An Approach to Summarizing Short Stories

Anna Kazantseva

The School of Information Technology and Engineering
University of Ottawa
`ankazant@site.uottawa.ca`

Abstract

This paper describes a system that produces extractive summaries of short works of literary fiction. The ultimate purpose of produced summaries is defined as helping a reader to determine whether she would be interested in reading a particular story. To this end, the summary aims to provide a reader with an idea about the settings of a story (such as characters, time and place) without revealing the plot. The approach presented here relies heavily on the notion of aspect. Preliminary results show an improvement over two naïve baselines: a lead baseline and a more sophisticated variant of it. Although modest, the results suggest that using aspectual information may be of help when summarizing fiction. A more thorough evaluation involving human judges is under way.

1 Introduction

In the course of recent years the scientific community working on the problem of automatic text summarization has been experiencing an upsurge. A multitude of different techniques has been applied to this end, some of the more remarkable of them being (Marcu, 1997; Mani et al. 1998; Teufel and Moens, 2002; Elhadad et al., 2005), to name just a few. These researchers worked on various text genres: scientific and popular scientific articles (Marcu, 1997; Mani et al., 1998), texts in computational linguistics (Teufel and Moens, 2002), and medical texts (Elhadad et al., 2002). All these genres are examples of texts characterized by rigid structure, relative abundance of surface markers and straightforwardness. Relatively few attempts have been made at summarizing less structured genres, some of them being dialogue and speech summarization (Zechner, 2002; Koumpis et al. 2001). The issue of summarizing fiction remains largely untouched, since a few very thorough earlier works (Charniak, 1972; Lehnert, 1982). The work presented here seeks to fill in this gap.

The ultimate objective of the project is stated as follows: to produce indicative summaries of short works of fiction such that they be helpful to a potential reader in deciding whether she would be interested in reading a particular story or not. To this end, revealing the plot was deemed unnecessary and even undesirable. Instead, the current approach relies on the following assumption: when a reader is presented with an extracted summary outlining the general settings of a story (such as time, place and who it is about), she will have enough information to decide how interested she would be in reading a story. For example, a fragment of such a summary, produced by an annotator for the story *The Cost of Kindness* by Jerome K. Jerome is presented in **Figure 1**. The plot, which is a tale of how one local family decides to bid a warm farewell to Rev. Cracklethorpe and causes the vicar to change his mind and remain in town, is omitted.

The data used in the experiments consisted of 23 short stories, all written in XIX – early XX century by main-stream authors such as Katherine Mansfield, Anton Chekhov, O.Henry, Guy de Maupassant and others (13 authors in total). The genre can be vaguely termed social fiction with the exception of a few fairy-tales. Such vagueness as far as genre is concerned was deliberate, as the author wished to avoid producing a system relying on cues specific to a particular genre. Average length of a story in the corpus is 3,333 tokens (approximately 4.5 letter-sized pages) and the target compression rate is 6%.

In order to separate the background of a story from events, this project relies heavily on the notion of aspect (the term is explained in **Section 3.1**). Each clause of every sentence is described in terms of aspect-related features. This representation is then used to select salient descriptive sentences and to leave out those which describe events.

The organization of the paper follows the overall architecture of the system. **Section 2** provides a generalized overview of the pre-processing stage of the project, during which pronominal and nominal *anaphoric references* (the term is explained in **Section 2**) were resolved and main characters were identified. **Section 3** briefly reviews the concept of aspect, gives an overview of the system and provides the linguistic motivation behind it. **Section 4** describes the classification procedures (machine learning and manual rule creation) used to distinguish between descriptive elements of a story and passages that describe events. It also reports results. **Section 5** draws some conclusions and outlines possible directions in which this work may evolve.

2 Data Pre-Processing

Before working on selecting salient descriptive sentences, the stories of the training set were analyzed for presence of surface markers denoting characters, locations and temporal anchors. To this end, the GATE Gazetteer (Cunningham et al., 2002) was used, and only entities recognized by it automatically were considered.

The findings were as follows. Each story contained multiple mentions of characters (an average of 64 mentions per story). Yet only 22 location markers were found, most of these being street names. The 22 markers were found in 10 out of 14 stories, leaving 4 stories without any identifiable location markers. Only 4 temporal anchors were identified in all 14 stories: 2 absolute (such as years) and 2 relative (names of holidays). These findings support the intuitive idea that short stories revolve around their characters, even if the ultimate goal is to show a larger social phenomenon.

