VarClass: An Open Source Language Identification Tool for Language Varieties

Marcos Zampieri[1], Binyam Gebrekidan Gebre[2]

Saarland University, Saarbrücken - Germany[1]
Max Planck Institute for Psycholinguistics, Nijmegen - Holland[2]
marcos.zampieri@uni-saarland.de, bingeb@mpi.nl

Abstract

This paper presents *VarClass*, an open-source tool for language identification available both to be downloaded as well as through a graphical user-friendly interface. The main difference of *VarClass* in comparison to other state-of-the-art language identification tools is its focus on language varieties. General purpose language identification tools do not take language varieties into account and our work aims to fill this gap. *VarClass* currently contains language models for over 27 languages in which 10 of them are language varieties. We report an average performance of over 90.5% accuracy in a challenging dataset. More language models will be included in the upcoming months.

Keywords: language identification, language varieties, n-grams

1. Introduction

Automatic language identification or simply language identification is a well-known task in NLP. State-of-the-art methods for language identification obtain performance above 95% accuracy (Brown, 2013). These methods are, however, general-purpose and neglect language varieties modelling pluricentric languages as unique classes. For instance, methods are trained to recognize English, French or Portuguese texts regardless of variety: American, Australian or British, French or Canadian. Brazilian or European, to name a few. It is, however, understood that language variation influences perception and allows speakers to identify texts or speech production (accent) that belong to a certain variety or dialect.

It is very common for language identification methods to perform almost perfectly when distinguishing languages which are typologically not closely related as well as when recognizing languages with unique character sets. In these scenarios, the distinction based on character n-gram models presents satisfactory results. This is, however, not the case of closely related languages, varieties or dialects, where distinction is based on very subtle differences that algorithms can be trained to recognize.

Even though general-purpose systems do not take language varieties into account, there are a number of situations in which identifying the language variety of a text might be important. For example, in NLP tasks such as machine translation or information retrieval identifying the variety of a given text can improve performance. For this reason, in recent years a couple of studies were published about the identification of varieties[1] and dialects as evidenced in section 2.1.. These studies, however, try to distinguish language varieties of the same language and to our knowledge none of them has been yet integrated into a real-world language identification system. In this work, we aim to go one step further and integrate language varieties into a real-world setting releasing the tool *VarClass*.

VarClass might be of interest not only to scholars working on language identification and computational linguistics but also to philologists and linguists who do not have programming skills to develop new tools or adapt existing ones. Researchers from different areas will be able to use *VarClass* through its web interface with only a few clicks.

2. Language Identification Methods

This section gives a brief overview on general purpose language identification methods starting from early approaches (Ingle, 1980) to character n-gram based methods (Beesley, 1988) and (Dunning, 1994) to the most recent publications (Brown, 2013). The main aim of this paper is to present a resource. Due to space limitations we will not go into detail on how other language identification methods work. Instead we will refer to a couple of studies which compare different methods and therefore provide an accurate picture of the strengths and weaknesses of each method.

There were a number of comparative studies study published over the years. The one by Grafenstette (1995) compares two language identification methods: a trigram approach inspired by the work of Beesley (1988) and Cavnar and Trenkle (1994) and the frequent word approach proposed by Ingle (1980). A couple of other comparative studies include the one by Vojtek and Belikova (2007) which compares two methods based on Markov processes. Another study (Padró and Padró, 2004) compared the performance of three methods: Markov models, trigram frequency vectors and n-gram based text categorization (Cavnar and Trenkle, 1994) and finally, Groethe et al. (2008) compared methods that used short words, frequent words and character n-grams.

The Internet is an interesting application for language identification. Documents available on the Internet are often unidentified regarding source language. Moreover, the

[1] For simplicity we grouped together methods designed to distinguish language varieties and similar languages. From an NLP point of view, the problems are similar. For more information see Clyne (1992) on pluricentric languages and Chambers and Trudgill (1998) on dialectology.

same document may be written in more than one language (code alternation), making it difficult for computer programs to process them. In the last years, a number of language identification methods were proposed for Internet data, (Martins and Silva, 2005), (Rehurek and Kolkus, 2009), (Tromp and Pechnizkiy, 2012) and (Vogel and Tresner-Kirsch, 2012).

Different classification algorithms have been tested for automatic language identification such as Monte Carlo sampling (Poutsma, 2001), Markov-based methods (Xafopoulos et al., 2004) and machine learning techniques (Combrinck and Botha, 1994). Although most language identification studies use supervised learning strategies, there were a couple of attempts to perform language identification using unsupervised methods (Amine et al., 2010). In this study, (Amine et al., 2010) propose a hybrid method that includes the popular k-means clustering.

