Similarity-based Multilingual Multi-Document Summarization

David Kirk Evans, Kathleen McKeown
Dept. of Computer Science
Columbia University
{devans, kathy}@cs.columbia.edu

Judith L. Klavans
Center for Advanced Study of Language
University of Maryland
jklavans@umd.edu

Abstract

We present a new approach for summarizing clusters of documents on the same event, some of which are machine translations of foreign-language documents and some of which are English. Our approach to multilingual multi-document summarization uses text similarity to choose sentences from English documents based on the content of the machine translated documents. A manual evaluation shows that 68% of the sentence replacements improve the summary, and the overall summarization approach outperforms first-sentence extraction baselines in automatic ROUGE-based evaluations.

1 Introduction

With the large amount of text available on the web, summarization has become an important tool for managing information overload. While multi-document summarization of English text has become more common, less attention has been paid to producing English summaries of foreign language text. Yet, use of foreign language on the web is growing rapidly (Grefenstette and Nioche, 2000), and with growing globalization many news events are covered by many countries.

Our multilingual multi-document summarizer takes as input a set of multiple documents on a particular topic, some of which are English, and some of which are machine translations of Arabic documents into English. Our task is to produce an English summary of the foreign language documents. This task was added to the 2004 Document Understanding Conference (Over and Yen, 2004) on summarization. One of the problems with extracting sentences from the machine translated text directly is that they can be ungrammatical and difficult to understand. Moreover, removing context makes the resulting summary hard to comprehend. Figure 1 shows an example of an Arabic sentence translated by IBM's statistical MT system, and the English sentence that our system suggests as a replacement.

Machine translated sentence: particular seven organizations Egyptian Organization for human rights today , Monday appealed to the Egyptian President Hosni Mubarak a cost-accounting the responsible for acts of torture which aimed villagers in upper Egypt during the investigation in the crimes killed in last August .
Similar English sentence: Seven Egyptian human rights groups appealed Sunday to President Hosni Mubarak to ensure that police officers they accuse of torturing hundreds of Christian Copts be brought to justice.

Figure 1: A system suggested replacement sentence for a machine translated Arabic sentence

In this paper, we introduce a new method to summarize machine translated documents using text similarity to related English documents. The summary is built by identifying the sentences to extract from the translated text, and replacing the machine translated sentences from the summary with similar sentences from the related English text when a good replacement can be found. The idea is to match content in the non-English documents with content in the English documents, improving the grammaticality and comprehensibility of the text by using similar English sentences.

We present different models for summarization using replacement and show their effectiveness in improving summarization quality. In addition to different metrics and thresholds for similarity, we investigate the utility of syntactic sentence simplification on the replacement English text, and sentence chunking on the machine translated Arabic text. We performed a manual evaluation of whether replacements of machine translated sentences by similar English sentences improve a summary on a sentence-by-sentence basis, as well as an evaluation of a similarity-based summarization system using the automatic ROUGE (Lin and Hovy, 2003) summary evaluation metric. We show that 68% of sentence replacements improve the resulting summary, and that our similarity-based system outperforms a state-of-the-art multi-document summarization system and first-sentence extraction baseline.

1.1 Related Research

Previous work in multilingual document summarization, such as the SUMMARIST system (Hovy and Lin, 1999) extracts sentences from documents in a variety of languages, and translates the resulting summary. Chen and Lin (Chen and Lin, 2000) describe a system that combines multiple monolingual news clustering components, a multilingual clustering component, and a summarization component. Their system clusters news from Chinese and English into topics, then the multilingual clustering component relates the clusters that are similar across languages. A summary is generated for each language based on scores from counts of terms from both languages. Our system differs by explicitly generating a summary in English using selection criteria from the non-English text.

Other work uses similarity-based approaches to summarization to guide selection (Radev et al., 2000), or to guide generation from similar content (Barzilay et al., 1999). Our work is original in using text from one language to guide selection exclusively on English text, thus improving comprehensibility of the summary.

