The Automated Acquisition of Topic Signatures for Text Summarization

Chin-Yew Lin and Eduard Hovy
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292, USA
{cyl,hovy}@isi.edu

Abstract
In order to produce a good summary, one has to identify the most relevant portions of a given text. We describe in this paper a method for automatically training topic signatures—sets of related words, with associated weights, organized around head topics and illustrate with signatures we created with 6,194 TREC collection texts over 4 selected topics. We describe the possible integration of topic signatures with ontologies and its evaluation on an automated text summarization system.

1 Introduction
This paper describes the automated creation of what we call topic signatures, constructs that can play a central role in automated text summarization and information retrieval. Topic signatures can be used to identify the presence of a complex concept—a concept that consists of several related components in fixed relationships. Restaurant-visit, for example, involves at least the concepts menu, eat, pay, and possibly waiter, and Dragon Boat Festival (in Taiwan) involves the concepts calamus (a talisman to ward off evil), moxa (something with the power of preventing pestilence and strengthening health), pictures of Chung Kuei (a nemesis of evil spirits), eggs standing on end, etc. Only when the concepts co-occur is one licensed to infer the complex concept; cat or moxa alone, for example, are not sufficient. At this time, we do not consider the interrelationships among the concepts.

Since many texts may describe all the components of a complex concept without ever explicitly mentioning the underlying complex concept—a topic—itself, systems that have to identify topic(s), for summarization or information retrieval, require a method of inferring complex concepts from their component words in the text.

2 Related Work
In late 1970's, DeJong (DeJong, 1982) developed a system called FRUMP (Fast Reading Understanding and Memory Program) to skim newspaper stories and extract the main details. FRUMP uses a data structure called sketchy script to organize its world knowledge. Each sketchy script is what FRUMP knows about what can occur in particular situations such as demonstrations, earthquakes, labor strikes, and so on. FRUMP selects a particular sketchy script based on clues to styled events in news articles. In other words, FRUMP selects an empty template whose slots will be filled on the fly as FRUMP reads a news article. A summary is generated based on what has been captured or filled in the template.

The recent success of information extraction research has encouraged the FRUMP approach. The SUMMONS (SUMMarizing Online News articles) system (McKeown and Radev, 1999) takes template outputs of information extraction systems developed for MUC conference and generating summaries of multiple news articles. FRUMP and SUMMONS both rely on prior knowledge of their domains. However, to acquire such prior knowledge is labor-intensive and time-consuming. For example, the University of Massachusetts CIRCUS system used in the MUC-3 (SAIC, 1998) terrorism domain required about 1500 person-hours to define extraction patterns (Riloff, 1996). In order to make them practical, we need to reduce the knowledge engineering bottleneck and improve the portability of FRUMP or SUMMONS-like systems.

Since the world contains thousands, or perhaps millions, of complex concepts, it is important to be able to learn sketchy scripts or extraction patterns automatically from corpora—no existing knowledge base contains nearly enough information. (Riloff and Lorenzen, 1999) present a system AutoSlog-TS that generates extraction patterns and learns lexical constraints automatically from preclassified text to alleviate the knowledge engineering bottleneck mentioned above. Although Riloff applied AutoSlog-TS

1 We viewed sketchy scripts and templates as equivalent constructs in the sense that they specify high level entities and relationships for specific topics.

2 An extraction pattern is essentially a case frame contains its trigger word, enabling conditions, variable slots, and slot constraints. CIRCUS uses a database of extraction patterns to parse texts (Riloff, 1996).
algorithms such as TextTiling (How and Lin, 1997) can be used to topic signatures and topic signatures are both trained on preclassified documents of specific topics and used to identify the presence of the learned topics in previously unseen documents. The main differences to our approach are: relevancy signatures require a parser. They are sentence-based and applied to text categorization. On the contrary, topic signatures only rely on corpus statistics, are document-based\(^3\) and used in text summarization.

In the next section, we describe the automated text summarization system SUMMARIST that we used in the experiments to provide the context of discussion. We then define topic signatures and detail the procedures for automatically constructing topic signatures. In Section 5, we give an overview of the corpus used in the evaluation. In Section 6 we present the experimental results and the possibility of enriching topic signatures using an existing ontology. Finally, we end this paper with a conclusion.

