Using Syntactic Information to Extract Relevant Terms for Multi-Document Summarization

Enrique Amigó **Julio Gonzalo** **Víctor Peinado** **Anselmo Peñas** **Felisa Verdejo**
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
c/ Juan del Rosal, 16 - 28040 Madrid - Spain
http://nlp.uned.es

Abstract

The identification of the key concepts in a set of documents is a useful source of information for several information access applications. We are interested in its application to multi-document summarization, both for the automatic generation of summaries and for interactive summarization systems.

In this paper, we study whether the syntactic position of terms in the texts can be used to predict which terms are good candidates as key concepts. Our experiments show that a) distance to the verb is highly correlated with the probability of a term being part of a key concept; b) subject modifiers are the best syntactic locations to find relevant terms; and c) in the task of automatically finding key terms, the combination of statistical term weights with shallow syntactic information gives better results than statistical measures alone.

1 Introduction

The fundamental question addressed in this article is: can syntactic information be used to find the key concepts of a set of documents? We will provide empirical answers to this question in a multi-document summarization environment.

The identification of key terms out of a set of documents is a common problem in information access applications and, in particular, in text summarization: a fragment containing one or more key concepts can be a good candidate to be part of a summary.

In single-document summarization, key terms are usually obtained from the document title or heading (Edmundson, 1969; Preston, 1994; Kupiec et al., 1995). In multi-document summarization, however, some processing is needed to identify key concepts (Lin and Hovy, 2002; Kraaij et al., 2002; Schlesinger et al., 2002). Most approaches are based on statistical criteria.

Criteria to elaborate a manual summary depend, by and large, on the user interpretation of both the information need and the content of documents. This is why this task has also been attempted from an interactive perspective (Boguraev et al., 1998; Buyukkokten et al., 1999; Neff and Cooper, 1999; Jones et al., 2002; Leuski et al., 2003). A standard feature of such *interactive summarization assistants* is that they offer a list of relevant terms (automatically extracted from the documents) which the user may select to decide or refine the focus of the summary.

Our hypothesis is that the key concepts of a document set will tend to appear in certain syntactic functions along the sentences and clauses of the texts. To confirm this hypothesis, we have used a test bed with manually produced summaries to study:

- which are the most likely syntactic functions for the key concepts manually identified in the document sets.
- whether this information can be used to automatically extract the relevant terms from a set of documents, as compared to standard statistical term weights.

Our reference corpus is a set of 72 lists of key concepts, manually elaborated by 9 subjects on 8 different topics, with 100 documents per topic. It was built to study *Information Synthesis* tasks (Amigo et al., 2004) and it is, to the best of our knowledge, the multi-document summarization testbed with a largest number of documents per topic. This feature enables us to obtain reliable statistics on term occurrences and prominent syntactic functions.

The paper is organized as follows: in Section 2 we review the main approaches to the evaluation of automatically extracted key concepts for summarization. In Section 3 we describe the creation of the reference corpus. In Section 4 we study the correlation between key concepts and syntactic function in texts, and in Section 5 we discuss the experimental results of syntactic function as a predictor to extract

key concepts. Finally, in Section 6 we draw some conclusions.

2 Evaluation of automatically extracted key concepts

It is necessary, in the context of an interactive summarization system, to measure the quality of the terms suggested by the system, i.e., to what extent they are related to the key topics of the document set.

(Lin and Hovy, 1997) compared different strategies to generate lists of relevant terms for summarization using Topic Signatures. The evaluation was extrinsic, comparing the quality of the summaries generated by a system using different term lists as input. The results, however, cannot be directly extrapolated to interactive summarization systems, because the evaluation does not consider how informative terms are for a user.

From an interactive point of view, the evaluation of term extraction approaches can be done, at least, in two ways:

- Evaluating the summaries produced in the interactive summarization process. This option is difficult to implement (how do we evaluate a human produced summary? What is the reference gold standard?) and, in any case, it is too costly: every alternative approach would require at least a few additional subjects performing the summarization task.
- Comparing automatically generated term lists with manually generated lists of key concepts. For instance, (Jones et al., 2002) describes a process of supervised learning of key concepts from a training corpus of manually generated lists of phrases associated to a single document.

