HEXTAC: the Creation of a Manual Extractive Run

Pierre-Etienne Genest, Guy Lapalme, Mehdi Yousfi-Monod
RALI-DIRO
Université de Montréal
P.O. Box 6128, Succ. Centre-Ville
Montréal, Québec
Canada, H3C 3J7
{genestpe,lapalme,yousfim}@iro.umontreal.ca

Abstract

This article presents an attempt to establish an upper bound on purely extractive summarization techniques. Altogether, five human summarizers composed 88 standard and update summaries of the TAC 2009 competition. Only entire sentences of the source documents were selected by the human "extractors", without modification, to form 100-word summaries. These summaries obtained better scores than any automatic summarization system in both linguistic quality and overall responsiveness, while still doing worse than any human abstractive summarizer.

1 Introduction

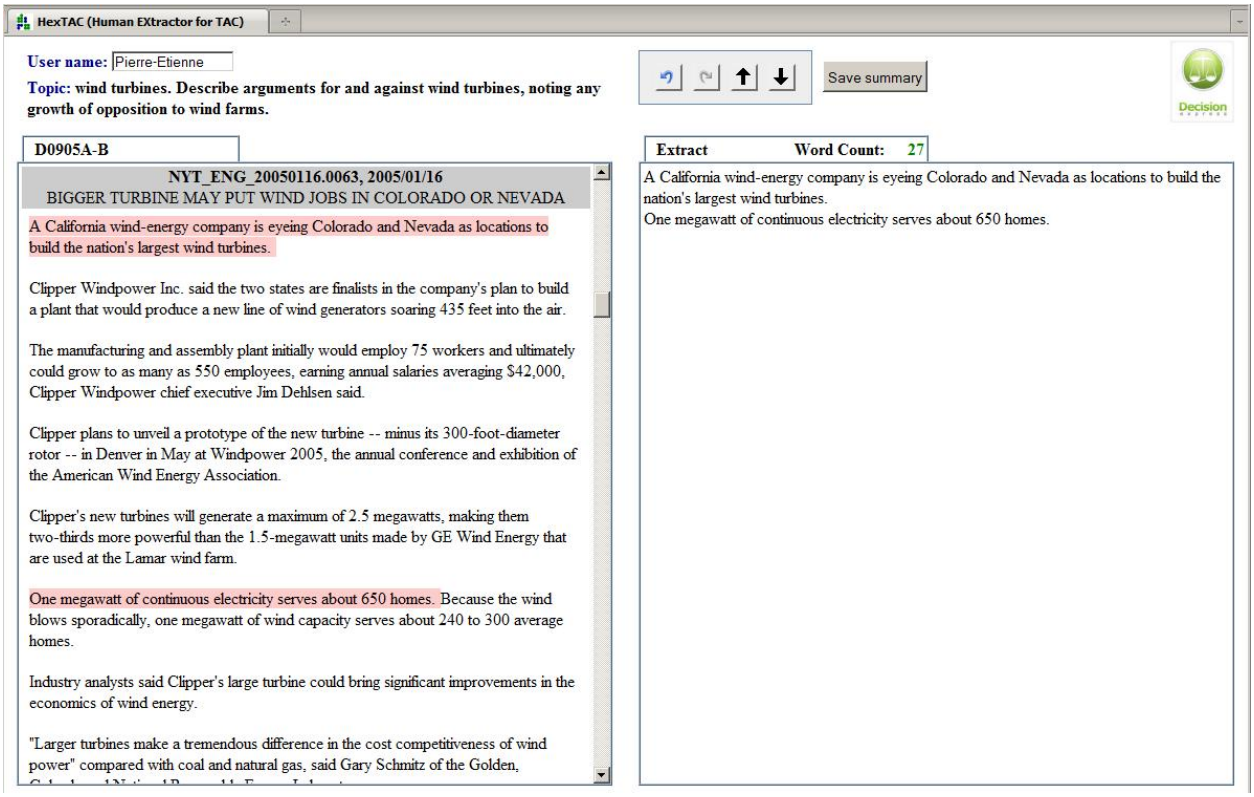
Year after year, notably at the Document Understanding Conference (DUC) and later the Text Analysis Conference (TAC), the best-performing summarization systems have been sentence extraction-based rather than abstractions of the source documents. However, in those conferences and in the literature, human-written model summaries are used for comparison and automatic evaluation. The model summaries are abstractive rather than extractive summaries. While these *gold standards* show how far computers are from achieving what humans can, it does not address the more restrictive – but probably no less interesting – question of how well one can solve the simpler problem of extracting sentences from documents for summarization. Some researchers in the summarization community even seem to consider this problem of extracting important sentences from groups of documents as *solved*!