2 Summarization Approach

Our approach relies on first translating the input documents (Arabic, in this work) into English and then using similarity at the sentence level to identify similar sentences from the English documents. As long as the documents are on the same topic, this similarity-based approach to multilingual summarization is applicable. This paper does not address the issue of obtaining on-topic document clusters in this paper; news clustering systems such as Google News[1], Columbia NewsBlaster[2], or News In Essence[3] demonstrate that this is feasible. The system architecture is:

1. Syntactically simplify sentences from related English documents, and possibly chunk machine translated Arabic sentences.
2. Produce a summary of the machine translated sentences using an existing sentence extraction summarization system.
3. Compute similarity between the summary sentences and sentences from similar English documents.
4. Replace Arabic sentences from summary with English sentences for those pairs with similarity over an empirically determined threshold.

Since the focus of this work is not extraction-based summarization, we used an existing state-of-the-art multi-document summarization system, DEMS (Schiffman et al., 2002), to select the sentences for the similarity computation process.

2.1 Sentence Simplification

Since it is difficult to find sentences in the related English documents containing exactly the same information as the translated sentences, we hypothesize that it may be more effective to perform similarity computation at a clause or phrase level. We ran the English text through sentence simplification software (Siddharthan, 2002) to reduce the English sentence length and complexity in the hope that each simplified sentence would express a single concept. The sentence simplification

[1] http://news.google.com/
[2] http://newsblaster.cs.columbia.edu/
[3] http://NewsInEssence.com/

software breaks a long sentence into two separate sentences by removing embedded relative clauses from a sentence, and making a new sentence of the removed embedded relative clause. This would allow a more fine-grained matching between the Arabic and English sentences, without including additional information from long, complex sentences that is not expressed in the Arabic sentence.

For example, for the following Arabic sentence,

> 1. had decided Iraq last Saturday halt to deal with the United Nations Special Commission responsible disarmament Iraqi weapons of mass destruction.

one similar English sentence found is:

> 2. Earlier, in Oman, Sultan Qaboos reportedly told Cohen that he opposed any unilateral U.S. strike against Iraq, which ended its cooperation with U.N. inspectors on Saturday.

That sentence simplifies to the following two sentences:

> 2a. Earlier, in Oman, Sultan Qaboos reportedly told Cohen that he opposed any unilateral U.S. strike against Iraq.
>
> 2b. Iraq ended its cooperation with U.N. inspectors on Saturday.

Using sentence simplification to break down the text allows us to match sentence 2b, without including 2a, which was not reported in the Arabic sentence.

We examined using two types of sentence simplification, *syntactic* and *syntactic with pronoun resolution*, and compared them to not using any sort of simplification. To limit the number of systems evaluated in the manual evaluation, we determined settings to use based on results from automated summary evaluation. In all of our experiments, syntactic simplification performed about 3% better on ROUGE scores than simplification with pronoun resolution, or not performing any simplification. Simplification with pronoun resolution did not always beat unsimplified text, possibly due to errors introduced by the pronoun resolution, which has a success rate of approximately 70%. We present results of the system using only syntactic simplification.

Similarly, we performed experiments for splitting the machine translated Arabic text. We investigated two methods for splitting Arabic text: one tags the text with TTT[4] and splits on verb groups, copying the previous noun group and verb group to the start of the next sentence. The other splits on verb groups and "and", "nor", "but", "yet" and ",", without performing the copying. In both cases, sentences with less than 3 tokens are filtered from the output. The copying method was approximately 3% better on the manual evaluation below, so we omit results from the other chunking method.

2.2 Similarity Computation

Text similarity between the translated and relevant text is calculated using Simfinder (Hatzivassiloglou et al., 2001). Simfinder is a tool for clustering text based on similarity computed over a variety of lexical and syntactic features. The features used in Simfinder are the overlap of word stems, nouns, adjectives, verbs, WordNet (Miller et al., 1990) classes, noun phrase heads, and proper nouns. Each feature is computed as the number of items in common between the two sentences normalized by the sentence length. The final similarity value is assigned via a log-linear regression model that combines each of the features using values learned from a corpus of news text manually labeled for similarity. No modifications were made to Simfinder to compensate for using machine translated text as input, although the machine translated text is quite different from the news text used to train Simfinder.

