DECREMENTAL FEATURE-BASED COMPACTION

BalaKrishna Kolluru, Heidi Christensen and Yoshihiko Gotoh

Department of Computer Science,
University of Sheffield, Sheffield S1 4DP, UK
{b.kolluru, h.christensen, y.gotoh}@dcs.shef.ac.uk

ABSTRACT

In this paper we present the results of repeated feature-based compaction applied as a part of DUC 2004 task 1 – 75-byte short summary, on a set of printed news stories. We report the performance of a system built using *tf*idf* and *named-entities* as the main features employed to retain the most relevant parts of text, while compacting it. Multiple stages of the compaction, whilst arriving at the final summary ensure that we retain the text based on informativeness of information from the already chosen information-rich zones. From the nature of the summaries produced at 75 bytes, we conjecture that there exists a certain threshold for compacting a news story, beyond which the quality and readability of a summary deteriorate.

1. INTRODUCTION

Ever since Luhn [1] published his work in 1958, extractive summarization of news stories has been actively pursued. In the many variations of this technique, a number of measures have been adopted, tested and evaluated. Prominent amongst these extraction techniques are paragraph extraction as implemented by Mitra *et al* [2], sentence extraction using document-sentence similarity measures [3], the use of co-reference chains for summarization [4] and summarization using lexical chains as reported by Barzilay and Elhadad [5].

In 2000, Knight and Marcu employed a technique that uses a parser and a context-free grammar score combination [6] to decide which word or a combination of words can be removed to tailor a compressed sentence, Jing [7] uses a number of features such as grammar and context to make a set of rules that guide the system to compress the sentence by reducing the seemingly irrelevant information. Daume and Marcu [8] took this approach to sentence compression further, by extending the algorithm to an entire document.

We seek to compress the most relevant parts of a news story be it broadcast news or printed text, using salient features observed at the phrase (or chunk) level: ie, the *compaction* of news stories. Our major interest is speech summarization and thence our system is built to incorporate in it both the traditional IR features such as *tf*idf* and speech specific features such as the confidence score (explained in detail in Section 5). We limit the scope of our system to IR features for the DUC 2004, whilst briefly discussing impact and nature of speech-related features.

In our implementation, we extend the concept of partial-parsing (chunking) for information extraction. For experiments in this paper, we have used Abney's partial-parser called Cass [9][1]. We chunk a phrase into multiple-level semantic segments and then estimate the relevance of each chunk based on its feature-value, deleting the lowest ranked chunks at each level, thus compressing the phrase by removing the irrelevant chunks.

In this paper we present the results of repeated compaction of news stories until the desired size of the summary is realised. As will be explained in detail in Section 3, we apply a multi-stage compaction at various levels of the selected news stories, using *tf*idf* and named-entity as main features. The news stories used for the experiments in this paper were from the DUC evaluation data (2003 data for development of the system and 2004 data for evaluation). Experimental results were evaluated using an automatic evaluation tool called ROUGE [12]. The model summaries against which the automatic summaries were rated are human-generated non-extractive summaries. We also provide a brief insight into the extension of this compaction technique on broadcast news (on automatic speech recognizer, ASR transcripts) with speech-related features.

[1] Why use a chunker?
With Cass [9] in conjunction with Brill's POS tagger [10], the levels of automata split a sentence in chunks at different levels as desired for our system. As explained by Abney in the documentation for Cass and the usage of Cass for IR techniques [11] along with its hand-in-hand approach with Brill tagger, which in turn could be trained both for speech and text data) made it a good candidate for experiments.

2. DATA

2.1. Development Data

For the development of our system compaction, we used the DUC 2003 test data containing about 620 news stories from Associated Press (APW), New York Times (NYT) and Xinhua (XIN). The news stories were paragraph delimited, had punctuation and case information.

Christensen *et al* [13] found that summary-worthy information in printed news stories mostly occurs at the beginning of the news story, thus we opted to select first paragraph of each news story. A preprocessor was used to select the first paragraph for compaction using the paragraph tag as the pointer. This selected text was then split into single phrases using a comma, an underscore, or a full stop as a delimiter. The compaction algorithm as explained in Section 3 was applied to this selected text.

2.2. Evaluation Set

We have used DUC 2004 test data containing 500 news stories from Associated Press (APW) and New York Times (NYT) to evaluate our approach. The news stories were case-sensitive and punctuated.

