Challenges in Evaluating Summaries of Short Stories

Anna Kazantseva
School of Information Technology
and Engineering,
University of Ottawa, Ottawa, Canada
ankazant@site.uottawa.ca

Stan Szpakowicz
School of Information Technology
and Engineering,
University of Ottawa, Ottawa, Canada
Institute of Computer Science,
Polish Academy of Sciences, Warsaw, Poland
szpak@site.uottawa.ca

Abstract

This paper presents experiments with the evaluation of automatically produced summaries of literary short stories. The summaries are tailored to a particular purpose of helping a reader decide whether she wants to read the story. The evaluation procedure includes extrinsic and intrinsic measures, as well as subjective and factual judgments about the summaries pronounced by human subjects. The experiments confirm the experience of summarizing more conventional genres: sentence overlap between human- and machine-made summaries is not a complete picture of the quality of a summary. In fact, in our case, sentence overlap does not correlate well with human judgment. We explain the evaluation procedures and discuss several challenges of evaluating summaries of works of fiction.

1 Introduction

In recent years the automatic text summarization community has increased its focus on reliable evaluation. The much used evaluation methods based on sentence overlap with reference summaries have been called into question (Mani 2001) as they provide only a rough approximation of semantic similarity between summaries. A number of deeper, more semantically-motivated approaches have been proposed, such as the factoid method (van Halteren and Teufel, 2003) and the pyramid method (Nenkova and Passonneau 2004). These methods measure similarity between reference and generated summaries more reliably but, unfortunately, have a disadvantage of being very labour-intensive.

This paper describes experiments in evaluating automatically produced summaries of literary short stories. It presents an approach that evaluates summaries from two different perspectives: comparing computer-made summaries to those produced by humans based on sentence-overlap and measuring usefulness and informativeness of the summaries by themselves – a step critical when creating and evaluating summaries of a relatively unexplored genre. The paper also points out several challenges specific to evaluating summaries of fiction such as questionable suitability of traditional metrics (those based on sentence overlap), unavailability of clearly defined criteria to judge "goodness" of a summary and a higher degree of redundancy in such texts.

We achieve these goals by performing a two-step evaluation of our summaries. Initially, for each story in the test set we compare sentence overlap between summaries which the system generates and those produced by three human subjects. These experiments reveal that inter-rater agreement measures tend to be pessimistic where fiction is concerned. This seems due to a higher degree of redundancy and paraphrasing in such texts. The second stage of the evaluation process seeks to measure usefulness of the summaries in a more tangible way. To this end, three subjects answered a number of questions, first after

Proceedings of the Workshop on Task-Focused Summarization and Question Answering, pages 8–15,
Sydney, July 2006. ©2006 Association for Computational Linguistics

Figure 1. Example of a summary produced by the system.

A MATTER OF MEAN ELEVATION. By O. Henry (1862-1910).

On the camino real along the beach the two saddle mules and the four pack mules of Don Señor Johnny Armstrong stood, patiently awaiting the crack of the whip of the arriero, Luis. These articles Don Johnny traded to the interior Indians for the gold dust that they washed from the Andean streams and stored in quills and bags against his coming. It was a profitable business, and Señor Armstrong expected soon to be able to purchase the coffee plantation that he coveted. Armstrong stood on the narrow sidewalk, exchanging garbled Spanish with old Peralto, the rich native merchant who had just charged him four prices for half a gross of pot-metal hatchets, and abridged English with Rucker, the little German who was Consul for the United States. […] Armstrong, waved a good-bye and took his place at the tail of the procession. Armstrong concurred, and they turned again upward toward Tacuzama. […] Peering cautiously inside, he saw, within three feet of him, a woman of marvellous, imposing beauty, clothed in a splendid loose robe of leopard skins. The hut was packed close to the small space in which she stood with the squatting figures of Indians. […] I am an American. If you need assistance tell me how I can render it. […] The woman was worthy of his boldness. Only by a sudden flush of her pale cheek did she acknowledge understanding of his words. […] " I am held a prisoner by these Indians. God knows I need help. […] look, Mr. Armstrong, there is the sea!

reading only the summary and then after reading the complete story. The set included both factual questions (e.g. *can you tell where this story takes place?*) and subjective questions (e.g. *how readable did you find this summary?*).

