Arabic/English Multi-document Summarization with CLASSY—The Past and the Future

Judith D. Schlesinger[1], Dianne P. O'Leary[2], and John M. Conroy[1]

[1] IDA/Center for Computing Sciences, Bowie MD 20715, USA
{judith,conroy}@super.org
[2] University of Maryland, CS Dept. and Inst. for Advanced Computer Studies, College Park MD 20742, USA
oleary@cs.umd.edu

Abstract. Automatic document summarization has become increasingly important due to the quantity of written material generated worldwide. Generating good quality summaries enables users to cope with larger amounts of information.

English-document summarization is a difficult task. Yet it is not sufficient. Environmental, economic, and other global issues make it imperative for English speakers to understand how other countries and cultures perceive and react to important events.

CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) is an automatic, extract-generating, summarization system that uses linguistic trimming and statistical methods to generate generic or topic(/query)-driven summaries for single documents or clusters of documents. CLASSY has performed well in the Document Understanding Conference (DUC) evaluations and the Multi-lingual (Arabic/English) Summarization Evaluations (MSE).

We present a description of CLASSY. We follow this with experiments and results from the MSE evaluations and conclude with a discussion of on-going work to improve the quality of the summaries–both English-only and multi-lingual–that CLASSY generates.

1 Introduction

Automatic multi-document summarization poses interesting challenges to the Natural Language Processing (NLP) community. In addition to addressing single document summarization issues such as determining the relevant information, pronoun resolution, and coherency of the generated summary, multi-document summary-generating systems must be capable of drawing the "best" information from a set of documents.

Automatic single document text summarization [11] has long been a field of interest, beginning in the 1950s, with a recent renaissance of activity beginning in the 1990s. System generated single document summaries for English are generally of good quality. Therefore, NIST ended single document summarization evaluation after the 2002 Document Understanding Conference (DUC). See [17] for DUC research papers and results over the years.

A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 568–581, 2008.

In contrast to the single document task, summarization of multiple documents written in English remains an ongoing research effort. A wide range of strategies to analyze documents in a collection and then synthesize/condense information to produce a multi-document summary have been explored by various research groups. System performance has improved but still lags behind human performance.

Nevertheless, environmental, economic, and other global issues make it imperative for English speakers to understand how other countries and cultures perceive and react to important events. Thus it is vital that English speakers be able to access documents in a variety of languages.

The quantity of non-English documents makes it impossible to expect quality (or, even, *any*) human translation. Therefore, we have come to rely on machine translation (MT) systems for translation to English. While MT systems continue to improve, generated translations remain difficult to read and understand, with critical words often omitted, and inconsistent translations for the same word in a document [5,6]. Translation of Arabic documents is particularly challenging due to errors introduced by incorrect sentence-splitting, tokenization, and lemmatization.

Volumes of documents in one or more languages may be summarized by:

- creating summaries in the original language(s) which can then be translated by either humans or MT systems to determine "importance".
- creating summaries of the (MT-translated) documents which can be used to determine which documents are important and should be translated by humans.

CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) is an automatic summarization system, developed for summarizing English documents. CLASSY uses trimming rules to shorten sentences in the document, identifies sentences as being more or less likely to be included in a summary, generates a summary for each document, selects sentences for a multi-document summary for a cluster of related documents, and finally organizes the selected sentences for the final summary.

Our approach to multi-lingual summarization is based on the second approach listed above: we use CLASSY to generate single or multi-document (cluster) summaries of MT-translated documents. The experiments presented in Sect. 4. helped determine the best way to accomplish this.

We participated in the two Multilingual Summarization Evaluations (MSE)[1], which evaluated summaries of document sets containing a mix of both English and Arabic documents. Both the Arabic source and the MT output were

[1] MSE 2005: *Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization Workshop* at the Annual Meeting of the Association of Computational Linguistics (ACL 2005), Ann Arbor Michigan, 25-30 June 2005. MSE 2006: *Multilingual Summarization Evaluation* at the 21st International Conference on Computational Linguistics (ACL 2006)/44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17-21 July 2006.

available, and either or both could be used. This paper describes our use of CLASSY and its success in these evaluations.

