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 evaluaton 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; *eat* 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[1] 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[2] (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).

to text categorization and information extraction, the concept of *relevancy signatures* introduced by her is very similar to the *topic signatures* we proposed in this paper. Relevancy 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 (Hovy 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':

> summarization = topic identification + topic interpretation + 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 counting and topic signatures to tackle the fusion problem.

Summary Generation: SUMMARIST can produce keyword and extract type summaries.

[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 (Hearst, 1997) can be used to find subtopic segments in text.

ABCNEWS.com : Delay in Handing Flight 990 Probe to FBI

NTSB Chairman James Hall says Egyptian officials want to review results of the investigation into the crash of EgyptAir Flight 990 before the case is turned over to the FBI.

Nov. 16 - U.S. investigators appear to be leaning more than ever toward the possibility that one of the co-pilots of EgyptAir Flight 990 may have deliberately crashed the plane last month, killing all 217 people on board.

However, U.S. officials say the National Transportation Safety Board will delay transferring the investigation of the Oct. 31 crash to the FBI - the agency that would lead a criminal probe - for at least a few days, to allow Egyptian experts to review evidence in the case.

Suspicions of foul play were raised after investigators listening to a tape from the cockpit voice recorder isolated a religious prayer or statement made by the co-pilot just before the plane's autopilot was turned off and the plane began its initial plunge into the Atlantic Ocean off Massachusetts' Nantucket Island.

Over the past week, after much effort, the NTSB and the Navy succeeded in locating the plane's two "black boxes," the cockpit voice recorder and the flight data recorder.

The tape indicates that shortly after the plane leveled off at its cruising altitude of 33,000 feet, the chief pilot of the aircraft left the plane's cockpit, leaving one of the two co-pilots alone there as the aircraft began its descent.

Figure 1: A Nov. 16 1999 ABC News page summary generated by SUMMARIST.

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 *tfidf* weights calculator, sentence length calculator, and sentence location identifier. We only activate the position module, the *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.

summary. We thus defined a topic signature as a family of related terms, as follows:

$$\begin{aligned} TS &= \{topic, signature\} \\ &= \{topic, < (t_1, w_1), ..., (t_n, w_n) >\} \end{aligned} \quad (1)$$

where *topic* is the target concept and *signature* is a vector of related terms. Each t_i is an term highly correlated to *topic* with association weight w_i. The number of related terms n can be set empirically according to a cutoff associated weight. We describe how to acquire related terms and their associated weights in the next section.

4.1 Signature Term Extraction and Weight Estimation

On the assumption that semantically related terms tend to co-occur, one can construct topic signatures from preclassified text using the χ^2 test, mutual information, or other standard statistic tests and information-theoretic measures. Instead of χ^2, we use *likelihood ratio* (Dunning, 1993) λ, since λ is more appropriate for sparse data than χ^2 test and the quantity $-2log\lambda$ is asymptotically χ^2 distributed[5]. Therefore, we can determine the confidence level for a specific $-2log\lambda$ value by looking up χ^2 distribution table and use the value to select an appropriate cutoff associated weight.

We have documents preclassified into a set $\mathcal{R}$ of relevant texts and a set $\tilde{\mathcal{R}}$ of nonrelevant texts for a given topic. Assuming the following two hypotheses:

Hypothesis 1 (H_1): $P(\mathcal{R}|t_i) = p = P(\mathcal{R}|\tilde{t_i})$, i.e. the relevancy of a document is independent of t_i.

Hypothesis 2 (H_2): $P(\mathcal{R}|t_i) = p_1 \neq p_2 = P(\mathcal{R}|\tilde{t_i})$, i.e. the presence of t_i indicates strong relevancy assuming $p_1 \gg p_2$.

and the following 2-by-2 contingency table:

	$\mathcal{R}$	$\tilde{\mathcal{R}}$
t_i	O_{11}	O_{12}
$\tilde{t_i}$	O_{21}	O_{22}

where O_{11} is the frequency of term t_i occurring in the relevant set, O_{12} is the frequency of term t_i occurring in the nonrelevant set, O_{21} is the frequency of term $\tilde{t_i} \neq t_i$ occurring in the relevant set, O_{22} is the frequency of term $\tilde{t_i} \neq t_i$ occurring in the nonrelevant set.

