Summarizing Texts at Various Levels of Detail

Marie-Francine Moens, Roxana Angheluta, Rik De Busser & Patrick Jeuniaux
Interdisciplinary Centre for Law & Information Technology
Katholieke Universiteit Leuven
Tiensestraat 41, B-3000 Leuven, Belgium
{marie-france.moens, roxana.angheluta, rik.debusser, patrick.jeuniaux}@law.kuleuven.ac.be

Abstract

Summarizing document texts at various levels of detail is required for many information selection tasks. For instance, when loading and visualizing documents on small screens of handheld devices, it is important to be able to dynamically compress texts. In this article we discuss a technique of generating hierarchical topic trees of a text and to use them in various ways to build summaries of a flexible length. For the topic tree building process we have implemented both a deterministic and probabilistic approach. We compare the results when the topic tree is used for automatic summarization.

Introduction

The technique of hierarchical topic segmentation allows condensing the information content of a text and building summaries of flexible length of it. It is particularly useful in situations where it is crucial that information is displayed as economically as possible. For instance, small screens (e.g., of mobile phones or PDAs) only allow a very limited amount of information to be displayed at the same time and it is therefore useful to generate a more dense representation of a long document, which allows the reader to easily locate relevant passages. Similarly, when skimming large amounts of texts on regular screens, text summarization makes it possible to cram more information in a limited visual space.

This article reviews texts summarization techniques that have been developed in the context of the project *Generic technology for information extraction from texts.* Our approach hierarchically segments a single text based on linguistic theories on the distribution of topic and comment in sentences and on patterns of thematic progression in text. The topic and the comment of a sentence are modeled both deterministically and probabilistically, the latter referring to the training of a probabilistic classifier. The segmentation tool is integrated in our summarization tool.

The article is organized as follows. The first section discusses the linguistic and cognitive background of our approach and sets the research goals. The next section discusses the methods and the algorithms. A third section describes our experiments, the results and their discussion. Before the conclusion, related research is described.

Background and Goals of Our Research

A text discusses a number of main topics. In the frame of these main topics a number of subtopics and more marginal topics are described. The writer of a text uses the following patterns of thematic progression: a hierarchy of topics, a sequence of topics and a semantic return, the latter referring to an abandoned topic that is picked up again in the discourse. The writer of a text is concerned about the correct interpretation of his or her text by the reader and uses explicit topicality cues to achieve this goal (van Dijk, 1988, p. 32 ff.). Apart from the ordering of thematic content through schemata that are typical for certain text genres (e.g., the lead of news stories signals the main topic) and the use of cue words or phrases (e.g., "speaking of") that introduce a topic or are an indication of its salience, the main surface cues for topicality are content words and their frequency and position. Among the content terms, noun phrases and their coreferents are the most important ones (Hahn, 1990). Two entities are considered as coreferents when they both refer to the same entity in the situation described by the text.

The position of a term within a sentence is another significant cue for judging topicality. A sentence is composed of a *sentence topic* and of additions to the topic (e.g., properties of the topic, relationships with other items, modifications of the topic) (Halliday, 1976; Hajièová & Sgall, 1988; Tomlin, Forrest, Pu & Kim, 1997). We find in the linguistic literature a number of universal cues for detecting the main topic of a sentence, the most important being initial sentence position and persistency of a topic across sentences (Givón, 1983, 1988, 2001; Gundel, 1988; Meinunger, 2000, p. 90). Noun-phrases are stressed when they occur in a sentence initial position. A main sentence topic usually occurs as the main topic in consecutive sentences and if the candidate topic or its referent occurs multiple times in a single sentence, this is an extra indication of its salience. Persistency can be expressed with the help of pronouns.

The patterns of thematic progression can be exploited to determine the salience of a sentence in a text. The goals of our research is to test an algorithm for hierarchical topic segmentation and for building a table of content of a text that gives an objective estimation of the salience of the text's topic and subtopics. An important focus of our research is the detection of the main topic of a sentence, for which we have developed both a deterministic and probabilistic approach and to test both approaches when building single document summaries. We perform a number of experiments from which we try to know whether the probabilistic approach outperforms the deterministic approach when the hierarchical topic tree is used for summarization, what properties of the topic tree are best exploited when building summaries, and whether the correct resolution of pronouns in the texts improves the summaries obtained with this technique.

