Automatic Summarization Based on Sentence Morpho-Syntactic Structure: Narrative Sentences Compression

Mehdi Yousfi-Monod, Violaine Prince

Automatic summarization based on sentence morpho-syntactic structure: narrative sentences compression

Mehdi Yousfi-Monod, Violaine Prince
LIRMM, UMR 5506
161 rue Ada, 34392 Montpellier Cedex 5, France
Email: {yousfi, prince}@lirmm.fr

Keywords: automatic summarization, sentence compression, syntactic analysis

Abstract: We propose an automated text summarization through sentence compression. Our approach uses constituent syntactic function and position in the sentence syntactic tree. We first define the idea of a constituent as well as its role as an information provider, before analyzing contents and discourse consistency losses caused by deleting such a constituent. We explain why our method works best with narrative texts. With a rule-based system using SYGFRAN's morpho-syntactic analysis for French [Cha84], we select removable constituents. Our results are satisfactory at the sentence level but less effective at the whole text level, a situation we explain by describing the difference of impact between constituents and relations.

1 Introduction

The amount of information available on the Web or in some compagnies, administrations and laboratories doesn't stop increasing, thus hardening information retrieval on such resources. Automatic summarization, aiming at considerably reducing the size of such data, appears to be a good solution to ease this search. It does so by introducing a smaller but relevant text, and thus shortens choice time of a request, concerning text relevance acceptance.

The main idea of our research is to find texts contraction bounds by sentence compression without major content loss. The originality of our approach is to rely on the constituents syntactic function and position in the syntactic tree to select deletable constituents.

In next section, we enumerate the main automatic summarization's approaches types, then we compare those working at a finer granularity level (section 2); we then outline our sentence compression method (section 3); we continue by illustrating the effectiveness of our approach with a prototype application based experimentation applied to story/short novel type texts (section 4); and finally we discuss about the results of this experiment and draw some perspectives (section 5).

2 Summarization by sentence compression

In this article, we only focus on sentence compression.

[KM02] tackles the sentence compression problem by using a *noisy-channel model* consisting in making the following assumption: "We look at a long string and imagine that it was originally a short string, and then someone added some additional, optional text to it. Compression is a matter of identifying the original short string.". The aim is then to locate this optional text and to remove it. To do so,the authors use a Bayesian probabilistic model trained on a corpus composed by documents with their summary.

[Sid02] focuses on detecting and removing relative clauses which are preceeded by clauses like $NP_1\ Prep\ \ NP_2$, where NP_1 and NP_2 are noun phrases and $Prep$ is a preposition. The purpose is to correctly attach the relative referent by choosing a wide or local attachment.

These two approaches based on textual units shorter than sentences do not take into account the sentences constituents syntactic function and position in the syntactic tree. In fact, function and position are naturally useful to help choosing the constituents to be removed. Moreover, such a technique is easily checked by human examination.

3 Compression by pruning the syntactic tree

The starting point of our approach was the insight that **the sentence constituents syntactic function and position in the syntactic tree plays a weighty role in the constituents importance for the text understanding**. This insight comes from logical grammatical analysis always taught and whose there are many well known manuals. Indeed, some adjective phrases, adverbials, etc, are not systematically needed to understand the main sentence's meaning,

This approach needs a sentence morpho-syntactic analysis tool (section 3.1) and a survey on constituents importance relative to their syntactic function and position in the syntactic tree (section 3.2). We present our system architecture in the section 3.3.

3.1 The morpho-syntactic analyser

Since our working language is French, our experiments have been run on this language. However, the same methods can be easily transposable to English or other languages for which syntactic parsers have been developed.

We use the French morpho-syntactic parser called SYGFRAN, based on the operational system SYGMART, both defined in [Cha84]. SYGFRAN uses a transformation rules set of structured elements, based on French grammar rules. It transforms a sentence (raw text) in a syntactic tree (structured element) enriched with information about constituents. This parser has the following advantages: **the fastness**: the analysis complexity is $O(k * n * log_2(n))$ where k is the rules number and n the text length. **the robustness**: SYGFRAN manages to produce a correct structure for at least 30% of the different cases of French sentences syntaxes, for other cases, SYGFRAN provides a partial *but workable* analysis. **the production of a syntactic tree**: much of the existing syntactic analysis systems only achieve a basic linear tagging and those providing a tree are not robust enough relatively to the body of existing syntactic constructions.

SYGFRAN takes a raw text input and produces a bracketed structure, corresponding to the morpho-syntactic tree of each text sentence, in which many variables are acquainted on the different constituents natures, syntactic functions, canonical forms, grammatical categories, tense, gender, number, etc.

3.2 Function and Position

The constituents deletion test is addressed by many French grammar works to help in syntactic function attachment of a constituent. The test is validated if the resulting sentence remains grammatically consistent. However, linguistic texts dealing with the constituent importance in the sentence according to their syntactic function are rather uncommon. Some recommendations are provided by linguists, but there is no fundamental rule.

