Microsoft Research at DUC2006: Task-Focused Summarization with Sentence Simplification and Lexical Expansion

Lucy Vanderwende, Hisami Suzuki, Chris Brockett
Microsoft Research
One Microsoft Way, Redmond, WA 98052
{lucyv,hisamis,chrisbkt}@ microsoft.com

Abstract

Our DUC2006 system comprised three main components: a task-focused extractive summarization system, sentence simplification, and lexical expansion of topic words. This paper details each of these components, together with experiments designed to quantify their individual contributions. We include an analysis of our results according to two independent human evaluation methods, the NIST evaluation and the Pyramid evaluation. Our system ranked first in terms of both overall mean score and averaged per-cluster mean ranking out of 22 systems in the Pyramid evaluation, and ranked third out of 35 systems in NIST content responsiveness.

1 Introduction

The DUC2006 task is to produce summaries of sets of documents in response to short topic statements that define what the summaries should address. The summaries are limited to 250 words in length. To evaluate the summaries produced by the participants' systems, henceforth peer summaries, DUC provides four human summaries, henceforth model summaries, for comparison. We participated in DUC2006 by submitting peer summaries, as well as by providing manual annotation for the Pyramid analysis (Nenkova and Passonneau, 2005).

Our contribution in DUC2006, System 10, builds on an earlier system, SumBasic (Nenkova and Vanderwende, 2005), which produces generic multi-document summaries; we will provide a brief description of SumBasic in Section 2. We will then describe each of the three main components that comprise our system: a task-focused extractive summarization system, sentence simplification, and lexical expansion of topic words. We will provide experiments designed to quantify the contributions of each component.

Our system, like most systems participating in DUC, is designed to produce extractive summaries by selecting sentences from the document set, either verbatim or with some simplification. No consideration is given to sentence ordering or cohesion other than that sentence ordering is determined exclusively as a result of the sentence selection process (see Section 2 for details). As (semi-) automated methods of evaluation become able to measure ordering and cohesion, we look forward to working on those aspects of summarization.

DUC2006 evaluation includes both automated metrics (ROUGE and BE) as well as metrics that are the result of human evaluation and annotation (NIST and Pyramid). In our discussions of evaluation in Sections 4.2 and 6, we will focus on the Pyramid evaluation, which measures the content overlap between the peer summary and the combined model summaries, as our primary goal was to maximize the semantic content of the summary as opposed to sentence cohesion. Out of 22 systems participating in the Pyramid evaluation, our system ranked first in terms of both overall mean Pyramid score and averaged per-cluster mean ranking. In another human evaluation, that of NIST content responsiveness, our system was rated third out of 35 systems, indicating that it succeeded in extracting important semantic content as judged by two independent human evaluation metrics.

2 Core System: SumBasic

SumBasic (Nenkova and Vanderwende, 2005) is a system that produces generic multi-document summaries. Its design is motivated by the observation that words occurring frequently in the document cluster occur with higher probability in the human summaries than words occurring less frequently. Sentence weights in SumBasic are assigned by computing the average of the word proabilities derived from the word frequency in the document set. Sentence selection in SumBasic is accomplished by iteratively selecting the word with the highest probability and finding the highest scoring sentence containing that word. In order not to select one sentence multiple times, SumBasic updates all weights for words in this sentence by squaring their probability, with the intention of modeling the probability that a word occurs twice in a summary. These two steps are repeated until the maximum summary length has been reached. The system resembles SUM_{avr} as recently described in Nenkova et al. (2006), except that the update function in SumBasic uses squaring rather than multiplication by a very small number.

3 SumFocus

To participate in DUC2006, we took a naïve approach to modifying SumBasic to produce topic-focused multi-document summaries, which we call SumFocus. We capture the information conveyed by the topic description by computing the word probabilities of the topic description. Having done so, we now compute the probability for each word as a linear combination of the unigram probabilities derived from the topic description, with back-off smoothing, and the unigram probabilities from the document, in the following manner (all other aspects of SumBasic remain unchanged):

$$\textit{WordScore} = (1-\lambda)*\textit{DocScore} + \lambda*\textit{TopicScore}$$

The optimal value of λ, 0.9, was empirically determined using the DUC2005 corpus, optimizing on ROUGE-2 scores (henceforth R-2).