The remainder of the paper is organized as follows. Section 2 contains a brief description of a sampling of other multi-lingual, multi-document summarization systems. Section 3 presents a description of CLASSY. Section 4 describes the experiments we ran and their success or failure. We then conclude with a discussion of current and future efforts to improve the generated summaries.

2 Related Work

There are many systems which summarize multi-lingual sets of documents, including languages such as Arabic, Chinese, Japanese, and Korean. We briefly describe four of these systems to indicate the breadth of work in this area.

Lakhas [5] is a summarization system that generates very short (headline) summaries. In contrast to many systems, Lakhas first summarizes the original Arabic documents and then applies MT to the summary only. While this eliminates the problems created by poor translations, it introduces its own myriad of difficulties related to Arabic sentence splitting, tokenization, and lemmatization. The scoring function is based on sentence position in the document, number of subject terms (i.e., words that appear in the headline) in the sentence, number of "indicative words" in the document (see the discussion of "signature terms" in Sect. 3.3), and the tf.idf value of each word in the sentence. This approach was very successful for very short (headline) summary generation (Task 3) in DUC 2004.

MEAD [15] is a platform for multi-document multi-lingual text summarization. It consists of multiple summarization algorithms including baselines (e.g., lead sentence) and both centroid-based and query-based methods. The MEAD architecture has four main components. First, each document is converted to an XML-based format. Then feature extraction is performed on each sentence of each document in a cluster, where the features are dependent on the selected summarization algorithm. Third, a composite score is calculated for each sentence. Finally, sentence scores may be refined based on considerations such as sentence repetition, sentence ordering, etc. MEAD currently supports Chinese and English summarization and can be extended to handle other languages.

The system described in [6] took an interesting approach. The DEMS summarizer [16] was first used to summarize a group of English and MT Arabic documents. DEMS produces summaries by extracting high-ranked sentences, where ranking is based on a set of features, some of which attempt to measure inherent importance of the thought. Text similarity measures [8] are then used to replace sentences chosen from the MT documents, which are generally ungrammatical and difficult to understand, with similar sentences from the English documents. This system performed quite well in DUC 2004, Task 3.

A multi-document, multi-lingual, theme-based summarization system based on modeling text cohesion (story flow) is presented in [7]. Some inherent text cohesion is specific to a particular story while some is specific to a particular

language, and these differ across stories and across languages. To exploit the story flow, an unsupervised modified K-means method was used to iteratively cluster multiple documents into different topics (stories) and learn the parameters of parallel Hidden Markov Story Models (HMSM), one for each story. Story models were compared within and across stories and within and across languages (English and Chinese). Experimental results support "one story, one flow" and "one language, one flow" hypotheses.

Twenty-five teams participated in MSE 2005 while only eight did in MSE 2006. These teams were from both industry and academia, from various parts of the world. For example, the 8 teams from 2006 came from China, England, India, Japan, Tunisia, and the US. The 2005 teams were similarly distributed. A conflict with other conferences seems to be the major cause in the drop in participation. While this was unfortunate, the small number of participants did enable a comprehensive human evaluation. Reports about 4 of the 2006 systems (including CLASSY) are available on-line at http://research.microsoft.com/~lucyv/MSE2006_reports.htm. Reports from 2005 are no longer available.

3 Description of CLASSY

CLASSY architecture consists of five steps: document preparation, sentence trimming, sentence scoring, redundancy reduction, and sentence ordering, discussed in the following sections. These discussions are limited to English except where Arabic is explicitly mentioned.

3.1 Document Preparation

Every document is transformed to the CLASSY internal format by performing sentence splitting and sentence typing.

