Adapting Freely Available Resources to Build an Opinion Mining Pipeline in Portuguese

Patrik Lambert[1] **and Carlos Rodríguez-Penagos**[2]
[1] Pompeu Fabra University, C. Roc Boronat 138
[2] Barcelona Media, Av. Diagonal 177
08018 Barcelona, Spain
patrik.lambert@upf.edu, carlos.rodriguez@barcelonamedia.org

Abstract

We present a complete UIMA-based pipeline for sentiment analysis in Portuguese news using freely available resources and a minimal set of manually annotated training data. We obtained good precision on binary classification but concluded that news feed is a challenging environment to detect the extent of opinionated text.

Keywords: Sentiment Analisys, UIMA, NLP pipeline, Opinion Mining

1. Introduction

This paper describes a natural language processing pipeline aiming at detecting opinionated words and the targets of these opinions in Portuguese. In order to build the training data for the statistical classifiers used to detect opinionated content, the pipeline performs a series of annotations mixing information crafted manually and obtained automatically. The successive annotation layers are implemented within the UIMA framework[1]. UIMA is an architecture for the processing of multimodal, unstructured information that supports the integration of processing modules. In this paper we focus on the freely available linguistic resources for Portuguese that we integrated in several components of the pipeline and on how we adapted these resources to meet our requirements. The aim is to provide a guide to building an opinion mining pipeline in Portuguese with freely available resources. We are aware of similar work on parts of this pipeline, such as part-of-speech tagging (Branco and Silva, 2004), parsing (Barreto et al., 2006; Silva et al., 2010) or sentiment analysis (Silva et al., 2009; Souza and Vieira, 2012), but not on a full UIMA-based pipeline, with specific requirements such as the use of an EAGLES-like part-of-speech tag set.

The rest of the paper is organised as follows. We first give a short description of each pipeline module and then focus on the components which required the most adaptation work. Section 3. reports on how we trained the PoS tagger, Section 4. deals with the training of the dependency parser and Section 5. is about the detection of entities and polar words. We finally perform an evaluation of the pipeline and conclude.

2. Pipeline Overview

Sentence detector This component ensures the segmentation of the input text in sentences. It is implemented with the OpenNLP[2] maximum-entropy-based sentence detector.

Tokenizer It ensures the segmentation of the input sentences in tokens. It is based on manually crafted rules.

Part-of-speech tagger Annotates each token with its most probable part-of-speech (PoS). It is implemented with the OpenNLP maximum-entropy-based PoS tagger. More details are given in Section 3.

Lemmatizer It gives the lemma (or base form) of a token. It is based on a dictionary of lemmas and rules (see Section 3.).

PoS processor It gives a simplified (and thus less specific) version of the tags annotated by the PoS tagger as well as a decomposition into morphosyntactic features. For example, the tag “AQ0FP0” (in a format mostly following Eagle’s recommendations), is written in two more general forms with one or two characters (“A”, “AQ”) and its morphosyntactic features are ”gen=f|num=p”, that is gender is feminine and number is plural. Less specific PoS tags and morphosyntactic features can be useful for the following pipeline components.

Chunker This component annotates noun phrases, based on rules which take the PoS tags into account.

Syntactic dependency analyser It provides tokens with annotations of the syntactic dependency relations with other tokens. They were implemented by adapting the DeSR[3] statistical-based tool (Attardi, 2006).

Named Entity recogniser This module performs named entity recognition and classification (NERC). It is based on the JULIE Lab Named Entity Tagger (JNET)[4], which is a machine learning tool implementing Conditional Random Fields (CRF).

Polar Word mapper It annotates words carrying an opinion, called sentiment words, opinion words, opinion cues or polar words. Polar words are obtained from word lists and dictionaries. This component implementation is based on the UIMA ConceptMapper

[1] http://uima.apache.org/

[2] http://opennlp.apache.org/

[3] https://sites.google.com/site/desrparser/

[4] http://www.julielab.de/Resources/Software/NLP_Tools.html

module, modified for simple PoS-based word sense disambiguation.

Quantifier and negations annotator This module identifies the tokens acting as quantifiers or negations. It was developed by using a dictionary of quantifiers and negations within UIMA ConceptMapper module.

3. OpenNLP POS Tagger

This section presents the training corpora adapted to train a POS tagger for Portuguese, as well as the POS tagger evaluation on part of these corpora. Since we need PoS tags with different levels of granularity in the separate components of our pipeline, a requirement was to use a tagset in line with EAGLES guidelines[5] (the first character gives the main category, the second one the subcategory, and the subsequent ones more specific attributes).

