Applied Artificial Intelligence, 19:119–136, 2005

ISSN: 0883-9514 print/1087-6545 online
DOI: 10.1080/08839510590887441

Taylor & Francis
Taylor & Francis Group

TOPIC DETECTION OF UNRESTRICTED TEXTS: APPROACHES AND EVALUATIONS

Yllias Chali □ *Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, Alberta, Canada*

□ *Topic detection and tracking refers to automatic techniques for locating topically related cohesive paragraphs in a stream of text. Most documents are about more than one subject, but many Natural Language Processing (NLP) and Information Retrieval (IR) techniques implicitly assume documents have just one topic. Even in the presence of a single topic within a document, the document may address multiple subtopics and various aspects of the primary topic. Hence, dividing documents into topically coherent units and discovering their topic might have many uses. We describe new clues that account for the topic of grouping of contiguous portions of the text. Those clues are based on general lexical resources, which make them applicable to unrestricted texts, and can have many uses such as helping users find answers to general questions in an information search task, or in question/answering systems, or in text summarization. We devise an algorithm for identifying these clues, and we report on the performance of these clues, as well as the improvements suggested by our experiments.*

Topic detection and tracking refers to automatic techniques for locating topically related cohesive paragraphs in a stream of text. The problem of topic detection and tracking involves five major tasks: (1) text segmentation: detect changes between topically cohesive sections; (2) topic tracking: keep track of texts similar to a set of example texts; (3) topic detection: build clusters of texts that discuss the same topic; (4) first text detection: detect if a text is the first text of a new, unknown topic; and (5) link detection: detect whether or not two texts are topically linked.

We devise an algorithm for detecting the topic of unrestricted texts based on an efficient use of lexical cohesion. The algorithm is based on the assumption that "the most often similar words are repeated in the text,

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) research grant and the Alberta Heritage Foundation for Science and Engineering Research under the Research Excellence Envelope funding.

Address correspondence to Yllias Chali, Department of Mathematics and Computer Science, University of Lethbridge, 4401 University Drive, Lethbridge Alberta, T1K 3M4, Canada. E-mail: chali@cs.uleth.ca

the more topical they are." Word similarities are detected using common lexical knowledge, and include a process of word sense disambiguation by eliminating those senses of the word that contribute less to the strength of the topic. The disambiguation process is indeed akin to the process of building the set of similar words. Word sense disambiguation method makes use of the information provided by machine readable dictionaries (McRoy 1992; Cowie et al. 1992; Miller et al. 1994; Agirre and Rigau 1997; Li et al. 1995; Mihalcea and Moldovan 1999b; 1999a), and heuristics (McRoy 1992; Bruce and Wiebe 1994; Ng and Lee 1996; Rigau et al. 1997). We investigate this problem in the context of topic detection.

We present also an evaluation of the algorithm for identifying the "aboutness" of a text segment, and we analyze the performance of the algorithm in terms of precision and recall, i.e., performance measures borrowed from the information retrieval domain.

MOTIVATION

Most documents are about more than one subject, but many NLP and IR techniques implicitly assume documents have just one topic. Even in the presence of a single topic within a document, the document may address multiple subtopics and various aspects of the primary topic. Dividing documents into topically coherent units, discovering, and threading their topic could be quite valuable in many applications where people need timely and efficient access to large quantities of information. For example, systems could alert users to new events and to new information about old enets. By examining one or two texts, a user could decide whether to pay attention to the remainder of an evolving thread. Similarly, a user could go to a large archive, find all the texts about a particular event, and learn how it evolved. Such automatic discovering and threading might have many applications and uses:

- In information retrieval, documents in many collections are likely to address multiple topics and various aspects of the primary topic. Indexing and clustering these documents based on topical words, instead of frequent phrases, can be exploited to improve the accuracy of an information retrieval system.
- In text summarization, the primary problem is detecting the relevant portions of texts. Characterizing those portions by their topics will improve the summarization task especially when the purpose of the summary is user-focused (Mani and Maybury 1999).
- In text understanding, the scope of several phenomena is intersentential, the topic can account of such a scope and hence can help in their

resolution, e.g., in resolving anaphora and ellipsis (Kozima 1993) and in cue phrases of discourse usage (Harabagiu 1999).
- In structuring text with regard to its discourse hierarchy (Halliday and Hasan 1976; Hahn 1990; Morris and Hirst 1991; Harabagiu 1999).
- In improving document navigation and hypertext links (Green 1997; Pratt et al. 1999).

In the reminder of the paper, we will focus on the main algorithm for deriving topic signatures of texts, independently from any application.

