SIMPLEX NPS CLUSTERED BY HEAD: A METHOD FOR IDENTIFYING SIGNIFICANT TOPICS WITHIN A DOCUMENT

Nina Wacholder[1]
Center for Research on Information Access
Columbia University
New York, NY 10027
nina@cs.columbia.edu

Abstract

This paper discusses 'head clustering', a novel, linguistically-motivated method for representing the aboutness of a document. First, a list of candidate significant topics consisting all simplex NPs is extracted from the document. Next, these NPs are clustered by head. Finally, a significance measure is obtained by ranking frequency of heads: those NPs with heads that occur with greater frequency in the document are more significant than NPs whose head occurs less frequently. An important strength of this technique is that it is in principle domain-general. Furthermore, the output can be filtered in a variety of ways, both for automatic processing and for presentation to users.

In order to evaluate the head clustering method, an experiment was conducted in which judges were asked to rate three lists as to whether they conveyed a sense of the content of the article. The judges agreed that the list of simplex NPs with repeated heads was more helpful in representing the content of the full document than a list of keywords with a frequency of greater than one or than a list of repeated word sequences.

Introduction

This paper describes a methodology for identifying significant topics in edited documents such as newspaper articles. For the purposes of this paper, a 'topic' is any event or entity explicitly referred to in the document, and a 'significant topic' is a topic central to what is sometimes called the aboutness of a document.

The notion 'significant', like the notion 'relevant', is both task and user dependent. What is significant for an application that answers specific questions is different from what is significant for an application that conveys the sense of particular documents; what is significant in a domain for a naive user may be quite different from what is significant to an expert. The goal of this work is to develop a general technique for identifying the topics referred to in a document and for ranking these topics in terms of their significance. The list can then be filtered in a variety of ways, depending on the requirements of the application.

A fundamental hypothesis of this research is that the head of a common NP makes more of contribution to the document as a whole than do modifiers and should therefore be treated in a way that gives it particular prominence. The intuitive justification for sorting simplex NPs by head is based on the fundamental distinction between heads and modifiers: if, as a practical matter, it is necessary to rank the contribution to the whole made by the set of words constituting an NP, the head is obviously the most important word, both from the point of view of syntax and semantics.

[1] This research was partly supported by NSF grant IRI-9712069, "Automatic Identification of Significant Topics in Domain Independent Full Text", Judith Klavans, PI; Nina Wacholder, co-PI.

The author extends thanks to Judith Klavans, Yael Ravin, David K. Evans, Luis Gravano and Eleazar Eskin for their thoughtful comments and suggestions.

The methodology described in this paper involves collecting a complete list of nominal elements which together constitute a simplified representation of the content of the document. These nominal elements are called simplex NPs. For common NPs (NPs whose head is a common noun (N)), a simplex NP is a maximal NP that includes premodifiers such as determiners and possessives but not post-nominal constituents such as prepositions or relativizers. Examples are *asbestos fiber* and *9.8 billion Kent cigarettes*. Simplex NPs can be contrasted with complex NPs such as *9.8 billion Kent cigarette with asbestos filters* where the head of the NP is followed by a preposition, or *9.8 billion Kent cigarettes sold by the company*, where the head is followed by a participial verb. An important property of these (English-language) simplex NPs is that the phrasal head is the last element.

This technique can be used in IR applications at indexing time. In addition, this method is useful for applications which require shallow language understanding in order to produce output that users will find satisfactory. Examples include:

- summarization or other techniques for conveying the content of a document.
- advanced information extraction where important entities in the document must be identified and linked so that information about the entity from different parts of the document can be merged.
- second stage information retrieval, where a subset of a larger corpus has been determined to be potentially relevant, perhaps by a statistically based system. The subset can then be further filtered in order to identify documents which are likely to be of interest for a particular query or which may provide the answer to a specific question.
- automatic or semi-automatic 'back-of-the-book' indexing of print and electronic texts.

In the next section, related work on methods for determining topic significance is reviewed. Then the problem of choosing candidate significant topics in the context of a particular document is addressed and the choice of simplex NPs as the unit of representation is justified. The method by which LinkIT, a software tool developed at Columbia Uniersity to identify significant topics in domain-independent full text, uses head clustering to identify significant topics is explained in some detail, using a sample newpaper article. [2]

Finally, the head clustering method is evaluated. Judges were asked to evaluate the helpfulness in conveying the content of a document of three lists: 1) a list of simplex NPs extracted from a document by LinkIT; 2) a list of stems which occur in the document more than once; and 3) a list of repeated sequences of words in the sample document. Judges agreed that the LinkIT output was superior.