In order to arrive at a 75-byte summary, we performed a series of experiments to arrive at selecting the nearest sentence break after first 300 characters for compaction, in lieu of the first paragraph as was used in our development set. Once this selection was made, we applied our compaction algorithm.

3. APPROACH

The approach can enumerated in 4 steps:

1. Select the information rich part of the text (first 300 characters for evaluation data and first paragraph for the development data).

2. Preprocess: Split the given text into various phrases based on punctuation marks (, . ! or _), then further split this text into smaller chunks.

3. Apply the multi-stage compaction algorithm to these chunks in decremental levels of granularity (explained in detail later in this section).

4. Post processing, like removing punctuation etc, to arrive at a 75-byte summary.

3.1. Principles

Our system was built with an intention of summarizing speech, broadcast news in particular, exploiting various features from both speech and text. However, we restrict our discussion in this section to text-related features and the speech-related features are described in Section 5. The aim is to arrive at an optimal combination of all these features.

Each phrase in the selected text (first paragraph or the nearest sentence around the 300-character) is tagged with its respective part-of-speech (POS) tag using the Brill tagger [10] trained on the Switchboard corpus. Each tagged phrase is then chunked using Abney's partial-parser (chunker) [9].

For the experiments discussed in this paper, we seek the following features:

1. term-relevance as in *tf*idf* score. The sum total of *tf*idf* for each word in a chunk *c* are added together to obtain a total *tf*idf* score for *c*. For instance, if the chunk is

   rained for weeks.

   then its *tf*idf* score will be

   $$tfidf_c = tfidf_{rained} + tfidf_{for} + tfidf_{weeks} \quad (1)$$

2. named-entity: We used GATE [14] to markup the named entities in the news stories.

To retain the significant information of the chunks from various phrases, we apply the compaction algorithm to chunker output levels, shown in Table 1. We increase the granularity of the chunker output at different stages of the compaction. We keep decrementing the length of the text until we arrive at 75-bytes.

3.2. Algorithm

We pursued a multi-stage repeated feature-based *compaction*, there by losing a (hopefully) irrelevant word or a set of words at every stage. For the results discussed in this paper, we initially selected approximately the first 300 characters, then calculated the chunk-relevance score (sum total of *tf*idf* for that chunk) to level 1 (as shown in table 1), compacting them to approximately 200 characters. These were further compacted 125 characters using the same compaction algorithm but to level 2 chunker output, and finally to 75 characters, which was DUC 2004 evaluation requirement for short summaries. We retained the chunks with maximum *tf*idf*. As shown in Figure 1 we repeated this for each stage of compaction applied at various levels

Level-1 chunker output
231 deaths have been blamed on Mitch the National Emergency Commission said Saturday El Salvador where 140 people died in flash floods
Level-2 chunker output
231 deaths have been blamed on Mitch, the National Emergency Commission said Saturday El Salvador where 140 people died in flash floods

Table 1. Chunker Output for a sentence from DUC 2004 data

of chunker output.

To conform with DUC 2004 requirement of 75-byte limit for short summary, we employed additional processing to increase the informativeness of our summary:

1. All the named-entities identified as PERSON in sequence, could be condensed to retaining the last word in that sequence, before applying the compaction algorithm. For example,

   President Bill Clinton

   would be condensed to

   Clinton

2. All numbers transcribed in letters were converted to digits, using a look-up table, where *twenty-one* would be converted to 21, before applying the compaction algorithm.

3. We opted to remove all the punctuation from our summaries, thus saving a few more characters after applying the compaction algorithm.

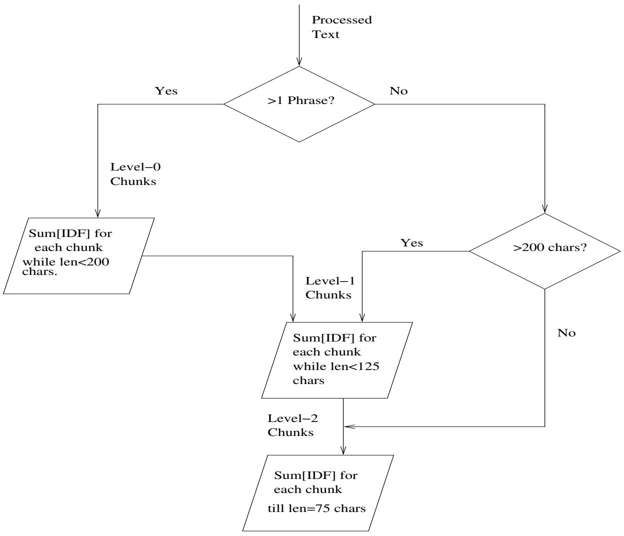
Processed Text
>1 Phrase?
Yes
No
Level−0 Chunks
Sum[IDF] for each chunk while len<200 chars.
Yes
>200 chars?
Level−1 Chunks
No
Sum[IDF] for each chunk while len<125 chars
Level−2 Chunks
Sum[IDF] for each chunk till len=75 chars

