Contrasting objective and subjective Portuguese texts from heterogeneous sources

Michel Généreux
Centro de Linguística da
Universidade de Lisboa (CLUL)
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
genereux@clul.ul.pt

William Martinez
Instituto de Linguística
Téorica e Computacional (ILTEC)
Avenida Elias Garcia, 147 - 5° direito
1050-099 Lisboa - Portugal
william@iltec.pt

Abstract

This paper contrasts the content and form of objective versus subjective texts. A collection of on-line newspaper news items serve as objective texts, while parliamentary speeches (debates) and blog posts form the basis of our subjective texts, all in Portuguese. The aim is to provide general linguistic patterns as used in objective written media and subjective speeches and blog posts, to help construct domain-independent templates for information extraction and opinion mining. Our hybrid approach combines statistical data along with linguistic knowledge to filter out irrelevant patterns. As resources for subjective classification are still limited for Portuguese, we use a parallel corpus and tools developed for English to build our subjective spoken corpus, through annotations produced for English projected onto a parallel corpus in Portuguese. A measure for the saliency of n-grams is used to extract relevant linguistic patterns deemed "objective" and "subjective". Perhaps unsurprisingly, our contrastive approach shows that, in Portuguese at least, subjective texts are characterized by markers such as descriptive, reactive and opinionated terms, while objective texts are characterized mainly by the absence of subjective markers.

1 Introduction

During the last few years there has been a growing interest in the automatic extraction of elements related to feelings and emotions in texts, and to provide tools that can be integrated into a more global treatment of languages and their subjective aspect. Most research so far has focused on English, and this is mainly due to the availability of resources for the analysis of subjectivity in this language, such as lexicons and manually annotated corpora. In this paper, we contrast the subjective and the objective aspects of language for Portuguese.

Essentially, our approach will extract linguistic patterns (hopefully "objective" for newspaper news items and "subjective" for parliamentary speeches and blog posts) by comparing frequencies against a reference corpus. Our method is relevant for hybrid approaches as it combines linguistic and statistic information. Our reference corpus, the Reference Corpus of Contemporary Portuguese (CRPC)[1], is an electronically based linguistic corpus of around 310 million tokens, taken by sampling from several types of written texts (literature, newspapers, science, economics, law, parliamentary debates, technical and didactic documents), pertaining to national and regional varieties of Portuguese. A random selection of 10,000 texts from the entire CRPC will be used for our experiment. The experiment flow-chart is shown in Figure 1. We define as objective short news items from newspapers that reports strictly a piece of news, without comments or analysis. A selection of blog post items and short verbal exchanges between member of the European parliament will serve as subjective texts.

2 Previous work

The task of extracting linguistic patterns for data mining is not new, albeit most research has so far dealt with English texts. Extracting subjective patterns represents a more recent and challenging task. For example, in the Text Analy-

[1] http://www.clul.ul.pt/en/resources/183-reference-corpus-of-contemporary-portuguese-crpc

Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), EACL 2012, pages 46–51,
Avignon, France, April 23 2012.

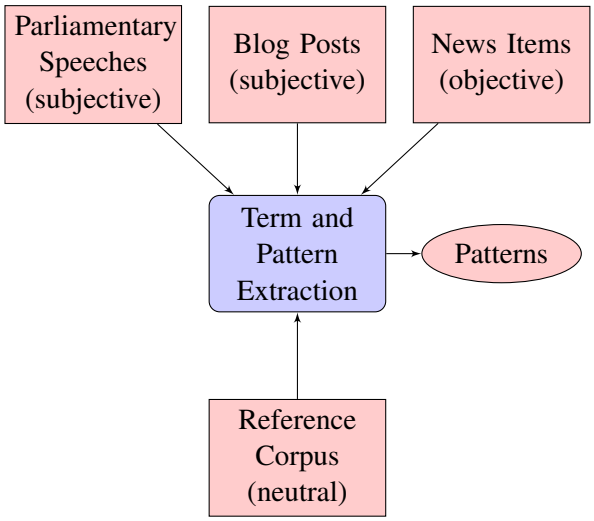

Figure 1: Experiment flow-chart.