Due to this fact, the data was pre-processed in such a way as to resolve pronominal and nominal *anaphoric references* to animate entities. The term *anaphora* can be informally explained as a way of mentioning a previously encountered entity without naming it explicitly. Consider examples *1a* and *1b* from *The Gift of the Magi* by O. Henri. *1a* is an example of pronominal anaphora, where the noun phrase (further NP) *Della* is referred to as an *antecedent* and both occurrences of the pronoun *her* as *anaphoric expressions* or *referents*. Example *1b* illustrates the concept of nominal anaphora. Here the NP *Dell* is the antecedent and *my girl* is the anaphoric expression (in the context of this story *Della* and the *girl* are the same person).

(1a)   Della *finished* her *cry and attended to* her *cheeks with the powder rag.*
(1b)   *"Don't make any mistake,* Dell,*" he said, "about me. I don't think there's anything […] that could make me like* my girl *any less.*

The author created a system that resolved 1st and 3rd person singular pronouns (*I*, *me*, *my*, *he*, *his* etc.) and singular nominal anaphoric expressions (e.g. *the man*, but not *men*). The system was implemented in Java, within the GATE framework, using Connexor Machinese Syntax parser (Tapanainen and Järvinen, 1997).

A generalized overview of the system is provided below. During the first step, the documents were parsed using Connexor Machinese Syntax parser. The parsed data was then forwarded to the Gazetteer in GATE, which recognized nouns denoting persons. The original version of the Gazetteer recognized only named entities and professions, but the Gazetteer was extended to include common animate nouns such as *man*, *woman,* etc. As the next step, an implementation based on a classical pronoun resolution algorithm (Lappin and Leass, 1994) was applied to the texts. Subsequently, anaphoric noun phrases were identified using the rules outlined

in (Poesio and Vieira, 2000). Finally, these anaphoric noun phrases were resolved using a modified version of (Lappin and Leass, 1994), adjusted to finding antecedents of nouns.

Table 1. Results of anaphora resolution.				
Type of anaphora	All	Correct	Incorrect	Error rate, %
Pronominal	597	507	90	15.07
Nominal	152	96	56	36.84
Both	749	603	146	19.49

A small-scale evaluation based on 2 short stories revealed results shown in **Table 1**. After resolving anaphoric expressions, characters that are central to the story were selected based on normalized frequency counts.

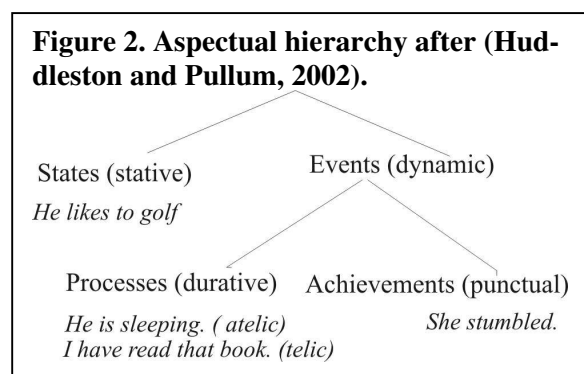# 3 Selecting Descriptive Sentences Using Aspectual Information

3.1 Linguistic definition of aspect

In order to select salient sentences that set out the background of a story, this project relied on the notion of *aspect*. For the purposes of this paper the author uses the term *aspect* to denote the same concept as what (Huddleston and Pullum, 2002) call the *situation type*. Informally, it can be explained as a characteristic of a clause that gives an idea about the temporal flow of an event or state being described.

A general hierarchy of aspectual classification based on (Huddleston and Pullum, 2002) is shown in **Figure 2** with examples for each type. In addition, aspectual type of a clause may be altered by multiplicity, e.g. repetitions. Consider examples *2a* and *2b*.

(2a)   She read a book.
(2b)   She usually read a book a day. (e.g. She

Figure 2. Aspectual hierarchy after (Huddleston and Pullum, 2002).

States (stative)
He likes to golf

Events (dynamic)

Processes (durative)
He is sleeping. (atelic)
I have read that book. (telic)

Achievements (punctual)
She stumbled.

used to read a book a day).

Example *2b* is referred to as *serial* situation (Huddleston and Pullum, 2002). It is considered to be a state, even though a single act of reading a book would constitute an event.

Intuitively, stative situations (especially serial ones) are more likely to be associated with descriptions; that is with things that are, or things that were happening for an extended period of time (consider *He was a tall man.* vs. *He opened the window.*).The rest of **Section 3** describes the approach used for identifying single and serial stative clauses and for using them to construct summaries.