Fig. 1. Illustration showing the implementation of multiple stages of compaction, applied to various stages of a chunker output.

4. RESULTS

We ran ROUGE [12] on 500 test documents as distributed by NIST. Figure 2 shows the unigram, bigram, trigram and 4-gram position of our system among all other system runs. The reference summaries are human-generated non-extractive summaries. For evaluation, human and automatic summaries had the same size limitations to avoid any skewed evaluations.

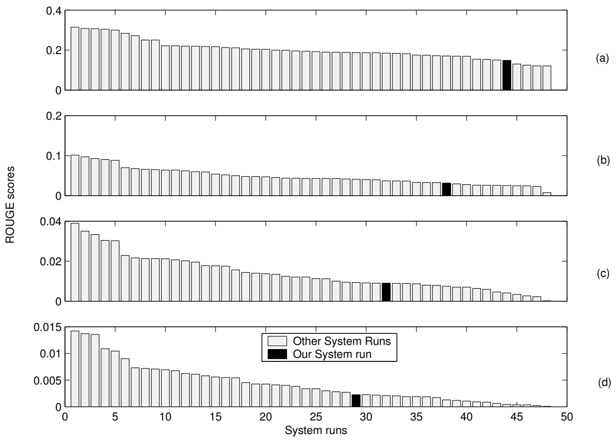

Fig. 2. Illustration showing the relative performance of our system wrt all the system runs that participated in DUC 2004 based on ROUGE scores. (a) is unigram, (b) is bigram, (c) is trigram and (d) is 4-gram listing.

The relative performance of our system, compared with other systems, improves from unigram through to 4-gram. We infer the following:

1. We opted to select relevant chunks rather than set of significant yet disjoint individual words from a news story, thus extracting the most significant part from a paragraph (or its equivalent) at a level higher than a word and lower than a sentence (or a paragraph). This could explain the relative improvement in the performance of the system from unigram to 4-gram.

2. We feel the performance of our system can be enhanced by incorporating co-reference as a salient feature because named-entities have a higher *tf*idf* on average. If the individual reference of each named-entity along with its occurrences through out a text were to be tapped, this will help enhance identify the significant part of a given text.

3. The low unigram score is a direct consequence of the mismatch between the evaluation system and our compaction algorithm. In other words, we sought to retain a chunk of words with high relevance from each phrase. Thus retaining the informativeness of a word as we retain the word with its context rather than collating a list of words with high relevance.

Table 2 show the precise ROUGE scores that our system run produced on the 500 news stories of DUC 2004.

ROUGE	*Avg*	*Med*	*Max*	*Min*
Unigram	0.14795	0.15043	0.17770	0.11322
Bigram	0.03108	0.03374	0.04120	0.01563
Trigram	0.00896	0.01041	0.01236	0.00265
4-gram	0.00225	0.00286	0.00313	0.00015
L	0.12845	0.13110	0.15504	0.09656
W-1.2	0.07849	0.08061	0.09366	0.05906

Table 2. ROUGE scores of the 500 documents (news stories). W-1.2 is the consecutive matches of length 1.2 and L is the first n-words in peer and model summaries

Table 3 shows a sample of the summaries as generated by our system. To subtabulate the results of our compaction algorithm we have tabulated the various summaries into three categories **The Good**, **The Bad** and **The Ugly** on the basis of coverage of the news story. While **The Good** had good coverage and moderate readability, **The Bad** had moderate coverage (sometimes very good readability) and **The Ugly** was poor at both. We could reason the following from the results:

1. One major factor for **The Ugly** is the occurrence of relatively rare words, sometimes named-entities, which often resulted in high *tf*idf* score for those chunks. We expect to overcome this drawback by using more features to determine the significance of a chunk.

