From Extractive to Abstractive Summarization: A Journey

Parth Mehta
Information Retrieval and Language Processing Lab
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, India
parth_me@daiict.ac.in

Abstract

The availability of large document-summary corpora have opened up new possibilities for using statistical text generation techniques for abstractive summarization. Progress in Extractive text summarization has become stagnant for a while now and in this work we compare the two possible alternates to it. We present an argument in favor of abstractive summarization compared to an ensemble of extractive techniques. Further we explore the possibility of using statistical machine translation as a generative text summarization technique and present possible research questions in this direction. We also report our initial findings and future direction of research.

1 Motivation for proposed research

Extractive techniques of text summarization have long been the primary focus of research compared to abstractive techniques. But recent reports tend to suggest that advances in extractive text summarization have slowed down in the past few years (Nenkova and McKeown, 2012). Only marginal improvements are being reported over previous techniques, and more often than not these seem to be a result of variation in the parameters used during evaluation using ROUGE, and some times due to other factors like a better redundancy removal module (generally used after the sentences are ranked according to their importance) rather than the actual algorithm. Overall it seems that the current state of the art techniques for extractive summarization have more or less achieved their peak performance and only some small improvements can be further achieved. In such a scenario there seem to be two possible directions of further research. One approach that could be used is making an ensemble of these techniques which might prove to be better than the individual methods. The other option is to focus on abstractive techniques instead.

A large number of extractive summarization techniques have been developed in the past decade especially after the advent of conferences like *Document Understanding Conference (DUC)*[1] and *Text Analysis Conference (TAC)*[2]. But very few inquiries have been made as to how these differ from each other and what are the salient features on some which are absent in others. (Hong et al., 2014) is first such attempt to compare summaries beyond merely comparing the ROUGE(Lin, 2004) scores. They show that many systems, although having a similar ROUGE score indeed have very different content and have little overlap among themselves. This difference, at least theoretically, opens up a possibility of combining these summaries at various levels, like fusing rank lists(Wang and Li, 2012), choosing the best combination of sentences from several summaries(Hong et al., 2015) or using learning-to-rank techniques to generate rank lists of sentences and then choosing the top-k sentences as a summary, to get a better result. In the next section we report our initial experiments and show that a meaningful ensemble of these techniques can help in improving the coverage of existing techniques. But such a scenario is not always guaranteed, as shown in the next section, and given that such fusion techniques do have a upper bound to the extent to which they can improve the summarization performance as shown by (Hong et al., 2015), an ensemble approach would be of limited interest.

Keeping this in mind we plan to focus on

[1] duc.nist.gov

[2] www.nist.gov/tac

both approaches for abstractive text summarization, those that depend on initial extractive summary and those that do not (text generation approach). Also availability of large document-summary corpora, as we discuss in section 3, has opened up new possibilities for applying statistical text generation approaches to summarization. In the next section we present a brief overview of the initial experiments that we have performed with an ensemble of extractive techniques. In section 3 we then propose further research directions using the generative approach towards text summarization. In the final section we present some preliminary results of summarizing documents using a machine translation system.

2 Fusion of Summarization systems

In this section we report some of our experiments with fusion techniques for combining extractive summarization systems. For the first experiment we consider five basic techniques mentioned in (Hong et al., 2014) for the simple reason that they are tested extensively and are simple yet effective. These systems include LexRank, the much popular graph based summarization technique(Erkan and Radev, 2004), and Greedy-KL(Haghighi and Vanderwende, 2009), which iteratively chooses the sentence that has least KL-divergence compared to the remaining document. Other systems are FreqSum, a word frequency based system, and TsSum, which uses topic signatures computed by comparing the documents to a background corpus. A Centroid based technique finds the sentences most similar to the document based on cosine similarity. We also combine the rank lists from these systems using the Borda count[3] and Reciprocal Rank Methods.