Methods

Algorithm

Our algorithm (Figure 1) uses generic topical cues for detecting the topic structure of a document text. In a pre-processing step the input text is tokenized, each token is tagged with its syntactic word class and sentences are chunked into simple noun phrases. The most representative terms of the chunks are identified. The topic segmentation algorithm relies on three processes. A first process detects the term distributions in a text. A second process concerns the detection of sentence topics. A third optional process detects terms in a text that corefer to the same entity. Based on the knowledge that is generated in these three processes, the text is segmented and a table of content is built. This table of content is used to build summaries of flexible length. We give a short overview of the topic segmentation algorithm (see also Moens & Dumortier (2003) where the algorithm including main sentence topic detection is extensively tested and evaluated) and focus upon its use in text summarization.

Input

The system that we implemented can process texts of various genres (Moens, Angheluta & Dumortier, 2004), but has been designed for relatively short texts up to a few pages of length. We assume that longer texts can be broken into hierarchical segments based on their heading structure (Yang and Wang, 2003).

Preprocessing: tokenization and morphosyntactic tagging

We break the text into sentences and words and tag each word with its word class (part-of-speech). For English texts we have used the LTCHUNK software, developed by Andrei Mikheev at the University of Edinburgh. LTCHUNK consists of a noun and verb phrase chunker built on top of a part-of-speech tagger (LTPOS) that uses a Hidden Markov Model disambiguation strategy (Mikheev, 1998). It assigns part-of-speech tags to individual tokens and orders them into noun phrase chunks. A noun phrase chunk (or shortly noun chunk) is a noun phrase that is not modified by another prepositional noun phrase.

The tagger allows identifying proper names. We also identify the head noun of each noun chunk by using a language-dependent heuristic and store a description of the noun chunk. The latter is composed of the head, the head and its proper name modifiers or the head with other modifiers in case these form a collocational phrase. Collocations are terms of which the constitutive parts typically co-occur more often than by chance and are characterized by limited compositionality, i.e., there is usually an element of meaning added to the collocation that cannot be predicted from the meanings of the composing parts (e.g., "joint venture"). A collocational phrase is identified with a certain significance level when the independence of the composing words of the collocational phrase can be rejected based on a training corpus. We used the likelihood ratio for a binomial distribution to reject or accept the hypothesis of independence (Dunning, 1993) and to detect collocations of bigrams or trigrams composed of nouns and composed of adjectives followed by nouns (for details on our implementation: see Moens & Angheluta, 2003).

Computation of term distribution

The *computation of term distribution* module is an assisting module which computes for each term its position in the text in terms of sentence number and position within the sentence.

Detection of the main sentence topic

The input of this module is the set of sentences that form the input text in reading order. Each sentence is broken into its noun chunks and for each noun chunk a number of features are extracted in accordance with the first four linguistic heuristics found in the list mentioned in section 2.

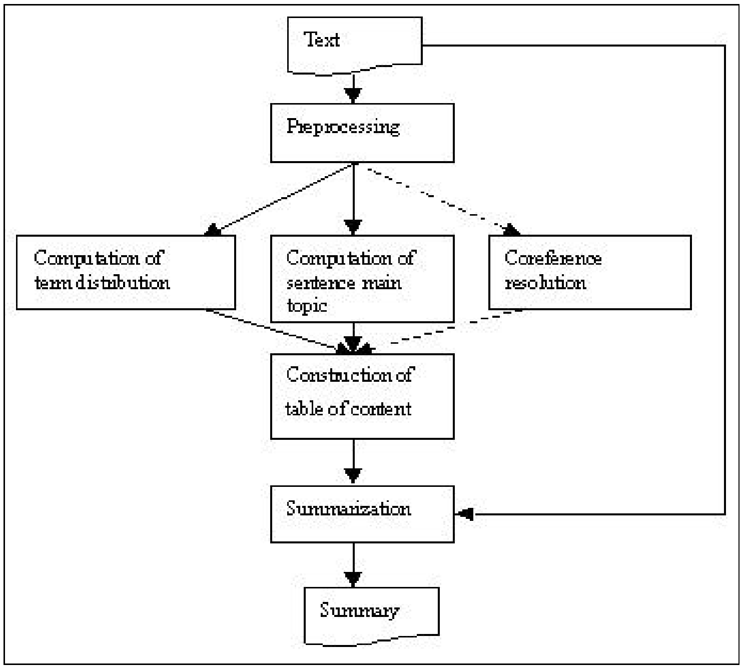
Text
Preprocessing
Computation of term distibution
Computation of sentence main topic
Coreference resolution
Construction of table of content
Summarization
Summary

Figure 1: Architecture of the single-document summarization

These features include:

- *First noun chunk* (*FNC*): The noun chunk is the first one in the sentence.
- *Noun chunk before verb phrase* (*NCBVP*): The noun chunk is located before the first verb phrase in the sentence.
- *Persistency* (*P*): The noun chunk has a member noun that also occurs in the previous sentence.
- *Persistent main topic* (*PMT*): The noun chunk has a member noun that is a member of the main topic of the previous sentence.
- *Multiple occurrences* (*MO*): The noun chunk has a member noun that occurs more than once in the current sentence.