3 SUMMARIST

SUMMARIST (How and Lin, 1999) is a system designed to generate summaries of multilingual input texts. At this time, SUMMARIST can process English, Arabic, Bahasa Indonesia, Japanese, Korean, and Spanish texts. It combines robust natural language processing methods (morphological transformation and part-of-speech tagging), symbolic world knowledge, and information retrieval techniques (term distribution and frequency) to achieve high robustness and better concept-level generalization.

The core of SUMMARIST is based on the following equation:

\[
\text{summarization} = \text{topic identification} + \text{topic interpretation} + \text{generation}.
\]

These three stages are:

**Topic Identification**: Identify the most important (central) topics of the texts. SUMMARIST uses positional importance, topic signature, and term frequency. Importance based on discourse structure will be added later. This is the most developed stage in SUMMARIST.

**Topic Interpretation**: To fuse concepts such as waiter, menu, and food into one generalized concept restaurant, we need more than the simple word aggregation used in traditional information retrieval. We have investigated concept

\(^3\)We would like to use only the relevant parts of documents to generate topic signatures in the future. Text segmentation algorithms such as TextTiling (Howarst, 1997) can be used to find subttopes in text.

Figure 1 shows an ABC News page summary about EgyptAir Flight 990 by SUMMARIST. SUMMARIST employs several different heuristics in the topic identification stage to score terms and sentences. The score of a sentence is simply the sum of all the scores of content-bearing terms in the sentence. These heuristics are implemented in separate modules using inputs from preprocessing modules such as tokenizer, part-of-speech tagger, morphological analyzer, term frequency and \textit{tfidf} weights calculator, sentence length calculator, and sentence location identifier. We only activate the position module, the \textit{tfidf} module, and the topic signature module for comparison. We discuss the effectiveness of these modules in Section 6.

4 Topic Signatures

Before addressing the problem of world knowledge acquisition head-on, we decided to investigate what type of knowledge would be useful for summarization. After all, one can spend a lifetime acquiring knowledge in just a small domain. But what is the minimum amount of knowledge we need to enable effective topic identification as illustrated by the restaurant-visit example? Our idea is simple. We would collect a set of terms\(^4\) that were typically highly correlated with a target concept from a preclassified corpus such as TREC collections, and then, during summarization, group the occurrence of the related terms by the target concept. For example, we would replace joint instances of table, menu, waiter, order, eat, pay, tip, and so on, by the single phrase restaurant-visit, in producing an indicative

\(^4\)Terms can be stemmed words, bigrams, or trigrams.
then the likelihood for \( H_1 \) is:
\[
L(H_1) = b(O_{11}; O_{11} + O_{12}, p)b(O_{21}; O_{21} + O_{22}, p)
\]
and for \( H_2 \) is:
\[
L(H_2) = b(O_{11}; O_{11} + O_{12}, p_1)b(O_{21}; O_{21} + O_{22}, p_2)
\]
The \(-2\log\lambda\) value is then computed as follows:
\[
-2\log \frac{L(H_1)}{L(H_2)}
\]
where \( N = O_{11} + O_{12} + O_{21} + O_{22} \) is the total number of term occurrence in the corpus, \( H(R) \) is the entropy of terms over relevant and nonrelevant sets of documents, \( H(R; T) \) is the entropy of a given term over relevant and nonrelevant sets of documents, and \( I(R; T) \) is the mutual information between document relevancy and a given term. Equation 5 indicates that mutual information is an equivalent measure to likelihood ratio when we assume a binomial distribution and a 2-by-2 contingency table.

To create topic signatures for a given topic, we:
1. classify documents as relevant or nonrelevant according to the given topic
2. compute the \(-2\log\lambda\) value using Equation 3 for each term in the document collection
3. rank terms according to their \(-2\log\lambda\) value
4. select a confidence level from the \( \chi^2 \) distribution table; determine the cutoff associated weight and the number of terms to be included in the signatures