sis Conference (TAC 2009), it was decided to withdraw the task of creating summaries of opinions, present at TAC 2008, the organizers having agreed on the difficulty of extracting subjective elements of a text and organize them appropriately to produce a summary. Yet, there is already some relevant work in this area which may be mentioned here. For opinions, previous studies have mainly focused in the detection and the gradation of their emotional level, and this involves three main subtasks. The first subtask is to distinguish subjective from objectives texts (Yu and Hatzivassiloglou, 2003). The second subtask focuses on the classification of subjective texts into positive or negative (Turney, 2002). The third level of refinement is trying to determine the extent to which texts are positive or negative (Wilson et al., 2004). The momentum for this type of research came through events such as TREC Blog Opinion Task since 2006. It is also worth mentioning recent efforts to reintroduce language and discursive approaches (e.g. taking into account the modality of the speaker) in this area (Asher and Mathieu, 2008). The approaches developed for automatic analysis of subjectivity have been used in a wide variety of applications, such as online monitoring of mood (Lloyd et al., 2005), the classification of opinions or comments (Pang et al., 2002) and their extraction (Hu an Liu, 2004) and the semantic analysis of texts (Esuli and Sebastiani, 2006). In (Mihalcea et al., 2007), a bilingual lexicon and a manually translated parallel corpus are used to generate a sentence classifier according to their level of subjectivity for Romanian. Although many recent studies in the analysis of subjectivity emphasize *sentiment* (a type of subjectivity, positive or negative), our work focuses on the recognition of subjectivity and objectivity in general. As stressed in some work (Banea et al., 2008), researchers have shown that in sentiment analysis, an approach in two steps is often beneficial, in which we first distinguish objective from subjective texts, and then classify subjective texts depending on their polarity (Kim and Hovy, 2006). In fact, the problem of distinguishing subjective versus objective texts has often been the most difficult of the two steps. Improvements in the first step will therefore necessarily have a beneficial impact on the second, which is also shown in some work (Takamura et al., 2006).

3 Creating a corpus of Subjective and Objective Portuguese Texts

To build our subjective spoken corpus (more than 2,000 texts), we used a parallel corpus of English-Portuguese speeches[2] and a tool to automatically classify sentences in English as objective or subjective (OpinionFinder (Riloff et al., 2003)). We then projected the labels obtained for the sentences in English on the Portuguese sentences. The original parallel corpus is made of 1,783,437 pairs of parallel sentences, and after removing pervasive short sentences (e.g. "the House adjourned at ...") or pairs of sentences with the ratio of their respective lengths far away from one (a sign of alignment or translation error), we are left with 1,153,875 pairs. A random selection of contiguous 20k pairs is selected for the experiment. The English sentences are submitted to OpinionFinder, which labels each of them as "unknown", "subjective" or "objective". OpinionFinder has labelled 11,694 of the 20k sentences as "subjective". As our experiment aims at comparing frequencies between texts, we have automatically created segments of texts showing lexical similarities using Textiling (Hearst, 1997), leading to 2,025 texts. We haven't made any attempt to improve or evaluate OpinionFinder and Textiling performance. This strategy is sensible as parliamentary speeches are a series of short opinionated interventions by members on specific

[2]European Parliament: `http://www.statmt.org/europarl/`

themes. The 11,694 subjective labels have been projected on each of the corresponding sentences of the Portuguese corpus to produce our final spoken corpus[3]. Note that apart from a bridge (here a parallel corpus) between the source language (here English) and the target language (here Portuguese), our approach does not require any manual annotation. Thus, given a bridge between English and the target language, this approach can be applied to other languages. The considerable amount of work involved in the creation of these resources for English can therefore serve as a leverage for creating similar resources for other languages.

We decided to include a collection of blog posts as an additional source of subjective texts. We gathered a corpus of 1,110 blog posts using BootCat[4], a tool that allows the harvesting and cleaning of web pages on the basis of a set of seed terms[5].

For our treatment of objectivity and how news are reported in Portuguese newspapers, we have collected and cleaned a corpus of nearly 1500 articles from over a dozen major websites (*Jornal de Notícias*, *Destak*, *Visão*, *A Bola*, etc.).

After tokenizing and POS-tagging all sentences, we collected all n-grams (n = 1, 2 and 3) along with their corresponding frequency for each corpus (reference (CRPC), objective (news items) and subjective (parliamentary speeches and blog posts)), each gram being a combination of a token with its part-of-speech tag (e.g. *falar*_V, "speak_V"). The list of POS tags is provided in appendix A.

[3]As our subjective spoken corpus has been built entirely automatically (Opinion Finder and Textiling), it is important to note that (Généreux and Poibeau, 2009) have verified that such a corpus correlates well with human judgements.

