NII at the 2006 Multilingual Summarization Evaluation

David Kirk Evans
National Institute of Informatics
Tokyo, Japan
devans@nii.ac.jp

ABSTRACT

In this paper I detail the implementation of an extraction-based summarization system that uses sentence clustering and named entity identification as main features for the 2006 Multilingual Summarization Evaluation. I discuss some of the failings of my system, and what can be done to improve it.

1. INTRODUCTION

I am interested in developing summarization systems that take advantage of the diversity of opinion and cultural background present in documents from multiple countries written in multiple languages. I would like to create systems that are able to present the user with factual information that is agreed upon by multiple sources, continuing a history of leveraging similarity and repetition in multi-document summarization, but also point out important differences between the documents. An interesting area to investigate with multilingual summarization is how opinions and viewpoints differ based on the country of origin, and in my research I would like to focus on automated methods to present and contrast this information across documents and languages.

1.1 Motivation / Conecpt

One area that I would like to explore is showing differences between documents. In the genre of news documents reports often cover conflicts, where there is a controversy and two or more sides represented in the coverage. I envision the role of summarization to be that of succinctly stating factual information that is agreed upon in the reporting, identifying the main points of controversy, and identifying and describing the position of the various parties involved in the conflict.

Recent research in multi-document summarization has leveraged repetition to identify “important” information that is agreed upon by multiple sources [5], but less work has focused on how documents differ. I would like to take a first step in this direction by identifying a conflict in the document set, and the various sides and positions that they take. A summary will then identify the point of conflict, and describe the parties involved and their viewpoints. I feel that multilingual summarization is a particularly interesting area to perform this kind of research because given domestic pro-national agendas it is likely that strongly differing positions will be held by participants from different countries over international disputes, and these might be easier to identify than more subtle positions that might occur domestically. Looking across countries also introduces cultural aspects that could identified using cultural and historical databases, tying in external resources in a limited manner.

2. SYSTEM DESCRIPTION

For this first summarization system developed at the start of my post-doctoral stay at the National Institute of Informatics I did not have much time to spend developing the system, and chose to focus on one aspect of the system. I wanted to first focus the system on named entities, and used open source tools to quickly build the system in approximately a week and a half.

Focusing on named entities is a first step to identifying different parties involved in conflict reported in the document set. By first identifying the named entities, and later identifying statements of opinion and the corresponding opinion holders, I can start to build a model of participants that are involved and their viewpoints.

In order to build a system quickly, I made use of the openNLP toolkit[1] for English part-of-speech tagging and named entity identification.

2.1 Approach

The general approach that I took for the summarization system is to take the English and machine translated English text as input, compute various features over the sentences, and perform a search over the summary hypothesis space to select the “best” summary consisting of extracted sentences with respect to the hypothesis evaluation function. This isn’t a particularly new approach to solving the summary sentence selection problem (see, for example [3, 4]), but it did allow me to dynamically evaluate the contribution of the features depending on the entire set of sentences selected for the hypothesis. The particular score for any

[1] http://opennlp.sourceforge.net/

```
1. Create empty candidate hypothesis set
2. Generate successor hypotheses from candidate hypothesis set
3. Score successor hypotheses
4. If successor hypothesis set is empty, proceed to 6
5. Set successor hypothesis set to candidate hypothesis set, go to 2
6. Sort candidate hypotheses, output best as summary
```

Figure 1: Summary generation algorithm

given sentence can change based on the other sentences under consideration for selection in the hypothesis. The basic algorithm is outlined in Figure 1.

2.2 Hypothesis Feature set

Seven features are computed for each sentence:

- Named Entity
- Cluster
- TF*IDF
- Sentence position
- Trigger word
- All lower case
- Quotes

The features are evaluated in the context of a hypothesis, a set of sentences and word count that represents either a complete summary, or a summary in the process of being completed. The Named Entity and Cluster feature scores can vary depending on the other sentences in the hypothesis, the others are static. The TF*IDF score for the summary is simply the term frequencies of all the words in the summary multiplied by their inverse document frequency taken from a large corpus of AP newswire text, the sentence position is inversely proportional to the sentence number, and the trigger word score is based on how many words are found in a small manually created trigger-word file based on a manual inspection of the training data. If a sentence's words are all lower case or it looks like the sentence is a quote, those features are set to 1, 0 otherwise.

