CLASSY 2011 at TAC: Guided and Multi-lingual Summaries and Evaluation Metrics

John M. Conroy Judith D. Schlesinger Jeff Kubina
IDA/Center for Computing Sciences Department of Defense
{conroy,judith}@super.org jeff.kubina@gmail.com

Peter A. Rankel Dianne P. O'Leary
University of Maryland
{rankel@math,oleary@cs}.umd.edu

Abstract

We present CLASSY's guided summarization as well as multi-lingual methods as submitted to TAC 2011. In addition, we describe improved metrics submitted to the AESOP task at TAC.

1 Introduction

CLASSY participated in both the guided summarization and multi-lingual summarization tasks of TAC 2011. We are happy with our performance, as described below. There is still room for improvement, however.

Our metrics for AESOP were also successful but, here again, there is still much more that can and should be done.

2 Guided Summarization

2.1 Linguistics

Data preparation consists of 1) sentence splitting, trimming, and categorization, and 2) query-term generation. Both of these tasks remained very consistent with 2010. The code for splitting and trimming is quite stable, other than some minor fixes when errors are discovered. We tag sentences to be in one of three categories: a) don't use for any reason; b) use for statistics only; and c) consider for use in the summary. For query terms, we used the aspects data structure we created last year ([4]) to generate a list of terms for each document set. Based on last year's results, we used a rich list of query terms influenced by every aspect for each associated category.

Our biggest effort this year was creating the "clean" data for other participants to use. We first prepared the 2010 data for testing and then the 2011 data. The clean data consisted of first sentence splitting and then identifying and flagging "boilerplate", the first step of our trimming task. No other trimming was performed.

The change in data being drawn from the KBP corpus rather than AQUAINT2 as in prior years revealed an unfortunate bug in the sentence splitter. This caused some datelines to be attached to the following sentence, even though they were separated by paragraph markers. E.g.,

$< P >$
QUARRYVILLE, Pennsylvania ....
$< /P >$
$< P >$
A man ...

became

QUARRYVILLE, Pennsylvania .... A man ....

This bug was unknown until reported by a user of the 2011 clean data. At that point, following the TAC rules, nothing could be done to correct the error. We found that sentences with these attached datelines occurred in 8 of our base summaries and 3 of our update summaries. We assume that our linguistic quality and overall responsiveness scores were impacted by this.

2.2 Algorithms

Once sentences have been defined, trimmed, and categorized by the linguistic step they are scored and selected. This year a tuned bigram version of CLASSY was submitted. CLASSY 2010 was unigram based while CLASSY 2011 uses stemmed bigrams with no stop words removed and the fully expanded set of query terms to cover aspects. The algorithm for sentence scoring and selection has three parts:

1. An estimate for the probability that a term will be included in a human generated summary is generated for each term. We call such terms "summary content terms" or SCTs for short. The score for a sentence is the expected number of SCTs divided by the length of the sentence.

2. A non-redundant subset of high scoring sentences is chosen using non-negative matrix factorization. See [10] for details. For update summaries, the term-sentence matrix is projected to minimize repetition of information in the base summary.

3. A subset of sentences selected in the second step is chosen to achieve the desired summary word length of 100 words. This is done using a branch and bound algorithm to approximately solve a knapsack problem.

The two submissions to the update task differed only in the first of the above three steps. Both methods were tuned using the automatic evaluation metrics, ROUGE-2 and Nouveau ROUGE-2 scoring (see [3]) on the TAC 2010 data.

- System 25 is a bigram version CLASSY 2010 ([4]) with weighting for the mixture model set to maximize automatic evaluation metrics as well as weighting for the projection used for the update summaries. The weighting was substantially changed to account for the bigrams and more extensive query set.
- System 42 used naïve Bayes term scoring based term features, which we now describe.

2.2.1 Naïve Bayes Model for Term Scoring

Following [2], we seek to estimate $P(t|\tau)$, the probability that term t will be included in a human-generated summary on topic τ. Instead of using a simple mixture model, we estimate the probability using machine learning based on TAC 2010 data. In system 42, as well as the multi-lingual task (see Section 3), we experimented with several machine learning approaches to computing estimates $\hat{P}(t|\tau)$, but found naïve Bayes to give the best performance. (In addition to naïve Bayes, we experimented with linear discriminant analysis and random forest classifiers.) For base summaries, the features used are:

- $\log(p)$, where p is the $p-$value of the Dunning signature term statistic ([5], [7]).
- Text rank as computed by term co-occurrence in sentences. Terms with a signature term $p-$value less than 0.001 were excluded unless they were identified as query terms ([8]).

For update summaries, the following two features were found to give the best performance on the TAC 2010 data:

- Text rank for the current cluster of documents.
- The log of the ratio of the text rank for a term in the current cluster to the text rank of the term in the base cluster. Both probabilities are smoothed by adding 0.01.

The term weights are computed via the posterior probability that a term would be included in 0, 1, 2, 3, or 4 human summaries and this 5-long vector is used to compute the expected value which is the maximum likelihood estimate of the probability that a term would be included in a human-generated summary.