Finally, we compare the two types of results with a surprising discovery: overlap-based measures and human judgment do not correlate well in our case.

This paper is organized in the following manner. **Section 2** briefly describes our summarizer of short stories. **Section 3.1** discusses experiments comparing generated summaries to reference ones based on sentence overlap. The experiments involving human judgment of the summaries are presented in **Section 3.2** and the two types of experiments are compared in **Section 3.3**. **Section 4** draws conclusions and outlines possible directions for future work.

2 Background: System Description

A detailed description of our summarizer of short stories is outside the scope of this paper. For completeness, this section gives an overview of the system's inner workings. An interested reader is referred to our previous work (Kazantseva 2006) for more information.

The system is designed to create a particular type of indicative generic summaries – namely, summaries that would help readers decide whether they would like to read a given story. Because of this, a summary, as defined here, is not meant to summarize the plot of a story. It is intended to raise adequate expectations and to enable a reader to make informed decisions based on a summary only. We achieve this goal by identifying the salient portions of the original texts that lay out the setting of a story, namely, location and main characters. The present prototype of our system creates summaries by extracting sentences from original documents. An example summary produced by the system appears in **Figure 1**.

The system works in two stages. First it attempts to identify important entities in stories (locations and characters). Next, sentences that are descriptive and set out the background of a story are separated from those that relate events of the plot. Finally, the system selects summary-worthy sentences in a way that favours descriptive ones that focus on important entities and occur early in the text.

The identification of important entities is achieved by processing the stories using a gazetteer. Pronominal and noun phrase anaphora are very common in fiction, so we resolve anaphoric expressions of these two types. The anaphora resolution module is restricted to resolving singular anaphoric expressions that denote animate entities (people and, sometimes, animals). The main characters are then identified using normalized frequency counts.

The next stage of the process attempts to identify sentences that set out the background in each story. The stories are parsed using the Connexor Machinese Syntax Parser (Tapanainen and Järvinen 1997) and sentences are split into clauses.

Each clause is represented as a vector of features that approximate its aspectual type. The features are designed to help identify state clauses (*John was a tall man*) and serial situations (*John always drops things*) (Huddleston and Pullum 2002, p. 123-124).

Four groups of features represent each clause: character-related, location-related, aspect-related and others. Character-related features capture such information as the presence of a mention of one of the main characters in a clause, its syntactic function, how early in the text this mention occurs, etc. Location-related features state whether a clause contains a location name and whether this name is embedded in a prepositional phrase. Aspect-related features reflect a number of properties of a clause that influence its aspectual type. They include the main verb's lexical aspect, the tense, the presence and the type of temporal expressions, voice, and the presence of modal verbs.

In our experiments we create two separate representations for each clause: fine-grained and coarse-grained. Both contain features from all four feature groups. The difference between them is only in the number of features and in the cardinality of the set of possible values.

Two different procedures achieve the actual selection process. The first procedure performs decision tree induction using C5.0 (Quinlan 1992) to select the most likely candidate sentences. The training data for this process consists of short stories annotated at the clause-level by the first author of this paper. The second procedure applies a set of manually created rules to select summary-worthy sentences.

The corpus for the experiments contains 47 short stories from Project Gutenberg (http://www.gutenberg.org) divided into a training set (27 stories) and a test set (20 stories). These are classical works written in English or translated into English by authors including O.Henry, Jerome K. Jerome, Katherine Mansfield and Anton Chekhov. They have on average 3,333 tokens and 244 sentences (4.5 letter-sized pages). The target compression rate was set at 6% counted in sentences. This rate was selected because it corresponded to the compression rate achieved by the first author when creating initial training and test data.

3 Evaluation: Experimental Setup

We designed our evaluation procedure to have easily interpreted, meaningful results, and keep the amount of labour reasonable. We worked with six subjects (different than the authors of this paper) who performed two separate tasks.

In Task 1 each subject was asked to read a story and create its summary by selecting 6% of the sentences. The subjects were explained that their summaries were to raise expectations about the story, but not to reveal what happens in it.

