Aspect extraction in sentiment analysis for portuguese language

Pedro Paulo Balage Filho

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)

Thesis Defense

Supervisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo
Commitee: Prof. Dr. Diego Raphael Amancio, ICMC-USP
Prof. Dr. Osvaldo Novais de Oliveira Junior, IFSC-USP
Profa. Dra. Flávia de Almeida Barros, UFPE

Outline

Outline

Outline

Importance of Sentiment Analysis

- Liu (2010) alleges that all textual information present in the world may be categorized in only two types: **facts** and **opinions**.
- **User generated content (UGC)** is now part of our daily lives
- Websites for product reviews have become an important resource to find opinions and influence users.
- The key of this transformation is to provide new methods to convert the raw unstructured data into structured information.

Study of Sentiment Analysis

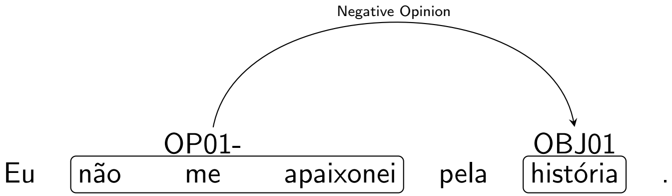
- The study of sentiment analysis is categorized in three levels: **text** level, **sentence** level, and **aspect** or entity level.
- Example for aspect level:

The Iphone is a good device. The battery is excellent. The quality is very good, but the price is not affordable.

Aspect	Evaluative Word	Sentiment
Iphone	good	Positive
battery (Iphone)	excellent	Positive
quality (Iphone)	very good	Positive
price (Iphone)	not affordable	Negative

Aspect-based Sentiment Analysis for Portuguese

- Freitas, Motta, Milidiú, and César (2012) compiled the corpus **ReLi** with book reviews.
- Example:



Research Gaps

1. Lack of understanding about opinions and sentiment analysis.
2. Difficulty to model linguistic phenomena.
3. Lack of research for Portuguese, thus few resources.

Outline

Objectives

1. To explore approaches based on frequency, relation and machine learning in aspect-based sentiment analysis and establish new benchmarks for the Portuguese.
2. To compare state of the art approaches for English with Portuguese corpora.
3. To investigate the use of syntax and semantics in Portuguese ABSA methods.
4. To develop new tools and lexicons for sentiment analysis.

Hypotheses

1. Deep linguistics knowledge such as syntax and semantics improve aspect-based sentiment analysis.
2. Aspect-based sentiment analysis approaches do not differ between English and Portuguese.
3. Corpora from different domains show different challenges.

Outline

Outline

Sentiment Analysis

Definition

Sentiment analysis, or **opinion mining**, is the area in Natural Language Processing that deals with the computational treatment of **opinion**, **sentiment** and **subjectivity** in text (Pang & Lee, 2008).

Opinion: it consists of two components, a target object and an associated sentiment.

Sentiment: generic term to designate every text expressing positive, negative, or neutral characteristics.

Subjectivity: the presence in the text of sentiment, points of view or personal beliefs.

Emotion: our subjective feelings and thoughts (Liu, 2012).

Opinion

- Formally, a sentiment or opinion, is defined by (Liu, 2010) as a quintuple (o_j, f_{jk}, oo_{ijkl}, h_i, l_l) where
  - o_j is an object,
  - f_{jk} is a feature of the object o_j,
  - oo_{ijkl} is the semantic orientation or polarity of the opinion on feature f_{jk} of object o_j,
  - h_i is the opinion holder and t_l is the time when the opinion is expressed by h_i.

Opinion

- Opinion can be classified into two types:

direct opinions: are those in which we have an evaluation or sentiment about an aspect present in an object referred into the text.

comparative opinion: expresses a relation between two or more objects.

Opinion

- A direct opinion may still take the explicit or implicit form.

Explicit direct opinion: the opinion is expressed explicitly in the sentence.

Implicit direct opinion: an inference of context and world knowledge is required to understand the expressed opinion.

Opinion

- Aspects can appear in two forms in the text: explicit aspects and implicit aspects.

Explicit aspects: aspects that are present in text.

Implicit aspects: aspects that are only perceptible through inference.

Outline

Lexical Generation

- Dictionary-based approaches (Hu & Liu, 2004; Kim & Hovy, 2004).
- Propagation approach exploits the relations between sentiment words and topics or product features (Qiu, Zhang, Hu, & Zhao, 2009).

Sentiment Classification

- Pang, Lee, and Vaithyanathan (2002) performed supervised classification in a movie reviews dataset.
- Wilson, Wiebe, and Hoffmann (2009) study features for supervised machine learning.
- Taboada, Brooke, Tofiloski, Voll, and Stede (2011) use a lexicon method to determine the polarity, or semantic orientation, for the individual words in the text.

