TALN 2011, Montpellier, 27 juin – 1er juillet 2011

A Collocation-Driven Approach to Text Summarization

Violeta Seretan
Institute for Language, Cognition and Computation
Human Communication Research Centre, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, United Kingdom
violeta.seretan@gmail.com

Résumé. Dans cet article, nous décrivons une nouvelle approche pour la création de résumés extractifs – tâche qui consiste à créer automatiquement un résumé pour un document en sélectionnant un sous-ensemble de ses phrases – qui exploite des informations collocationnelles spécifiques à un domaine, acquises préalablement à partir d'un corpus de développement. Un extracteur de collocations fondé sur l'analyse syntaxique est utilisé afin d'inférer un modèle de contenu qui est ensuite appliqué au document à résumer. Cette approche a été utilisée pour la création des versions simples pour les articles de Wikipedia en anglais, dans le cadre d'un projet visant la création automatique d'articles simplifiées, similaires aux articles recensées dans Simple English Wikipedia. Une évaluation du système développé reste encore à faire. Toutefois, les résultats préalables obtenus pour les articles sur des villes montrent le potentiel de cette approche guidée par collocations pour la sélection des phrases pertinentes.

Abstract. We present a novel approach to extractive summarization – the task of producing an abstract for an input document by selecting a subset of the original sentences – which relies on domain-specific collocation information automatically acquired from a development corpus. A syntax-based collocation extractor is used to infer a content template and then to match this template against the document to summarize. The approach has been applied to generate simplified versions of Wikipedia articles in English, as part of a larger project on automatically generating Simple English Wikipedia articles starting from their standard counterpart. An evaluation of the developed system has yet to be performed; nonetheless, the preliminary results obtained in summarizing Wikipedia articles on cities already indicated the potential of our collocation-driven method to select relevant sentences.

Mots-clés : résumé de texte automatique, résumé extractif, statistiques de co-occurrence, collocations, analyse syntaxique, Wikipedia.

Keywords: text summarization, extractive summarization, co-occurrence statistics, collocations, syntactic parsing, Wikipedia.

1 Introduction

Text summarization is a major NLP task, whose aim is "to present the main ideas in a document in less space" (Radev *et al.*, 2002). A specific type of summarization is the *extractive summarization*, which consists of selecting a subset of the original sentences of a document for verbatim inclusion in the summary; in contrast, the *abstractive summarization* consists of creating an abstract from scratch, by detecting the most important information in the document, appropriately encoding it, and, finally, rendering it using natural language generation techniques. Both extractive and abstractive approaches have been attempted over the past five decades since research on summarization started, with extractive approaches emerging as particularly suitable for large-scale applications. A survey on summarization methods can be found, for instance, in Das & Martins (2007).

The recent years have seen a growing interest in applying summarization to online content, and particularly to user-generated content, such as electronic mail messages, advertisements, and blogs, as well as to news articles. In addition, Wikipedia, the user-generated Internet encyclopedia,[1] is another online resource that is arguably very appealing from a summarization point of view. It currently contains over 3.6 million articles in English, and 18 million overall. Its comprehensive coverage makes it an important source of information used every day by a very broad audience. Since its content is continuously enriched, each article becomes more and more complex. Having an automatic means to produce article abstracts is therefore highly desirable for many readers.

The Simple English Wikipedia initiative[2] has recently been launched as an effort to improve the readability of Wikipedia articles and to provide shorter versions "presenting only the basic information". It resulted into the manual creation of shorter versions for almost 70'000 articles from the Ordinary English Wikipedia. Through the use of simpler words and simpler syntactic structures, these versions become accessible to a much broader audience, e.g., non-native speakers, children, and poor-literacy readers. The ultimate goal of our work is to create a text simplification system which can be used, in particular, to automatically generate simpler Wikipedia articles.