In Task 2 the subjects made a number of judgments about the summaries before and after reading the original stories. The subjects read a summary similar to the one shown in **Figure 1**. Next, they were asked six questions, three of which were factual in nature and three others were subjective. The subjects had to answer these questions using the summary as the only source of information. Subsequently, they read the original story and answered almost the same questions (see **Section 4**). This process allowed us to understand how informative the summaries were by themselves, without access to the originals, and also whether they were misleading or incomplete.

The experiments were performed on a test set of 20 stories and involved six participants divided into two groups of three people. Group 1 performed Task 1 on stories 1-10 of the testing set and Group 2 performed this task on stories 11-20. During Task 2 Group 1 worked on stories 11-20 and Group 2 – on stories 1-10.

By adjusting a number of system parameters, we produced four different summaries per story. All four versions were compared with human-made summaries using sentence overlap-based measures. However, because the experiments are rather time consuming, it was not possible to evaluate more than one set of summaries using human judgments (Task 2). That is why only summaries generated using the coarse-grained dataset and manually composed rules were evaluated in Task 2. We selected this version because the differences between this set of summaries and gold-standard summaries are easiest to interpret. That is to say, decisions based on a set of rules employing a smaller number of parameters are easier to track than those taken using machine learning or more elaborate rules.

On average, the subjects reported that completing both tasks required between 15 and 35 hours of work. Four out of six subjects were native speakers of English. Two others had a near-native and very good levels of English respectively. The participants were given the data in form of files and had four weeks to complete the tasks.

3.1 Creating Gold-Standard Summaries: Task 1

During this task each participant had to create extract-based summaries for 10 different stories. The criteria (making a summary indicative rather than informative) were explained and one example of an annotated story shown. The instructions for these experiments are available at <http://www.site.uottawa.ca/~ankazant/instructions.zip>.

Table 1 presents several measures of agreement between judges within each group and with the first author of this paper (included in the agreement figures because this person created the initial training data and test data for the preliminary experiments).

The measurement names are displayed in the first column of **Table 1**. *Cohen* denotes Cohen's kappa (Cohen 1960). *PABAK* denotes Prevalence and Bias Adjusted Kappa (Bland and Altman 1986). *ICC* denotes Intra-class Correlation Coefficient (Shrout and Fleiss 1979). The numbers *3* and *4* state whether the statistic is computed only for 3 subjects participating in the evaluation or for 4 subjects (including the first author of the paper).

Table 1. Inter-judge agreement.

Statistic	Group 1	Group 2	Average
Cohen (4)	0.50	0.34	0.42
Cohen (3)	0.51	0.34	0.42
PABAK (4)	0.88	0.85	0.87
PABAK (3)	0.89	0.86	0.87
ICC (4)	0.80 (0.78, 0.82)	0.67 (0.64, 0.70)	0.73 (0.71, 0.76)
ICC (3)	0.76 (0.74, 0.80)	0.6 (0.56, 0.64)	0.68 (0.65, 0.72)

As can be seen in **Table 1**, the agreement statistics are computed for each group separately. This is because the sets of stories that they annotated are disjoint. The column *Average* provides an average of these figures to give a better overall idea.

Cohen's kappa in its original form can only be computed for a pair of raters. For this reason we computed it for each possible pair-wise combination of raters within a group and then the numbers were averaged. The PABAK statistic was computed in the same manner using Cohen's kappa as its basis. ICC is the statistic that measures inter-rater agreement and can be computed for more than 2 judges. It was computed for all 3 or 4 raters at the same time. ICC was computed for a two-way mixed model and measures the average reliability of ratings taken together. The numbers in parentheses are confidence intervals for 99% confidence.

We compute three different agreement measures because each of these statistics has its weakness and distorts the results in a different manner. Cohen's kappa is known to be a pessimistic measurement in the presence of a severe class imbalance, as is the case in our setting (Sim and Wright 2005). PABAK is a measure that takes class imbalance into account, but it is too optimistic because it artificially removes class imbalance present in the original setting. ICC has weaknesses similar to Cohen's kappa (sensitivity to class imbalance). Besides, it assumes that the sample of targets to be rated (sentences in our case) is a random sample of targets drawn from a larger population. This is not

Figure 2. Fragments of summaries produced by 3 annotators for *The Cost of Kindness* by Jerome K Jerome.

Annotator A.

