

Sentence Compression as a Step in Summarization or an Alternative Path in Text Shortening

Mehdi Yousfi-Monod

University of Montpellier 2, CNRS
LIRMM, 161 rue Ada
34392 Montpellier Cedex 5
yousfi@lirmm.fr

Violaine Prince

University of Montpellier 2, CNRS
LIRMM, 161 rue Ada
34392 Montpellier Cedex 5
prince@lirmm.fr

Abstract

The originality of this work leads in tackling text compression using an unsupervised method, based on a deep linguistic analysis, and without resorting on a learning corpus. This work presents a system for dependent tree pruning, while preserving the syntactic coherence and the main informational contents, and led to an operational software, named COLIN. Experiment results show that our compressions get honorable satisfaction levels, with a mean compression ratio of 38 %.

1 Introduction

Automatic summarization has become a crucial task for natural language processing (NLP) since information retrieval has been addressing it as one of the most usual user requirements in its panel of products. Most traditional approaches are considering the sentence as a minimal unit in the summarization process. Some more recent works get into the sentence in order to reduce the number of words by discarding incidental information. Some of these approaches rely on statistical models (Knight and Marcu, 2002; Lin and Hovy, 2002; Hovy et al., 2005), while some other works use rule-based linguistically-motivated heuristics (McKeown et al., 2002; Dorr et al., 2003; Gagnon and Sylva, 2005) to improve the determination of the importance of textual segments. Considering a deeper linguistic analysis could considerably improve the quality of reduced sentences, we decided

to develop a sentence compression approach exclusively focused on linguistic heuristics. Our first work (Yousfi-Monod and Prince, 2005), slightly anterior to (Gagnon and Sylva, 2005), showed interesting results, and led to a deeper and more complete work fully detailed in (Yousfi-Monod, 2007). This paper sums up our approach.

2 Theoretical framework

The main hypothesis of this work leans on the observation that incident sentence constituents are often not as important as principal constituents. For instance, let us consider the temporal adverbial in the following sentence: "Taiwan elected on Saturday its first president". While the subject 'Taiwan' and the verb 'elected' are principal constituents of the sentence, 'on Saturday' is incident and can be removed without causing neither an agrammaticality nor a weighty content lost. Two aspects of the constituent significance: *Grammaticality* and *content*, are dealt with in this section.

2.1 Grammaticality preservation thanks to the syntactic function

The principal/incident constituent principle can be found in constituent or dependency grammar representations. The approach embedded in COLIN is based on such grammars, while adapting them to the relevant proprieties for sentence compression. As we aim at preserving sentences grammaticality, our first goal is to get a syntactic tree based on the grammatical importance, where for each node, a daughter node is an incident constituent which may be removed under certain conditions. We opted for the X-bar theory (Chomsky, 1970), which represents a sentence through a tree of constituents, composed by *heads* and *governed constituents* (also dependents). While a head is

grammatically mandatory, its dependents can often be removed, depending on some of their linguistic properties and/or those of their heads. Our goal is first to have a syntactic structure modeling based on constituents grammatical importance. Syntactic writing rules of the X-bar theory are focusing on sentence construction by placing specifiers, complements and adjuncts in the subtree of their constituent. While adjuncts are systematically removable, we have had to adopt a case-by-case study for specifiers and complements. For instance, in a noun phrase (NP), the article, if present, is a specifier, and it cannot be removed, while in an adjective phrase, the specifier is typically an adverb, which is removable. The removability of a complement depends on the subcategorisation properties of its head. On a clause scale, the dependents are not well defined in the X-bar theory and may include the subject and verbal groups, as, respectively, the specifier and the complement of the clause. Thus, the specifier (subject) cannot be removed. Our study has then consisted in a categorization of the X-bar's functional entities according to their removal property. We have decided (i) to consider mandatory specifiers as complements required by their head and (ii) to bring together optional specifiers and adjuncts in a different category: *Modifiers*¹.

We have defined two classes of functions: *Complements* (X-bar complements and mandatory specifiers) and *Modifiers* (X-bar adjuncts and optional specifiers). This syntactic function classification allows us to clearly define which sentence objects can be candidates for removal.

Nevertheless, the syntactic function information, although crucial, is not sufficient. One has to use other linguistic properties in order to refine the assessment of the constituent importance.

2.2 Important content preservation thanks to linguistic proprieties

Subcategorisation. For noun and clause heads, some of our complements have been identified as systematically mandatory in order to preserve the sentence coherence (subject, verbal group, articles...). Other heads (verb, adjective, adverb, preposition and pronoun) may admit optional or mandatory complements, depending on either the lexical head category or a particular head instance

¹We have chosen the term 'modifier' as its definitions in the literature fit quite well our needs.