In this paper, we present the first steps we have taken in this direction, by implementing a domain-specific extractive summarization system. Given a collection of manually created summaries in a given domain – in our specific scenario, these are Simple English Wikipedia articles from a given Wikipedia category – in the development stage we infer a special kind of "template" representing the most important information that should be included in the summary. Then, given a new document – in our case, an Ordinary English Wikipedia article from the same category – we match the "template" against it to detect the relevant sentences to select for the output. The summaries created in this way will later be fed into the simplification system proper, which will transform them so that they obey predefined constraints on length, lexical choice, and syntactic structure. In what follows, we motivate our approach (§2), describe the system (§3), present experimental results (§4), and provide concluding remarks (§5).

2 Motivation

The underlying motivation of our approach comes from the observation that *collocations* – understood here as typical, syntactically related word associations, such as *capital city*, *city built*, *named after*, *people live*, *metropolitan area*, *median income* – may be used to represent the gist of the semantic content of a document. Collocation statistics have, in fact, been used for detecting the words that are most representative of a text (Kilgarriff, 1996), for performing text classification (Williams, 2002), and for topic segmentation (Ferret, 2002).

Taking into account collocations rather than isolated words results in a more accurate representation of the content, due to the tendency of words to exhibit only one sense in a given collocation (Yarowsky, 1993). In contrast, the statistics on isolated words suffer from collapsing together distinct meanings of polysemous words, and therefore may lead to less accurate representations. Take the example of the word *area*, which has several meanings: geographical region, subject of study, extent of a surface enclosed between boundaries, etc. Collocational information, such as *metropolitan area*, *area of specialization*, and *total area* helps distinguishing between each of these meanings, whereas, in the absence of such information, all these distinctions are lost.

Over the past years, much efforts have been devoted to devising statistical measures for grading the strength of association between two words in a corpus in order to automatically extract collocations (Church & Hanks, 1990;

[1] `http://www.wikipedia.org/`
[2] `http://en.wikipedia.org/wiki/Simple_english_wikipedia`

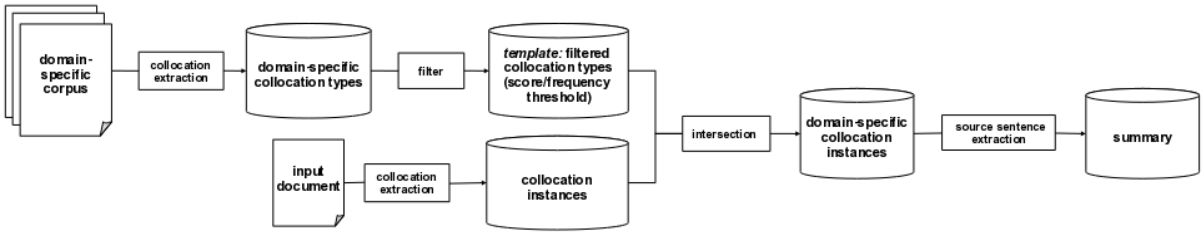

Figure 1: Architecture of the collocation-driven summarization system.

Dunning, 1993; Evert, 2004; Pecina, 2008). Relatively less investigated, however, was the problem of how to actually detect the collocation candidates on which these measures will be applied (Daille, 1994; Goldman *et al.*, 2001; Seretan, 2011). In practice, there are are basically two main approaches to this problem:

1. the *syntax-free approach*, which considers as a candidate any pair of (POS-filtered) words occurring at a short distance in the text. This light approach does not require advanced language tools (Ferret, 2002);
2. the *syntax-based approach*, which considers as a candidate only syntactically related word pairs, regardless of the distance between the component items. This approach requires syntactic parsing tools or, at least, partial/shallow analysis tools (Daille, 1994; Goldman *et al.*, 2001; Kilgarriff *et al.*, 2004; Evert, 2004; Pecina, 2008; Seretan, 2011).