RELATED WORK

Much research has been devoted to the task of structuring text, that is, dividing texts into units based on information within the text. Existing work falls roughly into one of the two categories: linear text segmentation aims to discover the topic boundaries, and discourse segmentation focuses on identifying relations between utterances. Methods for finding the topic boundaries include word repetition within a sliding window (Hearst 1997), lexical cohesion based on word similarity Morris and Hirst (1991) and Kozima (1993), entity repetition with regard to its position within the paragraph (Kan et al. 1998), word frequency algorithm and maximum entropy model (Reynar 1999), context vectors (Kaufmann 1999), feature induction model (Beeferman et al. 1999) divisive clustering (Choi 2000). On the other hand, discourse segmentation is fined-grained, Litman and Passonneau 1995), combines multiple knowledge sources for discourse segmentation using decision trees, and (Marcu 1997b) uses a rhetorical parsing (Marcu 1997a) and decision tree (Marcu 1999) to build up the discourse structure based on relations. Harabagiu (1999) devises a model of coherence structure based on the data provided by lexical paths from real world texts.

The systems for understanding "what the text is about" are based on world knowledge. These systems can be broadly categorized into three types: One System that rely on prior knowledge of their domains (DeJong 1982; Radev and Mckeown 1998). For instance, DeJong (1982) developed a system based on templates that organize its world knowledge in order to skim newspaper stories and extract the main details. Similarly, Radev and McKeown (1998) developed a system that takes template outputs of information extraction systems developed for Message Understanding Conference (MUC) and generates summaries of multiple news articles. Two Systems that learn patterns from preclassified texts of specific topics which are then used to identify the presence of the learned topics in previously unseen texts (Riloff and Lorenzen 1999; Lin and Hovy 2000). For example, Riloff and Lorenzen (1999) present a system that generates extraction patterns and learns lexical constraints automatically from preclassified texts. Similarly, Lin and Hovy

(2000) present a procedure to automatically acquire topic signatures from preclassified documents of specific topics and then use them to identify the presence of the learned topics in previously unseen documents. Three Systems based on commonly available resources, such as a thesaurus, which can be applicable for unrestricted texts (Chali 2001).

The first system relies on prior knowledge of their domains, i.e., they are domain-dependent. However, to acquire such prior knowledge is labor-intensive and time-consuming. The second systems reduce the knowledge engineering bottleneck. However, learning extraction patterns from corpora makes those systems domain-specific. The third systems use more general knowledge. We present a method which is based on commonly available resources, such as WordNet, and which can be applicable for unrestricted texts, and we investigate how accurate the systems based on this knowledge are.

The approach that we propose to pursue next is a step further to the approaches intending to identify the boundaries between paragraphs in a text where the text changes topic. In the sense that we attach topic signatures to contiguous portions of text between two boundaries in order to label them. We present a system that proceeds in two steps: (1) the input text is segmented at a spot where a topic shift is probable using *TextTiling* (Hearst 1997), *Segmenter* (Kan et al. 1998), or Choi's system (Choi 2000) and (2) Texical chains are extracted from each segment, using either WordNet or *Roget's Thesaurus*, as indicators of its topic.

LEXICAL CHAINS AS TOPIC SIGNATURES

In a text, a sequence of sentences tends to convey information about a certain topic, and by doing so, they use related words, providing the text with the quality of unity. Structural theories of text are concerned with identifying units of text that are about the "same thing." When this happens, there is a strong tendency for semantically related words to be used within that unit. The notion of cohesion, introduced by Halliday and Hasan (1976), is a property of sentences to "stick together" to function as a whole. It is achieved through the use of *grammatical cohesion*, i.e., reference, substitution, ellipsis and conjunction, and *lexical cohesion*, i.e., semantically related words. Lexical cohesion occurs not only between two terms, but among sequences of related words, called *lexical chains* (Morris and Hirst 1991). Lexical cohesion arises from the semantic connections between words. Therefore, deriving the cohesion structure of a text amounts to retrieving lexical chains (LCs):

$$LC = \{w_1, w_2, \ldots, w_n\} \tag{1}$$

where w_is are words, and on any pair of words holds a semantic relation.

Lexical chains that provide an easy-to-determine context to aid in the resolution of ambiguity and in the narrowing to specific meaning of a word tend to delineate portions of text that have a strong unity of meaning. We investigate how lexical chains can be used as an indicator of the text topic.

Lexical Resources

The method for measuring the semantic relation between words is dependent on the semantic representation used in the lexical database. Two lexical database resources were used for detecting the topic signatures.