Assuming a binomial distribution:

$$b(k; n, x) = \binom{n}{k} x^k (1-x)^{(n-k)} \quad (2)$$

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 $-2log\lambda$ value is then computed as follows:

$$\begin{aligned} &= -2log\frac{L(H_1)}{L(H_2)} \\ &= -2log\frac{b(O_{11}; O_{11}+O_{12}, p)b(O_{21}; O_{21}+O_{22}, p)}{b(O_{11}; O_{11}+O_{12}, p_1)b(O_{21}; O_{21}+O_{22}, p_2)} \\ &= -2((O_{11}+O_{21})log\ p + (O_{12}+O_{22})log\ (1-p) - \\ &\quad (O_{11}log\ p_1 + O_{12}log\ (1-p_1) + O_{21}log\ p_2 + O_{22}log\ (1-p_2))) \quad (3) \\ &= 2N \times (\mathcal{H}(\mathcal{R}) - \mathcal{H}(\mathcal{R}|\mathcal{T})) \quad (4) \\ &= 2N \times \mathcal{I}(\mathcal{R}; \mathcal{T}) \quad (5) \end{aligned}$$

where $N = O_{11} + O_{12} + O_{21} + O_{22}$ is the total number of term occurrence in the corpus, $\mathcal{H}(\mathcal{R})$ is the entropy of terms over relevant and nonrelevant sets of documents, $\mathcal{H}(\mathcal{R}|\mathcal{T})$ is the entropy of a given term over relevant and nonrelevant sets of documents, and $\mathcal{I}(\mathcal{R}; \mathcal{T})$ is the mutual information between document relevancy and a given term. Equation 5 indicates that mutual information[6] is an equivalent measure to likelihood ratio when we assume a binomial distribution and a 2-by-2 contingency table.

To create topic signature for a given topic, we:

1. classify documents as relevant or nonrelevant according to the given topic
2. compute the $-2log\lambda$ value using Equation 3 for each term in the document collection
3. rank terms according to their $-2log\lambda$ value
4. select a confidence level from the χ^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

[5] This assumes that the ratio is between the maximum likelihood estimate over a subpart of the parameter space and the maximum likelihood estimate over the entire parameter space. See (Manning and Schütze, 1999) pages 172 to 175 for details.

[6] 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).

TREC Topic Description

<num> Number: 151
<title> Topic: Coping with overcrowded prisons
<desc> Description:
The document will provide information on jail and prison overcrowding and how inmates are forced to cope with those conditions; or it will reveal plans to relieve the overcrowded condition.
<narr> Narrative:
A relevant document will describe scenes of overcrowding that have become all too common in jails and prisons around the country. The document will identify how inmates are forced to cope with those overcrowded conditions, and/or what the Correctional System is doing, or planning to do, to alleviate the crowded condition.
</top>

Test Questions

Q1	What are name and/or location of the correction facilities where the reported overcrowding exists?
Q2	What negative experiences have there been at the overcrowded facilities (whether or not they are thought to have been caused by the overcrowding)?
Q3	What measures have been taken/planned/recommended (etc.) to accommodate more inmates at penal facilities, e.g., doubling up, new construction?
Q4	What measures have been taken/planned/recommended (etc.) to reduce the number of new inmates, e.g., moratoriums on admission, alternative penalties, programs to reduce crime/recidivism?
Q5	What measures have been taken/planned/recommended (etc.) to reduce the number of existing inmates at an overcrowded facility, e.g., granting early release, transfering to uncrowded facilities?