As far as the use of collocations for semantic representation in concerned, it has been shown that the syntax-based approach is far more useful than the syntax-free approach in various natural language tasks, including synonymy detection and word sense disambiguation (Padó & Lapata, 2007). We adopt the syntax-based approach in our work, in which we exploit collocations for extractive text summarization. More precisely, we perform collocation extraction from a domain-specific development corpus and collocation detection in the input document, by relying on a syntax-based method we previously developed and described in Seretan (2011).

3 The Method

In this section, we describe our collocation-driven approach to extractive text summarization, as well as the methods and tools on which it relies. The system we present performs domain-specific summarization, rather than general-purpose summarization. Since the collocations extracted from a corpus are supposed to constitute the representative phrases of that corpus (§2), we apply our syntax-based collocation extraction method (Seretan, 2011) on a set of documents from a given Wikipedia category – like *city, actor* or *disease* – in order to detect which phrases are characteristic of that particular category. In principle, nothing in our approach prevents us from applying the same technique to open-domain corpora to perform general-purpose summarization; however, the question remains whether the collocations extracted from such corpora are useful for summarization at all, and this investigation is beyond the scope of our current work.

The main idea underlying our approach is that collocation statistics are suitable for detecting what a collection of articles is mainly about. In our particular scenario (§1), we consider Simple English Wikipedia articles on cities, and we want to infer what the typical content of such an article is. Once we have a collocation-based representation of the typical content, we detect the sentences from a new input article which realise that precise content. Finally, the summary is created by concatenating the resulting sentences in the order in which they appear in the original text, and, if necessary, removing duplicates. The system architecture is shown in Figure 1.

In the remainder of this section, we outline the collocation extraction step, which is at the core of our summarization method. Our collocation extractor (Seretan, 2011) has been built as an extension of Fips, a multilingual symbolic parser developed at the Language Technology Laboratory of the University of Geneva (Wehrli, 2007).[3]

This parser is based on generative grammar concepts and can be characterised as a strong lexicalist, bottom-up, left-to-right parser. For each input sentence, it builds a rich structural representation combining *a*) the constituent structure; *b*) the interpretation of constituents in terms of arguments (e.g., subject, objects); *c*) the interpretation of elements like clitics, relative and interrogative pronouns in terms of intra-sentential antecedents; and *d*) co-indexation chains linking extraposed elements (e.g., fronted NPs and *wh*-elements) to their canonical positions.

[3]Needless to say, any other parser and collocation extractor could be used; the method we present is not tailored to these specific tools.

According to the theoretical stipulations on which the parser relies, some constituents of a sentence may move from their canonical "deep" position to surface positions, due to grammatical transformations such as relativization or passivization (among many others). By accounting for such transformations and detecting long-distance dependencies, the parser can deal with situations in which the words in a collocation occur in a different order or are separated by additional material. Unlike syntax-free approaches, syntax-based approaches to collocation extraction do not impose a limit on the maximal distance at which the collocation components can be found in text, like the 5-word limit commonly used for English. This turns into a considerable advantage for those languages, like French or German, which exhibit a higher degree of word order freedom. The languages currently supported by Fips parser are: English, French, Italian, Spanish, German, and, partly, Greek and Romanian (other languages are under development).

We illustrate this discussion with a sentence from Simple English Wikipedia, shown in (1a) below, from which the collocation *to found – settlement* is detected by our extractor in spite of the fact that the component words are in the inverse order and they are separated by more than five words. The Fips parser retrieves the "deep" verb-object relation between *settlement* and *founded* from this passive construction with an embedded relative clause, as it creates a co-indexation chain (marked by *i*) between the empty trace of the canonical object, denoted by e_i, and the surface subject, the phrase DP_i headed by *settlement*. The (simplified) parser output is shown in (1b).

