Discourse-Based Summarization in DUC-2001

Daniel Marcu
Information Sciences Institute and
Department of Computer Science
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
marcu@isi.edu

1 Introduction

We focus in this paper on the following DUC-2001 related matters:

- Presenting the algorithm we used in order to summarize single documents.
- Presenting the algorithm we used in order to summarize collections of documents.
- Discussing perceived problems with the evaluation methodology that was employed by the NIST assessors.

2 Summarizing single documents

The single document summarization system that we used in DUC-2001 employed the following steps.

1. Derive the discourse structure of the text given as input

The discourse structure was derived using a version of the cue-phrase-based discourse parser described by Marcu [2000].

2. Determine the important sentences in the input document

The automatically produced discourse structure was used in order to determine the set of most important sentences in the input document that would yield a summary of at most 150 words. The important sentences were extracted from the discourse structure using the algorithm described by Marcu [2000].
At the end of this step, we created a list with all sentences in the input document, one sentence per line. The sentences that were considered important were marked.

3. Determine all co-reference links in the input document

We used CONTEX [Hermjakob, 2000], a syntactic parser developed in the context of the Webclopedia project, in order to determine for each sentence in a document the list of noun constructs used in that sentence. Each noun construct was paired with features such as gender, number, etc.
We implemented from scratch a co-reference resolution system that was used in order to resolve all third person pronouns in the input text. Each pronoun was assumed to co-refer with the closest preceding noun of the same gender and number.

4. Increase summary coherence and compactness

To increase the coherence and compactness of the summary, we modified the pool of important sentences by adding to, deleting from, and rewriting sentences in the pool. The following procedures were used, in the sequence presented below:

- Add sentences to the pool so as to avoid dangling discourse relations. For example, if a sentence in the pool of important sentences started with “Afterwards” or “But”, the preceding sentence was marked as important as well and added to the pool of important sentences.
- Remove from the pool of important sentences the sentences with less than five words.
- Remove the questions from the pool.
- Remove the quotes from the pool.
- Remove from the pool the sentences that contained only capitalized words.
- Remove from the pool the titles and subtitles.
- Remove the dates from the pool.
- Rewrite sentences by deleting overtly marked parenthetical units,

such as those delimited by long dashes.

At the end of this step, the pool of important sentences was no longer 150 words long. In some cases the corresponding summary was longer, in other cases it was shorter.

5. Generate summary

In the last step of the algorithm, we generated the summary by

- Printing first the title of the original document.
- Printing sentences from the pool of important sentences in the order of their occurrence in the text. Each third person pronoun that referred to an entity that was not mentioned already in the summary was replaced with the complete referring expression computed during step 3. The generation process stopped after printing approximately 100 words.

3 Summarizing document collections

The input for our multi document summarization system is a set of 100 word summaries (with no titles), which are produced by the single document summarizer described in the previous section. The summarization process follows these steps:

1. Pre-process the collection

During this step, we perform the following tasks:

- We compute the similarity between every pair of documents in a collection and between every sentence pair in all single document summaries in a collection.
- For each document and each single document summary sentence, we compute their average similarity scores. These average scores are used to determine the most important sentences in the collection. Following a method proposed by Hoey [1991] and Salton et al. [1994], we assume that documents that have high average similarity scores are more central to the collection (and hence, more important) than those that have low scores. By the same token, we assume that sentences that have high average similarity scores are more important than sentences that have low average similarity scores.
- For each single document summary sentence, we associate a date stamp, using the algorithm described by Filatova and Hovy [2001].

2. Select and order the sentences that summarize the collection

We estimate the goodness of a multi-document summary using the following heuristics.

Sentence-pair-specific:

- **Local order:** We assume that multi-document summaries that present sentences in the order they occurred in the individual documents are better than summaries that violate the original ordering.
- **No repetition:** We assume that multi-document summaries that do not contain the copy of the same sentence multiple times are better than summaries that contain such repeated occurrences.
- **Local date ordering:** We assume that a multi-document summary that reproduces a pair of sentences in the chronological order of the events described in those sentences is better than a summary that uses the reverse order.
- **Global document importance:** We assume that a multi-document summary that presents sentences from documents with high average similarity scores before sentences from documents with low average similarity scores is better than a summary that employs the reverse order.
- **Global sentence importance:** We assume that a multi-document summary that presents sentences with high average similarity scores before sentences with low average similarity scores are better than a summary that employs the reverse order.

