Investigating Machine Learning Approaches for Sentence Compression in Different Application Contexts for Portuguese

Fernando Antônio Asevedo Nóbrega and Thiago Alexandre Salgueiro Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
São Carlos – SP – Brazil
{fasevedo,taspardo}@icmc.usp.br
www.nilc.icmc.usp.br

Abstract. Sentence compression aims to produce a shorter version of an input sentence and it is very useful for many Natural Language applications. However, investigations in this field are frequently task focused and for English language. In this paper, we report machine learning approaches to compress sentences in Portuguese. We analyze different application contexts and the available features. Our experiments produce good results, outperforming some previously investigated approaches.

1 Introduction

The sentence compression task aims to produce a shorter version of a sentence [9] and it may be useful for many Natural Language applications. As an example, in compressive summarization, the systems produce summaries with the most relevant content of one or more related texts and they compress some sentences before their inclusion in the summaries [9, 11, 8, 10, 15, 2, 6].

The current sentence compression methods usually delete some tokens or arcs in a syntactic tree of the input sentence [9, 17, 13, 4, 2, 8, 16], and these methods frequently only use information from the input sentence, although the situations and compression applications are varied and have the potential to provide different features and hints to the task. In this paper, we have investigated sentence compression approaches based on machine learning techniques, employing different types of features in order to analyze different application contexts. We have performed experiments for texts written in Portuguese and show that our methods outperform some previously investigated approaches.

The paper is organized as follows. We briefly introduce the main related work in Section 2. Our dataset is presented in Section 3. Our methods and the investigated features are presented in Section 4. In Section 5, we show the evaluation methodology based on different application contexts and the results of our methods. Some final remarks are presented in Section 6.

2 Related Work

In one of the most used approaches in the area, [17] shows a sentence compression method based on deletions of segments in a syntactic tree using the Noisy-Channel framework. The authors report the problem of lack of datasets for the training process and improve their method by applying unsupervised approaches and some manually produced constraints.

[7] performs machine learning experiments for Portuguese by using a Decision Tree technique. They investigate syntactic and semantic features extracted by the PALAVRAS parser [3] and others based on the documents that the input sentence came from, as frequency of the token and position of the sentence.

[2] presents a compressive module for summarization based on Integer Linear Programming (ILP) in a similar way to [12]. Their system uses a bi-gram model as sentence representation and makes deletions over the arcs from the dependency tree. They use features based on the labels of the arcs in the tree, shallow features (frequency of the words, if the word is a stopword, its position in the sentence and document) and analysis of modifier tags (as negation, temporal words, and others). Furthermore, they use hard constraints, which avoid some arc deletions, in order to produce more grammatical sentences.

[16] investigates features based on two kinds of sentence representation, a list of tokens and a tree of syntactic dependencies, used with ILP. The authors say that this approach shows better results because these two kinds of knowledge complement each other.

[6] presents a sentence compression method based on token deletion using the deep learning framework. The authors defend the use of this approach because of the low performance of syntactic parsers. Their best results were achieved with embedding vectors obtained by a skip gram model [14] with 256 dimensions.

3 Dataset

We selected a dataset with 770 sentence pairs (25,966 tokens, including punctuation marks) of original sentences and their respective compressed versions from the Priberam Compressive Summarization Corpus (PCSC) [1]. In the PCSC corpus, there are 801 texts/documents organized into 80 clusters and 160 summaries (two for each cluster) that were manually made based on the compressive summarization approach.

When there were two or more compressed possibilities for the same source sentence in the corpus (there was a total of 78 cases like this), we have maintained only its shortest version. This way, we aim to produce a compression method that learns as many token deletions as possible. Aiming to maintain the data unbiased, we have also removed eventual duplicated pairs.

4 Our Methods

We handled the sentence compression task as a token deletion approach in a traditional classification problem of machine learning, in which we must to an-

swer the question "Should this token be deleted?" for each token in the input sentence.

We have initially experimented three sceneries of features (Sentence, Document and Summary) based on the diversity of available information on different applications. For instance, in compressive summarization, we may use information of the input sentence, its source text and the output summary being produced. On the other hand, in text simplification, we only have the input sentence and its text. Furthermore, we also have used background features that may be added in any of these sceneries.

