A New Annotated Portuguese/Spanish Corpus for the Multi-Sentence Compression Task

Elvys Linhares Pontes[1], **Juan-Manuel Torres-Moreno**[1,2], **Stéphane Huet**[1], **Andréa Carneiro Linhares**[3]
[1]CERI/LIA, Université d'Avignon et des Pays de Vaucluse, Avignon, France
[2]École Polytechnique de Montréal, Montréal, Canada
[3]Universidade Federal do Ceará, Sobral-CE, Brasil
{elvys.linhares-pontes, juan-manuel.torres, stephane.huet}@univ-avignon.fr
andrea.linhares@ufc.br

Abstract

Multi-sentence compression aims to generate a short and informative compression from several source sentences that deal with the same topic. In this work, we present a new corpus for the Multi-Sentence Compression (MSC) task in Portuguese and Spanish. We also provide on this corpus a comparison of two state-of-the-art MSC systems.

Keywords: Annotated Corpus, Multi-Sentence Compression, Multilingual Corpus.

1. Introduction

Among the various applications of Natural Language Processing, Automatic Text Summarization (ATS) aims at summarizing one or more texts automatically. Summarization systems identify relevant data and create a summary from key information. The (Multi-)Sentence Compression task can be seen as a subproblem of ATS with the objective to generate a shorter, informative and correct sentence from source sentence(s).

In many cases, state-of-the-art NLP systems are evaluated with experiments restrained to the English language, in part because there are a lot of available English resources for most NLP tasks. As regards Multi-Sentence Compression (MSC), the available resources are unfortunately limited; to our knowledge, only one dataset is freely available and it is confined to the French language (Boudin and Morin, 2013).

In this work, we present a new annotated corpus in the Portuguese and Spanish languages for the MSC task. Using this corpus, we evaluate two state-of-the-art systems and show that the use of several languages leads to more mitigated results on the superiority of one system than the use of the French corpus alone.

The remainder of this paper is organized as follows. In Section 2, we characterize MSC with respect to related tasks from the perspective of the available corpora. Section 3 describes the creation and the features of our corpus. In Section 4 we analyze the results achieved by state-of-the-art methods using our dataset. Finally, conclusions are set out in Section 5.

2. Related Work

Sentence Compression (SC) aims at producing a reduced grammatically correct sentence from a source sentence. SC can be used in the context of the abstractive summarization of documents, the generation of article titles or the simplification of complex sentences, using diverse methods (optimization, syntactic structure, deletion of words and/or generation of sentences). The corpora for SC can be divided in two categories: deletion-based and summarization-based SC.

In the case of SC by deletion of words, sentences are compressed by removing irrelevant words (Filippova et al., 2015; Ive and Yvon, 2016). Knight and Marcu (2002) developed a SC corpus by aligning abstracts and sentences extracted from the Ziff-Davis corpus, which is a collection of newspaper articles announcing computer products. Clarke and Lapata (2008) provided two manually created two-reference corpora for deletion-based compression. Filippova and Altun (2013), and Filippova et al. (2015) extracted and released deletion-based compressions by aligning news headlines to the first sentences. Finally, Ive and Yvon (2016) developed an English-French parallel corpus for the compression and simplification tasks.

SC by generations of sentences analyzes a whole sentence and generates a new shorter sentence with the core information of the source sentence (Rush et al., 2015; Ganitkevitch et al., 2011; Cohn and Lapata, 2008; Toutanova et al., 2016). Ganitkevitch et al. (2011) created a corpus of compression paraphrases composed of parallel English-English sentences obtained from multiple reference translations. Rush et al. (2015) produced compression pairs made up of the headline of each article and its first sentence; they released their code to extract data from the annotated Gigaword (Graff et al., 2011). Cohn and Lapata (2008) and Toutanova et al. (2016) describe two manually created abstractive compression corpora that are publicly available. The dataset presented in Cohn and Lapata (2008) comprises a single-reference sentence pairs for abstractive summary, while the corpus developed by Toutanova et al. (2016) has multiple references for short paragraph compressions.