Related work

In order to identify significant topics in a document, a significance measure is needed, i.e., a method for determining which concepts in the document are relatively important. In the absence of reliable full-scale syntactic parsing, frequency measures are often used to determine significance. One of the earliest statistical techniques for identifying significant topics in a document for use in creating automatic abstracts was proposed by Luhn (1958) who developed a method of making a list of stems and/or words, sometimes called keywords, removing keywords on a stoplist, and then calculating the frequency of the remaining keywords. This method, which is based on the intuition that frequency of reference to a concept is significant, can be usefully used to locate at least some important concepts in full text, especially when frequency of a keyword in a document is calculated relative to its frequency in a large corpus, as in standard information retrieval (IR) techniques (Salton 1989). However, the ambiguity of stems (*trad* might refer to *trader* or *tradition*) and of isolated words (*state* might be a political entity or a mode of being) means that lists of keywords have not usually been used to represent the aboutness of a document to human beings. Instead, techniques such as identifying sentences with multiple keywords have been used since Luhn for automatic creation of abstracts

[2] The LinkIT software is written by David K. Evans..

(Paice 1990).

Recently, the effort to develop techniques for domain-independent content characterization has been addressed by Boguraev and Kennedy (1997). They take as a starting point the question of the applicability to document characterization of the approach of Justeson and Katz (1995) to identifying technical terms in a corpus. Justeson and Katz developed a well-defined algorithm for identifying technical terminology, repeated multi-word phrases such as *central processing unit* in the computer domain or *word sense* in the lexical semantic domain. This algorithm identifies candidate technical terms in a corpus by locating NPs consisting of nouns, adjectives, and sometimes prepositional phrases. Technical terms are defined as those NPs, or their subparts, which occur above some frequency threshold in a corpus.

However, as Boguraev and Kennedy observe, the technical term technique is not simply adaptable to the task of content characterization of documents. For an open-ended set of documents and document types, there is no domain to restrict the technical terms. Moreover, patterns of lexicalization of technical terms in a corpus do not necessarily apply to individual documents, especially short ones. Boguraev and Kennedy therefore propose relaxing the notion of a technical term to include an exhaustive list of "discourse referents" in a wide variety of text documents, and determining which referents are important by some measure of discourse prominence.

With this approach, the concept of technical terms is greatly attenuated. Even in a technical document, technical terms do not constitute a complete list of all of the phrases in a document that contribute to its content, especially since technical terms are by definition multi-word. Moreover, a truly domain-general method should apply to both technical and non-technical documents. The relevant difference between technical and non-technical documents is that in technical documents, many of the topics which are significant to the document as a whole may be also technical terms.

Like the keyword and repeated word sequence methods for measuring topic significance, head clustering is statistical in that it relies on a frequency measure to provide an approximation of topic significance. However, instead of counting frequency of stems or repetition of word sequences, this method counts frequency of a relatively easily identified grammatical element, heads of simplex NPs.

In what follows, the head clustering methodology is described. First, simplex NPs are presented as a practical unit for 'gisting'. Next, these NPs are clustered by head. NPs whose heads have a greater frequency are ranked as being more important than NPs whose heads occur less frequently. In the evaluation of this method, discussed below, the head sorting method of determining topic significance is compared to the purely statistical keyword method and to the repeated word sequence.

Simplex NPs

On the simplifying assumption that nominal elements can be used to convey the gist of a document, simplex NPs, which are semantically and syntactically coherent, appear to be at the right level for content representation of expressions out of the context of the document. For common NPs (as mentioned above), a simplex NP is a maximal NPs that includes premodifiers such as determiners and possessives but not post-nominal constituents such as prepositions or relativizers. For proper names, a simplex NP is a name that refers to a single entity. For example, *Museum of the City of New York*, the name of an organization, is a simplex NP even though the organizational name incorporates a city name.