5 The Corpus

The training data derives from the Question and Answering summary evaluation data provided by TIPSTER-SUMMAC (Mani et al., 1998) that is a subset of the TREC collections. The TREC data is a collection of texts, classified into various topics, used for formal evaluations of information retrieval systems in a series of annual comparisons. This data set contains essential text fragments (phrases, clauses, and sentences) which must be included in summaries to answer some TREC topics. These fragments are each judged by a human judge. As described in Section 3, SUMMARIST employs several independent modules to assign a score to each sentence, and then combines the scores to decide which sentences to extract from the input text. One can gauge the efficacy of the mutual information is defined according to chapter 2 of (Cover and Thomas, 1991) and is not the pairwise mutual information used in (Church and Hanks, 1990).
of each module by comparing, for different amounts of extraction, how many 'good' sentences the module selects by itself. We rate a sentence as good simply if it also occurs in the ideal human-made extract, and measure it using combined recall and precision (F-score). We used four topics\(^7\) of total 6,194 documents from the TREC collection. 138 of them are relevant documents with TIPSTER-SUMMAC provided answer keys for the question and answering evaluation. Model extracts are created automatically from sentences containing answer keys. Table 1 shows TREC topic description for topic 151, test questions expected to be answered by relevant documents, and a sample relevant document with answer keys.

Table 1: TREC topic description for topic 151, test questions expected to be answered by relevant documents, and a sample relevant document with answer keys.

6 Experimental Results

In order to assess the utility of topic signatures in text summarization, we follow the procedure described at the end of Section 4.1 to create topic signature for each selected TREC topic. Documents are separated into relevant and non-relevant sets according to their TREC relevancy judgments for each topic. We then run each document through a part-of-speech tagger and convert each word into its root form based on the WordNet lexical database. We also collect individual root word (unigram) frequency, two consecutive non-stopword\(^8\) (bigram) frequency, and three consecutive non-stopwords (trigram) frequency to facilitate the computation of the \(-2\log\lambda\) value for each term. We expect high ranking bigram and trigram signature terms to be very informative. We set the cutoff associated weight at 10.83 with confidence level \(\alpha = 0.001\) by looking up a \(x^2\) statistical table.

Table 2 shows the top 10 unigram, bigram, and trigram topic signature terms for each topic\(^9\). Several conclusions can be drawn directly. Terms with high \(-2\log\lambda\) are indeed good indicators for their corresponding topics. The \(-2\log\lambda\) values decrease as the number of words in a term increases. This is reasonable, since longer terms usually occur less often than their constituents. However, bigram terms are more informative than unigram terms as we can observe: jail/prison overcrowding of topic 151, tobacco industry of topic 257, computer security of topic 258, and solar energy/power of topic 271. These automatically generated signature terms closely resemble or equal the given short TREC topic descriptions. Although trigram terms shown in the table, such as federal court order, philip morris irr, jet propulsion laboratory, and mobile telephone system are also meaningful, they do not demonstrate the closer term relationship among other terms in their respective topics that is seen in the bigram cases. We expect that more training data can improve the situation.

We notice that the \(-2\log\lambda\) values for topic 258 are higher than those of the other three topics. As indicated by (Mani et al., 1998) the majority of relevant documents for topic 258 have the query topic as their main theme; while the others mostly have the query topics as their subsidiary themes. This implies that it is too liberal to assume all the terms in relevant documents of the other three topics are relevant. We plan to apply text segmentation algorithms such as TextTiling (Hearst, 1997) to segment documents into subtopic units. We will then perform the topic signature creation procedure only on the relevant units to prevent inclusion of noise terms.

---

\(^7\)These four topics are:  
- topic 151: Overcrowded Prisons, 1211 texts, 85 relevant;  
- topic 257: Cigarette Consumption, 1727 texts, 126 relevant;  
- topic 258: Computer Security, 1701 texts, 49 relevant;  
- topic 271: Solar Power, 1555 texts, 59 relevant.

\(^8\)A relevant document only needs to answer at least one of the five questions.

\(^9\)We use the stopword list supplied with the SMART retrieval system.