2.3 Results

CLASSY submission 25 finished first in the base summary task, scoring the highest total score in overall responsiveness, the human judgement score that NIST uses to judge a summary's ability to respond to the need of the task.

The submission 25 update summaries finished 3rd, but our scores were statistically indistinguishable from the top performing system as measured by a paired Wilcoxon test.

Submission 42 was a new model which had a strong performance. While its rank was lower, its responsiveness scores were statistically indistinguishable from our best system.

3 Multilingual Summarization

Other than the MT-Arabic task of DUC 2004 and MSE (Multi-lingual Summarization Evaluation) using Arabic in both 2005 and 2006 ([10]), CLASSY has not been used for any language but English. The multi-lingual summarization pilot for TAC 2011 was our first "large-scale" endeavor with other languages.

3.1 Linguistics

Sentence splitting was the only linguistic task we used to process the multi-lingual data. We stripped FASST-E of all the code that dealt with the vagaries of English and created two new versions of the sentence splitter: FASST-CAP for languages that use alphabets with two cases of letters (Greek, and Latin (extended) for French and Czech) and FASST-ONE for languages that use alphabets with a single case (Hebrew, Arabic, and Devanagari for Hindi).

Splitting was done by focusing on the typical end characters of ".", "?", and "!". For Arabic and Hindi, respectively, we also used the characters specified in the unicode: the Arabic question mark and the single and double danda.

For English, we had identified a list of abbreviations that *do not typically* terminate a sentence. This includes honorifics such as "Dr.", "Prof.", etc. and other abbreviations such as "e.g.", "est.", etc. We used Google translate to try to determine if there were similar abbreviations using the ".", in each of the other 6 languages. This list was quite short for Hebrew, which typically uses "" or "" for abbreviations, to fairly rich for Greek, with the others falling somewhere in between. These lists can easily be modified if we learn that something is there in error or if we learn of an abbreviation that we didn't previously include. Of course, having "abbreviations" that don't really exist in the language won't usually hurt anything.

We have yet to analyze our output to determine how well our sentence splitting performed.

3.2 Algorithms

To adapt CLASSY to the 7 languages for the multi-lingual pilot, several steps were taken. First, backgrounds for each of the languages were collected so the Dunning G-statistic could be computed ([5], [7]). Second, a model training set had to be identified. For the background, Wikinews was a natural choice. The model training set posed a bit more of a challenge. The target length of the summaries was 250 words, which was the same as DUC 2005–2007; however, these summaries were "topic-focused" and no topic descriptions were to be provided for the multi-lingual pilot. To this end, CLASSY was trained on DUC 2005–2007 data, without the use of the topics.

Term scoring was limited to the naïve Bayes term weighting (Section 2.2.1) as it performed slightly better on the DUC data sets. The features that were found to perform well on the DUC 2005–2007 data include the two features used in the base summaries for the guided summary task as well a relevance feedback feature borrowed from the classic CLASSY mixture model, and simply, the normalized frequency of a term. More specifically, we used the following features:

1. $\log(p)$, where p is the $p-$value of the Dunning signature term statistic ([5], [7]).
2. Text rank as computed by term co-occurrence in sentences. Terms with a signature term $p-$value less than 0.001 are excluded ([8]).
3. $\log(P(t|S_0)$, log probability that a term occurs in a sentence in the cluster of documents to be summarized.
4. $\log(P(t|S_1)$, log probability that a term occurs in a sentence with 1 or more signature terms in the cluster of documents to be summarized.

The feature data and annotations for the DUC 2005–2007 data were then used to generate term scores for each of the 7 languages.

For sentence selection, the non-negative matrix factorization ([10]) and an integer programming method ([6]) were evaluated. While these methods gave comparable quality summaries as measured by the automatic metrics for the 100 word TAC 2010 data, the ILP approach was significantly better for the 250 word summaries of the DUC data!

3.3 Results

CLASSY finished 2nd or 3rd in 5 out 7 languages. (The seven languages were Arabic, Czech, English, French, Greek, Hebrew, and Hindi.) We were statistically tied for first in English, Greek, and Hindi.

In TAC 2009 ([1]), we noted that while the CLASSY adapted ILP achieved comparable ROUGE scores to the non-negative matrix factorization, it did so using more sentences, which we believe led to lower linguistic scores as measured in 2009. It is unclear if short sentences hurt CLASSY's multi-lingual performance in the multi-lingual pilot task.

4 AESOP

Our submissions to AESOP built on our TAC 2010 ([4]) approaches in that we computed two sets of features–those that are designed to correlate well with content and those that aim to measure linguistic quality. This year an improved set of content features were employed. The selection and weighting of the features was achieved using one of three linear algebraic methods to maximize Pearson correlation. The details are given in [9] and the next two sections present an overview of the approach and the results.

4.1 Feature Generation

Based on the outcome of AESOP 2010, it seemed that word bigrams produced the best results in predicting the content measure of a summary. In particular, ROUGE-2 dominated for correlation with the pyramid score. As such, we focused on variations of bigram scores for content measure. In all, we investigated six variations of bigrams, the first 2 of which were ROUGE.