- **Low redundancy:** We assume that non-redundant summaries are better than summaries that contain redundant information. (The redundancy score of a summary is computed by summing up the term $1 - similarity(s_i,s_{i+1})$ for each sentence pair (s_i,s_{i+1}) in the summary.)
- **Local cue-phrase-based coherence:** We assume that summaries that do not contain dangling discourse relations are better than summaries that contain such relations.

Sentence-specific:

- **Sentence length:** We assume that summaries that contain long sentences are better than summaries that contain short sentences. This is consistent with the findings reported by Marcu and Gerber [2001].
- **Average position in a single document:** We assume that summaries that contain sentences that occur in the beginning of single documents are better than summaries that contain sentences that occur towards the end of single documents. This is consistent with the findings reported by Hovy and Lin [1999].
- **Global date ordering:** We assume that summaries that contain sentences with recent date stamps are better than summaries that contain sentences with less recent date stamps.

Document-specific:

- **Coverage:** We assume that summaries that contain sentences from many documents are better than summaries that contain sentences from fewer documents.

For each of the heuristics above, we have implemented a scoring function that yields for a given summary a score between 0 and 1. The score of a summary is computed as a weighted sum of the scores corresponding to all heuristics.

In order to build multi-document summaries of arbitrary length, we start with the pool of sentences selected by the single document summarization system for each individual document in a collection and create a list of one-sentence-long "active" summaries. Initially, the list contains *n* summaries, one for each sentence selected as important by the single document summarization system. We iterate over all possible summaries of length two that can be created by appending one sentence to a summary from the list of "active" summaries. We keep only the top 100 summaries of highest score. We then create all possible summaries of length three that can be created by appending one sentence to a summary of length two. As before, we keep only the top 100 summaries of highest score. We continue this process until we create summaries that contain more than 400 words. The search procedure that corresponds to this step is the most computationally the most expensive one. For each document collection, the selection and ordering step takes a couple of hours of computation.

3. Resolve third person pronouns.

We resolve each third person pronoun to the noun/entity determined during the single document summarization process. If an noun/entity was used already in a multi-document summary, the pronoun is not replaced by the corresponding entity.

4. Rank headlines.

We rank the headlines of all documents according to the average similarity scores of the documents. Headlines of documents with high average similarity scores are considered more important than headlines of summaries with low average similarity scores.

5. Generate summaries.

We generate summaries according to the following rules:

- 50- and 100-word long multi-document summaries consist only of the top headlines, ranked according to the importance of the documents they correspond to and preceded by the phrase *"The k most_important_ headlines:"*.
- 200- and 400-word long multi-document summaries are divided into two parts. The first 100 words consist of the top headlines, ranked according to the importance of the documents they correspond to. The remaining 100 (300 words) are given by the multi-document summaries produced in

steps 1 to 3 that are closest in length to this threshold.

As an example, we show in Figure 1 the 200-word long multi-document summary that was generated automatically by our system for document collection d32f.

<multi size="200" docset="d32f" >
The 8 most_important_headlines:
-ALASKA TANKER PILOTED BY UNQUALIFIED OFFICE. EXXON UNABLE TO EXPLAIN CAPTAIN'S ABSENC. RISING WINDS STIR FEARS OF OIL SLICK DAMAGE
-EXXON SUBMITS STRATEGY ON ALASKA CLEANUP PLAN
-TANKER SPILLS OIL AFTER HITTING REEF OFF ALASKA
-FRESH OIL SHEEN SEEPS FROM EXXON VALDEZ
-WORKERS TRY TO UNLOAD TANKE. ENVIRONMENTALISTS CALL SPILL A DISASTER
-CHEMICALS FAIL TO BREAK UP LARGEST SPILL IN U.S. HISTORY
-CAPTAIN SHOULD HAVE BEEN PILOTING TANKER, EXXON REVEAL. DISASTER DECLARED
-EXXON RAISES VALDEZ CLEANUP COSTS TO $2 BILLIO. EARNINGS: THE OIL GIANT WILL TAKE ANOTHER $500-MILLION CHARGE OVER THE SPILL, BRINGING ITS TAB FOR THE YEAR TO $1.38 BILLION.