We were also motivated in further studying extractive methods because in some areas, namely in the judicial domain, extraction is a method of choice because summary sentences can be safely used as jurisprudence without worrying that the original might have been *interpreted* by the human abstracter.

With the support of Hoa Trang Dang and members of the TAC steering committee, our team created an extractive manual run for this year's Update Summarization task, called *Human EXtraction for TAC* (HEXTAC), which appeared in the 2009 competition as baseline number 3.

This experiment was designed with the goal of quantifying how well humans could perform at extraction and comparing the results with automatic summarizers. HEXTAC is thus an attempt to establish an upper bound on purely extractive summarization techniques. Five human extractors composed the 88 standard and update summaries for the TAC 2009 competition. Only entire unedited sentences in the source documents were selected, to create summaries of 100 words or less. In practice, this meant selecting about three to five sentences out of the 232 (on average) in each cluster, a tedious and finally quite harder task that we had originally anticipated. We are glad that we have developed computer systems for that!

The methodology and context of the experimentation are described in section 2. Section 3 presents the results and discusses them. We conclude with lessons that we learned during this exercise.



1. Pick one of your assigned topic to work on. Always begin with part A, the standard summary.
2. Read the topic and all 10 of the articles in full (to know all of the information and to have read each sentence at least once). All of the information in part A must be well remembered to avoid any repetition of information in part B.
3. Extract sentences that answer the topic and best summarize the source documents. Select preferentially sentences that can be understood on their own (avoid problems of referential clarity).
4. Refine your sentence selection to bring the summary under the limit of 100 words, while maximizing the information content.
5. Re-order the sentences of the extract to improve readability and save your work.
6. Make sure to complete the update summary – part B – immediately after writing the standard summary. Follow the same steps as for part A, with the added criterion that extracted sentences must avoid repetition of information included in part A articles.

Figure 1: Screen shot of the HEXTAC interface for human extractors and the guidelines given to the human extractors. The left part of the screen contains the texts of all documents from a cluster from which only full sentences can be selected and dragged into the right part to build the summary. The total number of words is updated as each sentence is added to the summary. Sentences added to the summary can be removed and or reordered using drag and drop. Undo and redo of extraction operations are possible.

2 Methodology and Context

2.1 Interactive Human Extraction Interface

In order to simplify the handling and creation of extractive summaries, we developed a browser-based interface. It enables the users to build a summary step by step in a convenient environment. The summarizers can access the data, check which ones they should work on, save their summaries, and consult or modify them later. In the background, the system logs the access times and other peripheral data.

The extractive summaries are created on a single, user-friendly page, shown at the top of figure 1. From the top down and left to right, it contains a user name box, a topic description, the articles and their meta-data (ID, date of publication and title), the editing tools, the save button, and the extractive summary area.

All articles of a given cluster are shown one after the other. The text of the articles has been previously segmented into sentences although the original paragraph structure of the articles is kept. When the user hovers over a part of the text, the sentence covered by the mouse pointer is highlighted and its number of words is shown. The total number of words in the summary should this sentence be added is also temporarily updated. This sentence can then be double-clicked to be put into the summary area. The selected sentences are building blocks for the summary. They can be later removed or re-arranged in any order desired, but they can never be modified by the user in any way (the text areas of the browsers are *read-only*). No summary of more than 100 words can be accepted by the system as a valid submission, though they can still be saved temporarily. The whole system works equally well with drag-and-drop and with double-clicking and using buttons. Undo and redo buttons are also included for convenience.

This interface is an adaptation of a summary revision interface that we have developed in a project dealing with judgements in cooperation with NLP Technologies[1] (Farzindar and Lapalme, 2004) (Chieze et al., 2008).

[1] http://www.nlptechnologies.ca

2.2 Experimental Context

There were 44 topics to answer in the TAC 2009 competition, with a standard and an update part for each – 88 total summaries. The human extraction task was divided unevenly between five computer scientists, all specialized in NLP with experience in automatic summarization, including the three authors, who volunteered to do this manual work. They all used the interactive interface, while following the specific guidelines shown at the bottom of Figure 1. The summaries were all composed within about a week and submitted five days after the deadline for the automatic runs. As our laboratory was also submitting automatic runs (IDs 10 and 8) developed by the first author, he only started working on the manual process once our automatic runs had been submitted.

Table 1 shows how many summaries were written by each human extractor (HE) and the average time in minutes it took him to complete one summary (Part A or B). A total of 30 man-hour were required to complete the 88 summaries.

