The Future of Multilingual Summarization: Beyond Sentence Extraction

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

ABSTRACT

In this paper I present a vision for the future of multilingual summarization that focuses on summarizing differences between documents: generating sentences that explain the main points of controversy in the document set, identifying different sides in the dialogue and the claims they support, and identifying how content differs across document boundaries (cultural, national, political, etc.). I propose integrating summarization with cross lingual information retrieval, and a visual system for interactive document clustering where summaries and extracted information change in response to the users' needs.

1. INTRODUCTION

I have always been enjoyed studying foreign languages. In high school I studied German, as an undergraduate I studied Japanese, and after graduating with a degree in computer science my goal for graduate school was to combine my love of foreign language study and text processing. I was fortunate enough to join the Columbia University Natural Language Processing group, where I was able to work with Professors Judith Klavans and Kathleen McKeown on text summarization. My thesis work was on summarization of Arabic and English text via (multilingual) similarity-based methods [1], and I am continuing this line of research while doing a post-doc stay at the National Institute of Informatics.

There has been a lot of interest in multilingual summarization recently, with a multilingual track in the Document Understanding Conference 2004 [10], workshops on multilingual summarization and question answering co-located with COLING and ACL in 2002 and 2003, and multilingual summarization evaluation workshops in 2005 and 2006. As machine translation systems improve and general usage becomes more common via web-based translation systems, integrating documents from other languages into search, summarization, question answering systems, etc. will become more important. In many cases, documents available in languages other than the user's native language will present new information that is not available to them normally, or with a different perspective that is important to better understand issues in our rapidly globalizing world.

In this position paper I will not focus on the problem of machine translation, which has been steadily improving, and in many cases has already been accepted by the public as useful for certain tasks, e.g., Google or Yahoo's web page translation tools. Instead I will focus on how to take advantage of the content of documents from different countries, cultures, and perspectives. I think summarization systems should focus on higher-level analysis to point out both important facts in a document set, and also interesting differences between documents across various boundaries (cultural, national, political, etc.) Existing sentence-extraction based methods are not sufficient for this task which aims to explain how document sets differ along a variety of axes.

As proposed by Kando [3] language processing for cross language information access can be seen to take place on a variety of layers. These layers proceed from the low level data storage level up to the conceptual layers of intention communication.

1. *pragmatic layer*: cultural & social asects, convention
2. *semantic layer*: concept mapping
3. *lexical layer*: morphology, syntax
4. *symbol layer*: character codes
5. *physical layer*: network, text files

Machine translation and summarization research span some areas of layers 2 and 3 to layer 5, but future research in multilingual summarization and information access will have to focus more on layers 1 and 2. To effectively present information across cultures, we will have to address concepts that do not translate well literally, or do not have parallel counterparts in different cultures.

In Section 2 I propose a move to more specific task-based models for summarization, and aspects that could be important for one information analysis type of task. Section 3 discusses the importance of considering cultural influence on document content for multilingual summarization, and Section 4 presents a vision of summarization as part of an interactive information management experience.

2. TASK-BASED SUMMARIZATION

Recent summarization evaluations have used automated or semi-automated approaches to summarization evaluation. Particularly, many systems have made use of ROUGE [6] to automate system parameter tuning, evaluating the results with ROUGE or possibly with the basic elements package [2]

SIGIR New Directions in Multilingual Information Access Workshop August 2006, Seattle, USA

or some form of pyramid evaluation [8]. While ROUGE and other methods have helped formalize summary evaluation immensely, there are still many difficulties evaluating "general audience" summaries. One of the main factors that makes summary evaluation difficult is that summary content can change greatly depending on the summary creator's interpretation of what is important. For general summary tasks, where the instruction to model summary creators is to write a summary that concisely conveys the important information contained in a set of documents, each summary writer has their own interpretation of what is important, and hence what is in their final summary. While the various summarization evaluation methods take this into account by more heavily weighting information that is shared between multiple model summaries, I think a fundamental change in the purpose of summarization will help create model data that is more consistent across evaluators, leading to more stable data that will reward high-risk, analysis-heavy approaches to summarization that currently are not used because simple techniques can still perform fairly well under the sorts of evaluations that are currently performed.