The system uses SimFinder [5] to cluster input sentences based on similarity. Sentences with similar content are clustered together, resulting in larger clusters for information that is reported on frequently in the document set. The cluster score for each hypothesis is based on the size of the cluster that each included sentence is a member of, with a large penalty for including sentences from the same cluster to avoid redundancy in the summary.

The emphasis of this summarization system is on using named entities to drive content selection. I would like to build a system that tracks sophisticated information about the named entities involved in the document set, performing cross-document and cross-lingual named entity disambiguation, identifying opinion statements and quotes, and tracking the named entities that hold the stated opinions. In this first system I did not have the time to implement such sophisticated techniques, but decided that using named entities to drive content selection would be an interesting approach to take to start my investigation of named entities in multi-lingual summarization.

I used the openNLP toolkit to part-of-speech tag and identify named entities in the input text. I used the default English models, and extracted all tagged named entities from the input documents. I did not perform any cross-document named entity disambiguation, and simply aggregated named entity counts across all documents based on exact string matches. I split the named entities up into five classes, person, location, organization, date, or time, and computed the number of occurrences of each named entity across all documents.

The named entity feature score for a sentence depends on the coverage of the named entities in the hypothesis to which the sentence is being considered for inclusion. For each named entity in the sentence, if it is already mentioned by one of the sentences in the hypothesis, the score is zero, otherwise the named entity feature score is incremented by the number of times the named entity is mentioned in the document set. The named entity classes are weighted such that people are more important than organizations, which are more important than locations, followed by dates and times. The motivation behind this feature scoring is that the summary should include the most important people and organizations in the document set, but once an entity has been included redundantly mentioning the named entity again should be avoided.

2.3 Hypothesis evaluation

Once all of the features for the sentences have been computed, the system performs a breadth-first search of the hypothesis space to find the best summary under the hypothesis evaluation metric. Ideally, this summary hypothesis evaluation metric would be able to rank potential summaries from poor to best, but of course the ranking is a difficult task and is only heuristic used for the search.

When creating the new hypotheses (step 2 in Figure 1) from the candidate hypothesis set, the feature scores for every sentence are computed for inclusion into the candidate hypothesis. A maximum of five new hypotheses are created and added to the successor hypothesis set by creating new hypotheses combining the current candidate hypothesis and each of the top five scoring sentences.

Due to time constraints the hypothesis evaluator is simply a linear sum of the feature scores for each sentence in the hypothesis, or 0 if the summary length exceeds the maximum summary target length. The hypothesis evaluation function was not tuned based on the training data.

In future versions of this summarization system, I would like to empirically determine a hypothesis evaluation function using ROUGE to evaluate a large set of candidate evaluation functions. If the process is computationally tractable I would like to use linear regression, otherwise, a standard search over a randomized space of evaluation functions should allow for optimizing this portion of the system.

After some number of iterations, all summary hypotheses are near the summary word limit, and it is not possible to create any new summary hypotheses. The summary hypotheses are sorted based on their scores, and the top summary is output. Since all evaluation will be performed using

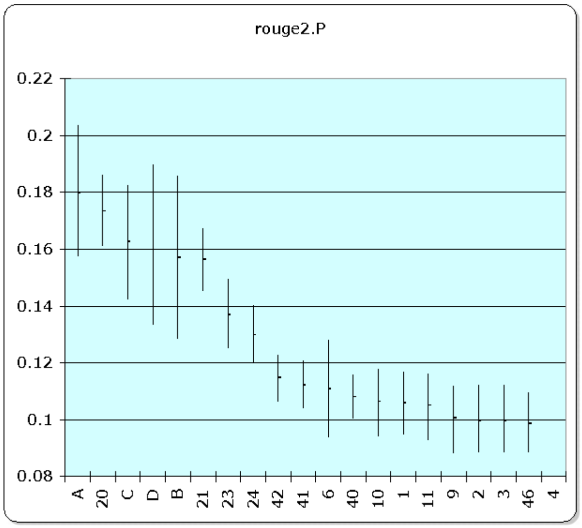

Figure 2: Rouge 2 Precision graph. My system is number 46.

