions. *arXiv preprint arXiv:1810.06638*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2016a. A sentence compression based framework to query-focused multi-document summarization. *arXiv preprint arXiv:1606.07548*.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1705–1714.

Zhiguo Wang, Haitao Mi, Wael Hamza, and Radu Florian. 2016b. Multi-perspective context matching for machine comprehension. *arXiv preprint arXiv:1612.04211*.

John Wieting and Kevin Gimpel. 2017. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.