(1) a. The *settlement* that became the City of Louisville was *founded* in 1778 by George Rogers Clark.

b. [$_{TP}$[$\boldsymbol{DP_i}$ The [$_{NP_j}$ **settlement** [$_{CP}$[$_{DP}e_j$] that [$_{TP}$[$_{DP}$ e_j][$_{VP}$ became [$_{DP}$ the [$_{NP}$ City [$_{PP}$ of[$_{DP}$ Louisville]]]]]]]]$_j$]$_i$ was [$_{VP}$ founded [$_{DP}$ $\boldsymbol{e_i}$][$_{PP}$ in [$_{DP}$ 1778]][$_{PP}$ by [$_{DP}$ George [$_{DP}$Rogers [$_{DP}$ Clark]]]]]]

In our scenario, since the development corpus considered is a subpart of Simple English Wikipedia it is not supposed to contain many cases of complex syntactic structures. Nonetheless, relying on an extraction methodology capable to deal with the morphosyntactic flexibility of collocations is important for our summarization system at subsequent stages, when an input document is processed. As the documents to summarize typically originates from the Ordinary English Wikipedia, they may contain complex sentence structures that challenge the process of collocation identification which is at the core of our sentence extraction strategy.

The collocation candidates identified from the development corpus with the help of the syntactic parser are scored with the *log-likelihood ratio* association measure (Dunning, 1993).

4 Experimental Results

In our first experiments, we considered the Wikipedia category of *cities*[4] and the task of automatically summarizing articles on cities. The development set consisted of 1'071 randomly selected Simple English Wikipedia articles from this category.[5] The corpus contains about 150'000 words (more precisely, there were 152'645 tokens detected, including punctuation). The total number of sentences is 8'665; the average sentence length 17.6% tokens, and the average file length 8.1 sentences.

The processing of this corpus consisted of parsing and collocation extraction. The percentage of sentences for which a full parse tree could be built is 79.7%. The relatively high number of unknown words in the corpus (6.8%) – mostly, proper nouns that are not part of the parser's lexicon – as well as the presence of headings and image captions have a slight but visible impact on the performance of the parser, whose normal sentence coverage is beyond 80% on journalistic corpora in English. When a full parse tree cannot be built for a sentence, the collocation extractor attempts to detect candidates from the partial structures output by the parser, thus considering at least local dependencies even if the more distant ones will not be retrieved.

After the development stage, a database of 12'858 collocation types specific to this particular domain has been created, corresponding to the 20'581 collocation instances identified in the corpus. This database represents the "template" that will be used for sentence selection when creating a summary. Note that these collocation types are raw extraction results, not filtered with respect to score or frequency, and they correspond to a complete and detailed collocational "profile" of the domain considered. By applying score and frequency thresholds, we obtain a less detailed profile, and, accordingly, a summary of coarser granularity. Our system can therefore be adapted to different summarization needs and can achieve different compression rates. Moreover, it is possible to customize

[4] `http://en.wikipedia.org/wiki/Category:City`

[5] As a consequence, about a third of these articles are "stub" articles, "too short to provide encyclopedic coverage of a subject" (`http://en.wikipedia.org/wiki/Wikipedia:Stub`). The size of the development set will be varied in future experiments.