System	Rouge-1	Avg-Rank
Centroid	0.3641	1.94
FreqSum	0.3531	1.48
Greedy-KL	**0.3798**	2.2
LexRank	0.3595	1.72
TsSum	0.3587	1.88
BC	0.3621	**2.5**
RR	0.3633	2.46

Table 1: Effect of Fusion

We evaluated the techniques based on ROUGE-1, ROUGE-2 and ROUGE-4 Recall (Lin, 2004) using the parameters mentioned in (Hong et al., 2014). We report only ROUGE-1 results due to space constraints. We also computed Average-Rank for each system. Average-Rank indicates the average number of systems that the given system outperformed. The higher the average-rank the more consistent a given system. When systems are ranked based on ROUGE-1 metric, both Borda and Reciprocal Rank perform better than four of the five systems but couldn't beat the Greedy-KL method. Both combination techniques outperformed all five methods when systems are ranked based on ROUGE-2 and ROUGE-4. Even in case where Borda and Reciprocal Rank did outperform all the other systems, the increase in ROUGE scores were negligible. These results are contrary to what has been reported previously (Wang and Li, 2012) as neither of the fusion techniques performed significantly better than the candidate systems. The only noticeable improvement in all cases was in the Average-Rank. The combined systems were more consistent than the individual systems. These results indicate that Fusion can at least help us in improving the consistency of the meta-system.

One clear trend we observed was that not all combinations performed poorly, and summaries from certain techniques when fused together performed well (on both ROUGE score and consistency). To further investigate this issue we conducted another experiment where we try to make an informed fusion of various extractive techniques.

Due to space constraints we report results only on two families of summarization techniques: one is a graph based iterative method as suggested in (Erkan and Radev, 2004) and (Mihalcea and Tarau, 2004) and the other is the 'Greedy approach' where we greedily add a sentence that is most similar to the entire document, remove the sentence from the document and repeat the process until we have the desired number of sentences. We then chose three commonly used sentence similarity measures: Cosine similarity, Word overlap and KL-Divergence. Several other similar approaches are possible, for example TsSum and FreqSum are related in the sense that each method rates a sentence based on the average number of important words in it, the difference being in the way in which importance of the word is computed.

[3] https://en.wikipedia.org/wiki/Borda_count

We perform this experiment in a very constrained manner and leave it to the future experimenting with other such possible combinations.

	Graph	Greedy	Borda
Cosine	0.3473	0.3313	0.3370
Word Overlap	0.3139	0.3229	0.3039
KLD	0.3248	0.3429	0.3121
Borda	**0.3638**	**0.3515**	-

Table 2: Effect of 'Informed' Fusion

We generate summaries using all the possible 6 combinations of two approaches and three sentence similarity metrics. We then combine the summaries resulting from a particular sentence similarity metric or from a particular sentence ranking algorithm. The results in table 2 show that techniques that have a similar ranking algorithm but use different sentence similarity metrics, when combined produce an aggregate summary whose coverage is much better than the original summary. The aggregate summaries from the systems that have different ranking algorithm but the same sentence similarity measure do not beat the best performing system. Figures in bold indicate the maximum score for that particular approach. We have tested this for several other ranking algorithms like centroid based and LSA based and sentence similarity measures. The hypothesis holds in most cases. We consider this experiment to be indicative of a future direction of research and do not consider it in any way to be conclusive. But it definitely indicates the difficulties that might be encountered when attempting to fuse summaries from different sources compared to the limited improvement in the coverage (ROUGE scores). This combined with availability of a larger training set of document-summary pairs, which enables us to use several text generation approaches, is our principle motivation behind the proposed research.