3.2   Overall system design

Selection of the salient background sentences was conducted in the following manner. Firstly, the pre-processed data (as outlined in **Section 2**) was parsed using Connexor Machinese Syntax parser. Then, sentences were recursively split into *clauses*. For the purposes of this project a *clause* is defined as a main verb with all its complements, including subject, modifiers and their sub-trees.

Subsequently, two different representations were constructed for each clause: one fine-grained and one coarse-grained. The main difference between these two representations was in the number of attributes and in the cardinality of the set of possible values, and not in how much and what kind of information they carried. For instance, the fine-grained dataset had 3 different features with 7 possible values to carry tense-related information: *tense*, *is_progressive* and *is_perfect*, while the coarse-grained dataset carried only one binary feature, *is_simple_past_or_present*.

Two different approaches for selecting descriptive sentences were tested on each of the representations. The first approach used machine learning techniques, namely C5.0 (Quinlan, 1992) implementation of decision trees. The second approach consisted of applying a set of manually created rules that guided the classification process. Motivation for features used in each dataset is given in **Section 3.3**. Both approaches and preliminary results are discussed in **Sections 4.1 - 4.4**.

The part of the system responsible for selecting descriptive sentences was implemented in Python.

3.3   Feature selection: description and motivation

57

Table 2. Description of the features in both datasets				
	Fine-grained dataset		Coarse-grained dataset	
Type of features	Number of features	Number of values	Number of features	Number of values
Character-related	9	16	4	6
Aspect-related	12	92	8	16
Location-related	2	4	2	4
Others	4	9	3	4
All	27	121	17	30

Features for both representations were selected based on one of the following criteria:

(Criterion 1)   a clause should 'talk' about important things, such as characters or locations

(Criterion 2)   a clause should contain background descriptions rather then events

The number of features providing information towards each criterion, as well as the number of possible values, is shown in **Table 2** for both representations.

The attributes contributing towards *Criterion 1* can be divided into *character-related* and *location-related*.

Character-related features were designed so as to help identify sentences that focused on characters, not just mentioned them in passing. These attributes described whether a clause contained a character mention and what its grammatical function was (subject, object, etc.), whether such a mention was modified and what was the position of a parent sentence relative to the sentence where this character was first mentioned (intuitively, earlier mentions of characters are more likely to be descriptive).

Location-related features in both datasets described whether a clause contained a location mention and whether it was embedded in a prepositional phrase (further PP). The rationale behind these attributes is that location mentions are more likely to occur in PPs, such as *from the Arc de Triomphe*, *to the Place de la Concorde.*

In order to meet *Criterion 2* (that is, to select descriptive sentences) a number of aspect-related features were calculated. These features were selected so as to model characteristics of a clause that help determine its aspectual class. The characteristics used were default aspect of the main verb of a clause, tense, temporal expressions, semantic category of a verb, voice and some properties of the direct object. Each of these characteristics is listed below, along with motiva-

tion for it, and information about how it was calculated.

It must be mentioned that several researchers looked into determining automatically various semantic properties of verbs, such as (Siegel, 1998; Merlo et al., 2002). Yet these approaches dealt with properties of verbs in general and not with particular usages in the context of concrete sentences.

Default verbal aspect. A set of verbs, referred to as *stative* verbs, tends to produce mostly stative clauses. Examples of such verbs include *be*, *like*, *feel*, *love*, *hate* and many others. A common property of such verbs is that they do not readily yield a progressive form (Vendler, 1967; Dowty, 1979). Consider examples *3a* and *3b*.

(3a)   She is talking. (a dynamic verb talk*)*
*(3b)   *She is liking the book. (a stative verb* like*)*

The default aspectual category of a verb was approximated using *Longman Dictionary of Contemporary English* (LDOCE). Verbs marked in LDOCE as not having a progressive form were considered stative and all others – dynamic. This information was expressed in both datasets as 1 binary feature.

Grammatical tense. Usually, simple tenses are more likely to be used in stative or habitual situations than progressive or perfect tenses. In fact, it is considered to be a property of stative clauses that they normally do not occur in progressive (Vendler, 1967; Huddleston and Pullum, 2002). Perfect tenses are feasible with stative clauses, yet less frequent. Simple present is only feasible with states and not with events (Huddleston and Pullum, 2002) (see examples *4a* and *4b*).