3.1. Training Data

To train the part-of-speech tagger we had two annotated corpora available as well as two lexicons with inflected forms and their possible lemma and part-of-speech.

- Bosque 8.0 (Afonso et al., 2002): this is a 221378 word corpus in the news domain in European Portuguese (news from Público, year 2000) and Brazilian Portuguese (news from Folha de São Paulo, year 1994). It is a part of the "floresta sintáctica" with manual syntactic analysis. This version was adapted to EAGLES format by Garcia and Gamallo (2010), and was used to train the PoS tagger distributed in Freeling 3.0 (Padró and Stanilovsky, 2012). The only adaptation we performed was to join some contractions and recognised clitics (see next bullet point).

- Mac Morpho (Aluísio et al., 2003): this is a 1.2 million word corpus in the news domain in Brazilian Portuguese (news from Folha de São Paulo, year 1994). It has its own tag set with 25 PoS tags, plus punctuation signs. We mapped it to EAGLES-like format with the help of a word form dictionary containing EAGLES-like tags. To build this dictionary, we merged two freely available word form dictionaries: (i) Unitex-PB[6] (Muniz, 2004), in DELAF format[7] and (ii) the dictionary of European Portuguese distributed in Freeling, which is a manually corrected adaptation of LABEL-LEX with EAGLES tags (Garcia and Gamallo, 2010). We first mapped Unitex-PB to EAGLES-like tags and then merged it with the EAGLES-like version of LABEL-LEX. In order to map Mac Morpho tags to EAGLES-like tags, we first determined manually the possible EAGLES-like categories of each tag in the Mac Morpho tag set, depending on the context. For each word of the corpus, we then looked the possible tags in the dictionary that matched these categories. In case several tags were possible, we kept the common characters of the matching tags, setting other characters to "0" (which indicates the corresponding information is not specified). For example, in the Mac Morpho corpus passage "Jersei_N atinge_V", "atinge" (reach) is tagged as verb, and there are two possible tags in our dictionary matching this category: VMM02S0 and VMIP3S0. Having no way to disambiguate between the imperative tense, second person ("__M02__") and the indicative present tense, third person ("__IP3__"), these characters are set to "0", yielding "VM000S0" as the final tag. In this way, the corpus morphosyntactic annotation is still under-specified, but much less so than with the original Mac Morpho tag set. Finally, clitics were recognised and some contractions (marked with a "|+" in the corpus) were joined. The decision to join or not a contraction was reached based on the use of split or joined variants in a corpus of news and blogs.

The statistics of the adapted corpora and lexicons used for training the PoS tagger are presented respectively in Tables 1 and 2.

Training corpus	Sentences	Words
Bosque 8.0 (Freeling)	7892	215428
Mac Morpho	50919	1150941

Table 1: Size of the corpora used for training the PoS tagger.

Lexicon	Entries	Lemmas
Freeling 3.0 (Label-lex)	1257689	96724
Unitex-PB	878651	60948
Merged lexicon	1587154	124858

Table 2: Size of the word form lexicons used. They contain information on the form, lemma and possible tags of the words.

3.2. Evaluation

For the purpose of evaluation we randomly extracted 10% of the Bosque and Mac Morpho corpora, and trained a tagger on the remaining 90%. Table 3 gives the precision obtained on each evaluation set (Bosque and Mac Morpho) for the PoS taggers trained on the Bosque corpus, on the Mac Morpho corpus and on the merging of both. A significant part of the errors are due to inconsistencies of the tags used in the training and evaluation corpus. For example, in Mac Morpho, some proper nouns have information of gender and number, but not in Bosque (for instance, Paulo is tagged as "NPMS000" in the training corpus and "NP00000" in the evaluation corpus. Although the tagger proposes correctly "NPMS000", it counts as an error. Another example are the common adjectives and nouns (such as "planeta", "estudante") which are always tagged as "common" genre in Mac Morpho and which have genre and number information in Bosque. To avoid counting these cases as errors,

[5] http://www.ilc.cnr.it/EAGLES/annotate/annotate.html

[6] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/publicacoes.html

[7] http://www.cis.uni-muenchen.de/people/lg3/ManuelUnitex.pdf

we implemented an evaluation on only the first category (1 character) of the tag (referred to as "eval1" in Table3). We can observe that the tagger found the right category of more than 95% of the words (with a recall of 100%), even when the training and evaluation corpora were different.