We currently use a Java-based sentence splitter, developed in-house and updated as needed. In addition, a post-processing phase that executes during tokenization (part of the sentence trimming task discussed in Sect. 3.2), corrects many of the sentence splitter's errors which result in either a single sentence erroneously being split into two or two sentences being run together. The main sources of sentence splitter errors are:

- foreign words, especially names that appear to be abbreviations of English words;
- less commonly used abbreviations not known to the sentence splitter;
- sentence termination punctuation embedded in parentheses or quotations;
- missing or bad punctuation; and
- ellipsis at sentence end.

Our sentence splitting is highly accurate, and the few errors that remain would require full parsing (which we do not perform) to detect.

After the initial sentence splitting step, all sentences are typed according to their potential usefulness in a summary. Sentences in headlines and other "title"

roles are given a type of 0; this indicates that they may be useful for determining "signature terms" (see Sect. 3.3) but should not be selected for the summary. Sentences in the textual portion of a document are given a type of 1, indicating that they may be selected for a summary. All other text is given a type of -1: do not use. The sentence trimming algorithms (see below) may modify a sentence type from 1 to either type 0 or type -1, based on sentence length or content, i.e., boilerplate.

3.2 Trimming Sentences

Our trimming code has been written so that it does not require any part-of-speech (POS) tagging or parsing in order to perform its task. This decision was made based on the computational demands of both POS-taggers and parsers as well as the fact that, as good as they have become, both tasks still introduce errors. Instead, we make extensive use of word lists, along with the position of commas, periods, or the sentence start and end, to identify most of the phrases or clauses we remove.

Our sentence trimming approach has been documented in [3,1]. We continue to improve the algorithms to minimize the errors that are made, since these errors result in ungrammatical or, worse, erroneous sentences. We have also been able to identify a larger set of phrases and clauses to eliminate. The sentence trimming we perform is:

1. Removal of extraneous words that appear in a sentence, including date lines, editor's comments, etc.
2. Removal at the start of a sentence of many adverbs, all conjunctions, and about 2000 phrases such as "As a matter of fact," and "At this point."
3. Removal of a small selection of words that occur mid-sentence, such as "however" and "also".
4. Removal of age references such as ", 51," or ", aged 24,".
5. Removal of gerund phrases (phrases starting with the -ing form of a verb) from the start, middle, or end of a sentence.
6. Removal of relative clause attributives (clauses beginning with "who(m)", "which", "when", and "where") wherever possible.
7. Removal of attributions, such as "police said", at the start or end of sentences.

Additional trims, including removing many parenthesized or dashed (–) "asides", remain to be added. Figure 1 shows an example of each of the last three trims in the above list.

3.3 Scoring Sentences

We give a brief overview of an approximate Oracle score, which estimates the fraction of *human abstract terms* a sentence contains. Details of this approach and its motivation can be found in [4,2].

More than 800 lives were lost when the 21,794 tonne ferry**, sailing from the Estonian capital Tallinn to Stockholm,** sank within minutes early yesterday morning in the Baltic Sea 40 km south west of the Finnish island of Uto.

a. Example of a gerund phrase to be removed.

The Menendez family lived in the Princeton Area until 1986**, when they moved to California**.

b. Example of a restricted relative-clause appositive to be removed.

The federal Government's highway safety watchdog said Wednesday that the Ford Bronco II appears to be involved in more fatal roll-over accidents than other vehicles in its class and that it will seek to determine if the vehicle itself contributes to the accidents.

c. Example of an attribution to be removed.

Fig. 1. Examples of phrase/clause eliminations

Instead of using term frequencies of the corpus, as done by [12], to infer highly likely terms in human summaries, we directly model the *set* of terms (vocabulary) that is likely to occur in a sample of human summaries.