3 Abstractive Summarization

Abstractive Summarization covers techniques which can generate summaries by rewriting the content in a given text, rather than simply extracting important sentences from it. But most of the current abstractive summarization techniques still use sentence extraction as a first step for abstract generation. In most cases, extractive summaries reach their limitation primarily because only a part of every sentence selected is informative and the other part is redundant. Abstractive techniques try to tackle this issue by either dropping the redundant part altogether or fusing two similar sentences in such a way as to maximize the information content and minimize the sentence lengths. We discuss some experiments we plan to do in this direction. An alternative to this technique is what is known as the *Generative approach* for text summarization. These techniques extract concepts (instead of sentences or phrases) from the given text and generate new sentences using those concepts and the relationships between them. We propose a novel approach of using statistical machine translation for document summarization. We discuss the possibilities of exploiting Statistical machine translation techniques, which in themselves are generative techniques and have a sound mathematical formulation, for translating a text in *Document Language* to *Summary Language*. In this section we highlight the research questions we are trying to address and issues that we might face in doing so. We also mention another approach we would like to explore which uses topic modeling for generating summaries.

3.1 Sentence Fusion

Most abstractive summarization techniques rely on sentence fusion to remove redundancy and create a new concise sentence. Graph based techniques similar to (Ganesan et al., 2010) and (Banerjee et al., 2015) have become very popular recently. These techniques rely on extractive summarization to get important sentences, cluster lexically similar sentences together, create a word graph from this cluster and try to generate a new meaningful sentence by selecting a best suited path from this word graph. Several factors like the linguistic quality of the sentence, informativeness, length of the sentence are considered when selecting an appropriate path form the word graph.

Informativeness of the selected path can be defined in several ways, and the choice defines how good my summary would be (at least when using ROUGE as a evaluation measure). In one of our experiments we changed the informativeness criteria from TextRank scores of words as used in the original approach in (Banerjee et al., 2015) to Log-Likelihood ratio of the words compared to a large background corpus as suggested in (Lin and Hovy, 2000). We observed that changing measure of informativeness produces a dramatic change in the quality of the summaries. We would like to continue working in this direction.

3.2 Summarization as a SMT problem

The idea is to model the text summarization problem as a Statistical Machine Translation (SMT) problem of translating text written in a *Document Language* to that in a *Summary Language*. Machine translation techniques have well defined and well accepted generative models which have been researched extensively over more than two decades. At least on the surface, the idea of modeling a text summarization problem as that of translation between two pairs of texts might enable us to leverage this progress in the field of SMT and extend it to abstractive text summarization, albeit with several modifications. We expect this area to be our primary research focus. While a similar approach has been used in the case of Question Answering (Zhang et al., 2014), to the best of our knowledge it has not yet been used for Document Summarization.

While the idea seems very intuitive and appealing, there are several roadblocks to it. The first and perhaps the biggest issue has been the lack of availability of a large training corpus. Traditionally SMT systems have depended on large volumes of parallel texts that are used to learn the phrase level alignment between sentences from two languages and the probability with which a particular phrase in the source language might be translated to another in the target language. The Text Summarization community on the other hand has relied on more linguistic approaches or statistical approaches which use limited amount of training data. Most of the evaluation benchmark datasets generated by conferences like DUC or TAC are limited to less than a hundred Document-Summary pairs and the focus has mainly been on short summaries of very few sentences. This makes the available data too small (especially when considering the number of sentences).

We hope to solve this problem partially using the *Supreme Court Judgments* dataset released by the organizers of *Information Access in Legal Domain Track*[4] at *FIRE 2014*. The dataset has 1500 Judgments with a corresponding summary known as a headnote, manually written by legal experts. The organizers released another dataset of additional 10,000 judgment-headnote pairs from the *Supreme court of India* spread over four decades, that are noisy and need to be curated. The average judgment length is 150 sentences while a headnote is 30 sentence long on an average. Using this we can create a parallel corpus of approximately 45,000 sentences using the clean data, and an additional 300,000 sentences after curating the entire dataset. This is comparable to the size of standard datasets used for training SMT systems.