2.3 System Implementation

Our summarization system can be run in multiple configurations.

1. Use DEMS to select Arabic sentences, retain only sentences that have similar English sentences, replacing them with the single most similar English sentence. If the summary is too short (less than 600 bytes,) delete it, and build a new summary using all Arabic sentences, sorted by similarity to English sen-

[4] http://www.ltg.ed.ac.uk/software/pos/

tences, and replacing each one by the single most similar English sentence.
2. Use DEMS to select Arabic sentences, replace only sentences above empirically determined threshold of 0.6 passing a cosine filter with similar English sentences, retain non-replaced Arabic sentences in the summary.
3. Use all Arabic sentences, sort by decreasing similarity to English sentences, and replace each one by all English sentences above an empirically determined threshold of 0.6 that pass a cosine filter. Machine translated sentences are kept if they do not pass the threshold.

Configuration 1 uses DEMS to select sentences, and maximizes the number of replacements made by re-running without DEMS if not enough similar sentences are found to make a large enough summary. Configuration 2 also uses DEMS to select sentences, but retains any machine translated sentences for which no suitable sentence replacements are found. Configuration 3 focuses on maximizing replacements by not using DEMS for selection, and builds a summary by taking the most similar English sentences, using only similarity to Arabic sentences to guide selection, removing any manually-constructed "intelligent" system from the selection task. All summaries are limited to 665 bytes since that was the size threshold that was used for the DUC evaluation. An evaluation of the different configurations of the system using ROUGE scores is presented in Section 3.3.

3 Evaluation

We performed evaluation at two levels: the sentence level to test the proposed sentence replacements of Arabic sentences from similar English sentences, and the summary level to evaluate quality of the full summaries that include these sentence replacements. At the summary level, we used the automated system, ROUGE, for evaluation. It allowed us to make rough distinctions between different models for constructing the full summary. However, this would not tell us whether a particular English sentence was a good replacement for a translated one and thus, we used a more time-consuming, manual evaluation to quantify how well replacement worked.

3.1 Evaluation data

We use the 2004 DUC corpus for both the sentence and summary level evaluations. The corpus contains 24 topics with relevant documents, some in English and some in Arabic, and machine translations of the Arabic documents into English from 2 different systems. As part of the corpus, each of the sets contains summaries by 4 human assessors who read manual translations of the Arabic documents. These 4 summaries are used as the reference models against which the automatic summaries are evaluated – note that the model summaries were created only with knowledge of the content from the Arabic documents, and not the English documents. In Section 3.3.1 we evaluate our similarity-based summarization system using ROUGE.

3.2 Sentence level evaluation

By replacing machine translated sentences with English sentences we run the risk of introducing false information that is not a good representation of the content of the Arabic sentences. We performed two evaluations that examined whether the sentence being replaced improved upon, or detracted from the overall meaning and understandability of the machine translated sentence being replaced.

The first evaluation examined IBM machine translated Arabic text sentences replaced with sentences from syntactically simplified related English sentences. Two systems were used to compute similarity for the sentence replacements, Simfinder, as described in Section 2.2, and a simple cosine-based similarity metric. For the 24 document sets, for each sentence in the summary, the top three most similar sentence replacements are evaluated by humans on a 5 point scale with reference to the understandability of the sentences, and the content with respect to the final summary:

1. improves the summary without changing the meaning
2. improves the summary but changes the meaning
3. is no better or worse than before
4. degrades the summary without changing the meaning

Arabic, Similarity type	% Good	# Sents
Full, Simfinder	59%	227
Chunked, Simfinder	56%	294
Chunked, Simfinder with filter	62%	250
Full, cosine chunked similarity	71%	21
Full, Simfinder chunked similarity	68%	151

Table 1: Percentage of good sentence replacements for sentence-by-sentence evaluation at 0.7 similarity threshold.