These features are used in both our deterministic and probabilistic classifiers for detecting the main sentence topic. At this moment the topic analysis does not go down to the clausal level, which implies that coordinated and subordinated clauses are not properly treated.

A deterministic approach In the deterministic approach (referred to as DET in the tables) we have implemented an algorithm based on heuristic rules and their priorities that selects the noun chunk representing the main topic of a sentence.

Schematic representation of the algorithm:

```
FOR each sentence $s_x$ of the input text
    IF a noun chunk $nc$ in sentence $s_x$ preceding the first verb has a member that is persistent
    ($nc.NCBVP = true$ and $nc.P = true$)
    OR
    IF a noun chunk $nc$ in sentence $s_x$ has a member that is persistent ($nc.P = true$)
    OR
    IF a noun chunk $nc$ in sentence $s_x$ has a member that occurs multiple times in $s_x$ ($nc.MO = true$)
        THEN select noun chunk $nc$
     ELSE select noun chunk $nc$ in initial position in sentence $s_x$ ($nc.FNC = true$)
```

In case of ties (e.g., two terms occur in the previous sentence), the algorithm gives higher priority to a noun chunk whose noun member is part of the main topic of the previous sentence (*nc.MT = true*) or to the first occurrence of the chunk in the current sentence. Sentences containing zero noun chunks do not have a main topic.

A probabilistic approach Although the discourse cues modeled in the deterministic algorithm are evidenced in the linguistic literature as being relevant for sentence topic detection, the implementation of the priority in which these cues apply is rather intuitive and the dependency of features that occur together is modeled quite simplistically. It seemed useful to us to model these phenomena probabilistically by using a training set of classified examples, in which the main topic of a sentence is assigned by humans. In previous experiments we tested different classifiers (e.g., maximum entropy modeling) and showed that a classifier that estimates the probability distributions of the various parameters involved in main sentence topic assignment from the training set and that combines the strengths of the various distributions to obtain an estimate of the overall distribution, while using the expectation maximization (EM) algorithm, obtained the best results (Moens & Dumortier, 2003). The estimate of the full distribution is used to find the most probable main topic of a sentence in a test example.

To automatically classify a noun chunk *nc* into the classes main topic or not main topic (*C)*, we wish to estimate a probability distribution indicating how likely the noun chunk is to be assigned to each possible class given its features:

$$P(C|FNC, NCBVP, P, PMT, MO)$$

It would be possible to calculate this distribution directly from the training data by counting the number of times a term as a main topic appears with this combination of features and dividing by the total number of times the combination of features appears. In many cases, we might never see a particular combination of features in the training data, and in other cases, we might see the combination only very rarely, providing a poor estimate for the probability.

We expect that the features interact in various ways and we cannot train directly on the full feature set. For this reason, we have built a classifier by combining probabilities from distributions conditioned on a variety of subsets of the features (e.g., $P(C|FNC,P)$, $P(\mathrm{C}|MT)$). We compute the relative occurrence of noun chunks with each possible combination of features, and choose the subset of features for which the occurrence with noun chunks is more than 1% in the training set, and compute the probability distribution conditioned on each chosen feature set from the training set. For our training set (see section 4), this resulted in the computation of ten distributions. To compute the general distribution we have to combine the strengths of various distributions.

$$
\begin{aligned}
&P(C|nc) \\
&= \lambda_1 P(C|FNC) + \lambda_2 P(C|NCBVP) + \lambda_3 P(C|P) + \lambda_4 P(C|PMT) + \lambda_5 P(C|MO) + \lambda_6 P(C|FNC,NCBVP) + \lambda_7 P(C|FNC,P) + \lambda_8 P(C|NCBVP,P) + \lambda_9 P(C|P,PMT) + \lambda_{10} P(C|FNC,NCBVP,P)
\end{aligned}
$$

$$\text{where } \sum_i \lambda_\mathrm{i} = 1$$

The geometric mean, when expressed in the log domain, is similar:

$$
\begin{aligned}
&P(C|nc) \\
&= \frac{1}{Z}\exp\{\lambda_1 \, log\ P\ (C|\ FNC) + \lambda_2 \, log\ P\ (C|NCBVP) + \lambda_3 \, log\ P\ (C|P) + \lambda_4 \, log\ P\ (C|PMT) + \\
&\lambda_5 \, log\ P\ (C|MO) + \lambda_6 \, log\ P\ (C|FNC,NCBVP) + \lambda_7 \, log\ P\ (C|FNC,\ P) + \lambda_8 \, log\ P\ (C|NCBVP,P) \\
&+ \lambda_9 \, log\ P\ (C|P,PMT) + \lambda_{10} \, log\ P\ (C|FNC,NCBVP,P)\ \} \qquad (4)
\end{aligned}
$$

where Z is a normalizing constant ensuring that $\sum_C P(C|nc) = 1$

We estimate the interpolation weights λ_i with the Expectation Maximization (EM) algorithm (Dempster, Laird & Rubin, 1977).

