Information Extraction and Summarization: Domain Independence through Focus Types

Min-Yen Kan and Kathleen R. McKeown
Department of Computer Science
Columbia University
New York, NY 10027, USA
{min,kathy}@cs.columbia.edu

April 2,, 1999

Abstract

We show how information extraction (IE) and summarization can be merged in a sequential pipeline, resulting in a new approach to domain-independent summarization. IE finds the document's terms and entities, that when processed by the methods shown, result in a more informative treatment of the document's topics.

1 Introduction

A critical goal in text summarization is the development of a satisfactory algorithm for summarizing long documents in a domain independent fashion. Unlike their shorter counterparts, long documents often exhibit complex discourse structure and more domain specificity, which can cause problems for current summarization techniques, such as sentence extraction. In long documents, sentence extraction can select sentences from distant locations in the article and thus, lack of coherence and the possibility of false implicatures from unrelated sentences being placed side by side is far greater than in shorter articles.

In this paper, we present a hybrid summarization system that merges information extraction (IE) with sentence extraction to reduce these errors. Like IE based summarization (also called *template based systems*), our summarizer looks for specific types of information to extract from the article which will serve as summary content. Unlike template based systems, we do not specify *a priori* the information to extract. Rather, the system dynamically determines the foci of the article, which in turn determine the specific information that will be extracted. The summary is first formed by extracting sentences from the document that contain the desired information, and later, modifying them.

The resulting system is domain independent, which allows us to summarize documents in any field without including any specific domain knowledge. This is achieved by analyzing terms and named entities, which are present in documents across all domains. From the analysis we can elevate a few salient entities or terms to *foci*, which represent the topics covered in the article.

For example, the entity "Jane Jacobs" refers to a person; an article with this focus will contain information about this person. Our system recognizes four major focus types: people, organizations, places and multiword terms. Each focus type and the interaction between foci suggest questions that may be answered by the text. These questions determine the information that will be extracted from the article. We use the question answering approach to improve sentence extraction in three ways:

1. The questions help the system select more appropriate sentences to extract;
2. The relationships between foci serve to reorder extracted sentences to make the resulting summary more coherent;
3. The descriptions of individual foci enable the system to find missing information and add it in.

Complementary to work by Jing [Jing1999], whose emphasis is on summary fluency, our approach focuses on ensuring summary informativeness. Other work on summarization at Columbia [Barzilay *et al.*1999, Radev and McKeown1998] focuses on multiple document summarization.

In the next section, we describe a classification hierarchy of summarization techniques that situates current systems and show how our strategy constitutes a new category. We then illustrate how each of the three tasks above can be accomplished, by following an example from IE output to summary. Finally, a short evaluation of the implemented system and a discussion of our findings conclude the paper.

2 The summarization hierarchy

Summarization systems can be broadly classified into two different categories: those that are template based and those that are text extraction based.

A template based system often produces a good summary if a document's domain is known. Template systems have been extensively researched in the past few decades [DeJong1982, Jacobs and Rau1990], in which articles are identified as belonging to a particular domain. The articles are then summarized by inserting extracted, domain-specific information into a text template, such as a company's name and the amount of its latest dividend. Current efforts in this arena, such as work by Radev and McKeown [Radev and McKeown1998], are considerably more sophisticated, using advanced techniques to dynamically add new text not present in the template. But when no template exists for a story, what then? Since there is an infinite variety of domains, we cannot simply exhaustively construct matching templates.

Text extraction systems avoid this problem by using empirical methods to find appropriate chunks of text for a summary. Most closely studied have been sentence level approaches [Brandow *et al.*1995, Kupiec *et al.*1995], since many argue that the sentence is the discourse unit with the best balance of semantic granularity and self contained cohesiveness. Sentence extraction has been augmented by adding other factors to compute final score of a sentence, as overviewed by Paice [Paice1990]: word importance as calculated by TF*IDF, position of the sentence in the document and the enclosing paragraph [Lin and Hovy1997], and the presence of cue phrases, among others. Selected sentences are given to the user as the summary of the article.