The Rev. Augustus Cracklethorpe would be quitting Wychwood-on-the-Heaththe the following Monday, never to set foot […] in the neighbourhood again. The Rev. Augustus Cracklethorpe, M.A., might possibly have been of service to his Church in, say, […] some mission station far advanced amid the hordes of heathendom. In picturesque little Wychwood-on-the-Heath […] these qualities made only for scandal and disunion. Churchgoers who had not visited St. Jude's for months had promised themselves the luxury of feeling they were listening to the Rev. Augustus Cracklethorpe for the last time. The Rev. Augustus Cracklethorpe had prepared a sermon that for plain speaking and directness was likely to leave an impression.

Annotator B.

The Rev. Augustus Cracklethorpe would be quitting Wychwood-on-the-Heaththe the following Monday, never to set foot […] in the neighbourhood again. The Rev. Augustus Cracklethorpe, M.A., might possibly have been of service to his Church in, say, [..] some mission station far advanced amid the hordes of heathendom. What marred the entire business was the impulsiveness of little Mrs. Pennycoop. Mr. Pennycoop, carried away by his wife's eloquence, added a few halting words of his own. Other ladies felt it their duty to show to Mrs. Pennycoop that she was not the only Christian in Wychwood-on-the-Heath.

Annotator C.

The Rev. Augustus Cracklethorpe would be quitting Wychwood-on-the-Heath the following Monday, never to set foot […] in the neighbourhood again. The Rev. Augustus Cracklethorpe, M.A., might possibly have been of service to his Church in, say, […] some mission station far advanced amid the hordes of heathendom. For the past two years the Rev. Cracklethorpe's parishioners […] had sought to impress upon him, [..] their cordial and daily-increasing dislike of him, both as a parson and a man. The Rev. Augustus Cracklethorpe had prepared a sermon that for plain speaking and directness was likely to leave an impression. The parishioners of St. Jude's, Wychwood-on-the-Heath, had their failings, as we all have. The Rev. Augustus flattered himself that he had not missed out a single one, and was looking forward with pleasurable anticipation to the sensation that his remarks, from his "firstly" to his "sixthly and lastly," were likely to create.

necessarily the case as the corpus was not compiled randomly.

We hope that these three measures, although insufficient individually, provide an adequate understanding of inter-rater agreement in our evaluation. We note that the average overlap (intersection) between judges in each group is 1.8% out of 6% of summary-worthy sentences.

All of these agreement measures and, in fact, all measures based on computing sentence overlap are inherently incomplete where fiction is concerned because any two different sentences are not necessarily "equally different". The matter is exemplified in **Figure 2**. It displays segments of summaries produced for the same story by three different annotators. Computing Cohen's kappa between these fragments gives agreement of 0.521 between annotators A and B and 0.470 between annotators A and C. However, a closer look at these fragments reveals that there are more differences between summaries A and B than between summaries A and C. This is because many of the sentences in summaries A and C describe the same information (personal qualities of Rev. Cracklethorpe) even though they do not overlap. On the other hand, sentences from summaries A and B are not only distinct; they "talk" about different facts. This problem is not unique to fiction, but in this context it is more acute because literary texts exhibit more redundancy.

Table 2. Sentence overlap between computer- and human-made summaries. Majority gold-standard.

Dataset	Prec.	Rec.	F
LEAD	25.09	30.49	27.53
LEAD CHAR	28.14	33.18	30.45
Rules, coarse-grained	34.14	44.39	38.60
Rules, fine-gr.	39.27	50.00	43.99
Machine learning, coarse-gr.	35.55	40.81	38.00
ML, fine-gr.	37.97	50.22	43.22

Tables 2-4 show the results of comparing four different versions of computer-made summaries against gold-standard summaries produced by humans. The tables also display the results of two baseline algorithms. The LEAD baseline refers to the version of summaries produced by selecting the first 6% of sentences in each story. LEAD CHAR baseline is obtained by selecting first

Table 3. Sentence overlap between computer- and human-made summaries. Union gold-standard.

Dataset	Prec.	Rec.	F
LEAD	36.53	17.97	24.09
LEAD CHAR	44.49	21.23	28.75
Rules, coarse-grained	52.41	30.96	38.92
Rules, fine-gr.	56.77	31.22	40.28
Machine learning, coarse-gr.	51.17	23.76	32.47
ML, fine-gr.	55.59	29.76	38.77

Table 4. Sentence overlap between computer- and human-made summaries. Intersection gold-standard.