In the context of the **Sentence**, we have only used information from the input sentences, as simple shallow features and features derived from the syntactic and semantic analyses. As shallow features[1], we have used: if the token is inside parentheses; if the token occurs in the beginning (if it is one of the 20% first tokens), ending (if it occurs after the 80% first tokens) or in the middle of the sentence (otherwise); if the token is a stopword; the two previous and next tokens; and the token itself. It is important to say that, for the two last features, we have used a stemmer in order to reduce the dimensionality of the machine learning model. For the remaining features, we have used information extracted by the PALAVRAS parser [3], as: the POS (Part of Speech) of the token and of the two previous and next tokens; available syntactic functions of the token in the dependency tree; and semantic information (named entity and semantic class labels presented by PALAVRAS) of the token. Furthermore, PALAVRAS expands syntactic contractions (*do = de + o; dele = de + ele; no = em + o*). Thus, in order to produce compressed sentences with the same tokens of the input sentences, we contract these expansions with a simple set of manually developed rules.

In the scenery of **Document**, in addition to the Sentence features, we have also extracted information from the documents of the sentences, as follows: the position of the sentence in the document; if the token occurs in the most relevant sentence in the document (which is the sentence with the most frequent words); and the token frequency in the document (normalized by the log).

For the **Summary** scenery, in addition to the features above, we also have used information based on the summarization process, as follows: if the token was used in previous selected sentences in the summary; and the sentence position in the summary. It is important to say that we have used the available summaries in the PSCP corpus in order to extract these features.

We have also experimented **Background** features based on an embedding vector representation of words, in which each word is represented by a numeric array of n dimensions with values trained in a big text corpus. We have used vectors of 256 dimensions made by a Skip gram model [14], as performed by [6], over 1,008,353 texts from the G1 portal[2]. We have applied this vector representation in order to calculate lexical similarities among tokens and use them as features (the similarity of the token for the two previous and next tokens).

[1] Those that require limited linguistic processing.

[2] It is a famous news web portal in Brazil, at g1.globo.com

Here, the idea is to identify pairs of terms that are very similar between each other and, therefore, may be simplified, as: names of companies (e.g., Microsoft Corporation may be simplified to Microsoft) and names of famous people (e.g., President Dilma Rousseff may be simplified to President Dilma or only Dilma), and others.

Finally, since the sentence compression process is interpreted as a sequence of token deletions from the sentence and, therefore, the decision if we will keep or remove a token probably affects the next decisions in the sentence, we have also included features that indicate if the two previous tokens in the **Context** of the target token were removed.

We have experimented 5 machine learning algorithms from different approaches, as follows: Decision Tree; Logistic Regression; MultiLayer Perceptron (MLP); Naïve Bayes; and Support Vector Machine (SVM). Furthermore, we have also experimented an Ensemble approach [5], in which a set of methods are used in order to classify the inputs by using a voting system (weighted or not). In this paper, our Ensemble method is simply composed of all the previously mentioned methods with a weighted voting strategy based on the evaluated f-measure values of the methods.

5 Evaluation

We used the ten-fold cross-validation strategy in order to perform our experiments. We report the traditional Precision ($\frac{|\text{correctly classified tokens}|}{|\text{compressed sentence}|}$), Recall ($\frac{|\text{correctly classified tokens}|}{|\text{original sentence}|}$)[3] and F-measure (F-1) metrics.

We contrast our methods with the system presented by Kawamoto and Pardo [7][4], which investigated the Decision Tree framework for Portuguese language, and the compression method used by Almeida and Martins [2]. Table 1 shows the evaluation results, in which we organize the methods on the rows and the scores on the columns. In the first column, we present the groups of features that were used. For instance, the Logistic Regression approach, using Sentence, Background, and Context features, presents a 0.887 f-measure.

One may see that Logistic Regression produced the best results, largely outperforming the baseline methods. We believe that Almeida and Martins method, which is based on the ILP approach, did not show good results probably because of the size of the dataset, since we train and evaluate this method with a cross-validation methodology. It is also interesting to see that, although there are differences in performance for the different application sceneries, the results are not very different. This may happen due to the fact that all the sceneries probably use the same main features (the ones related to the Sentence scenery).

Finally, a factor that we did not explore and that may be important is the compression rate. We simply adopted the compressed sentences in the corpus,

[3] Where $|\bullet|$ is the size of $\bullet$.

[4] To the best of our knowledge, it was the first sentence compression investigation for Portuguese.

