Topical and rhetorical term extraction for text description

Nicolas Hernandez and Brigitte Grau
LIMSI/CNRS - LIR Laboratory
Paris Sud University
BP 133, F-91403 ORSAY CEDEX (France)
Hernandez—Grau@limsi.fr

Abstract

We present a dynamic summarization method that combines text segmentation and topical and rhetorical text description to rapidly skim through texts. The speed reading is made thanks to presentation to the user of relevant expressions from the text, labelled with topical and rhetorical specification. In particular, we describe a method to automatically acquire terminological resources from a scientific corpus without specific knowledge resources requirement.

1. Introduction

Typically information retrieval (IR) systems return ranked lists of documents. Users have then to read some of them to know if they really fulfill their needs. Unfortunately, electronic document presentation often obeys to the same presentation as paper documents and leads to a linear browsing.

This remark led us to conceive an automatic text structuring allowing text visualization and intra-document navigation that fills the gap between user’s information need and IR result presentation. In order to rapidly offering a maximum of information, text visualization has to be concise while being both indicative and informative. Our approach is based on works of [7, 9, 1] in automatic summarization, especially for the aspect of the prior semantic annotation of text fragments to lead the summarization. But in our case, we want to emphasize the dynamic aspect of text exploration. The indicative view is given by the presentation of the main topics of the text, while the informative part is satisfied by presenting both the different segments that develop them (embellished with their location, their relevance in regards of the rest of the text) and the rhetorical role of segments in which they occur; in other words which topic is developed while another is in progress and which relation they entertain. In addition, the link between each entity is always maintained, allowing a user to navigate from a view to another.

Figure 1. System overview

Figure 1 depicts an overview of our system. Beside a text structuring process that segments texts into thematic units as in [4], our system describes these units by their salient terms, in particular topical and rhetorical ones. We associate each segment with a general topic it is about, and the local topic it particularly develops. Topic descriptors correspond to salient noun phrases as proposed by [1], except that we extend this notion to categorize them from a topic structure point of view. Moreover, we enhance segment characterization by its rhetorical role [3, 9]. For this purpose, we propose an automatic method non-knowledge based to automatically acquire terminological resources. This kind of approach works well with scientific and technical documents.

2 Text Topic Segmentation

Our segmentation method is based on numerical criteria that rely on the fact that topic elaboration leads to the repetition of specific terms, in particular in technical and scientific
texts. The recognition that several parts of text are linked to the same topic relies on the distribution and the recurrence of words. When a word is often repeated throughout a text, it is not relevant for our purpose, whereas its reiteration in a limited zone is highly significant to characterize the topic segment. The general principle of the method takes up a principle applied by several existing systems and aims at pointing topic break [4]. It consists of representing a text fragment by a vector of lemmatized words. A weight is associated depending on their frequency in the text fragment and all over the text. The weight is computed using the tf.idf scheme. The scalar product of these vectors permits to merge or to separate the adjacent text fragments they describe according to whether it is below a threshold or not. In our case, we choose the paragraph as the minimal topic unit of text [6].

3 Description of text and text segments

In a dynamic summarization perspective, our system aims at giving a user keys for text reading. In particular, our desire is to inform her/him about text contents, and text or segment role; it means to notify her/him about text topics and relations of segments in text structure.

The originality of our description task rests on the inheritance principle between a text and the features of some specific linguistic expressions contained in text. As a definition, we call descriptors expressions which play a role in text description and make easier their understanding. So segments inherit properties of its descriptors and a text structure is created from the main global topic(s) towards the very local ones.

3.1 Descriptor categories

For term categorization, we foreground the two following features: “The frequency” - number of occurrences - and the “lexical distribution” - location of the occurrences - are relevant hints to make it possible to determine labels of terms. We consider this proposal as true at least at the macroscopic level (i.e. segment, document, corpus levels), even if we agree that other information (e.g. grammatical role, position in the sentence, etc.) are necessary at the microscopic level (i.e. sentence, clause levels).

This lead us to distinguish two main types of descriptor labels: the topical and the generic ones (also called meta-descriptors). The former can be subdivided into global and local thematic labels. Table 1 gives a more complete overview of descriptors categories. The topic descriptors are considered inside a single document, and they correspond specifically to “what a text is about”. Meta-descriptors correspond to the lexicon shared by the majority of the documents of our corpus except the stop-words. In our particular case, working with a scientific corpus, the generic lexicon is composed of such words : “example, hypothesis, analyze, problem, etc.”.

