Quotation Extraction for Portuguese

William Paulo Ducca Fernandes, Eduardo Motta, Ruy Luiz Milidiú

Departamento de Informática – Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro – RJ – Brazil

{wfernandes, emotta, milidiu}@inf.puc-rio.br

Abstract. *Quotation extraction consists of identifying quotations and their authors. In this work, we present a Quotation Extraction system for Portuguese that is based on Entropy Guided Transformation Learning, a supervised Machine Learning algorithm. This is the first system that uses a Machine Learning approach for Portuguese. In order to train and evaluate the proposed system, we build the* GLOBOQUOTES *corpus, with news extracted from the* GLOBO.COM *portal. Our system obtains an $F_{\beta=1}$ score of 79.02% for the subtask of associating a quotation to its author. For the whole Quotation Extraction task, the observed $F_{\beta=1}$ score value is 66.03%. These findings indicate that the overall extraction quality is highly dependant on the quotation identification subtask.*

1. Introduction

Quotation extraction [Sarmento and Nunes 2009] consists of identifying quotations from a text and associating them to their authors. In this work, we propose a Quotation Extraction system that handles direct and mixed quotations for Portuguese based on *Entropy Guided Transformation Learning* (ETL) [dos Santos and Milidiú 2009] algorithm.

ETL solves the main bottleneck of *Transformation Based Learning* [Brill 1995], which is the automatic generation of template rules. ETL is applied successfully in several computational linguistic problems [dos Santos et al. 2010, Milidiú et al. 2008].

Since we employ a supervised Machine Learning algorithm, we need an annotated corpus to train and evaluate the system. In order to accomplish this task, we build the GLOBOQUOTES corpus, with news extracted from the GLOBO.COM portal. We generate the golden features for entities, coreferences, quotations and associations between quotations and authors. Moreover, we include the part-of-speech (POS) annotation in the corpus using a state-of-the-art tagger [dos Santos et al. 2008], also based on ETL. After producing the annotations, we separate the GLOBOQUOTES corpus into two sets, training set and test set.

Quotation Extraction has been previously approached using different techniques and for several languages. The *NewsExplorer*[1] system, based on lexical rules, extracts quotations from multilingual news [Pouliquen et al. 2007]. The *Sapiens* system, based on syntatic rules, extracts quotations from news wires in French [Clergerie et al. 2009]. The *verbatim*[2] system, based on speech act rules, extracts quotations for Portuguese [Sarmento and Nunes 2009]. The EVRI portal[3] offers a Quotation Extraction API for En-

[1]http://press.jrc.it/NewsExplorer
[2]http://irlab.fe.up.pt/p/verbatim
[3]http://www.evri.com

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 204–208,
Cuiabá, MT, Brazil, October 24–26, 2011.

glish news feeds [Liang et al. 2010]. Their approach is based on rules that use several linguistic features automatically provided by standard auxiliary processors.

Our proposal differs from previous work since we use Machine Learning to automatically build specialized rules instead of human derived rules. We train our system for Portuguese, although our approach is language independent. In Table 1, we present the proposed system quality evaluated on the test set. Our system obtains an $F_{\beta=1}$ score of 79.02% for the subtask of associating a quotation to its author. For the whole Quotation Extraction task, the observed $F_{\beta=1}$ score value is 66.03%. These findings indicate that the overall extraction quality has a strong dependency on the quotation identification subtask. The proposed system performance cannot be directly compared to previous work, since the corresponding corpora are not publicly available.

Table 1. Quotation Extraction performance on the test set

Subtasks	*Precision (%)*	*Recall (%)*	$F_{\beta=1}$*(%)*
quotation association	79.02	79.02	79.02
quotation identification and association	64.79	67.32	66.03

The remaining of this work is organized as follows. In section 2, we describe the Quotation Extraction task. In section 3, we present the corpus as well as the adopted annotation. We report our modeling in section 4 and experiments in section 5. Finally, in section 6, we present our conclusions and future work.