When a word is presented in isolation, the structural information provided by the ordered juxtaposition of the words that combine with it to form a meaningful unit is lost, as in the distinction between *unit* and *central processing unit*. This information may not be important in large scale information retrieval systems, but it is important to people.

On the other hand, a list of all of the nominals in a document is impractical because it is bulky and repetitive, in part because of embedding. For example, in the 115 word excerpt in Figure 1 (Wall Street Journal 0003, Penn Treebank) 37 Ns are italicized, and 43 NPs and 5 pronouns are bracketed.[3]

[3] The full text of this article is in Appendix A.

[[[A *form*] of [*asbestos*]] once used to make [Kent *cigarette filters*]] has caused [[[[a high *percentage*] of [cancer *deaths*]] among [[a *group*] of [*workers]]*] exposed to [it] more than [30 *years*]] ago, [*researchers*] reported.

[[The *asbestos fiber*], [*crocidolite*]], is unusually resilient once [it] enters [the *lung*], with [[even brief *exposures*] to [it]] causing [[*symptoms*] that show up [*decades*] later], [*researchers*] said. [[*Lorillard Inc.*], [[the *unit*] of [New York-based *Loews Corp.*]] that makes [*cigarettes*]], stopped using [*crocidolite*] in [[its] *Micronite cigarette filters*] in 1956.

Although [preliminary *findings*] were reported more than [a *year*] ago, [the latest *results*] appear in today's [[[*New England Journal of Medicine*], [[a *forum*] likely to bring [new *attention*] to [the *problem*]]].

Figure 1

Compared to simplex NPs, complex NPs (e.g., *symptoms that crop up decades later*) are difficult to identify by automatic means and are also difficult for people to interpret, especially out of context. For example, the expression *information about medicine for babies* is ambiguous: in [[information about medicine] [for infants]], the information is for infants; in [information about [medicine for infants]], the medicine is for infants.

In contrast, simplex NPs form a coherent unit, with less structural ambiguity. Furthermore, simplex NPs can be relatively reliably extracted by a finite state grammar from text that has been tagged with part-of-speech by a state-of-the-art system. Figure 2 shows the simplex NPs extracted by LinkIT from the excerpt in Figure 1.

S1 1-2 (1) A form
S1 4-4 (2) asbestos
S1 9-11 (3) Kent cigarette filters
S1 14-16 (4) a high percentage
S1 18-19 (5) cancer deaths
S1 21-22 (6) a group
S1 24-24 (7) workers
S1 30-31 (9) 30 years
S1 33-33 (10) researchers
S2 35-37 (11) The asbestos fiber
S2 38-38 (12) crocidolite
S2 45-46 (14) the lungs
S2 48-50 (15) even brief exposures
S2 54-54 (17) symptoms
S2 56-56 (18) show
S2 58-58 (19) decades
S2 60-60 (20) researchers
S3 62-63 (21) Lorillard Inc.
S3 64-65 (22) the unit
S3 67-71 (23) New York-based Loews Corp.
S3 74-75 (24) Kent cigarettes
S3 78-78 (25) crocidolite
S3 81-83 (27) Micronite cigarette filters
S3 87-88 (28) preliminary findings
S3 93-94 (29) a year
S3 96-98 (30) the latest results
S3 101-101 (31) today 's
S3 102-106 (32) New England Journal of Medicine
S3 107-108 (33) a forum
S3 112-113 (34) new attention
S3 115-116 (35) the problem

Figure 2

The list of simplex NPs in Figure 2 was created by LinkIT, a tool developed at Columbia University to identify significant topics in domain-independent full text. The input to LinkIT is text which has been pre-processed and tagged with part-of-speech by Mitre's publicly available *Alembic* Workbecn (Aberdeen et al. 1995). One of LinkIT's components is a finite state grammar which extracts simplex NPs. As it processes the text, LinkIT stores the sentence number and token span of simplex NP, and assigns it a unique identifier reflecting the order in which it appeared in the document.

However, identifying all of the simplex NPs in a document is still not adequate for conveying the gist of a document because not all candidate significant topics are in fact significant. An assumption underlying Justeson and Katz' notion of technical terms is that technical terms have distinguished usage in some domain. This is distinctly not the case for the complete list of simplex NPs in a document. For a list to be useful, additional filtering is needed.

