CLASSY Arabic and English Multi-Document Summarization

John M. Conroy
IDA/Center for Computing Sciences
conroy@super.org

Dianne P. O'Leary
University of Maryland
oleary@cs.umd.edu

Judith D. Schlesinger
IDA/Center for Computing Sciences
judith@super.org

1 Introduction

Our Multilingual Summarization Evaluation entries for MSE-2006 were based upon an improved version of our CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) system. Our two entries were systems 20 and 21 and represented approaches based upon extracts from a) only English documents and b) English and the translated Arabic documents (full clusters). This paper presents a brief review of the method we used, including adaptations made for MSE 2006.

An analysis of the results of our efforts using ROUGE is also discussed.

2 Description of Our System

2.1 Splitting Sentences

The sentence splitter we used in previous versions of our system was a commercial product, which performed POS (Part of Speech) tagging along with sentence splitting, for which we did not have source code. It proved ineffective in our work because we could not make changes in response to splitting errors. Our new Java-based sentence splitter, developed in-house, can be updated as needed, and a post-processing phase tries to correct errors due to:

- erroneous splits on foreign words, especially on names, that appear to be abbreviations of English words;
- erroneous splits on less commonly used abbreviations;
- erroneous splits due to missing or bad punctuation; and
- erroneous splits due to ellipsis at sentence end.

Unfortunately, some errors remain that require full parsing (which we do not perform) to detect.

2.2 Trimming Sentences

Prior to 2006, we used POS tags to help identify individual words, phrases, and/or clauses to be trimmed from each sentence. This POS-tagger, like all available, made errors that caused poor trimming and was the slowest component in our system. In some cases it failed entirely, returning no tagged file.

Because of this, we rewrote our trimming code to eliminate dependence on POS tags. 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 to remove.

Our sentence trimming tasks basically remain as they have been[4].

1. We remove extraneous words that appear in a sentence, including date lines, editor's comments, and so on.
2. We remove adverbs and conjunctions, including phrases such as "As a matter of fact," and "At this point," that occur at the start of a sentence.
3. We remove a small selection of words that occur in the middle of a sentence, such as "however" and "also". Some of these require commas; some do not.
4. For 2006, we added the removal of age references such as ", 51," or ", aged 24,".
5. We remove gerund phrases (phrases starting with the -ing form of a verb) from the start, middle, or end of a sentence.
6. We remove relative clause attributives (clauses beginning with "who(m)", "which", "when", and "where") wherever possible.
7. We remove attributions, such as "police said", at the start or end of sentences when the text is not a quote.

2.3 Scoring Sentences

The method we used to score sentences is new this year. We use an approximate Oracle score, which approximates the fraction of *human abstract terms* a sentences contains. Details of this approach and its motivation can be found in [3] [2]. We give a brief overview here and cite two significant improvements on the approach which are new.

Instead of using term frequencies of the corpus, as done by [6], to infer highly likely terms in human summaries, we propose to directly model the *set* of terms (vocabulary) that is likely to occur in a sample of human summaries. The following description is taken directly from [3].

We model human variation in summary generation with a unigram bag-of-words model of the terms. 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 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. We produce a computable *approximate oracle score* ([2]) to substitute for this score.

If we were given a set of human abstracts for a topic τ, we can readily compute the maximum-likelihood estimate of $P(t|\tau)$. Suppose we are given h sample summaries generated independently. Let $c_{it}(\tau) = 1$ if the i-th summary contains the term t and 0 otherwise. Then the maximum-likelihood estimate of $P(t\tau)$ is given by

$$\hat{P}(t|\tau) = \frac{1}{h}\sum_{i=1}^{h} c_{it}(\tau).$$

We define $\hat{\omega}$ by replacing P with $\hat{P}$ in the definition of ω. Thus, $\hat{\omega}$ is the maximum-likelihood estimate for ω, given a set of h human summaries. This *maximum likelihood oracle score*, which allows us to compute the expected number of abstract terms in a sentence, was shown to achieve ROUGE-2 performance exceeding that of the humans on DUC-2005 data ([2]). We use this oracle score as a guide to develop approximate scores when the human abstracts are not known.