2.1 Information Analysis Scenario

I see summarization as a step in an information-processing workflow, and not necessarily as a final end-product itself. Summarization can be an important aid to information analysts of all sorts. As an example, imagine an analyst that is interested in investigating the benefits and risks involved in a potential investment in an unfamiliar foreign country. In such a scenario, I see many areas where computer-aided information processing technology are applicable:

1. Cross lingual information retrieval to identify relevant documents.
2. Multilingual summarization to provide an overview of the information in the analyst's native language.
3. Summarization focusing on different aspects of information analysis.
4. Interactive question-answering systems for detailed investigation of specific topics.

Cross-language information retrieval is important to identify documents that present opinions and viewpoints that would not be found in the media coming from one source only. By opening up document search to multiple languages, the analyst is presented with information that could not be found in their native language, allowing access to opinions that might vary wildly from the predominating viewpoint presented in the media available from one country. As newswire syndication becomes more commonplace, the variety of viewpoints that are available from the news outlets of any one county diminishes, and gaining access to media from different countries (with different sets of syndication networks) is much more important to understand the different facets of a given topic.

In step 2 above, I still think that a "general overview" type of summary is useful for an analyst to present the content of the document set, but I also feel that more specific types of summarization are needed to support information analysis tasks. Different summaries can be created to summarize

- factual information
- points of controversy
- opinions and reaction
- culturally-relative information

Focusing on specific tasks, such as identifying key factual information (how many people were killed, who was arrested, etc.), identifying the controversy in a set of documents, summarizing the viewpoints of different sides in the controversy and their supporting information, and identifying information important to one group or another are all higher level tasks that are not suited to sentence extraction. Across languages and culture the information that is presented, and how it is presented, can also vary greatly, and automatic analysis of bias based on the document source, as well as identification of what different sides in a controversy view as important are both useful tasks that can aid information analysts.

2.2 Factual Information Identification

Factual information identification is an important technique for summarization. For analysts that are investing specific areas, for example tracking business mergers and personnel movement, or terrorist attacks, a succinct summary of the relevant facts (what companies are involved in the merger, how much money is involved; what position is being filled and by whom) is clearly of great benefit. Past research has looked at using MUC-style templates as input for summarization [12], and I think the idea of identifying factual information for use in a summary is a good one, but it is difficult to apply MUC-style systems to an open domain. More research needs to be done on the general approach of identifying factual snippets of text (as has been done in the question answering community) and models for determining what types of information are important in a given context. Using a document set to dynamically derive concepts that are important to the domain, and identification of facts based on analysis of the text expressed in the documents would go a long way towards developing systems that can find factual information in an open domain and then use that information as input to drive sentence re-generation around the extracted text snippets.

2.3 Automatic controversy identification

I would like to see summarization grow from taking advantage of similarity and repetition in documents to analyzing the differences between documents. Taking advantage of repetition is important currently because it is assumed to be a strong indicator of importance, which is what most summarization evaluations are currently trying to measure. I would like to see summarization evaluation stress different areas to make pragmatic analysis and semantic processing more important.

One task that I think is useful and achievable in the next few years is automatic identification of controversy.

Identifying sentences that are controversial, or sentences that support one side of a controversy or criticise another is a task that can be approached with common tools employed today for summarization, classification, or categorization. Succinctly describing the point of contention in a set of articles is more difficult though, as it is unlikely that a sentence that clearly describes the controversy will be readily available in the input document set. As the tasks demanded of summarization systems become more complex, sentence extraction will be less effective, and more attention will have to be shifted to natural language generation and alternative forms of summarization such as table and list generation. While open-domain natural language generation has proven to be difficult without complicated deep-semantic representations, generation that targets specific questions such as

"What is the main controversy in this set of documents?" should be achievable in an open domain.