Sentiment Analysis for Portuguese

- Elaboration of corpus (Sarmento, Carvalho, Silva, & De Oliveira, 2009; Carvalho, Sarmento, Silva, & de Oliveira, 2009; Scopim et al., 2012)
- Sentiment lexicons (Silva, Carvalho, Costa, & Sarmento, 2010; Pasqualotti, 2008; Souza et al., 2011; Balage Filho, Pardo, & Aluísio, 2013)
- Sentiment classification at the document level (Afonso et al., 2011; Amancio, Fabbri, Oliveira Jr, Nunes, & da F Costa, 2010; Souza & Vieira, 2012)

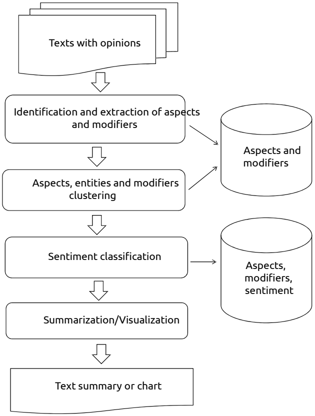
Outline

Definitions

- According with Liu (2010), the aspect-based sentiment analysis is composed by three main tasks:

1. **Aspect extraction**: task responsible for extracting aspects and their modifiers.
2. **Group entity, aspects and modifiers**: task consists in grouping entities, aspects, and modifiers.
3. **Sentiment classification**: determine whether the opinion is positive, negative or neutral.

Task


Texts with opinions
Identification and extraction of aspects and modifiers
Aspects and modifiers
Aspects, entities and modifiers clustering
Sentiment classification
Aspects, modifiers, sentiment
Summarization/Visualization
Text summary or chart

Definitions

- According to Pontiki et al. (2014), the activity of aspect-based sentiment analysis is analyzed for different tasks on different datasets from different perspectives.
- They propose a joint assessment to be held in **Semantic Evaluation Workshop** (SemEval) comprising four subtasks:
  1. **Aspect Term Extraction** (Opinion Target Extraction)
  2. Aspect Term Polarity
  3. Aspect Category Detection
  4. Aspect Category Polarity

Outline

Aspect Term Extraction

- There are different techniques for aspect extraction (Liu, 2012).
    1. Frequency-based
    2. Relation-based
    3. Machine learning
    4. Topic modeling

Aspect Term Extraction

- There are different techniques for aspect extraction (Liu, 2012).
    1. **Frequency-based**
    2. Relation-based
    3. Machine learning
    4. Topic modeling

Frequency-based works

- Hu and Liu (2004) extract the most frequent nouns and noun phrases as candidate for aspects.
- Popescu and Etzioni (2005) remove candidates based on the Pointwise Mutual Information (PMI) relation between the aspect and the candidates.
- Blair-goldensohn et al. (2008) refined the candidates to the ones which obeying certain syntactic patterns.

Aspect Term Extraction

- There are different techniques for aspect extraction (Liu, 2012).
  1. Frequency-based
  2. **Relation-based**
  3. Machine learning
  4. Topic modeling

Relation-based works

- Hu and Liu (2004) collects infrequent aspects associated with words of sentiment.
- Zhuang, Jing, and Zhu (2006) uses a dependency parser to indicate the relationship between aspects and their targets.

Aspect Term Extraction

- There are different techniques for aspect extraction (Liu, 2012).
  1. Frequency-based
  2. Relation-based
  3. **Machine learning**
  4. Topic modeling

Machine learning for aspect extraction

- Jin, Ho, and Srihari (2009) apply a hidden and lexicalized Markov chain to learn patterns to extract expressions of aspects and opinions.
- Jakob and Gurevych (2010) trained a Conditional Random Field (CRF) for the extraction of aspects and opinions.

Aspect Term Extraction

- There are different techniques for aspect extraction (Liu, 2012).
  1. Frequency-based
  2. Relation-based
  3. Machine learning
  4. **Topic modeling**

Topical modeling for aspect extraction

- Mei, Ling, Wondra, Su, and Zhai (2007) propose a joint model based on Probabilistic Latent Semantic Analysis (pLSA) for the extraction of aspects and sentiments.
- Li, Huang, and Zhu (2010) proposed two joint models for extraction, one for sentiment and the other for aspects that depend on this sentiment.

Outline

Aspect-based Sentiment Analysis for Portuguese

- Carvalho, Sarmento, Teixeira, and Silva (2011) describes the construction of a corpus in the political domain with the annotation of opinions and their targets.
- Silva, Carvalho, Sarmento, Magalhães, and Oliveira (2009) developed the system OPTIMISM to recognize and analyse the sentiment towards entities in the political domain.

Aspect-based Sentiment Analysis for Portuguese

- Chaves, de Freitas, Souza, and Vieira (2012) presents the PIRPO tool for rating sentiment in online evaluations for the hotel sector. It uses an ontology in the hotel domain and a lexicon of sentiment (Souza et al., 2011) for extracting and classifying aspects contained in the text.
- Ribeiro, Junior, Meira, and Pappa (2012) presents a polarity classification system for aspects in vehicle evaluation texts using lexical- and machine learning-based classifiers. The extraction of aspects was not covered by the work.