(a lexical entry). Indeed, prepositions are systematically requiring a complement², while other heads must be considered on a case-by-case basis. Once we get the subcategorisation information for a head, we are able to determine whether its complement(s) can be removed without causing an incoherence.

Other linguistic proprieties. We identified several other linguistic clues that may help assessing the importance of dependents. We do not detail our analysis here for space reasons, refer to (Yousfi-Monod, 2007) for the full description. These clues include lexical functions, fixed expressions, type of the article (definite or indefinite), parenthetical phrases, detached noun modifiers, the dependent constituent position in the sentence, negation and interrogation.

3 COLIN's compressor: System architecture and implementation

3.1 Architecture

We assume we have a raw text as an input, which may be the output of an extract summarizer, and we have to produce a compressed version of it, by reducing as many sentences as we can, without deleting a single one.

Syntactic analysis. This step consists in using a syntactic analyzer to produce, from the source text, dependent trees according to our syntactic model (heads, complements, modifiers). In order to handle the important content assessment, the parser uses linguistic resources including subcategorisation information, lexical functions, and the other linguistic properties (section 2.2), and then enriches the trees with this information.

Pruning and linearization. The trees will be then pruned according to a set of compression rules defined from our theoretical analysis. Several set of rules can be defined according to (i) the desired importance conservation, (ii) the desired compression rate, (iii) the confidence in syntactic analysis results, (iv) the trust in the identified linguistic clues, (v) the textual genre of the source text. In order to get effective rules, we have first defined a relatively reliable kernel of rules. Then we have decided to define and test, during our evaluation

²Accordingly to the X-bar structure as well as ours: The preposition is the head of the prepositional syntagm.

described in the next section, several rules configurations, taking into account each of the five points, in order to find the most effective ones. Rules tag each tree node (complements and modifiers) which will be removed, then trees are pruned and linearized to get back new sentences, compressed.

3.2 Implementation

The first step in our implementation was to select a parser satisfying our syntactic requirements as much as possible. SYGFRAN (described in (Yousfi-Monod, 2007)), is the one that has been chosen as: (i) It produces constituent trees very close to our model, (ii) it has a good syntactic coverage, (iii) it has a very good parsing complexity ($O(n \cdot \log(n))$, with n the size of the data in words), and (iv) its author and developer, Jacques Chauché, works in our research team at LIRMM³, which considerably eases the adaptation of the syntactic model to ours. SYGFRAN consists in a set of grammar networks, each of them containing several set of transformational rules. COLIN and SYGFRAN's rules are implemented with the parser SYGMART, a tree transducers system (Chauché, 1984). COLIN's rules are split into several grammars including (i) a basic anaphora resolution, (ii) a tagging of candidate nodes⁴, (iii) a pruning of tagged constituents and a linearization of leaves.

4 Evaluation, experimentation and results

This section sums up the validation process used for our approach. Our evaluation protocol is manual intrinsic, focuses on facilitating the evaluator's task and is inspired from (Knight and Marcu, 2002)'s one. For space reasons, we do not detail the protocol here, a full description of the protocol as well as the experimentation is available in (Yousfi-Monod, 2007).

Setting up. As our approach deeply relies on syntactic properties, which are not always properly detected by current parsers, we decided to manually improve the syntactic analysis of our evaluation corpus. Otherwise, the evaluation would have more reflected the users' satisfaction about parsing than their opinion about the quality of our importance criteria. In order to assess the genre influ-

ence, we selected three genres: Journalistic, narrative and scientific. We composed 5 texts per genres, each of them contained about 5 paragraphs, 16 sentences and 380 words, thus a total of 240 sentences. We decided to test the importance of different clause modifiers, i.e. adverbial phrases, according to their type. We considered the following types: Temporal, locative and other ones. So, while keeping the core rules for each rules configuration, we tested the removal of (i) all adverbials, (ii) temporal adverbials, (iii) locative adverbials, (iv) only other adverbials (keeping temporal and place ones).

We got 25 active users participating to the evaluation, who were mainly composed of PhD students and PhDs, working in NLP or computational linguistic domains, and being fluent in French. Some of them did a set of manual compressions, used to compare the quality with COLIN compressions in the scoring stage of the evaluation. 59 text compressions were done, corresponding to about 3,9 compressions per text. In the scoring stage, judges gave about 5,2 notations per compressed paragraph for manual and automatic compressions.

Results. Tables 1 and 2 present the results of respectively obtained average compression rates⁵ and paragraph scorings⁶, per genre. For COLIN's evaluation, we only display the rules configuration which has obtained the best results for the compression rate relatively to the paragraph scoring, i.e. the rules configuration (iv).