3.2 Characterization: the lexical intersection method

From a theoretical point of view, the operation which is realized to obtain the various categories is a lexical intersection between all the units of a text, children of the same depth level, going from the upper level - the corpus - to the lower one - the sentences in the segments.

![Figure 2. Various units intersection](image)

For example in our corpus of scientific gender, the descriptors common to most documents of the corpus are labelled as “meta-descriptor”; the lexical intersection of the

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Frequency</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global theme</td>
<td>High in text</td>
<td>Spread in text</td>
</tr>
<tr>
<td>Local theme</td>
<td>High in text</td>
<td>Local in text</td>
</tr>
<tr>
<td>Meta-descriptor</td>
<td>High in corpus</td>
<td>Spread in corpus</td>
</tr>
<tr>
<td>Situation-word</td>
<td>Weak in corpus</td>
<td>Fortuitous</td>
</tr>
<tr>
<td>Stop-word</td>
<td>High in all</td>
<td>Spread in all</td>
</tr>
</tbody>
</table>

Table 1. Descriptor categories
segments of a document corresponds to terms spread all over the text or over several segments and are annotated as “global topic descriptor”, and lastly the lexical intersection of the sentences of a same segment corresponds to terms only occurring inside a segment and so called “local topic labels” (see figure 2).

4 Term extraction and characterization

Since descriptor feature complexity depends on the text unit in scope, we examined different linguistic expressions for topic descriptors and generic descriptors.

For generic descriptors, only examined at the corpus level in term of frequency and lexical distribution, we did not suppose a specific type of phrases, and so, we worked with n-grams, which are consecutive sequence of “n” words; this approach offers us a rich range of phrases and prevents us to abusively remove some interesting and not yet observed phrases. Topic descriptors, since it is observed at a document level, require more features. For this reason, we focus our attention on noun phrases. This choice has also been led by some previous works, emphasizing the thematic role of these phrases [1].

4.1 Corpus

We work with a heterogeneous thematic corpus, whose type is specified, here technical and scientific. Each document deals with a specific subject which makes it to be distinguished from others. First, our system has been conceived for French processing. Currently, we adapt it to process English texts too. As an English corpus we choose to use the Computation and Language (cmp-lg) collection, part of the TIPSTER SUMMAC effort. Its markup permits us to easily filter headings, abstracts and references. In order to light the various processes of our system we worked on 76 documents selected by chance. Each of them contains about 5000 tokens (words and punctuation).

4.2 Meta-descriptors acquisition

To perform term acquisition, we based our approach on statistical filters (see [2] for a more complete synthesis). Our specificity arises from combining lists of expressions holding different frequency and lexical distribution properties to filter each other.

4.2.1 Considering “n-grams” as input

For meta-descriptor acquisition we choose to consider n-grams as linguistic expressions. The main advantage is that recall of generic terms depends only on the maximal size of the n-grams and not on syntactic pattern. In our experiments, we handled n-grams from 1 to 5 word-components. Nevertheless the main drawback of n-grams is that, even filtered by its frequency, they do not always correspond to a fixed phrase. We set that a fixed phrase is a phrase which is normally fixed or change very little. If a n-gram is not a fixed phrase, it can be a flexible phrase, it means a slight variation of a fixed one; or it can be no such phrases at all.

4.2.2 Basic algorithm

The base of our algorithm lies on a lexical intersection operation. Indeed to determine generic terms, we actually perform a mathematical intersection of expressions embedded at a same level unit. For meta-descriptors documents are considered as units.

- first, we extract from each document a unique instance of each expressions;
- second, we merge these expressions into the same list;
- the frequency of an expression in this list corresponds to the number of documents in which it occurs. A threshold can be set to filter the less frequent expressions.

So far, lists of extracted meta-descriptors remain quite satisfactory and, even in noisy situation, it can still be used as a tool to assist a manual extraction or validation.

4.2.3 Automatic selection

Since we wanted to provide an automatic process, we implemented a heuristic to fully exploit the n-grams and determine the fixed ones. This heuristic is similar to the lexical intersection mechanism :

- first, considering the n-grams indifferently, we merge all lists of unique n-grams of each document into a single list.
- then, the frequency of each form is used to order them and to filter them.

Next step is the heart of the process. We assume that whether two expressions have got a close frequency and that one of them can contains the other, then probably these expressions are only one (the contained one being a variation of the container one).