Aspect-based Sentiment Analysis for Portuguese

- Fernandes (2010) presents the HowGood tool that performs the analysis of sentiment at the aspect level using frequent nouns using manually filtered aspects of interest
- da Silva (2010) adapted the previous system using the SentiWordNet (Esuli & Sebastiani, 2006) lexic using the Google Translate tool to translate the terms from English to Portuguese.

Aspect-based Sentiment Analysis for Portuguese

- Siqueira and Barros (2010) present a process of extraction of aspects in the analysis of sentiment for texts in Portuguese in the e-commerce domain. The WhatMatter system, described by them, performs four steps: identifies frequent nouns, identifies relevant nouns, maps aspect indicators, and removes unrelated nouns.

Aspect-based Sentiment Analysis for Portuguese

- Nobre et al. (2016) shows intial studies for an aspect-based sentiment analysis system in the e-commmerce domain using the machine learning approach with a conditional random field following the approach taken by (Balage Filho & Pardo, 2014).
- Vargas and Pardo (2017) studies and propose an ontology to group aspects in Portuguese.

Outline

State-of-art for SemEval ABSA 2015 - Opinion Target Expression (OTE)

- San Vicente, Saralegi, and Agerri (2015) present the EliXa system first ranked in the OTE task in the 2015 edition, with 70.05% for f-score.
- This system used machine learning based on the *Averaged Perceptron* algorithm with the following machine learning characteristics: n-grams; part-of-speech label; n-grams of prefixes and suffixes; Brown clusters and word embeddings.

State-of-art for SemEval ABSA 2015 - Opinion Target Expression (OTE)

- Toh and Su (2015) show the NLANGP system, the second highest score for the OTE task in the 2015 edition, reached the f-score of 67.11%.
- This system was based on the machine learning algorithm Conditional Random Fields with the following characteristics: the word itself; the head of the syntactic constituents (obtained from a dependency parser); lists of names (extracted based on frequency from a corpus); and Brown clusters.

State-of-art for SemEval ABSA 2016 - Opinion Target Expression (OTE)

- Toh and Su (2016) present the system NLANGP that was the best system in the OTE task with the f-measure score of 72.34%.
- The system, an enhancement of the same system that won the second place in the previous edition of the competition, brings the addition of a new learning feature based on the extracted probability of a recurrent neural network.

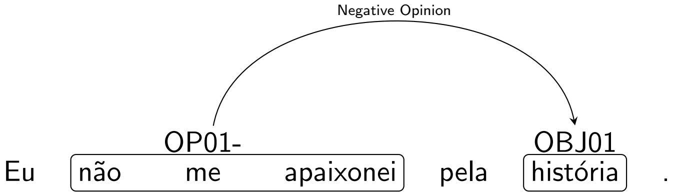
State-of-art for SemEval ABSA 2016 - Opinion Target Expression (OTE)

- Xenos, Theodorakakos, Pavlopoulos, Malakasiotis, and Androutsopoulos (2016) present the system AUEB, second in the OTE task in the 2016 edition, reached the score of 70.44% through a system based on the algorithm *Conditional Random Fields* with the following set of characteristics: part-of-speech tags; lexicon, list of aspects and word embeddings.

Outline

ReLi

- We use the corpus **ReLi** (Freitas et al., 2012) extracted from a social plataform for sharing opinion on books.
- Composed of 2,056 reviews from 13 different books with about 200 comments each.



ReLi

- According to Freitas et al. (2012), the main difficulty in the process of noting the corpus was to distinguish subjective information from factual information.

Tabela: Distribution of expressions of opinion according to the number of word (Freitas et al., 2012)

N-gram size	Frequency
1-3	69%
4-6	15%
7+	15%

SemEval ABSA Dataset

- Restaurant and Laptop domains.
- 1000 review texts (approx., 6K sentences) with fine-grained human annotations (opinion target expressions, aspect categories and polarities).

PALAVRAS

- The syntax parser PALAVRAS (Bick, 2000) allows automatic part-of-speech tagging and syntactical analysis of texts in Portuguese.
- The analyzer outputs a tokenized text, the part-of-speech annotations, a dependency parsing of each sentence and, a semantic type for some words.

Outline

Methodology

- Approch ReLi and SemEval datasets with classic and advanced methods for aspect extraction.
- Explore frequency, relation and machine learning-based approaches.
- Study how the inclusion of syntax and semantic could help the algorithms.
- Release new tools for the community.

Outline

Basic steps for Aspect Extraction for Frequency-based

1. Compute all word frequencies tagged as aspects in the training set.
2. Select a threshold for cut.
3. Tag all the aspects which repeat in the test set.