ROUGE, sentence order does not impact summary scores, so I have not spent any time on determining sentence order.

3. RESULTS

The 2006 Multilingual Summarization Evaluation used the ROUGE package [7] to evaluate peer summaries, and reported both ROUGE-2 and ROUGE-SU4 scores, run with the same settings used for the Document Understand Conference in 2006. Figures 2 to 5 show the results in graphical form. In general, my system (number 46) performed poorly. In Section 3.1 I will discuss some causes that led to this poor performance, and discuss future work that should improve performance in Section 4.

For both the precision and recall variants of ROUGE-2 my system performed in the bottom range of systems. Taking the confidence intervals into account, there are approximately four groupings of systems, system 20 and 21, followed by 23 and 24, then systems 42 to 11, and finally systems 9 through 46. The last set of systems also have a large overlap in their confidence intervals, so the separation there is not as clear as between the top scoring systems.

For both the precision and recall variants of ROUGE-SU4 my system also performed at the bottom of the group. The differences in the SU4 metric are a bit stronger, with systems 21, 23, and 24 clearly better than all lower groups. Systems 20 and 21 again perform very well, followed again by 23 and 24. The next group contains systems 42 to 3, with my system, 46, performing statistically significantly worse than some (but not all) of the systems in the larger grouping.

In all the cases, there was one system, system 4, that scored 0 under each metric. When I reviewed a selection of the peer summaries, there was clearly some sort of bug with system 4 that interspersed control characters between each character in the summary, causing the poor performance. Without the control characters, the system would have received some non-zero score.

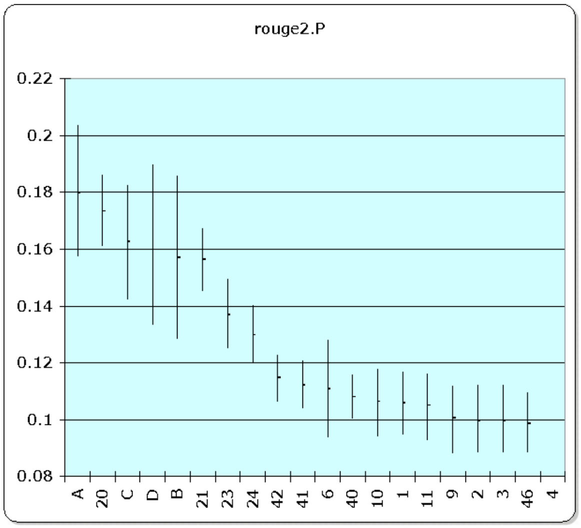

Figure 3: Rouge 2 Recall graph. My system is number 46.

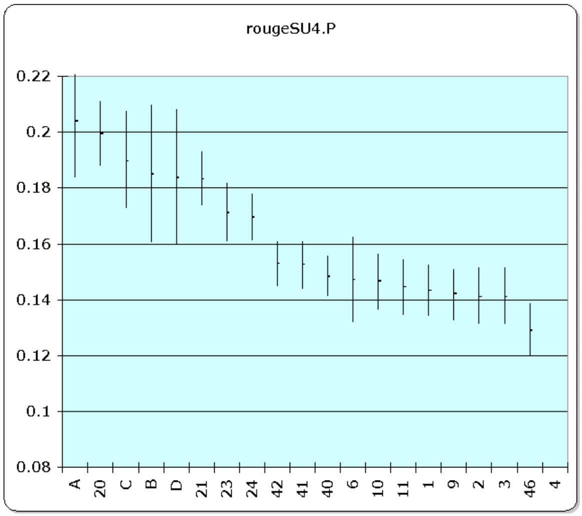

Figure 4: Rouge SU4 Precision graph. My system is number 46.