Roget's Thesaurus, a collection of words and phrases arranged according to the ideas they express, is organized as a hierarchy from major classes to categories, and further to their subdivisions in paragraphs with words having the same part-of-speech and groups indicate a structure that induces semantic classes and relations among them. Furthermore, semantic similarity is indicated by reference pointers between categories, paragraphs, and groups of words. Figure 1 illustrates the organization of *Roget's Thesaurus*.

The implementation of the mechanism that retrieves concepts is performed by an *index function*, returning for any word a list of words suggesting related sub-senses, along with the category and the paragraph number of each of these. Figure 2 illustrates the index entry for the word *lid*.

WordNet encodes the lexical concepts as synsets (i.e., synonym sets of words). Moreover, WordNet returns the list of synsets containing a name grouped along the same part-of-speech and ordered by the frequency in the Brown corpus. WordNet employs also a hierarchical representation for nouns and verbs. There are 11 noun hierarchies and 558 verb hierarchies in WordNet 1.6 made possible by the *isa* semantic relations, but also encodes semantic relations that cross these hierarchies (e.g., the *is-part*, *is-member*, *has-stuff* semantic relations between noun synsets and the *entail* and *cause-to* semantic relations between verb synsets). Figure 3 illustrates the representation for the word *lid*.

Topic Signatures

Devising an algorithm for building lexical signatures by grouping sets of words that are semantically related requires the definition of how these words are related. Identities, synonyms, and hypernyms/hyponyms, which together define a tree of *isa* relations between words, are the relations among words that might cause them to be grouped into the same topic signature. Specifically, words may be grouped when:

- Two nouns instances are identical, and used in the same sense.
  (1) The house on the hill is large. The house is made of wood.

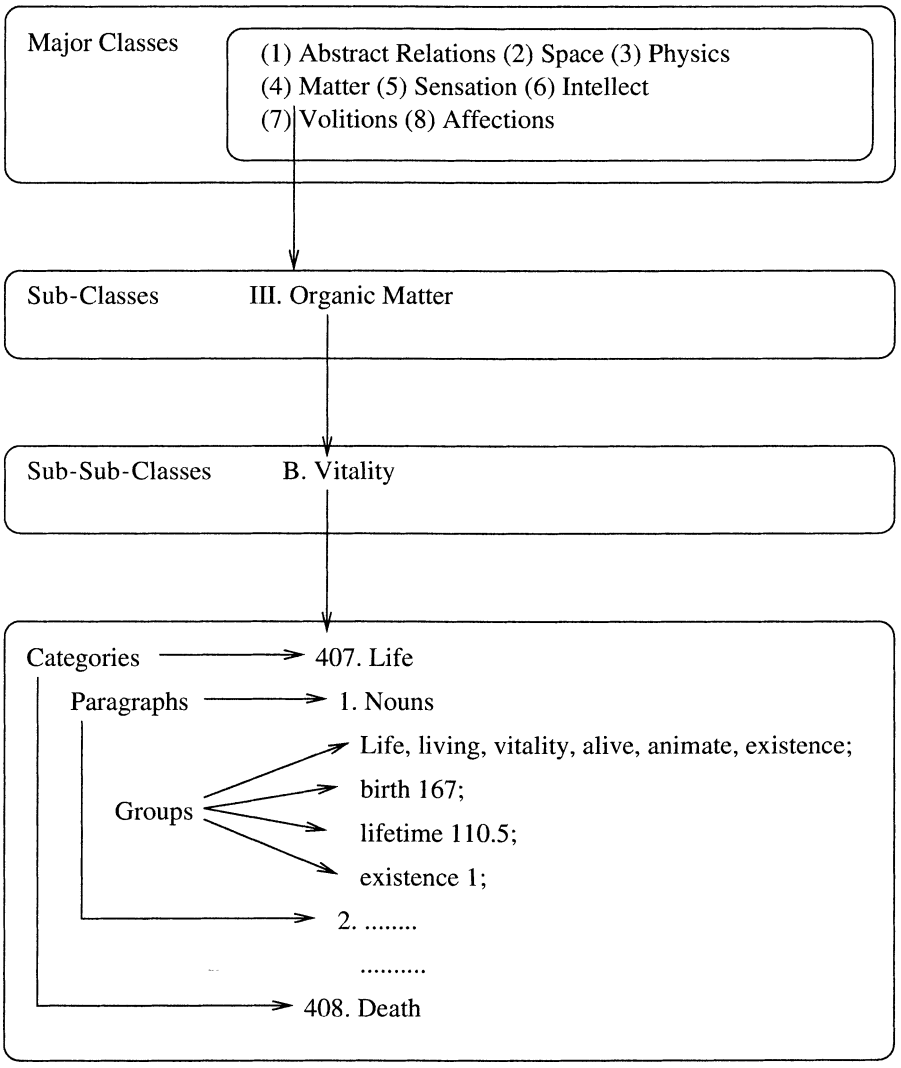

FIGURE 1 Semantic structure of *Roget's Thesaurus*.