We model human variation in summary generation with a unigram language model. In particular, let $P(t|\tau)$ be the probability that a human will select term t in a summary given a topic τ. We define the *oracle score* for a sentence x to be

$$\omega(x) = \frac{1}{|x|} \sum_{t \in T} x(t) P(t|\tau)$$

where $|x|$ is the number of distinct terms that sentence x contains, T is the universal set of all terms used in the topic τ and $x(t) = 1$ if the sentence x contains the term t and 0 otherwise. This score depends on knowledge of human abstracts. Since this information is not available, we substitute a computable *approximate* oracle score [2].

In the absence of human abstracts, we view the *signature terms* as "samples" from idealized human summaries. A signature term is a term which occurs significantly more than expected in the document [9,2]). We use the Porter stemmer [14], which greatly improves the correlation of signature terms with human abstract terms. We define the signature term approximation to the oracle score for a sentence's expected number of human abstract terms as

$$\omega_s(x) = \frac{1}{|x|} \sum_{t \in T} x(t) P_s(t|\tau)$$

where $|x|$, T, and $x(t)$ are defined above, and $P_s(t|\tau) = 1$ if t is a signature term and 0 otherwise (a characteristic function).

The score is built upon an estimate of the probability that a term t will be included in a human summary given a topic τ. This probability is denoted $P(t|\tau)$.

It is approximated using the signature terms and the distribution of the terms in the relevant document cluster.

We estimate our target probability by a mixture of two distributions: the characteristic for the signature terms and the probability that a term occurs in the sentences to be considered for extraction:

$$P_{qs\rho}(t|\tau) = \frac{1}{2}s_t(\tau) + \frac{1}{2}\rho_t(\tau)$$

where $s_t(\tau)$=1 if t is a signature term for topic τ and 0 otherwise, and $\rho_t(\tau)$ is the maximum likelihood estimate of the probability that term t occurs in a sentence in the topic τ. Note that the mixture weights are balanced: both are set to 1/2. We found no statistical improvement in the performance of the approximate oracle score when other weights were used.

The correlation between the oracle score and the approximate oracle score is very strong. Figure 2 is a histogram of the Pearson correlation coefficients for 25 multi-lingual clusters from the MSE data sets.

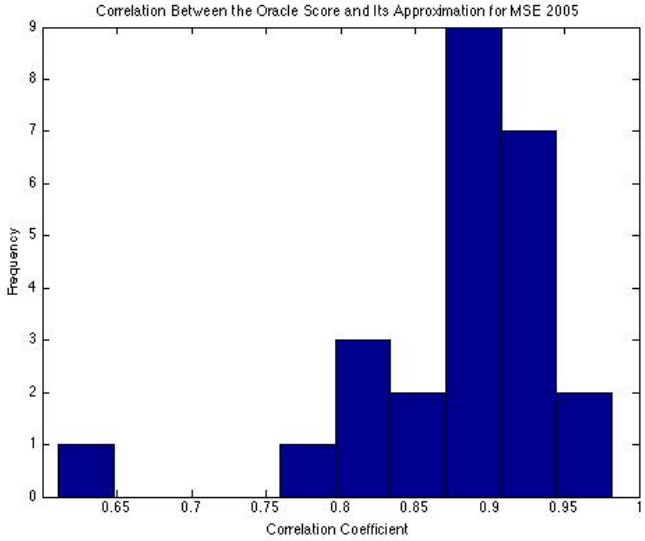

Fig. 2. Pearson correlation coefficients for 25 MSE multi-lingual clusters

3.4 Reducing Redundancy of the Selected Sentences

To reduce redundancy in the sentences chosen for inclusion in the summary, we have a three-step process.

1. Sentences are ordered by score, and enough sentences are chosen to produce a summary 9 times as long as desired. The length was chosen empirically, based on training on MSE 2005 data.