Although the probabilistic model is trained on example texts, it has been proven that the learned probability distribution could be successfully ported to different types of texts written in the same language (Moens & Dumortier, 2003).

Coreference resolution

Topics in a text might be referred to by other terms such as pronouns, synonyms or hypernyms. Ideally, our algorithm should include a tool for noun phrase coreference resolution, which detects when two noun phrase entities (noun chunks, proper noun phrases and pronouns) refer to the same entity in the situation described in the text. Coreference resolution also includes the resolution of entities that act as anaphora and cataphora in the text. An anaphoric expression is a textual element whose interpretation depends on the meaning of another textual element with a more descriptive phrasing and found earlier in the text. An anaphor and its noun phrase of reference are said to be coreferents; a cataphoric expression refers to an element further in the text. Sidner (1983) has demonstrated that anaphors (especially pronouns) are important in detecting the topical item of a text passage. We have developed a tool for coreference resolution based on a fuzzy clustering of the entities found in the text and relying on a set of linguistic features (Mitra, Angheluta, Jeuniaux & Moens, 2003). In the experiments described below we have not yet integrated this tool, since we are currently in the process of evaluating and refining it. However, we perform tests in order to measure the effect of coreference resolution and more specifically of the resolution of pronouns upon the performance of our summarization tool.

Construction of the table of content

The algorithm for calculating the topic hierarchy consists of two main steps.

A first step consists of the detection of topically coherent segments. For detecting coherent segments, the head noun of the noun chunk that is designated as main topic of the sentence (further called segment topic term) is selected. A coherent segment is formed by consecutive sentences that have the same segment topic term or by one isolated sentence, if none of the neighboring sentences has the same segment topic term. A coherent segment in this framework is different from a segment spanned by a lexical chain (Barzilay & Elhadad, 1999). We only take into account the terms that designate the main topic of a sentence, where lexical chains might consider all content terms. Lexical chains also consider related terms (e.g., synonyms, hypernyms), which we do not, except when we would consider noun phrase coreference resolution.

In the second step, the topical relationships between segments are computed (i.e., hierarchical and sequential relationships and semantic returns); each segment is described by key terms; and a table of content is built. To detect the relationships between topical segments the term distributions and a number of heuristics are taken into account. For instance, a current topical segment is a subtopic of another topical segment when the terms of the other topical segment occur in the current segment or in the nearby context of the current segment. In this way the level of topicality of each segment can be computed. The shape of the table of content depends on how many different main topics of a sentence the text has and the number of semantic returns by which a main sentence topic is reintroduced in the discourse after other topics have been discussed.

Each topical segment in the hierarchy is described by its textual boundaries by means of character pointer positions and by a set of one or more key terms. The key terms are composed of the segment topic terms and are sometimes augmented by comment terms. Terms from comments are only added to the segment term when the comment terms are the cause of a subtopic.