Since sentence selection is controlled by choosing the words with the highest probability, which we call “best words”, it is in principle possible for the best words to come from either the document or from the topic description. In practice, however, the best word is nearly always a word from the topic description due to the very high value of λ. For DUC2005 overall, 618 document words were identified as best on the basis of topic statements and only 22 independently on the basis of frequency alone. For DUC2006 overall, all 600 document words flagged as best were matched by the topic statements. We did add a small stopword list of topic words based on DUC2005 data (*describe*, *discuss*, *explain*, *identify*, *include*, *including*, *involve*, *involving*), which might account for document words being chosen as best words in DUC2005 but not in DUC2006. We note that the topic statements in DUC2006 appear to contain more instructions than in DUC2005. This is certain to have been a factor and suggests additional candidate stopwords (e.g., *concerning*, *note*, *specify*, *give*, *examples*, and *involved*).

ROUGE	System	DUC2005	DUC2006
R-1	SumBasic	0.25605	0.30599
	SumFocus	0.25358	0.30153
R-2	SumBasic	0.03642	0.05453
	SumFocus	0.04054	0.06053
R-SU4	SumBasic	0.06631	0.08769
	SumFocus	0.06975	0.09224

Table 1. ROUGE average recall scores, not using stopwords, for SumBasic and SumFocus.

Table 1 compares the performance of SumBasic and SumFocus on both DUC2005 and DUC2006 tasks[1]. As Table 1 shows, the adaptation of SumBasic to SumFocus yields higher R-2 and R-SU4 scores, though the confidence intervals reported by the ROUGE tool overlap slightly. When we compute ROUGE scores for each cluster individually, however, it is clear that at least some clusters are negatively impacted by the topic focus.

4 Sentence Simplification

Our goal is to create a summarization system that produces summaries with as much content as possible that satisfies the user. Since summaries are extractive, we view sentence simplification (also known as *sentence shortening* or *sentence compression*) as a means of creating more space within which to capture important content.

The most common approach to sentence simplification has been to shorten the sentences to be used in the summary deterministically. For example, the CLASSY system (Conroy et al., 2005) incorporates a heuristic component for sentence simplification that pre-processes the sentences used in

[1] For all experiments in this paper, we compute ROUGE without jack-knifing, which allows us to use 4 model summaries instead of 3 for more stable results. Note that DUC reports ROUGE numbers computed with jack-knifing for better differentiation between system and human summaries.

Pattern	Example
Noun appositive	One senior, Liz Parker, had slacked off too badly to graduate.
Gerundive clause	The Kialegees, numbering about 450, are a landless tribe, sharing space in Wetumka, Okla., with the much larger Creek Nation, to whom they are related.
Nonrestrictive relative clause	The return to whaling will be a sort of homecoming for the Makah, whose real name _ which cannot be written in English _ means "people who live by the rocks and the seagulls."
Intra-sentential attribution Lead adverbials and conjunctions	Separately, the report said that the murder rate by Indians in 1996 was 4 per 100,000, below the national average of 7.9 per 100,000, and less than the white rate of 4.9 per 100,000.

Table 2. Syntactic patterns for sentence simplification (underlined parts are removed)

their sentence selection component. Columbia University's summarization system uses a syntactic simplification component (Siddharthan et al., 2004), the results of which are sent to their sentence clustering component. Daumé and Marcu (2005a) reports that post-processing to delete adverbs and attributive phrases boosts ROUGE scores in the Multilingual Summarization Evaluation, though this post-processing was not found to be useful in DUC2005 (Daumé and Marcu, 2005b), possibly because the summary length is 250 words rather than 100 words.