Head clustering

The intuitive justification for sorting simplex NPs by head is based on the fundamental linguistic distinction between head and modifier: a head makes a greater contribution to the syntax and semantics of a grammatical constituent than does a modifier. This linguistic insight can be extended to the document level: if, as a practical matter, it is necessary to rank the contribution to a whole document made by the sequence of words constituting an NP, the head is more important than the other words in the phrase. A variation of this observation has been recognized by Strzalkowski (1997) and others, who have used the distinction between heads and modifiers for query expansion. In this section, we propose using the head-modifier distinction to determine concept significance.

Since simplex NPs have been defined so that the head is always the last element, the first step in LinkIT's processing of the list of simplex NPs is to rank them by frequency of head. The Ns that occur as heads of simplex NPs three times or more in wsj_0003 are listed in Figure 3.

workers (9)	*asbestos (8)*	*filter(s) (8)*
researchers (8)	*fiber (5)*	*crocidolite (5)*
factory(4)	*years (4)Talcott (4)*	
deaths (3)	diseases(3)	cigarettes (3)

Figure 3

By itself, Figure 3 is a simple representation of the content of the document. To allow the reader to make an independent judgement, the full text of this article appears in Appendix A. This list, in combination with the structural information discussed above, can be used at indexing time for IR applications.

The document as a contained world

The challenge in preparing an abbreviated representation of an article is to identify heuristics which make it possible to represent to the user the sense in which in which an author used an expression in the document, without performing full sense disambiguation. In an important sense, every document can be viewed as forming its own 'self-contained' world. A document is written to get across a particular idea or set of ideas. The task of the author, at least in documents intended for public distribution, is to convey to the reader what general knowledge is assumed and to inform the reader of the context so that ambiguous expressions can be easily identified. These references are governed by certain standard conventions.

For example, in an edited document such as a newspaper article, the first reference to a named entity such as a person, place or organization typically uses a relatively full form of the name in a version which is sufficient to disambiguate the reference for the expected audience. Later in the document, the same entity is usually referred to by a shorter, more ambiguous form of the name (Wacholder and Ravin 1997). An article might first refer to Columbia University or, (more formally) Columbia University in the City of New York, and later refer only to Columbia. Without the initial disambiguating reference, Columbia by itself is quite ambiguous. It might be a city (Columbia, MD), a bank (Columbia Savings and Loan) or one of many other entities. Nominator, a module which identifies proper names developed at the IBM TJ Watson Research categorizes them, and links expressions in the same document which refer to the same entity successfully exploited this property of documents (Wacholder et al. 1997). Nominator first builds a list of proper names in each document and then applies heuristics in order to link names which refer to the same entity (e.g., *Hillary Clinton* and *Ms. Clinton*, but not *Bill Clinton*). This technique produces reliable links between references to the same entity in a document.

Common NPs also manifest a pattern of referential linking in documents, although it is more subtle and complicated than the proper name behavior. Any article of more than minimal length contains repeated references to important concepts. In general, when a word appears as a head of an NP in a document, it is used in the same sense throughout the document, especially in articles of newspaper length. Some of the references to the head are elliptical and therefore very ambiguous, at least out of context, but some of the references are usually fuller and therefore more specific and more informative. For example, in the *Wall Street Journal* article that is used as the primary example throughout most of this paper, the most frequent head of simplex NPs is the *workers*. Six of the nine references to *workers* are not preceded by an adjective or noun which delimits the intended sense of workers; however, one of these references is to the more specific *asbestos workers*. The different references to a concept implicitly or explicitly refer to each other and collectively form an abstract construct that conveys the sense that the author (presumably) intended to convey. (See Kameyama (1997) for a discussion of the importance of establishing all referential links within a document for information extraction applications, so that information about these entities can be merged.) When simplex NPs are clustered by head, NPs with the same head are likely to refer to the same concept, if not to the same entity. For example, in the sentence "Those worker got a pay raise but the other workers did not", the same sense of *worker* is used in both NPs, but the workers referred to are different.

Sorting by final word of the name is a simplification for proper names which, in contrast to common NPs, do not have a head in the sense that there is a single word which is semantically and semantically the most important. However, clustering proper names by the final word, as if it were the head, is satisfactory for certain kinds of proper names, including human ones. For example, Talcott can reasonably be considered the head of both *James Talcott* and *Dr. Talcott*. We are currently in the process of refining the head clustering procedure to handle organizations and other categories of proper names that have different naming conventions (Wacholder et al. 1997). In contrast to common NPs and proper NPs, reference to concepts in the form of pronominal anaphors contribute no references to new entities and therefore will not be discussed in this paper.