Aspect Extraction for Frequency-based

Tabela: Distribution of aspects in the ReLi corpus

Extracted aspects	Word frequency in Corpus	Relative Freq. word in corpus	Freq. as aspect	Relative Freq. as aspect
livro (book)	2779	1.07%	916	33.0%
história (story)	864	0.33%	208	24.1%
leitura (reading)	409	0.16%	112	27.4%
personagens (story characters)	321	0.12%	85	26.5%
crepúsculo (Twilight)	260	0.10%	62	23.8%
narrativa (narrative)	141	0.05%	61	43.3%
final (final)	193	0.07%	57	29.5%
romance (romance)	274	0.11%	55	20.1%
obra (book)	251	0.10%	48	19.1%
ele (him)	1053	0.40%	43	4.1%

Aspect Extraction for Frequency-based

Tabela: ReLi Aspect Extraction Scores for Frequency Methods

Run	ReLi corpus – Aspect Extraction		
	Precision	Recall	F-measure
All aspects	7,14%	**82,26%**	13,13%
Stopwords cut	14,32%	79,11%	24,25%
Frequency cut	30.27%	55.51%	39,17%
Relative frequency cut	**36,44%**	78,25%	**49,73%**

Aspect Extraction for Frequency-based

Tabela: Distribution of aspects in the SemEval ABSA 2015 corpus

Extracted aspects	Word frequency in Corpus	Relative Freq. word in corpus	Freq. as aspect	Relative Freq. as aspect
food	190	1.03%	158	83.2%
service	127	0.69%	117	92.1%
place	135	0.73%	82	60.7%
restaurant	82	0.44%	29	35.4%
staff	33	0.18%	27	81.8%
pizza	42	0.23%	26	61.9%
atmosphere	26	0.14%	21	80.8%
sushi	32	0.17%	20	62.5%
decor	19	0.10%	16	84.2%
ambience	13	0.07%	13	100.0%

Aspect Extraction for Frequency-based

Tabela: Extraction results of aspects in the SemEval ABSA 2016 corpus using frequency based methods

Method	Precision	Recall	F-score
1. All aspects	50,88%	**62,31%**	56,02%
2. Stopwords cut	50,88%	**62,31%**	56,02%
3. Frequency cut	50,88%	**62,31%**	56,02%
4. Relative frequency cut	**60,35%**	58,77%	**59,55%**

Observations

- The relative frequency of the word in the SemEval ABSA corpus is higher in comparision to the ReLi corpus.
- This indicates that the words labeled as aspects in the ReLi corpus are used in a more varied manner and therefore lead to greater difficulty in the extraction process.
- For comparison to the state of the art, only possible in the SemEval ABSA corpus, the best system in the 2015 competition obtained an f-measure of 70.00% (San Vicente et al., 2015) while with frequency-based approaches achieved **59,55%**
- This observation leads us to consider that, even though they are simple methods, the frequency-based methods have good results.

Outline

Relatation-based approach

- **patterns in extracting aspects** of texts is presented in (Nasukawa & Yi, 2003).
- Patterns involve the relation between the verb, aspect and opinion.
- In out approach, we extracted patterns

Sentence	O livro é bom	Eu adorei o livro
Learned Patterns		
Verb	*ser*	*adorar*
Aspect	*livro*	*livro*
Opinion	*bom*	*adorar*
Polarity	*positiva*	*positiva*

Relatiton-based approach

Tabela: Extraction results of aspects in the ReLi corpus using the method of (Nasukawa & Yi, 2003)

Method	Precision	Recall	F-score
Patterns with lemma	**22,60%**	**17,85%**	**19,94%**
Patterns with PoS	8,30%	17,85%	11,33%

Outline

Machine Learning methods

- The methods based on machine learning allow to feed the learning algorithms with sets of features extracted from the text and let the algorithm itself decide the most important set of characteristics for each case.
- The learning algorithms that best fit the problem of aspect extraction are the algorithms belonging to the class of **sequential learning**. Examples of tasks inside this class are: part-of-speech tagging (Silfverberg, Ruokolainen, Lindén, & Kurimo, 2014), shallow parsing (Sha & Pereira, 2003), entity recognition (Finkel, Grenager, & Manning, 2005), among others.

Machine Learning methods

- The biggest challenge resides in **feature engineering**.
- experiments are reported using the CRF algorithm through the framework **CRFSuite** (Okazaki, 2007), used in conjunction with the machine learning library **Scikit-Learn** (Pedregosa et al., 2011).

Features Evaluated

word: we use the word or token extracted as a form of lexicalization of the learning model;

lemma: the use of the lemma (as opposed to the word) brings generalization to learning, which usually improves learning for morphologically rich languages;

Part-of-Speech : The PoS label enriches learning by generalizing the word by its function;

Features Evaluated

dependency relation: the syntactic relation obtained through a dependency parser add context to the learning model. It provides the syntactic function for the word in the sentence;

head of the syntactic relation: The head of a syntactic relation is the word that governs the syntactic relation towards the word in analysis. With this, we were able to capture the modification and effect relationships between words in the sentence, thus adding context.

