A Computational Approach to the Study of Portuguese Newspapers Published in Macau

Marcos Zampieri[1,2], Shervin Malmasi[3], Octavia-Maria Şulea[1,2,4], Liviu P. Dinu[4]
Saarland University, Germany[1]
German Research Center for Artificial Intelligence (DFKI)[2]
Harvard Medical School, Boston, MA, USA[3]
University of Bucharest, Romania[4]

Abstract

This paper investigates the application of text classification methods to investigate diatopic variation in Portuguese journalistic texts. We compare the language used in Portuguese newspapers written in Brazil, Macau, and Portugal under the assumption that the more similar language varieties are, the more difficult it is for algorithms to discriminate between them. We present two sets of experiments: in the first one we use original texts and in the second one we use texts with blinded named entities to remove country-specific expressions. Our results indicate that the language of Portuguese newspapers published in Macau is substantially more similar to the language used in European newspapers than that used in Brazilian newspapers.

1 Introduction

Portuguese is a pluricentric language (co-)official language in nin countries and in Macau, a special administrative region of China with a population of around 650,000 people. Portuguese is spoken by a minority of the Macanese population (between 5% and 10%). It coexists with Cantonese, spoken by over 90% of the population, and Macanese, a Portuguese-based creole. In Macau, Portuguese is used for official communication in street signs, official documents, and media including a Lusophone TV channel, radios, and newspapers.

In this paper we propose the use of text classification methods to study the language used in Portuguese newspapers published in Macau in comparison to newspapers published in other Lusophone countries, namely: Brazil and Portugal. There have been a number of studies on the differences between Portuguese language varieties and on Portuguese and Portuguese-based creoles in Macau [Baxter, 1992; 1996; Amaro, 2016], however, to the best our knowledge, no study has been carried out in order to investigate the current language of journalism in Macau. A similar study [Zampieri and Gebre, 2012] has shown that Brazilian and European newspaper texts use substantially different language and that a system trained on character and word n-grams can distinguish between them with 99.8% accuracy.

Our work is related to recent studies which apply text classification methods for discriminating between texts written in different national language varieties or dialects [Lui and Cook, 2013; Maier and Gómez-Rodrıguez, 2014; Malmasi and Dras, 2015a; Malmasi *et al.*, 2015]. It has been argued that such experiments are useful to level out differences between corpora for further linguistic analysis [Zampieri *et al.*, 2013; Ciobanu and Dinu, 2016]. We agree with this claim and we analyze the most informative lexical features used in our experiments in Section 4.1.

The question we aim to answer in this paper is:

- Are there substantial differences between the language used in newspapers published in Macau and those published in other Lusophone countries?

In addition to the historical cooperation and exchange between Macau and Portugal in areas such as trade and culture, many Portuguese speakers currently living in Macau are actually Portuguese expats. The hypothesis we would like to test is whether, despite the great geographical distance between Macau and Portugal, both of the aforementioned factors influence journalists based in Macau to use a language that is similar to the European Portuguese standard.

2 Methods

2.1 Corpus

In this paper, we use the three Portuguese sub-corpora from Brazil, Macau, and Portugal (hereafter BR, MO, and PT) available in the dataset of the 2015 edition of the Discriminating between Similar Languages (DSL) shared task [Zampieri *et al.*, 2015], the DSL Corpus Collection (DSLCC) version 2.1 [Tan *et al.*, 2014].

The DSLCC is a collection of journalistic texts compiled from multiple sources, including previously released corpora, containing short text excerpts sampled from various newspapers.[1] According to the information provided by the authors of the DSLCC, Macanese texts were compiled from two newspapers: *Tribuna de Macau* and *Hoje Macau*.[2,3]

The three Portuguese sub-corpora combined contain a total of 54,000 excerpts (documents) and each document contains between 20 and 100 tokens. Table 1 presents the number of documents and types in each sub-corpus.

[1] A list of sources is available in [Tan *et al.*, 2014].

[2] http://jtm.com.mo/

[3] http://hojemacau.com.mo/

	Tokens	Types	Documents
BR	602,684	41,419	18,000
MO	547,479	32,547	18,000
PT	582,420	36,313	18,000
Total	1,732,583	-	54,000

Table 1: Number of Tokens, Types, and Documents in the DSLCC BR, MO, PT sub-corpora

2.2 Computational Approach

We approach the task using a text classification system based on a linear SVM classifier implemented in LIBLINEAR [Fan *et al.*, 2008]. SVMs proved to deliver very good performance in discriminating between language varieties, achieving first place in both the 2015 [Malmasi and Dras, 2015b] and 2014 [Goutte *et al.*, 2014] editions of the DSL shared task.[4]

We use unigram and bigram language models to capture lexical and lexico-syntactic differences between the newspapers published Brazil, Macau, and Portugal. Unigram language models have been used to discriminate between Brazilian and European Portuguese newspapers texts with results of over 99% accuracy by [Zampieri and Gebre, 2012].[5] In this study researchers pointed out that lexical variation and orthographic differences play an important role in the task.