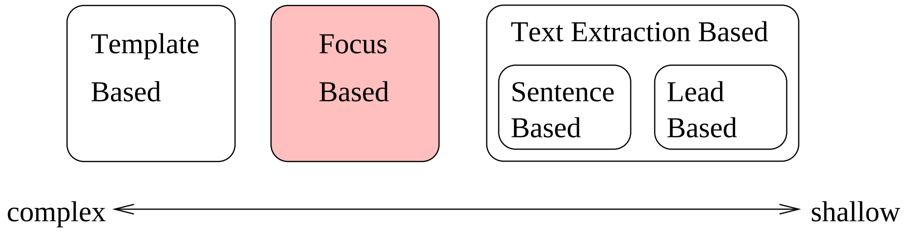

Figure 1: Positioning of our system in the complexity hierarchy

Although these two categories of automatic summarization seem fairly unrelated, they are connected as shown by Hovy and his contributors [Hovy1998] by virtue of a close cousin, information extraction (IE). In fact, a hybrid of the two categories that utilizes IE technology claims a middle ground that allows us to establish a summarization methodology hierarchy. Figure 1 shows this classification, with the axis representing the depth of knowledge needed from the source document. Our approach constitutes a middle ground, since we will show that it relies on knowing the list of the terms and named entities, which are by nature domain independent.

3 System overview

Our system was built to produce short, four to five sentence summaries of *long* (defined here as exceeding 1500 words in length) journalistically styled documents. This specific task lends itself to some optimization of general methods, especially since the length of the inputs and outputs are specified. A long input article presents difficulty with its more complex discourse structure. A short summary length forces the summary to be indicative, since we cannot hope to represent all of the many topics touched on in the full article. The key is to find only the most important point and bring together its most salient aspects and relationships.

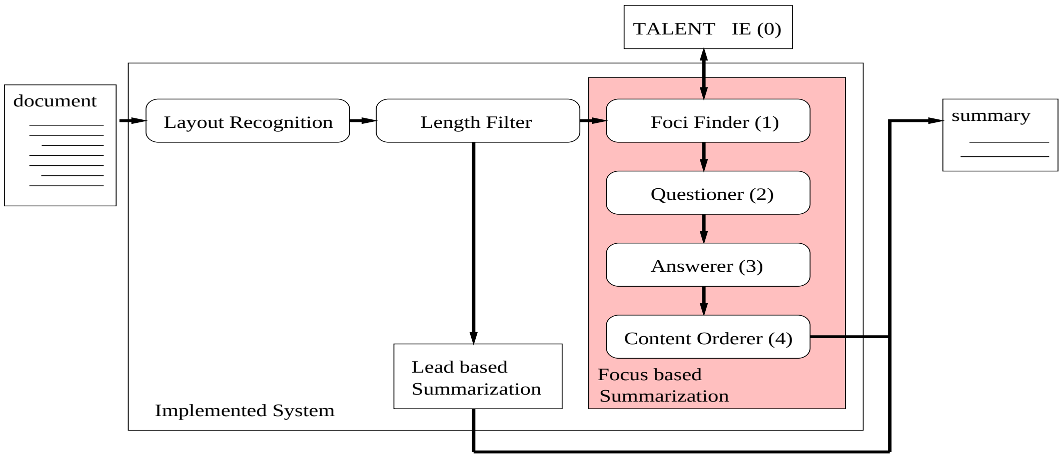

Figure 2: System architecture

Our implemented system, shown in Figure 2, is designed to address these issues. Documents first are analyzed by a layout recognition preprocessor [Kan1999] to highlight special sections and remove tables and lists. Next, a length filter determines the overall length of the preprocessed document; if it is less than the 1500 word *long* threshold, the document is routed to a simple lead based approach to complete the summary.

If the document is a *long* document, we invoke our focus based summarization approach (the shaded box in the figure) that augments sentence extraction with an analysis of IE output. This involves four steps. First, given IE output, the article's foci determined (step 1). Next, the system dynamically determines questions to ask based on the focus types present (2) and the text found to answer them are validated and extracted (3). In the final steps, we determine which pieces of information are important enough to be ordered into a coherent summary (4).