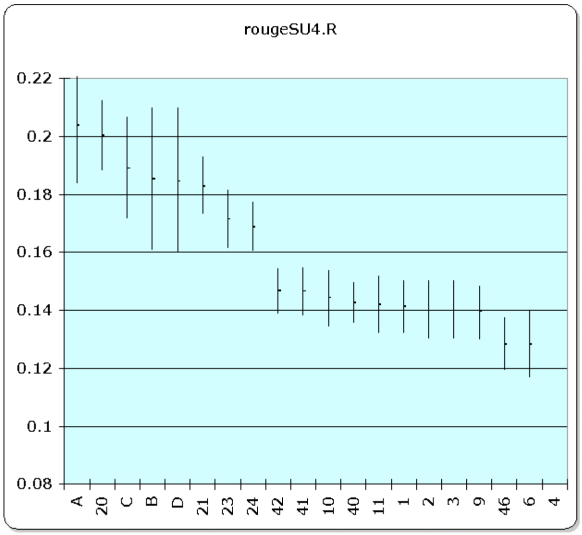

Figure 5: Rouge SU4 Recall graph. My system is number 46.

3.1 Failure Analysis

In order to understand why my system performed poorly, I read through a selection of peer systems' output, comparing them to my system's output and reference judge A's summaries. In general, I was pleased that my system did not include much redundancy in the generated summaries, due to the sentence clustering and preference to not select sentences from the same clusters. Some of the peer summaries exhibited more redundancy than my system, although as a biased judge, perhaps my evaluation is not objective.

I included system 21, and it was clear to me that its summaries were very good, clearly better than my system's summaries. For individual sets, I did feel that my system performed as well or better than other systems in the mid-range of scores. Often though, the summaries from my system exhibited three problems:

1. My system often selected useless "Attribution sentences".
2. My system selected short sentences with no real content.
3. My system sometimes selected sentences that were poorly written (from the machine translated text) when similar well-written English sentences existed in the document set.

By "attribution sentence", I mean a sentence that is used primarily to identify a reporter, but carries no other information. Due to the importance of named entities in my sentence evaluation, my system often selected short attribution sentences that identify the story's reporter. For example, in one set, 44004, about hostages being freed from a terrorist group, my system selected three useless sentences out of five total sentences:

> I'm Lisa Mullins. Reporter John Mc Clain speaking to us from Manila. Reporters John Mc Clain is following this stand off for the BBC. The text: Manila AFP, Reuters Philippine President Joseph Estrada announced yesterday that 12 of 17, holding 77 hostages Philippine Abu Sayyaf Islamic fundamentalism in the island of Jolo in the south of the country is now in the hands of the Philippine armed forces following violent fighting left 12 killed in its ranks. With the recovery of the 12 evangelists, the Abu Sayyaf rebels still hold five hostages _ an American, three Malaysians and a Filipino.

Set	Attribution sentences	Short sentences
44004	3	0
44005	1	0
44006	0	1
44008	1	1
44009	2	1
44011	2	2
44013	0	1
44014	3	0
44015	2	0
44016	1	0
44017	1	0
44020	2	0
44021	1	1
44022	0	0
44023	1	0
44024	1	0

Table 1: Count of attribution and short content-free sentences per set. Sets with neither are not included.

- That's our news summary.
- How densely populated?
- , AFP, Don't despair.
- Yes, that's correct.

Figure 6: Examples of short, non-content bearing sentences selected by the system.

The sentences were selected based on the named entities, "Lisa Mullins", "John Mc Clain", "BBC", and "Manilla", but clearly have no important content for the summarization task. My system would greatly benefit from a simple check of whether a sentence is simply an attribution of some sort, and giving a low weight to such sentences. The remaining two sentences in the summary are reasonable sentences that capture some of the important content.

Of the twenty-four sets, thirteen of them contained a total of twenty-one "attribution sentences", an average of 1.6 per set that contained such sentences. Table 1 lists the number of attribution and short sentences included for each set where these problems were evident.

Another problem is that the summaries tended to contain short sentences with little content. A selection of such sentences is given in Figure 6. This is likely due to the search process, where the summary is not penalized for adding a short sentence if it will fit. While it would be possible to modify the evaluation function to avoid adding short sentences, it would be better to detect short sentences with low information content and automatically have them evaluate to a low score during the hypothesis evaluation. While adding a single short, four-word sentence doesn't really hurt the summary in the ROUGE scoring, had the sentence evaluated to a lower score, perhaps some other combination of longer sentences that are better overall (although not better than the small contribution from the four word sentence in the current implementation) would have been selected.

My system also would sometimes select sentences that had minor grammatical or readability problems that came from the machine translated source text. I'm not sure how much influence this has in lowering the scores, but reading over the summaries was more difficult when compared to other well-written English sentences that conveyed approximately the same information. In future versions of the system I will add an explicit "translated text" feature that gives a slight negative weight to the overall sentence score to penalize the system for selecting such sentences.