We will, therefore, use the second approach, evaluating the quality of automatically generated term lists by comparing them to lists of key concepts which are generated by human subjects after a multi-document summarization process.

3 Test bed: the ISCORPUS

We have created a reference test bed, the ISCORPUS[1] (Amigo et al., 2004) which contains 72 manually generated *reports* summarizing the relevant information for a given topic contained in a large document set.

For the creation of the corpus, nine subjects performed a complex multi-document summarization task for eight different topics and one hundred relevant documents per topic. After creating each topic-oriented summary, subjects were asked to make a list of relevant concepts for the topic, in two categories: relevant entities (people, organizations, etc.) and relevant factors (such as "ethnic conflicts" as the origin of a civil war) which play a key role in the topic being summarized.

These are the relevant details of the ISCORPUS test bed:

3.1 Document collection and topic set

We have used the Spanish CLEF 2001-2003 news collection testbed (Peters et al., 2002), and selected the eight topics with the largest number of documents manually judged as relevant from the CLEF assessment pools. All the selected CLEF topics have more than one hundred documents judged as relevant by the CLEF assessors; for homogeneity, we have restricted the task to the first 100 documents for each topic (using a chronological order).

This set of eight CLEF topics was found to have two differentiated subsets: in six topics, it is necessary to study how a situation evolves in time: the importance of every event related to the topic can only be established in relation with the others. The *invasion of Haiti by UN and USA troops* is an example of such kind of topics. We refer to them as "Topic Tracking" (TT) topics, because they are suitable for such a task. The other two questions, however, resemble "Information Extraction" (IE) tasks: essentially, the user has to detect and describe instances of a generic event (for instance, *cases of hunger strikes* and *campaigns against racism in Europe* in this case); hence we will refer to them as IE summaries.

3.2 Generation of manual summaries

Nine subjects between 25 and 35 years-old were recruited for the manual generation of summaries. All subjects were given an in-place detailed description of the task, in order to minimize divergent interpretations. They were told they had to generate summaries with a maximum of information about every topic within a 50 sentence space limit, using a maximum of 30 minutes per topic. The 50 sentence limit can be temporarily exceeded and, once the 30 minutes have expired, the user can still remove sentences from the summary until the sentence limit is reached back.

3.3 Manual identification of key concepts

After summarizing every topic, the following questionnaire was filled in by users:

- Who are the main people involved in the topic?

[1] Available at http://nlp.uned.es/ISCORPUS.

- What are the main organizations participating in the topic?
- What are the key factors in the topic?

Users provided free-text answers to these questions, with their freshly generated summary at hand. We did not provide any suggestions or constraints at this point, except that a maximum of eight slots were available per question (i.e., a maximum of $8X3 = 24$ key concepts per topic, per user).

This is, for instance, the answer of one user for a topic about the invasion of Haiti by UN and USA troops:

People	**Organizations**
Jean Bertrand Aristide	ONU *(UN)*
Clinton	EEUU *(USA)*
Raoul Cedras	OEA *(OAS)*
Philippe Biambi	
Michel Josep Francois	

Factors
militares golpistas *(coup attempting soldiers)*
golpe militar *(coup attempt)*
restaurar la democracia *(reinstatement of democracy)*

Finally, a single list of key concepts is generated for each topic, joining all the answers given by the nine subjects. These lists of key concepts constitute the gold standard for all the experiments described below.

3.4 Shallow parsing of documents

Documents are processed with a robust shallow parser based in finite automata. The parser splits sentences in chunks and assigns a label to every chunk. The set of labels is:

- `[N]`: noun phrases, which correspond to names or adjectives preceded by a determiner, punctuation sign, or beginning of a sentence.
- `[V]`: verb forms.
- `[Mod]`: adverbial and prepositional phrases, made up of noun phrases introduced by an adverb or preposition. Note that this is the mechanism to express NP modifiers in Spanish (as compared to English, where noun compounding is equally frequent).
- `[Sub]`: words introducing new subordinate clauses within a sentence (*que, cuando, mientras*, etc.).
- `[P]`: Punctuation marks.

This is an example output of the chunker:

Previamente `[Mod]` ,`[P]`el presidente Bill Clinton `[N]` había dicho `[V]` que `[Sub]` tenemos `[V]` la obligacion `[N]` de cambiar la política estadounidense `[Mod]` que `[Sub]` no ha funcionado `[V]` en Haití `[Mod]`.`[P]`

Although the precision of the parser is limited, the results are good enough for the statistical measures used in our experiments.