	Jour.	Narr.	Scien.	Mean
Manual	36 %	17 %	23 %	25 %
COLIN	38 %	35 %	41 %	38 %

Table 1: Average compression rates.

	Jour.	Jour.	Scien.	Mean
Manual	4,03	3,67	3,41	3,7
COLIN	3,7	3	3	3,23

Table 2: Average paragraph scorings.

The compression rate proposed by COLIN is quite better than the manual one for a quality scoring just below the latter. COLIN is obviously

³<http://www.lirmm.fr>

⁴We tag trees before pruning them as COLIN can work in a semi-automatic mode (not presented here) where a user can modify the tagging.

⁵A higher percentage means a shorter compressed text.

⁶Scores are between 1 and 5, a value of 1 means a completely unsatisfying compression, while a value of 5 means a very good compression for the judge.

far better in compression time, with about 5 seconds per document versus between 200 and 300 seconds for the manual compressions. COLIN's compression-quality-time ratio is therefore really better than the manual compressions. Each genre obtained a good compression rate as well as a correct quality scoring, particularly for the journalistic one. Note that our results could have been improved if they weren't sensibly degraded because of an imperfect parsing, despite some focused improvements we did on it.

A performance comparison with similar approaches was an issue for our approach for at least two reasons: (i) As our parser is exclusively for French, we had to do comparisons with French tongue systems only. The system presented in (Gagnon and Sylva, 2005) is the only that matches this constraint. (ii) Our evaluation protocol drastically differs from traditional ones in several points: 1. Having a single human judge who compresses sentences produces compressions which are too much subjective to the latter, that's why each of our texts were compressed about 4 times by different humans. Evaluating compressions rise the same issue of subjectivity, so each of our compressions were evaluated about 5 times. 2. We consider assessing the quality of separated compressed sentences is harder and less relevant for evaluators than assessing full paragraphs as we did. 3. Text genre has an actual influence on NLP approaches, thus we took into account this factor in our evaluation, as described above, while the above cited system extracted random sentences in a single corpus. For all these reasons, we haven't been able to perform a comparison with the above cited system yet.

5 Conclusion

In this paper we have addressed the task of sentence compression based on a deep linguistic analysis. The system we developed, called COLIN, theoretically relies on a constituents and dependencies sentence tree pruning, removing those branches which could be cut without jeopardizing the sentence construction, or tempering too strongly with the sentence meaning. A careful study of syntactic properties, lexical functions, verbs arguments has led us to design several different configurations in which the sentence compression quality could degrade if compression goes too far. The appreciation of a compression quality has

been here demonstrated as a user satisfaction protocol. If COLIN has been able to shorten texts by an average 38%, humans were not able to remove more than 25%. At the same time, the satisfaction mean score is 3.23 over 5, whereas the same users attribute to human compressors a satisfaction mean score of 3.7, really not so much more.

References

- Chauché, Jacques. 1984. Un outil multidimensionnel de l'analyse du discours. In *Coling'84*, pages 11–15.
- Chomsky, Noam. 1970. Remarks on nominalization. In *R. Jacobs and P. Rosenbaum (eds.) Reading in English Transformational Grammar*, pages 184–221, Waltham: Ginn.
- Dorr, Bonnie, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *In R. Radev & S. Teufel (Eds.), Proceedings of the HLT-NAACL 2003 Workshop on Text Summarization*. *Omnipress*, pages 1–8.
- Gagnon, Michel and Lyne Da Sylva. 2005. Text summarization by sentence extraction and syntactic pruning. In *Computational Linguistics in the North East (CLiNE'05)*, Université du Québec en Outaouais, Gatineau, 26 August.
- Hovy, Eduard H., Chin-Yew Lin, and Liang Zhou. 2005. A be-based multi-document summarizer with sentence compression. In *the Multilingual Summarization Evaluation Workshop at the ACL 2005 conference*.
- Knight, Kevin and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence archive*, 139(1):91–107, July.
- Lin, Chin-Yew and Eduard H. Hovy. 2002. Automated multi-document summarization in neats. In *the DARPA Human Language Technology Conference*, pages 50–53.
- McKeown, K., D. Evans, A. Nenkova, R. Barzilay, V. Hatzivassiloglou, B. Schiffman, S. Blair-Goldensohn, J. Klavans, and S. Sigelman. 2002. The columbia multi-document summarizer for duc 2002.
- Yousfi-Monod, Mehdi and Violaine Prince. 2005. Automatic summarization based on sentence morpho-syntactic structure: narrative sentences compression. In *the 2nd International Workshop on Natural Language Understanding and Cognitive Science (NLUCS 2005)*, pages 161–167, Miami/USA, May.
- Yousfi-Monod, Mehdi. 2007. *Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus*. Ph.D. thesis, University of Montpellier II, Montpellier, November.