Given this data is only semi-parallel and aligned at document level and not at sentence level, the next issue is extracting pairs of source sentence and target sentence. The exception being that both the source sentence and target sentence can actually be several sentences instead of a single sentence, the possibility being higher in case of the source than the target. This might seem to be a classic example of the problem of extracting parallel sentences from a comparable corpus. But there are several important differences, the biggest one being that it is almost guaranteed that several sentences from the text written in *Document Language* will map to a single sentence in the *Summary Language*. This itself makes this task more challenging compared to the already daunting task of finding parallel sentences in a comparable corpora. Another notable difference is that unlike in case of SMT, the headnotes (or the *Summary Language*) are influenced a lot by the stylistic quality of its author. The nature of headnotes seems to vary to a large extent over the period of four decades, and we are in the process of trying to figure out how this affects the sentence alignment as well as the overall translation process. The other major difference can actually be used as leverage to improve the quality of sentence level alignment. The headnotes tend to follow a general format, in the sense that there are certain points about the Court judgment that should always occur in the headnote and certain phrases or certain types of sentences are always bound to be excluded. How to leverage this information is one of the research questions we plan to address in the proposed work.

Another issue that we plan to address in particular is how to handle the mismatch between lengths of a sentence (i.e. multiple sentences considered to be a single sentence) in the *Document Language* when compared to the *Summary Language*. Currently two different languages do vary in the average sentence lengths, for example Ger-

[4] http://www.isical.ac.in/~fire/2014/legal.html

man sentences are in general longer than English. But in our case the ratio of sentence lengths would be almost 3:1 with the *Document Language* being much longer than their *Summary Language* counterparts. While most current translation models do have a provision for a penalty on sentence lengths which can make the target sentence longer or shorter, the real challenge lies in finding phrase level alignments when either the source sentence or the target sentence is too long compared to the other. This leads to a large number of phrases having no alignment at all which is not a common phenomenon in cases of SMT.

In effect we propose to address the following research questions:

- Exploring the major challenges that one might face when modeling Summarization as a Machine translation problem ?
- How to create a sentence aligned parallel corpus from a given document and its handwritten summary ?
- How to handle the disparity in lengths of sentence of *Document Language* and *Summary Language* ?
- How to reduce the sparsity in training data created due to the stylistic differences present within the Documents and Summaries ?

3.3 Topic model based sentence generation

The graph based approaches of sentence fusion mentioned above assumes availability of a number of similar sentences from which a word graph can be formed. It might not always be easy to get such similar sentences, especially in case of single document summarization. We wish to explore the possibility of using topic modeling to extract informative phrases and entities and then use standard sentence generation techniques to generate representative sentences.

4 Preliminary experiment

We would like to conclude by reporting results of a very preliminary experiment wherein we used simple cosine similarity to align sentences between the original Judgments and the manually generated headnotes (summaries). For a small training set of 1000 document-summary pairs, we compute the cosine similarity of each sentence in the judgment to each sentence in the corresponding headnote. Sentences in the judgment which do not have a cosine similarity of at least 0.5 with any sentence in the headnote are considered to have no alignment at all. The remaining sentences are aligned to a single best matching sentence in the headnote. Hence each sentence in the judgment is aligned to exactly one or zero sentences in the headnote, while each sentence in the headnote can have a many to one mapping. All the judgment sentences aligned to the same headnote sentence are combined to form a single sentence, thus forming a parallel corpus between *Judgment Language* and *Headnote Language*. Further we used the Moses[5] machine translation toolkit to generate a translation model with the source language as the *judgment (or the document Language)* and the target language as the *headnote (or summary language)*. Since we have not yet used the entire training data, the results in the current experiment were not very impressive. But there are certain examples worth reporting, where good results were indeed obtained.

4.1 Example Translation

Original: There can **in my opinion** be no escape from the conclusion that section 12 of the Act by which **a most important protection or** safeguard conferred on the subject by the Constitution has been taken away is not **a valid provision** since it **contravenes** the very provision in the Constitution under which the Parliament derived its competence to enact it.
Translation: There can be no escape from the conclusion that section 12 of the Act by which safeguard conferred on the subject by the Constitution has been taken away is not valid since it contravened the very provision in the Constitution under which the Parliament derived its competence to enact it.

