# 5
# Experiments

In this chapter, we present corpus composition statistics and the adopted annotation. We also present the experimental setup and the observed quality of the ETL model with its subtasks and the Structured Perceptron model. Moreover, we analyze the residual errors for our best model.

## 5.1
## Corpus

Since there is no publicly available quotation corpus for Portuguese, we have built the GLOBOQUOTES corpus with golden annotations for named entities, coreferences, quotations and associations between quotations and authors. This corpus is based on news pieces in Portuguese from the GLOBO.COM portal.

During the annotation process, we produced guidelines for named entity annotation, coreference annotation and association between quotation and author annotation. We present the annotation guidelines in Appendix A.

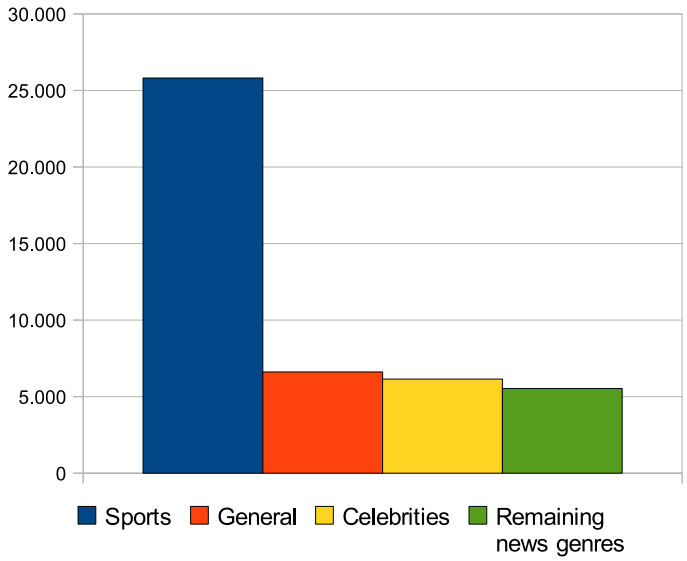


Figure 5.1: GLOBO.COM corpus distribution of the main news genres

PUC-Rio - Certificação Digital Nº 1012671/CA

### 5.1.1 Statistics

The *raw corpus* is composed of 10 news genres, dated from August, 2007 to August, 2008. It has more than 44,000 pieces of news, amassing 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 remaining genres – *Arts*, *Economy*, *Education*, *Politics*, *Science*, *Technology* and *World* – represent all together 12.6% of the corpus. In Figure 5.1, we present a chart with the distribution of the main news genres. In Figure 5.2, we show a chart with the distribution of the remaining news genres.

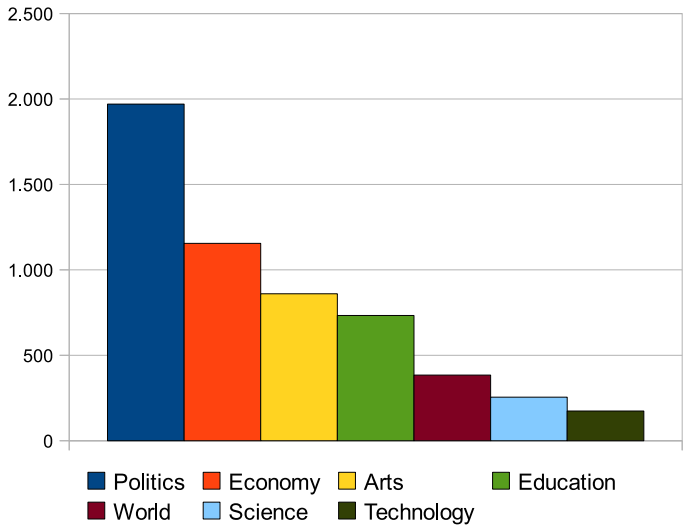


Figure 5.2: GLOBO.COM corpus distribution of the remaining news genres

GLOBOQUOTES is a random sample of 685 pieces of news from the raw corpus. This sample preserves the original distribution by news genres.

In Figure 5.3, we present a chart with the distribution of distances from the quotations to their authors in GLOBOQUOTES. The categories are indicators of relative distance of the d± type. For instance, the *3-* category indicates that the quotation author is the third coreference before the quotation. In that chart, we see the most frequent category is *1+*, i.e., most frequently, the quotation author is the first coreference after the quotation.

