Summarizing with *Roget's* and with FrameNet

Terry COPECK, Alistair KENNEDY,
Martin SCAIANO, Diana INKPEN, Stan SZPAKOWICZ
School of Information Technology and Engineering
University of Ottawa
800 King Edward Avenue
Ottawa, Ontario, Canada K1N 6N5
{terry,akennedy,diana,szpak}@site.uottawa.ca
mscai056@uottawa.ca

Abstract

We submitted runs from two different systems for the update summary task at TAC 2009. The first system refined its use of *Roget's Thesaurus*, moving beyond 2008's semantic relatedness to compute an entropy-based uniqueness measure, with improved results in summary construction. The other system, our first use of deeper semantic knowledge, represents sentences as FrameNet types in a conceptual graph. Pairwise similarity comparisons identify the sentences most central to the document collection content and best candidates for a summary. Our AESOP submission suggests that together, the group of TAC participants tend to select summary-worthy sentences.

1 Introduction

In 2009 the University of Ottawa NLP research group once again used the summarization task provided by the Text Analysis Conference to focus our efforts in this area and to assess the performance of our software on the task. And once again we submitted two quite different systemsimplementing algorithms developed by graduate students. One system builds on and moves beyond the design which its author used last year, while the other is the first attempt by someone in our group to perform extractive summarization using deeper semantic knowledgeof the content of a text.

Two additional summarization task runs were submitted under the name of a consultancy operated by a group member. Where each of the two main University of Ottawa submissions expresses a plausible and comprehensive approach to the task, the two latter runs were derivative. One is based on a slight reworking of the system our group used at DUC in previous years, while the other blended the sentence rankings of the three other submissions made this year, a practice we have employed in the past (Copeck *et al*. 2006, 2007). The first serves as an internal benchmark, while the second continues our investigation of the results of using multiple approaches to produce a summary. The results of these submissions appear in the charts without further comment.

We also made a single submission to the AESOP task. This work was not intended to offer a practical methodology for evaluating summaries, but rather to complement the efforts of others in this area by adding to our general knowledge about evaluating summary quality with information about the performance of TAC participants as a group.

Finally we once again augmented the corpus of TAC test documents annotated with Summary Content Unit (SCU) information we developed and maintain, with the data from this year's conference. This corpus is available to TAC participants on request to NIST.

1.1 System Framework

Our practice of submitting different systems to the TAC conference from one year to the next requires us to change the summarization algorithm in use repeatedly. This has led us over years of repeated participation to develop a modular system architecture which facilitates making that change easily.

Processing begins by extracting the <TEXT>elementfrom each XML document provided by NIST. Limited efforts are made to improve results by correcting certain punctuation errors and replacing encodings with the characters they signify. The extracted string is then broken into sentences. The boundary detection algorithm we use appears accurate enough that the SCU-marked corpus we maintain is intelligible to others. Document text bodies are then provided to the summarization module in 'normalized' format, with one sentence per line and a blank line signifying a paragraph break. The sentence list is in document order, with documents concatenated in the order in which they are provided by NIST.

The summarization module reorders this list on the basis of how suitable it deems each sentence for use in the sort of summary desired, assigning a numerical value to each rank. The summarization framework then constructs the actual summary by including as many of the top-ranked sentences as the limit of words allows. Only complete sentences are used and at present no pruning is done. Should the count of words in the selected top-ranked sentences fall too far short of the number allowed, an attempt is made to pack the summary by adding lower-ranked shorter sentences.

The system framework next tries to make the summary more fluent by replacing certain pronouns with their referents when it is reasonably confident in its identification of them. Sentences in the summary are then reordered to maximize the number of content words in co-occurring in adjacent pairs—a sort of lexical chaining. The summary construction procedure also tries to remove any extraneous material such as wire-service tags it can distinguish, correct unmatched single or double quotes, and fix errors in capitalization at the sentence beginning.

The system framework described here is relatively stable. Modifications between conferences are generally made only in response to faults highlighted by the new data encountered. Although the framework deals with subjects of research interest (sentence boundary detection, reference resolution, how to make text more fluent), these are not issues which we actively investigate. What we *are* focused on is the core question of extractive summarization. The following two sections describe our group's work on this task for TAC 2009.

2 *Roget*'s Thesaurus-Based Sentence Ranking

At last year's TAC competition we experimented with using the 1911 public domain edition of*Roget*'s *Thesaurus*[1] (Kennedy & Szpakowicz, 2008), to enhance tf.idf based sentence ranking (Copeck et al. 2008). While that approach was moderately successful, this year we tested a new method using *Roget's Thesaurus*which produces multiple scores for each sentence based on that sentence's relatedness to a variety of concepts. This series of scores is useful not only for ranking sentences, but also for maximizing information contained in the summary.