- Two nouns instances are used in the same sense (i.e., synonyms).
  (2) The car is fast. My automobile is faster.
- The senses of two nouns instances have a hypernym/hyponym relation between them.
  (3) John owns a car. It is a Toyota.
- The senses of two nouns instances are siblings in the hypernym/hyponym tree.
  (4) The truck is fast. The car is faster.

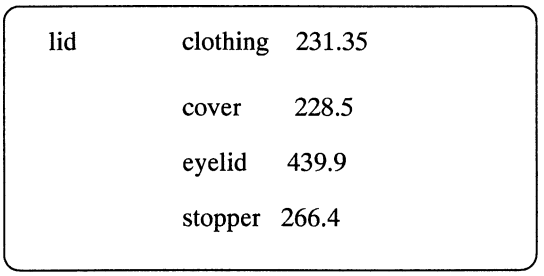

FIGURE 2 Index entry in *Roget's Thesaurus* for the word *lid*.

In computing the topic signatures, the noun instances must be grouped according to the above relations, and each noun instance must be used in one sense, if it corresponds to several different word senses.

SYSTEM OVERVIEW

The overall architecture of the system is shown in Figure 4. It consists of two independent modules organized as a pipeline. The following sections describe in details the two modules.

TEXT SEGMENTING

The linear segmentation task is motivated by the observation that comprehension of longer texts benefits from automatic chunking of cohesive sections. This task involves breaking input text into segments that represent some meaningful grouping of contiguous portions of the text. The input text is divided into a linear sequence of adjacent segments and segment boundaries are found at various paragraph separations that identify one or more subtopical shifts.

Multi-paragraph subtopic segmentation should be useful for many text analysis tasks, including information retrieval and summarization. Specifically, text segmentation is interesting for the following purposes:

- Segmentation is intended to identify the boundaries between paragraphs in a text where the text changes topic. Thus, a text can comprise merely a single segment, or perhaps several different segments, when it touches on several different topics.
- It helps in processing the user needs when they are specified as terms in the sense that only segments that are relevant to the terms specified

```
Synonyms/Hypernyms (Ordered by Frequency) of noun lid

3 senses of lid

Sense 1
eyelid, lid
      => protective fold
         => flap
            => animal tissue
               => tissue
                  => body part
                     => part, piece
                        => entity, something

Sense 2
lid
      => top, cover
         => covering
            => artifact, artefact
               => object, physical object
                  => entity, something

Sense 3
hat, chapeau, lid
      => headdress, headgear
         => clothing, clothes, apparel, vesture, wearing apparel, wear
            => covering
               => artifact, artefact
                  => object, physical object
                     => entity, something
            => consumer goods
               => commodity, goods
                  => object, physical object
                     => entity, something
```

FIGURE 3 WordNet synset representation for the word *lid*.

by the user are chosen (Reynar 1999, Chali et al. 1999). When the topically coherent units (i.e., text segments) are represented by a set of topical clues, a content-based process of matching the user's terms against the segment's clues will determine the relevancy of the segments, i.e., the segments with the highest matches are selected as answers to the user's query.

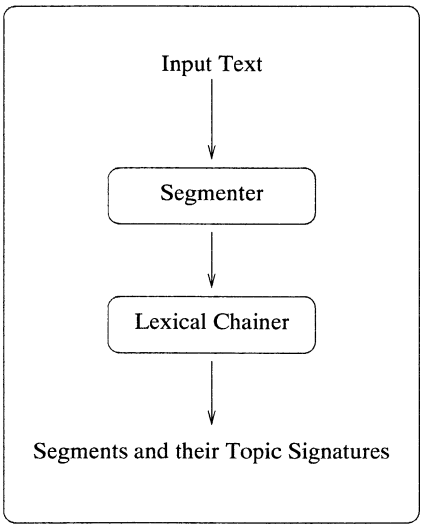

FIGURE 4 System overview.

The segmentation step is based on three different methods: the *TextTiling* system (Hearst 1997), the *Segmenter* system (Kan et al. 1998), and Choi's system (Choi 2000).

TextTiling compares blocks of text based on vocabulary overlap to identify topic boundaries. Thus, repetition is used to identify where topical segment start and end. The algorithm divides a document into fixed-length text windows (e.g., 20 words). Adjacent blocks of windows are compared for similarity based on a vocabulary overlap measure. The similarity scores are then plotted against gaps between blocks, and the resulting gaps are then sorted by how large a positive change in similarity occurs on either side of the gap. The system assigns topic segment boundaries to the gap with the largest similarity change.

