Porting a Summarizer to the French Language

Rémi Bois[1] Johannes Leveling[2] Lorraine Goeuriot[2] Gareth Jones[2] Liadh Kelly[2]
(1) LINA, Université de Nantes, France *
(2) CNGL, DCU, Dublin, Ireland
remi.bois@etu.univ-nantes.fr, {lgoeuriot, lkelly, gjones, jleveling}@computing.dcu.ie ,

Résumé. Nous présentons dans cet article l'adaptation de l'outil de résumé automatique REZIME à la langue française. REZIME est un outil de résumé automatique mono-document destiné au domaine médical et s'appuyant sur des critères statistiques, syntaxiques et lexicaus pour extraire les phrases les plus pertinentes. Nous décrivons dans cet article le système REZIME tel qu'il a été conçu et les différentes étapes de son adaptation à la langue française. Les performances de l'outil adapté au français sont mesurées et comparées à celle de la version anglaise. Les résultats montrent que l'adaptation au français ne dégrade pas les performances de REZIME, qui donne des résultats équivalents dans les deux langues.

Abstract. We describe the porting of the English language REZIME text summarizer to the French language. REZIME is a single-document summarizer particularly focused on summarization of medical documents. Summaries are created by extracting key sentences from the original document. The sentence selection employs machine learning techniques, using statistical, syntactic and lexical features which are computed based on specialized language resources. The REZIME system was initially developed for English documents.In this paper we present the summarizer architecture, and describe the steps required to adapt it to the French language. The summarizer performance is evaluated for English and French datasets. Results show that the adaptation to French results in a system performance comparable to English.

Mots-clés : Résumé automatique, multilangue, domaine médical.

Keywords: single-document summarization, multilingual, medical domain.

1 Introduction

The rapid growth of online text resources is producing information overload, where individuals cannot make use of all the available information. This is particularly the case in specialised domains such as medicine and biomedicine, where finding relevant information is critical. As stated by Afantenos (Afantenos *et al.*, 2005), "the number of scientific journals in the fields of health and biomedicine is unmanageably large, even for a single speciality". This makes it very difficult for scientists to follow the evolution of their speciality. A similar situation exists for the general public seeking medical information from the internet and other electronic resources. Automatic summarization for the medical domain is a possible solution that can alleviate this problem.

In this paper we describe REZIME, an automatic summarizer designed to create efficient summaries of documents from the medical domain. REZIME generates single-document summaries, built from sentences containing key material extracted from documents. Sentence selection is based on machine learning techniques, using statistical, syntactic and lexical features computed with specialized language resources. The REZIME system was initially developed for English documents (Nguyen & Leveling, 2013). While the architecture of REZIME is language-independent, porting it to another language requires adaptation and/or translation of its linguistic resources (i.e. syntactic, lexical and terminological resources). In this paper we present the different steps required to adapt it to another language and focus on adapting the summarizer to the French language, and show its comparable effectiveness with the original English language system.

The rest of this paper is organized as follows : After a brief description of the main studies in single-document summarization in Section 2, we describe the REZIME architecture in Section 3. The main steps involved in its adaptation to French are presented in Section 4. The evaluation of the French summarizer and the comparison of its performances to the English one are shown in Section 5. We conclude with remarks on future work in Section 6.

∗. This study has been conducted while Rémi was an intern in CNGL.

2 Related work

Automatic summarization is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that" (Radev *et al.*, 2002). The two main approaches to text summarization are extractive summarization and abstractive summarization. Extractive summarization consists of selecting parts of the original text that contain the most important information which then form the summary. Extractive methods rely mostly on machine learning, using a set of features to rank sentences (Mani & Bloedorn, 1998). Abstractive summarization aims at rephrasing the content of the texts, generating sentences that do not necessarily appear in the original document (Barzilay & Mckeown, 2005).

Summarization for the medical domain poses unique challenges, as stated by Afantenos et al, "uniqueness of medical documents is due to their volume, their heterogeneity, as well as due to the fact that they are the most rewarding documents to analyse, especially those concerning human medical information due to the expected social benefits" (Afantenos *et al.*, 2005).