This lead us to give a computational definition of a fixed phrase. A n-gram linguistic expression is called fixed phrase from a computational point of view :

(1) if it does not exist other expressions which could contain it in its surrounding “neighborhood”;

...
This “neighborhood” is expressed according to the frequency of the phrase in scope. Empirically, we set it to 20% of the frequency of the phrase. The table 2 shows this filtering task with $F$ for the phrase frequency, $N$ for the frequency of the more distant neighbor considered in the 20%.

<table>
<thead>
<tr>
<th>$F$</th>
<th>Expressions</th>
<th>$N$</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>example ,</td>
<td>126</td>
<td>Not a candidate term</td>
</tr>
<tr>
<td>142</td>
<td>for example</td>
<td>113</td>
<td>Not a candidate term</td>
</tr>
<tr>
<td>135</td>
<td>for example</td>
<td>108</td>
<td>Not a candidate term</td>
</tr>
<tr>
<td>130</td>
<td>. for example</td>
<td>104</td>
<td>Not a candidate term</td>
</tr>
<tr>
<td>127</td>
<td>. for example</td>
<td>101</td>
<td>Is a candidate term !</td>
</tr>
<tr>
<td>102</td>
<td>example of</td>
<td>81</td>
<td>Is a candidate term !</td>
</tr>
<tr>
<td>84</td>
<td>example , the</td>
<td>67</td>
<td>... ?</td>
</tr>
<tr>
<td>76</td>
<td>an example</td>
<td>60</td>
<td>... ?</td>
</tr>
</tbody>
</table>

Table 2. Automatic filtering

According to the fixed phrase definition, our algorithm filters the 1-grams. After removing the stop-expressions, we re-insert those occurring at least one in one fixed phrase. Stop-expressions are combinations of stop-words (determiner, pronoun, preposition and punctuation) and usual thematic words (as such “linguistic, syntax, phrase, etc.”). The list of thematic words was made by hand looking at most frequent 1-gram list. Eventually, stop-expressions are removed of the final list.

We performed several tests with frequency threshold and neighborhood radius. Moreover we made a test by filtering all punctuation signs since the beginning. Figure 3 depicts the quantity of n-grams for a given frequency. The upper curve stands for n-grams with punctuation.

So by arbitrarily fixing, different frequencies, we obtain various quantity of fixed phrases. Table 3 informs about the number of fixed phrases we retrieve according to various frequency threshold, with a neighborhood radius of 0.2.

<table>
<thead>
<tr>
<th>N-grams type</th>
<th>N-grams total</th>
<th>$F$ 15</th>
<th>$F$ 10</th>
<th>$F$ 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punctuation</td>
<td>1057548</td>
<td>301</td>
<td>613</td>
<td>2040</td>
</tr>
<tr>
<td>Without</td>
<td>930286</td>
<td>210</td>
<td>411</td>
<td>1504</td>
</tr>
</tbody>
</table>

Table 3. Number of fixed phrases with various frequency $F$

We first remarked the quality of the extracted list. A second statement concerned the variety of types of descriptors we extract. Although a deeper study needs to be performed to specifically categorize all of them, it is important to notice that these phrases could play a role in building the rhetorical structure of scientific texts.

| by mean of combination of the representation of the basis for identification of have be implement available . as far as a collection the sequence of the default sort of the model the importance of property of the solution be show to point of view the result this approach be rate of in this section , be obtain by a theory of a result of the meaning of the this mean that in this paper we focus on the method the goal of the ability to so as in the sense in particular as follow : the output of in case of it be necessary be implement in as oppose to a consequence an algorithm the degree see section the condition on the long the figure suggest a small number of in the same way |
| the sequence of the default sort of the model the importance of property of the solution be show to point of view the result this approach be rate of in this section , be obtain by a theory of a result of the meaning of the this mean that in this paper we focus on the method the goal of the ability to so as in the sense in particular as follow : the output of in case of it be necessary be implement in as oppose to a consequence an algorithm the degree see section the condition on the long the figure suggest a small number of in the same way |

Table 4. Excerpt of our final list of fixed phrases

The presence of neighbors depends on the maximal $n$-size of n-grams, the threshold of the minimal frequency for n-gram instance, and the neighborhood radius. We do not have a solid theory to determine the optimum threshold. Nevertheless they should be set according to the frequency and the complexity of a phrase.
4.3 Identification of topic descriptors

Topic descriptor retrieval and identification is proceeded in three successive steps (see figure 4). We take the work of [1] as a starting point. The first stage aims at retrieving referring expressions (nominal and pronominal expressions) referring to the entities of the text (i.e. “what the text is about”). The second stage is a robust anaphora resolution system (ARS), which permits to improve the frequency and the distribution measures of each concept of the text, and on the other hand to identify the representative forms of a concept by filtering anaphoric forms. The third stage aims at characterizing the topic relevance of identified entities according to the topic structure of the text.