(4a)   She likes writing.
*(4b)   *She writes a book. (e.g. now)*

In the fine-grained dataset this information was expressed using 3 features with 7 possible values

(whether a clause is in present, past or future tense, whether it is progressive and whether it is perfective). In the coarse-grained dataset, this information was expressed using 1 binary feature: whether a clause is in simple past or present tense.

Temporal expressions. Temporal markers (often referred to as temporal adverbials), such as *usually*, *never*, *suddenly*, *at that moment* and many others are widely employed to mark the aspectual type of a sentence (Dowty, 1982; Harkness, 1987; By, 2002). Such markers provide a wealth of information and often unambiguously signal aspectual type. For example:

(5a)   She read a lot <u>tonight.</u>
(5b)   She <u>always</u> read a lot. (Or She <u>used to</u> read a lot.)

Yet, such expressions are not easy to capture automatically. In order to use the information expressed in temporal adverbials, the author analyzed the training data for presence of such expressions and found 295 occurrences in 10 stories. It appears that this set could be reduced to 95 templates in the following manner. For example, the expressions *this year*, *next year*, *that long year* could all be reduced to a template <*some_expression*> *year*. Each template is characterized by 3 features: type of the temporal expression (location, duration, frequency, enactment) (Harkness, 1987); magnitude (year, day, etc.); and plurality (year vs. years). The fine-grained dataset contained 3 such features with 14 possible values (type of expression, its magnitude and plurality). The coarse-grained dataset contained 1 binary feature (whether there was an expression of a long period of time).

Verbal semantics. Inherent meaning of a verb also influences the aspectual type of a given clause.

(6a)   She <u>memorized</u> that book by heart. (an event)
(6b)   She <u>enjoyed</u> that book. (a state)

Not surprisingly, this information is very difficult to capture automatically. Hoping to leverage it, the author used semantic categorization of the 3,000 most common English verbs as described in (Levin, 1993). The fine-grained dataset contained a feature with 49 possible values that corresponded to the top-level categories described in (Levin, 1993). The coarse-grained dataset contained 1 binary feature that carried this informa-

tion. Verbs that belong to more than one category were manually assigned to a single category that best captured their literal meaning.

Voice. Usually, clauses in passive voice only occur with events (Siegel, 1998). Both datasets contained 1 binary feature to describe this information.

Properties of direct object. For some verbs properties of direct object help determine whether a given clause is stative or dynamic.

(7a)   She wrote a book. (event)
(7b)   She wrote books. (state)

The fine-grained dataset contained 2 binary features to describe whether direct object is definite or indefinite and whether it is plural. The coarse-grained dataset contained no such information because it appeared that this information was not crucial.

Several additional features were present in both datasets that described overall characteristics of a clause and its parent sentence, such as whether these were affirmative, their index in the text, etc. The fine-grained dataset contained 4 such features with 9 possible values and the coarse-grained dataset contained 3 features with 7 values.

4   Experiments

4.1   Experimental setting

The data used in the experiments consisted of 23 stories split into a training set (14 stories) and a testing set (9 stories). Each clause of every story was annotated by the author of this paper as summary-worthy or not. Therefore, the classification process occurred at the clause-level. Yet, summary construction occurred at the sentence-level, that is if one clause in a sentence was considered summary-worthy, the whole sentence was also considered summary-worthy. Because of this, results are reported at two levels: clause and sentence. The results at the clause-level are more appropriate to judge the accuracy of the classification process. The results at the sentence level are better suited for giving an idea about how close the produced summaries are to their annotated counterparts.

The training set contained 5,514 clauses and the testing set contained 4,196 clauses. The target compression rate was set at 6% expressed in terms of sentences. This rate was selected because it approximately corresponds to the average compression rate achieved by the annotator

(5.62%). The training set consisted of 310 positive examples and 5,204 negative examples, and the testing set included 218 positive and 3,978 negative examples.

Before describing the experiments and discussing results, it is useful to define baselines. The author of this paper is not familiar with any comparable summarization experiments and for this reason was unable to use existing work for comparison. Therefore, a baseline needed to be defined in different terms. To this end, two naïve baselines were computed.

Intuitively, when a person wishes to decide whether to read a certain book or not, he opens it and flips through several pages at the beginning. Imitating this process, a simple lead baseline consisting of the first 6% of the sentences in a story was computed. It is denoted LEAD in **Tables 3** and **4**. The second baseline is a slightly modified version of the lead baseline and it consists of the first 6% of the sentences that contain at least one mention of one of the important characters. It is denoted LEAD CHAR in **Tables 3 and 4**.