4 System modules

4.1 Step 0: Information Extraction (IE) – finding all terms and entities in the document

We use an IE engine as a first pass selector for terms and entities that are likely to be article topics, since topics – whether concrete ("Standard and Poor"), abstract ("divine right"), action ("course registration") or chronology ("World War II") – can and do appear nominalized as entities and terms.

```
24      Jane Jacobs (PERSON): Loc 1|2|3|6|8|12|14|16|18|21|25|33|47|52|54|54|57|75|76|83|86|87|95|98|
9       Toronto (PLACE): Loc 10|15|20|43|46|47|54|55|99|
7       Ideas That Matter (PERSON): Loc 9|12|12|18|29|29|91|
6       Dee W. Hock (PERSON): Loc 82|83|86|87|88|91|
5       Zeidler Roberts Partnership (ORG): Loc 43|49|50|51|96|
4       Tyler (UNAME): Loc 1|6|8|32|
... ...
1       New Denver (UNAME): Loc 100|
```

Figure 3: Some TALENT output: left column is term/entity frequency; type is shown in parenthesis

As input to the first step in the summarization system, we use IBM's information extraction tool, TALENT [Wacholder *et al.*1997, Justeson and Katz1995] that recognizes generic named entities, such as organizations, people and places, as well as multiword terms and untyped names. TALENT also links together partial references to their full canonical forms, similar to [Aberdeen *et al.*1995], and also reports each term/entity's type and location. TALENT output, post-

```
[foci list]
1.      Jane Jacobs (PERSON)                    290
2.      Toronto (PLACE)                         129
3.      Ideas That Matter (PERSON)              115
4.      Tyler (TERM)                            88
```

Figure 4: Foci found in *Maclean's* article, right column is the calculated measure of importance.

processed to highlight frequency and location information, is shown in Figure 3, for a particular example article from *Maclean's* on the revolutionary urban planner, Jane Jacobs.

4.2 Step 1 : Finding Foci – what is the article about?

Figure 4 shows the topics in the article, as determined by this first stage. "Jane Jacobs" is the prominent focus by a wide margin, and the unnamed entity "Tyler" has been promoted as a focus, substituting for the more frequently occurring person "Dee W. Hock". Let's now examine how these foci were determined.

Once we have the IE output, we need to select the salient *foci*, the specific terms and entities that will appear in the summary. A weighting system was employed that utilizes factors in ranking each entity, including its type, frequency, as well as the centroid and variance of its occurrences. All factors were normalized such that the maximum possible value is 1.0.

TERM/ENTITY TYPE (e.g. whether a term refers to a person) is an important factor, since it is needed to balance the weighting between the different types. This is necessary because some types of terms happen to be topics more often, and others less, as indicated by our weighting in Table 1. For example, PLACES often indicate a setting of a story rather than an actual topic. Similarly, topics like "city planning" are TERMS, but will involve occurrences of organizations and people that actually have less import, so we assign TERM occurences a higher relative weight.

Besides the obvious importance of FREQUENCY, the CENTROID and VARIANCE of occurrences were also selected as factors. The centroid factor models the position metric's [Lin and Hovy1997] judgment that the article's beginning is more important. The variance factor captures the influence of the particular term or entity over the course of the entire ar-

PLACE	PERSON	ORGANIZATION	MULTIWORD TERM	OTHERS
0.6	0.9	0.9	1.0	0.7

Table 1: Weights for Term/Entity Type Bias as indicated by an empirical study of 15 NANTC articles

ticle. Thus, a term that occurs early on in a document and continues throughout is more salient than one that is only mentioned at the beginning. Weights for all factors were then established by an empirical study of 15 *long* articles, selected from the North American News Text Corpus from the LDC [Consortium1997] that yielded good results.