Summarizer ID	# summaries	Avg. time (min)
HE1	18	17
HE2	18	16
HE3	12	27
HE4	24	24
HE5	16	17
Average	18	20

Table 1: Number of summaries out of the 88 composed by each human extractor and the average time in minutes it took them.

2.3 Feedback from Participants

Following the experiment, we met with the human extractors who participated in the HEXTAC experiment to receive feedback on their experience.

The foremost opinion was that the interactive interface made everything a lot easier. According to the feedback, this tool saved a lot time and even helped in organizing thoughts. Using text editors and copy-paste would have made this task an even greater chore than it already was to some.

The extractors felt some frustration because of the inability to make even the smallest of textual modi-

fications to the sentences. Cutting down one or two words in one sentence would, in some cases, have permitted them to fit their preferred choice of sentences into the summary. Also, some sentences had great content but could not be included because of an unresolved personal pronoun anaphora or relative time reference, which would be easy for a human – and in some cases for a machine as well – to resolve.

The topic queries also caused some headaches, because they often times asked for a broad description or a list of several related events/opinions/etc., whereas the articles would only offer sentences with one piece of information at a time. Choosing which sentences to extract became a huge difficulty in those circumstances and, in general, subjective choices of what content to prioritize in the very limited space has been a big issue. At times, the tradeoff between quality of content and linguistic quality was also difficult to deal with.

Most extractors complained about the time commitment and the repetitiveness of the task. It was reported that doing several summaries in a row might decrease the level of attention to details of the extractors. On the other hand, many felt that the more extracts they completed, the easier the task became.

3 Results and Discussion

3.1 TAC 2009 Scores

HEXTAC is considered a baseline in TAC 2009, and it has run ID 3. Table 2 shows the scores of pyramid, linguistic quality and overall responsiveness for HEXTAC, the best score obtained by an automatic system, and the average of human abstractors.

Part A	Pyramid	Ling. Qual.	Ov. Resp.
Abstracts	0.683	8.915	8.830
HEXTAC	0.352	7.477	6.341
Best Auto	0.383	5.932	5.159
Part B			
Abstracts	0.606	8.807	8.506
HEXTAC	0.324	7.250	6.114
Best Auto	0.307	5.886	5.023

Table 2: Scores for HEXTAC when compared to the best automatic systems and the humans abstracts for parts A and B.

The overall responsiveness score is significantly higher for HEXTAC than for any automatic summarizer. This might come to a surprise to some, since we used pure extraction whereas the best systems often use sentence compression and/or reformulation. This superiority probably comes from the much higher linguistic quality of HEXTAC summaries, while the pyramid scores were on par with the best systems of the competition.

The human extracts still receive far lower scores than abstracts in all evaluation metrics, as expected. The difference in performance is easily understandable given the severe limitations that pure extraction puts on our summarizers. In particular, the amount of content that can be included in an extract is much less than in an abstract, as shown by the pyramid scores. The difference in linguistic quality probably arises because of some sentences with unresolved references that were still included by extractors, and mostly because pure extraction does not grant the flexibility required to create text that flows as nicely as abstracts can. We notice that the difference in linguistic quality between extracts and abstracts is much less noticeable than the difference in pyramid scores, thus hinting that good language quality can still be achieved without even any modification to the sentences.

We believe that these evaluation results can be interpreted as a soft upper-bound on what can theoretically be done by purely extractive methods. "Soft" because the extractors were not as competent as professional summarizers probably would have been and we have strong reasons to believe better extracts than those submitted exist. The known tradeoff between content and linguistic quality could play a role here, for example. The variations in the performance of the different extractors and the low inter-extractor agreement are other indicators that better extracts could likely be written. Nevertheless, the gap between the manual extracts and abstracts is so large that we can safely claim – now with numerical results to show for – that the performance of pure extraction summarization will never come close to what can be achieved by abstraction.

On the other hand, the results show that even using pure extracts, there is still significant improvements that can be made to improve the quality of the summaries we create automatically. It seems

that perhaps a lot of progress could still be made in aspects that increase linguistic quality like sentence ordering and avoiding redundancy, unresolved references, bad grammar in reformulated sentences, etc.

3.2 Inter-Extractor Agreement

We computed the inter-extractor agreement on a small sample of 16 summaries that have been written twice. On average, each extract has 0.58 sentence in common with one written by an other extractor, on an average of 3.88 sentences per summary. This gives roughly a 15% chance that a sentence selected by one extractor is also selected by another one working on the same topic. We consider this level of agreement to be very low, although it can be expected because of the redundancy in the source documents of a multi-document summarization corpus. Indeed, we have observed that some sentences were even repeated verbatim in more than one article of the same cluster, not to mention all the sentences which were nearly identical and had the same information content.