The head clustering technique provides a way to situate the entities referred to in the document in the context of related entities so their sense is comprehensible to users who have not actually read a document. The full list of simplex NPs which have these heads appears in Figure 4. Examination of this list suggests that it provides a more explicit representation of the content of the article than does the list in Figure 3.

(48) the workers
(83) workers
(100) the workers
(104) any asbestos workers
(144) 160 workers
(152) Workers
(161) Workers
(169) those workers
(2) asbestos
(41) asbestos
(43) no asbestos
(67) asbestos
(115) asbestos
(118) asbestos
(141) asbestos
(143) cancer-causing asbestos

(3) Kent cigarette filters
(27) Micronite cigarette filters
(70) the filters
(72) filter
(74) the filters
(112) the cigarette filters
(147) the Kent filters
(160) filters

(10) researchers
(20) researchers
(47) the researchers
(61) researchers
(92) the researchers

(1) The asbestos fiber
(28) needle-like fibers
(35) More common chrysotile fibers
(57) acetate fibers
(58) the dry fibers

(12) crocidolite
(25) crocidolite
(116) crocidolite
(129) crocidolite
(151) the crocidolite
(103) paper factory
(145) a factory
(150) the factory
(165) the factory

(9) 30 years
(29) a year
(39) years
(178) 35 years

(56) James A. Talcott
(59) Dr. Talcott
(97) Dr. Talcott
(122) Dr. Talcott

(5) cancer deaths
(88) 18 deaths
(99) lung cancer deaths

(84) asbestos-related diseases
(96) asbestos-related diseases
(171) asbestos-related diseases

(24) Kent cigarettes
(51) the Kent cigarettes

(73) 9.8 billion Kent cigarettes

Figure 4

For example, *filter* is the head of eight simplex NPs. Four of these have adjective and nominal premodifiers: *Micronite cigarette filters*, *Kent cigarette filters*, *the cigarette filters* and *Kent filters*. In the absence of other references to specific kinds of filters, the correct and accurate generalization is that the kinds of filters discussed in this document are cigarette filters, rather than coffee filters or oil filters. *Asbestos workers* and *cancer-causing asbestos*, the most specific NPs with the head *workers* and *asbestos* respectively, as measured by number of content words preceding the head, accurately characterizes the property of the workers and of asbestos that is most important for this document. Similarly, the most specific simplex NP suggests that the type of factory under discussion is a paper factory.

For the head *fiber*, there are five different premodifiers. While it is impossible to determine from the list here which of these types of fiber are the same and which are different, the variety of premodifiers suggests that types of fibers are being discussed in this document.

For *researchers* and *crocidolite*, this technique provides no further specification of information, but merely does the same thing that a count of the occurrence of these strings in the document would yield, along with the additional information that these words are repeatedly used as heads in the document and therefore are more likely to be candidate significant topics than a word like Kent which is used five times, but only as a modifier.

Evaluation

The technique proposed in this paper is a general purpose one that can be used in a variety of ways to identify significant topics in a document. In the long run, the practical value of this technique will be judged by its utility in NLP applications such as run-time indexing for information retrieval, automatic summarization and back-of-the-book indexing.

However, an initial evaluation, as well as useful suggestions for refining the technique, has been obtained from human users. In the evaluation, three articles were presented to five individuals; none had any experience in NLP (though one was a professional librarian and indexer). The judges uniformly ranked the list of clustered NPs was ranked most highly, with an average rating of 3.15; the keyword list was ranked second, with an average rating of 2.45, and the list of word sequences was ranked last, with an average rating of 1.92.

The evaluation was conducted as follows. In order to determine whether the list of significant topics output by the headed clustering technique conveys the sense of a document, judges were asked to compare it to two other kinds of output: a list of keywords of frequency of more than one and a list of repeated sequences of words. The keyword list was chosen because it has become a standard in many NLP applications. It therefore establishes a baseline for comparison, even though lists of keywords are not generally used to represent document aboutness. The list of repeated word sequences is similar in its use of repeated phrases, except that it uses a variation of relies notions of technical terms and technical prominence rather than on repeated heads.