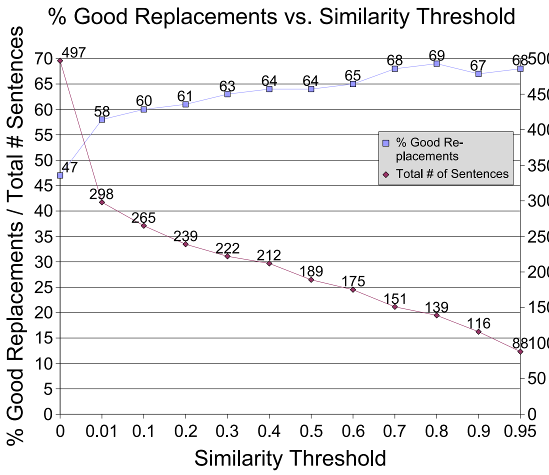

Figure 2: Threshold-Precision curve for MT Arabic to simplified English replacements using chunked Arabic similarity values.

5. degrades the summary and changes the meaning

Six humans performed the evaluation, with each sentence pair being marked by two evaluators. Average agreement between evaluators was 70% on whether replacement improved or degraded the summary, with a Kappa of 0.41.

The second sentence evaluation examined chunked IBM machine translated Arabic text and syntactically simplified related English sentences. The machine translated Arabic sentences were split using the TTT tagging software, splitting on verb groups and copying the previous noun and verb group. In addition to the 1-5 scale above, each sentence pair was labeled as to whether the replacement sentence was "related" or "not related" to the machine translated sentence, where related is an indicator that the sentences are on the same topic. We later use this feature to learn filters to improve performance.

3.2.1 Sentence level evaluation results

We computed evaluation results by examining, for a given similarity threshold, how many proposed sentence replacements have a similarity higher than the threshold. Those sentences that are above the threshold and marked by evaluators with category 1 or 2 are marked as "Good" sentence replacements, while those from category 3, 4, or 5 are poor replacements. Table 1 shows the results of the two evaluations at a similarity threshold of 0.7, chosen as a good trade-off between quality and number of sentences over the threshold. Since the cosine similarity metric resulted in very few replaced sentences in the full MT Arabic evaluation, only the Simfinder scores were evaluated using chunked Arabic. The cosine metric performed uniformly poorly for similarity thresholds below 0.5, at about 32% good replacements, and then improved to 70%–100% from 0.7–1.0, but only very few sentences were found at these levels. The cosine metric had fewer than 5% of the number of sentences that Simfinder found from thresholds of 0.7 and above.

In the first evaluation, replacing full machine translated Arabic text with syntactically simplified related English text based on similarities computed by Simfinder, about 59% of the replacements were judged to improve the summary at a similarity threshold of 0.7. One can improve that percentage by increasing the threshold, but that reduces the total number of sentences that are replaced (see Figure 2.) With more than half of the sentence replacements helping the summary, we were interested in the effect splitting the Arabic sentences might have; by focusing on smaller parts of the Arabic sentences, we hypothesized that we could find better matches to just that part of the sentence in the related English text.

The initial results using chunked Arabic text using Simfinder similarity, were not as good as using full Arabic sentences: 56% at a 0.7 similarity threshold. With shorter sentences, there was a problem of lack of context, as 26% of the sentences over the 0.7 similarity threshold were labeled as "not related" by evaluators. Looking at the results for only sentences labeled "related", results were much better: 70% good replacements at the same threshold.