Segmenter extracts occurrences of terms and links them together according to a proximity metric. The paragraphs are weighted according to the positional relationship with each term link. Then, local maxima are examined in order to arrive to the final topic boundaries.

Choi's system takes a list of tokenized sentences as input. A sentence is represented by a set of word frequencies, and a cosine similarity measure is computed between each pairs of sentences. A local ranking process based on the number of neighboring elements with a lower similarity value is applied to each pair. Then, a clustering based on Reynar (1999) is used to locate the topic boundaries.

Segmentation is followed by the characterization of the segment in terms of lexical chains as clues of the segment topic.

LEXICAL CHAINING

The steps of the algorithm for the lexical chain computation are as follows:

1. We select the set of candidate words. To this end, we run a part-of-speech tagger (Brill 1992) on a text segment, and only the open class words that function as noun phrases or proper names are chosen.
2. The set of the candidate words are exploded into senses, the senses are given by the thesaurus in use, and at this step all the senses of the same word are considered. In the actual implementation, we are using two different thesauri; *Roget's Thesaurus* (Chapman 1988) and WordNet thesaurus Miller et al. 1993). From this step each word sense is represented by distinct sets (see Figure 5) considered as levels. The first one constitutes the set of synonyms and antonyms, the second one constitutes the set of first hypernyms/hyponyms and their variations (i.e., meronyms/holonyms, etc.), and so on. In our experiments, we consider two levels up and two levels down in the dictionary hierarchy.
3. We find the semantic relatedness among the set of word senses according to their representations. A semantic relationship exists between two word senses if comparing two sense representations (see Figure 5) of two distinct words, a matching exists, i.e., a non-empty intersection exists between the sets of word senses. To each semantic relatedness is associated a measure that indicates the length of the path taken in the matching with respect of the levels of the compared two sets.
4. We build up chains which are sets such as:

$$\{(word_1\ [sense_{11},\ sense_{12}, ...]),\ (word_2[sense_{21},\ sense_{22}, ...]), ...\} \qquad (2)$$

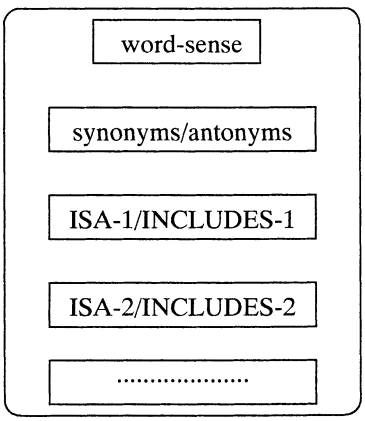

FIGURE 5 Word sense representation.

in which $word_i$-$sense_{ix}$ is semantically related to $word_j$-$sense_{jy}$ for $i \neq j$, and x and y correspond to the senses of $word_i$ and $word_j$, respectively.

5. We retain longest chains relying on the following preference criterion:

$$\begin{aligned} &word\ repetition \gg \\ &synonym/antonym \gg \\ &\quad ISA-1/INCLUDES-1 \gg \\ &\quad\quad ISA-2/INCLUDES-2 \gg \\ &\quad\quad \ldots \end{aligned} \tag{3}$$

In our implementation, this preference is handled by assigning scores to each pair of semantically related word senses in the chain, and then adding up those pairwise scores. Hence, the score of a chain is based on its length and on the type of relationship holding among its members.

$$score(LC) = \sum_{i=1}^{n} (score(pairwiseR_i)) \tag{4}$$

where $pairwiseR_i$ is a semantic relationship holding between two word senses, and n is their total number.

For instance, we assign the score of 5 to word repetitions, the score of 4 to synonyms/antonyms, the score of 3 to first hypernyms/hyponyms, and the score of 2 to second hypernyms/hyponyms. These scores are chosen to reflect the preference formula and prove to be fruitful in practice. The total score of a chain is the sum of all pairwise scores, and the chain with the highest score is preferred to the chain with the lowest score.

Consider the words *breathing*, *equipment*, *heater*, and *smoke*. The word *heater* has two senses:

(5)#{heater, warmer}
(6)#{fastball, heater, smoke, hummer, bullet}

The two following chains are built:

(7)#{equipment, heater}
(8)#{breathing, heater, smoke}

However, chain (8) is retained over chain (7), since the score of chain (8) is 10 (i.e., $3+3+4$, breathing and heater are first hyponyms, breathing and smoke are first hyponyms, and heater and smoke are synonyms) compared to the score of chain (7) which is 2 (i.e., equipment and heater are second hypernyms). We mention that in chain (8) the word *heater* is in sense (6) and in chain (7) the word *heater* is in sense (5). Also, by retaining

chain (8) over chain (7), we disambiguate the word *heater* since the sense {*fastball, heater, smoke, hummer, bullet*} in chain (8) is preferred over the sense {*heater, warmer*} in chain (7).