A Long Beach-bound Exxon oil tanker ran aground on a reef Friday and spilled an estimated 8.4 million gallons of crude oil into Alaska's Prince William Sound, a pristine Pacific waterway heavily used by kayakers, fishermen and tourists.
Exxon Corp. on Wednesday increased its estimate of the total 1989 costs of cleaning up the massive Alaskan oil spill to $2 billion and said it would take another $500-million charge in the fourth quarter to cover costs from what is now the most expensive environmental disaster in history.
</multi>

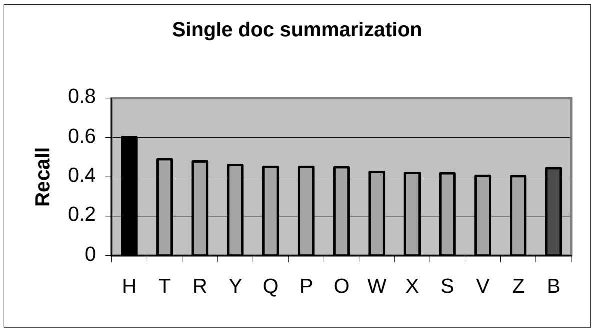
Figure 1 : Multidoc summary example.

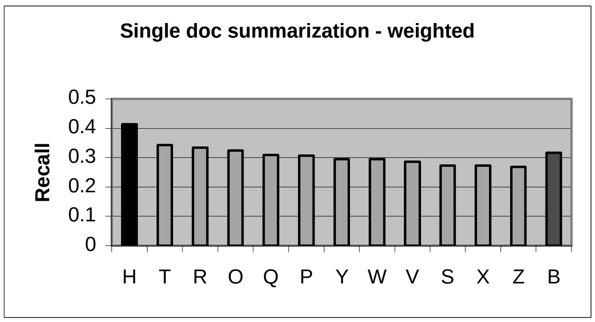
4 Problems with the DUC-2001 evaluation

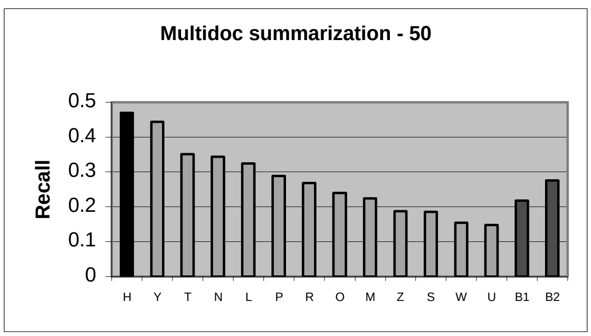
We have used the judgments produced by the NIST analysts in order to evaluate internally the performance of all systems. In the process, we became aware of some problems that we believe have hampered the DUC evaluation enterprise. Instead of presenting evaluation results similar to those produced by NIST for all participants in the evaluation, we are going to focus in the rest of this paper on enumerating the negative aspects that pertain to the DUC evaluation. Addressing these problems may yield different results.

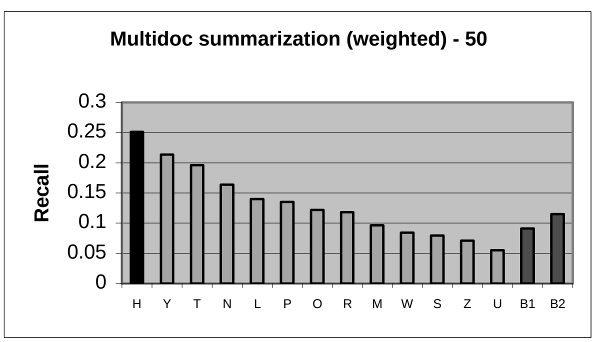
4.1 Different recall metrics yield different rankings

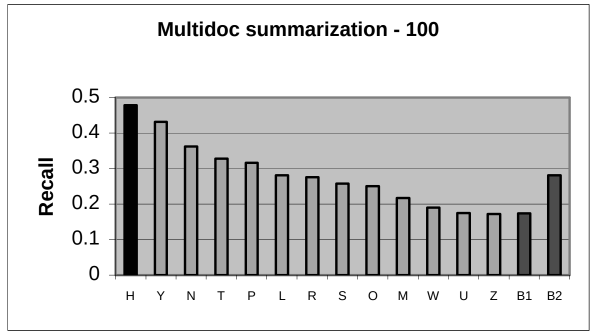
We estimated the ability of summarization systems to identify important information in single and multiple documents using two recall metrics. One metric estimates the recall by dividing the number of units marked with peer units by the number of units in the model summary. The other recall is weighted, i.e., it gives high credit to the unit pairs in the model and peer summaries that were judged to have a high degree of overlap (a score of 4 in the evaluation schema employed by DUC-2001) and low credit to the units that have a low degree of overlap (a score of 1 in the evaluation schema employed by DUC-2001). The chart pairs presented below that depict evaluation results across all compression rates and all documents yield a fairly consistent ranking among summarizers for summaries up to 100 words long. However, for summaries that are 200 and 400-word long, depending on the recall metric one chooses, one obtains quite different rankings of the participating systems. (In all charts, H corresponds to the average human performance level, B, B1, and B2 to the baselines, and the other letters to the participating systems.)