Summary construction

By computing the topic hierarchy, we have constructed a table of content of the text, which has a tree-like structure and which indicates the topics and more detailed subtopics in the text (see Figure 2a). For each topic, the text segment that covers the topic is represented by its boundaries, i.e., begin and end positions in the text. Instead of describing each segment with a few topic terms, we extract the first sentence of each segment to form the summary. The topic tree allows us to build summaries at various levels of detail. The segments up to a certain level of topicality can be chosen to form the basis of the summary (e.g., Figure 2b and 2c). Or alternatively, the segments with largest length (i.e., that cover the largest segments as indicated by the character pointers in the text) can be selected (e.g., Figure 3). From each chosen segment a representative sentence is extracted. We choose in the experiments described here the first sentence of a selected segment, but alternative ways of selecting content from the segment might be explored. If the summary has to fit a certain length requirement, the topic tree can be processed in a breadth-first way or the largest segments can be chosen up until the extracted sentences sum to the required number of words. The extracted sentences are then shown

```
a) Gilbert power hurricane 0    4063
       injuries hurricane      172     1591
               hour trees      311     1591
                       Service Kingston        445     622
                               shock   553     622
                       Skies hurricane 623     1591
                               flights 740     811
                               Flights Miami   812     1101
                                       People  961     1101
                               report winds    1102    1439
                                       warnings rain   1222    1439
                                               Jamaica 1322    1439
                               interests       1440    1591
       warnings Cayman 1687    2013
               winds   1863    2013
       people  2014    2117
       coast   2118    2549
               Radio   2343    2549
       Washington hurricane    2550    3047
               Ross    2747    3047
                       reports 2996    3047
       spokesman       3048    3256
       storm   3257    3353
       Sheets Gilbert  3354    3794
               Residents       3523    3647
               Service 3648    3794
       interests       3795    3927
       Gerrish 3928    4063


b) Hurricane Gilbert slammed into Kingston on Monday with torrential rains
and 115 mph winds that ripped roofs off homes and buildings, uprooted trees
and downed power lines.
No serious injuries were immediately reported in the city of 750,000
people, which was hit by the full force of the hurricane around noon.
Hurricane warnings were issued Monday for the south coast of Cuba east of
Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued
for the Dominican Republic.
Most of Jamaica's 2.3 million people stayed home, boarding up windows in
preparation for the hurricane.

c) Hurricane Gilbert slammed into Kingston on Monday with torrential rains
and 115 mph winds that ripped roofs off homes and buildings, uprooted trees
and downed power lines.
No serious injuries were immediately reported in the city of 750,000
people, which was hit by the full force of the hurricane around noon.
Hurricane warnings were issued Monday for the south coast of Cuba east of
Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued
for the Dominican Republic.
Most of Jamaica's 2.3 million people stayed home, boarding up windows in
preparation for the hurricane.
The popular north coast resort area, on the other side of the mountains,
was expected to receive heavy rain but not as much damage from the
hurricane as the south coast, where officials urged residents to seek
higher ground.
In Washington, the Navy reported its bases at Guantanamo Bay, Cuba, and
Roosevelt Roads, Puerto Rico, had taken various precautionary steps but
appeared to be safe from the brunt of the hurricane.
```

Figure 2: a) Hierarchical topic tree of document AP880912-0137.S set d061j (DUC-2002 corpus) made with the probabilistic approach to main sentence topic detection, the summary sentences are selected by considering the topicality level as a criterion of topical salience of a segment; b) 100-word summary; c) 150-word summary.

```
a) Hurricane Gilbert slammed into Kingston on Monday with torrential rains
and 115 mph winds that ripped roofs off homes and buildings, uprooted trees
and downed power lines.
No serious injuries were immediately reported in the city of 750,000
people, which was hit by the full force of the hurricane around noon.
For half an hour, the hurricane lashed the city, tearing branches from
trees, blowing down fences and whipping paper through the air.
Skies brightened, the winds died down and people waited for an hour before
the second blow of the hurricane arrived.

b) Hurricane Gilbert slammed into Kingston on Monday with torrential rains
and 115 mph winds that ripped roofs off homes and buildings, uprooted trees
and downed power lines.
No serious injuries were immediately reported in the city of 750,000
people, which was hit by the full force of the hurricane around noon.
For half an hour, the hurricane lashed the city, tearing branches from
trees, blowing down fences and whipping paper through the air.
Skies brightened, the winds died down and people waited for an hour before
the second blow of the hurricane arrived.
In Washington, the Navy reported its bases at Guantanamo Bay, Cuba, and
Roosevelt Roads, Puerto Rico, had taken various precautionary steps but
appeared to be safe from the brunt of the hurricane.
Sheets said Gilbert was expected next to sweep over the Cayman Islands, on
its westward track, and in two to three days veer northwest into the
southern Gulf of Mexico.
```

Figure 3: Document AP880912-0137.S set d061j (DUC-2002 corpus) summaries are made with the probabilistic approach to main sentence topic detection; the summary sentences are selected by considering the segment length as a criterion of topical salience of a segment a) 100-word summary; b) 150-word summary.

in reading order. In our experiments described below, we do not further reduce the content of the extracted sentences (cf. Dorr, Zajic & Schwartz, 2003), but this is an approach that could be considered if a strong reduction of the content is required.

In this paper and in our experiments below, we focus on generic summaries. A *generic or topic-general summary* is a summary that truly reflects the main content of the original text. A *viewpoint-oriented summary* summarizes text according to a certain viewpoint, which might express the information need of the user in a retrieval system. The topic tree allows selecting certain topical viewpoints and computing their salience, which is useful for building view-point oriented summaries at certain levels of topical detail (Moens, Angheluta & Dumortier, 2004).

Experiments, Results and Discussion

An extensive evaluation of the use of the deterministic topic segmentation algorithm in different summarization tasks (e.g., topic-general, viewpoint-oriented, single and multi-document summarization) can be found in Moens, Angheluta and Dumortier (2004) and clearly shows the validity of this technique. In this cited article we also discuss the integration of the topic segmentation with other summarization techniques.