Recent language identification studies include (Lui and Baldwin, 2012), who developed a tool called *langid.py*, (Takçı and Güngör, 2012) who propose a centroid-based classification approach for language identification reporting results of 97.5% accuracy and finally Brown (2013) who presented a language identifier for over 1,100 languages.

2.1. Models for Similar Languages, Dialects and Varieties

As previously mentioned, general-purpose methods for automatic language identification were substantially explored over the years. The same is, however, not true for methods designed to deal specifically with similar languages, varieties and dialects. One of the first studies (Ljubešić et al., 2007) proposed a computational model for the identification of Croatian texts reporting 99% recall and precision in three processing stages. One of these processing stages, includes a list of forbidden words (blacklist) that appear only in Croatian texts, making the algorithm perform better. An improved version of the blacklist classifier was recently published (Tiedemann and Ljubešić, 2012).

Huang and Lee (2008) presented a bag-of-words approach to distinguishing Chinese texts from the mainland and Taiwan reporting results of up to 92% accuracy. (Ranaivo-Malancon, 2006) presents a semi-supervised character-based model to distinguish between Indonesian and Malay, two closely related languages from the Austronesian family.

Two shared tasks are worth mentioning: the DEFT2010[2] shared task held in 2010 in Montreal (Grouin et al., 2010) and the Discriminating between Similar Languages (DSL)[3] held at the VarDial Workshop in 2014. The DEFT2010 shared task combined language variety discrimination and temporal text classification. Systems that participated in the DEFT2010 shared task had to classify French journalistic texts with respect to their geographical location as well as the decade in which they were published. The DSL shared task aims to provide a dataset and an evaluation methodology to evaluate systems discriminating 13 different languages and language varieties in 6 groups of languages as follows: A (Bosnian, Croatian, Serbian), B (Brazilian Portuguese, European Portuguese), C (Indonesian, Malaysian), D (Czech, Slovakian), E (Peninsular Spain, Argentine Spanish) and F (American English, British English).

As evidenced in this section, it was only in the last few years that a couple of studies were published about the automatic identification of similar languages, varieties and dialects. To our knowledge, none of them has been yet integrated into a real-world identification setting or an open-source tool and our work aims to fill this gap.

2.2. Web-based Language Identifiers

One of the new aspects of our work is to provide an online language identifier through a user-friendly interface that can be used by anyone. There are, of course, other language identifiers on the web as for example the one used by Google Translator which is able to identify over 75 languages from complete texts to just a few words. None of them, to our knowledge, take language varieties into account.

One of the tools that identify languages available online is TextCat[4]. TextCat is an implementation of the previously discussed algorithm proposed by Cavnar and Trenkle (1994). The tool contains language models from 76 languages and can be adapted or customized to user's needs as it allows users to train the system with one's own data.

Other online language identifiers worth mentioning are the one developed by Xerox[5] which is able to discriminated between over 80 languages, the one by Translated Labs[6] containing 102 language models. Lingua::Identify[7] is a language identification tool available for download but does not contain a web interface.

3. Methods

The algorithm behind *VarClass* is an adapted version of the likelihood algorithm described in (Zampieri and Gebre, 2012) and later tested in different scenarios. In this paper we applied a simple discriminative model with results reaching up to 99,8% in distinguishing Brazilian from European Portuguese texts. The algorithm was also tested to identify French (Canada and France) achieving 99,0% accuracy and Spanish (Spain, Argentina), 96.2% (Zampieri et al., 2012).

The algorithm uses a simple likelihood function calculated over smoothed language models. These language models can be obtained using character, words or even POS categories (Zampieri et al., 2013)[8]. The likelihood function is calculated as described in equation 1.

$$P(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \quad (1)$$

[2] http://www.groupes.polymtl.ca/taln2010/deft.php

[3] http://corporavm.uni-koeln.de/vardial/sharedtask.html

[4] http://odur.let.rug.nl/vannoord/TextCat/

[5] http://open.xerox.com/Services/LanguageIdentifier

[6] http://labs.translated.net/language-identifier/

[7] http://search.cpan.org/ ambs/Lingua-Identify-0.51/

[8] The performance of the algorithm using POS tags and disregarding word forms is substantially lower than the one obtained using characters or words. Therefore these features are used mostly for contrastive studies rather than in real-world NLP applications.

N is the number of n-grams in the test text, n_i is the ith n-gram and L stands for the language models. Given a test text, we calculate the probability for each of the language models. The language model with highest probability determines the identified language of the text.