To improve similarity computation with these short sentences, we investigated predicting the "related" feature given the sentences. We used a machine learning approach, computing a variety of features between the two sentences, and then using the WEKA machine learning framework to induce learners for the "related" class. We computed 22 features over the two sentences, including cosine similarity, jaccard similarity, length differential features, tf*idf differential features, longest common substring features, overlap on verbs, non-communicative verb overlap, and overlap on proper nouns. The best resulting classifier used only the cosine feature, and had about 76% accuracy at predicting the "related" class. Integrating the "related" class prediction, sentences are only replaced if they are predicted to belong to the "related" class, and have a similarity above the similarity threshold. This improved results using chunked Arabic text to 62% at a 0.7 similarity threshold, as shown in Table 1.

The approach of chunking the Arabic text into smaller units had disappointing results; some of the Arabic chunks were so small that computing meaningful similarity separately was difficult. However, if we could combine the comparison of chunks within the context of the full sentence, perhaps this would give an improvement. Breaking the Arabic sentence into chunks better differentiates the similarity values; with longer sentences, due to the greater number of words similarity values tend to increase, but by looking at smaller chunks, the incidental similarity decreases. An Arabic sentence that is not really very similar to an English sentence might have two chunks that taken together by chance have enough words in common to appear similar, but when broken into chunks do not have high individual similarity values.

The final system shown in Table 1 replaces full machine translated Arabic sentences with syntactically simplified English sentences, but uses similarity values for the Arabic sentences from the chunks it contains. When evaluating replacement of a full Arabic sentence by an English sentence, we retrieved the similarity values for each chunk in the Arabic sentence to the English sentence, yielding a set of similarity values for each Arabic sentence. We used the maximum similarity value of all the possible chunks, checked whether the similarity value was above the threshold, and performed replacement if so. This model avoided the problems caused by the very short Arabic chunks by choosing the maximum similarity score from all chunks; the very short chunks are therefore ignored. At the same time, it avoided some of the problems with comparing overly long sentences where false matches are suggested for replacement simply due to the larger quantity of words. The full threshold–precision chart for this run is shown in Figure 2. Of all the approaches, this one performs the best; with approximately 68% of the replacements being judged as improving the summary.

Using similarity values computed as a function of the similarity of sub-sections of the Arabic sentence allows comparisons of sentences by the propositions they contain and thus, shows an improvement over using similarity values from the entire sentence. While taking the maximum of the chunk similarity values performed best, taking the minimum or average also performed better than using similarity values from the full sentence.

3.3 Summary level evaluation

We evaluated our similarity-based summarization system using ROUGE,[5] a system for summary evaluation that compares system output to multiple reference summaries. We include results from two baseline systems: a first-sentence system, and runs of the DEMS system without replacement.

The first-sentence summarization baseline takes the first-sentence from each document in the set until the maximum of 665 bytes is reached. If the first-sentence was already included from each document in the set, the second sentence from each document is included in the summary, and so on. Two baseline summaries were generated; one for the relevant English documents only, and one for

[5] version 1.2.1, -b 665

Similarity System	ROUGE-L
System Config 1	0.25441
System Config 2	0.19348
System Config 3	0.21936
1st Sentence Baseline	
Related English	0.23973
IBM translations	0.22118
DEMS Baseline	
Related English	0.16197
IBM translation	0.21966

Table 2: Summary evaluation results.

the IBM translated documents alone. The IBM translation baselines give us an idea of scores for summaries drawn from the same content as the reference summaries, while the relevant English baselines tell us how well summaries generated without any knowledge from the Arabic text score. Our similarity-based system was run with simplified English sentences and full machine translated Arabic sentences.

3.3.1 Summary level evaluation results

Table 2 lists the results using the ROUGE-L evaluation metric along with the results of the four baseline runs. The ROUGE-L score is a longest common substring score from the ROUGE system, which rates summaries based on n-gram overlap between the system summary and multiple reference summaries. Evaluations with ROUGE in the past have demonstrated that the score often fails to show statistical significance between scores for evaluated systems. In DUC04 on the multilingual system task, the 95% confidence interval split the 11 participating systems into two main groups; the bottom group containing three systems and the top group containing everyone else. One could argue for a third group containing the top system only, which was statistically significantly better than the bottom six systems when taking the 95% confidence interval into effect. It is not a surprise, then, that the results for the three versions of our system and the baselines also fall within the 95% confidence interval. As the only automated method for summarization, ROUGE is often, nonetheless, used to roughly rank different approaches. Even if the similarity-based systems do not beat the baselines by statistically significant margins, replacing the machine translated text with English text does improve the readability of the summary.