In the research that we discuss in this paper the topic trees are generated with a deterministic and probabilistic approach. We also take into account different ways of choosing important segments from a text and try to estimate the effect of a correct coreference resolution.

We randomly selected 25 news story texts and their 100-words man-made model summaries from the

Document Understanding Conference 2002 (DUC-2002) corpus. Our SUMMA system built 100-words summaries from the texts, based on both the deterministic and probabilistic approach to main sentence topic detection and in a separate experiment the system built summaries with the probabilistic approach from texts in which the pronouns were manually resolved. For the probabilistic approach, we trained on 60 texts randomly selected from the DUC-2003 corpus, which were annotated with their main sentence topic by a researcher with a linguistic background. The task was to mark the noun chunk that most reflected the aboutness of a sentence in the context of a text. The collocational phrases were trained per set to which a document belongs. A set contains about 10 related documents.

For comparing two summaries (e.g., a model summary with a system-made summary) we use evaluation measures used in DUC-2002 (Over & Ligett, 2002). Coverage is computed as the completeness judgments in terms of 0%, 20%, 40%, 60%, 80% and 100% for a system-created summary as compared to the model summary. The tool of Lin and Hovy (2002) breaks the system-made summary into units and allows computing the coverage as the average completeness of each unit compared with selected sentences of the model summary. The aim is to maximize coverage while minimizing the length of the summary. *Length-adjusted coverage* is defined as the weighted sum of coverage and brevity, where coverage is twice as important as brevity and brevity is zero when the summary exceeds the target length. The *length-adjusted coverage with penalty* corrects the coverage with the factor “target length /actual length” which additionally punishes when the summary is over the target length.

The results of a summarization that uses the level of topicality in the table of content as a criterion to select important segments and their representative sentences are shown in Table 1. The results are given in terms of mean coverage (MC) and mean length-adjusted coverage (MLAC) computed over the 25 texts. We compare the model summaries with the system-made summaries built with the deterministic approach (DET-our-evaluation). Because we participated in DUC-2002 with our system, we show also the evaluation done by the evaluation team of the Document Understanding Conference evaluation (Angheluta, De Busser & Moens, 2002). For this participation we have only implemented our deterministic approach to main sentence topic detection. We also compare the model summaries with the system-made summaries built with the probabilistic approach (PROB-EM-our-evaluation). In DUC-2002, two summaries were made for each text by two different humans, one of which was randomly chosen as model summary. We also show the comparison of the two DUC-2002 human-made summaries (MAN-DUC-evaluation), a value which indicates inter-human agreement.

We acknowledge that the constructing of human-made summaries and their comparison with system-made summaries is rather subjective. In addition, results of summarizing 25 texts can not be very conclusive. However, we have indications of the following findings. From these results we cannot conclude that the summarization based on a probabilistic approach of main sentence topic detection performs better than a deterministic approach. The hierarchical topic segmentation that is based on a probabilistic modeling of the main sentence topic provides about equal summaries in terms of mean coverage and mean length-adjusted coverage if we take the average of the two evaluations of the deterministic approach done by two different humans. If the summaries are made from texts in which the pronouns (except for genitive cases) are manually resolved, the mean coverage of the summaries could not be improved (PROB-EM-our-evaluation resulted in 0.403 and 0.269 for MC and MLAC respectively). This is somehow logical because less than 30% of the topic trees changed as a result of coreference resolution, which resulted in summaries which were only different by maximum one sentence with the summary of the same text without noun phrase coreference resolution.

Because we do not have ideal single-document summaries of other lengths, we did not perform an evaluation. Figure 2 (b and c) shows the summaries of 100 and 150 words. However, in a small experiment we made summaries of 150 words instead of summaries of 100 words and could improve the mean coverage and mean length-adjusted coverage to 0.559 and 0.371 respectively. If we can now reliably reduce the sentences by 33 %, we achieve coverage that is comparable to and even better than the one of summaries that are manually made. So, it seems to us very worthwhile to develop technologies that condense sentences to content that is salient with regard to the complete discourse.

	MC	MLAC
DET-our-evaluation	0.416	0.284
DET-DUC-evaluation	0.386	0.264
PROB-EM-our-evaluation	0.402	0.279
MAN-DUC-evaluation	0.464	0.309

Table 1: Mean coverage (MC) and mean length-adjusted coverage (MLAC) for the summarization of the 25 texts taking into account the level of topicality as a criterion of topical salience of a segment.

	MC	MLAC
PROB-EM-our-evaluation	0.378	0.258

Table 2: Mean coverage (MC) and mean length-adjusted coverage (MLAC) for the summarization of the 25 texts taking into account the length of a topical segment as a criterion of topical salience of a segment.