Training	Tagger	Eval B	Eval MM	Eval 1 B	Eval 1 MM
B		94.0	81.6	96.6	94.8
B	+dict	94.7	84.2	97.3	95.6
MM		83.9	96.5	95.8	98.2
MM	+dict	85.1	93.2	95.2	97.1
B+MM		88.1	96.2	96.7	98.1
B+MM	+dict	91.0	91.4	97.0	98.0

Table 3: Precision of Part-of-speech tagger for different training and evaluation corpora. "B" and "MM" refer respectively to Bosque 8.0 and Mac Morpho corpora. In "Eval", the precision is calculated based on the full tag whereas in "Eval1", only the first letter of the tag is taken into account. "+dict" indicates that our word form dictionary was used to restrict the possible choices of the tagger.

4. Syntactic Dependency Analysis

To train the DeSR statistical dependency parser, we used the corpus distributed in the shared task on multilingual dependency parsing at the tenth CoNLL conference (Buchholz and Marsi, 2006). This corpus is actually the Bosque corpus (version 7.3). Our problem with this corpus is that PoS tags are coarse grained, indicating only the main category, and in a format different from EAGLES guidelines. To be able to use this dependency parsing training data to parse unseen text tagged with our PoS tagger, we thus had to convert the PoS tags to EAGLES-like format. A section of the CoNLL-X shared task corpus is actually included in the Bosque 8.0 corpus mentioned earlier. The two versions differ in the segmentation of multi-words, contractions and clitics, plus some other differences such as typos which had been corrected in a version and not the other or differences in punctuation. We were nevertheless able to automatically map the two versions, yielding a 126k word corpus with syntactic dependency annotations and EAGLES-like PoS tags. For the fragment of the CoNLL-X shared task corpus not included in Bosque 8.0 (81k words), we had no manually EAGLES-like tags and we automatically re-tagged it with our PoS-tagger.[8] Having rich PoS tags, we also regenerated the PoS features (such as genre, person, number) with our PoS processor.

5. Entity Recognition and Annotation

5.1. Named Entities

We took as NERC training corpus a merge of the Bosque 8.0 corpus, which also contains annotations of named entities, and of the "Colecção dourada do HAREM" (HAREM gold collection) corpus (Rocha and Santos, 2007).

Training	training #	test #	prec	rec	f1
All	14986	2997	0.47	0.10	0.16
Opinionated	4170	834	0.30	0.55	0.39

Table 4: Recognition of opinionated sentences in Portuguese news corpus. This table shows the number of training and test sentences as well as the precision (prec), recall (rec) and F-measure (f1) values.

Training	training #	test #	prec	rec	f1
All	14986	2997	0.10	0.40	0.16
Opinionated	4170	834	0.63	0.68	0.65

Table 5: Polar classification of opinionated sentences in Portuguese news corpus. The table indicates the number of training and test sentences as well as the precision, recall and F1 values.

5.2. Quantifiers and Negations

The dictionary of quantifiers and negations was manually written by a Spanish linguist and a native Brazilian Portuguese teacher.

5.3. Polar words

To detect polar words, we used the second version of the SentiLex dictionary (Silva et al., 2012), developed by P. Carvalho, M. Silva and J. Ramalho at Lisboa University. This dictionary is richer that many similar resources in other languages, allowing the polarity of a word to depend on its PoS tag. For example, "luto" as a verb means a fight, an effort, which has a positive polarity, whereas as a noun it means "mourning", which has a negative polarity.

6. Pipeline Evaluation

In the framework of the "Social Media" Spanish project, we evaluated this pipeline on news (mostly of Portugal and in a minor proportion, of Brazil). Note that in the news domain, the distribution of opinionated segments is low. In the languages analized along with Portuguese (English, Catalan and Spanish) an average 35% of the sentences have some subjective aspect (not necessarily opinionated, whereas in our Portuguese corpus these kind of analyzable sentences was even lower. Learning from this sparse dataset is necessarily a very hard problem.

We performed ablation tests to determine an adequate combination of attributes for the training vectors. The most complete model included annotations for lemmas, polar words and PoS tags. We trained a model using all examples (opinion-bearing ones and other, non-opinionated) and another one using only opinionated examples. The metrics obtained in a 10% of documents from training with the other 90% for are shown in Table 4. The model trained with all examples has better precision, but very low recall, while the one trained exclusively on the opinionated sentences suffer from low precision but has a high recall. A harmonic F1 of almost 40% points was achieved.