Some systems for medical summarization offer multilingual summaries, for example using English documents to create French summaries (Lenci *et al.*, 2002). Multilingual summarization has been explored in different works, however most of this work has concentrated on single documents and has been evaluated on news corpora (Dalianis *et al.*, 2004), (Litvak *et al.*, 2010). Most approaches for French summarization aim at providing multi-lingual or cross-lingual summaries (Torres-Moreno *et al.*, 2001), (Fernandez *et al.*, 2008), (Boudin & Torres-Moreno, 2009).

3 A single document summarizer for medical text

REZIME is a single-document summarizer based on sentence extraction that was initially designed for the English language, with specific features developed for effective summarization in the medical domain. Its workflow is typical of extractive summarization systems, such as the ScandSum system (Dalianis *et al.*, 2004).

The medical summaries generated by REZIME are presented to patients and medical professionals. Therefore, the readability and clarity of the summaries is critical. In the design of REZIME, we opted for an extractive approach, as non-extractive ones may lead to incoherent sentences, potentially giving false information to the user about the advice given in the full document.

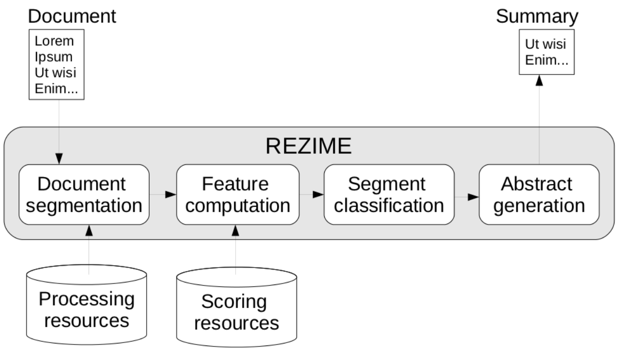

FIGURE 1: REZIME Workflow

For summarization, each document is processed in a workflow consisting of 4 steps, illustrated in Figure 1. First, the document is preprocessed, extracting its structure (paragraphs, sentences, tokens). Each sentence is then represented by a vector of features, that can either be statistical or linguistic. We used seven term checking features, that check for the presence or absence of particular significant words or phrases in sentences. We check for the presence of pre-defined basic words, that can help to create different summaries for a general audience or for professionals. We search for cue phrases such as "importantly" or "in summary" as they are good indicators of whether a sentence should be included in a summary. We also count the overlap of a sentence with the title terms as a feature. We use a naive Named Entity recognition as another feature. It checks for the number of capitalized words (the first word of the sentence is excluded). We chose this naive approach as the summarizer is designed to work online and process summaries on the fly, so speed and robustness are significant issues. We use preposition detection as they often give context to sentences. Finally, we use pronoun detection and count punctuation marks to discard sentences that contain too many of them making these hard to interpret independently. Most of these features are commonly used in summarization for newspaper articles (Lin, 1999).

We use five non term checking features. We incorporate a cluster keyword feature, a method proposed by Luhn to score sentences according to the number of significant words they contain (Luhn, 1958). We use the global bushy feature, which generates inter-document links based on similarity of paragraphs (Salton *et al.*, 1997). We also count the number of terms in each sentence and use it as a feature, assuming that too long or too short sentences are a liability. We use the position of a sentence in a paragraph as another feature, as sentences occurring early in a paragraph usually give context for all the paragraph. Finally, we use the TF-ISF, which is similar to TF-IDF, but on a sentence level. Each sentence is treated like a document.

For the medical domain, two domain-specific features are included : namely the affix presence feature, and the domain term feature. The former checks for terms containing medical-related affixes (896 affixes), while the later checks for terms appearing in a list of medical terms (5799 terms).

These features are then aggregated through a machine learning algorithm, in which the selection factors are combined in a weighting linear summation. Since development of the best machine learning algorithm is still in progress, we set all weights to 1 for this work. Finally, the selected sentences are post processed to provide a readable summary.

Some of these linguistic features rely on language dependent resources. Their adaptation to the French language, is described in the following section.

4 Porting REZIME to the French Language

In this section, we describe the resources used by REZIME and their adaptation to the French language. We assume that methods used in summarization are generally independent of the language used (e.g. keyword or phrase matching, keyword clustering or the global bushy algorithm). The major challenges of porting a summarizer to a new language relate mainly the availability of the necessary linguistic resources in this language. Development of these resources can generally be achieved by different methods : (1) using an existing resource in the target language ; (2) manually translating an existing resource ; (3) automatically translating an existing resource ; or (4) creating a new resource.