In this approach once again we make use of Open Roget's Thesaurus, specifically the semantic distance function *semDist*. *SemDist* takes two words and gives a score from 0..16 indicating how related these two words are, 16 being the highest score. This function is used to measure the relatedness between words in the query, ie the topic information request, and words in a sentence.

2.1 Sentence Ranking

A sentence is ranked based on its similarity to the query. To do this the distance between each word $w_j$in a sentence S is measured against each word $q_i$in the query Q. For each sentence a series of scores x_1 .. x_n is created where x_i corresponds to the maximum score of any word in S to the query term q_i.

$$x_i = \max semDist(w_j, q_i)$$

A single score for a sentence *score(S)* is generated by taking the sum of the scores x_1 .. x_n for *S.*

$$score(S) = \sum_{i=0}^{n} x_i$$

[1] We used Open Roget's Thesaurus: http://rogets.site.uottawa.ca/

Score(S) can then be used to rank sentences in order of their relevance to the query. This score will be called a "sentence score" as opposed to the "uniqueness scores" that are discussed later. To create a score for an entire summary *SUM* we take the sum of the sentence scores $score(S_i)$ for each sentence $S_i$which appears in *SUM*. This creates a score *score(SUM).* In this way we produce an overall score for each summary.

2.2 Sentence Ranking Evaluation

Evaluation is done on a SCU-labeled corpus. This corpus is produced by our group. We take the sentences from the summaries generated by previous years'peer systems and map each sentence where possible back to its original location in the corpus. This results in a partially annotated corpus in which sentences can be marked as either having SCUs (along with SCU ids and weights), having no SCUs (appeared in summaries with no SCU assigned), or unknown (sentences not appearing in any system summary). For evaluation purposes unknown sentences are ignored.

To evaluate this ranking we calculated the Macro Average Precision or MAP, which is the average precision score achieved across all summaries. In computing this value sentences containing one or more SCUs are positive and sentences known to contain no SCUs are negative. The average is taken from processing the amalgamation of DUC/TAC data from five tasks across the years: 2005, 2006, 2007, 2007 Update Pilot, and 2008. When this method was implemented with the 1911 *Roget'sThesaurus* a MAP score of 60.9% was achieved. The MAP score from last year's system is 55.8% on these data sets, while computing cosine similarity on a tf.idf representation of them produces a MAP score of 55.4%. This represents a noticeable improvement over last year's data. For purposes of comparison, a summary composed of randomly-selected sentences will give a MAP of about 49.6% while ranking sentences by length gives a MAP of 58.2%. That said, selecting sentences based on length alone would result in summaries of just one or two sentences.

2.3 Maximizing Information

Selecting sentences that bring new information to a summary is harder than simply ranking sentences based on their relatedness to a set of concepts. It is also desirable to minimize the amount of redundant information presented in a summary.However, even if two sentences overlap somewhat in content, it does not mean that they do not each contain significant unique information as well. Our goal is to maximize unique information rather than to eliminate redundancy. To do this we measure the entropy of the scores x_1 .. x_n in our summary. Because our measurement of information is performed after the summary is generated, it cannot guide summary generation To deal with this we generate several summaries and pick one that has a good balance of information and a high score from the sentence-ranking component.

The method we employ uses the entropy of the similarity scores x_1 .. x_n.As described above, each sentence is given a set of scores x_1 .. x_n to determine how close the sentence is to each word in the query. A set of scores $SUMx_1$ .. $SUMx_n$ can be calculated for the summary itself where $SUMx_i$ is the sum of all x_i scores from each sentence S_i in the summary.

The assumption behind this approach is that a summary which describes just one part of the query well is more likely to contain redundant information. A summary which contains information related to all parts of the query will have a higher entropy score and is less likely to contain redundant information.

The information entropy of $SUMx_1$ .. $SUMx_n$ is calculated thus:

$$H(SUM) = -\sum_{i=0}^{n} p(x_i)\log_b p(x_i)$$

Where the *probability* $p(x_i)$ is defined as:

$$p(x_i) = \frac{SUMx_i}{score(SUM)}$$

The result, *H(X),* is the entropy of our summary.