Multi-Sentence Compression (MSC), also known as Multi-Sentence Fusion, is a variation of SC. MSC aims at analyzing a cluster of similar sentences to generate a new sentence, which is shorter than the average length of source sentences and has the key information of the cluster (Barzilay and McKeown, 2005; Filippova, 2010). MSC enables summarization and question-answering systems to gener-

Characteristics	French		Portuguese		Spanish	
	Source	Reference	Source	Reference	Source	Reference
#tokens	20,224	2,362	17,998	1,425	30,588	3,694
#vocabulary (tokens)	2,867	636	2,438	533	4,390	881
#sentences	618	120	544	80	800	160
avg. sentence length (tokens)	33.0	19.7	33.1	17.8	38.2	23.1
type-token ratio	38.8%	50.1%	33.7%	67.9%	35.2%	43.4%
sentence similarity [0,1]	0.46	0.67	0.51	0.59	0.47	0.64

Table 1: Statistics of the corpora.

ate outputs combining fully formed sentences from one or several documents. Various corpora have been developed for MSC and are composed of clusters of similar sentences from different source news in English, French, Spanish or Vietnamese languages (Barzilay and McKeown, 2005; Filippova, 2010; Boudin and Morin, 2013; Thadani and McKeown, 2013; Luong et al., 2015). Filippova's corpus as well as Boudin and Morin's contain clusters of similar sentences, each cluster composed of at least 7 or 8 sentences, whereas the datasets introduced in (McKeown et al., 2010) and (Luong et al., 2015) have only a pair of source sentences per cluster. McKeown et al. (2010) collected 300 English sentence pairs taken from newswire clusters using Amazon's Mechanical Turk. Likewise, the dataset built by Luong et al. (2015) contains 250 Vietnamese sentences divided into 115 groups of similar sentences with 2 sentences per group. Thadani and McKeown (2013) presented an English corpus with 1,858 clusters having between 2 and 4 sentences; this dataset was built using automatic methods from annotations made for the DUC[1] and TAC[2] evaluations. The corpora presented in (McKeown et al., 2010), (Boudin and Morin, 2013) and (Luong et al., 2015) are publicly available, but among these three datasets only the second one is more suited to multi-document summarization or question-answering tasks because the documents to analyze are usually composed of many similar sentences.

3. Dataset Description

We introduce a novel annotated corpus collected from Portuguese and Spanish Google News.[3] This corpus is composed of clusters of similar sentences along with reference compressions for each cluster. The data are described in the following subsections. Table 1 summarizes the characteristics of the corpus and Table 2 shows a small example of our Portuguese dataset.

3.1. Source Sentences

In keeping with the methodology introduced by Filippova (2010), we collected links from Google News in Spanish and Portuguese between July and September 2016. These links redirect international news sites in Spanish (*La Jornada, Milenio, El Economista, BBC Mundo, El Colombiano, El País, CNN en español*, etc.) and in Portuguese (*G1, Uol Notícias, Estadão, O Globo, etc.*). Each cluster is composed of related sentences describing a specific event and was chosen among the first sentence from different articles about Science, Sports, Economy, Health, Business, Technology, Accidents/Catastrophes, General Information and other subjects. During the collection period, sentences were gathered among news threads that had at least 8 different sources. The source sentences of each cluster were manually selected so that they best describe the news, while sentences dealing with less relevant information were discarded. Each source sentence is composed of at least 8 tokens and a verb. In order to ensure the variability of source sentences inside a cluster, we removed all duplicated sentences, by assuming that sentences were too similar when the cosine similarity[4] computed from one-hot vectors was higher that 0.8. We used the TreeTagger system[5] to tag the source sentences with Parts-of-Speech.

3.2. Reference Compressions

Like in (Filippova, 2010; Boudin and Morin, 2013), reference compressions are edited by human annotators, all native Portuguese or Spanish speakers, who analyzed the most relevant facts of a cluster and generated a condensed sentence of this cluster. We suggested that the annotators should use the same vocabulary and n-grams as the source sentences and only select the most relevant information about the topic. We also recommended that they should generate compressions that are shorter than the length average of the source sentences. The following sections provide details about the Portuguese and Spanish parts of the corpus and, as a matter of comparison, briefly recalls the main characteristics of the French corpus built by Boudin and Morin.

3.2.1. Portuguese Dataset

The Portuguese corpus is composed of 40 clusters. Each cluster has at least 10 similar sentences by topic and 2 reference compressions made by 2 human annotators. This corpus contains 17,998 tokens and has a vocabulary of 2,438 tokens. Source sentences have an average of 33.1 tokens per sentence with a standard deviation of 9.9 tokens. The Type-Token Ratio (TTR) indicates the reuse of tokens in the cluster and is defined by the number of unique tokens divided by the number of tokens in each cluster; the lower

[1] http://duc.nist.gov

[2] http://www.nist.gov/tac

[3] The Spanish and Portuguese MSC datasets are freely available, under GPL license on the DOI website: http://dev.termwatch.es/~fresa/CORPUS/MSF2/.