Sample Answer Keys

<DOCNO> AP891027-0063 </DOCNO>
<FILEID>AP-NR-10-27-89 0615EDT</FILEID>
<1ST_LINE>r a PM-ChainedInmates 10-27 0335</1ST_LINE>
<2ND_LINE>PM-Chained Inmates,0344</2ND_LINE>
<HEAD>Inmates Chained to Walls in Baltimore Police Stations</HEAD>
<DATELINE>BALTIMORE (AP) </DATELINE>
<TEXT>
<Q3>Prisoners are kept chained to the walls of local police lockups for as long as three days at a time because of overcrowding in regular jail cells, police said.</Q3>
Overcrowding at the <Q1>Baltimore County Detention Center</Q1> has forced police to ...
</TEXT>

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

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[8], and a sample relevant document with answer keys markup.

[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.

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 nonrelevant 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[9] (bigram) frequency, and three consecutive non-stopwords (trigram) frequency to facilitate the computation of the $-2log\lambda$ 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 χ^2 statistical table.

Table 2 shows the top 10 unigram, bigram, and trigram topic signature terms for each topic [10]. Several conclusions can be drawn directly. Terms with high $-2log\lambda$ are indeed good indicators for their corresponding topics. The $-2log\lambda$ 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 rjr*, *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 $-2log\lambda$ 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.

[9]We use the stopword list supplied with the SMART retrieval system.

[10]The $-2log\lambda$ values are not comparable across ngram categories, since each ngram category has its own sample space.

Topic 10 Signature Terms of Topic 151 — Overcrowded Prisons					
Unigram	$-2log\lambda$	Bigram	$-2log\lambda$	Trigram	$-2log\lambda$
jail	461.044	county jail	160.273	federal court order	45.960
county	408.821	early release	85.361	comply consent decree	35.121
overcrowding	342.349	state prison	74.372	dekalb county sheriff	35.121
inmate	234.765	state prisoner	67.666	gov. jo frank	35.121
sheriff	154.440	day fine	61.465	joe frank harris	35.121
state	151.940	jail overcrowding	61.329	prisoner county jail	35.121
prisoner	148.178	court order	60.090	state prison county	28.043
prison	145.306	local jail	56.440	't put prison	26.341
city	133.477	prison overcrowding	55.373	county jail state	26.341
overcrowded	128.008	central facility	52.909	hold local jail	26.341
Topic 10 Signature Terms of Topic 257 — Cigarette Consumption					
Unigram	$-2log\lambda$	Bigram	$-2log\lambda$	Trigram	$-2log\lambda$
cigarette	476.038	tobacco industry	80.768	philip morris rjr	28.061
tobacco	313.017	bn cigarette	67.429	rothmans benson hedge	26.969
smoking	284.198	philip morris	54.073	lung cancer death	22.214
smoke	159.134	cigarette year	48.045	qtr firm chg	21.418
rothmans	156.675	rothmans international	44.434	qtr qtr firm	21.418
osha	148.372	tobacco smoke	44.269	bn bn bn	20.226
seita	126.421	sir patrick	40.455	consumption bn cigarette	20.226
ban	113.849	cigarette company	39.399	great american smokeout	20.226
smoker	104.110	cent market	36.223	lung cancer heart	20.226
bat	79.903	tax increase	36.223	malaysian singapore company	20.226
Topic 10 Signature Terms of Topic 258 — Computer Security					
Unigram	$-2log\lambda$	Bigram	$-2log\lambda$	Trigram	$-2log\lambda$
computer	1159.351	computer security	213.331	jet propulsion laboratory	98.854
virus	927.674	graduate student	178.588	robert t mo	98.854
hacker	887.377	computer system	146.328	cornell university graduate	79.081
morris	666.392	research center	132.413	lawrence berkeley laboratory	79.081
cornell	385.684	computer virus	126.033	nasa jet propulsion	79.081
university	305.958	cornell university	108.741	university graduate student	79.081
system	290.347	nuclear weapon	107.283	lawrence livermore national	69.195
laboratory	287.521	military computer	106.522	livermore national laboratory	69.195
lab	225.516	virus program	106.522	computer security expert	66.196
nuclary	128.515	west german	82.210	security center bethesda	49.423
Topic 10 Signature Terms of Topic 271 — Solar Power					
Unigram	$-2log\lambda$	Bigram	$-2log\lambda$	Trigram	$-2log\lambda$
solar	484.315	solar energy	268.521	division multiple access	31.347
mazda	308.015	solar power	94.210	mobile telephone service	31.347
leo	276.932	christian aid	86.211	british technology group	23.510
iridium	258.705	leo system	70.535	earth height mile	23.510
pavilion	203.811	mobile telephone	70.535	financial backing iridium	23.510
pound	128.121	iridium project	62.697	global mobile satellite	23.510
tower	126.353	real goods	61.901	handheld mobile telephone	23.510
lookout	125.406	science park	54.859	mobile satellite system	23.510
inmarsat	109.728	solar concentrator	54.859	motorola iridium project	23.510
boydston	78.373	bp solar	31.347	active solar system	15.673