In these approaches, simplification operations apply to all sentences equally, and the core sentence selection component has only either the original or the shortened sentence available to choose from. For this reason, simplification strategies have so far remained very conservative, probably to avoid possible oversimplification. However, this may not be an optimal approach, because the best simplification strategy is not necessarily the same for all sentences. For example, we might want to delete material *X* from a sentence only if *X* is already covered by another sentence in the summary; otherwise retain it.

An alternative approach to sentence simplification is to provide multiple shortened sentence candidates for the summarization engine to choose from. For example, Zajic et al. (2005)'s Multi-Document Trimmer (MDT) uses a syntactic trimmer (Dorr et al., 2003), which was initially developed for headline generation, to generate multiple trimmed versions of the sentences in the document cluster. Each of these trimmed candidates are given to a feature-based sentence selection component, which includes the redundancy score of the sentence given the current state of the summary and the number of trimming operations as features.

Our approach to sentence simplification is based on the same underlying idea as MDT's: we apply a small set of heuristics to a parse tree to create alternatives, after which both the original sentence and (possibly multiple) simplified versions are available for selection. Unlike MDT, original and alternative simplified sentences are provided for selection without differentiation in our system, i.e, without keeping any link between them. This is because we believe that a multi-document summarization engine is inherently equipped with the ability to handle redundancy, and the simplified alternatives only add to the redundancy. SumBasic's method for updating the unigram probabilities given the sentences already selected allows the sentence alternatives to be considered independently, while maintaining minimum redundancy.[2] Given that this approach to sentence simplification allows the sentence selection component to make the optimal decision among alternatives, we can now pursue more aggressive simplification, as the original non-simplified version is always available for selection. The approach is extensible beyond simplification operations to include novel sentence rewrites in a summary, as the candidate generation works independently of sentence selection, and only word probability is needed to compute the sentence score.

[2] Note, however, that the probability update by SumBasic must be computed based on the original document cluster so that it reflects the probability distribution of the words in the original document set.

Reason: parser error

LONDON _ British aviation authorities on Wednesday formally ruled the Concorde supersonic airliner unfit to fly unless its manufacturers took steps to prevent the problems that led to last month's fatal Air France Concorde crash near Paris.

Le Figaro newspaper on Wednesday quoted Gayssot, the transport minister, as raising the possibility that the ban on Air France Concorde flights could remain in place until the Accident and Inquiry Office releases a preliminary report on the crash at the end of August.

Reason: deleted material contains important information

PARIS _ **French investigators looking into the crash** last month of an Air France Concorde said Thursday it was probable that a 16-inch piece of metal found on the runway caused a tire to blow out, sending debris from the tire through fuel tanks and triggering a fire that brought down the plane.

The sleek, needle-nosed aircraft could cross the Atlantic at an altitude of 60,000 feet and at 1,350 mph, **completing the trip from London to New York in less than four hours** -- half the time of regular jets.

Table 3. Examples of full sentences chosen instead of their simplified counterpart. The underlined portion of text was deleted in the simplified sentence. Text corresponding to an SCU is indicated by boldface.

4.1 Syntax-Based Simplification Filter

Our simplification component consists of heuristic templates for the elimination of syntactic units based on parser output. Each sentence in the document cluster is first parsed using a broad-coverage English parser (Ringger et al., 2004). We then run a filter on the parse tree that eliminates certain nodes from the parse tree when the node matches the patterns provided heuristically. Table 2 lists the syntactic patterns we used for DUC2006 submission. These patterns are inspired by and similar to those discussed in Dunlavy et al. (2003), with the difference that we made use of a full-fledged parser for the extraction these patterns rather than employing a shallow parsing approach. For the first three patterns in Table 2 (noun appositive, gerundive clause and non-restrictive relative clause), the parser returns a node label corresponding exactly to these patterns; we simply deleted the nodes with these labels. For the identification of intra-sentential attribution, we added specific conditions for detecting the verbs of attribution (*said* in Table 2), its subject (*the report*), the complementizer (*that*) and any adverbial expressions if any, and deleted the nodes when conditions are matched. In the case of sentence-initial adverbials, we deleted only manner and time adverb expressions, using the features returned by the parser. Currently, we apply all these patterns simultaneously and create one simplified sentence per input; in principle, however, it is also possible to generate multiple simplified candidates.