2. Quotation Extraction

We decompose the Quotation Extraction task into two subtasks: quotation identification and quotation association to its respective coreference set label, that identifies the quotation author. In Figure 1, we show an illustrative example with three coreference sets. Coreferences are in **bold** and quotations in *italic*. Each coreference is tagged with an integer subscript that indicates its respective coreference set label. Similarly, each quotation is tagged with its respective author coreference set label.

> **Nélio Machado**$_1$, que defende **Daniel Dantas**$_2$, considerou '*estranha*'$_1$ a acusação de que **Dantas**$_2$ teria cogitado subornar **o juiz**$_3$. '*Isso é o fim da picada. Completamente sem fundamento e bem no dia em que o* ***Daniel***$_2$ *vai prestar depoimento. Estou inclinado a pedir suspeição* ***dele***$_3$ *[****Fausto de Sanctis***$_3$*]. Acho muito estranho, tem conteúdo de mais armação do que qualquer outra coisa'*$_1$ disse **ele**$_1$.

Figure 1. Quotation extraction subtasks

This work purpose is to identify quotations and their corresponding authors. Named entity recognition [Zhou and Su 2002, Solorio and López 2005, Sarmento 2006, Bick 2006] and coreference resolution [Souza et al. 2008, Cuevas and Paraboni 2008] are classical NLP tasks. Hence, we use as input to our quotation extractors golden annotation for both named entities and coreferences.

3. Corpus

Considering there is no publicly available quotation corpus for Portuguese, we build the GLOBOQUOTES corpus, with golden annotation for named entities, coreferences, quotations and associations between quotations and authors. This corpus is based on news from the GLOBO.COM portal.

The *raw corpus* is composed by 10 news genres, dated from August, 2007 to August, 2008. It has more than 44,000 news, totalizing more than 13.5 million tokens. The predominant genre is *Sports*, accounting for 58.5% of the corpus. Next, we have *General* with 15%, and *Celebrities* with 13.9%. The other genres – *Arts*, *Science*, *Economy*, *Education*, *World*, *Politics* and *Technology* – represent all together 12.6% of the corpus.

GLOBOQUOTES is a random sample of 685 news from the raw corpus. This sample preserves the original distribution by news genre. The corpus information is codified in a per token basis.

4. Modeling

In our modeling, we adopt a token classification approach. We decompose the Quotation Extraction task into two subtasks: *quotation identification* and *association between quotation and author*. We further decompose the quotation identification subtask into three identification subtasks that are sequentially solved: *quote beginning*, *quote end* and *quote bounds*, that corresponds to match a quote beginning with a quote end. Given the solution of the first subtask, the last two subtasks, quote end and bounds, are easily solved by a simple heuristic given by a set of regular expressions. We use a 5-fold cross-validation over the training set to calibrate the model. In order to evaluate it, we use the test set.

Since ETL is an error-driven learning algorithm, we must provide a baseline system for each subtask. Each baseline is given by a set of regular expressions manually constructed. We tackle the *quote beginning identification* subtask using the POS and NE features applying the ETL algorithm. We tackle the *quote end identification* subtask using the word, POS, NE and quote start features and the baseline system. We tackle *quote bounds* subtask using the quote start and quote end features and the baseline system. We tackle the *association between quotation and author* subtask using the POS, quote and coreference features applying the ETL algorithm.

5. Experiments

In this section, we present the experimental setup and results. In order to evaluate the proposed system, we separate the annotated corpus into training set, with 802 quotations, and test set, with 205 quotations.

Table 2. Performance of the Quotation Extraction task

Model	*Precision (%)*	*Recall (%)*	$F_{\beta=1}$*(%)*
ETL	64.79	67.32	66.03
Baseline	48.62	51.71	50.12

In Table 2, we present the performance of the whole system, along with the baseline system. When we use the automatically identified quotations, our system achieves

Table 3. Performance of the association between quotation and author subtask

Model	*Precision (%)*	*Recall (%)*	$F_{\beta=1}$*(%)*
ETL	79.02	79.02	79.02
Baseline	57.56	57.56	57.56

an $F_{\beta=1}$ score of 66.03%. In Table 3, we show the performance when using the golden annotation of quotations. In this setup, we obtain an $F_{\beta=1}$ score of 79.02%. These findings indicate that the overall extraction quality has a strong dependency on the quotation identification subtask.