4 Distribution of key concepts in syntactic structures

We have extracted empirical data to answer these questions:

- Is the probability of finding a key concept correlated with the distance to the verb in a sentence or clause?
- Is the probability of finding a key concept in a noun phrase correlated with the syntactic function of the phrase (subject, object, etc.)?
- Within a noun phrase, where is it more likely to find key concepts: in the noun phrase head, or in the modifiers?

We have used certain properties of Spanish syntax (such as being an SVO language) to decide which noun phrases play a subject function, which are the head and modifiers of a noun phrase, etc. For instance, NP modifiers usually appear after the NP head in Spanish, and the specification of a concept is usually made from left to right.

4.1 Distribution of key concepts with verb distance

Figure 1 shows, for every topic, the probability of finding a word from the manual list of key concepts in fixed distances from the verb of a sentence. Stop words are not considered for computing word distance. The broader line represents the average across topics, and the horizontal dashed line is the average probability across all positions, i.e., the probability that a word chosen at random belongs to the list of key concepts.

The plot shows some clear tendencies in the data: the probability gets higher when we get close to the verb, falls abruptly after the verb, and then grows steadily again. For TT topics, the probability of finding relevant concepts immediately before the verb is 56% larger than the average (0.39 before the verb, versus 0.25 in any position). This is true not only as an average, but also for all individual TT topics. This can be an extremely valuable result: it shows a direct correlation between the position of a term in a sentence and the importance of the term in the topic. Of course, this direct distance to the verb should be adapted for languages with different syntactic properties, and should be validated for different domains.

The behavior of TT and IE topics is substantially different. IE topics have smaller probabilities overall, because there are less key concepts common to all documents. For instance, if the topic is “cases of hunger strikes”, there is little in common between

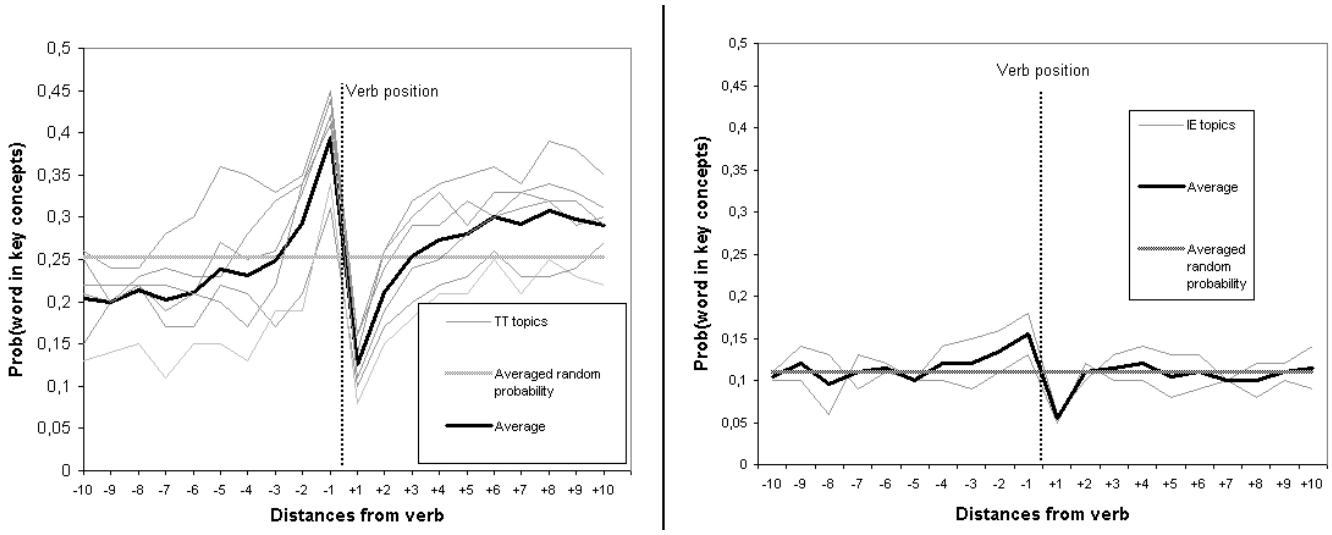

Figure 1: Probability of finding key concepts at fixed distances from verb

all cases of hunger strikes found in the collection; each case has its own relevant people and organizations, for instance. Users try to make abstraction of individual cases to write key concepts, and then the number of key concepts is smaller. The tendency to have larger probabilities just before the verb and smaller probabilities just after the verb, however, can also be observed for IE topics.