The shortened sentences are then cleaned up in terms of punctuation and capitalization before being made available to the selection component, along with the original, non-simplified sentences in the document cluster.

Though our DUC2006 submission implements only a small number of simplification patterns described above, its effect was quite extensive. Of all the sentences selected in summary, 43.4% of them were the result of simplification. This resulted in adding on average another sentence to the summary: the average number of sentence in a summary increased from 11.32 to 12.52 when we used sentence simplification.

It is also interesting to note that 33.6% of the original non-simplified sentences (19% of all sentences) in summary had been selected even though a simplified counterpart was also available. Manual evaluation of one cluster (D0631D) established that four out of eleven sentences were non-simplified despite the availability of a simplified alternative. These four sentences are shown in Table 3. Of these, two were incorrectly parsed as noun appositives, resulting in an unexpectedly large portion of the text being deleted and rendering the simplified version less likely to be selected. The other two sentences present interesting cases where the deleted portion of text (indicated in boldface in Table 3) included important content, corresponding to Summary Content Units (SCUs) of weight 3 and 4, respectively, according to the Pyramid evaluation. These manual examinations suggest that the best simplification strategy may not be the same for all cases and that there is an

	With Simplification	No Simplification
R-1	0.2848	0.2815
R-2	0.0460	0.0457
R-SU4	0.0772	0.0771

Table 4. Effect of sentence simplification on ROUGE (recall, not using stopwords)

advantage in using the summarizer itself to choose the best sentence alternative given the context.

4.2 Effect of Sentence Simplification

In this section, we examine the effect of sentence simplification using various evaluation metrics. We must however keep in mind that it is difficult to isolate the effect of sentence simplification in the DUC2006 submission results.

ROUGE evaluation. Table 4 summarizes the results of experiments that attempt to isolate the effect of sentence simplification using ROUGE metrics on DUC2005 and 2006 data combined. We compared the performance of SumBasic with and without sentence simplification. The improvement by sentence simplification is statistically significant ($p < .05$) in ROUGE-1 according to the Wilcoxon Matched-Pairs Signed-Rank Test, but not significant in R-2 and R-SU4.

NIST evaluation. Three of the five NIST linguistic quality questions are relevant to sentence simplification: grammaticality, non-redundancy and referential clarity, along with the responsiveness question. Table 5 shows the scores of our system in DUC2006 relative to the peers, along with the rank of our system in 35 peer systems.

Sentence simplification undoubtedly contributes to our low score for grammaticality. The parser-based simplification filter produces some ungrammatical sentences and sentences with degraded readability, often due to misplaced punctuation marks. However, the grammaticality score may be swayed by whether a system allowed incomplete last sentences. Only 2 systems with incomplete last sentences scored higher on grammaticality than systems with complete last sentences.

Referential quality is also negatively affected by the current simplification filter, as the deletion of intra-sentential attribution can result in deleting the antecedent of pronouns in the summary. On the other hand, it is encouraging to note that our methods perform well on non-redundancy and content responsiveness. In particular, providing sentence alternatives did not increase redundancy, even though the alternatives were not explicitly linked, suggesting that the method of updating unigram weights given context used in SumFocus is robust enough to handle the greater redundancy introduced by providing simplified alternatives.

	MSR	Avg. peer	MSR rank
Grammaticality	3.12	3.58	31
Non-redundancy	4.42	4.23	10
Referential Clarity	2.64	3.11	31
Responsiveness	2.94	2.54	3

Table 5. NIST evaluation results

Pyramid evaluation. The goal of sentence simplification in summarization is to improve *content selection* by removing duplicate or irrelevant content from extractive summaries, and so the Pyramid evaluation is the most relevant evaluation for our purposes.