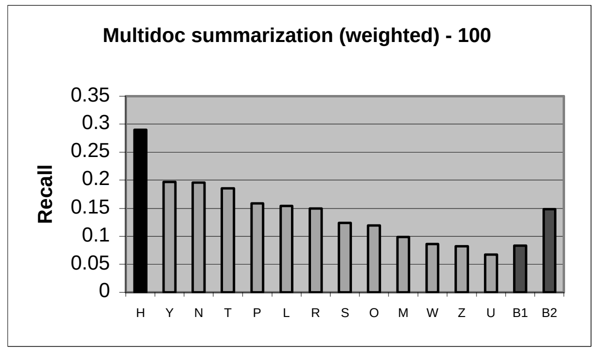

Single doc summarization
Recall
0.8
0.6
0.4
0.2
0
H T R Y Q P O W X S V Z B

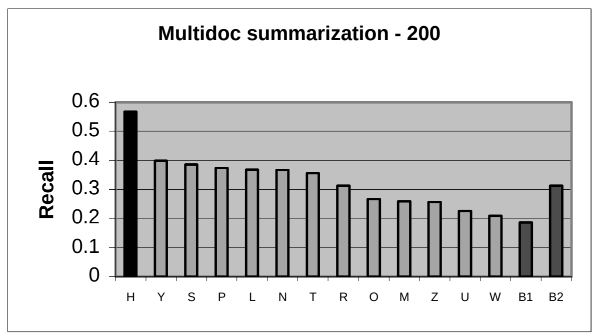

Single doc summarization - weighted
Recall
0.5
0.4
0.3
0.2
0.1
0
H T R O Q P Y W V S X Z B

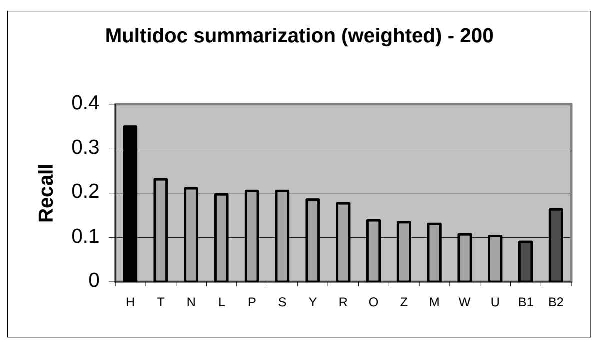

Multidoc summarization - 50
Recall
0.5
0.4
0.3
0.2
0.1
0
H Y T N L P R O M Z S W U B1 B2

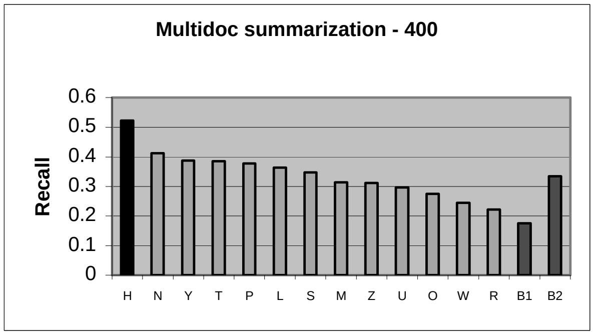

Multidoc summarization (weighted) - 50
Recall
0.3
0.25
0.2
0.15
0.1
0.05
0
H Y T N L P O R M W S Z U B1 B2

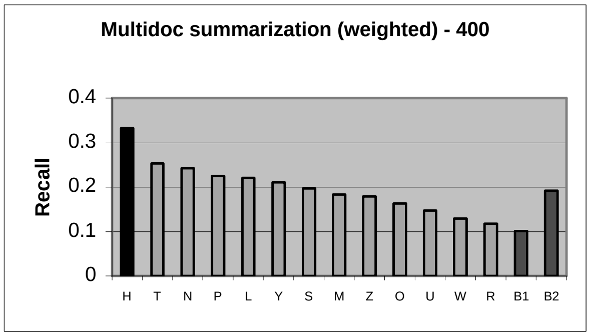

Multidoc summarization - 100
Recall
0.5
0.4
0.3
0.2
0.1
0
H Y N T P L R S O M W U Z B1 B2