### 5.1.2 Annotation

The corpus information is codified on a per token basis. In Table 5.1, we show an example which illustrates the corpus annotation for its eight basic features. The first one is the word. Next, we have the POS annotation provided

PUC-Rio - Certificação Digital Nº 1012671/CA

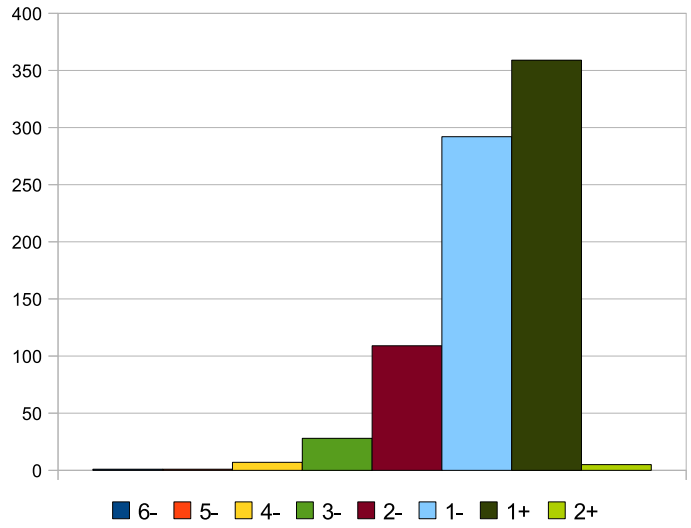


Figure 5.3: Distribution of distances from quotations to their authors in GloboQuotes

by a state-of-the-art tagger (11). We use the IOB format (31) to annotate two kinds of named entities (NE): person and organization. Each coreference is tagged with its corresponding set label. In the quote start (QS) feature, we use $S$ to mark its starting token. Similarly, we use $E$ in the quote end (QE) feature. In the Quote feature, we assign the $q$ tag to tokens which belong to a quote. We use the PQ feature to associate a quotation to its author by utilizing a relative distance tag of type $r$d±. For instance, the tag *r2-* indicates that the quotation author is the second coreference before the quotation.

Table 5.1: Annotated corpus excerpt

| *Word* | **Kubica** | **:** | **‘** | **Fiquei** | **atrás** | **de** | **dois** | **carros** | **tops** | **’** |
|---|---|---|---|---|---|---|---|---|---|---|
| *POS* | NPROP | : | ’ | V | PREP | PREP | NUM | N | ADJ | ’ |
| *NE* | I-PER | O | O | O | O | O | O | O | O | O |
| *Coref* | ref00 | _ | _ | _ | _ | _ | _ | _ | _ | _ |
| *QS* | _ | _ | _ | S | _ | _ | _ | _ | _ | _ |
| *QE* | _ | _ | _ | _ | _ | _ | _ | _ | E | _ |
| *Quote* | _ | _ | _ | q | q | q | q | q | q | _ |
| *PQ* | _ | _ | _ | r1- | r1- | r1- | r1- | r1- | r1- | _ |

## 5.2 Experimental Setup

In order to assess the proposed models, we separate the annotated corpus into a training set and a test set. We show the annotated corpus set sizes in Table 5.2.

PUC-Rio - Certificação Digital Nº 1012671/CA

Table 5.2: Annotated corpus set sizes

| *Part* | *#Feeds* | *#Sentences* | *#Tokens* | *#Quotations* |
|---|---|---|---|---|
| Training | 552 | 7,963 | 174,415 | 802 |
| Test | 133 | 1,834 | 41,613 | 205 |

In our work, we create three models to tackle the Quotation Extraction task. The first model is Structured Perceptron based on weighted interval scheduling problem (SP-WIS). In this model, we find a maximum-weight subset of non-overlapping tasks, where each task represent 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.

In order to calibrate the ETL model, we use a 5-fold cross-validation over the training set in the quote beginning identification subtask and association between quotation and author subtask. By testing a wide range of initial parameter values, the best combination found is a window size of 5 and a rule threshold of 2.

Since the Structured Perceptron is an online algorithm and the order in which examples are processed influences the learned model, we calibrate the model running a 5-fold cross-validation 5 times over the training set. By testing a wide range of initial parameter values, the best combination found is the number of epochs of 65, loss weight of 10 and template size between 2 and 4, removing root.