The similarity-based summarization system in configuration 1 performs better than all the baselines, whether over the related English text, or the IBM machine translated text. By out-performing the first sentence baseline and DEMS on the machine translated text, we infer that the similarity system is able to choose sentences from the related English text that are relevant to the content summarized by the humans who read the manual translations of the Arabic text. In contrast, simply running first sentence extraction and DEMS on the related English text does not perform as well; using the machine translated Arabic text to guide selection of related English sentences gives an improvement in performance over the related English baselines. The similarity-based system even outperforms DEMS when run over the manual translations.

Of the three system configurations, the first performs the best. In this evaluation, this configuration builds a summary using all Arabic sentences and replaces them with the most similar English sentence because DEMS selection resulted in too few sentences. Using DEMS for selection in configuration 2 resulted in summaries containing mostly machine translated text, since few sentences pass the required threshold level and filters, but did not perform as well as the DEMS baseline since sentences were sorted by similarity, resulting in different sentences in the truncated summary. Configuration 3 also contained some machine translated sentences, and did not perform as well as configuration 1, which only contained English text.

4 Conclusions

In this paper, we presented a summarization system that summarizes machine translated Arabic text using the Arabic sentences to guide selection of English sentences from a set of related articles. Syntactic sentence simplification on the related English text improves overall summarizer performance, and a hand evaluation of the sentence replacements show that 68% of the replacements improve the summary.

The results from the ROUGE metric show that the similarity-based summarization approach outperforms DEMS and the first-sentence extraction baseline. It is interesting that a state-of-the-art summarization system run over the relevant English articles performs worse than the similarity-based summarization systems run over the same data. This clearly demonstrates that the similarity-based selection system driven by the machine translations is able to select the good sentences from the relevant text.

5 Future Work

In the process of performing our manual evaluation, often there was different content in the Arabic and English texts, and finding similar content for some subset of the sentences was just not possible. In our ongoing work, we are expanding on the idea of summarizing two different sets of documents by looking at not just what is similar between them, but also what is different. Instead of just using the similarity values as we have done here, we cluster the sentences, and identify sentence clusters that contain information exclusive to the Arabic documents, information exclusive to the English documents, and information that is similar between the two. The clusters with similar sentences can be summarized using the approach in this paper. For the other clusters, we are working on an approach to generate indicative summaries that point out the differences. Given that summaries that point out both similarities and differences are quite different from the model summaries currently used in DUC, future work will also need to develop strategies to evaluate these summaries. We are also developing a multilingual text similarity computation system which takes Arabic and English text as input, only performing machine translation after the sentence clusters have been formed.

References

Regina Barzilay, Kathy McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Association of Computational Linguistics*, Maryland, June.

Hsin-Hsi Chen and Chuan-Jie Lin. 2000. A multilingual news summarizer. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 159–165.

G. Grefenstette and J. Nioche. 2000. Estimation of english and non-english language use on the WWW. In *Proceedings of RIAO'2000, Content-Based Multimedia Information Access*, pages 237–246, Paris, 12–14.

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

E.H. Hovy and Chin-Yew Lin. 1999. Automated text summarization in summarist. In I. Mani and M. Maybury, editors, *Advances in Automated Text Summarization*, chapter 8. MIT Press.

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

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, 4(3):235–244.

Paul Over and J. Yen. 2004. An introduction to duc 2004 intrinsic evaluation of generic news text summarization systems. National Institute of Standards and Technology.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Proceedings of ANLP/NAACL 2000 Workshop*, pages 21–29, April.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*, March.

Advaith Siddharthan. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL'02)*, pages 60–65, Philadelphia, USA.