Features Evaluated

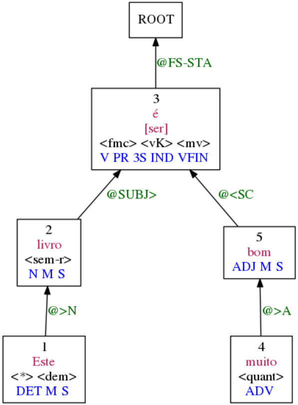


Figura: Example of syntatic parser produced by PALAVRAS

Features Evaluated

hierarchical clusters of words: the clusters used here are extracted through the methodology defined by (Brown, Desouza, Mercer, Pietra, & Lai, 1992); In this technique, words are organized hierarchically into clusters according to their meaning obtained through the context of their use in a corpus;

Features Evaluated

word representation vectors: word representation vectors capture context and semantic relations between words expressed by their numerical values within a vector space of representation. The generation methodology of the representation vectors adopted here is based on the work of (Mikolov & Dean, 2013), known as **Word2Vec**;

Features Evaluated

In order to obtain clusters and representation vectors using the Word2Vec technique, a corpus of book reviews was automatically built by crawling the Skoob.com website, which was the source for the elaboration of the text. ReLi corpus. This corpus consists of 343,000 reviews representing the entire collection of reviews of the site as of November 20, 2015.

Features Evaluated

semantic category from parser PALAVRAS: the syntactic parser PALAVRAS provides together with its analysis the semantic category of some words. This information is tied to its grammar used for sentential analysis.

Features Evaluated

Tabela: Semantic tags present in the ReLi corpus annotated by the parser PALAVRAS

Label	Explanation	Freq.	Relative Freq.	Exemples
sem-r	que pode ser lido	5573	9.34%	livro(50.3%), história(13.6%), leitura(5.9%), romance(3.9%)
am	abstrato	3634	6.09%	amor(6.4%), tempo(6.2%), partido(5.9%), poder(3.1%), atenção(2.2%)
ac	abstrato contável	3338	5.60%	amor(7.0%), coisa(10.3%), parte(4.9%), verdade(4.2%)
per	período do tempo	2889	4.84%	história(26.3%), vida(14.8%), ano(7.9%), tempo(7.8%), dia(3.5%)
sem-c	produto da cognição	2075	3.48%	obra(10.2%), fim(5.8%), visão(4.4%), trama(4.0%), opinião(3.7%)
HH	grupo de humanos	1830	3.07%	sociedade(9.9%), parte(9.0%), grupo(7.2%), família(6.3%), governo(4.5%)
H	humano	1790	3.00%	pessoas(22.3%), amor(13.0%), criança(4.9%), tipo(4.7%)
temp	temporal	1710	2.87%	ano(13.4%), tempo(13.1%), final(8.9%), vez(8.4%), fim(7.1%)
percep-f	que pode ser sentido	1667	2.79%	forma(15.2%), verdade(8.5%), realidade(7.7%), nome(4.9%), pena(4.7%)
act	ação	1298	2.18%	ação(3.2%), carinho(2.5%), geração(2.3%), manipulação(1.8%), prisão(1.8%)

Results

Tabela: Results of machine learning using CRF in the ReLi corpus with PALAVRAS Parser

Experiment	Features	Precision	Recall	F-score
1	Word	57,30%	15,50%	24,40%
2	Word+PoS	56,90%	15,40%	24,20%
3	Lemma+PoS+Head	58,50%	20,10%	29,90%
4	Lemma+PoS+Head+Sem	**60,40%**	**24,70%**	**35,10%**
5	Lemma+Pos+Head+Sem+ clusters+Word2Vec	**60,40%**	**24,70%**	**35,10%**

Dependency Parser using the Universal TreeBank

- The Universal Dependencies have been used for many tasks in Natural Language Processing (Manning et al., 2014).
- The MaltParser was trained using the Universal Dependency Corpus (Nivre et al., 2016).

Results

Tabela: Results of machine learning using CRF in the ReLi corpus with syntactic annotations of Universal TreeBank

Experiment	Features	Precision	Recall	F-score
1	Word	**59,00%**	15,10%	24,10%
2	Word+PoS	54,40%	16,10%	24,80%
3	Lemma+PoS+Head	57,10%	**20,00%**	**29,70%**

Results

Tabela: Results of machine learning using CRF in the corpus SemEval ABSA 2016

Experiment	Features	Precision	Recall	F-score
1	Word	77,20%	46,80%	58,30%
2	Word+PoS	75,30%	55,40%	63,80%
3	Lemma+PoS+Head	**78,00%**	57,50%	66,20%
4	AUEB (Xenos et al., 2016)	71,82%	69,12	70,44%
5	NLANGP (Toh & Su, 2016)	75,49%	**69,44%**	**72,34%**

Discussion

- Results for SemEval show the machine learning could achieve a score closer to the state of the art.
- Results for the ReLi show the frequency methods are extremely efficient.
- This difference is possibly due to the annotation criteria and the domain/genre of the reviews.