We evaluate the performance of our method using standard NLP evaluation metrics such as precision (P), recall (R), and f-score (F) for multi-class classification and accuracy (A) for binary classification settings. All results are presented using k-fold cross-validation, with $k = 10$. We consider random baseline as the baseline performance for this task. This means 33% accuracy for classification sets containing three classes and 50% accuracy for binary classification settings.

3 Results

In our first experiment, we apply the aforementioned SVM classifier to discriminate between the three corpora. We report precision, recall, and f-score in Table 2.

Features	Class	P	R	F
Unigrams	BR	0.87	0.90	0.88
	MO	0.78	0.76	0.77
	PT	0.76	0.75	0.75
	Average Scores	0.80	0.80	0.80
Bigrams	BR	0.82	0.91	0.86
	MO	0.78	0.72	0.75
	PT	0.74	0.72	0.73
	Average Scores	0.78	0.78	0.78

Table 2: Three-way classification (BR, MO, PT)

The classifier achieves an average performance of 80% f-score using unigrams and 78% f-score using bigrams suggesting that the lexical differences are the most informative information in this task. We observed that performance varies substantially among the classes. Using unigrams performance is higher for the Brazilian class (88% f-score) than for Macau (77% f-score) and Portugal (75% f-score). We investigate the performance variation by analyzing the confusion matrix of the word unigram results in Figure 1.

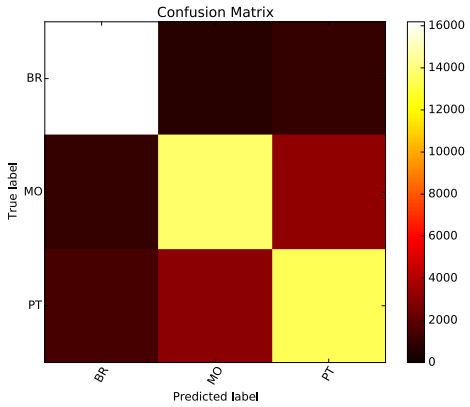

Figure 1: Confusion matrix for three-way classification

The confusion matrix shows that there is substantial confusion between MO and PT texts whereas BR texts are always the easiest to identify. To investigate this further we conduct binary classification experiments in which the algorithm is trained to choose between texts from only two countries at a time. We present accuracy results in Table 3.

Features	Newspapers	Accuracy
Unigrams	BR vs. PT	0.91
	BR vs. MO	0.93
	MO vs. PT	0.79
Bigrams	BR vs. PT	0.89
	BR vs. MO	0.91
	MO vs. PT	0.77

Table 3: Binary classification results

The algorithm achieves very good performance, 91% accuracy, when discriminating between BR and PT texts using unigrams which corroborates the findings by [Zampieri and Gebre, 2012]. The method achieves even higher performance, 93% accuracy, discriminating between BR and MO texts using unigrams. We investigate the reasons for this high performance in Section 4.1. The performance of the classifier discriminating between MO and PT texts is substantially lower than the other two. These outcomes confirm our hypothesis that currently Macanese newspaper texts are similar to the European Portuguese standard.

It is well known that named entities (NE) such as people, places, and organization play an important role in this task. To investigate the influence of NEs in classification we propose a second round of experiments presented next section.

3.1 The Influence of Named Entities

It is safe to assume that texts published in Portugal are more likely to refer to *Lisbon* and to the *European Union* than

[4] See [Goutte *et al.*, 2016] for a comprehensive evaluation.

[5] It should be noted that in [Zampieri and Gebre, 2012] the authors use full texts containing up to 500 tokens.

Macanese texts and that texts from Brazil are very likely to include names of famous Brazilian people and places. We observed this phenomenon in the analysis of the most informative features obtained in the experiments described in Section 3. To diminish the influence of country-specific expressions in classification we replicate an experiment proposed in the second edition of the DSL shared task. The experiment consists of substituting most named entities in text by placeholders *#NE#*. The DSL approach to named entity removal addresses only capitalized proper nouns and all words which are not capitalized are left in text.[6] Below is an example of how texts are represented before and after this substitution:

(1) Compara este sistema às indulgências vendidas pelo Clero na Idade Média, quando os fiéis compravam a redenção das suas almas dando dinheiro aos padres.