To estimate $P(t|\tau)$, we view the signature terms as "samples" from idealized human summaries. Loosely, a signature term is a term which occurs significantly more than expected ([5, 1]). Here, in contrast to previous versions of CLASSY, we used the Porter stemmer [7], which significantly improves the signature terms correlation with human abstract terms. As such, we expect that the set of these terms may approximate the underlying set of human summary terms.

We define the signature term approximation of the oracle score of a sentence's expected number of human abstract terms as

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

$P_s(t|\tau) = 1$ if t is a signature and 0 otherwise. We denote $|x|$ as the number of distinct terms sentence x contains, T is the universal set of all terms and $x(t) = 1$ if the sentence x contains the term t and 0 otherwise.

In [3] we improved the estimate for the probability that a given term is an abstract term by using pseudo-relevance feedback; however, we found that when the simple approximation was coupled with the new redundancy removal method, pseudo-relevance feedback was not necessary.

2.4 Reducing Redundancy of the Selected Sentences

To reduce redundancy in the sentences chosen for inclusion in the summary, we have a three-step process that we are using for the first time in MSE 2006.

1. We order the sentences by score and choose enough sentences to produce a summary 9 times as long as desired. This is more sentences than we previously considered and the length was chosen empirically, based on training on MSE 2005 data.

2. We replace the term-sentence matrix A for these sentences by using singular value decomposition to compute a rank-k approximation $\tilde{A}$ to the matrix, where $k = \max(1, \lfloor 0.65\, n \rfloor)$ and n is the number of sentences under consideration. This use of latent semantic indexing (LSI) is new to CLASSY. The LSI portion can be viewed as a method of improving the approximate oracle score, as the column sums of A are the approximate oracle scores for the top scoring sentences. To the extent that these sentences represent the *main ideas* of the document, LSI projects the sentences onto the subspace of these ideas. The column sums of $\tilde{A}$ can be then viewed as *refined* approximate oracle scores for the sentences. We conjecture that it is this refinement that makes the simple pseudo-relevance feedback superfluous.

3. We then choose the sentences for inclusion using a matrix decomposition of $\tilde{A}$. Previously, we used a pivoted-QR decomposition [1] to identify sentences that provide distinct information. In tests on the DUC 2005 data, we found that a nonnegative-QR decomposition worked better, so this was used in MSE 2006.

The nonnegative-QR decomposition proceeds as follows:

> Begin with an empty summary. As long as the summary length is shorter than desired, choose the largest remaining column and include its sentence in the summary. Subtract a multiple of this column from each remaining column in order to account for duplicate coverage of terms. Continue until the desired summary length is reached.

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

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

and the multiples that are subtracted make the remaining columns orthogonal to the column chosen. In this year's entry, we measure size using the 1-norm:

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

and after the orthogonalization, 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.

3 CLASSY Submissions

Our submissions for MSE 2006 used the given data in two different ways. For both submissions, we used both the English documents and the machine translations of the Arabic documents to obtain signature terms. Submission 21 then chose sentences from all of these documents using the algorithm described in the previous section.

For submission 20, as in our second submission last year, we mitigated the effects of machine translation by choosing sentences from the English documents only, although the signature terms were the same as for submission 21.

4 Results

Our submissions, 20 and 21, rank first and second among peer systems in each of the ROUGE-based evaluations (ROUGE-1, ROUGE-2, and ROUGE-SU4). (See Tables 1, 3, and 2 for human scores as well as those of some other submissions.) Remarkably, submission 20's ROUGE scores were better than 3 of the humans for ROUGE-2 and ROUGE-SU4 and 2 of the humans for ROUGE-1, and within the 95% confidence intervals for those humans who outscored the system. While our submission 21 was always outside the 95% confidence interval of system 20, it was always within the 95% confidence interval of at least 2 of the 4 human model summaries.