2.4 Opinion identification

Recently there has been much work on identifying opinions and polarity of opinions in text [4, 13]. These sorts of techniques can also play a large role in summarization. Since I am focusing on using summarization to identify controversy in the input document set, a complementary task is to identify the main actors in the controversy, and what viewpoints and opinions they hold. Opinion and polarity identification can be useful as features to help identify sentences that are involved in controversy, but are also very useful to build up supporting and criticizing information for different sides involved in the discussion.

For certain types of information analysts, a "summary" consisting of a list of people involved in the document set, a brief description of who they are, and a list of the opinions and positions that they hold with respect to the controversy discussed in the document set would be a very useful type of summary.

3. MULTILINGUAL MULTIDOCUMENT SUMMARIZATION

While Sections 2 to Sections 2.4 are applicable to monolingual summarization, multilingual summarization is particularly interesting and challenging because of differences in assumed common background that come from a shared cultural heritage. I am particularly interested in how document content, presentation, and perception change across the boundaries of countries and cultures. As an obvious example in the news media, the source of information plays a large role in the bias with which information is presented, and the content that is presented. Contrasting the news stories on American conservative news agency Fox News' website[1] and the Middle-Eastern based Al-jazeera News' website[2] is an interesting exercise in examining the role that bias plays in news selection. Particularly with multilingual data that represents viewpoints from different communities, countries, and cultures, I think taking advantage of the diversity of viewpoints and opinions is a strength that can be used to improve summarization.

At the National Institute of Informatics I am developing a multilingual summarization corpus with documents in English, Japanese, Chinese, and Korean. Based on the relevance judgments from the NTCIR CLIR (cross lingual information retrieval) task [5], I have selected 24 topics that are likely to have controversial contents, and am having separate multi-document summaries created for the documents in each language (Chinese, Japanese, English, and sometimes Korean.) By summarizing each language independently it will be possible to analyze the differences in content between the languages instead of focusing on what information is shared across the documents in different languages. Initial investigation of the data shows that the viewpoints can vary greatly in documents from different languages (although the variation has more to do with document source than document language.) Over the next two years, I plan to research methods for automatically identifying the source of the controversy, and building summaries that identify support for each side of the issue, and criticisms that each sides offers for opposing views.

I am hopeful that customisation of statistical machine translation techniques to the problem, coupled with using carefully selected extracted text as input will result in generation of high quality sentences without relying on manually creating templates for generation.

In the context of multilingual multi-document summarization systems is it possible to build models that can predict the role that cultural, and other bias, plays in content selection and presentation? Along with opinion identification and polarity detection, can we build systems that analyze across documents what content is selected due to a predisposed preference based on the country or culture, and what content is deliberately not selected for the same reason? Automatically pointing out exactly these omissions and deliberate wording choices across sets of documents could be very helpful for dealing with information in this quickly globalizing world.

4. INTERACTIVE MULTILINGUAL SUMMARIZATION

While most summarization systems currently use text as the medium for summary presentation, I believe that more dynamic formats and more attention to information presentation will be important areas in the future. I see multilingual summarization as a tool that will be used in conjunction with CLIR to interactively explore and analyze data.

Continuing with the hypothetical information analyst scenario that I presented earlier, an analyst could search the web using a CLIR system, with the result documents clustered in an interactive visual browser. The results could be clustered according to a variety of criteria chosen by the user. With large result sets keywords and content terms differentiating the clusters can be displayed allowing for query refinement in a visual manner accessible to the novice and expert alike. When the query has been sufficiently refined to return a smaller document set, more sophisticated criteria such as controversy and opinion identification as discussed in Section 2 to label clusters with the main entities and a brief summary of their positions and the controversy discussed in the document cluster. By choosing to re-cluster the documents based on language, the user can see how the points of discussion differ, and what role culture plays in determining the dominant topics.