2.4 *Final System*

These measures are not applied during the process of building a summary. Instead a variety of possible summaries are produced and the measures are used to select the one most likely to be best. Accordingly, all possible summaries of 8..100 words are generated from the top 10 ranked sentences (ranking based on *score(S)* values). The summary with the top sentence score and the summary with the highest uniqueness score are identified in the resulting set of summaries and a maximum score *maxScore* and maximum uniqueness *maxH* are determined.. We use these maximum sentence and uniqueness scores to normalize the sentence and uniqueness scorescalculated for every summary generated. These normalized scores are then combined to produce each summary's final score:

$$finalScore(SUM) = \frac{H(SUM)}{\max H} + \frac{score(SUM)}{\max Score}$$

The summary with the highest final score is used in our submission.

2.5 *Evaluation*

Performance of this entropy-based uniqueness detection system was evaluated by generatingthe 80..100 word summary sets for all topics in the 2005, 2006, 2007, 2007 Update, and 2008 data[2]. Sentences whose SCU count is unknown were not included. Each summary was evaluated on the total SCU score, the number of unique SCUs, the number of redundant SCUs, the number of positive sentences, and the number of negative sentences.

As a baseline, results are provided for a system which uses the sentence score, but does not include any uniqueness score. Summaries for this baseline are generated by greedily selecting sentences in order of sentence score. Table I shows that the system we develop using an entropy-based uniqueness measure outperformed the baseline on almost every criterion.

[2] The 2005-07 data were originally intended to build 250-word summaries.

2.6 *Results*

Although using an entropy-based uniqueness score does generate summaries with slightly more redundant SCUs, the total SCU count and number of unique SCUs is actually higher. Further, the total SCU score increased when this uniqueness measure was used. The entropy-based method also produces summaries with fewer sentences, approximately 60-80 fewer than the baseline method. This is not surprising, since it may sometimes select summaries with as few as 80 words, while the greedy method will attempt to add a new sentence in such a case. Although the method under development selected fewer sentences, it retained almost as many positive sentences as the greedy method; it was mostly negative sentences that were lost.

It should be noted that our method of ranking sentences will favor longer sentences to some degree, since the longer a sentence is the more likely it is that one of its words w_i will have a high similarity score with q_i. This is actually advantageous as summaries with fewer sentences tend to be more readable. This is apparent in our results where our system was ranked 6th overall in readability and 4th when baseline systems are excluded.

In terms of our SCU rating we were near the middle of the pack. This is a bit disappointing, but an indifferent SCU score may be a result of selecting longer sentences. Those sentences have a good chance of containing SCUs,but they take up more room in a summary and allow fewer to be included.

Table I shows the average scores of all TAC 2009

		Baseline	**Entropy-Based Uniqueness Detection**
Total Score		1762	1851
SCU Count	Total	715	743
	Unique	605	630
	Redundant	110	113
Sentence Count	Positive	435	421
	Negative	312	242

Table 1: Evaluation Results for Baseline and Submission Systems

participants on the responsiveness measure, which we consider to be the most significant measure of summary quality. Here we ranked 10th overall, or 8th when the baseline systems are excluded.

3 Summarizing Using Deeper Semantic Knowledge

This system employs a knowledge representation based on conceptual graphs (Sowa, 1984) where type information is provided by the frame taxonomy in FrameNet (Fillmore *et al*. 2004). The motivation for the system is to work towards natural language understanding. Although the system is at present underdeveloped, we found that much could be learned from attempting this task even using shallow semantics.

The core of the system is the frame-labeling moduledescribed by Scaiano and Inkpen (2009). Thiscomponent identifies and labels frames and frame elements in text using a dependency parse tree and machine learning techniques. Frames allow for events, objects, or ideas to be represented in a uniform way. The system presented by Scaiano and Inkpen has subsequently been improved in performance, and now supports some co-reference resolution. It also now uses conceptual graphs as the central knowledge representation.

3.1 Method

After each sentence from a document set has been parsed and its semantic representation built, it is compared to representations of the sentences composing the topic information request and to each previously-processed sentence. Each comparison produces a similarity value; these values are summed for each sentence, as it compares to every other sentence. We hypothesize that the accumulated value measures how effectively this sentence represents the most frequent ideas or themes appearing in the document collection. Sentences with the highest values are used to produce the summary.

3.2 Comparison

Our present sentence comparison uses shallow semantics techniques. The knowledge representation is a graph: nodes represent concepts such as people, locations, and frames; edges represent relations such as semantic roles (frame elements). When comparing two graphs, all the concepts in the first graph are compared to all the concepts in the second graph. The comparison of the concepts differs depending on the concept type, as follows.