The scores obtained individually by each human extractor, on average, were very different for each one and in each metric, as can be seen in Table 3.

	Pyramid	Ling. Qual.	Overall Resp.
HE1	0.278	8.222	7.556
HE2	0.297	7.611	5.333
HE3	0.340	7.000	5.917
HE4	0.378	7.583	7.125
HE5	0.392	6.063	4.125

Table 3: Average scores for each human extractor.

The small sample size can partly explain the high variance of the scores between human extractors. Some summaries were harder to complete than others, because of the topic or the available sentences. Also, the extractors have had different types of experiences with summarization, they possessed different levels of knowledge on the topics given to them, and a different level of proficiency in English, which was not the native language of any of them.

3.3 HEXTAC as a ROUGE model

As part of our experiment, we ran the automatic summarization evaluation engine ROUGE on all the runs except for the baselines and human extracts, using HEXTAC as the model – we call this HEXTAC-ROUGE. We wanted to see how this evaluation would compare to the ROUGE evaluation based on 4 human abstraction models, with jack-knifing (the ROUGE metric used in TAC). The correlation coefficients between HEXTAC-ROUGE, ROUGE, and the overall responsiveness scores of all the participating systems (runs 4 through 55) are given in Table 4. All the ROUGE scores use ROUGE2.

	Part A	Part B
HEXTAC-ROUGE–ROUGE	0.80	0.85
HEXTAC-ROUGE–O. Resp.	0.78	0.91
ROUGE–O. Resp.	0.97	0.94

Table 4: Correlation coefficients between HEXTAC-ROUGE, ROUGE, and the overall responsiveness scores.

HEXTAC-ROUGE is fairly well correlated to both ROUGE and the overall responsiveness scores, with correlation coefficients between 78 and 91%. This shows that HEXTAC summaries are potential models for extractive systems to compare themselves with, obtaining better evaluation scores, as we have seen before. We believe that training a sentence selection engine on the manual extracts, using HEXTAC-ROUGE, is easier and more straightforward than training on ROUGE scores obtained from abstracts, because the model sentences can be found in the source documents.

4 Conclusion

The HEXTAC experiment presents a successful, reusable approach to human sentence extraction for summarization. We have developed a comprehensive methodology with detailed guidelines, and we now have a better idea of how much time is required to complete the summaries. We have observed that an interactive interface such as the one we used is an invaluable tool, in part because it reduced the amount of time spent on writing each extractive summary, thus keeping our extractors happier.

Viewed as an upper-bound on purely extractive summarization techniques, the competition results for HEXTAC lead to two main conclusions. First,

that significant improvements to current sentence selection engines and sentence ordering schemes can still be made since the current automatic summarizers do not achieve results comparable to those of human extracts yet. Second, that since there are large, now quantifiable gaps between the scores of human abstracts and extracts – mostly in the amount of content that can be included –, developing techniques to extract smaller segments than sentences or to compress or reformulate sentences is essential to make great improvements to the current techniques in the long-term.

We view the HEXTAC extracts as an interesting alternative to using the ROUGE scores based on abstracts for sentence selection engine training. The main attraction lies in the fact that the training is supervised through data that corresponds to the same challenge. Similarly, sentence ordering could perhaps be trained using HEXTAC summaries to supervise the learning.

More comprehensive information from humans, in the form of sentence evaluation, would lead to even much more valuable information for the purpose of supervised training. Humans could list all the sentences in the cluster that could potentially be of use in a summary, excluding anything with low content or bad linguistic form. They could then rate the sentences in that list and identify which ones are redundant and could not be included together. While this would be a monumentally larger amount of work, the gathered data would be more directly usable and the inter-annotator agreement would likely increased, improving the reliability of the data.

5 Acknowledgements

Great thanks to Atefeh Farzidar, president of NLP Technologies, who accepted that we adapted the revision interface for his project. Thanks to Fabrizio Gotti and Florian Boudin for their valuable contribution as human extractors.

References

Emmanuel Chieze, Atefeh Farzindar, and Guy Lapalme. 2008. Automatic summarization and information extraction from canadian immigration decisions. In *Proceedings of the Semantic Processing of Legal Texts Workshop*, pages 51–57. LREC 2008, may.

Atefeh Farzindar and Guy Lapalme. 2004. Legal texts summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out, Conference held in conjunction with ACL04*, Barcelona, Spain, jul.