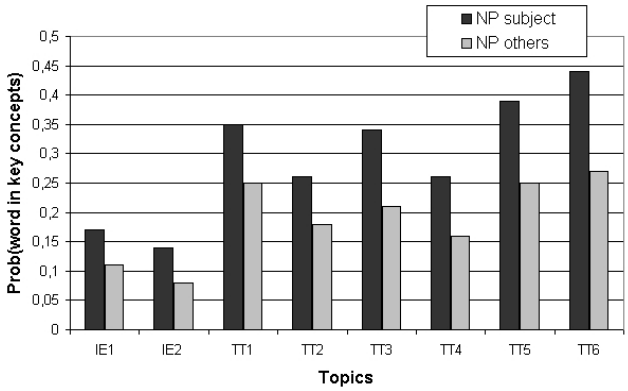

Figure 2: Probability of finding key concepts in subject NPs versus other NPs

4.2 Key Concepts and Noun Phrase Syntactic Function

We wanted also to confirm that it is more likely to find a key concept in a subject noun phrase than in general NPs. For this, we have split compound sentences in chunks, separating subordinate clauses ([Sub] type chunks). Then we have extracted sequences with the pattern [N][Mod]*. We assume that the sentence subject is a sequence [N][Mod]* occurring immediately before the verb. For instance:

El presidente [N] en funciones [Mod] de Haiti [Mod] ha afirmado [V] que [Sub]...

The rest of [N] and [Mod] chunks are considered as part of the sentence verb phrase. In a majority of cases, these assumptions lead to a correct identification of the sentence subject. We do not capture, however, subjects of subordinate sentences or subjects appearing after the verb.

Figure 2 shows how the probability of finding a key concept is always larger in sentence subjects. This result supports the assumption in (Boguraev et al., 1998), where noun phrases receive a higher weight, as representative terms, if they are syntactic subjects.

4.3 Distribution of key concepts within noun phrases

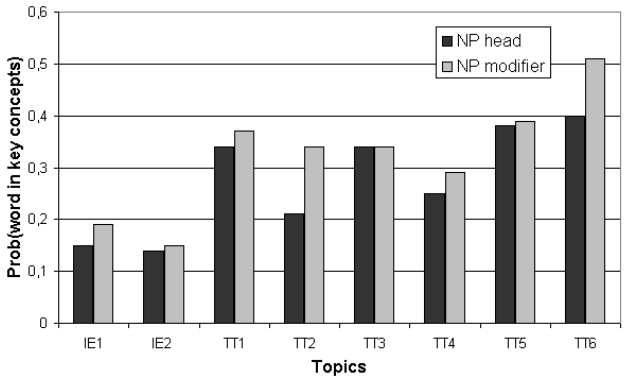

Figure 3: Probability of finding key concepts in NP head versus NP modifiers

For this analysis, we assume that, in `[N][Mod]*` sequences identified as subjects, `[N]` is the head and `[Mod]*` are the modifiers.

Figure 3 shows that the probability of finding a key concept in the NP modifiers is always higher than in the head (except for topic TT3, where it is equal). This is not intuitive a priori; an examination of the data reveals that the most characteristic concepts for a topic tend to be in the complements: for instance, in "the president of Haiti", "Haiti" carries more domain information than "president". This seems to be the most common case in our news collection. Of course, it cannot be guaranteed that these results will hold in other domains.

5 Automatic Selection of Key Terms

We have shown that there is indeed a correlation between syntactic information and the possibility of finding a key concept. Now, we want to explore whether this syntactic information can effectively be used for the automatic extraction of key concepts.

The problem of extracting key concepts for summarization involves two related issues: a) What kinds of terms should be considered as candidates? and b) What is the optimal weighting criteria for them?