(2) Compara este sistema às indulgências vendidas pelo #NE# na #NE# #NE# quando os fiéis compravam a redenção das suas almas dando dinheiro aos padres.

We use texts produced after NE substitution firstly in a three way classification setting involving BR, MO, and PT texts. We used word unigrams as features because these were the best performing features presented in the last section. In Table 4 we include a column *Diff.* which contains the difference between the f-scores obtained using original texts and texts with NE substitution in percentage points.

	P	R	F	Diff.
BR	0.83	0.86	0.84	-4 pp
MO	0.70	0.72	0.71	-6 pp
PT	0.70	0.65	0.67	-8 pp
Average Scores	0.74	0.74	0.74	-6 pp

Table 4: Blind NE - Three-way classification (BR, MO, PT)

In this setting we observed that BR texts were again the easiest to identify. However, the performance of the algorithm identifying PT texts dropped 8 percentage points. We repeat the binary experiments without NEs and present results in terms of accuracy in Table 5.

Newspapers	Accuracy	Diff.
BR vs. PT	0.88	-3 pp
BR vs. MO	0.90	-3 pp
MO vs. PT	0.74	-5 pp

Table 5: Blind NE - Binary classification results

The results obtained by the classifier when discriminating between MO and PT texts were worse than when using the settings featuring the BR class (Table 5). Moreover, we observed a performance drop of 3 percentage points for the two settings including BR texts whereas the result obtained by the algorithm discriminating between MO and PT texts without NE were 5 percentage points lower. This once again suggests that BR texts are substantially different from both MO and PT texts. Finally, we observed a performance drop from 3 to 8 percentage points in all settings which confirms that named entities play an important role in this task. However, as mentioned earlier in this section, not all named entities were removed from texts, and in the most informative features we find a number of expressions which are country-specific. We discuss this in more detail in the next section.

[6]We used the script provided by the DSL shared task organizers: https://github.com/alvations/bayesmax/tree/master/bayesmax

4 Discussion

4.1 Feature Analysis

Our results show that BR newspapers can be identified with over 90% accuracy. Our analysis of the most informative features indicate that this is mostly due to orthographic conventions, for example mute consonants used in Portugal *(director)* and not used in Brazil *(diretor)*, and because of words that are more frequently used in Brazil than in other Portuguese speaking countries, for example *você* (EN: *you*). The top ten most informative features are presented in Table 6.

The method discriminates between texts from PT and MO using original texts and texts without most NEs with 79% and 74% accuracy respectively. This represents above chance (50%) performance, but it is still substantially worse than the performance obtained by the classifier when discriminating between texts from BR and MO or BR and PT. According to our feature analysis, texts from PT and MO were identified mostly relying on country-specific words. Using original texts, half of the top ten most informative features in PT texts are names of Portuguese regions or cities such as *Leiria, Aveiro, Braga, Algarve*, and *Minho* are among the top-10 most informative lexical features in MO texts after NE removal we find *patacas* (the Macanese currency), *chinesa*, and *macaense*, all of them country-specific.

Answering the question posed in the introduction, our classification results and the analysis of the most informative features suggest that there are substantial differences between the language used in BR and MO newspapers in terms of orthography and lexicon, but not between MO and PT texts. Although the SVM classifier was able to discriminate between MO and PT texts with above chance performance, the algorithm did not achieve very high performance and it relied mostly on NEs and country-specific expressions rather than on lexical or orthographic variation. MO and PT texts use the same orthography which is different from that used by BR texts. The assumption that newspapers texts from MO and PT are very similar was confirmed and we would like to investigate this in future work using other sets of features.

4.2 Visualization - Cluster Dendograms

To test the validity of the results obtained using supervised classification methods, we apply hierarchical clustering to obtain dendograms of the three language varieties. For this purpose we used the top 100 overused unigrams in MO articles (without NE removal) which in the supervised setup were the most helpful features for the MO class in distinguishing between BR, MO, and PT texts.