Table 2: Top 10 signature terms of unigram, bigram, and trigram for four TREC topics.

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), F. F-score is defined by:

$$F = \frac{(1+\beta^2)PR}{\beta^2 P + R}, \text{ where}$$

$$R = \frac{N_{me}}{N_m} \quad (6)$$

$$P = \frac{N_{me}}{N_e} \quad (7)$$

N_{me} : *# of sentences extrated that also appear in the model summary*

N_m : *# of sentences in the model summary*

N_e : *# of sentences extracted by the system*

β : *relative importance of R and P*

We assume equal importance of recall and precision and set β to 1 in our experiments. The baseline (position) module scores each sentence by its position in the text. The first sentence gets the highest score, the last sentence the lowest. The baseline method is expected to be effective for news genre. The *tfidf* module assigns a score to a term t_i according to the product of its frequency within a document j (tf_{ij}) and its inverse document frequency ($idf_j = log\frac{N}{df_j}$). N is the total number of documents in the corpus and df_j is the number of documents containing term t_i.

The topic signature module scans each sentence, assigning to each word that occurs in a topic signature the weight of that keyword in the topic signature. Each sentence then receives a topic signature score equal to the total of all signature word scores it contains, normalized by the highest sentence score. This score indicates the relevance of the sentence to the signature topic.

SUMMARIST produced extracts of the same texts separately for each module, for a series of extracts ranging from 0% to 100% of the original text.

Although many relevant documents are available for each topic, only some of them have answer key

markups. 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 reserve 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 unigram 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 select 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 indicates 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 identification at the lexical level. Although the automatically 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 representations for topic signatures. Automated entity recognizers can be used to classify unknown entities into their appropriate SENSUS concept nodes. We are also investigating other approaches to automatically 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 automatically acquire topic signatures and valuated 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 identification), but group related terms together under one target concept (topic interpretation). Topic identification 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 inverse 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 automatically identify highly correlated terms and use them to improve information retrieval performance has been a main research issue since late 1960's. Recent 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 vehicle for a first approximation to world knowledge. We are now busy creating a large number of signatures to overcome the world knowledge acquisition problem and use them in topic interpretation.

8 Acknowledgements

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

References

Eneko Agirre, Olatz Ansa, Eduard Hovy, and David Martínez. 2000. Enriching very large ontologies using the www. In *Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI)*.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. 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. *Computational Linguistics*, 19:61–74.

Marti Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64.

Eduard Hovy and Chin-Yew Lin. 1999. Automated text summarization in SUMMARIST. In Inderjeet 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)*.