4. FUTURE WORK

One of the first things I would like to work on in my system is cross-document named entity disambiguation. I only performed exact string matching across documents to aggregate counts for named entities, which hurt the named entity scoring used to select sentences. The idea is to select important named entities first, where importance is determined by repetition of the named entity, but as named entity references vary greatly even within a document [9] I need to disambiguate names within and across documents. I feel that would more accurately reflect the important named entities, which would help improve sentence selection.

In the far future, I would like to process non-English text internally in the system, continuing with my thesis work on multi-lingual text similarity for summarization [1]. To that end, I also think it is important to think about cross-lingual named entity disambiguation, including transliteration of names between languages [6, 10].

Finally, while it does not seem to be important in ROUGE-based evaluations, I would like to add some logic for sentence ordering. While most of the summaries that my system generated had the key content for the summary, the sentence were often out of order and confusing to read. I will add simple measures such as ordering sentences based on document time stamp, relative order from within the extracted document, and so on.

5. CONCLUSIONS

In this paper I presented the design and implementation of a clustering summarizer using named entities for the 2006 Multilingual Summarization Evaluation. I presented some of the failings of the system, and how I plan to improve it.

Using named entities as a focus seemed to work well for a certain class of summarization topics, such as reporting on political figures. It does not seem to be as suitable for other types of topics, such as reports about natural disasters. In the 2004 Document Understanding Conference, the Columbia University team routed a document set to a different summarization system based on analysis of the type of set [8]. Using a summarization strategy that is tailored to a specific task seems to be a good approach, if there is some way to identify what strategy would work well for a given document set.

In previous work, I have preferred to use English sentences when both English and machine translated sentences are available for inclusion in a summary [2]. I did not take that approach with this system, although I will add that ability in a future version. Unfortunately, I don't think that systems are rewarded in this evaluation for using content from the Arabic text. I think it is becoming more and more important to consider non-English text in information processing tasks in the rapidly globalizing world that we live in, but current evaluation tasks are not set up to require systems to use non-English text. I radical change from the "create a general audience summary of these documents" might be needed to shift the emphasis away from English document processing. New evaluation tasks, such as identifying parties involved in conflicts and their positions and viewpoints, or some other sort of radical alteration to the task would be an interesting direction that could help move the community away from sentence-extraction based summarization systems.

6. REFERENCES

[1] David Kirk Evans. *Identifying Similarity in Text: Multi-Lingual Analysis for Summarization.* PhD thesis, Columbia University, 2005.

[2] David Kirk Evans and Kathleen McKeown. Identifying similarities and differences across english and arabic news. In *International Conference on Intelligence Analysis*, McLean, VA, May 2005.

[3] Elena Filatova and Vasileios Hatzivassiloglou. Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization*, Barcelona, Spain, July 2004.

[4] Elena Filatova and Vasileios Hatzivassiloglou. A formal model for information selection in multi-sentence text extraction. In *Proceedings of COLING*, Geneva, Switzerland, August 2004.

[5] V. Hatzivassiloglou, J. L. Klavans, M. Holcombe, R. Barzilay, M.Y. Kan, and K.R. McKeown. Simfinder: A flexible clustering tool for summarization. In *NAACL'01 Automatic Summarization Workshop*, 2001.

[6] Kevin Knight and Jonathan Graehl. Machine transliteration. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 128–135, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[7] Chin-Yew Lin and E.H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May 2003.

[8] Ani Nenkova, Sasha Blair-Goldensohn, David Kirk Evans, Andrew Hazen Schlaikjer, Advaith Siddarthan, Vasileios Hatzivassiloglou, Kathleen McKeown, and Becky Passonneau. Columbia university at duc 2004. In *Document Understanding Conference (DUC)*, Boston, MA, May 2004.

[9] Ani Nenkova and Kathleen McKeown. References to named entities: a corpus study. In *Short Paper Proceedings of NAACL-HLT*, 2003.

[10] S. Wan and M. Verspoor. Automatic english-chinese name transliteration for development of multilingual resources. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1352–1356, Montreal, Canada, 1998.