2. The approximate oracle score is simply the sum of the elements in the corresponding column of the (signature) term-sentence matrix A. To improve this score, we replace A by the rank-k matrix $\tilde{A}$ computed using the singular value decomposition. We choose $k = \max(1, \lfloor 0.65\, n \rfloor)$ where n is the number of sentences under consideration. This latent semantic indexing (LSI) improves the approximate oracle score, since it gives partial credit for closely related terms that are not literally in the sentence. This is an attempt to move from a term-based oracle to an idea-based one: to the extent that the sentences represent the *main ideas* of the document, LSI projects the sentences onto a subspace of these ideas. The column sums of $\tilde{A}$ can be then viewed as *refined* approximate oracle scores for the sentences.
3. Sentences are then chosen for inclusion using a pivoted-QR decomposition of the matrix $\tilde{A}$. The pivoted-QR decomposition proceeds as follows:
   (a) Begin with an empty summary.
   (b) As long as the summary length is shorter than desired, choose the largest remaining column and include its sentence in the summary.
   (c) Subtract a multiple of this column from each remaining column in order to account for duplicate coverage of terms.
   (d) Continue until the desired summary length is reached.

In the usual pivoted-QR decomposition, the size of a column is measured by its Euclidean norm; the norm of a vector q with entries q_i is computed as

$$\|q\| = \left(\sum_i |q_i|^2 \right)^{1/2}.$$

The multiples that are subtracted make the remaining columns orthogonal to the column chosen. In our latest system, we use a nonnegative-QR decomposition. We measure size using the 1-norm

$$\|q\| = \sum_i |q_i|,$$

and after subtracting off the multiple, we replace any negative entries in the matrix by zero to avoid having well-covered terms increase the length of the column and thus make the sentence appear to be more important than it is. In tests on the MSE 2005 data, we found that this works better than the standard pivoted-QR decomposition to identify sentences that provide distinct information.

4 CLASSY Experiments

Four experiments that we ran for the Multilingual Summarization Evaluations and afterward are discussed here. Data for MSE consisted of clusters of related documents where each cluster contained some number of English and Arabic documents. Machine translated versions of the Arabic documents were also available. Figure 3 shows the ROUGE-2 scores for both human and system-generated summaries.

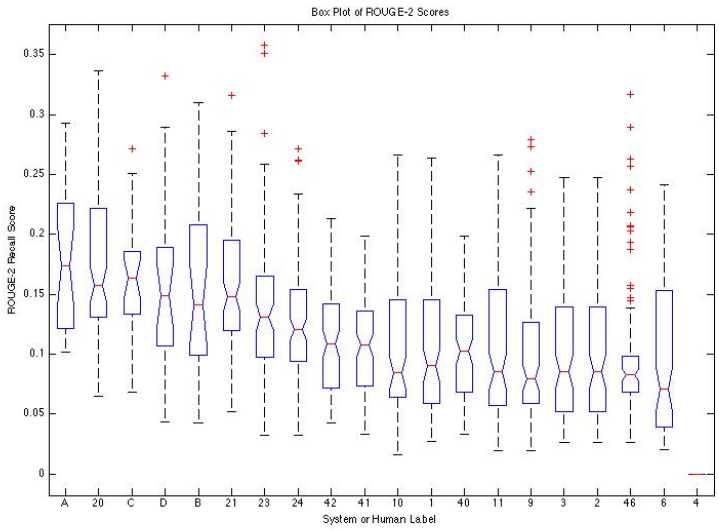

Fig. 3. Box Plot, sorted by mean, of ROUGE-2 Scores for All Humans and Systems. Letters represent human summary scores. Experiment 2 (Sect. 4.2) below is system 20; experiment 3 (Sect. 4.3) is system 21.

The experiments differed in which documents were used to select sentences for the summary and which documents were used to compute signature terms.

4.1 Experiment 1—English and Arabic Source Documents

The 2004 Document Understanding Conference (DUC '04, [13]) included two tasks to generate very short ($\leq$ 75 bytes, i.e., headlines) and short ($\leq$ 665 bytes) generic summaries using both MT-generated and "related" English background documents. The Lakhas system [5] used the original Arabic documents, rather than the translations, to generate headlines, which were then translated to English. Lakhas outperformed all the other systems that participated in this task.