Using a computer with an Intel Core i7 processor of 2.8GHz and 6GB of RAM, it takes 1 minute and 30 seconds to create the ETL model and 2 seconds to evaluate it. To create the Structured Perceptron model, it takes 30 seconds and to evaluate it, 1 second.

## 5.3 Quality Results

In Table 5.3, we present the quality of our models assessed in the test set. SP-WIS obtains an $F_{\beta=1}$ score of 76.80%, which is an error reduction of 31.70% compared to the baseline system. Comparing to ETL, SP-WIS reduces errors by 19.28%. The performance of our models cannot be directly compared to previous work, since the corresponding corpora are not publicly available.

We also show the quality of ETL subtasks. We present the performance of the quote beginning identification subtask in Table 5.4, together with the

PUC-Rio - Certificação Digital Nº 1012671/CA

Table 5.3: Performance of the Quotation Extraction task

| *Model* | *Precision (%)* | *Recall (%)* | $F_{\beta=1}$*(%)* |
|---|---|---|---|
| SP-WIS | 83.24 | 71.49 | 76.80 |
| ETL | 69.44 | 73.17 | 71.26 |
| Baseline | 64.35 | 67.80 | 66.03 |

baseline system. The proposed model reduces errors by 80.18% compared to the baseline system.

Table 5.4: Performance of the quote beginning identification subtask

| *Model* | *Precision (%)* | *Recall (%)* | $F_{\beta=1}$*(%)* |
|---|---|---|---|
| ETL | 86.70 | 92.20 | 89.36 |
| Baseline | 30.18 | 99.51 | 46.31 |

For the quote end identification subtask, we present the performance of our baseline system in Table 5.5. When using the golden annotation of the quote start feature, we obtain an $F_{\beta=1}$ score of almost 100%.

Table 5.5: Performance of the quote end identification subtask

| *Setup* | *Precision (%)* | *Recall (%)* | $F_{\beta=1}$*(%)* |
|---|---|---|---|
| goldenQuoteStart | 98.62 | 98.25 | 98.44 |
| quoteStart | 85.71 | 90.73 | 88.15 |

Table 5.6: Performance of the quote bounds subtask

| *Setup* | *Precision (%)* | *Recall (%)* | $F_{\beta=1}$*(%)* |
|---|---|---|---|
| goldenQuoteStart, goldenQuoteEnd | 100.00 | 100.00 | 100.00 |
| quoteStart, quoteEnd | 85.25 | 90.24 | 87.68 |

We present the performance of our baseline system in Table 5.6 for the quotation bounds subtask. When using the golden annotation of the quote start and quote end features, we obtain an $F_{\beta=1}$ score of 100%.

PUC-Rio - Certificação Digital Nº 1012671/CA

## 5.4 Error Analysis

We analyse the residual errors for our best model, the SP-WIS. They are divided into two categories, precision errors and recall errors. After a diligent error analysis, we have not identified any pattern for precision errors. However, for recall errors, we have identified two frequent patterns which our model generally does not solve. We present examples of those patterns in Figure 5.4. The first one, presented in examples 1 to 3, is *a quotation* followed by *a verb of speech* and *a period*. The second one, presented in examples 4 and 5, is *a quotation* followed by *a period*. If our model learns to classify those patterns, recall errors will be considerably reduced.

1. “Quem prometeu comprovar as denúncias com documentos foi a Denise Abreu. Ela não apontou uma testemunha, um documento sequer”, disse.
2. ‘A idéia não é uma manifestação para provocar o caos. É para mostrar que tem ciclistas disputando espaço de maneira desigual’, afirmou.
3. –Se o Barça foi mal nestes dois últimos anos, não é por culpa de um ou dois jogadores, mas por culpa de todos os que fizemos parte da estrutura profissional – explica.
4. “Agora o Canal do Cunha virou uma questão nacional. Não é mais um problema do Rio de Janeiro. É um problema do país. Portanto, eu espero brevemente que essas obras comecem”.
5. ‘Agora, temos de identificar os funcionários públicos e privados que pediam os atestados porque isso prejudica alguém’.

Figure 5.4: Several examples of SP-WIS recall errors

PUC-Rio - Certificação Digital Nº 1012671/CA