FREQUENCY	TERM/ENTITY TYPE	VARIANCE	CENTROID
24.0	5.0	3.0	2.0

Table 2: Weights for each factor type in determining salient foci

As one can see from Table 2, the most important factor by far was FREQUENCY, which agrees with the general literature. Newly correlated with topicality was TERM/ENTITY TYPE, which is significant in determining the ranking of the extracted information. CENTROID and VARIANCE also helped, but only had marked effects in instances where the number of occurrences were relatively low.

4.3 Step 2 : Questioner – what information should we extract?

Once the algorithm has chosen the article's foci, we analyze their types to determine what possible questions might be answered in the text. Figure 5 shows some questions that we might expect to be answered from the example article.

We can enumerate these questions since each focus type has particular properties that may be expanded in the text. The concept is similar to template based approaches, but with the unique difference that we are working with generic entities: people, places, terms and organizations appear in texts across all domains. Take an identified focus of the *Maclean's* article, "Jane Jacobs", a PERSON, for instance. Figure 6 lists the unary relations we can expect to find for all PERSON foci, which are determined *a priori*: her identity, her age, and possibly what she said or did.

More interesting questions can be asked by examining the relationships between different pairs of focus types, shown in Figure 7. This corresponds

```
[questions]
1.      Does Jane Jacobs live in Toronto?
2.      Did Jane Jacobs visit Toronto?
3.      Does Ideas That Matter live in Toronto?
4.      How old is Jane Jacobs?
5.      What did Jane Jacobs say, if anything?
6.      What did Jane Jacobs do, if anything?
7.      Does the story occur in Toronto?
...
16.     What did Ideas That Matter say or do, if anything?
```

Figure 5: Some questions posed for the *Maclean's* article

Focus Type (X)

1. Person	2. Organization	3. Place	4. Multiword Term
A. Who is X? B. What did X say? C. What did X do? D. How old is X?	E. Is X a nonprofit or governmental or corporate agency? F. What did X say? G. What did X do?	H. Is X the setting of the story? I. Is X the governing agency of X?	J. What does X mean? K. Are there synonyms for X?

Figure 6: Some unary relations found in texts

roughly to the notion of *named relations*, in which two foci enter into a defined relationship. This differs from all template based approaches because the target information is decided *dynamically*, based on the specific pairs of focus types that are found in the article.

As a starting point for finding the answers to these questions, all sentences containing focus occurrences are retrieved and are used as basis for the summary, the list of sentences to be pruned by the later stages. Even at this early stage, by limiting the sentence extraction to foci, we guarantee that the system can select sentences that are both topical and tightly bound, which fulfills the first task of IE integration as stated in the introduction.

4.4 Step 3 : Answerer – Identifying sentences and phrases that contain needed information

Now that we have focused on a small group of sentences, we can escalate the amount of effort used to analyze them. We parse them to find the grammatical relationships (as in Boguraev and Kennedy, 1997), by passing each of the extracted sentences to IBM's English Slot Grammar [McCord1990].

Focus Type (Y) \ Focus Type (X)	Person	Organization	Place	Multiword Term
Term	a. X developed Y b. X uses Y c. X does Y d. X is a type of Y	m. X developed Y n. X uses Y o. X does Y p. X is a Y	u. Y developed at X v. X is a type of Y	y. X and Y are synonyms/ antonyms z. X and Y are hypernyms/ hyponyms
Place	e. X lives in Y f. X visited Y g. X is now in Y	q. X is located at Y r. X is interested in Y	w. X and Y are adjacent x. X and Y are subparts/ superparts of each other	
Organ- ization	h. X works for Y i. X heads Y j. X is against/supports Y's policies	s. X and Y are allies/ competitors t. X is a subsidiary of Y		
Person	k.X and Y are friends/ enemies l. X or Y work for each other			

Figure 7: Generic binary relationships between two focus types

Figures 9 and 10 shows the patterns that we attempt to detect to determine whether a sentence or phrase answers a question involving a single focus type (hereafter, *unary* relation) or a pair of focus types (binary relation). Unary relationships are simpler to detect, since they use shallower features than the binary ones. As to be expected, verb analysis [Levin1993, Klavans and Kan1998, Dang *et al.*1998] plays a large role in the detection of binary relationships.