Based on this result, we decided to experiment with using the original Arabic documents, rather than the MT translations, for one of our submissions to MSE. The Arabic documents were tokenized as 6-grams[2]. Signature tokens for each set of Arabic documents in a cluster were computed against an Arabic corpus. Independently, signature words were computed for the English documents in each cluster. Both the original Arabic and English sentences were scored using our summarization algorithms with the appropriate set of signature terms. When an Arabic sentence was selected for the summary, it was replaced with the corresponding sentence from the MT version of the document.

[2] 6-grams were chosen as a "reasonable" character length for tokens without creating too much of a computational load. For future efforts, we will most likely use white space splits for token identification.

We had expected this submission to perform well and were surprised when it scored lower than using the MT translations of the documents (described in Sect. 4.3). However, it still scored better than all submissions from other participants. We hypothesize that any gain we had from using the original Arabic was more than offset by the substitution of sentences from the noisy machine translations. This is consistent with results seen in Sects. 4.2, 4.3, and 4.4.

4.2 Experiment 2—English Documents Only

For this experiment, we used both the English and machine translations of the Arabic documents to compute signature terms for each cluster. Using the Arabic translations to compute the signature terms gave us larger clusters, which can improve the quality of the signature terms. However, in order to mitigate the noisy effects of machine translation, we chose sentences from the English documents only.

This English-only submission ranked first among all participating systems in MSE. Remarkably, the ROUGE [10] automatic evaluation system scores were better than 3 of the 4 human-generated summaries for ROUGE-2 and ROUGE-SU4, and 2 of the 4 human-generated summaries for ROUGE-1, and within the 95% confidence intervals for those humans who outscored the system. While CLASSY's performance is impressive, there are three points to remember. First, while the ROUGE performance measures have been shown to correlate well with human evaluation [10], they clearly are not a replacement. (We will address human evaluation in Sect. 4.5.) Second, the performance of the humans was limited by the poor quality of the translated documents. Third, we were able to exploit the fact that every Arabic document in a cluster had a closely related English document which, of course, is not always the case. Figure 4 shows a human-generated summary along with the CLASSY summary for the same document set.

4.3 Experiment 3—English and Translated Arabic Documents

Signature terms were computed identically as for Experiment 2 for this experiment. In this case, however, we used both the English documents and the machine translations of the Arabic documents to select sentences for the summary.

This English/MT-Arabic submission ranked second among all participating systems in MSE. While it was *always* outside the 95% confidence interval of the English-only submission on each of the ROUGE scores, it was always *within* the 95% confidence interval of at least 2 of the 4 human-generated summaries. We conjecture that the quality of the machine translation degraded both our summaries and the human summaries in a similar way.

4.4 Experiment 4—English Only

For this experiment, we used *only* the English documents for both signature term computation as well as summary selection. The purpose of this experiment was to measure the impact of using the translated Arabic documents for signature

Bombs exploded outside churches in Jakarta and five other Indonesian cities and towns on Christmas Eve, killing at least 14 people, injuring dozens and worsening the tension between Muslims and Christians. There were no immediate claims of responsibility, but religious violence and tensions have been rising throughout the predominantly Muslim country. Christians make up less than 10 percent, mostly ethnic Chinese, of Indonesia's 210 million people. President Abdurrahman Wahid asked Christians not to be provoked and blamed the attacks on forces intent on destabilizing the government". The Christmas celebrations coincide with the final days of Ramadan, Islam's month of fasting.