There are several possible answers to the first question. Previous work includes using noun phrases (Boguraev et al., 1998; Jones et al., 2002), words (Buyukkokten et al., 1999), n-grams (Leuski et al., 2003; Lin and Hovy, 1997) or proper nouns, multi-word terms and abbreviations (Neff and Cooper, 1999).

Here we will focus, however, in finding appropriate weighting schemes on the set of candidate terms. The most common approach in interactive single-document summarization is using *tf.idf* measures (Jones et al., 2002; Buyukkokten et al., 1999; Neff and Cooper, 1999), which favour terms which are frequent in a document and infrequent across the collection. In the iNeast system (Leuski et al., 2003), the identification of relevant terms is oriented towards multi-document summarization, and they use a likelihood ratio (Dunning, 1993) which favours terms which are representative of the set of documents as opposed to the full collection.

Other sources of information that have been used as complementary measures consider, for instance, the number of references of a concept (Boguraev et al., 1998), its localization (Jones et al., 2002) or the distribution of the term along the document (Buyukkokten et al., 1999; Boguraev et al., 1998).

5.1 Experimental setup

A technical difficulty is that the key concepts introduced by the users are intellectual elaborations, which result in complex expressions which might even not be present (literally) in the documents. Hence, we will concentrate on extracting lists of terms, checking whether these terms are part of some key concept. We will assume that, once key terms are found, it is possible to generate full nominal expressions using, for instance, *phrase browsing* strategies (Peñas et al., 2002).

We will then compare different weighting criteria to select key terms, using two evaluation measures: a *recall* measure saying how well manually selected key concepts are covered by the automatically generated term list; and a *noise* measure counting the number of terms which do not belong to any key concept. An optimal list will reach maximum recall with a minimum of noise. Formally:

$$R = \frac{|\mathcal{C}_l|}{|\mathcal{C}|} \qquad Noise = |\mathcal{L}_n|$$

where $\mathcal{C}$ is the set of key concepts manually selected by users; $\mathcal{L}$ is a (ranked) list of terms generated by some weighting schema; $\mathcal{L}_n$ is the subset of terms in $\mathcal{L}$ which do not belong to any key concept; and $\mathcal{C}_l$ is the subset of key concepts which are represented by at least one term in the ranked list $\mathcal{L}$.

Here is a (fictitious) example of how R and $Noise$ are computed:

$\mathcal{C} = \{$Haiti, reinstatement of democracy, UN and USA troops$\}$
$\mathcal{L} = \{$Haiti, soldiers, UN, USA, attempt$\}$
$\rightarrow$ $\mathcal{C}_l = \{$**Haiti**, **UN** and **USA** troops$\}$ $\quad R = 2/3$
$\mathcal{L}_n = \{$soldiers,attempt$\}$ $\quad Noise = 2$

We will compare the following weighting strategies:

TF The frequency of a word in the set of documents is taken as a baseline measure.

Likelihood ratio This is taken from (Leuski et al., 2003) and used as a reference measure. We have implemented the procedure described in (Rayson and Garside, 2000) using unigrams only.

OKAPImod We have also considered a measure derived from Okapi and used in (Robertson et al., 1992). We have adapted the measure to consider the set of 100 documents as one single document.

TFSYNTAX Using our first experimental result, TFSYNTAX computes the weight of a term

as the number of times it appears preceding a verb.