1. ROUGE-2, (R2) the consecutive bigram score.
2. ROUGE-SU4, (SU4) the bigram score that allows for a skip distance of up to 4 words.
3. Bigram coverage score (Coverage). This score is similar to ROUGE-2 but does not take the frequency that the bigram occurs in either the model summaries or in the summary to be scored. A credit of $\frac{i}{n}$ for a bigram is given if i out of n model human summaries included that term.
4. Unnormalized ROUGE-2 (Bigram). The score is essentially ROUGE-2 without the normalization for the length of the summaries.
5. Bigram coverage, as measured by a point to point (Coverage P2P). This score is similar to the 3rd score; however, it is computed comparing one summary to another as opposed to one summary to 3 or 4 summaries.
6. Unnormalized ROUGE-2 as measured by a point to point comparison (Bigram P2P). This score is a point to point version of score 4.

4.2 Feature Selection and Weighting

TAC 2009 and 2010 were used to train the model. All $2^{13} - 1$ subsets of features were considered and three methods used to compute weights, as was done in TAC 2010 ([4]). Each method was trained using average system performance and average predictors. The three methods for computing the weights were:

1. Canonical Correlation (canon) which computes optimal linear combinations of features and the predictors (pyramid, responsiveness, and readability) to maximize the Pearson correlation.

2. Robust least squares (robust) solves a least squares problem to predict one of the response vectors of pyramid score, responsiveness, or readability. This method is robust to outliers.

3. Non-negative least squares (nonneg). A least square predictor for one of the response vectors of pyramid score, responsiveness, or readability that restricts itself to non-negative weights.

4.3 Results

For the update summaries, in both the "all peers" and "no models" subtasks, the CLASSY metrics were the best at predicting pyramid scoring, a human judgement of content. The CLASSY metrics significantly outperformed ROUGE in the "all peers" evaluations. "All peers" means that both human and machine systems are evaluated by the automatic metrics.

For the corresponding 6 results for base summaries, at least 2 CLASSY metrics scored within the 95% confidence interval for all but one category, for predicting overall responsiveness in the all peers case.

5 Conclusions and Future Efforts

"Classic" CLASSY again performed well at the guided summary task. The new approach of term scoring, while competitive, did not perform as well. This approach combined with an ILP was a strong competitor in the multi-lingual pilot.

A long-standing task has been to handle anaphora. Our current strategy of not selecting any sentence beginning with a pronoun works well to maintain accuracy, continuity, and readability but means we eliminate some very good sentences from selection. We hope to finally turn our attention to this task.

We need to evaluate our sentence splitting for languages other than English and strengthen the algorithms to account for any errors we find. We would like to extend our linguistic processing to non-English languages. For some languages, we suspect that our lead and medial phrase trimming will be quite useful. Sentence structure is a strong predictor of this. We also need to investigate why CLASSY did not work as well with some languages as with others.

Further investigation is needed into what can make the machine learning stronger for both the guided summarization and multi-lingual task. This year, the "metric-gap" was narrowed and, for the first time, several metrics significantly outperformed ROUGE in

the "all peers" task. Further work is needed to predict readability of machine generated summaries, which, hopefully, will yield metrics to improve summarization.

References

[1] John M. Conroy and Judith D. Schlesinger. CLASSY 2009: Summarization and Metrics. In *TAC 2009 Workshop Proceedings, http://www.nist.gov/tac/publications/index.html*, 2009.

[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*, pages 152–159, Sydney, Australia, July 2006.

[3] John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. Nouveau-ROUGE: A Novelty Metric for Update Summarization. *Computational Linguistics*, 37(1):1–8, 2011.

[4] John M. Conroy, Judith D. Schlesinger, Peter A. Rankel, and Dianne P. O'Leary. Guiding CLASSY Toward More Responsive Summaries. In *TAC 2010 Workshop Proceedings, http://www.nist.gov/tac/publications/index.html*, 2010.

[5] T. Dunning. "Accurate Methods for Statistics of Surprise and Coincidence". *Computational Linguistics*, 19:61–74, 1993.

[6] Dan Gillick, Benoit Favre, and Dilek Hakkani-Tür. The ICSI Summarization System at TAC 2008. In *TAC 2008 Workshop Proceedings, http://www.nist.gov/tac/publications/index.html*, 2008.

[7] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[8] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Text. In *EMNLP*, pages 404–411, Barcelona, Spain, July 2004. ACL.

[9] Peter A. Rankel, John M. Conroy, and Judith D. Schlesinger. Better metrics to automatically predict the quality of a text summary. *SIAM Data Mining Text Mining Workshop 2012, MDPI Journal of Algorithms, http://www.mdpi.com/journal/algorithms*, 2012.

[10] Judith D. Schlesinger, Dianne P. O'Leary, and John M. Conroy. Arabic/English Multi-document Summarization with CLASSY–The Past and the Future. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, pages 568–581, Haifa, Israel, February 2008. Springer.