6. Conclusions

Quotation extraction consists of identifying quotations from a text and associating them to their authors. The Quotation Extraction task is decomposed into two subtasks: quotation identification and association between quotation and author.

In this work, we propose a Quotation Extraction system for Portuguese. The system is based on *Entropy Guided Transformation Learning* algorithm.

In order to train and evaluate the proposed system, we build the GLOBOQUOTES annotated corpus. We produce golden features for entities, coreferences, quotations and associations between quotations and authors. Also we include part-of-speech tags in the corpus using a state-of-the-art tagger. To the best of our knowledge, this is the first corpus with annotations which let one identify quotations and associate them to their authors produced for Portuguese.

We intend to improve our results by improving the results of the subtasks of quote beginning identification and association between quotation and author, creating new derived features and applying other classification algorithms.

References

Bick, E. (2006). Functional aspects in Portuguese NER. In *Proceedings of the 7th International Conference on Computational Processing of the Portuguese Language*, pages 80–89.

Brill, E. (1995). Tranformation-based Error-driven Learning and Natural Language Processing: a case study in Part-of-Speech Tagging. In *Computational Linguistics*, pages 543–565.

Clergerie, E. V. L., Sagot, B., Stern, R., Denis, P., and Recourcé, G. (2009). Extracting and visualizing quotations from news wires. In *Proceedings of LTC 2009*, pages 522–532.

Cuevas, R. R. M. and Paraboni, I. (2008). A machine learning approach to portuguese pronoun resolution. In *IBERAMIA*, pages 262–271.

dos Santos, C. N. and Milidiú, R. L. (2009). Entropy Guided Transformation Learning. In *Foundations of Computational Intelligence (1)*, pages 159–184.

dos Santos, C. N., Milidiú, R. L., Crestana, C. E. M., and Fernandes, E. R. (2010). ETL Ensembles for Chunking, NER and SRL. In *Eleventh International Conference on*

Computational Linguistics and Intelligent Text Processing, volume 6008 of *Lecture Notes in Computer Science*, pages 100–112. Springer Berlin.

dos Santos, C. N., Milidiú, R. L., and Renteria, R. P. (2008). Portuguese part-of-speech tagging using Entropy Guided Transformation Learning. In *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language*, pages 143–152.

Liang, J., Dhillon, N., and Koperski, K. (2010). A large-scale system for annotating and querying quotations in news feeds. In *Proceedings of the 3rd International Semantic Search Workshop*, pages 7:1–7:5.

Milidiú, R. L., dos Santos, C. N., and Duarte, J. C. (2008). Portuguese corpus-based learning using ETL. *Journal of the Brazilian Computer Society*, pages 17–27.

Pouliquen, B., Steinberger, R., and Best, C. (2007). Automatic detection of quotations in multilingual news. In *Proceedings of the Internation Conference Recent Advances in Natural Language Processing*, pages 25–32.

Sarmento, L. (2006). Siemês: a named entity recognizer for Portuguese. In *Proceedings of the 7th International Conference on Computational Processing of the Portuguese Language*, pages 90–99.

Sarmento, L. and Nunes, S. (2009). Automatic extraction of quotes and topics from news feeds. In *4th Doctoral Symposium on Informatics Engineering*.

Solorio, T. and López, A. L. (2005). Learning named entity recognition in Portuguese from Spanish. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, pages 762–768.

Souza, J. G., Gonçalves, P. N., and Vieira, R. (2008). Learning coreference resolution for portuguese texts. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, pages 153–162, Berlin, Heidelberg. Springer-Verlag.

Zhou, G. and Su, J. (2002). Named entity recognition using an HMM-based chunk tagger. In *Proceedings the 40th Annual meeting of the ACL*, pages 647–655.