\(^{10}\)The \(-2\log\lambda\) values are not comparable across ngram categories, since each ngram category has its own sample space.
6.1 Comparing Summary Extraction Effectiveness Using Topic Signatures, Tfidf, and Baseline Algorithms

In order to evaluate the effectiveness of topic signatures used in summary extraction, we compare the summary sentences extracted by the topic signature module, baseline module, and tfidf modules with human annotated model summaries. We measure the performance using a combined measure of recall ($R$) and precision ($P$), and F-score is defined by:

$$F = \frac{1 + \beta^2}{\beta^2 P + R},$$

where

$$R = \frac{N_{out}}{N_m},$$

$$P = \frac{N_{out}}{N_s},$$

$N_{out}$: number of sentences extracted from the last 5 documents

$N_m$: number of sentences in the model summary

$N_s$: number of sentences of the system

$\beta$: relative importance of $R$ and $P$

We assume equal importance of recall and precision and set $\beta$ to 1. We evaluate the baseline (position) module scores each sentence by its position in the text. We compare the topic signature module with the tfidf module.

| Topic | Signature Terms of Topic 151 | Overcrowded Prisons |
|-------|----------------------------|----------------------|
| rad | $<$tag>rad</tag> | $<$tag>rad</tag> |
| gun | $<$tag>gun</tag> | $<$tag>gun</tag> |
| early | $<$tag>early</tag> | $<$tag>early</tag> |
| late | $<$tag>late</tag> | $<$tag>late</tag> |
| state | $<$tag>state</tag> | $<$tag>state</tag> |
| county | $<$tag>county</tag> | $<$tag>county</tag> |

Table 2: Top 10 signature terms of unigram, bigram, and trigram for four TREC topics.
marks. The number of documents with answer 
keys are listed in the row labeled: "# of Relevant 
Docs Used in Training". To ensure we utilize all 
the available data and conduct a sound evaluation, 
we perform a three-fold cross validation. We re-
served one-third of documents as test set, use the rest as 
training set, and repeat three times with non-
overlapping test set. Furthermore, we use only uni-
gram topic signatures for evaluation.

The result is shown in Figure 2 and Table 3. We 
find that the topic signature method outperforms the 
other two methods and the tfidf method performs 
poorly. Among 40 possible test points for four topics 
with 10% summary length increment (0% means se-
lect at least one sentence) as shown in Table 3, the 
topic signature method beats the baseline method 
34 times. This result is really encouraging and indi-
icates that the topic signature method is a worthy 
addition to a variety of text summarization methods.

6.2 Enriching Topic Signatures Using 
Existing Ontologies

We have shown in the previous sections that topic 
signatures can be used to approximate topic iden-
tification at the lexical level. Although the au-
tomatically acquired signature terms for a specific 
topic seem to be bound by unknown relationships as 
shown in Table 2, it is hard to image how we can 
enrich the inherent flat structure of topic signatures 
as defined in Equation 1 to a construct as complex 
as a MUC template or script.

As discussed in (Agirre et al., 2000), we propose 
using an existing ontology such as SENSUS (Knight 
and Luk, 1994) to identify signature term relations. 
The external hierarchical framework can be used to 
generalize topic signatures and suggest richer rep-
resentations for topic signatures. Automated entity 
recognizers can be used to classify unknown enti-
ties into their appropriate SENSUS concept nodes. 
We are also investigating other approaches to auto-
matically learn signature term relations. The idea 
mentioned in this paper is just a starting point.

7 Conclusion

In this paper we presented a procedure to automati-
\cally acquire topic signatures and evaluated the 
effectiveness of applying topic signatures to extract topic 
relevant sentences against two other methods. The 
topic signature method outperforms the baseline and the 
tfidf methods for all test topics. Topic signatures 
can not only recognize related terms (topic identifi-
cation), but group related terms together under one 
target concept (topic interpretation). Topic identifi-
cation and interpretation are two essential steps in 
a typical automated text summarization system as 
we present in Section 3.

Topic signatures can also been viewed as an in-
verse process of query expansion. Query expansion 
intends to alleviate the word mismatch problem in 
information retrieval, since documents are normally 
written in different vocabulary. How to automati-
cally identify highly correlated terms and use them 
to improve information retrieval performance has 
been a main research issue since late 1960's. Re-
cent advances in the query expansion (Xu and Croft, 
1996) can also shed some light on the creation of 
topic signatures. Although we focus the use of topic 
signatures to aid text summarization in this paper, 
we plan to explore the possibility of applying topic 
signatures to perform query expansion in the future.

The results reported are encouraging enough to 
allow us to continue with topic signatures as the ve-
hicle for a first approximation to world knowledge. 
We are now busy creating a large number of signa-
tures to overcome the world knowledge acquisition 
problem and use them in topic interpretation.

8 Acknowledgements

We thank the anonymous reviewers for very use-
ful suggestions. This work is supported in part by 
DARPA contract N66001-97-9538.