[4] The cosine similarity between two vectors u and v associated with two sentences is defined by $\frac{u \cdot v}{||u|| \, ||v||}$ in the [0,1] range.

[5] Website: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Source sentences :
A Tesla fez uma oferta de compra à empresa de serviços de energia solar SolarCity por mais de 2300 milhões de euros.
A Tesla Motors , fabricante de carros elétricos , anunciou aquisição da SolarCity por US$ 2,6 bilhões .
A fabricante de carros elétricos e baterias Tesla Motors disse nesta segunda-feira (1) que chegou a um acordo com a SolarCity para comprar a instaladora de painéis solares por US$ 2,6 bilhões , em um grande passo do bilionário Elon Musk para oferecer aos consumidores um negócio totalmente especializado em energia limpa , informou a Reuters .
Reference compressions :
A Tesla Motors anunciou acordo para comprar a SolarCity por US$ 2,6 bilhões .
A fabricante Tesla Motors vai adquirir a instaladora de painéis solares da SolarCity .

Table 2: Small example of our Portuguese dataset.

the TTR, the greater the reuse of tokens in the cluster. The sentence similarity represents the average cosine similarity of the sentences in a cluster. Using these metrics, references have an average length of 17.8 tokens and a standard deviation of 1.5 tokens, while the Portuguese source corpus has a TTR of 33.7%. The Portuguese annotators generated the compressions with a TTR of 67.9% and a sentence similarity of 0.59. Finally, the average compression ratio between the reference and source sentences is 54%.

3.2.2. Spanish Dataset

The Spanish part is also composed of 40 clusters. It has 30,588 tokens and a vocabulary of 4,390 tokens. Each cluster has 20 similar sentences on the same topic and 4 reference compressions made by 4 human annotators. Source sentences have an average of 38.2 tokens per sentence with a standard deviation of 10.7 tokens and an average TTR of 35.2%. Reference compressions contain the same vocabulary as source sentences while keeping an average size of 23.1 tokens, a standard deviation of 2.4 tokens and a TTR of 43.4%. The sentence similarity between the compressions is 0.64. The average compression rate is 61%.

3.2.3. French Dataset

We used in the following experiments the French corpus developed by Boudin and Morin (2013). This corpus also has 40 clusters composed of 618 sentences (33 tokens on average). The clusters are composed of 15 sentences on average and the TTR of the corpus is 38.8%. Reference compressions have a compression rate of 60%.

4. Experimental Evaluation

We used our corpus to provide a more thorough evaluation of state-of-the-art approaches for MSC than the study on the French corpus alone. We tested on our dataset a simple baseline, as well as (Filippova, 2010) and (Boudin and Morin, 2013) methods. Filippova modeled clusters of similar sentences as Word Graphs based on the cohesion of tokens and their Part-of-Speech (PoS). Inspired by the good results of the Filippova's method, Boudin and Morin used the TextRank method as a re-rank method to analyze the sentences generated by Filippova's method in order to produce well punctuated and hopefully more informative compressions. The baseline system creates a Word Graph (WG) like Filippova's method, but this time all arcs have the same weight. Then, the system generates a compression represented by the shortest path in the WG that has at least 8 tokens. Algorithms were implemented using the Python programming language and the takahe[6] library.

4.1. Automatic and Manual Metrics

The most important features of MSC are informativeness and grammaticality. Informativeness is the percentage of the main information retained in the compression, while grammaticality analyzes whether a sentence is correct or not.

References are assumed to contain the most important information. Thus we calculated informativeness scores based on the common information between the output of the MSC system and the references using ROUGE (Lin, 2004). In particular, we used the f-measure metrics ROUGE-1, ROUGE-2 and ROUGE-SU4. Like in Boudin and Morin (Boudin and Morin, 2013), ROUGE metrics are calculated using stop words removal and stemming.[7]

We also led a manual evaluation with 4 native speakers for each language. The native speakers of each language judged the compression in two aspects: informativeness and grammaticality. In the same way as (Filippova, 2010; Boudin and Morin, 2013), the native speakers evaluated the grammaticality in a 3-point scale: 0 point for an ungrammatical compression, 1 point for compression with minor mistakes; and 2 points for a correct compression. The informativeness evaluation process is similar for grammaticality: 0 point if the compression is not related to the main topic, 1 point if the compression misses some relevant information and 2 points if the compression conveys the gist of the main event.

4.2. Results with Automatic Metrics

Table 3 shows f-score ROUGE scores for the French, Portuguese and Spanish datasets.[8] Boudin and Morin's system generated better compressions with higher ROUGE scores than Filippova's and the baseline for all datasets.