- Concepts represented by frames are compared only to other concepts that are also frames. Frames of identical type are assigned a comparison value of 1. Frames are also

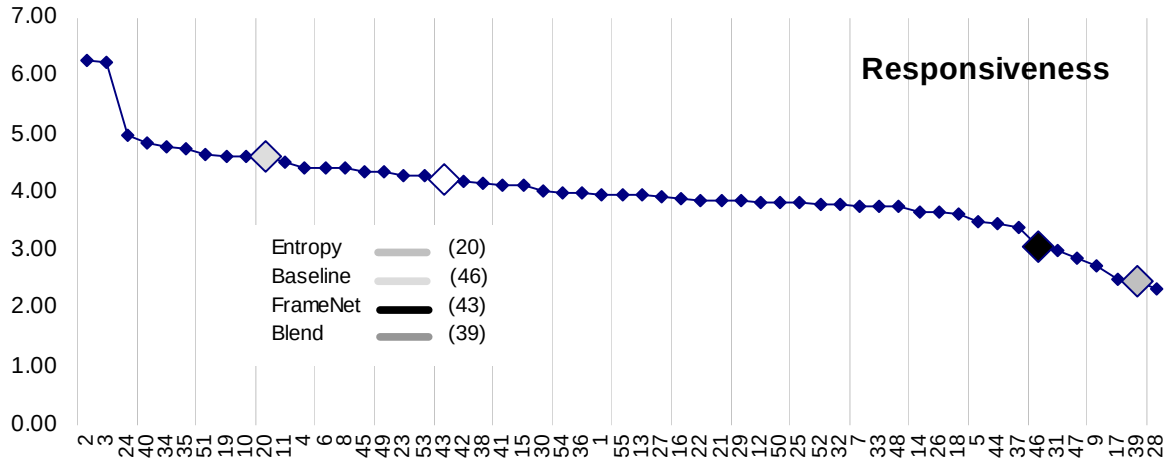

Figure 1: University of Ottawa Update Summary Responsiveness Ranking

comparable if one is an ancestor of the other through either the inheritance or perspective relationships in the FrameNet ontology. The further the frames are separated by relationships, the more general the ancestor becomes compared to the descendant, and thus less likely to be describing exactly the same event. For each relationship that separates the two frames, a decay value is applied to the initial ranking of 1.

- Named locations, such as London and Europe, are compared using the WordNet ontology (Fellbaum 1998), to see if one location contains another, or if they are synonyms.
- Non-location named entities are compared by counting the number of matching words in their names. The lower of the two percentages of matching words becomes their comparison value.

3.3 Conclusions

The system's output ranked very low. While the results were poor, the experience of preparation and process immediately highlighted a number of ways in which we could improve our system for next year. Enhancements being considered are:

- Using a more complete representation including temporal and modal logic;
- Application of other ontologies and knowledge to improve the comparison;
- Comparison of frames or predicates should also include comparison of the related roles;
- Removal of sentences from the summary with redundant information. Our current method is likely to pick sentences with identical frames and concepts;
- Empirical testing to assign tuned comparison and decay values.

4 AESOP

In addition to the update summary task, the TAC 2009 summarization track turned the tables on its participants, in a manner of speaking, by setting them the task of Automatically Evaluating the Summaries of Peers (AESOP). Indeed, the test data for this task were the submissions we made to the same year's update task.

Although we at the University of Ottawa had no insight into the problem sufficient to prompt us to develop software, one related question did suggest itself: given that most of our peers perform the extractive summarization that employs the common coin of a shared set of candidate sentences, to what degree do our most frequent selections measure up well? Do we as a group tend to pick good sentences, and does the rate at which a sentence is used in peer summaries correlate with its suitability to be used in a summary?