a. A Human-Generated Summary

Bombs exploded outside churches in Jakarta and five other Indonesian cities and towns on Christmas Eve, killing at least 14 people, injuring dozens and worsening the already difficult relations between Muslims and Christians throughout the fractured archipelago. Most of the bombs were planted in cars parked outside targeted churches _ including Jakarta's Roman Catholic cathedral, near the presidential palace and the capital's main mosque. Most of Indonesia's religious violence has been in the Moluccan islands, where about 5,000 Christians and Muslims have been killed over the past two years. Four of the dead Sunday were police officers who tried to

b. CLASSY-Generated Summary

Fig. 4. Example of Human- and CLASSY-generated Summaries

term computation (see Sect. 4.2). We computed a two-sided rank sum test, to test if the median ROUGE-2 Recall scores for both Experiment 2 and this experiment are equal for the MSE data. Forty-seven (47) of the scores from Experiment 2 were higher than their corresponding score from this experiment, 42 were less, and 7 were equal. The overall significance is a p-value of 0.2435, which means we cannot reject the null hypothesis that the medians are equal. Therefore, we conclude that while using the translated Arabic documents to compute signature terms did improve the ROUGE-2 scores, the improvement is not statistically significant.

4.5 Human Assessment

In addition to the automatic evaluation with ROUGE, a human evaluation was done. Human assessors read all the documents (both English and translated Arabic) for each cluster and then assigned each of the human- and machine-generated summaries to one of 5 equivalence classes–Unacceptable, Somewhat acceptable, Acceptable, Good, and Excellent (1 to 5, respectively)–describing *overall responsiveness* to the information presented in the documents in a cluster. Figure 5 is a scatter plot of the ROUGE-2 versus average overall responsiveness, the human evaluation score. The 8 machine systems and 4 human summaries

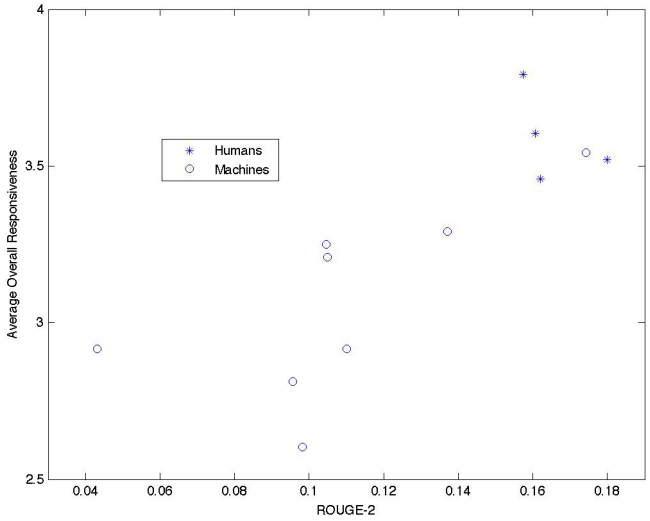

Fig. 5. Scatter Plot of ROUGE-2 Scores vs. Human Evaluation of Responsiveness for MSE systems. Our system is represented by the circle farthest to the right and to the top.

scores are displayed; CLASSY (system 20 in the scatter plot) is the only system to score at human levels of performance.

5 Conclusion and Continuing Efforts

Using the translated Arabic in conjunction with English to compute signature terms but then selecting sentences from *only* the English documents was a very successful approach. This perhaps indicates, as we have previously conjectured, that the Arabic documents in these collections did not provide any information beyond that contained in the English documents.

The summaries which used all documents for both computing signature terms *and* sentence selection, were statistically worse in a number of the ROUGE measures. We can only conclude that the inclusion of the MT sentences degraded the summary. With this said, this method scored second among all the submissions in all ROUGE measures.

These results indicate that when presented with a combination of documents in both English and Arabic (or, we suspect, any other language), that CLASSY, using signature terms computed from both English and the MT-versions of the Arabic documents, generates very good quality summaries.

A great deal more remains to be done. We realize that non-English documents will not always be as similar to "comparable" English documents as with the MSE data set. We would like to continue working with the original Arabic documents to better exploit them for the information and perspective that they contain, as compared to the English documents. We would also like to find

ways to improve the machine translations of the documents in order to more effectively use the translated content of the documents.