We first consider all texts from the same language variety as one cohesive dataset and merge them into one document, thus obtaining three large documents corresponding to

Rank	BR Orig.	BR No-NE	MO Orig.	MO No-NE	PT Orig.	PT No-NE
1	equipe	equipe	Macau	patacas	concelho	concelho
2	prefeito	prefeito	patacas	território	euros	euros
3	projeto	time	território	chinesa	freguesia	freguesia
4	fato	fato	Serviços	actriz	autarquia	autarquia
5	time	projeto	actriz	chinês	Leiria	portagens
6	você	você	China	residentes	Aveiro	distrital
7	diretor	diretor	chinesa	casinos	Braga	âmbito
8	ações	equipes	Chan	macaense	Algarve	autarcas
9	atividades	gol	Kong	territórios	Minho	orçamento
10	atual	ação	Território	casino	Novas	algarvia

Table 6: Overuse of lexical features in BR, MO, and PT texts

our three sub-corpora. Figure 2 shows the dendogram when hierarchical clustering with the average method and Euclidian distance is applied to the three merged sub-corpora.[7]

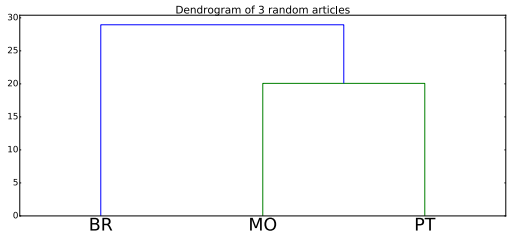

Figure 2: Dendogram of articles merged by language variety

In this simple dendogram we can see that the MO and PT datasets are displayed in the same branch of the dendogram whereas the BR dataset stands out. To confirm this, we subsequently carried out hierarchical clustering in five iterations over 16 randomly sampled articles from each sub-corpora (48 documents in total).[8] Figure 3 shows one of these iterations. MO articles are generally clustered much faster and closer to PT articles than to BR ones.

The use of hierarchical clustering confirms the results obtained using supervised text classification and indicates that the language used in articles published in Macau is substantially more similar to the language used in Portuguese newspapers than that used in Brazilian newspapers.

5 Conclusion

In this paper we proposed a supervised text classification approach to the study of language variation in Portuguese newspapers. Along with text classification we carried out a concise yet informative linguistic analysis of the most informative features in classification and we experimented with hierarchical clustering and dendograms to confirm the results obtained in the text classification experiments.

We focused on journalistic texts published in Macau in comparison to those published in Brazil and Portugal. We used short excerpts of texts available in the DSL Corpus Collection (DSLCC) version 2.1 [Tan *et al.*, 2014]. Our results confirmed our initial hypothesis that the language used in Portuguese newspapers published in Macau is much more similar to the language used in texts published in Portugal than to the one used in Brazilian newspapers.

We provided quantitative and qualitative evidence that journalistic texts published in Brazil and Macau differ substantially from each other. Our SVM classifier using a unigram language model can discriminate between texts from these two countries with 93% accuracy. Our results indicate that the same is not true for texts from Portugal and Macau. Texts from these two corpora could not be easily distinguished from each other by the SVM classifier. The analysis we carried out on the most informative features indicate that the main differences between texts published in Macau and Portugal captured by the classifier are country-specific expressions such as place names, currency name, etc.

5.1 Future Work

Our paper is, to the best of our knowledge, the first attempt to study the language of Portuguese newspapers published in Macau using NLP methods. We would like to test other features in future work such as word trigrams, POS tags and other forms of delexicalized text representations [Lui *et al.*, 2014] to investigate whether there are specific grammatical constructions prominent in Macanese journalism that are not used so often in the other two Portuguese varieties. In future work we would also like to investigate variation in the style of texts published in the newspapers from these three countries. To this end we are using readability metrics such as sentence length and lexical density.

Finally, we would like to investigate whether native speakers of different Portuguese varieties are able to discriminate Macanese texts from Brazilian and European texts. [Ács *et al.*, 2015; Goutte *et al.*, 2016] report that classification algorithms are able to obtain higher performance than humans distinguishing between Brazilian and European texts. We would like to investigate if this is true for Macanese texts as well.

Acknowledgments

Liviu P. Dinu was supported by UEFISCDI, PNII-IDPCE-2011-3-0959.

[7] We used the linkage and dendogram methods in SciPy [Jones *et al.*, 2001] and matplotlib to obtain this image.

[8] The number of documents used in this step was defined to optimize the dendogram visualization.