4.2   Experiments with the rules

The first classification procedure consisted of applying a set of manually designed rules to produce descriptive summaries. The rules were designed using the same features that were used for machine learning and that are described in **Section 3.3**.

Two sets of rules were created: one for the fine-grained dataset and another for the coarse-grained dataset. Due to space restrictions it is not possible to reproduce the rules in this paper. Yet, several examples are given in **Figure 4**. (If a rule returns *True*, then a clause is considered to be summary-worthy.)

Figure 4. Examples of manually composed rules.

Rule 1
if a clause contains a character mention as subject or object **and** a temporal expression of type enactment (*ever, never, always*)
            return True
Rule 2
if a clause contains a character mention as subject or object **and** a stative verb
            return True
Rule 3
if a clause is in progressive tense
            return False

The results obtained using these rules are presented in **Table 3**. They are discussed along with the results obtained using machine learning in **Section 4.4**.

4.3   Experiments with machine learning

As an alternative to rule construction, the author used C5.0 (Quilan, 1992) implementation of decision trees to select descriptive sentences. The algorithm was chosen mainly because of the readability of its output. Both training and testing datasets exhibited a 1:18 class imbalance, which, given a small size of the datasets, needed to be compensated. *Undersampling* (randomly removing instances of the majority class) was applied to both datasets in order to correct class imbalance.

This yielded altogether 4 different datasets (see **Table 4**). For each dataset, the best model was selected using 10-fold cross-validation on the training set. The model was then tested on the testing set and the results are reported in **Table 4**.

Table 3. Results obtained using rules (summary-worthy class)						
Dataset	Level	Precision,%	Recall,%	F-score,%	Kappa	Overall error rate,% (both classes)
Baseline LEAD	Clause	19.92	23.39	21.52	16.85	8.87
Baseline LEAD CHAR	Clause	8.93	25.69	13.25	6.01	17.47
Fine-grained	Clause	34.77	40.83	37.55	33.84	17.73
Coarse-grained	Clause	32.00	47.71	38.31	34.21	7.98
Baseline LEAD	Sent.	23.57	24.18	23.87	19.00	9.24
Baseline LEAD CHAR	Sent.	22.93	23.53	23.23	18.31	9.24
Fine-grained	Sent.	41.40	42.48	41.94	38.22	6.99
Coarse-grained	Sent.	40.91	41.18	41.04	37.31	7.03

4.4 Results

The results displayed in **Tables 3** and **4** show how many clauses (and sentences) selected by the system corresponded to those chosen by the annotator. The columns *Precision, Recall* and *F-score* show measures for the minority class (summary-worthy). The columns *Overall error rate* and *Kappa* show measures for both classes.

Although modest, the results suggest an improvement over both baselines. Statistical significance of improvements over baselines was tested for p = 0.001 for each dataset-approach. The improvements are significant in all cases.

The columns *F-score* in **Tables 3** and **4** show f-score for the minority class (summary-worthy sentences), which is a measure combining precision and recall for this class. Yet, this measure does not take into account success rate on the negative class. For this reason, Cohen's kappa statistic (Cohen, 1960) was also computed. It measures the overall agreement between the system and the annotator. This measure is shown in the column named *Kappa*.

In order to see what features were the most informative in each dataset, a small experiment was conducted. The author removed one feature at a time from the training set and used the decrease in F-score as a measure of informativeness. The experiment revealed that in the coarse-grained dataset the following features were the most informative: 1) the position of a sentence relative to the first mention of a character; 2) whether a clause contained character mentions; 3) voice and 4) tense. In the fine-grained dataset the findings were similar: 1) presence of a char-

acter mention; 2) position of a sentence in the text; 3) voice; and 4) tense were more important than the other features.

It is not easy to interpret these results in any conclusive way at this stage. The main weakness, of course, is that the results are based solely on the annotations of one person while it is generally known that human annotators are likely to exhibit some disagreement. The second issue lies in the fact that given the compression rate of 6%, and the objective that the summary be indicative and not informative, more that one 'good' summary is possible. It would therefore be desirable that the results be evaluated not based on overlap with an annotator (or annotators, for that matter), but on how well they achieve the stated objective.

5 Conclusions

In the immediate future the inconclusiveness of the results will be addressed by means of asking human judges to evaluate the produced summaries. During this process the author hopes to find out how informative the produced summaries are and how well they achieve the stated objective (help readers decide whether a story is potentially interesting to them). The judges will also be asked to annotate their own version of a summary to explore what inter-judge agreement means in the context of fiction summarization.