For both of these, we would like to improve basic non-English language tasks such as sentence splitting and lemmatization. Arabic presents serious challenges for these tasks, as do other languages. Early experiments suggest, however, that improvements to these would yield significant improvements to both the MT and summarization tasks.

We would also like to evaluate each of the components of CLASSY on languages other than English. For example, we do not know if the method we use for redundancy removal will be effective on non-English languages. Our trimming methods are truly language dependent. We would like to identify a class of trims that are "universal" for all languages, even when they appear quite different in different languages. We also need to compile trims that are useful for a single language or class of languages.

Acknowledgments

The authors thank Michael Graham for the work he did in developing the sentence splitter that we are now using. We also wish to thank the organizers of both MSE 2005 and 2006 for all their hard work in creating the data sets and performing the evaluation.

The work of Dianne O'Leary was partially supported by NSF Grant CCF-0514213.

References

1. Conroy, J.M., et al.: Left-Brain Right-Brain Multi-Document Summarization. In: Procs. of 2004 Document Understanding Conference, Boston, MA (2004), `http://duc.nist.gov/`
2. Conroy, J.M., Schlesinger, J.D., O'Leary, D.P.: Topic-focused Multi-document Summarization Using an Approximate Oracle Score. In: Procs. of 2006 ACL/COLING Conference, Sydney, Australia (2006)
3. Conroy, J.M., Schlesinger, J.D., Stewart, J.G.: CLASSY Query-based Multi-document Summarization. In: Procs. of 2005 Document Understanding Conference, Vancouver, BC (2005), `http://duc.nist.gov/`
4. Conroy, J.M., et al.: Back to Basics: CLASSY 2006. In: Procs. of 2006 Document Understanding Conference, New York (2006), `http://duc.nist.gov/`
5. Douzidia, F.S., Lapalme, G.: Lakhas, an Arabic Summarization System. In: Procs. of 2004 Document Understanding Conference, Boston, MA (2004), `http://duc.nist.gov/`
6. Evans, D.K., McKeown, K., Klavans, J.L.: Similarity-based Multilingual Multi-document Summarization. Technical Report CUCS-014-05, Department of Computer Science, Columbia University (2005)
7. Fung, P., Ngai, G.: One Story, One Flow: Hidden Markov Story Models for Multilingual Multidocument. ACM Trans. on Speech and Language Processing (TSLP) 3(2) (2006)

8. Hatzivassiloglou, V., et al.: Simfinder: A Flexible Clustering Tool for Summarization. In: Procs. of NAACL 2001 Workshop on Automatic Summarization, Pittsburg, PA (2001)
9. Lin, C.-Y., Hovy, E.: The Automated Acquisition of Topic Signatures for Text Summarization. In: Procs. of 18th Intl. Conference on Computational Linguistics (COLING 2000), Saarbrücken, Germany (2000)
10. Lin, C.-Y., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: Procs. of North American Chapter of the Association for Computational Linguistics, vol. 1, Edmonton, Alberta (2003)
11. Mani, I., Maybury, M.T. (eds.): Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)
12. Nenkova, A., Vanderwende, L.: The Impact of Frequency on Summarization. Technical Report MSR-TR-2005-101, Microsoft Research (2005)
13. Over, P., Yen, J.: An Introduction to DUC-2004: Intrinsic Evaluation of Generic News Text Summarization Systems. In: Procs. of 2004 Document Understanding Conference, Boston, MA (2004), http://duc.nist.gov/
14. Porter, M.: An Algorithm for Suffix Stripping. Program 14(3), 130–137 (1980)
15. Radev, D., et al.: MEAD - A Platform for Multidocument Multilingual Text Summarization. In: Procs. of Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004)
16. Schiffman, B., Nenkova, A., McKeown, K.: Experiments in Multidocument Summarization. In: San Diego, C.A. (ed.) Procs. of Human Language Technology Conference, San Diego, CA (2002)
17. Document Understanding Conferences Publications. http://www-nlpir.nist.gov/projects/duc/pubs.html.