Dataset	Prec.	Rec.	F
LEAD	12.55	37.36	18.78
LEAD CHAR	15.97	46.14	23.73
Rules, coarse-grained	19.66	62.64	29.92
Rules, fine-gr.	23.10	76.92	35.53
Machine learning, coarse-gr.	19.14	53.85	28.24
ML, fine-gr.	21.36	69.23	32.64

6% of sentences that contain a mention of an important character. The improvements over the baselines are significant with 99% confidence in all cases.

By combining summaries created by human annotators in different ways we create three distinct gold-standard summaries.

The majority gold-standard summary contains all sentences that were selected by at least two judges. It is the most commonly accepted way of creating gold-standard summaries and it is best suited to give an overall picture of how similar computer-made summaries are to man-made ones.

The union gold standard is obtained by considering all sentences that were judged summary-worthy by at least one judge. Union summaries provide a more relaxed measurement. Precision for the union gold standard gives one an idea of how many irrelevant sentences a given summary contains (sentences not selected by any of three judges are more likely to prove irrelevant).

The *intersection* summaries are obtained by combining sentences that all three judges deemed to be important. Intersection gold standard is the strictest way to measure the goodness of a summary. Recall for intersection gold standard tells one how many of the most important sentences were included in summaries by the system (sentences selected by all three judges are likely to be the most important ones).

It should be noted, however, that the numbers in **Tables 2-4** do not give a complete picture of the quality of the summaries for the same reason that the agreement measures do not reveal fully the extent of inter-judge agreement: sentences that are not part of the reference summaries are not necessarily equally unsuitable for inclusion in the summary.

3.2 Human Judgment of Computer-Made Summaries: Task 2

In order to evaluate one summary in Task 2, a participant had to read it and to answer six questions using the summary as the only source of information. The participant was then required to read the original story and to answer another six questions. The questions asked before and after reading the original were the same with one exception: question Q4 was replaced by Q11 (see **Table 6**.) The subjects were asked not to correct the answers after the fact.

Table 5. Answers to factual questions.

Id	Question	After summary only		After reading the original	
		Mean	Std. dev	Mean	Std. dev.
Q1, Q7	Please list up to 3 main characters in this story, in the order of importance (scale: -1 to 3)	2.28	0.64	2.78	0.45
Q2, Q8	State where this story takes place. Be as specific as possible (scale: -1 to 3)	1.78	1.35	2.60	0.91
Q3, Q9	Select a time period when this story takes place.(scale: 0 or 1)	0.53	0.50	0.70	0.46

Table 6. Answers to subjective questions.

Id	Question (scale: 1 to 6)	After summary only		After reading the original	
		Mean	Std. dev	Mean	Std. dev.
Q4	How readable do you find this summary?	4.43	1.39	N/A	N/A
Q5, Q10	How much irrelevant information does this summary contain?	4.27	1.41	4.51	1.16
Q11	How complete is the summary?	N/A	N/A	4.53	1.25
Q6, Q12	How helpful was this summary for deciding whether you would like to read the story or not?	4.52	1.37	4.6	1.21

Three of the questions were factual and three others – subjective. **Table 5** displays the factual questions along with the resulting answers. The participants had to answer questions Q1 and Q2 in their own words and question Q3 was a multiple-choice question where a participant selected the century when the story took place. Q1 and Q2 were ranked on a scale from -1 to 3. A score of 3 means that the answer was complete and correct, 2 – slightly incomplete, 1 – very incomplete, 0 – a subject could not find the answer in the text and -1 if the person answered incorrectly. Q3 was ranked on a binary scale (0 or 1).

Questions Q3-Q7 asked the participants to pronounce a subjective judgment on a summary. These were multiple-choice questions where a participant needed to select a score from 1 to 6, with 1 indicating a strong negative property and 6 indicating a strong positive property. The questions and results appear in **Table 6**.

The results displayed in **Tables 5** and **6** suggest that the subjects can answer simple questions based on the summaries alone. They also seem to indicate that the subjects found the summaries quite helpful. It is interesting to note that even after reading complete stories the subjects are not always capable of answering the factual questions with perfect precision.