Most importantly, note that binary relationships are typically filled by sentence level constituents, whereas unary relations are more frequently filled by noun appositives and relative clauses. In fact, this is such a general division between unary and binary relations that we can cast it as a rule, summarized as:

Relationship with one focus (Unary) == Phrase unit that attaches to focus' occurrence == Can be inserted

Relationship with two foci (Binary) == Sentence unit with relation as matrix verb == Can be reordered

In our current prototype, we have implemented detection methods for the unary and binary relationships indicated by the asterisks in the two figures.

```
[answers]
1.      Binary: [20].                      Does Jane Jacobs live in Toronto?
2.      No evidence.                       Did Jane Jacobs visit Toronto?
3.      No evidence.                       Does Ideas That Matter live in Toronto?
4.      Unary: [2].                        How old is Jane Jacobs?
5.      Unary: [57].                       What did Jane Jacobs say, if anything?
6.      Unary: [3 17 87].                  What did Jane Jacobs do, if anything?
7.      Unary: [47].                       Does the story occur in Toronto?
...     ...                                ...
16.     Binary: [29 (Jane Jacobs)].        What did Ideas That Matter say or do, if anything?
```

Figure 8: Questions answered for the *Maclean's* article. Sentence numbers indicated in brackets

Entity or Term Type (X)

1. Person	2. Organization	3. Place	4. Multiword Term
A. look for appositive relative clause, or "be" verb* B. look for communication* verbs* C. look for action verbs* D. look for appositive number*	E. (same as A)* adding cue phrases for each type F. (same as B)* G. (same as C)*	H. look for byline structure and object position occurences** I. look for active verbs	J. (same as A)* or by "defining" verbs K. try other terms with same head or with different word order

Figure 9: Detecting Unary Relationships (*=implemented, **=partially implemented)

4.5 Step 4 : Content Ordering – Putting it all together

To assess which answers are worthwhile enough to put in the summary, we reuse the focus importance scores derived earlier in Section 4.2. Two heuristics govern which answers are selected for inclusion into the final summary: 1) binary relations take precedence over unary ones and 2) relationships involving the first occurrence of the focus' canonical form are favored over ones using variant forms.

To order the sentences and phrases in a way that is meaningful, we first lay out the sentence level (binary) units, and then merge in the phrasal (unary) components at appropriate locations.

- **Sentence ordering.** Problems often occur in sentence extraction approaches when sentences are presented in the preserved order of the original story, that can result in false implicatures and unexplained references without the intervening material. To deal with this problem,

Focus Type (X)

Focus Type (Y)	Person	Organization	Place	Term
Term	a. find "make, develop" verbs* b. find "use" verbs* c. find "perform, do" verbs* d. find Y as appositive of X	m. (same as test a)* n. (same as test b)* o. (same as test c)* p. (same as test d)*	u. (same as test a)* v. (same as test d)*	y. find X as appositive description of Y, or in parenthetical expression z. check subsumption of X's words in Y in appositive
Place	e. find "reside" verbs* f. find "go, visit" verbs* g. find "current" time modifiers	q. find passive "locate" verb, Y as address form, or (test e)* r. <unspecified as of now>	w. coordinated X Y* x. X superordinates Y*	
Organ-ization	h. find "work" verb. can be in appositive form* i. find officer titles in appositive or "lead" verbs** j. find "conflict/support" verbs*	s. look for coordination* t. look for appositive or "sub/super" modifier**		
Person	k. find coodination (and test j) l. find coordination (and test h)			

Figure 10: Patterns for Detecting Binary Relationships (*=implemented, **=partially implemented)

we order our extracted sentences in such a way to enable smooth transitions between foci. Starting with the focus of highest importance, the sentences that represent binary relationships are ordered such that each references a focus from the previous one. We traverse the foci in the order of rank importance, as established in section 4.2, resulting in a skeleton summary.