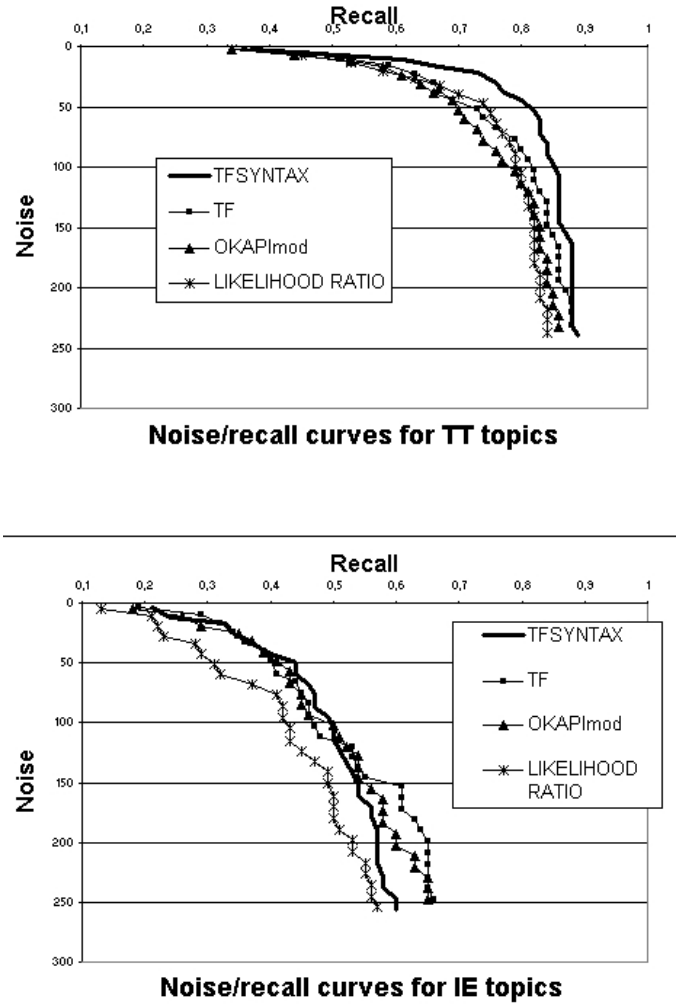

Figure 4: Comparison of weighting schemes to extract relevant terms

5.2 Results

Figure 4 draws Recall/Noise curves for all weighting criteria. They all give similar results except our TFSYNTAX measure, which performs better than the others for TT topics. Note that the TFSYNTAX measure only considers 10% of the vocabulary, which are the words immediately preceding verbs in the texts.

In order to check whether this result is consistent across topics (and not only the effect on an average) we have compared recall for term lists of size 50 for individual topics. We have selected 50 as a number which is large enough to reach a good coverage and permit additional filtering in an interactive summarization process, such as the iNeast terminological clustering described in (Leuski et al., 2003).

Figure 5 shows these results by topic. TFSYNTAX performs consistently better for all topics except one of the IE topics, where the maximum likelihood measure is slightly better.

Apart from the fact that TFSYNTAX performs better than all other methods, it is worth noticing that sophisticated weighting mechanisms, such as Okapi and the likelihood ratio, do not behave better than a simple frequency count (TF).

6 Conclusions

The automatic extraction of relevant concepts for a set of related documents is a part of many models of automatic or interactive summarization. In this paper, we have analyzed the distribution of relevant concepts across different syntactic functions, and we have measured the usefulness of detecting key terms to extract relevant concepts.

Our results suggest that the distribution of key concepts in sentences is not uniform, having a maximum in positions immediately preceding the sentence main verb, in noun phrases acting as subjects and, more specifically, in the complements (rather than the head) of noun phrases acting as subjects. This evidence has been collected using a Spanish news collection, and should be corroborated outside the news domain and also adapted to be used for non SVO languages.

We have also obtained empirical evidence that statistical weights to select key terms can be improved if we restrict candidate words to those which precede the verb in some sentence. The combination of statistical measures and syntactic criteria overcomes pure statistical weights, at least for TT topics, where there is certain consistency in the key concepts across documents.

Acknowledgments

This research has been partially supported by a research grant of the Spanish Government (project Hermes) and a research grant from UNED. We are indebted to J. Cigarrán who calculated the Okapi weights used in this work.

References

E. Amigo, J. Gonzalo, V. Peinado, A. Peñas, and F. Verdejo. 2004. Information synthesis: an empirical study. In *Proceedings of the 42th Annual Meeting of the ACL*, Barcelona, July.

B. Boguraev, C. Kennedy, R. Bellamy, S. Brawer, Y. Wong, and J. Swartz. 1998. Dynamic Presentation of Document Content for Rapid On-line Skimming. In *Proceedings of the AAAI Spring*