Table 7. Spearman rank correlation coefficient between sentence overlap measures and human judgments.

Question	Prec.	Rec.	F
Q1(main characters)	0.09	0.29	0.17
Q2(location)	0.21	0.18	0.22
Q3(time)	0.38	0.28	0.34
Q4(readability)	0.47	0.31	0.50
Q5(relevance)	0.31	0.19	0.34
Q10(relevance)	0.60	0.40	0.59
Q11(completeness)	0.40	0.29	0.40
Q12(helpfulness)	0.59	0.41	0.61

3.3 Putting Sentence Overlap and Human Judgment Together

In order to check whether the two types of statistics measure the same or different qualities of the summaries, we explored whether the two are correlated.

Table 7 displays the values of Spearman rank correlation coefficient between median values of answers for questions from Task 2 and measurements obtained by comparing computer-made summaries against the majority gold-standard summaries. All questions, except Q10 (relevance) and Q11 (completeness) are those asked and answered using the summary as the only source of information. Sentence overlap values (F-score, precision and recall) were discretized (banded) in order to be used in this test. These results are based on the values obtained for 20 stories in the test set – a relatively small sample – which prohibits drawing definite conclusions. However, in most cases the correlation coefficient between human opinions and sentence overlap measurements is below the cut-off

Table 8. ANOVA F-values between sentence overlap measures and human judgments.

Question	Prec.	Rec.	F
Q1(main characters)	0.60	0.61	0.58
Q2(location)	2.58	1.94	2.36
Q3(time)	1.11	0.67	0.97
Q4(readability)	2.10	0.90	1.60
Q5(relevance)	4.55	3.75	4.28
Q10 (relevance)	6.33	3.46	5.15
Q11(completeness)	3.11	4.22	3.43
Q12(helpfulness)	4.53	2.54	3.72

value with 99% confidence, which is 0.57 (the exceptions are highlighted). This suggests that in our case the measurements using sentence overlap as their basis are not correlated with the opinions of subjects about the summaries.

We also performed a one-way ANOVA test using human judgments as independent factors and sentence-overlap based measures as dependent variables. The results are in line with those obtained using Spearman coefficient. They are shown in **Table 8**. The F-values which are statistically significant with 99% confidence are highlighted (the cut-off value for questions Q4-Q12 is 4.89, for Q1 and Q2 – 6.11 and for Q3 – 8.29).

4 Conclusions and Future Work

This paper presented an experimental way of evaluating automatically produced summaries of literary short stories.

In the course of our experiments we have remarked a few issues pertinent to evaluating summaries of short fiction. Firstly, higher degree of redundancy of sentences in texts makes measures based on sentence overlap not very enlightening when evaluating extracted summaries. Secondly, at least in our corpus, the sentence overlap-based measures do not correlate well with those measuring opinions of humans about summaries.

This work is exploratory, and as such raises more questions than it answers. In order to evaluate summaries of literary works in a meaningful and reliable way one needs to define criteria which make such summaries suitable or not suitable for a particular purpose. We will explore this issue in our future work. We also intend to apply the pyramid method of evaluating summaries to extracted summaries produced by the human annotators.

5 Acknowledgements

The authors would like to express their gratitude to Connexor Oy and especially to Atro Voutilainen for their kind permission to use Connexor Machinese Syntax parser free of charge for research purposes.

References

J. Bland and D. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476):307-310.

J. Cohen, 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960; 20:37-46..

R. Huddleston and G. Pullum. 2002. *The Cambridge Grammar of the English Language Usage*, 123-124. Cambridge University Press.

A. Kazantseva. 2006. *Proc Student Research Workshop* at EACL 2006, 55-63.

I. Mani. 2001. *Automatic Summarization*. John Benjamins B.V.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *Proc Human Language Technology Conference and NAACL.*

J. Quinlan. 1992. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Pub., San Mateo, CA.

P. Shrout and J. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979; 86:420–428

J. Sim and C. Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 2005(85-3): 257-268.

P. Tapanainen and T. Järvinen. 1997. A non-projective dependency parser. *Proc 5th Conference on Applied Natural Language Processing,* 64-71.

H. Van Halteren and S. Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. *HLT/NAACL-2003 Workshop on Automatic Summarization,* 57-64.