The results for sentiment classification of the phrases (once selecting the ones that actually are opinionated) are better,

[8] This number is significantly less than what is available in CONLLX data sets for other languages such as Spanish or Catalan, but DeSR is able to produce mostly correct parses.

as shown in Table 5, suggesting that the difficult step in this genre is determining where the opinion is being expressed rather than doing a binary classification once the polar character has been recognised. Newswire text is not ideal for measuring opinion polarity due to the way attitude is expressed, as opposed to the more explicit expression of speaker position in blogs or in social media channels as a whole. Although the lexicon used is of a very high quality, it was not adapted by us to the genre or domain. In our experience, this is a vital step to ensure any kind of significant results in Sentiment Analysis tasks. Doing this involves, at a minimun, doing a manual review of the top frequency words in the document collection in order to trim or expand the polar lexicon.

7. Conclusions

We described how we built a UIMA-based opinion mining pipeline in Portuguese from freely available resources and using EAGLES-like part-of-speech tags. One of the main challenges we faced was ensuring that tagsets, segmentation criteria and other variables were similar all along the processing modules and training dataset so that robustness and homogenity were achieved. Integrating processing modules within a common architecture and framework helps achieve this goal. To enquire about the availability of the adapted resources, please contact the authors.

To evaluate the pipeline ultimate output, we automatically detected opinionated sentences in a news corpus and performed classification of opinionated sentences. The results suggest that the difficult step in the news genre is to determine where the opinion is being expressed rather than doing a binary classification once the polar character has been recognised, but also that adapting a polar lexicon to each genre and thematic domain is a vital step to ensure a proper balance between precision of classification and recall in the extraction phase.

8. Acknowledgments

The research leading to these results has received funding from the CDTI through the CENIT project Social Media and from the Seventh Framework Program of the European Commission through the Intra-European Fellowship (CrossLingMind-2011-300828) Marie Curie Actions.

9. References

Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá(c)tica: a treebank for Portuguese. In *Proc. of the International Conference on Linguistic Resources and Evaluation (LREC)*, pages 1698–1703, Las Palmas, Spain.

Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., and Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In *Proceedings of the 6th international conference on Computational Processing of the Portuguese language (PROPOR).*, pages 110–117, Faro, Portugal. Springer-Verlag.

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 166–170, New York City, June. Association for Computational Linguistics.

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M. F., Nunes, F., and Silva, J. (2006). Open resources and tools for the shallow processing of Portuguese: the TagShare project. In *Proc. of the International Conference on Linguistic Resources and Evaluation (LREC)*, Genoa, Italy.

Branco, A. and Silva, J. (2004). Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In *Proc. of the International Conference on Linguistic Resources and Evaluation (LREC)*, Lisbon, Portugal.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York, USA. Association for Computational Linguistics.

Garcia, M. and Gamallo, P. (2010). Análise morfossintáctica para português europeu e galego: Problemas, soluções e avaliação. *Linguamática*, 2(2):59–67.

Muniz, M. C. M. A. (2004). A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, Instituto de Ciências Matemáticas de São Carlos, USP.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey.

Rocha, P. and Santos, D. (2007). Disponibilizando a Colecção Dourada do HAREM através do projecto AC/DC. In Santos, D. and Cardoso, N., editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, Linguateca.*, pages 307–326.

Silva, M. J., Carvalho, P., Sarmento, L., de Oliveira, E., and Magalhaes, P. (2009). The design of optimism, an opinion mining system for portuguese politics. *New Trends in Artificial Intelligence: Proceedings of EPIA*, pages 12–15.

Silva, J., Branco, A., Castro, S., and Reis, R. (2010). Out-of-the-box robust parsing of Portuguese. In Pardo, T. A. S., Branco, A., Klautau, A., Vieira, R., and Strube de Lima, V. L., editors, *Computational Processing of the Portuguese Language*, volume 6001 of *Lecture Notes in Computer Science*, pages 75–85. Springer.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In *Lecture Notes in Computer Science (LNCS) / Lecture Notes in Artificial Intelligence (LNAI), International Conference on Computational Processing of Portuguese (PROPOR)*, Coimbra, Portugal.

Souza, M. and Vieira, R. (2012). Sentiment analysis on twitter data for portuguese language. In Caseli, H., Villavicencio, A., Teixeira, A., and Perdigão, F., editors, *Computational Processing of the Portuguese Language*, volume 7243 of *Lecture Notes in Computer Science*, pages 241–247. Springer.