Option (3) is in most cases not viable, especially for domain-specific resources in languages other than English. Option 4) is costly, both in terms of time and manual labour. Thus, for each resource, we had to choose between automatic translation, or the identification of an existing alternative resource.

4.1 Resources used in the REZIME System

The resources used by the summarizer can be divided into two categories : the document processing resources and the scoring resources. Table 1 gives a list of the specific resources used and the category they belong to.

Category	Resource name	Use
Preprocessing resources	poss_sent_end	Indicates which tokens can end a sentence
	bad_sent_start	Indicates an impossible beginning for a sentence
	bad_sent_end	Indicates an impossible end for a sentence
	abbreviations	List of common abbreviations
Feature computation resources	sections	Potential title of sections in a paper
	cue_phrases	Keywords indicating that a sentence may be important
	medical_dict	Affixes indicating a medical term (eg patho-, -trophy)
	medical_terms	List of medical terms
	SpacheWordList	List of most common words
	stopwords	List of stopwords

TABLE 1: Resources for the Summarizer

The first category includes keywords used for sentence boundary detection. They indicate how and where to split a text into paragraphs, a paragraph into sentences, and a sentence into tokens. Most of these resources include generic vocabulary and/or punctuation marks. The *poss_sent_end* resource contains punctuation marks that can appear at the end of a sentence (e.g.. '.', ' ?'). The abbreviations resource is composed of frequent acronyms and abbreviations (e.g.. 'Mr.', 'Dr.'). Most of the resources from that category are language independent (e.g. punctuation marks), at least among European languages. The remaining resources can either be translated from one language to another or replaced with alternatives in the target language.

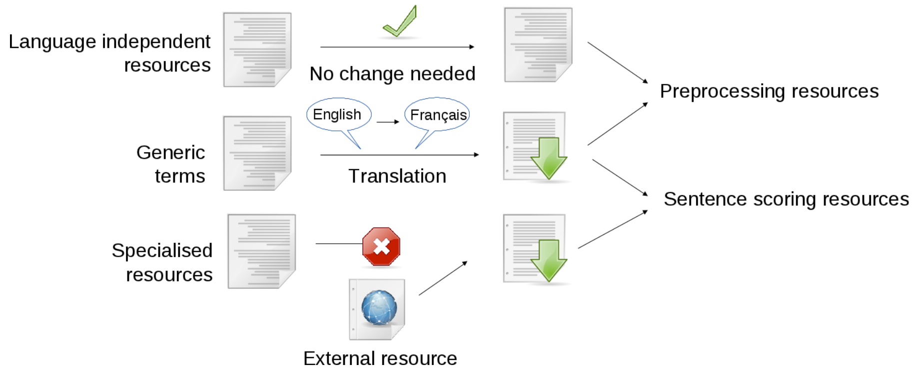

FIGURE 2: Strategies to port resources to the French language

The second category contains the resources used to represent sentences as vectors of features. These indicate elements from sentences that should be included in a summary. The *cue_phrases* resource is composed of such phrases (e.g.. 'in conclusion', 'the most important'). The *SpacheWordList* resource lists frequent and general terms, that should be preferred in a summary intended for non-specialists (e.g.. 'after', 'again'). Two scoring resources are specialized for the medical domain. Affixes for medical terms which allow recognition of most of the scientific terms as new terms have essentially been created by using Greek and Latin elements (Andrews, 1948). Nearly 900 of these affixes are gathered in our *medicat_dict* resource (e.g.. 'patho-', '-trophy'). The *medical_terms* list consists of a several thousands of medical terms, drug names and symptoms, collected from `www.medterms.com`.

These resources are mainly language dependent. Therefore they needed to be adapted to the French language. As described in Figure 2, this requires a terminological and syntactical adaptation (translation), or sometimes involves including new resources. We describe this process in the following sections.

4.2 Porting Resources to the French language

4.2.1 General Resources

This category includes the resources allowing REZIME to split the paragraphs into sentences and the sentences into words (detailed in Table 1. These resources are not specialized resources and need at most a simple translation. Some of these files did not need any modification as they contained only punctuation (*poss_sent_end*, *bad_sent_start* and *bad_sent_end*). The most significant work involved the general language abbreviations and acronyms (*abbreviations*) and which were manually translated by a native French speaker.