[6]Website: `http://www.florianboudin.org/publications.html`

[7]http://snowball.tartarus.org/

[8]Although we used the same system and data as (Boudin and Morin, 2013) for the French corpus, we were not able to reproduce exactly their results. The ROUGE scores given in their article are close to ours for their system: 0.6568 (ROUGE-1), 0.4414 (ROUGE-2) and 0.4344 (ROUGE-SU4), but using Filippova's system we measured higher scores than them: 0.5744 (ROUGE-1), 0.3921 (ROUGE-2) and 0.3700 (ROUGE-SU4).

Method	French			Portuguese			Spanish		
	RG-1	RG-2	RG-SU4	RG-1	RG-2	RG-SU4	RG-1	RG-2	RG-SU4
Baseline	0.3681	0.1904	0.1758	0.3199	0.1273	0.1309	0.2700	0.0990	0.0984
Filippova (2010)	0.6384	0.4423	0.4297	0.5388	0.2971	**0.2938**	0.5004	0.2983	**0.2847**
Boudin and Morin (2013)	**0.6674**	**0.4672**	**0.4602**	**0.5532**	**0.3029**	0.2868	**0.5140**	**0.2960**	0.2801

Table 3: ROUGE f-scores measured on the French, Portuguese and Spanish datasets. The best ROUGE results are in bold.

Table 4 provides statistics on the length and the compression ratio of the sentences generated by the systems. The baseline system output the shortest compressions, which translated into the worst ROUGE scores. For the three tested datasets, Filippova's method generated shorter compressions with a smaller standard deviation than Boudin and Morin's system. Let us note that for this last system the lengths of the outputs are less regular across the three languages.

The Portuguese and Spanish languages derive from Latin and are closely related languages. However, they differ in many details of their grammar and lexicon. Moreover, the datasets produced for the three languages are unlike according to several features. First, our corpus contains a smaller (Portuguese corpus) and a larger (Spanish corpus) dataset in terms of sentences than the original French corpus. Besides, the compression rates of the three datasets (see Section 3.) indicates that the Portuguese source sentences have more irrelevant tokens. The sentence similarity (Table 1, last line) describes the variability of sentences in the source sentences and in the references, and reflects here that the sentences are slightly more diverse for the Portuguese corpus. It can be noticed that the references are more similar too each other than source sentences since they only retain the main information. Finally, the French corpus has a TTR of 38.8% whereas the Portuguese and Spanish datasets have TTRs of 33.7% and 35.2%, respectively.

The baseline system generated the shortest compression because all arcs of the WG have the same weights. However, this system analyzes neither the grammaticality nor the most used n-grams in the clusters. Consequently, the baseline system generated compressions with the worst ROUGE scores.

4.3. Human Evaluation

ROUGE only analyzes the overlapping between the candidate compression and the references. Since this analysis is not reliable enough, we led a further manual evaluation to study the informativeness and grammaticality of compressions, as described in Section 4.1.. Given the poor results of the baseline with ROUGE, we only analyzed the Filippova's and Boudin and Morin's methods (Table 5).

We measured inter-rater agreement on the judgments we collected, obtaining values of Fleiss' kappa of 0.418, of 0.305 and 0.364 for French, Portuguese and Spanish respectively. These results show that human evaluation is rather subjective. Questioning evaluators on how they proceed to rate sentences reveals that they often made their choice by comparing outputs for a given cluster. As the differences of the grammaticality and the informativeness scores for the methods are not statistically significant, we move our investigation on the average and standard deviation of the results. Both methods generated compressions of good quality (scores higher than 1) for all datasets, especially for the French and the Portuguese parts where scores above 1.5 for grammaticality and above 1.2 for informativeness were obtained. Filippova's method generated more correct compressions (except for the Portuguese corpus where both methods obtained almost the same results), which shows that the re-ranking step tends to moderately deteriorate grammaticality. By contrast, Boudin and Morin's method consistently improves informativeness, which validates the interest of integrating the analysis of key phrases inside candidate compressions. This reranking method combines the cohesion score of Filippova and the relevance of key phrases[9] to generate more informative compression. This method selects the path of Word Graph that has relevant key phrases even if this path has a lower cohesion quality.

All in all, Boudin and Morin's method generated more informative but also longer compressions than Filippova's, CR showing a relative increase of 18% between both systems (Table 4).

5. Conclusion and Future Work

Multi-Sentence Compression aims to generate a short informative text summary from several sentences with related and redundant information. This task can be used in the domain of multi-document summarization or question answering to provide more informative and concise texts.