4.3.1 Referring expression retrieval

The retrieval process of referring expressions consists of extracting noun phrases and pronouns by using syntactic patterns. We distinguish the simple noun phrases from complex noun phrases; the first composed only of nouns and adjectives, and the last of nominal phrases admitting the preposition “of”.

4.3.2 Anaphora Resolution System

This system aims at identifying text entities. First our system looks at the local level to identify local anaphoric relation. Second, it binds salient entities at the global level.

We specifically payed attention at anaphoric expressions which are most likely to be a reprise of an entity previously introduced in the text. Indeed some of them are more ambiguous than others. So we focused our attention on noun anaphors inserted by demonstrative determiners and pronoun anaphors such as personal, reflexive and demonstrative ones. We filtered most usual non-personal pronouns and only the noun phrase were considered as potential antecedent.

At the local level, the system follows a pipeline architecture made up of three stages, as follows:

1. First we select the potential antecedents for a given anaphor by examining its current and prior last sentences.
2. Then we rank the antecedent candidates as the most probable to refer to the anaphor;
3. And eventually we filter them according to their agreement with the anaphor in scope.

In comparison with [8, 1], we inverted these two last stages. Since it exists an antecedent for the anaphors we got in concern, our system should always determine an antecedent. So filters are ordered by constraint degrees. The first antecedent which validates the highest degree is considered as the anaphor’s antecedent.

Selection, filtering and ranking tasks are based on morphological indicators (gender/number agreements), lexical indicators (identical semantic head), syntax indicators (parallel grammatical relations, kinds of grammatical relation), discursive indicators (distance between candidate antecedents and an anaphor).

The ranking task determines two different weights; one is inherent to each expression and the other according to the type of the anaphor in focus. Most of those indicators were computed by robust heuristics (e.g. grammatical role -Subject, Obj- are assigned according to the position of the referring expression compared to the verb of the clause and according to its position compared to a preposition).

At the end of the local pass, we get set of expressions referring to a same entity. Some of them are considered as salient entities whether they validate a minimal inherent weight. This threshold is set to consider some grammatical relations (such as subject or object), repetition, and demonstrative expressions as salient. The entity properties are the mean of its individual expressions.

The global pass attempts to match entities according to their preferred form. It means the most frequently observed.
4.3.3 Characterization of topic descriptors

This part concerns only salient entities. We characterize them with topic labels according to their role in topic text structure. We count the number of topic segments in which an entity occurs to characterize it. The presence in only one segment corresponds to a local thematic label. The presence in at least two segments is equivalent to a global thematic label. In this last case, certain terms can benefit of a global thematic label “of all the text” if they appear in the greatest number of segments.

Our future objective will be to build a text topic hierarchy according to elaboration or association relations.

5 Text visualization

Currently, due to our running task for adapting the process to English, we do not have a full text description in English. The example in figure 5 is a text fragment extracted from an article of a French scientific news paper called “la Recherche”. The article deals with “vin jaune” or “yellow wine”. The author wonders which molecule is responsible of its specific flavour. In particular, the fragment presents the most straight method of wine analysis.

Figure 5. Example of segment description

Thanks to our term labelling process, we identify in this text segment “wine/vin” as playing a global thematic role; which sounds as normal because the entire text is about wine. In the same way, we identify mixture/mélange and element/composé as playing local theme role. Both reinforce the role played by terms identified as indicative label, it means analysis/analyse and technique/technique. As a note, we point out that a global theme and a meta-label term occur in the first sentence of the segment and that they describe correctly the text fragment since we can type it as a description of technical analysis of wine.

6 Conclusion and future work

Our purpose is to facilitate access to the text contents. We provide a set of text analysis in particular text structure, rhetorical segment role and topic identification. Each one of these analysis can be considered as many different abstract views of a text. The user can so access to the text by visualize such or such view and navigate from one to another.

For example, a scenario of text exploration could follow the different degrees of topic elaborations [5] and segment description could lie on presentation of linguistic expressions in which different relevant descriptors occur.

Besides this, it seems important to examine more specifically meta-descriptors that we acquire, in particular by studying collocation interactions in order to automatically perform a better understanding of a text.

The final version of this paper will relate to an evaluation part by comparing topic descriptors and meta-descriptors to titles of sections of documents.

References