References

Enrico Agirre, Olatz Ansa, Eduard Hovy, and David 
Martinez. 2000. Enriching very large ontologies 
using the www. In Proceedings of the Workshop 
on Ontology Construction of the European Con-
ference of AI (ECAI).

Kenneth Church and Patrick Hanks. 1990. Word 
association norms, mutual information and lexicog-
raphy. In Proceedings of the 28th Annual Meeting 
of the Association for Computational Linguistics 
(ACL-90), pages 76-83.

Thomas Cover and Joy A. Thomas. 1991. Elements 
of Information Theory. John Wiley & Sons.

Gerald DeJong. 1982. An overview of the FRUMP 
system. In Wendy G. Lehnert and Martin H. 
Ringle, editors, Strategies for natural language 
processing, pages 149-76. Lawrence Erlbaum 
Associates.

Ted Dunning. 1993. Accurate methods for the 
statistics of surprise and coincidence. Computa-
tional Linguistics, 19:61-74.

Marc Hearst. 1997. TextTiling: Segmenting text 
into multi-paragraph subtopic passages. Compu-
tational Linguistics, 23:33-61.

Eduard Hovy and Chin-Yew Lin. 1999. Automated 
text summarization in SUMMARIST. In Inder-
jeet Mani and Mark T. Maybury, editors, Advances 
in Automatic Text Summarization, chapter 8, pages 81-94. MIT Press.

Kevin Knight and Steve K. Luk. 1994. Building a 
large knowledge base for machine translation. In 
Proceedings of the Eleventh National Conference 
on Artificial Intelligence (AAAI-94).
Figure 2: F-measure vs. summary length for all four topics. Topic signature clearly outperforms tfidf and baseline except for the case of topic 258 where performance for the three methods are roughly equal.

Table 3: F-measure performance difference compared to baseline method in percentage. Columns indicate at different summary lengths related to full length documents. Values in the baseline rows are F-measure scores. Values in the tfidf and topic signature rows are performance increase or decrease divided by their corresponding baseline scores and shown in percentage.

|        | 0%    | 10%   | 20%   | 30%   | 40%   | 50%   | 60%   | 70%   | 80%   | 90%   | 100%  |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Baseline | 0.54  | 0.36  | 0.36  | 0.29  | 0.25  | 0.22  | 0.21  | 0.24  | 0.27  | 0.29  | 0.32  |
| tfidf   | 0.52  | 0.34  | 0.34  | 0.27  | 0.23  | 0.20  | 0.19  | 0.22  | 0.25  | 0.27  | 0.31  |
| Baseline | 0.54  | 0.36  | 0.36  | 0.29  | 0.25  | 0.22  | 0.21  | 0.24  | 0.27  | 0.29  | 0.32  |
| tfidf   | 0.52  | 0.34  | 0.34  | 0.27  | 0.23  | 0.20  | 0.19  | 0.22  | 0.25  | 0.27  | 0.31  |
| Baseline | 0.54  | 0.36  | 0.36  | 0.29  | 0.25  | 0.22  | 0.21  | 0.24  | 0.27  | 0.29  | 0.32  |
| tfidf   | 0.52  | 0.34  | 0.34  | 0.27  | 0.23  | 0.20  | 0.19  | 0.22  | 0.25  | 0.27  | 0.31  |

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Leo Obrat, Thérèse Firmin, Michael Chzanowski, and Beth Sundheim. 1998. The TIPSTER SUMMAC text summarization evaluation final report. Technical Report MTR98W00000138, The MITRE Corporation.

Christopher Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Kathleen McKeown and Dragomir R. Radev. 1999. Generating summaries of multiple news articles. In Inderjeet Mani and Mark T. Maybury, editors, Advances in Automatic Text Summarization, chapter 24, pages 381-389, MIT Press.

Ellen Riloff and Jeffrey Lorenzen. 1999. Extraction-based text categorization: Generating domain-specific role relationships automatically. In Tomek Strzalkowski, editor, Natural Language Information Retrieval. Kluwer Academic Publishers.

Ellen Riloff. 1996. An empirical study of automated dictionary construction for information extraction in three domains. Artificial Intelligence Journal, 85, August.

SAIC. 1998. Introduction to information extraction. http://www.macsai.com.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 4-11.