So we have proceeded in the following way. We have considered these recommendations as working assumptions and we have tried to support them empirically. Mel'čuk, in his contemporary French analysis, speaks about syntactic functions known as *governement* (in the aftermath of Chomsky's works). Are **governors**, constituents being considered as indispensable to the grammatical coherence and to the sentence semantics. The sentence subject and its verbal group are governors in a grammatical coherence viewpoint.

We have noted three constituent categories likely to be deleted, according to their syntactic function and their position: adverbials, epithets and appositions. As we can see, they have a medium granularity level. Appositions, when transformed in relative clauses (noun complement) get a wider granularity level, thus they increase the final compression ratio.

Adverbials. We have noticed that the most important adverbials where *temporal* and *purpose* ones. They do answer the questions we deem the most important namely "When ?" and "In which purpose ?" In the case where a location adverbial is present after the verb "to be", deletion cannot be done. "to be" is a particular verb, and must be cautiously dealt with.

However, if several location adverbials are consecutive, all but one can be deleted without major content loss : "*John is in the car, in the car park, near to the sweet shop.*" At last, adverbials located in interrogative sentences appear to be extremely important since they do issue the question.

Epithets. Adjectives, adjective phrases and some relative clauses (noun complement) have an epithet function. In a way similar to adverbials, when an epithet si located after the verb "to be", and more generally after a stative verb, its importance considerably increases, making deletion impossible.

Also, we have noticed that when the epithet is located in a noun phrase in which the determiner is a definite article, then its deletion is difficult. The reason is the definite article is used to speak about a specific entity and , thus the noun epithet allows to differenciate this entity from others.

Appositions. Apposition may be of different types and might appear asa noun phrase, a pronoun, a relative clause, a present participle clause, a past participle clause, an infinitive clause.In the first three cases, constituents can be easily deleted. Participle clauses

can be deleted too, but with a more important content loss. In the latter case, deleting the clause appears to be more difficult, because the infinitive clause systematically provides an important information completing the subject.

3.3 Architecture

Our system architecture is outlined in figure 1. It relies on all considerations provided in the preceding section about the importance of constituents in a sentence. It is based on a parser output in the form of syntactic trees, and produces as an output, a text coloration of the deletable segments according to this constituents hierarchy. This is the way the system works: source text is fed to SYGFRAN, which in turn produces syntactic trees. Then, the textual segment selection/coloration module uses the following informations to accomplish the selection: the source text, the syntactic trees and variables/values provided by SYGFRAN, the size/loss ratio threshold not to exceed, being provided by the user or defined by the application type and the constituents selection rules set to achieve the different constituents selection iterations until the size/loss ratio is satisfied. The selected constituents are then deleted.

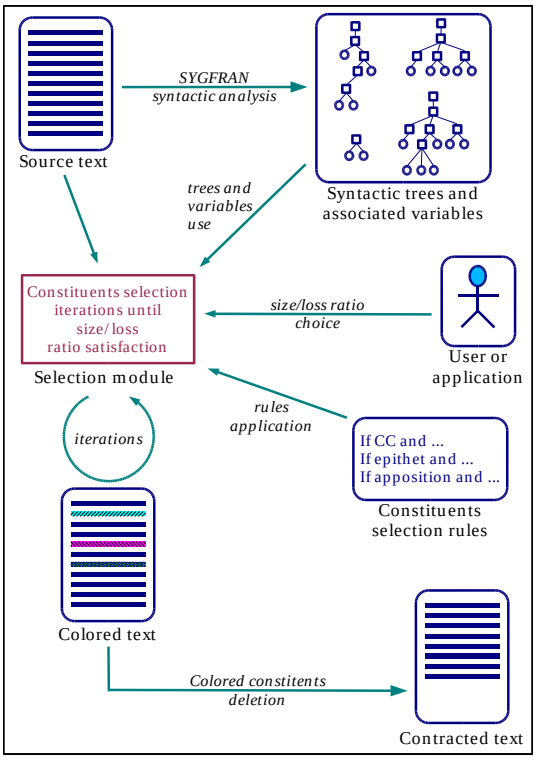

Figure 1: From the source text to the compressed text: our sentence compression system

4 Experiment

We have have implemented a part of our theory in a computer program to assess the effectiveness of such an approach. We have defined a system using basic rules, based on our experimental survey's results (section 3.2):

Our current propotype only performs one iteration. The first step consists in coloring deletable constituents. A color is assigned to each constituent type. So it is easy to assess rules quality on the processed text before actually deleting these constituents.

In the second step, colored segments are deleted to produce the summary. The chosen text is a French Haitian story. We have chosen a French text because the current rules set of SYGFRAN allows it to analyze only French sentences. The reason of choosing this story is that SYGFRAN produces a correct syntax for all the sentences of this text and because it is a well-sized, good representative of what is a narrative text. The coloration result of a story part is presented in the figures 2 (the orignial French version) and 3 (the English translated one).