Our system ranked first of the 22 systems participated in the Pyramid evaluation with the overall mean Pyramid score of 0.257. Though it is difficult to isolate the effect of sentence simplification to our Pyramid performance, the component at least contributed to making room for one more sentence per summary on average. We note that the Pyramid scores in general remain very low for all peer systems: even in our system, the average mean score of 0.257 means that only 25% of the weighted SCUs attainable in the 250-word summary were actually included in the summary. Making room for more content and removing redundant material by simplifying sentences is therefore a promising operation for extractive summarization systems.

5 Lexical Expansion

We also explored the impact of augmenting the unigram probabilities of SumFocus with lexical expansions supplied by morphological variants and synonyms drawn from both hand-crafted thesauri and dynamically-learned sources. Expansions were applied only when choosing the “best words” drawn only from the topic statements for sentence selection. A uniform default lambda of 0.5 was determined by hand inspection of R-2 results tuned on the DUC 2005 data, and applied to the cumulative scores of matches of all expansion types, using the following formula, where d is the document score, and e is the score for each expansion type:

		No Simplification	With Simplification	Significance
R-1 Recall	SumBasic	0.30599	0.30886	
	SumFocus w/o Expansion	0.30153	0.29641	
	SumFocus + Expansion	0.30092	**0.29891**	p <= 0.0494
R-2 Recall	SumBasic	0.05453	0.05517	
	SumFocus w/o Expansion	0.06053	0.05711	
	SumFocus + Expansion	0.06064	**0.05935**	p <= 0.0029
R-SU4 Recall	SumBasic	0.08760	0.08719	
	SumFocus w/o Expansion	0.09224	0.08841	
	SumFocus + Expansion	0.09185	**0.08986**	p <= 0.0067

Table 6. Lexical expansion with and without sentence simplification (DUC2006 data, not using stopwords)

$$WordScore = (1-\lambda) * d + \lambda \sum_{i}^{n} e_i$$

Table 6 shows selected ROUGE recall scores. The SumFocus + Expansion rows reflect the composite application of expansions (and correspond to MRF+GEO+ENC+WA in Table 7). Without sentence simplification, the scores are statistically indistinguishable from unexpanded SumFocus. When sentence simplification is employed, however, lexical expansions appear to offset the small degradation introduced by combining SumFocus with simplification. Although these improvements are small, and are not significant at the 95% confidence level reported by the ROUGE tool, p-scores computed using the Wilcoxon Matched-Pairs Signed-Ranks Test suggest that the overall improvement over the SumBasic baseline can be interpreted as significant. Table 7 presents the individual and combined contributions of different expansion components in conjunction with sentence simplification as measured by R-2.

5.1 Morphological Variants

Morphologically-derived forms (MRF) were looked up in the version of the American Heritage Dictionary used by our parser, thereby allowing us to obtain pairs such as “develop” ⬄ “development.” We also used a small inventory of paired geographical names (GEO) and their adjective counterparts (e.g., “United States” ⬄ “American”, “Tanzania” ⬄ “Tanzanian”). Simple counts of the number of times the forms were encountered: if both forms occurred in the topic text, both received a boost. Since topic words were not lemmatized, only exact matches were considered, with the result that few instances of morphological derivation were reflected in the final selection. Inclusion of geographical variants did yield a statistically significant boost as measured by the Wilcoxon test.

5.2 Learned Lexical Pairs

One of our primary objectives in utilizing the lexical expansions was to investigate the potential impact of a 65,335-pair synonym list (shown as WA in Table 7) that we automatically acquired from clustered news articles available on the World Wide Web, our hypothesis being that a thesaurus derived from news data might prove more useful than static general-domain synonym resources. Starting with an initial dataset of 9.5 million sentences in ~32,400 clusters, we created a monolingual bitext of 282,583 aligned sentence pairs using a simple word-based edit distance heuristic. A Support Vector Machine trained on 10,000 tagged sentence pairs was also used to extract an additional ~20,000 sentences.