- **Partial reference resolution and phrase insertion.** When the skeleton summary is completed, we correct misreferences for any focus. The initial reference to each focus is checked to see whether the canonical form is used; if not, the reference is replaced. This guarantees that the introduced form is in its normal form. Integrating any unary relationship phrases is the final step in constructing the summary. Appositives, relative clauses and other unary relations are attached to instances of the focus in subject position, whenever possible. These insertions have a minimal chance of creating discordancy in the text, since they are unary relationships that only involve the particular focus being described. Taken together, partial reference resolution and phrase insertion are used to merge identified missing information into the summary.

In theory, each focus' importance and its grammatical roles in the document's sentences are all we would need to determine the resulting summary order. However, proximal sentences are often dependent on each other for referencing and causal relationships. To account for this cohesive effect of adjacent sentences, any pair of sentences that originally occurred close together (defined as within 3 sentences and in the same paragraph) are grouped together as a unit and are internally positioned in their original order. Conceptually, this is similar to passage retrieval [**?**].

The completed summary for the example, with the relationships enumerated, is shown below:

Original Sentence Number	Relations Represented (uppercase indicates unary relations; lowercase are binary relations)	Text of Sentence
1	b. Jane Jacobs ⇒ Tyler J. Tyler D. Jane Jacobs	The late afternoon sun filters through the autumn leaves and rests for a gentle moment on Jane Jacobs's face as Jacobs, **at 81**, savors a mouthful of Tyler pudding, a concoction of eggs, granulated sugar, milk and a little flour baked in a pie crust.
20	e. Jane Jacobs ⇒ Toronto	Jacobs works where she lives in a three-storey brick house in Toronto's Annex area, a tree-lined residential pocket on the edge of the University of Toronto and half a block from the hurly-burly of Bloor Street.
12	k. Jane Jacobs ⇒ Ideas That Matter C. Ideas That Matter C. (upgraded unary from sent 18) Ideas That Matter ⇒ Jane Jacobs	Billed as an "International gathering to create and share knowledge," Jane Jacobs: Ideas that Matter, **who may be turning Jacobs into a celebrity**, began Sept. 20.
54	C. (upgraded unary) Jane Jacobs ⇒ Toronto	By writing a newspaper article castigating city planners for attempting to "Los Angelize" Toronto, "the most hopeful and healthy city in North America, still unmangled, still with options," Jacobs galvanized a group of local citizens into forming Stop Spadina, a protest in which Jacobs played a major role as a political strategist.

5 Evaluation

5.1 Experimental design

We evaluated the entire implemented summarization system (the outer box in figure 2), which has the focus based summarization algorithm at its core.

To judge the implemented system's performance, we performed a ranking evaluation that tests our performance against two other systems. This type of evaluation nicely avoids the difficult problem of having to produce canonical summaries from human subjects [Jing *et al.*1998] and having to reconcile different, but equally ideal summaries. We chose the lead based method and a TF*IDF based method as the two competing techniques.

We collected a set of ten new *long* test articles for the evaluation, which we partitioned into two sets of five. Because of time limitations, we designed the experiment to take each subject through only one of the two test article sets. Some subjects that had more time were asked to perform the summarization evaluation on both sets. The order of the articles was randomized, and each article was passed to each of the three summarization engines to produce a short, four sentence summary. The three summaries were shown to the human subjects in a random order on a single screen. Subjects were asked to rank the summaries in order of best to worst for informativeness and fluency. The original article was accessible to them via a hyperlink. A total of 38 subjects evaluated one set of articles. Because of randomization and several subjects terminating the study early, we ended up having between 13 and 19 judgments per article.

5.2 Results

To assess whether there was any significance, we used a non-parametric test, Friedman Analysis of Variances (χ_r^2). The evaluations show very little statistical significance bias toward any of the three summary types, as marked by asterisks in table 5.2. In fact, in the few cases in which statistical significance was reached, the simplest lead based method was usually favored.