2. The assumption that the most relevant information occurs in the first 300 characters is a very unreliable heuristic, more so in longer news stories.

3. Usage of co-reference resolution would certainly improve **The Bad** to **The Good**, as it would partly account for dangling anaphora.

5. EXTENSION TO BROADCAST NEWS SUMMARIZATION

We are working on porting this approach used for DUC 2004 to speech. Speech summarization involves a number of additional operations, such as automatic sentence boundary detection, automatic topic detection and confidence estimation, all of which come with a certain error, thus making summarization that much more interesting.

We are currently working on broadcast new stories from about 114 30-minute news broadcasts from the TDT–2 broadcast news corpus[2], totalling 43 hours of speech [13]. Each programme spanned 30 minutes as broadcast, reduced to around 22 minutes once advert breaks were removed, and contains on average 7–8 news stories, giving 855 stories in total. In addition the estimated word error rate (WER) of the ASR transcripts was 32.0%. All transcripts have been segmented at two levels: 1) utterance boundaries (fully automatic) and 2) story boundaries (the individual news stories were hand-segmented as part of the TREC/SDR evaluations).

We seek to apply the same compaction as discussed earlier, but with more features. We are looking at incorporating confidence score as one of the features. These are confidence scores output by the recognizer based on posterior probabilities from the acoustic model recurrent networks and MLPs [16]. In other words this is probability of how correctly a speech recognizer has identified a hypothesised word.

Spoken language is less grammatical than written language, thus affecting the performance of tools such as part-of-speech taggers and partial-parsers. However, our style of compaction does not take into account grammatical accuracy in identifying or splitting a sentence into chunks.

[2] The TDT–2 [15] corpus has been used in the NIST Topic Detection and Tracking evaluations and in the TREC–8 and TREC–9 spoken document retrieval (SDR) evaluations.

The Good
· Negotiations to form next government have become deadlocked and opposition
· announcing authorisation for the European common currencys use in trade
· Spain and Portugal put finishing touches Saturday on an IberoAmerican summit
· Americas economic history is being rewritten In energy as in businesses
· 2 giants in the energy patch were in merger talks is biggest sign yet that
The Bad
· the House voted Friday to condemn the NaziSoviet nonaggression pact of 1939
· a UN war crimes tribunal on Monday convicted 3 prison officials and guards
· We want to be sure that the only aim of the trial is to show the truth and
· FBI agents this week began questioning relatives of the victims of bombing
· And that excuse is bin Laden the man Washington calls Enemy No 1 and
The Ugly
· A joint statement by the IMF and the Brazilian government said the 2 sides
· in dollars cents But Ewing president of National Basketball Association
· and fighting for rebel leader Ernest Wamba dia Wamba said on Friday Besides
· he warned the Norths communist leaders not to squander a chance to achieve
· 500 Palestinian delegates next week US Secret Service agents have arrived

Table 3. Summaries generated by our system for DUC 2004 data. Each line corresponds to a 75-byte summary for a new story.

Also under the microscope is the usage of named-entity relevance. By this we mean, ability to distinguish the significant named-entities from the insignificant ones. A *tf*idf* kind of measure just for named-entities. For example, we would like to identify *Clinton* or *Blair* from *Deborah Wang* who is a reporter presenting the news story.

5.1. Results

The preliminary observations of our experiments with speech look promising. We applied the same compaction algorithm described in Section 3, but we started the compaction from first 500 characters of the broadcast news story and decremented it to about 250 characters before finally arriving to about 150 characters. We opted to use ROUGE for evaluation, however our gold-standard summary was a human-extractive $\sim$ 150 character summary. That is we picked out the most significant sentence(s) and resized them to about 150 characters by deleting insignificant words like *the, a, an*.

We had a ROUGE unigram score ranging from 0.38 to 0.6 on a set of 7 news stories, averaging about 609 words each. However, given that the reference summaries for the ROUGE were human-extractive summaries for the news stories, this high ROUGE score is not very significant, although it underlines accuracy of *tf*idf* application in our algorithm. Here are a sample of the type of summaries we got from our preliminary experiments:

> *this morning at high tide waves lapped over the bar some of the hardest rate dropped in the area of santa cruz south highway woman in monterrey is completely*

> *federal agents continued calm in the area ron murphy north carolina today for the thirty one year old white male after the blast that killed one person*

From the kind of results we infer the following:

1. Obviously, WER plays a very important role and the readability of a summary very much depends on it.

2. We expect the ROUGE scores to be significantly lower if measured against a human non-extractive summary compared to the ROUGE results on extractive summaries, especially given the WER.