We also made summaries by using the length of a segment as a criterion for selecting the most important segments and their representative sentences (cf. Figure 3). The results are shown in Table 2. By taking into account the length of a segment, the mean coverage and mean length adjusted coverage could not be improved.

Related Research

Topic segmentation is usually an important step when summarizing texts (e.g., Barzilay & Elhadad, 1999; Mani & Bloedorn, 1999; Chali, Kolla, Singh & Zhang, 2003). Linear segmentation algorithms are often used for this task. From the segments important terms and possibly their related terms are identified and sentences that contain these terms are extracted to form the summary. Famous linear topic segmentation algorithms have been developed for instance by Hearst (1997), Ponte and Croft (1997), Kan, Klavans and McKeown (1998), and Choy (2000). We use hierarchical topic segmentation in a quite similar framework (except that we do not detect related terms and use them to select sentences, but select sentences as a first sentence of the segment). The advantage of using a hierarchical topic segmentation is that content can be chosen at different levels of detail. The research into hierarchical topic segmentation is quite limited. Yaari (2000) presents a hierarchical clustering of sentences. This approach requires that a subtopic sentence contains the terms or their coreferents of its more general topics, which is not always the case. Interesting to mention is the computation of salience of a sentence based on a rhetorical structure analysis of a text (Marcu, 2000), an approach we might in the future incorporate in our technology.

Hierarchical topic segmentation is acknowledged to be very valuable for generic and view-point oriented text summarization, text searching and navigation (Zizi & Beaudouin-Lafon, 1995; Kan, McKeown & Klavans, 2001; Yang & Wang, 2003) and allows zooming in and out on the content of a text, which is very helpful when displaying texts on small screens or when skimming large amounts of texts on regular ones. Current approaches that zoom in and out on the content of documents exploit their layout structure, but ignore hierarchical, sequential and semantic return relationships between topical segments that can be found in unstructured natural language texts (e.g., Yang & Wang, 2003) or allow just selecting the text based on key terms or summary sentences that contain terms that are highly weighted because of their high frequency in this text and low frequency in a reference corpus (e.g., Buyukkokten, Garcia-Molina & Paepcke, 2001).

Conclusions

In this article we have shown that hierarchical topic segmentation is useful for text summarization. The segmentation relies on discourse patterns of topic-comment and on patterns of thematic progression. Learning the topics of a sentence based on training examples with a probabilistic classifier that estimates various probability distributions and combines these where interpolation weights that are chosen with the Expectation Maximization algorithm, gives about equal summarization performance than a deterministic approach. In the future we want to expand the proposed techniques with other valuable approaches for content recognition of texts such as coreference resolution, reliance on the rhetorical relations between text spans when they are signaled in the text, and condensing sentences at different levels of detail in order to enhance the text's compression.

Acknowledgments

The research is sponsored by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), Belgium. We also want to thank Donna Harman, Paul Over and their team (National Institute of Standards and Technology, USA) for the availability of single-document summaries in the frame of the Document Understanding Conference 2002.

Bibliographical References

Angheluta, R., De Busser, R. & Moens, M.-F. (2002). The use of topic segmentation for automatic summarization. In *Proceedings of the ACL-2002 Post-Conference Workshop on Automatic Summarization*.

Barzilay, R. & Elhadad, M. (1999). Using lexical chains for text summarization. In I. Mani & M.T. Maybury (Eds.), *Advances in Automatic Text Summarization* (pp. 111-121). Cambridge, MA: MIT Press.

Buyukkokten, O., Garcia-Molina, H. & Paepcke, A. (2001). Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Proceedings of the World Wide Web Conference 10, May 2-5 2001 Hong Kong* (pp. 652-662). New York: ACM.

Chali, I., Kolla, M., Singh, N. & Zhang, Z. (2003). The University of Lethbridge text summarizer at DUC-2003. In D. Radev & S. Teufel (Eds.), *Proceedings of the Text Summarization Workshop and 2003 Document Understanding Conference May 31 and June 1, 2003* (pp. 148-152). Gaithersburg, MD: NIST.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, series B, 39 (1), 1-38.

Dorr, B., Zajic D. & Schwartz, R. (2003). Hedge Trimmer: A parse-and-trim approach to headline generation. In R. Radev & S. Teufel (Eds.), *Proceedings of the HLT-NAACL 2003 Workshop on Text Summarization* (pp. 1-8). Omnipress.

Dunning, T (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, 61-74.

Givón, T. (1983). Introduction. In T. Givón (Ed.). *Topic Continuity in Discourse: A Quantitative Cross-Language Study* (pp. 1-41). Amsterdam: John Benjamins.