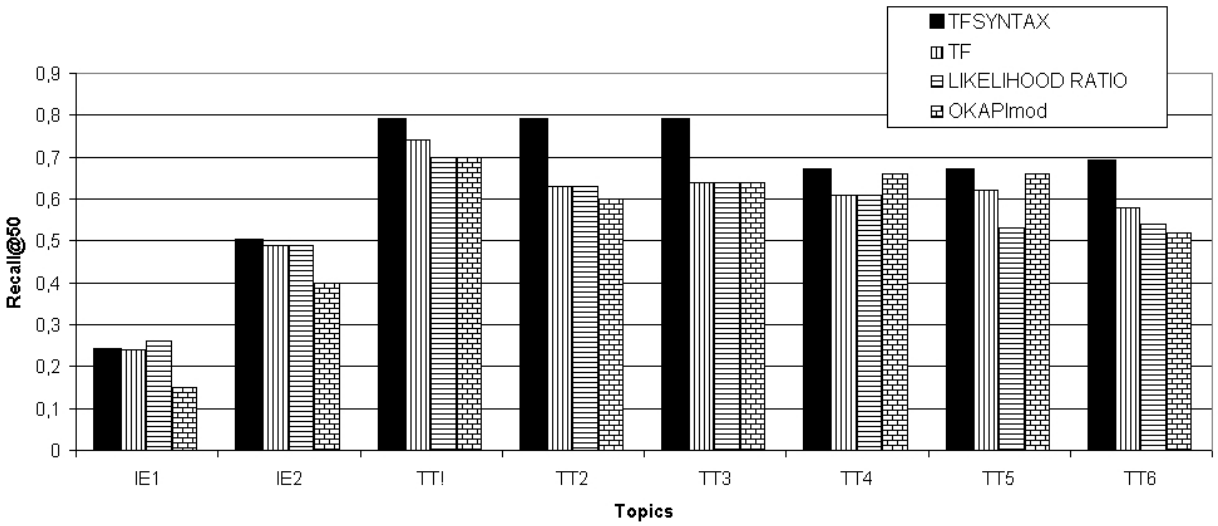

Figure 5: Comparison of weighting schemes by topic

1998 Symposium on Intelligent Text Summarization, Stanford, CA.

O. Buyukkokten, H. García-Molina, and A. Paepcke. 1999. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In *Proceedings of 10th International WWW Conference*.

T. Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

H. P. Edmundson. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

S. Jones, S. Lundy, and G. W. Paynter. 2002. Interactive Document Summarization Using Automatically Extracted Keyphrases. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, Big Island, Hawaii.

W. Kraaij, M. Spitters, and A. Hulth. 2002. Headline Extraction based on a Combination of Uni- and Multi-Document Summarization Techniques. In *Proceedings of the DUC 2002 Workshop on Multi-Document Summarization Evaluation*, Philadelphia, PA, July.

J. Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *Proceedings of SIGIR'95*.

A. Leuski, C. Y. Lin, and S. Stubblebine. 2003. iNEATS: Interactive Multidocument Summarization. In *Proceedings of the 4lst Annual Meeting of the ACL (ACL 2003)*, Sapporo, Japan.

C.-Y. Lin and E.H. Hovy. 1997. Identifying Topics by Position. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*, Washington, DC.

C. Lin and E. Hovy. 2002. NeATS in DUC 2002. In *Proceedings of the DUC 2002 Workshop on Multi-Document Summarization Evaluation*, Philadelphia, PA, July.

M. S. Neff and J. W. Cooper. 1999. ASHRAM: Active Summarization and Markup. In *Proceedings of HICSS-32: Understanding Digital Documents*.

A. Peñas, F. Verdejo, and J. Gonzalo. 2002. Terminology Retrieval: Towards a Synergy between Thesaurus and Free Text Searching. In *IBERAMIA 2002*, pages 684–693, Sevilla, Spain.

C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors. 2002. *Evaluation of Cross-Language Information Retrieval Systems*, volume 2406 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin-Heidelberg-New York.

S. Preston, K.and Williams. 1994. Managing the Information Overload. *Physics in Business*, June.

P. Rayson and R. Garside. 2000. Comparing Corpora Using Frequency Profiling. In *Proceedings of the workshop on Comparing Corpora*, pages 1–6, Honk Kong.

S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. 1992. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30.

J. D. Schlesinger, M. E. Okurowski, J. M. Conroy, D. P. O'Leary, A. Taylor, J. Hobbs, and H. Wilson. 2002. Understanding Machine Performance in the Context of Human Performance for Multi-Document Summarization. In *Proceedings of the DUC 2002 Workshop on Multi-Document Summarization Evaluation*, Philadelphia, PA, July.