5 Discussion

With our current rule set, our method has allown us to delete approximatively 34% of the full text. We can note a light discursive content and coherence loss, which is more than satisfactory relatively to current automatic summarizers. Moreover, the grammatical consistency is preserved. We think our rules can be more refined, but there is a lack of linguistic information in this domain. For this text, SYGFRAN provides us correct syntactic trees, but variable values are not systematically true and full. For adverbials, SYGFRAN only specifies the object semantics for the temporal and locative ones. The other somehow lack semantic information.

Selecting rules of deletable constituents can be more refined according to constituent function and especially to text types. Concerning this subject, we project to carry out experiments on more texts dealing with more different types. However, sentence compression is not sufficient to produce a summary of a satisfying size in most application cases. As we have already seen, compression greatly depends on the text type. So we consider our intra-sentential approach as one of the tasks to perfom in the automatic summary production, in complement with other approaches working at a granularity level at least as big as sentences.

Au bout d'un moment elle bougea et marmonna: "Quelle sorte de nuit est-ce donc pour durer si longtemps ?" Mais elle se rendormit parce qu'il faisait aussi noir qu'au cœur de la nuit dans la maison. Finalement elle se réveilla en sursaut et se mit à chercher ses vêtements. Courant de tous côtés, elle arracha ce que Maui avait fourré dans les fentes. Mais c'était le jour! Le grand jour! Le soleil était déjà haut dans le ciel ! Elle s'empara d'un morceau de tapa pour se couvrir et se sauva de la maison, en pleurant à la pensée d'avoir été ainsi trompée par ses propres enfants . Sa mère partie, Maui bondit près du store qui se balançait encore de son passage et regarda par l'ouverture. Il vit qu'elle était déjà loin, sur la première pente de la montagne. Puis elle s'arrêta, saisit à pleines mains un arbuste de tiare Tahiti, le souleva d'un coup: un trou apparut, elle s'y engouffra et remit le buisson en place comme avant.
Maui jaillit de la maison aussi vite qu'il put, escalada la pente abrupte, trébuchant et tombant sur les mains car il gardait les yeux fixés sur l'arbuste de tiare. Il l'atteignit finalement, le souleva et découvrit une belle caverne spacieuse qui s'enfonçait dans la montagne.

Légende : compléments_circonstanciels, proposition_au_gérondif, propositions_relatives, groupes_adjectivaux.

Figure 2: Our text coloration/compression, original French version

After a moment, she stirred and muttered; "what type of a night it is to be so long" ?
But she went back to sleep because it was as dark in the house as in the core of the night . Finally she woke up with a start and began to look for her clothes. Running everywhere she tore up what Maui had slipped into the holes. It was day ! The full bright day ! The sun was already high up in the sky! She took a piece of tapa to cover herself and fled from home,weeping at the thought that she had been so deceived by her own children.
His mother gone Maui jumped close to the window shade that was still moving after her and looked through the openning. He saw that she was already far away ,on the first slope of the mountain. Then she stopped, grabbed a Tahiti tiara bushtree with her whole arms and lifted it up completely : a hole appeared, and she rushed in and then put the bushtree back like before.
Maui sprang up from the house as quickly as possible, climbed up the abrupt slope , stumbling and falling on his hand, because his eyes were kept on the tiara bushtree.
He finally reached it, lifted it up and found a beautiful spacious cave that went deep under the mountain.

Legend : adverbials, gerund_clause, relative_clause, adjective_group.

Figure 3: Our text coloration/compression, English translated version

6 Conclusion

Current automatic summarization approaches use information such as term frequency, lexical relations, POS tags, probabilistical learning engines, texts rhetorical structure, however, none of them use both **constituents syntactic function and position in the syntactic tree** as our is able. Our approach has started by a survey on the sentence constituents importance. The deletion criterion evaluates the contents and coherence loss generated by constituents deletion. The selection criterion is based on constituents syntactic function and position in the syntactic tree. Narrative texts (novels, stories, ...) appeared to be the most suitable for such an approach. We have modeled a sentence compression system based on constituents deletion. The creation of a rule system based on our model has allowed us to assess the feasibility of such an approach. We have first colored the constituents according to selection rules, in order to judge the relevance of each rule. Our method managed to delete approximatively 34% of the test text, while preserving a good grammatical coherence. We thus conclude that our compression could be useful when used as one of the tasks of a wider automatic summarization process, either as a first-phase running summarization, or as a post-phase, after having removed larger chunks of text. We plan to augment accuracy of text sentences pruning by running our system on important narrative text corpora, find heuristics for wider portions of text deletion based on rethorical information use text types or domains to introduce specific summary rules (scientific articles in which titles might help to delete wide portions of text). All this, naturally, will be sorted out and put into a more sophisticated system to provide a better set-up for summarization by compression.

REFERENCES

[Cha84] Jacques Chauché. Un outil multidimensionnel de l'analyse du discours. In *Coling'84*, pages 11–15, Standford University, California, 1984.

[KM02] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence archive*, 139(1):91–107, Jully 2002.

[Sid02] Advaith Siddharthan. Resolving relative clause attachment ambiguities using machine learning techniques and wordnet hierarchies. In *5th National Colloquium for Computational Linguistics in the UK (CLUK 2002)*, pages 45–49, 2002.