Multilingual Multi-Document Summarization: content selection and redundancy treatment based on lexical-conceptual knowledge *

Yasmin Vizeu Camargo[1,2]

[1] Interinstitutional Center for Computational Linguistics (NILC), São Carlos, Brazil
[2] Graduate Program in Linguistics, Federal University of São Carlos, Brazil
yvizeu@gmail.com

Abstract. On Multilingual Multi-Document Summarization (MMS), one of the aims is to construct a summary in the user's language, from a collection of at least two news texts that address the same subject, one of them being in the user's language and another in a foreign language. These summaries, when extractives, are composed by the sentences of the source-texts that convey the main information of the collection. The selection of sentences can be based on different strategies. A very promising strategy is the one that uses the frequency of occurrence of the lexical concepts in the collection. Therefore, the aim of this project is to propose content selection methods that consider relations (such as hyponymy and hyperonymy, for example) between the concepts of the collection. Furthermore, another aim is to treat the redundancy between the selected sentences based on conceptual (not lexical) overlap. Thus, we expect to improve the informativeness and gramaticality of multilingual extracts.

Keywords: Multilingual Multi-Document Summarization · Lexical-Conceptual Knowledge · Content Selection

1 Introduction

Multilingual Multi-Document Summarization (MMS) has been gaining attention recently. In this Natural Language Processing (NLP) application, one of the goals is to identify the central information of a collection of at least two journalistic texts that address the same subject — one in the user's language and the other in a foreign language — and to present it in the form of a summary in the user's language. MMS is a challenging task, since it involves the classic issues of Automatic Multi-Document Summarization (AMS) (cohesion, coherence, redundancy, etc.) and also the multiplicity of languages.

To summarize multilingual collections, the earlier works in the literature have adapted extractive methods originally developed for monolingual AMS. In extractive summarization, the source-sentences are ranked according to a relevance

* Supported by Coordination for the Improvement of Higher Education Personnel (CAPES).

criterion or parameter and the most well-ranked are concatenated for the summary production (MANI, 2001) [1]. As such parameters are superficial, that is, based on simple linguistic knowledge (e.g., lexical statistics), or hybrids, that is, based on the combination of superficial and deep knowledge (e.g., lexical semantics), the methods in AMS can only be applied to collections in a monolingual collections. Thus, MMS comprises an initial phase consisting of machine translation (MT) of the texts in the foreign language into the user's language, producing a monolingual multi-document collection (EVANS et al., 2004 [2]; ROARK, FISHER, 2005 [3]; EVANS et al., 2005 [4]; TOSTA et al., 2013 [5]). About the collection, Roark and Fisher[3] extract sentences from the machine-translated and the original texts in the user's language. Consequently, the extracts present ungrammatical sentences and disfluencies resulting from MT. Evans et al. [2] [4] only extract sentences from the translated texts, and replace them with similar ones from the text in the user's language. Thus, this method avoids the MT problems, but the content selection does not take into account the information from the text in the user's language.

More recently, deep MMS methods have been developed, which process the collection of source-texts without the complete machine translation of the texts in the foreign language (TOSTA, 2014 [6]; DI-FELIPPO et al., 2016 [7]). These methods select the sentences based on the frequency of the nominal concepts of the collection, avoiding the redundancy between them through the word overlap measure. The extracts generated by these methods are more informative and have less grammatical problems than those generated by baseline methods for Portuguese, which include the MT of the texts in foreign language and the selection of the sentences based on word frequency or sentence position in the source-texts (TOSTA et al., 2013 [5]). In such scenario, the goal of this project is to refine the application of lexical-conceptual knowledge in MMS to produce extracts in Portuguese.

2 Goals and Hypothesis

Considering the promising results of the investigations on the application of lexical-conceptual knowledge in MMS, the goal is to propose methods of content selection that include: (i) different weights (relevance scores) for the concepts linked by super-subordinate relation (also called hyperonym, hyponym or ISA) (i.e., relation that links more general concepts like "vehicle" to specific ones like "car") and (ii) identification of redundancy based on concept overlap. The goal (i) was specified based on the hypothesis that superordinate concepts (or hyperonyms) helps to identify generic content in the collection, which is important to generate informative and generic extracts, as can be seen in the examples below, extracted from the corpus CM2News, wich will be used in this research. The excerpts are from news about a dengue epidemic that affected the city of Campo Grande in 2013.

(1) Heavy rains have increased the risk of reproduction of the mosquito which

transmits the **disease**.

(2) The federal authorities said that nationwide two people had died from **dengue** fever.