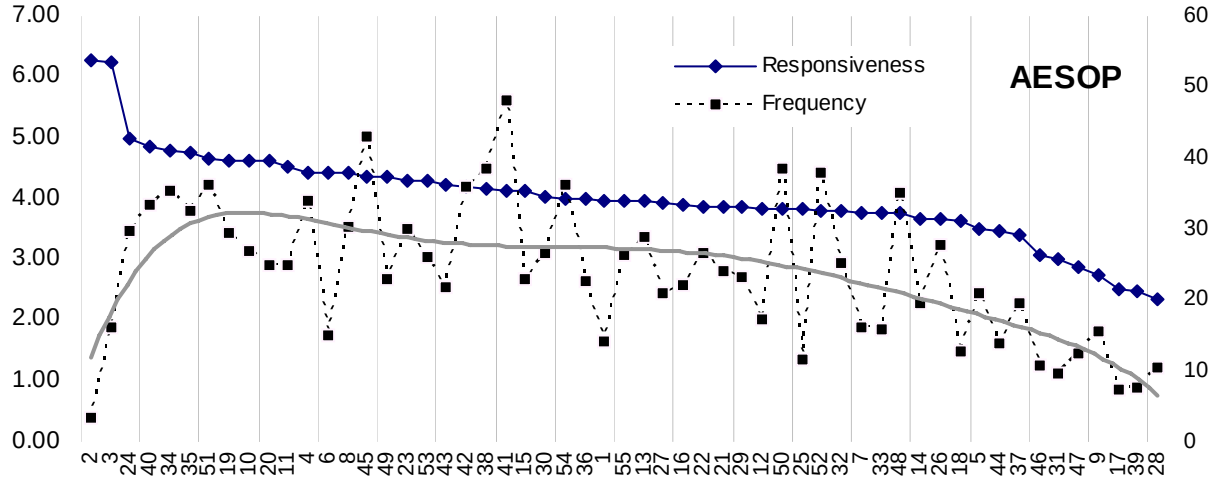

Figure 2: TAC 2009 Peers Responsiveness and Sentence Frequency Ranking

To address the question we proceeded in this way. The documents which make up a topic set were read and their sentences stored in a hash. Each summary of this document set was then processed. Although most peer summaries are formatted with one sentence per line, some are presented as paragraphs. Extra blank lines and irregular punctuation also occur and require correction. The sentences in the summary were then matched against the hash of topic document sentences and each match recorded. Some peers end summaries with whatever fragment of a sentence will fit under the limit; to the degree these matched a document sentence non-trivially they too were recorded. This operation augmented the sentence hash with summary frequency data; its most-used sentences having the highest counts. Peer summaries were then processed a second time and ranked with the sum of the frequency counts of the sentences they employ. A peer's overall rank is the average of its rank on all topics. Generic and update topic sets were not distinguished.

4.1 AESOP Results

Figure 2 shows the results. A polynomial trendline based on the highly-variable sentence frequency data correlates on inspection fairly closely with a peer's overall responsiveness score. Ignoring the special cases of peers 2 and 3, this suggests that we peers as a group are indeed tending to pick summary-worthy sentences. Those of us who tend to use sentences used by others tend to produce more responsive summaries.

The data for peers 2 and 3 corroborate this analysis. Peer 2 is a randomly-chosen model summary. Written by a human author, this highly responsive summary will only by chance match sentences in the topic set. The peer 3 data is even more pertinent. This summary is composed of the best sentences in the document set as determined by human assessors—it is our gold standard. Although it better matches frequently-used sentences than does peer 2, its low frequency score shows that collectively we still have some distance to go—many of the sentences in the collection most suitable to use in a summary are not yet often picked.

A final observation. Averaged across 88 generic or update summaries, the best extractive summary is almost as responsive as one written by a human author[3]. This is wind in the sails for those who are committed to summarization by sentence extraction.

Acknowledgment

Partial support for this work comes from the Natural Sciences and Engineering Research Council of Canada.

References

Alistair Kennedy and Stan Szpakowicz. 2008. "Evaluating Roget's Thesauri". In Proceedings of ACL-08: HLT, Columbus, Ohio, USA, Association for Computational Linguistics, 416-424.

Christine Fellbaum. 1998. WordNet: An electronic lexical database. MIT Press.

Charles J. Fillmore, Josef Ruppenhofer and Collin Baker. 2004. FrameNet and representing the link between semantic and syntactic relations. *In* Churen Huan and Winfried Lenders, editors, *Frontiers in Linguistics, volume I of Language and Linguistics Monograph Series B. Institute of Linguistics, Academia Sinica, Taipei*, 19-59.

John F. Sowa. 1984.Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA.

Martin Scaiano and Diana Inkpen. 2009. Automatic Frame Extraction from Sentences. *In Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligencel*, Springer Berlin/ Heidelberg, 110-120.

Terry Copeck, Anna Kazantseva, Alistair Kennedy, Alex Kunadze, Diana Inkpen and Stan Szpakowicz. 2008. Update Summary Update. In, *Proceedings of the Workshop on Automatic Summarization* (TAC 2008), Gaithersburg, Maryland, USA, 2008.

Terry Copeck, Diana Inkpen, Anna Kazantseva, Alistair Kennedy, Darren Kipp and Stan Szpakowicz. 2007. Catch What You Can. In *Proceedings of the Workshop on Automatic Summarization* (DUC 2007), HLT/NAACL-2007.

[3] with sentences in random order, which will tend to lower its responsiveness score

Terry Copeck, Diana Inkpen, Anna Kazantseva, Alistair Kennedy, Darren Kipp, Vivi Nastase and Stan Szpakowicz. 2006. Leveraging DUC. In *Proceedings of the Workshop on Automatic Summarization* (DUC 2006), HLT/NAACL-2006.