In the lexical chaining method, the relationship between the words in a chain is pair-wise mutual, that is, each word-sense has to be semantically related to every other word-senses in the chain. The order of the open class words in the document does not play a role in building up the chains. However, it turned out that the number of lexical chains could be extremely large, and thus problematic, for large segments of text. To cope with this, we reduced the word-sense representation to synonyms only when we have long text segments. This reduction has another benefit, in the sense that a lexical chain based only on synonyms could be better than one based on ISA-2/INCLUDES-2. This reduction also has to narrow down the set of lexical chains stemming from a single segment in the case when there are too many.

We show the output of the lexical chaining on a fragment of text (9). (10) and (11) are the lexical chains computed using WordNet and *Roget's Thesaurs* respectively. For the sake of clarity, we do not show the sense numbers of each word since those numbers depend on the dictionary.

(9) A series of explosions and fire shut down electricity generation at the world's largest solar power plant near here Wednesday. Thick plumes of black smoke spiraled into the clear desert air when one of four natural gas-fired heaters used to back up the solar heating system exploded. A short time later, a second natural gas heater caught fire and exploded as the first of 75 firefighters and 25 pieces of equipment were arriving at the site, about 140 miles northeast of Los Angeles. "We had a series of explosions, more than two," said Capt.Sharon Sellers of the San Bernardino County Fire Department. "Out first units got on-scene at 9:16 a.m. and a second explosion occurred at that point, then a series of the during the entire incident," Sellers said. "There was a mushroom cloud. The heat was real intense and there were explosions," said an inmate from the Boron Federal Prison Camp who was pressed into service to help fight the fire. He would not identify himself. Sellers said two workers at the plant suffered minor breathing problems and were treated at Barstow Community Hospital. Operated by LUZ International Ltd. of Los Angeles, the $280-million Harper Lake solar plant began generating electricity on Dec. 28 and produces 80 megawatts, enough power to serve 115,000 people. The company operates eight such plants in the California desert. Combined, they generate 274 megawatts, which is sold to Southern California Edison Co. An Edison spokesman said there was no interruption of electric service to its customers. "We had two oil heaters on line and were bringing up the third

and fourth oil heaters when this explosion occurred," LUZ International spokeswoman Kathleen Flanagan said in Los Angeles. While no flames were visible 1 1/2 hours after the fire began shortly before 9 a.m., San Bernardino County firefighters had difficulty reaching the blaze deep within the generating equipment. "There is fire up there somewhere still heating that oil," Sellers said. The blaze was contained, but continued to burn late Wednesday. Cause of the fire was unknown, but fire officials ruled out arson and said it probably resulted from an equipment malfunction. While Flanagan said she could not immediately estimate the cost of the blaze, the Fire Department said a single natural gas heater costs $500,000. One was destroyed and a second was heavily damaged. Flanagan said the black smoke from an estimated 15,000 gallons of burning synthetic oil was not any more toxic than smoke from natural crude or refined oil and was not carcinogenic. But that report was disputed by Capt. Clyde Gamma of the California Department of Forestry and Fire Protection. He identified the synthetic oil as Therminol and said it is cancer-causing. Flanagan said the plant could resume generating electricity by Monday. But she said the backup natural gas-fired heaters would not be used. "We will be operating strictly in the solar mode," she said. For solar generation, large curved mirrors are used to concentrate the sun's energy onto synthetic oil, which flows through an insulated steel pipe. The hot oil boils water into steam that drives conventional electrical turbines. Sellers said LUZ International had a fire about two years ago at another solar plant at Daggett and that explosions continued five hours into the incident. Stammer reported from Los Angeles and Harris from Barstow.