Outline

Objectives

1. To explore approaches based on frequency, relation and machine learning in aspect-based sentiment analysis and establish new benchmarks for the Portuguese.
2. To compare state of the art approaches for English with Portuguese corpora.
3. To investigate the use of syntax and semantics in Portuguese ABSA methods.
4. To develop new tools and lexicons for sentiment analysis.

Hypotheses

1. Deeper linguistics knowledge such as syntax and semantics improve aspect-based sentiment analysis.
    - Confirmed. PALAVRAS semantic tags helped to achieve a better score.
2. Aspect-based sentiment analysis approaches do not differ between English and Portuguese.
    - The methods could be easily used interchangeably between languages.
3. Corpora from different domains show different challenges.
    - The differences in domain and aspect distribution impact the choice of the method.

Contributions

This research, unprecedented for the Portuguese language, has resulted in **relevant contributions** to the area of research, both theoretical and practical. Some of these are:

1. The exploration of methods based on the frequency and proposal of a variation that overcame the classical methods of this approach;
2. The exploration of a classic method based on relation and the proposal of automation of its application, by learning automatic patterns of occurrence of aspects;
3. The research of linguistic standards in aspect-based sentiment analysis in Portuguese;

Contributions

This research, unprecedented for the Portuguese language, has resulted in **relevant contributions** to the area of research, both theoretical and practical. Some of these are:

5. The exploration of methods based on machine learning and its enrichment with linguistic information of a syntactic and semantic nature, producing better results than the original methods;
6. Linguistic characterization of semantic nature of the most frequent aspects in Portuguese language;
7. UTB-based syntactic parser training, providing a new tool for the research area;
8. Pre-processing and availability of the ReLi corpus with syntactic and semantic information.

Outline

Publications

- Pedro Paulo Balage Filho, Thiago Pardo, e Sandra Aluísio. 2013. **An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis.** Em Sandra Maria Aluísio e Valéria Delisandra Feltrim, editores, Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL). Sociedade Brasileira de Compu- tação, Fortaleza-CE, Brazil, páginas 215–219

Publications

- Pedro Paulo Balage Filho e Thiago Alexandre Salgueiro Pardo. 2013. **NILC_USP: A hybrid system for sentiment analysis in twitter messages.** Em Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, páginas 568–572

Publications

- Pedro Paulo Balage Filho e Thiago Alexandre Salgueiro Pardo. 2014b. **NILC_USP: Aspect extraction using semantic labels.** Em Preslav Nakov e Torsten Zesch, editores, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for 29Análise de Sentimentos Orientada a Aspectos para o Português Computational Linguistics and Dublin City University, Dublin, Ireland, páginas 433–436

Publications

- Pedro Paulo Balage Filho e Thiago Alexandre Salgueiro Pardo. 2014a. **BuscaOpinioes: Searching for opinions over the internet.** Em Proceedings of the 11th International Conference on Computational Processing of Portuguese Language. Software Demonstration. São Carlos-SP, Brazil, páginas 1–3

Publications

- Nathan Siegle Hartmann, Lucas Vinicius Avanço, Pedro Paulo Balage Filho, Magali Duran, Maria das Graças Volpe Nunes, Thiago Alexandre Salgueiro Pardo, e Sandra Maria Aluísio. 2014. **A large opinion corpus in portuguese: Tackling out-of-vocabulary words.** Em Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland, páginas 3865–3871

Publications

- Roque Lopez, Thiago Pardo, Lucas Avanço, Pedro Paulo Balage Filho, Alessandro Bokan, Paula Cardoso, Márcio Dias, Fernando Nóbrega, Marco Cabezudo, Jackson Souza, Andressa Zacarias, Eloize Seno, e Ariani Di Felippo. 2015. **A qualitative analysis of a corpus of opinion summaries based on aspects.** Em Proceedings of The 9th Linguistic Annotation Workshop. Association for Computational Linguistics, Denver, Colorado, USA, páginas 62–71

Publications

Obrigado.

Aspect extraction in sentiment analysis for portuguese language

Pedro Paulo Balage Filho

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)

Thesis Defense

Supervisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo
Commitee: Prof. Dr. Diego Raphael Amancio, ICMC-USP
Prof. Dr. Osvaldo Novais de Oliveira Junior, IFSC-USP
Profa. Dra. Flávia de Almeida Barros, UFPE

References I

Afonso, B., Batista, D., Bovnjak, M., Carvalho, P., Correia, P., Couto, F., . . . Teixeira, J. (2011, May). *Notas sobre a Realização e Qualidade do Twitómetro* (Tech. Rep.). University of Lisbon, Faculty of Sciences,LASIGE.

