6
Conclusions

Quotation extraction consists of identifying quotations from a text and associating them to their authors. We propose a Quotation Extraction system for the Portuguese language.

Quotation Extraction has been previously approached for several languages by using rule-based systems. Our proposal differs from previous work since we use Machine Learning to automatically build specialized rules instead of human-derived ones. Machine Learning models usually present stronger generalization power compared to human-derived ones due to their capacity to adapt to different writing styles. In human-derived models, even small changes in the writing style may need several modifications in the human-derived rule set. In addition, we are able to easily adapt our model into other languages, needing nothing but a list of verbs of speech for a given language. The previously proposed systems would probably need a rule set adaptation to correctly classify the quotations, which would be time consuming.

In our work, we create three models to tackle the Quotation Extraction task. The first model is Structured Perceptron based on the weighted interval scheduling problem (SP-WIS). In this model, we find a maximum-weight subset of non-overlapping tasks, where each task represents a combination of quotation and author candidate. The second model is ETL. In this model, we divide the original task into two subtasks, quotation identification and association between quotation and author. ETL predicts the corresponding label sequence for each subtask. The last model is a baseline system, created using a rule-based approach.

SP-WIS presents the best quality compared to ETL and the baseline system. The performance of our models cannot be directly compared to previous work, since the corresponding corpora are not publicly available.

We analyze the residual errors for our best model, the SP-WIS. After a diligent error analysis, we have identified two frequent patterns which our model generally does not identify. If our model learns to classify those patterns, recall errors will be considerably reduced.

In this work, we have built the GloboQuotes annotated corpus.

We produce golden features for named entities, coreferences, quotations and associations between quotations and authors. Also, we have included part-of-speech tags in the sample corpus by 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.

In future work, we intend to enhance the size of GloboQuotes, producing golden annotation for named entities, coreferences, quotations and associations between quotations and authors. This annotation would certainly enhance the quality of our models. Moreover, we could apply our system to other languages. Our only need is a list of verbs of speech in a given language. In addition, we may prepare our system to identify indirect quotations, the only quotation type our system is not yet prepared to deal with. In order to do this, we need to annotate indirect quotations in GloboQuotes. Futhermore, we could modify SP-WIS in order to make it identify two quotation patterns that are part of recall errors at present. Finally, for the Structured Perceptron algorithm, we could change the optimization problem as an attempt to improve the quality of our system.