(10) a.#{blaze, fire}
b. {breathing, heater, smoke}
c. {california, los_angeles}
d. {company, ltd.}
e. {county, department}
f. {crude_oil, oil}
g. {desert}
h. {difficulty, problem}
i. {electricity}
j. {equipment, mode}
k. {equipment, unit}
l. {explosion, fire}
m. {fire, protection}
n. {gas_heater, heater, oil_heater, smoke}
o. {international}
p. {monday, wednesday}

q. {plant, worker}
r. {firefighters}
s. {barstow}
t. {luz}
u. {flanagan}
v. {generating}

(11) a.#{air, difficulty, line}
b. {air, line, pipe, series, unit}
c. {air, line, mode}
d. {air, line, piece, report}
e. {cause, energy}
f. {cause, world_power}
g. {county, department}
h. {fire, protection}
i. {international}
j. {monday, wednesday}
k. {cloud, electricity, energy, power}
l. {cloud, mushroom}
m. {company, system, units}
n. {difficulty, problems}
o. {energy, heating}
p. {plant, worker}
q. {firefighters}
r. {barstow}
s. {luz}
t. {flanagan}

Lexical chains are computed for each text segment. They are sets of clues reflecting the topic of the text segment.

EVALUATION

Since there is no formal method to evaluate the quality of the system, we relied on the following experiments. We carried out an evaluation of the system to assess its quality. We selected randomly ten texts from the Brown corpus as test corpora. We segmented them using Choi's segmenter because it is more precise than the two others (cf. Choi 2000). This gave us a sample of 112 text segments. Then we computed the lexical chains for each of these segments using both of the thesauri and our system, and using a system which considers just the word repetitions. This constituted the baseline for our experiments. Finally, we presented the text segments and the lexical chains to five judges.[1] We asked all the judges:

1. To read carefully the text segment.
2. To read carefully each lexical chain and answer the following question:

$$\text{"}\textit{Is the chain's topic present in the segment?}\text{"} \tag{5}$$

3. After reading all the lexical chains corresponding to one segment and answering the previous question, we asked them to answer the following question:

$$\text{"}\textit{Is the segment's topic covered by all its chains?}\text{"} \tag{6}$$

We considered the answer of *yes* or *no* given by the majority of the judges. Related to the information retrieval measures, the answers to the first question correspond to precision (i.e., how many lexical chains are good among all the computed lexical chains) and the answers to the second question correspond to recall (i.e., how much of the segment's topic is covered by the lexical chains). Precision and recall are computed according to Eqs. (7) and (8), respectively.

$$\textit{Precision} = \frac{\textit{number of answer yes to question } 5}{\textit{total number of questions } 5} \tag{7}$$

$$\textit{Recall} = \frac{\textit{number of answer yes to question } 6}{\textit{total number of questions } 6} \tag{8}$$

We notice that the *total number of question v* corresponds to the total number of chains, and the *total number of question vi* corresponds to the total number of segments.

The results of our evaluation are shown in Table 1.

DISCUSSION AND FUTURE WORK

The results generated by our system are much better than the results generated by the baseline system. This experiment shows also that the

TABLE 1 Results of the Evaluation

	Precision	Recall
Word Repetition	27.44%	32.81%
Using Roget's Thesaurus	38.75%	54.67%
Using WordNet	44.56%	63.83%

entire system is more accurate using WordNet than *Roget's Thesaurus*. This is due on one hand to the number of entries in the thesaurus (i.e., 99,642 synsets and 121,962 unique words in WordNet as of version 1.6 compared to *Roget's Thesaurus* 1035 categories and 46,500 unique words as of version 7.1). On the other hand, the classification into categories in *Roget's* are more general abstractions compared to the organization into synsets defined in WordNet. Indeed, WordNet represents the largest public available lexical resource to date.

Lexical chains have been proposed by Morris and Hirst (1991) as indicators of the structure of text. Barzilay and Elhadad (1997) investigate the production of summaries based on lexical chaining. The summaries are built using scoring, which is based on chain length, and the extraction of significant sentences is based on heuristics using chain distribution. For example, choose the sentence that contains the first appearance of a chain member in the text. In the paper, we investigate the production of lexical chains to account for the topic of the text segment.

The described algorithm for the lexical chaining was implemented in C++. Its primary purpose is to extract from the text segments meaningful clues as indicators of the segment's topic. This technique has many uses in the processing and searching of information.

The results reported in this paper suggest that we may refine the process of lexical chaining. Instead of choosing any content word tagged as noun or proper noun as candidate for the computation of the chains, it seems that restricting the set of candidate words will improve the precision of the chains.

CONCLUSION

Topic detection and identification is an important area of research, addressing many application needs. It presents new and interesting technical challenges.

We presented an algorithm for detecting the topic of unrestricted texts based on an efficient use of lexical chains acquired from common lexical knowledge. The results show that the algorithm is promising for many applications where efficient access to large quantities of information is needed.

NOTE

1. Given the labor intensive nature of the task, we could not select more judges. The human judges were graduate students in computer science. All of the subjects had good reading and comprehension skills in English.