The evaluation was conducted as follows. A list of simplex NPs clustered by head was output by LinkIT for each of the three articles. The list included all clusters whose head occurred in the document as a head of more than one simplex NP; duplicates were removed. For wsj_0003, 10 of the 32 simplex NPs considered significant because their heads occurred more than once are shown in Figure 5. The number in brackets is the frequency of occurrence.

workers [4]
the workers [2]
any asbestos workers [1]
160 workers [1]
those workers [1]

asbestos [7]
no asbestos [1]
cancer-causing asbestos [1]

Kent cigarette filters [1]
Micronite cigarette filters [1]
the filters [2]
filter [1]
the cigarette filters [1]
the Kent filters [1]
filters [1]

Figure 5

The keyword list was produced from the list of term frequency produced for wsj_0003 by the SMART system; keywords that occurred in the document only once were removed. The ten most frequent keywords (out of a total of 32) are shown in Figure 6.

asbesto [14]
work [11]
filt [8]
canc [7]
research [6]
cigaret [6]
make [5]
lorillard [5]
kent [5]
fibe [5]

Figure 6

The list of repeated sequences of words was output by *termer*, an implementation by Min Yen-Kan of Katz and Justeson's technical term algorithm. All word sequences which occurred more than once in the document were listed and capitalization was added where appropriate. There were six repeated word pairs in this document, all of which are listed in Figure 7.

Kent cigarette [4]
cigarette filter [3]
Dr. Talcott [3]
cancer death [2]
lung cancer [2]
U.S. [2]

Figure 7

Although care was taken to make the lists as equivalent as possible, not all the differences could be balanced out while maintaining faithfulness to the reliability of the method. For example, the repeated sequence method consistently produces a shorter list than does the clustered NP technique, and keyword technique produces the longest list.

The judges were asked to study the three lists, compare them to each other and to the text of the article and then rank each one on a scale of 1 to 5, where 1 indicated that the list provided no idea of the content of the article and 5 indicated that the list provided an excellent idea of the article content. The results of the evaluation are shown in Figure 8.

AVERAGE LIST RANKINGS			
	Clustered	Keywords	Repeated
article	NPs		sequences
wsj_0003	3.45	2.49	1.8
wsj_0013	2.85	2.2	1.9
wsj_0015	3.15	2.65	2.05
summary	3.15	2.45	1.92

Figure 8

The judges' preference for keywords over technical terms was surprising, given the claim made above that phrases are more informative than keywords. However, in informal discussion, judges confirmed that the coherent expressions in the clustered NP list and the repeated word sequence list were more meaningful than the stems and isolated words in the keyword list. However, the fact that the repeated sequence list was significantly shorter than the other two made it less helpful than the other two and was responsible for the relatively low scores that this list received. This suggests that better results might be obtained from a list in which the list of clustered NPs is further filtered to include only simplex NPs with content-bearing modifiers; for example, instead of the simplex NPs whose head is *workers* shown in Figure 7, only *asbestos workers* would be listed.

This evaluation suggests that the head clustering method does in fact produce a set of plausible signficant topics.

Summary

In conclusion, it appears that the head clustering technique is a promising one for a variety of applications. Moreover, since head clustering has a grammatical basis, the method discussed in this paper is in principle domain general. In fact, the code for recognizing simplex NPs in Wall Street Journal articles did not have to be modified in order to handle abstracts of National Science Foundation grant applications, a quite different genre and domain than newspaper articles. The method described in this paper therefore merits further study.

We plan to take this research in several directions. First we are exploring the applicability of head clustering to other types of documents and to documents that are longer than newspaper articles or proposal abstracts. Second, we are undertaking qualitative and quantitative analysis of the significant topics identified by the method described in this paper and evaluation of their usefulness, in comparison with other techniques for identifying significant topics. Finally, LinkIT output is being used in a variety of research applications.

REFERENCES

Aberdeen, J., J. Burger, D. Day, L. Hirschman, and M. Vilain (1995) "Description of the *Alembic* system used for MUC-6". In *Proceedings of MUC-6*, Morgan Kaufmann. Also, Alembic Workbench, http://www.mitre.org/resources/centers/advanced_info/g04h/workbench.html.