In this paper, we presented a new annotated corpus in two languages that extends the French data made available in (Boudin and Morin, 2013). We also compared two state-of-the art systems on this new dataset. We hope this corpus will help the NLP community to develop and validate multi-language methods for multi-sentence compression.

In order to extend the multi-language resources to more diverse languages, we plan to create a similar MSC dataset for Arabic. We also want to use our corpus to test other competitive MSC systems, such as the one based on integer linear programming we introduced in (Linhares Pontes et al., 2016).

6. Acknowledgments

This work was partially financed by the European Project CHISTERA-AMIS ANR-15-CHR2-0001.

[9]Key phrases are multi-word phrases composed of the syntactic pattern $(ADJ)^*(NPP|NC)^+(ADJ)^*$, which ADJ are adjectives, NPP are proper nouns and NC are common nouns. French, Portuguese and Spanish have similar syntactic patterns.

Method	French		Portuguese		Spanish	
	Length	CR	Length	CR	Length	CR
Baseline	9.6 ± 1.5	29%	9.5 ± 1.2	29%	9.1 ± 0.3	24%
Filippova (2010)	16.9 ± 5.1	51%	17.3 ± 5.3	52%	16.5 ± 6.4	43%
Boudin and Morin (2013)	19.7 ± 6.9	59%	22.9 ± 6.3	69%	23.4 ± 8.4	61%

Table 4: Length (average and standard deviation of tokens) and compression ratio (CR) of system outputs.

Method	French		Portuguese		Spanish	
	Gram.	Info.	Gram.	Info.	Gram.	Info.
Filippova (2010)	$\mathbf{1.65 \pm 0.58}$	1.25 ± 0.76	1.61 ± 0.64	1.51 ± 0.66	$\mathbf{1.51 \pm 0.69}$	1.02 ± 0.72
Boudin and Morin (2013)	1.56 ± 0.62	$\mathbf{1.48 \pm 0.68}$	$\mathbf{1.66 \pm 0.62}$	$\mathbf{1.70 \pm 0.59}$	1.30 ± 0.76	$\mathbf{1.16 \pm 0.82}$

Table 5: Manual evaluation of compressions (ratings are expressed on a scale of 0 to 2). All results are statistically equivalent.

7. Bibliographical References

Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328, September.

Boudin, F. and Morin, E. (2013). Keyphrase extraction for N-best reranking in multi-sentence compression. In *NAACL*, pages 298–305.

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research (JAIR*, 31:399–429.

Cohn, T. and Lapata, M. (2008). Sentence compression beyond word deletion. In *COLING*, pages 137–144.

Filippova, K. and Altun, Y. (2013). Overcoming the lack of parallel data in sentence compression. In *EMNLP*, pages 1481–1491.

Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence compression by deletion with LSTMs. In *EMNLP*, pages 360–368.

Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. In *COLING*, pages 322–330.

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Durme, B. V. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *EMNLP*, pages 1168–1179.

Ive, J. and Yvon, F. (2016). Parallel sentence compression. In *COLING, Technical Papers*, page 1503–1513.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Workshop Text Summarization Branches Out (ACL'04)*, pages 74–81.

Linhares Pontes, E., Gouveia da Silva, T., Linhares, A. C., Torres-Moreno, J.-M., and Huet, S. (2016). Métodos de otimização combinatória aplicados ao problema de compressão multifrases. In *Anais do XLVIII Simpósio Brasileiro de Pesquisa Operacional (SBPO)*, pages 2278–2289.

Luong, A. V., Tran, N. T., Ung, V. G., and Nghiem, M. Q. (2015). Word graph-based multi-sentence compression: Re-ranking candidates using frequent words. In *Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pages 55–60.

McKeown, K., Rosenthal, S., Thadani, K., and Moore, C. (2010). Time-efficient creation of an accurate sentence fusion corpus. In *HLT-NAACL*, pages 317–320.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389.

Thadani, K. and McKeown, K. (2013). Supervised sentence fusion with single-stage inference. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP*, pages 1410–1418.

Toutanova, K., Brockett, C., Tran, K. M., and Amershi, S. (2016). A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *EMNLP*, pages 340–350.

8. Language Resource References

Boudin, Florian and Morin, Emmanuel. (2013). *Keyphrase Extraction for N-best Reranking in Multi-Sentence Compression*. NAACL (2013). Available on https://github.com/boudinfl/lina-msc.

Graff, David and Cieri, Christopher and Kong, Junbo and Chen, Ke and Maeda, Kazuaki. (2011). *English Gigaword*. Linguistic Data Consortium, 5th, ISLRN 911-942-430-413-0.