[4]`http://bootcat.sslmit.unibo.it/`

[5]In an attempt to collect as much opinionated pages in Portuguese as can be, we constraint BootCat to extract pages written in Portuguese from the following web domains: `communidades.net`, `blogspot.com`, `wordpress.com` and `myspace.com`. We used the following seed words, more or less strongly related to the Portuguese culture: *ribatejo*, *camões*, *queijo*, *vinho*, *cavaco*, *europa*, *sintra*, *praia*, *porto*, *fado*, *pasteis*, *bacalhau*, *lisboa*, *algarve*, *alentejo* and *coelho*.

4 Experiments and Results

4.1 POS and n-grams

In our experiments we have compared all the n-grams (n = 1, 2 and 3) from the objective and subjective texts with the n-grams from the reference corpus. This kind of analysis aims essentially at the identification of salient expressions (with high *log-odds* ratio scores). The log-odds ratio method (Baroni and Bernardini, 2004) compares the frequency of occurrence of each n-gram in a specialized corpus (news, parliamentary speeches or blogs) to its frequency of occurrence in a reference corpus (CRPC). Applying this method solely on POS, we found that objective texts used predominantly verbs with an emphasis on past participles (PPT/PPA, *adotado*, "adopted"), which is consistent with the nature of reported news. In general, we observed that subjective texts have a higher number of adjectives (ADJ, *ótimo*, "optimum"): parliamentary speeches also include many infinitives (INF, *felicitar* "congratulate"), while blogs make use of interjections (ITJ, *uau*, "wow"). Tables 1, 2 and 3 show salient expressions for each type of texts. These expressions do not always point to a distinction between subjectivity and objectivity, but also to topics normally associated with each type of texts, a situation particularly acute in the case of parliamentary speeches. Nevertheless, we can make some very general observations. There is no clear pattern in news items, except for a slight tendency towards the use of a quantitative terminology ("save", "spend"). Parliamentary speeches are concerned with societal issues ("socio-economic", "biodegradable") and forms of politeness ("wish to express/protest"). In blog posts we find terms related to opinions ("pinch of salt"), wishes ("I hope you enjoy"), reactions ("oups") and descriptions ("creamy").

4.2 Patterns around NPs

The n-gram approach can provide interesting patterns but it has its limits. In particular, it does not allow for generalization over larger constituents. One way to overcome this flaw is to chunk corpora into noun-phrases (NP). This is the approach taken in (Riloff and Wiebe, 2003) for English. In Riloff and Wiebe (2003), the patterns for English involved a very detailed linguistic analysis, such as the detection of grammatical functions as well

PORTUGUESE	ENGLISH
detetado_PPA	"detected"
empatado_PPT	"tied"
castigado_PPT	"punished"
ano_CN perdido_PPA	"lost year"
triunfa_ADJ	"triumph"
receção_CN	"recession"
podem_V poupar_INF	"can save"
vai_V salvar_INF	"will save"
deviam_V hoje_ADV	"must today"
ameaças_CN se_CL concretizem_INF	"threats materialize"
andam_V a_DA gastar_INF	"go to spend"
ano_CN de_PREP desafios_CN	"year of challenges"
contratações_CN de_PREP pessoal_CN	"hiring of staff"

Table 1: Salient expressions in news.

PORTUGUESE	ENGLISH
socioeconómicas_ADJ	"socio-economic"
biodegradveis_ADJ	"biodegradable"
infraestrutural_ADJ	"infra-structural"
base_CN jurídica_ADJ	"legal basis"
estado-membro_ADJ	"member state"
resolução_CN comun_ADJ	"common resolution"
gostaria_V de_PREP expressar_INF	"wish to express"
gostaria_V de_PREP manifestar_INF	"wish to protest"
adoptar_INF uma_UM abordagem_CN	"adopt an " approach"
agradecer_INF muito_ADV sinceramente_ADV	"thank very sincerely"
começar_INF por_PREP felicitar_INF	"start by congratulate"
senhora_CN comissária_CN	"Commissioner"
senhora_CN deputada_CN	"Deputy"
quitação_CN	"discharge"
governança_CN	"governance"

Table 2: Salient expressions in parliamentary speeches.

as active or passive forms. Without the proper resources needed to produce sophisticated linguistic annotations for Portuguese, we decided to simplify matters slightly by not making distinction of grammatical function or voice. That is, only NPs would matter for our analysis. We used the NP-chunker Yamcha[6] trained on 1,000 manually annotated (NPs and POS-tags) sentences. The main idea here remains the same and is to find a set of syntactic patterns that are relevant to each group of texts, as we did for n-grams previously, each NP becoming a single 1-gram for this purpose. It is worth mentioning that NP-chunking becomes particularly challenging in the case of blogs, which are linguistically heterogeneous and noisy. Finally, log-odds ratio once again serves as a discriminative measure to highlight relevant patterns around NPs. Tables 4, 5 and 6 illustrate salient expressions from the three specialized corpora, presenting some of them in context.