(3) Uma mulher de 45 anos morreu em decorrência da **doença** na segunda-feira (7).

(4) A cidade de Campo Grande (MS) vive uma epidemia de **dengue** : a prefeitura diz já ter notificado 3.456 casos nos dez primeiros dias do ano, mais que o triplo de casos registrados em 2012 (1.079).

Considering the relation between hyperonymy and hyponymy (verified based on WordNet.Pr) between the concepts "disease" and "dengue", in English, and "doença" and "dengue" (in Portuguese), the sentences above would be scored with values greater than just scored based on their frequency in the text — as done in Tosta [6] — and, thus, would have their content selected to compose the extract, corroborating the hypothesis that a higher score for hyperonymic concepts generates a selection of more general content and, therefore, more interesting in terms of informativeness for Multilingual Multi-Document Summarization.

The goal (ii) was traced under the hypothesis that the similarity between sentences is more properly calculated based on the occurrence of distinct expressions of a same concept (synonymy), which is not possible through the word overlap measure.

3 Methodology

The methodology of this work is composed of five main stages: (i) selection and/or construction of the corpus, (ii) annotation of the corpus, (iii) score and ranking of source-sentences, (iv) content selection and redundancy treatment and (v) construction and evaluation of the extracts.

For this project, we will use the CM2News corpus (TOSTA, 2014 [6]), which has 20 bilingual collections (Portuguese-English), accounting for 40 documents and 24,724 words. The collections or clusters cover different domains, such as politics, health, science, entertainment, etc. Since CM2News is a small corpus, there is the possibility of extending the resource with the inclusion of new bilingual collections. In this case, the texts will be compiled from the web based on the guidelines established by Tosta (2014) [6].

The second step consists in annotating the hyperonymy relations in the corpus. If the corpus is extended, the nominal concepts of the new collections should also be tagged. In this case, we will use the editor called MulSen (Multilingual Sense Estimator) [8], which allows (i) the identification of nouns by using a morphosyntactic tagger, (ii) the translation of the Portuguese nouns into English, which is necessary considering that Princeton WordNet [9] is our conceptual

repository, and (iii) the annotation of nouns with their correspondent synsets from WordNet.

In the third step, we will score the source-sentences based on their concepts (synsets) and produce a rank. The scoring yields a ranking in which the sentences with the most frequent and generic concepts of the collection are in the top positions. The fourth step consists in treating redundancy by the application of a concept overlap measure to be specified.

Finally, the fifth stage consists in producing and evaluating the extracts. To generate the extracts, we will simply juxtapose the sentences selected from the rank, ordering them according to their position in their corresponding source texts. Regarding informativeness, we will apply the traditional automatic ROUGE measure [10], which is mandatory in the area. To analyze the linguistic quality of the extracts, we will use the 5 criteria of Document Understanding Conference (DUC) [11].

References

1. Mani, I.: Automatic summarization. John Benjamins Publishing, Amsterdam (2001).
2. Evans, D. K., Klavans, J. L.; McKeown, K. R.: Columbia NewsBlaster: multilingual news summarization on the web. North American Chapter of The ACL: Human Language Technologies, 1–04 (2004).
3. Roark, B.; Fisher, S.: OGI OHSU baseline multilingual multi-document summarization system. Multilingual Summarization Evaluation (MSE), (2005).
4. Evans, D. K. et al.: Similarity-based multilingual multi-document summarization. Technical Report CUCS-014-05, (2005).
5. Tosta, F. E. S. et al. Estudo de métodos clássicos de sumarização automática no cenário multidocumento multilíngue. In: Workshop de IC em Tecnologia da Informação e da Linguagem Humana, 34–36. Fortaleza, Brasil (2013).
6. Tosta. F. E. S.: Aplicação de conhecimento léxico-conceitural na sumarização multidocumento multilíngue. Dissertação de Mestrado, Universidade Federal de São Carlos, São Carlos (2014).
7. Di-Felippo, A. et al.: Applying lexical-conceptual knowledge for multilingual multi-Document summarization. In: PROPOR, 2016, vol. 9727, pp. 38–49. Springer, Tomar (2016).
8. Multilingual Sense Estimator, https://www.icmc.usp.br/pessoas/taspardo/sucinto/resources.html. Last accessed 11 May 2018
9. Fellbaum, C.: Wordnet: an electronic lexical database. Computational Linguistics, (1998).
10. Lin, C-Y.: Automatic evaluation of summaries using N-gram coocurence statistics. In: Language Technology Conference. Edomonton, Canada (2003).
11. Dang, H. T.: Overview of DUC 2005. In: Document Understanding Conference. (2005).