Givón, T. (1988). The pragmatics of word-order: predictability, importance and attention. In M. Hammond, E. Moravcsik & Jessica Wirth (Eds.), *Studies in Syntactic Typology* (pp. 243-284). Amsterdam: John Benjamins.

Givón, T. (2001). *Syntax: An Introduction.* Amsterdam: John Benjamin.

Gundel, J. (1988). Universals of topic-comment structure. In M. Hammond, E. Moravcsik & J. Wirth (Eds.), *Studies in Syntactic Typology* (pp. 209-239). Amsterdam: John Benjamins.

Hahn, U. (1990). Topic parsing: accounting for text macro structures in full-text analysis. *Information Processing & Management* , 26 (1), 135-170.

Hajièová, E. & Sgall, P. (1988). Topic and focus of a sentence and the patterning of a text. In J.S. Petöfi (Ed.), *Text and Discourse Constitution: Empirical Aspects, Theoretical Approaches* (pp. 70-96). Berlin: Walter de Gruyter.

Halliday, M.A.K. (1976). Theme and information in the English clause. In G.R. Kress & M.A.K. Halliday (Eds.), *Halliday: System and Function in Language* (pp. 174-188). London: Oxford University Press.

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

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

Kan, M.-Y., McKeown, K.R. & Klavans, J.L. (2001). Domain-specific informative and indicative summarization for information retrieval. In D. Harman & D. Marcu (Eds.), *Proceedings of DUC 2001 Workshop on Text Summarization* (http://www-nlpir.nist.gov/projects/duc/pubs.html/#2001).

Kononenko, I., Kononenko, S., Popov, I. & Zagorulko, Y. (2000). Information extraction from non-segmented text. In *RIAO'2000 Content-Based Multimedia Information Access Paris April 12-14 2000.*

Mani, I. & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. In I. Mani & M.T. Maybury (Eds.). *Advances in Automatic Text Summarization* (pp. 357-379). Cambridge, MA: The MIT Press.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA*:* The MIT Press.

Meinunger, A. (2000). *Syntactic Aspects of Topic and Comment.* Amsterdam: John Benjamins.

Mikheev, A. (1998). *Part-of-Speech Guessing Rules : Learning and Evaluation* (http://www.ltg.ed.ac.uk/software/pos).

Mitra, R. , Angheluta, R., Jeuniaux, P. & Moens, M.-F. (2003). Progressive fuzzy clustering for noun phrase coreference resolution. In *Proceedings of the Fourth Dutch-Belgian Information Retrieval Workshop DIR-2003.*

Moens, M.-F. & Angheluta, R. (2003). Concept extraction from legal cases: The use of a statistic of coincidence. In *Proceedings of the Eight International Conference on Artificial Intelligence and Law* (pp. 142-146). New York: ACM.

Moens, M.-F., Angheluta, R. & Dumortier, J. (2004). Generic technologies for single- and multi-document summarization. *Information Processing & Management* (in press).

Moens, M.-F. & Dumortier, J. (2003). Using patterns of thematic progression for building a table of content of a text. Technical Report.

Over P. & Ligett, W. (2002). Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization systems. Gaithersburg, MD: NIST.

Ponte, J.M. & Croft, B.W. (1997). Text segmentation by topic. In *Proceedings of the first European Conference on Research and Advanced Technology for Digital Libraries* (pp. 120-129).

Salton, G., Singhal, A., Buckley, C. & Mitra, M. (1996). Automatic text decomposition using text segments and text themes. *Hypertext'96*, 53-65.

Sidner, C.L. (1983). Focusing in the comprehension of definite anaphora. In M. Brady & R.C. Berwick (Eds.), *Computational Models of Discourse* (pp. 267-330). Cambridge, MA: The MIT Press.

Tomlin, R.S., Forrest, L., Pu, M.M. & Kim, M.H. (1997). Discourse semantics. In T.A. van Dijk (Ed.), *Discourse as Structure and Process* (*Discourse Studies: A Multidisciplinary Introduction* 1) (pp. 63-111). London: SAGE.

Van Dijk, T.A. (1988). *News as Discourse*. Hillsdale, NJ: Lawrence Erlbaum.

Yaari, Y. (2000). NLP-assisted exploration of texts. In *Proceedings RIAO'2000 Content-Based Multimedia Information Access Paris, April 12-14, 2000.* Paris: CID-CASIS.

Yang, C. & Wang, F.L. (2003). Fractal summarization for mobile devices to access large documents on the Web. In *Proceedings of the International World Wide Web Conference, May 20-24, 2003 Budapest, Hungary*. New York: ACM.

Zizi, M. & Beaudouin-Fafon, M. (1995). Hypermedia exploration with interactive dynamic maps. *International Journal Human-Computer Studies*, 43, 3, 441-464.