The highlighted parts in the original sentence are the ones that have been changed in the corresponding translation. We can attribute the exclusion of *'in my opinion'* solely to the language model of the Summary Language. Since the summaries are in third person while many statements in the original judgment would be in first person, such a phrase which is common in the Judgment will never occur in the headnote. Similarly the headnotes are usually written in past tense and that might account for changing *'contravenes'* to *'contravened'*. We are not sure what the reasons might be behind the other changes. We plan to do an

[5] `www.statmt.org/moses`

exhaustive error analysis on the results of this experiment, which will provide further insights and ideas. We have reported some more examples in the appendix section.

Although not all translations are linguistically correct and many of them don't make much sense, we believe that by using a larger training corpus (which we are currently curating) and a better technique for creating a sentence aligned corpus the results can be significantly improved. Also currently the target sentences are not much shorter than their source, and we need to further work on that issue. Overall the idea of using SMT for document summarization seems to be promising and worth pursuing.

References

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Multi-document abstractive summarization using ilp based multi-sentence compression. In *24th International Joint Conference on Artificial Intelligence (IJCAI)*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

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*, pages 340–348. Association for Computational Linguistics.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*.

Kai Hong, Mitchell Marcus, and Ani Nenkova. 2015. System combination for multi-document summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 107–117. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer.

Dingding Wang and Tao Li. 2012. Weighted consensus multi-document summarization. *Information Processing & Management*, 48(3):513–523.

Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question retrieval with high quality answers in community question answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 371–380. ACM.

A Additional Examples

- The underlined parts in the original sentence are the ones that are correctly omitted in the target sentence. The striked out part in the original sentences are wrongly missing in the translation, affecting the comprehensibility of the sentence.
- The striked out parts in the Translation are the ones that are misplaced in the sentence. Boldfaced parts in the Translation are the ones newly introduced.
- The boldfaced parts in the Expected Translations are the corrections that are made compared to the actual translation.

Original:

The Act provides for levy of **two kinds of taxes called** the general tax and the special tax by the two charging sections 5 and 10 respectively. ~~**Seervai attempted to make out**~~ that the provisions of the charging sections 5 and 10 fixing Rs. 30000 and Rs. 5000 as the minimum taxable turnover for general tax and special tax respectively were found discriminatory and void under article 14 read with article 13 of the Constitution and he gave us several tables of figures showing how the imposition of the tax actually works **out in practice in hypothetical cases**.

Translation:

The Act provides for the levy of the general tax and special tax by the two charging sections 5 and 10 respectively. that the provisions of the charging sections 5 and 10 fixing Rs. 30000 and Rs. 5000 as the minimum taxable turnover for general tax and special tax respectively are discriminatory and void under ~~**art**~~ of the Constitution and he gave the several tables of figures showing how the imposition of the tax actually works.

Expected Tranlsation:

The Act provides for the levy of the general tax and special tax by the two charging sections 5 and 10 respectively. **Seervai attempted to make out** that the provisions of the charging sections 5 and 10 fixing Rs. 30000 and Rs. 5000 as the minimum taxable turnover for general tax and special tax respectively are discriminatory and void under **article 14 read with article 13** of the Constitution and he gave the several tables of figures showing how the imposition of the tax actually works.

Original:

The **learned trial** magistrate **believed the prosecution evidence rejected the pleas raised by the defence convicted** the appellants of the charge framed and sentenced them to undergo simple imprisonment for two months each. ~~**The appellate court**~~ confirmed the conviction of the appellants but reduced their sentence from simple imprisonment for two months to a fine of Rs. 50 or in default simple imprisonment for one month each.

Translation:

The Magistrate **found** the appellants of the charge framed and sentenced them to undergo simple imprisonment for two months ~~**guilty**~~. confirmed the conviction of the appellants but reduced their sentence from simple imprisonment for two months to a fine of Rs. 50 or in default simple imprisonment for one month each.

Expected Tranlsation:

The Magistrate found the appellants **guilty** of the charge framed and sentenced them to undergo simple imprisonment for two months. **The appellate court** confirmed the conviction of the appellants but reduced their sentence from simple imprisonment for two months to a fine of Rs. 50 or in default simple imprisonment for one month each.