Multidoc summarization (weighted) - 100
Recall
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
H Y N T P L R S O M W Z U B1 B2


Multidoc summarization - 200
Recall
0.6
0.5
0.4
0.3
0.2
0.1
0
H Y S P L N T R O M Z U W B1 B2


Multidoc summarization (weighted) - 200
Recall
0.4
0.3
0.2
0.1
0
H T N L P S Y R O Z M W U B1 B2





4.2 Recall and precision metrics do not account for units that are important but are not in the model summary

As part of the evaluation process, judges were asked to determine which units in an automatically produced summary were important, but nevertheless not present in the model summary. The DUC systems often include in the their outputs such units, which under a correct evaluation schema would have to be accounted for in the computation of the recall and precision figures. Unfortunately, none of the two recall schemas discussed above accounts for these units. This is quite unfortunate as it penalizes especially the systems that find large amounts of important information that is not present in the model summary.

For example, our single document summarization system found more such information than any other system. Its average score that reflects this important information that is not accounted for by the model summary was 0.429; in contrast, the average score for all other systems was 0.29. Although our system found more unaccounted for important information than any other system, its ability to find important information is not reflected by traditional recall and precision metrics.

4.3 Precision figures are misleading

We have come to believe that precision is not a useful metric in the DUC evaluation, as all systems produced summaries of bounded lengths. To explain why this is the case, consider the model summary below and two peer summaries of approximately equal length, which convey the same information.

Model summary:

[Officials at Southern Co. conspired to cover up their accounting for spare parts.]

Peer summaries:

A. [A grand jury has been investigating whether officials at Southern Co. conspired to cover up their accounting for spare parts to evade federal income taxes.]
B. [Allegedly, in order to evade federal income taxes,] [officials at Southern Co. conspired to cover up their accounting for spare parts.] [A grand jury has been investigating this.]

Both peer summaries reproduce the important information in the model summary and some additional information. Yet, peer summary A consists of 1 unit, while peer summary B of 3 units. Computing precision at the unit level would yield a precision of 1.00 for peer summary A and a precision of 1/3= 0.33 for peer summary B. This is counterintuitive, as both peer summaries are approximately equal in length and have the same semantic content.

We believe that systems that include shorter sentences in the summaries or long sentences with clearly marked clauses are put at a disadvantage when any unit-based precision metric is employed. For example, our system, system Y, produced a total of 190 units in all multidoc summaries of length 50. By contrast, the average number of units produced by the other systems was 67. Although our system produced summaries of the same length with the other system, it is systematically penalized by the precision metric for producing more

units. As a consequence, precision figures for our system are systematically lower.
As all systems produce summaries of bounded length, we believe that precision figures are irrelevant in the context of the DUC evaluation.

4.4 Grammaticality, cohesion, and organization judgments look suspicious.

We found the grammaticality, cohesion, and organization judgments highly suspicious. For example, for the single document baseline, the average scores across all judgments were 3.19, 2.88, and 3.04 respectively. Given that these baselines were created by taking the first 100 words in a document, it is very likely that they were both grammatical, cohesive, and coherent. The multidocument summaries produced by humans faired better with respect to their grammaticality but exhibited the same levels of performance when it came to cohesion and coherence. The fact that these summaries received so low scores is disturbing. We believe that in order to make these results reliable, future evaluations will need to be carried out only after employing extensive training with the NIST assessors in order to ensure higher consistency with respect to these judgments.

4.5 Grammaticality, cohesion and coherence in non-naratives.

The notions of grammaticality, cohesion, and coherence mean different things in different textual contexts. The grammar of headlines is different from the grammar of texts. List environments are cohesive and coherent in a different way than narrative texts are. For example, text A is ungrammatical, incohesive, and incoherent as a narrative, but grammatical, cohesive and coherent when presented as in B.

A. Biking on the seashore. Hiking in the mountains. Playing bridge with my friends. Dancing.

B. The things I like most are:
   - Biking on the seashore.
   - Hiking in the mountains.
   - Playing bridge with my friends.
   - Dancing.