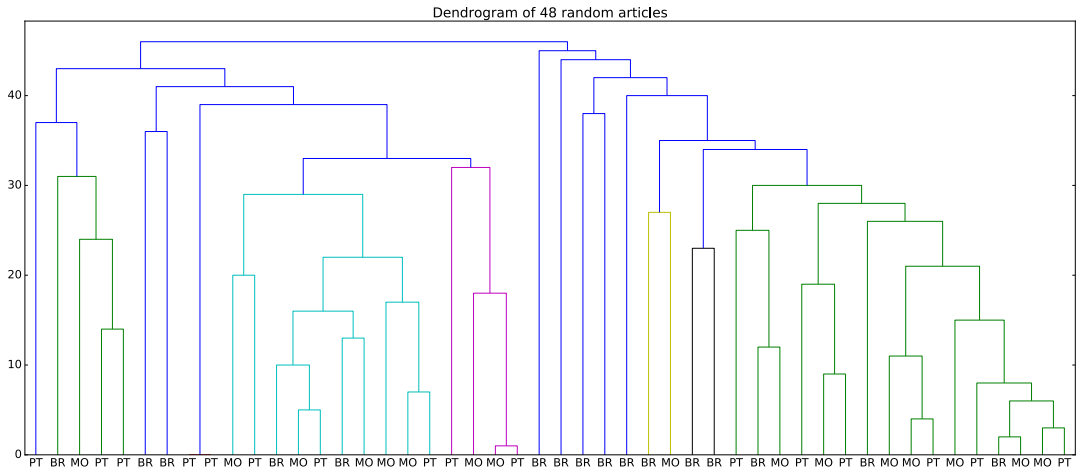

Figure 3: Hierarchical clustering of random sampled articles

References

[Ács *et al.*, 2015] Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. A Two-level Classifier for Discriminating Similar Languages. In *Proceedings of LT4VarDial*, 2015.

[Amaro, 2016] Vanessa Amaro. Linguistic Practice, Power and Imagined Worlds: The Case of the Portuguese in Postcolonial Macau. *Journal of Intercultural Studies*, 37(1):33–50, 2016.

[Baxter, 1992] Alan N Baxter. Portuguese as a pluricentric language. *Pluricentric languages: differing norms in Different Nations*, (62):11, 1992.

[Baxter, 1996] Alan N Baxter. Portuguese and Creole Portuguese in the Pacific. *Atlas of languages of intercultural communication in the Pacific, Asia, and the Americas*, 3:299, 1996.

[Ciobanu and Dinu, 2016] Alina Maria Ciobanu and Liviu P. Dinu. A Computational Perspective on Romanian dialects. In *Proceedings of LREC*, 2016.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Goutte *et al.*, 2014] Cyril Goutte, Serge Léger, and Marine Carpuat. The NRC System for Discriminating Similar Languages. In *Proceedings of VarDial*, 2014.

[Goutte *et al.*, 2016] Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of LREC*, 2016.

[Jones *et al.*, 2001] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001.

[Lui and Cook, 2013] Marco Lui and Paul Cook. Classifying English Documents by National Dialect. In *Proceedings of ALTA*, 2013.

[Lui *et al.*, 2014] Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. Exploring Methods and Resources for Discriminating Similar Languages. In *Proceedings VarDial*, 2014.

[Maier and Gómez-Rodrıguez, 2014] Wolfgang Maier and Carlos Gómez-Rodrıguez. Language Variety Identification in Spanish Tweets. In *Proceedings of LT4CloseLang*, 2014.

[Malmasi and Dras, 2015a] Shervin Malmasi and Mark Dras. Automatic Language Identification for Persian and Dari Texts. In *Proceedings of PACLING*, 2015.

[Malmasi and Dras, 2015b] Shervin Malmasi and Mark Dras. Language Identification Using Classifier Ensembles. In *Proceedings of LT4VarDial*, 2015.

[Malmasi *et al.*, 2015] Shervin Malmasi, Eshrag Refaee, and Mark Dras. Arabic Dialect Identification Using a Parallel Multidialectal Corpus. In *Proceedings of PACLING*, 2015.

[Tan *et al.*, 2014] Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In *Proceedings of BUCC*, 2014.

[Zampieri and Gebre, 2012] Marcos Zampieri and Binyam Gebrekidan Gebre. Automatic Identification of Language Varieties: The Case of Portuguese. In *Proceedings of KONVENS*, 2012.

[Zampieri *et al.*, 2013] Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. N-Gram Language Models and POS Distribution for the Identification of Spanish Varieties. In *Proceedings of TALN*, 2013.

[Zampieri *et al.*, 2015] Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. Overview of the DSL Shared Task 2015. In *Proceedings of LT4VarDial*, 2015.