REFERENCES

Agirre, E. and G. Rigau. 1995. A proposal for word sense disambiguation using conceptual distance. In *Proceedings of the First International Conference on Recent Advances in Natural Language Processing*, Velingrad, Bulgaria.

Barzilay, R. and M. Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter Meeting of the Association for Computational Linguistics, Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain.

Beeferman D., A. Berger, and J. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning, Special Issue on Natural Language Processing* 34(13):177–210.

Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*, pages, 152–155, Trento, Italy.

Bruce, R. and J. Wiebe. 1994. Word sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–146, Las Cruces, New Mexico, USA.

Chali, Y. 2001. Topic detection using lexical chains. In *Proceedings of the Fourteenth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 552–558, Budapest, Hungary. Lecture Notes in Computer Science 2070, Springer-Verlag.

Chali, Y., S. Matwin, and S. Szpakowicz. 1999. Query-biased text summarization as a question-answering technique. In *Proceedings of AAAI Symposium on Question-Answering Systems*, pages 52–56, Cape Cod, Massachusetts. AAAI Press.

Chapman, R. 1988. *Roget's International Thesaurus*. London: Longman.

Choi, F. Y. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics*, pages 26–33, Seattle, Washington, USA.

Cowie, J., L. Guthrie, and J. Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the Fifth International Conference on Computational Linguistics*, pages 157–161, Nantes, France.

DeJong, G. 1982. An overview of the frump system. *Strategies for Natural Language Processing*, G. Lehnert and M. H. Ringle, In eds. 76–49. Lawrrence Erlbaum Associates.

Green, S. J. 1997. *Automatically Generating Hypertext by Computing Semantic Similarity*. Ph.D. thesis, Department of Computer Science, University of Toronto.

Hahn, U. 1990. Topic parsing: Accounting for text macrostructures in full text analysis. *Information Processing and Management* 26:135–170.

Halliday, M. and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Harabagiu, S. 1999. From lexical cohesion to textual coherence: A data driven perspective. *Journal of pattern Recognition and Artificial Intelligence* 13(2):247–265.

Hearst, M. A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64.

Kan, M.-Y., K. R. McKeown, and J. L. Klavans. 1998. Linear segmentation and segment relevance. In *Proceedings of 6th International Workshop of Very Large Corpora (WVLC-6)*, pages 197–205, Montréal, Canada.

Kaufmann, S. 1999. Cohesion and collocation: Using context vectors in text segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (Students Session)*, pages 591–595, College Park, Maryland, USA.

Kozima, H. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 286–288, Columbus, Ohio, USA.

Li, X., S. Szpakowicz, and S. Matwin. 1995. A wordNet based algorithm for word semantic sense disambiguation. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1368–1374, Montréal, Canada.

Lin, C. Y. and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of 18th International Conference in Computational Linguistics*, pages 494–501. Saarbrücken, Germany.

Litman, D. and R. J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 108–115, Cambridge, Massachusetts, USA.

Mani, I. and M. Maybury. 1999. *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press.

Marcu, D. 1997a. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the European Chapter of the Association for Computational Linguistics*, pages 96–103, Madrid, Spain.

Marcu, D. 1997b. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, Department of Computer Science, University of Toronto.

Marcu, D. 1999. A decision-based approach to rhetorical parsing. In *Proceedings of The 37th Annual Meeting of the Association for Computational Linguistics*, pages 365–372, College Park, Maryland, USA.

McRoy, S. 1992. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics* 18(1):1–30.

Mihalcea, R. and D. I. Moldovan. 1999a. An automatic method for generating sense tagged corpora. In *proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 461–466, Orlando, Florida, USA.

Mihalcea, R. and D. I. Moldovan. 1999b. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 152–158, College Park, Maryland, USA.

Miller, G., M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 240–243, Princeton, New Jersey, USA.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, 1993. *Five Papers on WordNet*. CSL Report 43, Cognitive Science Laboratory, Princeton University.

Morris, J. and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17(1):21–48.

Ng, H. T. and H. B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, USA.

Pratt, W., M. A. Hearst, and L. M. Fagan. 1999. A knowledge-based approach to organizing retrieved documents. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 80–85, Orlando, Florida, USA.

Radev, D. R. and K. R. Mckeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics* 24(3):469–500.

Reynar, J. C. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 357–364, College Park, Maryland, USA.

Rigau, G., J. Atserias, and E. Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of Joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* pages 48–55, Madrid, Spain.

Riloff, E. and J. Lorenzen. 1999. Natural language information retrieval, Generating Domain-Specific Role Relationships Automatically. ed. Tomek, pages 1–30, Strzalkowski, Kluwer Academic Publishers.