Submission	Mean	95% CI Lower	95% CI Upper
A	0.47131	0.44753	0.49559
C	0.46207	0.43823	0.48536
20	0.45054	0.43694	0.46381
B	0.44935	0.41933	0.47789
D	0.44504	0.41404	0.47708
21	0.43035	0.41910	0.44245
23	0.42354	0.41166	0.43606
24	0.41970	0.41016	0.43012
41	0.38728	0.37694	0.39808
42	0.38555	0.37517	0.39583
40	0.38149	0.37199	0.39129
10	0.38125	0.37032	0.39152
11	0.37395	0.36165	0.38559
9	0.37288	0.36136	0.38374
1	0.36716	0.35610	0.37836
3	0.36572	0.35464	0.37766
2	0.36572	0.35464	0.37766
46	0.35302	0.34485	0.36149
6	0.34116	0.32536	0.35668
4	0.03526	0.03271	0.03783

Table 1: Average ROUGE-1 Recall

5 Conclusion and Future Efforts

We are very pleased with our system's performance. Using the translated Arabic in conjunction with English to compute signature terms but then selecting sentences from the English documents (submission 20) was a very successful approach. This perhaps indicates, as we have previously conjectured, that the Arabic documents did not provide any information beyond that contained in the English documents.

As in MSE 2005, our submission (21), which used all documents for computing signature terms and sentence selection, was statistically worse in a number of the ROUGE measures. We can only conclude that the inclusion of the machine translation sentences degraded the summary. With this said, this method scored second among all the submissions in all ROUGE measures.

References

[1] John M. Conroy, Judith D. Schlesinger, and Jade Goldstein. Three classy ways to perform arabic and english multi-document summarization. In *Multi-Lingual Summarization Evaluation 2005*, volume http://www.isi.edu/ cyl/MTSE2005/MSE2005/papers/index.html, 2005.

[2] John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the ACL'06/COLING'06*, 2006.

[3] John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary, and Jade Goldstein. Back to basics: Classy 2006. In *Proceedings of the 2006 Document Understanding Workshop, New York*, 2006.

[4] Judith D. Schlesinger John M. Conroy and Jade Goldstein Stewart. Classy query-based multi-document summarization. In *Proceedings of the 2005 Document Understanding Workshop, Boston*, 2005.

Submission	Mean	95% CI Lower	95% CI Upper
A	0.17986	0.15791	0.20378
20	0.17419	0.16176	0.18731
C	0.16211	0.14214	0.18169
D	0.16075	0.13382	0.19102
B	0.15732	0.12879	0.18615
21	0.15612	0.14529	0.16785
23	0.13718	0.12558	0.14891
24	0.12948	0.11989	0.13971
42	0.11015	0.10223	0.11806
41	0.10779	0.09982	0.11564
10	0.10498	0.09300	0.11633
1	0.10451	0.09356	0.11527
40	0.10380	0.09669	0.11117
11	0.10346	0.09173	0.11436
9	0.09903	0.08708	0.11024
3	0.09898	0.08798	0.11122
2	0.09898	0.08798	0.11122
46	0.09813	0.08803	0.10887
6	0.09561	0.08277	0.10872
4	0.00000	0.00000	0.00000

Table 2: Average Recall of ROUGE-2

[5] C.Y. Lin and E. Hovy. The automatic acquisition of topic signatures for text summarization. In *DUC 02 Conference Proceedings*, 2002.

[6] Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. Technical Report MSR-TR-2005-101, Microsoft Research, 2005.

[7] Martin Porter. An algorithm for suffix stripping. In *Program*, volume 14(3), pages 130–137, 1980.

Run	Mean	95% CI Lower	95% CI Upper
A	0.20385	0.18402	0.22534
20	0.20034	0.18854	0.21247
C	0.18903	0.17191	0.20661
B	0.18526	0.16092	0.20991
D	0.18456	0.16048	0.20975
21	0.18270	0.17332	0.19288
23	0.17150	0.16172	0.18153
24	0.16884	0.16069	0.17738
42	0.14682	0.13892	0.15443
41	0.14661	0.13849	0.15459
10	0.14439	0.13462	0.15386
40	0.14253	0.13594	0.14971
11	0.14208	0.13225	0.15189
1	0.14128	0.13222	0.15025
3	0.13987	0.13047	0.15036
2	0.13987	0.13047	0.15036
9	0.13961	0.13024	0.14842
46	0.12828	0.11973	0.13745
6	0.12826	0.11703	0.13972
4	0.00666	0.00610	0.00722

Table 3: Average recall of ROUGE-SU4