An interactive visual system for result display also will allow for more compact information display techniques and on-demand summarization. Systems such as Google News[3], NewsInEssence[4] [11], and Columbia NewsBlaster[5] [7] summarize news documents, but view it more as a batch process. In a interactive system, sophisticated techniques such as cross-document co-reference analysis can be used to identify major entities in the document set, re-write references in a compact manner following journalistic conventions [9] to list important named entities in a cluster. If the user would like more information about an entity, they can click on the name, bringing up a more full description of the entity along with other related information, such as viewpoints

[1] http://www.foxnews.com/
[2] http://english.aljazeera.net/
[3] http://news.google.com/
[4] http://www.newsinessence.com/
[5] http://NewsBlaster.cs.columbia.edu/

and opinions that the entity is known to have. Re-clustering documents that involve the entity could result in a new set of clusters, each with a brief summary of the content and controversy which can also be expanded on with further user interaction.

4.1 Natural Language Generation

Sentence extraction is often a successful technique for general informative content summarization when used with news documents due to the journalistic convention of summarizing the main point of the article early in the lead paragraph. As the summarization community branches out into new mediums and moves away from "general" summaries to more task-focused summaries it will become more important to focus on natural language generation. In the interactive clustering summarization scenario I present, informative summaries about each cluster are necessary as well as meta-information about how the clusters differ. As the clusters are generated on demand, it is unlikely that sentences that describe how the content differs between these clusters will exist, and some form of generation will be necessary to present these differences.

Especially with multilingual summarization, where documents are likely to come from different countries and cultures, we would like to be able to operate at a higher layer of processing to identify high-level pragmatic differences between documents.

5. CONCLUSION

I have presented a vision of summarization as part of a larger workflow for information analysts investigating a new area previously unknown to them. Generating summaries that contain both important factual information shared by the documents and information about how documents differ across language and cultural boundaries help analysts become familiar with all aspects of an issue, and not just what is reported in their national media. The types of summarization envisioned here do not lend themselves well to sentence extraction, so more attention must be paid to natural language generation. In addition to the technology needed to build the summarization component of such a workflow, it is also important to concentrate on actual usage scenarios, and to build systems that people can easily use.

Focusing on ease-of-use and exposing summarization systems as part of a larger interactive document search and exploration system will help improve visibility for the technology. As more people use systems that can clearly show how information differs across language and source, they can become more aware of what they are missing by using only one or few sources for their information needs. Integrating culturally-sensitive multilingual summarization with on-demand cross-lingual information retrieval would be a boon for improving global relations.

6. REFERENCES

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

[2] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, 2006.

[3] Noriko Kando. Evaluation – the way ahead: A case of the ntcir. In *Proceedings of 25th ACM-SIGIR Workshop on Cross Language Information Retrieval: A Research Roadmap*, pages 72–77, Tampere, Finland, Aug 2002.

[4] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the COLING conference*, Geneva, Switzerland, 2004.

[5] Kazuaki Kishida, Kuang hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, and Sung Hyon Myaeng. Overview of clir task at the fifth ntcir workshop. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, December 2005. National Institute of Informatics.

[6] 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.

[7] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of HLT 2002 Human Language Technology Conference*, San Diego, CA, 2002.

[8] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: the pyramid method. In *Proceedings of the Human Language Technology / North American chapter of the Association for Computational Linguistics conference*, May 2004.

[9] Ani Nenkova, Advaith Siddharthan, and Kathleen McKeown. Automatically learning cognitive status for multi-document summarization of newswire. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, October 2005.

[10] Paul Over and J. Yen. An introduction to duc 2004 intrinsic evaluation of generic new text summarization systems, 2004. National Institute of Standards and Technology.

[11] Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Demo Presentation, Human Language Technology Conference*, San Diego, CA, March 2001.

[12] Dragomir R. Radev and Kathleen McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, 1998.

[13] Franco Salvetti, Stephen Lewis, and Christoph Reichenbach. Impact of lexical filtering on overall opinion polarity identification. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, California, USA, March 2004.