1. **It is the capital of the Languedoc-Roussillon region, as well as the Hérault department.**
2. **Montpellier is the 8th largest city of the country, and is also the fastest growing city in France over the past 25 years.[1]**
3. **Montpellier is one of the few large cities in France without a (Gallo-)Roman background.**
4. In the Early Middle Ages, the nearby episcopal town of Maguelone was the major settlement in the area, but raids by pirates encouraged settlement a little further inland.
5. Montpellier, first mentioned in a document of 985, was founded under a local feudal dynasty, the Guillem counts of Toulouse, who joined together two hamlets and built a castle and walls around the united settlement.
6. The city became a possession of the kings of Aragon in 1213 by the marriage of Peter II of Aragon with Marie of Montpellier, who brought the city as her dowry.
7. **In 1432, Jacques Cœur established himself in the city and it became an important economic centre, until 1481 when Marseille took over this role.**
8. At the time of the Reformation in the sixteenth century, many of the inhabitants of Montpellier became Protestants (or Huguenots as they were known in France) and the city became a stronghold of Protestant resistance to the Catholic French crown.
9. Louis XIV made Montpellier capital of Bas Languedoc, and the town started to embellish itself, by building the Promenade du Peyrou, the Esplanade and a large number of houses in the historic centre.
10. **After the French Revolution, the city became the capital of the much smaller Hérault.**
11. **Geography Montpellier seen from Spot satellite**
12. The city is situated on hilly ground 10 kilometres (6 mi) inland from the Mediterranean coast on the River Lez.
13. **The name of the city, which was originally Monspessulanus, is said to have stood for mont pelé (the naked hill, because the vegetation was poor), or le mont de la colline (the mount of the hill)**
14. **The city is built on two hills, Montpellier and Montpelliéret, thus some of its streets have great differences of altitude.**
15. Montpellier has a Mediterranean climate (Koppen Csa), with mild, somewhat wet winters, and very warm, rather dry summers.
16. **The whole metropolitan area had a population of 600,000 in 2006.**
17. **In 2008, the estimated population of the metropolitan area was 533,000.[citation needed]**
18. **In 2009, it was estimated that the population of the city of Montpellier had reached 265,000.[3]**
19. It was suppressed during the French Revolution but was re-established in 1896.
20. The school of law was founded by Placentinus, a doctor from Bologna university, who came to Montpellier in 1160, taught there during two different periods, and died there in 1192.
21. **The French Revolution did not interrupt the existence of the faculty of medicine.**
22. **The city is home to a variety of professional sports teams :**
23. **The city is a high-place for the cultural events since there is a lot of students.**
24. International relations See also : List of twin towns and sister cities in France Sign on the Esplanade Charles de Gaulle, showing Montpellier's sister cities Twin towns — Sister cities
25. **The capital of the American state of Vermont was named Montpelier because of the high regard held by the Americans for the French who aided their Revolutionary War against the British.**

FIG. 2 – Sample output the Wikipedia article on the city of Montpellier. Sentences in bold constitute a subsumed summary of coarser granularity.

the system by taking into account to type of information a user is interested in. Suppose that a user is mostly interested in economic aspects related to cities, such as the GDP, the median income, the distribution of income by gender, and the major economic players in a city. Customised summaries displaying such kind of information can easily be obtained by filtering the collocation template so that in includes only predefined words of interest.

Among the typical collocations in the city template we found: *be city, have population, people live, county seat, known as, be capital city*,[6] *large city, city population, close to, area of city, most important, city name, most famous, located on coast*, etc. Many of the collocations in the template contain emphasising words, like *oldest, important, major*. This result confirms the importance of using cue words in summarization, a strategy used since the early work of Edmundson (1969).

Figure 2 shows the summary created for the Wikipedia article on the city of Montpellier[7] when the score threshold is set to 10; the sentences in bold show a coarser-grained summary obtained when the threshold is set to 20. The compression rate is 14.3% (25/175) in the first case, and 9.1% (16/175) in the second case. This example illustrates the potential of our method to select relevant sentences for the summary.[8] A comparison against baseline summaries created by selecting the first sentence of each Wikipedia subsection will be provided in the near future.

5 Conclusion

We proposed a novel approach to summarization, motivated by the intuition that collocation statistics are able to capture the gist of the information content in documents from a given domain, and by the fact that syntactically related co-occurrences represent a better way to model lexical meaning than surface co-occurrences (Padó & Lapata, 2007). We relied on a syntax-based collocation extraction method to build a summarization system, in which a collocation-driven content template is inferred from a development corpus and then matched against an input document. A proper evaluation of the system has yet to be performed; nonetheless, our preliminary analysis

[6] Our collocation extractor detect complex collocations made up of more than two words when one of the two items is a complex lexical item present in the parser's lexicon (in this case, *capital city* is part of the lexicon).