From the grammaticality, cohesion, and coherence scores assigned to the output produced by our system, it appears that the NIST judges decided to employ criteria for narratives to non-narrative texts. For example, the summary in Figure 1 was assigned by the NIST assessors a grammaticality score of 2, and cohesion and coherence scores of 1!

4.6 Formatting

The pre-processing of the summaries in order to enable their evaluation in the SEE interface puts at a disadvantage the systems that employ textual formatting devices. For example, our system presented the headlines in upper case, as a bullet list, one headline per line. And the rest of the summary as normal narrative (see Figure 1). However, NIST assessors saw the summary as shown in Figure 2. Evaluating a non-formatted summary can decrease the chance that human assessors treat the list environments differently and apply different grammaticality, cohesion, and coherence judgments as they move from one type of environment to another.

4.7 Stability and reliability of the evaluation schema

We believe the most important weakness of the evaluation schema employed by NIST concerns the lack of evaluation of the evaluation protocol. The current results do not seem to enable one determine

- whether one human judge makes consistent judgments when assessing the performance of the same summarization system at different moments in time. (This amounts to assessing the stability of the evaluation schema).
- whether two or more human judges agree on their assessments. (This amounts to assessing the reliability of the evaluation schema).

Unless the evaluation schema employed by NIST is both stable and reliable, no conclusions can be derived in conjunction with DUC-2001.

<multi size="200" docset="d32f" >
The 8 most_important_headlines: -ALASKA TANKER PILOTED BY UNQUALIFIED OFFICE. EXXON UNABLE TO EXPLAIN C APTAIN'S ABSENC. RISING WINDS STIR FEARS OF OIL SLICK DAMAGE -EXXON SUBMITS STRATEGY ON ALASKA CLEANUP PLAN -TANKER SPILLS OIL AFTER HITTING REEF OFF ALASKA -FRESH OIL SHEEN SEEPS FROM EXXON VALDEZ -WORKERS TRY TO UNLOAD TANKE. ENVIRONMENTALISTS CALL SPILL A DISASTER -CHEMICALS FAIL TO BREAK UP LARGEST SPILL IN U.S. HISTORY -CAPTAIN SHOULD HAVE BEEN PILOTING TANKER, EXXON REVEAL. DISASTER DECLARED -EXXON RAISES VALDEZ CLEANUP COSTS TO $2 BILLIO. EARNINGS: THE OIL GIANT WILL TAKE ANOTHER $500 -MILLION CHARGE OVER THE SPILL, BRINGING ITS TAB FOR THE YEAR TO $1.38 BILLION. A Long Beach -bound Exxon oil tanker ran aground on a reef Friday and spilled an estimated 8.4 million gallons of crude oil into Alaska's Prince William Sound, a pristine Pacific waterway heavily used by kayakers, fishermen and tourists. Exxon Corp. on Wednesday increased its estimate of the total 1989 costs of cleaning up the massive Alaskan oil spill to $2 billion and said it would take another $500-million charge in the fourth quarter to cover costs from what is now the most expensive environmental disaster in history.
</multi>

Figure 2 : Mul tidoc summary example with no formatting.

5 References

[Filatova and Hovy, 2001]. Filatova Elena and Eduard Hovy. Assigning Time -Stamps to Event Clauses. *Proceedings of the ACL -2001 Workshop on Spatial and Temporal Reasoning.* Toulouse, France. July 2001.

[Hermjakob, 2000]. Ulf Hermjakob. Rapid Parser Development : A Machine Learning Approach to Korean. *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2000).* Seattle, WA., May 2000.

[Hoey, 1991]. Michael Hoey. *Patterns of Lexis in Text.* Oxford University Press. 1991.

[Hovy and Lin, 1999]. Eduard Hovy and Chin-Yew Lin. Automated Text Summarization in SUMMARIST. In Advances in Automatic Text Summarization, Inderjeet Mani and Mark Maybury editors, pp. 81-94. The MIT Press, 2000.

[Marcu, 2000]. Marcu Daniel. *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press. November 2000.

[Marcu and Gerber, 2001]. Marcu Daniel and Laurie Gerber. An Inquiry into the Nature of Multidocument Abstract, Extracts, and Their Evaluation. Proceedings of the NAACL'01 Workshop on Text Summarization. Pittburgh, PA, 2001.

[Salton et al., 1994] Salton Gerard, Chris Buckley and Amit Singhal. Automatic Analysis. Theme Generation and Summarization of Machine -Readable Texts. Science (264), pp. 1421-1426, 1994.