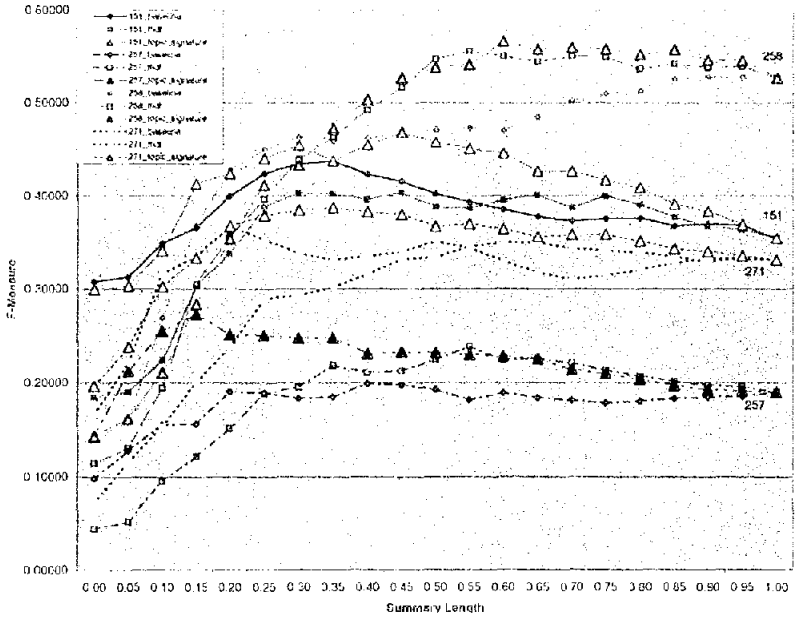

Figure 2: F-measure vs. summary length for all four topics. Topic signature clearly outperform *tfidf* and baseline except for the case of topic 258 where performance for the three methods are roughly equal.

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
151_baseline	0.308	0.349	0.400	0.434	0.423	0.402	0.385	0.373	0.376	0.370	0.355
151_tfidf	-39.93	-35.82	-15.26	-7.22	-6.35	-3.50	2.55	3.76	3.83	-0.75	0.00
151_topic_sig	-2.76	-2.19	+6.02	+4.58	+7.48	+13.77	+15.63	+14.17	+8.66	+3.50	0.00
257_baseline	0.098	0.155	0.191	0.184	0.199	0.193	0.189	0.181	0.181	0.185	0.190
257_tfidf	-55.11	-38.56	-20.59	+6.54	+6.06	+16.44	+18.34	+21.08	+14.46	+7.00	0.00
257_topic_sig	+45.53	+64.06	+31.86	+34.91	+16.07	+20.40	+20.60	+18.01	+12.48	+4.24	0.00
258_baseline	0.141	0.270	0.428	0.463	0.462	0.471	0.470	0.502	0.512	0.528	0.527
258_tfidf	-18.84	-27.88	-16.57	5.21	+6.61	+16.13	+16.90	+9.56	+4.74	+1.92	0.00
258_topic_sig	+1.55	-21.82	-17.19	-6.56	+8.96	+14.16	+20.40	+11.44	+7.74	+3.43	0.00
271_baseline	0.167	0.316	0.368	0.338	0.335	0.351	0.341	0.310	0.321	0.333	0.332
271_tfidf	-50.75	-51.35	-35.25	-13.34	-5.99	1.83	5.96	+10.97	15.76	-1.22	0.00
271_topic_sig	+17.97	-4.40	+6.09	+13.67	+14.10	+4.83	+10.26	+15.70	+9.65	+2.29	0.00

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.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Leo Obrst, Thérèse Firmin, Michael Chrzanowski, and Beth Sundheim. 1998. The TIPSTER SUMMAC text summarization evaluation final report. Technical Report MTR98W0000138, The MITRE Corporation.

Christopher Manning and Hinrich Schütze. 1999. *Foudations 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.muc.saic.com.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11.