Amancio, D. R., Fabbri, R., Oliveira Jr, O. N., Nunes, M. G., & da F Costa, L. (2010). Distinguishing between positive and negative opinions with complex network features. In *Proceedings of the 2010 workshop on graph-based methods for natural language processing* (pp. 83–87).

References II

Balage Filho, P. P., & Pardo, T. A. S. (2014, 23–24 August). NILC_USP: Aspect extraction using semantic labels. In P. Nakov & T. Zesch (Eds.), *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)* (pp. 433–436). Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Balage Filho, P. P., Pardo, T., & Aluísio, S. (2013, 21–23 October). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In S. M. Aluísio & V. D. Feltrim (Eds.), *Proceedings of the 9th brazilian symposium in information and human language technology (stil)* (pp. 215–219). Fortaleza-CE, Brazil: Sociedade Brasileira de Computação.

References III

Bick, E. (2000). *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework*. University of Arhus.

Blair-goldensohn, S., Neylon, T., Hannan, K., Reis, G. A., Mcdonald, R., & Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In *Nlp in the information explosion era.*

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467–479.

References IV

Carvalho, P., Sarmento, L., Silva, M. J., & de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceeding of the 1st international cikm workshop on topic-sentiment analysis for mass opinion* (pp. 53–56). ACM. Retrieved from `http://dl.acm.org/citation.cfm?id=1651471`

Carvalho, P., Sarmento, L., Teixeira, J., & Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2* (pp. 564–568). Portland, Oregon: Association for Computational Linguistics. Retrieved from

References V

http://xldb.lasige.di.fc.ul.pt/xldb/publications/Carvalho:LiarsAndSaviors:2011_document.pdf

Chaves, M. S., de Freitas, L. A., Souza, M., & Vieira, R. (2012). Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. In *International conference on application of natural language to information systems* (pp. 296–301).

da Silva, N. G. R. (2010). *Bestchoice: Classificação de sentimento em ferramentas de expressão de opinião* (Tese de graduação,). Universidade Federal de Pernambuco, Recife, 2010. 7, 17.

References VI

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings the 11th meeting of the european chapter of the association for computational linguistics (eacl-2006)* (Vol. 6, pp. 417–422). Citeseer. Retrieved from `http://gandalf.aksis.uib.no/lrec2006/pdf/384_pdf.pdfhttp://acl.ldc.upenn.edu/eacl2006/main/papers/13_1_esulisebastiani_192.pdf`

Fernandes, F. M. d. M. (2010). *Um Framework para Análise de Sentimento em Comentários sobre Produtos em Redes Sociais.* Recife-PE, Brasil.

References VII

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)* (pp. 363–370). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/P05-1045` doi: 10.3115/1219840.1219885

Freitas, C., Motta, E., Milidiú, R., & César, J. (2012). Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. *ENCONTRO DE LINGUÍSTICA DE CORPUS*, *11*, 3.

References VIII

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 168–177).

Jakob, N., & Gurevych, I. (2010). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1035–1045). Retrieved from `http://portal.acm.org/citation.cfm?id=1870759`

References IX

Jin, W., Ho, H. H., & Srihari, R. K. (2009). OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1195–1203). New York, New York, USA: ACM New York, NY, USA. Retrieved from `http://portal.acm.org/citation.cfm?id=1557019.1557148` doi: 10.1145/1557019.1557148

Kim, S., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on computational linguistics* (pp. 1367–es).

References X

Li, F., Huang, M., & Zhu, X. (2010). Sentiment Analysis with Global Topics and Local Dependency. In M. Fox & D. Poole (Eds.), *Aaai*. AAAI Press. Retrieved from http://dblp.uni-trier.de/db/conf/aaai/aaai2010.html#LiHZ10

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*, 627–666.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers. Retrieved from https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf doi: 10.2200/S00416ED1V01Y201204HLT016

References XI

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Acl (system demonstrations)* (pp. 55–60).

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on world wide web* (pp. 171–180).

Mikolov, T., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

References XII

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on knowledge capture* (pp. 70–77).

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., … others (2016). Universal dependencies v1: A multilingual treebank collection. In *Lrec.*

Nobre, G., Justino, A., Tadao, F., Nunes, D., Takabayashi, D., & Küllian, R. (2016). Booviews: Aspect-based sentiment analysis on product reviews combining svm and crf in portuguese. In *Proceedings of the first student research workshop* (pp. 1–6).