Although limited to relatively simple syntactic patterns, this approach reveals a number of salient linguistic structures for the subjective texts. In parliamentary speeches, forms of politeness are clearly predominant ("ladies and <NP>", "thank <NP>" and "<NP> wish to thank"). Unfortunately, the patterns extracted from blog posts are pervaded by "boiler-plate" material that were not filtered out during the cleaning phase and parasite the analysis: "published by <NP>", "share on <NP>" and "posted by <NP>". However, opinions ("<NP> is beautiful") and opinion primer ("currently, <NP>") remain present. News items are still characterized mainly by the absence of subjective structures (markers), albeit quantitative expressions can still be found ("spent").

Obviously, a statistical approach yields a certain number of irrelevant (or at best "counter-intuitive") expressions: our results are no exception to this reality. Clearly, in order to reveal insights or suggest meaningful implications, an external (human) evaluation of the patterns presented in this study would paint a clearer picture of the relevance of our results for information extraction and opinion mining, but we think they constitute a good starting point.

5 Conclusion and Future Work

We have presented a partly automated approach to extract subjective and objective patterns in se-

[6] `http://chasen.org/~taku/software/yamcha/`. Our evaluation of the trained chunker on Portuguese texts lead to an accuracy of 86% at word level.

PORTUGUESE	ENGLISH
direto_ADJ	"direct"
cremoso_ADJ	"creamy"
crocante_ADJ	"crispy"
atuais_ADJ	"current"
coletiva_ADJ	"collective"
muito_ADV legal_ADJ	"very legal"
redes_CN sociais_ADJ	"social networks"
ups_ITJ	"oups"
hum_ITJ	"hum"
eh_ITJ	"eh"
atualmente_ADV	"currently"
atrações_CN	"attractions"
tenho_V certeza_CN	"I am sure"
é_V exatamente_ADV	"this is exactly"
café_CN da_PREP+DA manhã_CN	"morning coffee"
pitada_CN de_PREP sal_CN	"pinch of salt"
espero_V que_CJ gostem_INF	"I hope you enjoy"

Table 3: Salient expressions in blogs.

Some NP-patterns in context
• *fiquemos_V à_PREP+DA <NP>* "we are waiting for <NP>" *E também não **fiquemos à <espera da Oposição>** mais interessada em chegar ao* Poder. "And also we are not waiting for an opposition more interested in coming to power."
• *revelam_V <NP> gastámos_V* "revealed by <NP> we spent" *O problema é que, como **revelam <os dados da SIBS, na semana do Natal> gastámos** quase 1300 euros por segundo.* "The problem is that as shown by the data of SIBS, in the Christmas week we spent nearly 1300 Euros per second."
• *<NP> deviam_V hoje_ADV* "<NP> must today" *E para evitar males maiores, <**todos os portugueses (ou quase todos)**> **deviam hoje** fazer . . .* "And to avoid greater evils, all the Portuguese (or almost all) should today make . . .
Other NP-patterns
• *<NP> gostámos_V quase_ADV* "<NP> spent almost"
• *precisa_V daqueles_PREP+DEM <NP>* "need those <NP>"

Table 4: NP-patterns in news

lected texts from the European Parliament, blog posts and on-line newspapers in Portuguese. Our work first shows that it is possible to built resources for Portuguese using resources (a parallel corpus) and tools (OpinionFinder) built for English. Our experiments also show that, despite our small specialised corpora, the resources are good enough to extract linguistic patterns that give a broad characterization of the language in use for reporting news items and expressing subjectivity in Portuguese. The approach could be favourably augmented with a more thorough cleaning phase, a parsing phase, the inclusion of larger n-grams (n $>$ 3) and manual evaluation. A fully automated daily process to collect a large-scale Portuguese press (including editorials) and blog corpora is currently being developed.

Acknowledgments

We are grateful to Iris Hendrickx from CLUL for making available the POS-tagger used in our experiments.

References

Asher N., Benamara F. and Mathieu Y. Distilling opinion in discourse: A preliminary study. In Coling 2008, posters, pages 710, Manchester, UK.