Brockett and Dolan (2005) provide further detail concerning the methods applied to extract this data. Identical words were deleted from the sentence pairs and the remainder matched using a log likelihood-ratio-based Word Association technique (Moore, 2001), using formula given in Moore (2004), modified here for readability:

$$LLR(t,s) = \sum_{t \in \{1,0\}} \sum_{s \in \{1,0\}} C(t,s) \log \frac{p(t \mid s)}{p(t)}$$

where t and s are variables ranging over the presence (1) or absence (0) of the words under consideration, and $C(t,s)$ is the observed joint count for their values. The probabilities are maximum likelihood estimates.

The extracted word pairs were filtered to remove typographical errors, numerical mismatches and other artifacts of unnormalized news data, and

System	R-2
Baseline (SumFocus w/o expansion)	0.05711
+WN	0.05699
+WA	0.05736
+MRF+GEO+WN	0.05859
+MRF+GEO+ENC+WA+WN	0.05885
+MRF+GEO	0.05918
+MRF+GEO+ENC	0.05918
+MRF+GEO+ENC+WA	**0.05935**
+MRF+GEO+WA	**0.05935**

Table 7. Contributions of lexical expansion components (DUC2006 data, not using stopwords)

scores were converted log-linearly into pseudo-probabilities ranging between 1.0 and 0.0. In Table 7, the contribution of the Word Association pairs appears slightly positive, although the gains are not statistically significant. Further research will be needed to determine how such dynamically-acquired, potentially domain-relevant resources can be best utilized in summarization.

5.3 Static Thesauri

Table 7 also shows the impact of expansions taken from two static thesauri. In the submitted system, we deployed a list of 125,054 word pairs found in the *Encarta Thesaurus* (ENC) (Rooney, 2001) for which heuristic weights had been precomputed. This thesaurus had no measurable impact. The table additionally shows the effect of simple synonym expansion (raw occurrence counts in conjunction with the uniform lambda) using WordNet 2.0 (Fellbaum, 1998), which we did not include in the system submitted to DUC2006. WordNet generally appears to degrade performance, though not significantly. Better motivated similarity weights might achieve different results.

6 Observations on Pyramid and NIST

In the sections above, we presented various system settings using the ROUGE metrics for comparison. Results obtained from human evaluations are potentially more diagnostic and can inform future directions. Unfortunately, neither NIST nor Pyramid, being one-time human evaluations, lend themselves to comparing system settings. Nevertheless, since both measure content directly, these are the metrics we focus on, given our primary goal of maximizing summary content. Accordingly, we report NIST and Pyramid metrics only for the system we submitted, which is SumFocus with sentence simplification and with lexical expansion of the topic words.

Cluster	Rank acc. to Pyramid score	MSR SCUs / Average SCUs attainable	NIST content (5=very good)
D0601	1	0.3056	1
D0603	2	0.1818	3
D0605	6	0.0976	2
D0608	1	0.3200	3
D0614	10	0.1739	5
D0615	1	0.2500	1
D0616	3	0.3043	4
D0617	1	0.3409	2
D0620	2	0.3103	3
D0624	1	0.5000	3
D0627	15	0.1429	1
D0628	15	0.1613	1
D0629	21	0.0435	2
D0630	13	0.1935	3
D0631	1	0.5625	4
D0640	12	0.2564	2
D0643	6	0.3077	2
D0645	8	0.1786	3
D0647	6	0.1600	2
D0650	3	0.2750	2

Table 8. Pyramid and NIST results for system 10, including MSR's rank per cluster according to Pyramid evaluation, average number of unweighted SCUs out of the average number of SCUs attainable for each cluster, and the NIST content responsiveness score, which is on a scale of 5 to 1

Our system ranked first in the overall mean Pyramid score of the 22 systems that participated in the Pyramid evaluation. It must be noted, however, that the maximum Pyramid score for each cluster differs (Nenkova and Passonneau, 2004), so the average rank is a better method to compare systems, presented in Table 8. Our system is ranked first for 5 out of 20 clusters, and is in the top 3 for half of the clusters. Our per-cluster mean ranking (5.90) is the best among the 22 system. though our performance across the clusters, shown in Table 8, is not evenly distributed.