Informativeness			Fluency		
Question No.	χ_r^2	Best Summary Type	Question No.	χ_r^2	Best Summary Type
1	2.0	TF*IDF	1	17.2	TF*IDF
2	16.1	Lead	2	27.1*	Lead
3	0.4	Foci	3	13.6	TF*IDF
4	1.2	Lead	4	18.5*	TF*IDF
5	2.0	TF*IDF	5	10.5	TF*IDF
6	4.5	Lead	6	4.6	Foci
7	17.2*	Lead	7	7.0	Foci
8	15.9	Lead	8	14.2	Foci
9	10.3	Lead	9	4.1	None
10	4.2	TF*IDF	10	2.9	TF*IDF

Table 5.2: χ_r^2 results, ($* = p < .05$)

6 Discussion and future work

Unfortunately, the results are inconclusive. Comments from the evaluators included that the experimental design could have been controlled for different levels of domain expertise, and that using a task-based evaluation scheme may have been more appropriate. Many subjects also felt that reading the full article was a requirement to correctly judge the adequacy of the summary, but time constraints forced many to skip or only skim the long source articles. Evaluation of summarization systems is an open problem, but progress towards a standard is being pursued. We feel that a redesign of the experiment is needed before any conclusions can be drawn.

Our question answering model currently relies on sentence extraction as a first approximation. We are working on improving it such that will extract only the answers to the questions, bypassing sentence extraction entirely.

7 Conclusions

In this paper, we have shown how incorporating an analysis of a document's named entities and terms can produce a more informative summary by: 1) improving information selection, 2) highlighting relationships between article foci, 3) identifying descriptions of each foci that can be added. These contributions make fundamental progress in realizing a robust, domain-independent algorithm for summarizing long texts. Our current efforts are integrating work by Jing [Jing1999], which will help us further refine the coherence and conciseness of the summary.

References

[Aberdeen *et al.*1995] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain. MITRE: Description of the Alembic System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 141–155. Morgan Kaufmann Publishers, Inc., Columbia, Maryland, USA, November 1995.

[Barzilay *et al.*1999] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, 1999.

[Brandow *et al.*1995] Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.

[Consortium1997] Lingustic Data Consortium. North American News Text Corpus. CD-Rom, October 1997.

[Dang *et al.*1998] Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenweig. Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of COLING/ACL 98*, pages 293–299, Montral, Qubec, Canada, August 1998.

[DeJong1982] Gerald DeJong. An Overview of the FRUMP System. In W. G. Lehnert and M. H. Ringle, editors, *Strategies for Natural Language Processing*, pages 149–176. Erlbaum, Hillsdale, NJ, 1982.

[Hovy1998] E.H. Hovy. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 1998.

[Jacobs and Rau1990] P.S. Jacobs and L.F. Rau. SCISOR:Extracting Information from on-line news source. *Communications of the ACM*, 33(11):88–97, 1990.

[Jing *et al.*1998] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Summarization Evaluation Methods: Experiments and Analysis. In *The Working Notes of AAAI Symposium on Intelligent Summarization*, Stanford University, CA, March 1998.

[Jing1999] Hongyan Jing. The Decomposition of Human-Written Summary Sentences. Submitted to ACL-99, 1999.

[Justeson and Katz1995] John S. Justeson and Slava M. Katz. Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

[Kan1999] Min-Yen Kan. Layser: a layout parser. Poster at the 1999 AAAI Fall Symposium on Using Layout for the Generation, Understanding or Retrieval of Documents, 1999.

[Klavans and Kan1998] Judith L. Klavans and Min-Yen Kan. Role of Verbs in Document Analysis. In *Proceedings of COLING/ACL 98*, pages 680–686, Montral, Qubec, Canada, August 1998.

[Kupiec *et al.*1995] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *ACM SIGIR*, pages 68–73, 1995.

[Levin1993] Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, Illinois, 1993.

[Lin and Hovy1997] Chin-Yew Lin and Eduard Hovy. Identifying topics by position. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 283–290, Washington, DC, USA, 1997.

[McCord1990] Michael McCord. *English Slot Grammar*. IBM, 1990.

[Paice1990] C. D. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1):171–186, 1990.

[Radev and McKeown1998] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469:500, September 1998.

[Wacholder *et al.*1997] Nina Wacholder, Yael Ravin, and Misook Choi. Disambiguating proper names in text. In *Proceedings of the Applied Natural Language Processing Conference*, March 1997.