4.2.2 Adaptation of feature extraction resources

This category includes resources used to represent sentences as feature vectors, in order to choose which ones should be included in a summary, as shown in Figure 1.

These cannot be automatically translated with a high enough translation quality due to the presence of specialized vocabulary or specific linguistic features (medical terms, cue phrases). As shown in 2, resources could either be (1) kept as they are ; (2) translated ; or (3) replaced by a French resource.

1. The list of medical affixes *medical_dict* did not have to be translated. These are Latin and Greek affixes (e.g. cephal-, coron-, -phage), that are also widely used to create French medical terms.
2. We manually translated *sections* and *cue_phrases*. These resources were short lists and a manual translation was the most cost-effective solution (e.g. significant).
3. Some of the resources being very large collections, we could not translate them manually. As similar ones were available in French, we replaced *medical_terms*, *stopwords* and *SpacheWordList*. The medical_terms resource has been created thanks to a European project [1]. It contains 1840 medical terms, available in 8 languages, including French and English. The stopwords list is a short list of 126 terms. The SpacheWordList is composed of 1750 words from the general language [2].

1. `http://users.ugent.be/~rvdstich/eugloss/welcome.html`
2. `http://fr.wiktionary.org/wiki/Wiktionnaire:Liste_de_1750_mots_fran\%C3\%A7ais_les_plus_courants`

5 Evaluation and Comparison of the English and the French Summarizers

The goal of this experiment is to assess the effect of porting the summarizer to another language, by investigating the performance of the summarizer system adapted to French relative to the original English summarizer.

While an evaluation on a parallel test collection in two languages would have been suitable, the absence of such a resource led us to test our system on two independent collections, both issued from the medical domain, but on two different specialities (the English one is generic, the French one deals with endoscopy). Our assumption is that the topical focus will not affect the results significantly, since the resources come from the same domain. In this section, we describe the two corpora used for our experiments and the encouraging results we obtained.

5.1 Evaluation Test Collections

To evaluate the similarity of the two systems, we used comparable corpora from the medical domain. We created a corpus for the French language from scientific articles on a medical speciality journal : Acta Endoscopica. The English corpus was composed of articles from BioMed Central (BMC) [3]. Assuming a manually written abstract is a good quality summary of the scientific articles, they are used as gold reference for automatic summarization to evaluate our system (da Cunha & Wanner, 2005).

We limit the size of the summaries and obtain summaries of about 230 words, comparable in length to our manual abstracts. Our corpora are composed of about 100 documents. Table 2 provides some statistics on the corpora.

	French	English
Origin	Acta Endoscopica	BioMed central
Corpus size (documents)	104	100
Abstract length (words)	220	250
Summary length (words)	218	230

TABLE 2: Statistics about the French and English corpora and summaries.

5.2 Results

We report results obtained with the ROUGE measure (Lin, 2004). We use ROUGE-2 and ROUGE-3 with and without considering stopwords, as they are the most widely used ROUGE scores. We are interested in changes that porting to French causes to the behaviour of the REZIME system. Figure 3 shows the experimental results.

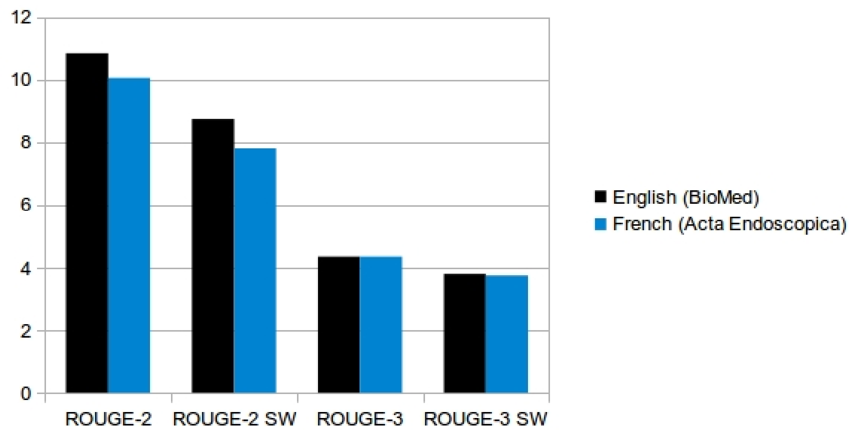

FIGURE 3: ROUGE scores for the English and the French summaries with REZIME

We observe that the ROUGE-3 scores are almost identical for the two systems. The ROUGE-2 score drops by one point at most, we believe this is mostly due to the differences between corpora, the French corpora belonging to a more specific