Features	Method	Precision	Recall	F-measure
	Kawamoto and Pardo [7]	0.616	0.577	0.596
	Almeida and Martins [2]	0.491	0.460	0.475
Sentence + Background + Context	Decision Tree	0.870	0.885	0.878
	Ensemble	0.736	0.783	0.759
	Logistic Regression	0.882	0.892	**0.887**
	Naïve Bayes	0.658	0.457	0.558
	MLP	0.729	0.764	0.746
	SVM	0.869	0.872	0.870
Document + Background + Context	Decision Tree	0.857	0.875	0.866
	Ensemble	0.855	0.880	0.867
	Logistic Regression	0.881	0.892	**0.887**
	Naïve Bayes	0.693	0.471	0.582
	MLP	0.774	0.793	0.784
	SVM	0.869	0.872	0.870
Summary + Background + Context	Decision Tree	0.859	0.855	0.857
	Ensemble	0.881	0.891	0.886
	Logistic Regression	0.882	0.892	**0.887**
	Naïve Bayes	0.693	0.471	0.582
	MLP	0.776	0.808	0.792
	SVM	0.869	0.872	0.870

Table 1. Evaluation of the compressed sentences produced by our methods in different sceneries

without explicitly modeling the compression rate as a parameter. However, it is known that the number of deleted tokens is directly influenced by the desired size of the summary that contains the compressed sentences. Therefore, our machine learning is probably biased by this, but so far we have not analyzed its impact in our results.

6 Final Remarks

We have investigated 6 machine learning techniques for sentence compression in three different application sceneries for texts in Portuguese. For each application context, we have presented a different set of features. In general, we have produced good results that outperformed some previous approaches for Portuguese, but we believe that there is still room for improvements. Future work includes the investigation of some hard constraints to avoid the production of ungrammatical compressed sentences, as performed by many previous investigations [9, 8, 2, 16, 6], as well as the investigation of more sophisticated methods.

Acknowledgments

The authors are grateful to CAPES and FAPESP for supporting this work.

References

1. Almeida, M.B., Almeida, M.S.C., Figueira, A.F.T.M.H., Mendes, P., Pinto, C.: A new multi-document summarization corpus for european portuguese. In: Language Resources and Evaluation Conference (LREC'14), Reykjavik, Iceland (2014) 1–7
2. Almeida, M.B., Martins, A.F.T.: Fast and robust compressive summarization with dual decomposition and multi-task learning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. (2013) 196–206
3. Bick, E.: The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)
4. Cohn, T., Lapata, M.: Sentence compression beyond word deletion. In: Proceedings of the International Conference on Computational Linguistics. (2008) 137–144
5. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple Classifier Systems. Volume 1857 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2000) 1–15
6. Filippova, K., Alfonseca, E., Colmenares, C., Kaiser, L., Vinyals, O.: Sentence compression by deletion with LSTMs. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. (2015) 360–368
7. Kawamoto, D., Pardo, T.A.S.: Learning sentence reduction rules for brazilian portuguese. In: Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science – NLPCS. (2010) 90–99
8. Klein, T.B.K.D.G.D.: Jointly learning to extract and compress. In: Proceedings of the International Conference on Computational Linguistics. (2011) 10
9. Knight, K., Marcu, D.: Statistics-based summarization – step one: Sentence compression. In: Proceedings of the AAAI. (2000) 703–711
10. Li, C., Liu, F., Weng, F., Liu, Y.: Document summarization via guided sentence compression. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2013) 490–500
11. Madnani, N., Zajic, D., Dorr, B., Ayan, N.F., Lin, J.: Multiple alternative sentence compressions for automatic text summarization. In: Proceedings of the Document Understanding Conference. (2007) 8
12. Martins, A.F.T., Smith, N.A.: Summarization with a joint model for sentence extraction and compression. In: Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing. (2009) 1–9
13. McDonald, R.: Discriminative sentence compression with soft syntactic evidence. In: Proceedings of the AAAI. (2006) 297–304
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the Advances in Neural Information Processing Systems. (2013) 3111–3119
15. Qian, X., Liu, Y.: Fast joint compression and summarization via graph cuts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2013) 1492–1502
16. Thadani, K., McKeown, K.: Sentence compression with joint structural inference. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. (2014) 65–74
17. Turner, J., Charniak, E.: Supervised and unsupervised learning for sentence compression. In: Proceedings of the 43rd Annual Meeting on Association for Computational. (2005) 290–297