Our system ranked third out of 35 systems in NIST content responsiveness. On a per-cluster basis, providing a system ranking according the NIST content score is less than instructive since there are only 5 values that can be given. Initial investigation showed there is no correlation between Pyramid score and NIST content score, but we have reached no conclusion yet. It is certainly unexpected that for cluster D0601, MSR is ranked highest among the peers in the Pyramid evaluation, with a relatively high degree of SCU overlap, while receiving a poor NIST content score for the same cluster.

7 Future Work

The Pyramid annotation shows us that that only rarely does a peer summary match 50 percent or more of the content in the combined model summaries. A detailed analysis of the percentage of SCUs per weight must still be done, but anecdotally, our system matches only half of the high-scoring SCUs, i.e., those SCUs that were found in all of the model summaries. Clearly, finding methods to model this data more closely offers opportunities for improving the overall content of summaries. One direction will lead us to find more sophisticated methods of computing the relation between a topic word or any of its lexical expansions, the document words, and the target summaries. And, as we continue to pursue extractive summarization, we will also expand our system to take full advantage of sentence simplification component. In particular, we plan to include more drastic simplifying and rewrite operations (such as splitting coordinated clauses) and produce multiple candidates per sentence in order to explore the full potential for the proposed approach.

Acknowledgments

We are indebted to Ani Nenkova, who conceived of SumBasic during her internship at Microsoft Research, and to Arul Menezes for implementing several features described here. We also wish to thank the NIST team that supports DUC for the NIST evaluation, and the DUC community who contributed to the Pyramid evaluation. We also thank members of the Butler-Hill Group, especially Ben Gelbart, for their assistance with the Pyramid creation and annotation.

References

Brockett, C. and W. B. Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of The Third International Workshop on Paraphrasing (IWP2005*) .

Conroy, J.M., J. Schlesinger and J. Goldstein Stewart. 2005. CLASSY Query-Based Multi-Document Summarization. In *Proceedings of DUC 2005*.

Daumé III, H., and D. Marcu. 2005a. Bayesian Multi-Document Summarization at MSE. In *Proceedings of MSE 2005*.

Daumé III, H., and D. Marcu. 2005b. Bayesian Summarization at DUC and a Suggestion for Extrinsic Evaluation. In *Proceedings of DUC 2005*.

Dorr, B., D. Zajic and R. Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of HLT-NAACL 2003 Text Summarization Workshop*, pp.1-8.

Dunlavy, D., J. Conroy, J. Schlesinger, S. Goodman, M. Okurowski, E., O'Leary and H. van Halteren. 2003. Performance of a Three-Stage System for Multi-Document Summarization. In *Proceedings of DUC 2003*.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Moore, R. C. 2001. Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships among Words. In *Proceedings*, *Workshop on Data-driven Machine Translation*.

Moore, R. C. 2004. Association-Based Bilingual Word Alignment. In *Proceedings, Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, Ann Arbor, Michigan.

Nenkova, A. and L. Vanderwende. 2005. *The impact of frequency on summarization*. MSR-TR-2005-101.

Nenkova, A., L. Vanderwende, and K. McKeown. 2006. A Compositional Context Sensitive Multi-document Summarizer. In *Proceedings of SIGIR 2006*.

Nenkova, A. and R. Passonneau. 2004. Evaluating content selection in summarization: The Pyramid method. In *Proceedings of the HLT-NAACL 2004.*

Ringger, E., R.C. Moore, E. Charniak, L. Vanderwende and H. Suzuki. 2004. Using the Penn Treebank to Evaluate Non-Treebank Parsers. In *Proceedings of LREC 2004*.

Rooney K. (ed.) 2001. *Encarta Thesaurus*. Bloomsbury Publishing.

Siddharthan, A., A. Nenkova and K. McKeown. 2004. Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. In *Proceedings of COLING 2004*.

Zajic, D., B. Dorr, J. Lin, C. Monz, and R. Schwartz. 2005. A Sentence-Trimming Approach to Multi-Document Summarization. In *Proceedings of DUC2005*.