3. http://www.biomedcentral.com

medical area, with terms that do not appear in our list of medical terms. ROUGE-3 gives lower results than ROUGE-2, as would be expected, since it is based on 3-grams instead of bigrams.

This experiment shows that when porting the summarizer to another language similar summarization performance can be obtained for each language. The lack of parallel data and the consequent use of varied evaluation collections could affect the results, but additional experiments would be required to investigate this potential effect.

6 Conclusion and Future Work

In this paper, we described methods to adapt an extractive summarizer to another language. We presented results showing that single document summarization systems based on extraction can be successfully ported to an alternative language. Our adaptation technique is based on selecting and generating adequate resources by translation, which is an adequate approach even when the system deals with a specific context like the medical domain.

We also created a new summarization evaluation collection, consisting of French documents from the medical domain.

As part of future work we plan to investigate whether the technique which has been shown to be successful for French is also applicable to other pairs of languages, and to other domains.

Acknowledgments

This work is supported in part by the Khresmoi project (257528) and SFI (07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.

Références

AFANTENOS S., KARKALETSIS V. & STAMATOPOULOS P. (2005). Summarization from medical documents : a survey. *Artificial intelligence in medicine*, **33**(2), 157–177.

ANDREWS E. (1948). A history of scientific English. *The American Journal of the Medical Sciences*, **215**(1), 120.

BARZILAY R. & MCKEOWN K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, **31**, 297–328.

BOUDIN F. & TORRES-MORENO J.-M. (2009). Résumé automatique multi-document et indépendance de la langue : une première évaluation en français. In *TALN 2009 – Session posters*.

DA CUNHA I. & WANNER L. (2005). Towards the automatic summarization of medical articles in Spanish : Intergration of textual, lexical, discursive and syntactic criteria. *Crossing Barriers in Text Summarization Research, RANLP, Borovets*.

DALIANIS H., HASSEL M., STOCKHOLM K., SMEDT K. D., LISETH A., COGNIT T. C. L., WEDEKIND J. & COPENHAGEN C. (2004). Porting and evaluation of automatic summarization. In *Nordisk Sprogteknologi*, p. 107–121.

FERNANDEZ S., SANJUAN E. & TORRES-MORENO J. M. (2008). Enertex : un système basé sur l'énergie textuelle. In *TALN 2008*, p. 99–108.

LENCI A., BARTOLINI R., CALZOLARI N., AGUA A., BUSEMANN S., CARTIER E., CHEVREAU K. & COCH J. (2002). Multilingual summarization by integating linguistic resources in the MLIS-MUSI project. In *LREC*, volume 2, p. 1464–1471.

LIN C.-Y. (1999). Training a selection function for extraction. In *CIKM 1999*, p. 55–62 : ACM.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*, p. 74–81.

LITVAK M., LAST M. & FRIEDMAN M. (2010). A new approach to improving multilingual summarization using a genetic algorithm. In *ACL 2010*, p. 927–936.

LUHN H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, **2**(2), 159–165.

MANI I. & BLOEDORN E. (1998). Machine learning of generic and user-focused summarization. In *AAAI/IAAI*, p. 821–826.

NGUYEN D. & LEVELING J. (2013). Exploring domain-sensitive features for extractive summarization in the medical domain. In *Natural Language Processing and Information Systems*, volume 7934 of *LNCS*, p. 90–101.

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

SALTON G., SINGHAL A., MITRA M. & BUCKLEY C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, **33**(2), 193–207.

TORRES-MORENO J.-M., VELÁZQUEZ-MORALES P. & MEUNIER J.-G. (2001). Cortex : un algorithme pour la condensation automatique de textes. *ARCo 2001*, **2**.