[7] `http://en.wikipedia.org/wiki/Montpellier`. Accessed: April 7, 2011.

[8] Note that, as any extractive system, the results of our system suffer from the presence of "dangling anaphora" whose antecedents are lost (e.g., *It* from sentence 1, referring to Montpellier, and from sentence 19, referring to the University of Montpellier).

of abstracts generated for Wikipedia articles on cities suggest that collocational information is indeed useful for selecting relevant sentences for inclusion in a summary.

Until now, collocational information has seldom been taken into account in previous work on summarization. Aone *et al.* (1999) considered noun-noun collocations along with single words in their term frequency approach, in which sentences are selected depending on the number of frequent words they contain. Our work considered more varied and more flexible syntactic configurations, and made use of association measures instead of raw frequency. Barzilay & Elhadad (1997) also took into account noun sequences, and, in addition, lexical chains made up of semantically related words, retaining sentences with a high density of chain elements. Our work considered, in contrast, only syntactically related word combinations, thus eliminating the need for word sense disambiguation heuristics. Another difference is that our method allows to control the length and detail of the summary produced. More recently, Toutanova *et al.* (2007) considered information on word co-occurrences as a feature in learning a sentence scoring function, but without relying on a syntax-based approach as in our case.

Our future work will focus on applying the proposed method to other domains and evaluating its performance; building an online version of the extraction system; and using this system in our larger text simplification project.

Acknowledgement

This work has been supported by the Swiss National Science Foundation (grant no. PA00P1_131512). We thank Kristian Woodsend for sharing with us his Simple English Wikipedia dataset.

References

AONE C., OKUROWSKI M. E., GORLINSKY J. & LARSEN B. (1999). A trainable summarizer with knowledge acquired from robust NLP techniques. In I. MANI & M. T. MAYBURY, Eds., *Advances in Automatic Text Summarization*, p. 71–80.

BARZILAY R. & ELHADAD M. (1997). Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization Workshop (ISTS'97)*, p. 10–17, Madrid, Spain.

CHURCH K. & HANKS P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

DAILLE B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.

DAS D. & MARTINS A. F. T. (2007). A Survey on Automatic Text Summarization. `http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf`.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

EDMUNDSON H. P. (1969). New methods in automatic extracting. *Journal of the Association for Computing Machinery*, **16**, 264–285.

EVERT S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, University of Stuttgart.

FERRET O. (2002). Using collocations for topic segmentation and link detection. In *Proceedings of the 19th International Conference on Computational linguistics (COLING 2002)*, p. 260–266, Taipei, Taiwan.

GOLDMAN J.-P., NERIMA L. & WEHRLI E. (2001). Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, p. 61–66, Toulouse, France.

KILGARRIFF A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, p. 33–40, Sussex, UK.

KILGARRIFF A., RYCHLY P., SMRZ P. & TUGWELL D. (2004). The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, p. 105–116, Lorient, France.

PADÓ S. & LAPATA M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.

PECINA P. (2008). *Lexical Association Measures: Collocation Extraction*. PhD thesis, Charles University in Prague.

RADEV D. R., HOVY E. & MCKEOWN K. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, **28**(4), 399–408.

SERETAN V. (2011). *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Dordrecht: Springer.

TOUTANOVA K., BROCKETT C., GAMON M., JAGARLAMUDI J., SUZUKI H. & VANDERWENDE L. (2007). The PYTHY summarization system: Microsoft Research at DUC 2007. In *Proceedings of DUC 2007*, Rochester, USA.

WEHRLI E. (2007). Fips, a "deep" linguistic multilingual parser. In *ACL 2007 Workshop on Deep Linguistic Processing*, p. 120–127, Prague, Czech Republic.

WILLIAMS G. (2002). In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, **7**(1), 43–64.

YAROWSKY D. (1993). One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*, p. 266–271, Princeton.