More remote plans include possibly tackling the problem of summarizing the plot and dealing more closely with the problem of evaluation in the context of fiction summarization.

Table 4. Results obtained using machine learning (summary-worthy class)							
Dataset	Training dataset	Level	Precision, %	Recall, %	F-score, %	Kappa	Overall error rate, %
Baseline LEAD		Clause	19.92	23.39	21.52	16.85	8.87
Baseline LEAD CHAR		Clause	8.93	25.69	13.25	6.01	17.47
Fine-grained	original	Clause	28.81	31.19	29.96	25.96	7.58
Fine-grained	undersampled	Clause	39.06	45.87	42.19	38.76	6.53
Coarse-grained	original	Clause	34.38	30.28	32.22	28.73	6.63
Coarse-grained	undersampled	Clause	28.52	33.49	30.80	26.69	7.82
Baseline LEAD		Sent.	23.57	24.18	23.87	19.00	9.24
Baseline LEAD CHAR		Sent.	22.93	23.53	23.23	18.31	9.24
Fine-grained	original	Sent.	38.93	37.91	38.41	34.57	7.22
Fine-grained	undersampled	Sent.	41.4	42.48	41.94	38.22	6.99
Coarse-grained	original	Sent.	42.19	35.29	38.43	34.91	6.72
Coarse-grained	undersampled	Sent.	37.58	38.56	38.06	34.10	7.46

References

Tomas By. 2002. *Tears in the Rain*. Ph.D. thesis, University of Sheffield.

Eugene Charniak. 1972. *Toward a Model of Children's Story Comprehension*. Ph.D. thesis, Massachusetts Institute of Technology.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement,* (20): 37–46.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia.

David Dowty. 1982. Tense, Time Adverbs, and Compositional Semantic Theory. *Linguistics and Philosophy*, (5), p. 23-59.

David Dowty. 1979. *Word Meaning and Montague Grammar.* D. Reidel Publishing Company, Dordrecht.

Noemie Elhadad, Min-Yen Kan, Judith Klavans, and Kathleen McKeown. 2005. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine* 33(2): 179-198.

Janet Harkness. 1987. Time Adverbials in English and Reference Time. In Alfred Schopf (ed.)*, Essays on Tensing in English, Vol. I: Reference Time, Tense and Adverbs*, p. 71-110. Tübingen: Max Niemeyer.

Rodney Huddleston and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language Usage,* p. 74-210. Cambridge University Press.

Konstantinos Koumpis, Steve Renals, and Mahesan Niranjan. 2001. Extractive summarization of voicemail using lexical and prosodic feature subset selection. In *Proeedings of Eurospeech*, p. 2377–2380, Aalborg, Denmark.

Herbert Leass and Shalom Lappin. 1994. An algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4): 535-561.

Wendy Lehnert. 1982. Plot Units: A Narrative Summarization Strategy. In W. Lehnert and M. Ringle (eds.). *Strategies for Natural Language Processing*. Erlbaum, Hillsdale, NJ.

Beth Levin. 1993. *English Verb Classes and Alternations.* The University of Chicago Press.

Longman Dictionary of Contemporary English. 2002. Pearson Education.

Inderjeet Mani, Eric Bloedorn and Barbara Gates. 1998. Using Cohesion and Coherence Models for Text Summarization. In *Working Notes of the Workshop on Intelligent Text Summarization*, p. 69-76. Menlo Park, California: American Association for Artificial Intelligence Spring Simposium Series.

Daniel Marcu. 1997. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD Thesis, Department of Computer Science, University of Toronto.

Paola Merlo, Suzanne Stevenson, Vivian Tsang and Gianluca Allaria. 2002. A Multilingual Paradigm for Automatic Verb Classification. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*(ACL'02), Philadelphia.

Massimo Poesio and Renata Vieira. 2000 . An Empirically Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4): 525-579.

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

Eric V. Siegel. 1998. *Linguistic Indicators for Language Understanding: Using machine learning methods to combine corpus-based indicators for aspectual classification of clauses.* Ph.D. Dissertation. Columbia University.

Pasi Tapanainen and Timo Järvinen. 1997 A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, p. 64-71.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles-experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4): 409–445.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, p. 97- 145.

Klaus Zechner. 2002. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics* 28(4):447-485.