References XIII

Okazaki, N. (2007). *Crfsuite: a fast implementation of conditional random fields (crfs).* Retrieved from `http://www.chokkan.org/software/crfsuite`

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the acl-02 conference on empirical methods in natural language processing - volume 10* (pp. 79–86). Morristown, NJ, USA: Association for Computational Linguistics. Retrieved from

References XIV

http://dx.doi.org/10.3115/1118693.1118704 doi: 10.3115/1118693.1118704

Pasqualotti, P. R. (2008). *Reconhecimento de expressões de emoções na interação mediada por computador* (Master's thesis, UNIVERSIDADE DO VALE DO RIO DOS SINOS). Retrieved from http://www.dominiopublico.gov.br/download/texto/cp057721.pdfhttp://bdtd.unisinos.br/tde_arquivos/1/TDE-2008-09-02T064908Z-574/Publico/PauloRobertoPasqualotti.pdf

References XV

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, *12*, 2825–2830.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014, August). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)* (pp. 27–35). Dublin, Ireland: Association for Computational Linguistics and Dublin City University. Retrieved from `http://www.aclweb.org/anthology/S14-2004`

References XVI

Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 339–346. Retrieved from http://dl.acm.org/citation.cfm?id=1220618

Qiu, L., Zhang, W., Hu, C., & Zhao, K. (2009). SELC : A Self-Supervised Model for Sentiment Classification. *Information Storage and Retrieval*, *26*(6), 929–936. Retrieved from http://portal.acm.org/citation.cfm?id=1645953.1646072 doi: 10.1016/S0010-4825(96)00034-0

References XVII

Ribeiro, S., Junior, Z., Meira, W., & Pappa, G. L. (2012). Positive or negative? using blogs to assess vehicles features. *Encontro Nacional de Inteligência Artificial*, 1–12.

San Vicente, I., Saralegi, X., & Agerri, R. (2015, June). Elixa: A modular and flexible absa platform. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)* (pp. 748–752). Denver, Colorado: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/S15-2127`

References XVIII

Sarmento, L., Carvalho, P., Silva, M. J., & De Oliveira, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. In *Proceedings of the 1st international cikm workshop on topic-sentiment analysis for mass opinion* (pp. 29–36).

Scopim, D., et al. (2012). *Estudo de padrões lexicais em textos opinativos* (Mestrado em Linguística)). Universidade Federal de São Carlos.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1* (pp. 134–141).

References XIX

Silfverberg, M., Ruokolainen, T., Lindén, K., & Kurimo, M. (2014, June). Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 259–264). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/P14-2043`

Silva, M. J., Carvalho, P., Costa, C., & Sarmento, L. (2010). *Automatic expansion of a social judgment lexicon for sentiment analysis* (Tech. Rep.). University of Lisbon Faculty of Sciences LASIGE.

References XX

Silva, M. J., Carvalho, P., Sarmento, L., Magalhães, P., & Oliveira, E. (2009, October). The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics. *New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence*, 565–576. Retrieved from `http://paginas.fe.up.pt/~niadr/PUBLICATIONS/2009/epia2009-OPTIMISM-submitted.pdfhttp://epia2009.web.ua.pt/onlineEdition/565.pdf`

Siqueira, H., & Barros, F. (2010). A feature extraction process for sentiment analysis of opinions on services. In *Proceedings of international workshop on web and text intelligence.* São Bernardo do Campo-SP, Brazil.

References XXI

Souza, M., & Vieira, R. (2012). Sentiment analysis on twitter data for portuguese language. In *International conference on computational processing of the portuguese language* (pp. 241–247).

Souza, M., Vieira, R., Busetti, D., Chishman, R., Alves, I. M., et al. (2011). Construction of a portuguese opinion lexicon from multiple resources. In *8th brazilian symposium in information and human language technology* (pp. 59–66).

References XXII

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011, June). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. Retrieved from http://www.sfu.ca/~mtaboada/docs/Taboada_etal_SO-CAL.pdfhttp://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00049 doi: 10.1162/COLI_a_00049

Toh, Z., & Su, J. (2015, June). Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)* (pp. 496–501). Denver, Colorado: Association for Computational

References XXIII

Linguistics. Retrieved from
http://www.aclweb.org/anthology/S15-2083

Toh, Z., & Su, J. (2016, June). Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 282–288). San Diego, California: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/S16-1045

Vargas, F. A., & Pardo, T. A. S. (2017, march). *Estudo Empírico sobre Agrupamento e Organização Hierárquica de Aspectos para Mineração de Opinião.* (Tech. Rep.). São Carlos - SP.

References XXIV

Wilson, T., Wiebe, J., & Hoffmann, P. (2009, September 1). Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, *35*(3), 399–433. doi: 10.1162/coli.08-012-R1-06-90

Xenos, D., Theodorakakos, P., Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2016, June). Aueb-absa at semeval-2016 task 5: Ensembles of classifiers and embeddings for aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 312–317). San Diego, California: Association for Computational Linguistics.

References XXV

Retrieved from
http://www.aclweb.org/anthology/S16-1050

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th acm international conference on information and knowledge management* (pp. 43–50).