Boguraev, Branimir and Christopher Kennedy (1997) "Technical terminology for domain specification and document characterization". In *Information extraction: A multidisciplinary approach to an emerging information technology*, edited by Maria Teresa Pazienza, pp. 73-96. Lecture Notes in Computer Science Series, Springer-Verlag, Berlin.

Hirschman, Lynette and Marc Vilain (1995) "Extracting Information from the MUC", ACL 95 tutorial.

Justeson, John S. and Slava M. Katz (1995) "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering* 1(1):9-27.

Kameyama, Megumi "Recognizing referential links: an information extraction perspective" cmp--lg/9707009.

Kennedy, Christopher and Branimir Boguraev (1996) "Anaphora for everyone: pronominal anaphora resolution without a parser". In *Proceedings of COLING-96*, Copenhagen, Denmark.

Luhn, H.P. (1958) "The automatic creation of literature abstracts", *IBM Journal of Research and Development*, 2(2):159-165.

Mani, Inderjeet, T. Richard Macmillan, Susann Luperfoy, Elaine P. Lusher and Sharon J. Laskowski (1995) "Identifying unknown proper names in newswire text", *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, MA.

Paice, Chris D. (1990) "Constructing literature abstracts by computer: techniques and prospects". *Information Processing & Management* 26(1):171-186.

Palmer, David D. and David S. Day (1997) "A statistical profile of the Named Entity Task". In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp.194-202, Morgan Kaufmann Publishers.

Penn Treebank. Wall Street Journal, 1988. Treebank, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Radev, Dragomir and Kathleen R. McKeown (1997) "Building a generation knowledge source using Internet-accessible newswire", *Proceedings of the ANLP*, ACL, Washington, DC.

Salton, Gerald (1989) *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA.

Strzalkowski, Thomas (1997) "Building effective queries in natural language information retrieval", *Proceedings of the ANLP*, ACL, Washington, DC., pp.299-306.

Wacholder, Nina, Yael Ravin and Misook Choi (1997) "Disambiguation of proper names in text", *Proceedings of the ANLP*, ACL, Washington, DC., pp. 202-208.

Appendix A: Complete text of wsj 0003

A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago, researchers reported.

The asbestos fiber, crocidolite, is unusually resilient once it enters the lung, with even brief exposures to it causing symptoms that show up decades later, researchers said. Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

A Lorillard spokeswoman said, "This is an old story. We're talking about years ago before anyone heard of asbestos having any questionable properties. There is no asbestos in our products now."

Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes. "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute. Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University.

The Lorillard spokeswoman said asbestos was used in "very modest amounts" in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956. From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold, the company said.

Among 33 men who worked closely with the substance, 28 have died -- more than three times the expected number. Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer. The total of 18 deaths from malignant mesothelioma, lung cancer and asbestosis was far higher than expected, the researchers said.

"The morbidity rate is a striking finding among those of us who study asbestos-related diseases," said Dr. Talcott. The percentage of lung cancer deaths among the workers at the West Groton, Mass., paper factory appears to be the highest for any asbestos workers studied in Western industrialized countries, he said.

The finding probably will support those who argue that the U.S. should regulate the class of asbestos including crocidolite more stringently than the common kind of asbestos, chrysotile, found in most schools and other buildings, Dr. Talcott said.

The U.S. is one of the few industrialized nations that doesn't have a higher standard of regulation for the smooth, needle-like fibers such as crocidolite that are classified as amphobiles, according to Brooke T. Mossman, a professor of pathlogy at the University of Vermont College of Medicine. More common chrysotile fibers are curly and are more easily rejected by the body, Dr. Mossman explained.

In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.

About 160 workers at a factory that made paper for the Kent filters were exposed to asbestos in the 1950s. Areas of the factory were particularly dusty where the crocidolite was used.

Workers dumped large burlap sacks of the imported material into a huge bin, poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used to make filters. Workers described "clouds of blue dust" that hung over parts of the factory, even though exhaust fans ventilated the area.

"There's no question that some of those workers and managers contracted asbestos-related diseases," said Darrell Phillips, vice president of human resources for Hollingsworth & Vose. "But you have to recognize that these events took place 35 years ago. It has no bearing on our work force today."