3. Given the high accuracy of named-entity tagging (up to 93% precision [17]) on broadcast news, named-entity relevance would be a very reliable feature.

6. CONCLUSION

The results indicate that our compaction approach is feasible, although there is plenty of room for improvement. This system is only based on two features, *tf*idf* and a variation of *named entity* identification, in combination with the premise that the first few hundred words hold the summary-worthy information. Where it fails, a detailed examination of the nature of summaries indicated that the

75-byte limit was a bit too taxing on summary quality particularly for those news stories which violated our premise on the basis of which our compaction was implemented.

We are currently working on embedding confidence scores and named-entity relevance score along with *tf*idf* in equation (1) with the speech data, i.e. speech recognizer output. In order to automate the process in totality we are employing automatic sentence and topic boundary detection on the speech data. We are also engaged in statistical analysis of the speech data to arrive at the optimal combination of all these features and possible expansion of our premises.

Acknowledgements

We would like to acknowledge the usage of Brill tagger [10] and Abney's partial-parser [9] as components of our compaction approach. Also, we would like to thank Steve Renals, Louise Guthrie, Joseph Guthrie, Horacio Saggion and Mirella Lapata for their valuable contributions during the discussions we have had with them, while implementing the system. This work was supported by EPSRC grant GR/R42405 (S3L: Statistical Summarization of Spoken Language).

7. REFERENCES

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, pp. 155–164, April 1958.

[2] Mandar Mitra, Amit Singhal, and Chris Buckley, "Automatic text summarization by paragraph extraction," in *ACL workshop on Intelligent Scalable Text Summarization*, 1997, pp. 39–46.

[3] Jaime Carbonell and Jade Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *the proceedings of SIGIR*, August 1998.

[4] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas, "Using coreference chains for text summarization," in *ACL Workshop on Coreference and its Applications*, June 1999.

[5] Regina Barzillay and Micheal Elhadad, "Using lexical chains for text summarization," in *ACL workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 1997, pp. 10–17.

[6] Kevin Knight and Daniel Marcu, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," *Artificial Intelligence*, vol. 139, pp. 91–107, July 2002.

[7] Hongyan Jing, "Sentence reduction for automatic text summarization," in *6th Applied Natural Language Processing Conference (ANLP)*, Seattle, Washington, May 2000.

[8] Hal Daume III and Daniel Marcu, "A noisy-channel model for document compression," in *the Conference of the Association of Computational Linguistics (ACL 2002).*, 2002.

[9] Steve Abney, "Cass: A fast robust partial-parser," http://www.vinartus.net/spa/.

[10] Eric Brill, *A Corpus Based Approach to Language Learning*, Ph.D. thesis, University of Pennsylvania, 1993.

[11] Ken Church, Steve Young, and Gerrit Bloothooft, Eds., *Corpus-based Methods in Language and Speech*, chapter Tagging and Partial Parsing, Kluwer Academic Publishers, 1996.

[12] Chin-Yew Lin, "Recall oriented understudy of gisting evaluation," http://www.isi.edu/∼cyl/ROUGE.

[13] Heidi Christensen, Yoshihiko Gotoh, BalaKrishna Kolluru, and Steve Renals, "Are extractive text summarisation techniques portable to broadcast news?," in *Automatic Speech Recognition and Understanding*, Virgin Islands, US, November-December 2003.

[14] "General architecture for text engineering, university of sheffield," http://www.gate.ac.uk.

[15] C. Cieri, D. Graff, and M. Liberman, "The TDT-2 text and speech corpus," in *Proceedings of DARPA Broadcast News Workshop*, 1999.

[16] Gethin Williams, *Knowing What You Dont Know: Roles for Confidence Measures in Automatic Speech Recognition*, Ph.D. thesis, University of Sheffield, 1999.

[17] Yoshihiko Gotoh and Steve Renals, "Information extraction from broadcast news," in *Philosophical Transactions of the Royal Society of London, series A*, vol. 358, issue 1769, pp. 1295–1310. April 2000.