Banea C., Mihalcea R., Wiebe J. and Hassan S. Multilingual subjectivity analysis using machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Honolulu, Hawaii, October 2008.

Baroni M. and Bernardini S. Bootcat : Bootstrapping corpora and terms from the web. In Proceedings of LREC 2004, p. 1313-1316.

Esuli A. and Sebastiani F. Determining term subjectivity and term orientation for opinion mining. In EACL 2006.

Généreux M. and Poibeau T. Approche mixte utilisant des outils et ressources pour l'anglais pour l'identification de fragments textuels subjectifs français. In DEFT'09, DÉfi Fouilles de Textes, Atelier de clôture, Paris, June 22nd, 2009.

Hearst M. TextTiling: Segmenting text into multi-paragraph subtopic passages. In Computational Linguistics, pages 33–64, 1997.

Hu M. and Liu B. Mining and summarizing customer reviews. In ACM SIGKDD.

Some NP-patterns in context
• *também_ADV <NP> gostaria_V* "also <NP> would like" *Senhor Presidente ,* ***também <eu> gostaria*** *de felicitar a relatora, . . .* "Mr President, I would also like to congratulate the rapporteur, . . ."
• *senhoras_ADJ e_CJ <NP>* "ladies and <NP>" *Senhor Presidente , Senhora Deputada McCarthy,* ***Senhoras e <Senhores Deputados>****, gostaria de começar . . .* "Mr President, Mrs McCarthy, Ladies and gentlemen, let me begin . . ."
• *agradecer_INF à_PREP+DA <NP>* "thank <NP>" *Gostaria de* ***agradecer à <minha colega, senhora deputada Echerer>****, pela . . .* "I would like to thank my colleague, Mrs Echerer for . . ."
Other NP-patterns
• *<NP> desejo_V agradecer_INF* "<NP> wish to thank"
• *aguardo_V com_PREP <NP>* "I look forward to <NP>"
• *associar_INF aos_PREP+DA <NP>* "associate with <NP>"
• *considero_V ,_PNT <NP>* "I consider, <NP>"

Table 5: NP-patterns in parliamentary speeches

Some NP-patterns in context
• *publicada_V por_PREP <NP>* "published by <NP>" ***Publicada por <Joaquim Trincheiras>*** *em 07:30* "Posted by Joaquim Trenches at 07:30"
• *partilhar_INF no_PREP+DA <NP>* "share on <NP>" ***Partilhar no <Twitter>*** *. . .* "Share on Twitter " . . .
• *postado_PPA por_PREP <NP>* "posted by <NP>" ***Postado por <Assuntos de Polícia>*** *às 13:30.* "Posted by Police Affairs at 13:30."
Other NP-patterns
• *<NP> por_PREP lá_ADV* "<NP> over there"
• *<NP> deixe_V <NP>* "<NP> let <NP>"
• *atualmente_ADV ,_PNT <NP>* "currently, <NP>"
• *<NP> é_V linda_ADJ* "<NP> is beautiful"

Table 6: NP-patterns in blogs

Kim S.-M. and Hovy E. Identifying and analyzing judgment opinions. In HLT/NAACL 2006.

Lloyd L., Kechagias D. and Skiena S. Lydia: A system for large-scale news analysis. In SPIRE 2005.

Mihalcea R., Banea C. and Hassan S. Learning multilingual subjective language via cross-lingual projections. In ACL 2007.

Pang B., Lee L. and Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In EMNLP 2002.

Riloff E. and Wiebe J. Learning extraction patterns for subjective expressions. In Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing, Sapporo, JP.

Riloff E., Wiebe J. and Wilson T. Learning subjective nouns using extraction pattern bootstrapping. In W. Daelemans & M. Osborne, Eds., Proceedings of CONLL-03, 7th Conference on Natural Language Learning, p. 2532, Edmonton, CA.

Takamura H., Inui T. and Okumura M. Latent variable models for semantic orientations of phrases. In EACL 2006.

Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In ACL 2002.

Wilson T., Wiebe J. and Hwa R. Just how mad are you? Finding strong and weak opinion clauses. In Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence, p. 761-769, San Jose, US.

Yu H. and Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In EMNLP 2003.

A List of POS-tags

ADJ (adjectives), ADV (adverbs), CJ (conjunctions), CL (clitics), CN (common nouns), DA (definite articles), DEM (demonstratives), INF (infinitives), ITJ (interjections), NP (noun phrases), PNT (punctuation marks) PPA/PPT (past participles), PREP (prepositions), UM ("um" or "uma"), V (other verbs).