At the moment, *VarClass* contains language models for 27 languages (10 of them are language varieties) and can be accessed through a web interface[9]. More language models will be incorporated in the tool within the next few months. A complete list with the languages available along with their respective ISO-3166 code is available next.

Language	Country	ISO
Albanian	Albania	ALB
Bosnian	Bosnia	BIH
Bulgarian	Bulgaria	BGR
Croatian	Croatia	HRV
Czech	Czech Republic	CZE
Dutch	Netherlands	NLD
English	United States	USA
English	United Kingdom	GBR
French	Canada	CAN
French	France	FRA
German	Germany	DEU
Greek	Greece	GRC
Indonesian	Indonesia	IDN
Italian	Italy	ITA
Macedonian	Macedonia	MKD
Malay	Malaysia	MYS
Portuguese	Brazil	BRA
Portuguese	Portugal	PRT
Romanian	Romania	ROU
Serbian	Serbia	SRB
Slovakian	Slovakia	SVK
Spanish	Argentina	ARG
Spanish	Mexico	MEX
Spanish	Peru	PER
Spanish	Spain	ESP
Swedish	Sweden	SWE
Turkish	Turkey	TUR

Table 1: Language Models

To calculate the language models we compiled corpora from different sources such as the DSL Corpus Collection (Tan et al., 2014), the SETimes Corpus (Tyers and Alperen, 2010), OPUS (Tiedemann, 2012) and Wikipedia[10] data.

In *VarClass*, users can choose to discriminate between languages using words, characters or both. Although discrimination based on POS tags have been tested, in *VarClass* we did not implement this function as it would substantially worsen performance.

4. Evaluation

The performance of the algorithm has been discussed in the aforementioned studies (Zampieri and Gebre, 2012) and it has been tested in different language settings. When distinguishing the 27 classes presented here, *VarClass* achieved a performance of 90.5% F-Measure using character trigrams. To evaluate *VarClass* we used a test set containing 5,400 documents corresponding to 200 documents per class. The test set was not previously included in the training stage and featured documents of up to 300 characters. The performance varied depending on the language and reached 1.0 for Albanian to 0.68 for British English. The average scores obtained for the four pluricentric languages and 10 language varieties in terms of Precision, Recall and F-Measure are presented next:

Language	Classes	P	R	F
English	2	0.784	0.725	0.753
Spanish	4	0.864	0.825	0.844
Portuguese	2	0.971	0.970	0.970
French	2	0.980	0.977	0.979

Table 2: Discriminating Pluricentric Languages

5. Conclusion

This paper presented a new resource, the *VarClass* language identification tool. This work is a first step towards the evaluation of language varieties into broader language identification settings. We believe that this tool fills an important gap among other language identification tools that do not take language varieties into account. The tool can be used by linguists and computational linguists interested in language identification, contrastive linguistics as well as by other users not related to the NLP and linguistics research community.

As shown in preliminary work (Zampieri and Gebre, 2012; Zampieri et al., 2012), our algorithm presents performance comparable to similar language identificaiton algorithms which do not include language varieties. Our results show that it is possible to distinguish language varieties in real-world settings without substantial performance loss.

We evaluated *VarClass* using texts of up to 300 characters. An open question that our experiments leave is the performance of this tool when identifying the language of very short texts. Our initial investigation suggests that *VarClass* is able of identifying the language of texts that contain more than 3 words (or 15 characters) with satisfactory performance. This variable should be investigated more carefully in future work.

Acknowledgement

We would like to thank Jörg Tiedemann, Nikola Ljubešić, Sascha Diwersy and Liling Tan for interesting discussions and important feedback throughout the development of this software. The name *VarClass* was a suggestion of Sascha Diwersy to whom we once more express our sincere gratitude. We would also like to thank the anonymous reviewers who provided us important feedback.

6. References

Abdelmalek Amine, Zakaria Elberrichi, and Michel Simonet. 2010. Automatic language identification: an al-

[9] http://corporavm.uni-koeln.de/varclass

[10] http://www.wikipedia.org/

ternative unsupervised approach using a new hybrid algorithm. *International Journal of Computer Science and Applications*, 7:94–107.

Kenneth Beesley. 1988. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the Annual Conference of the American Translators Association*, pages 57,54.

Ralf Brown. 2013. Selecting and weighting n-grams to identify 1100 languages. In *Proceedings of the 16th International Conference on Text Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence (LNAI 8082)*, pages 519–526, Pilsen, Czech Republic. Springer.

William Cavnar and John Trenkle. 1994. N-gram-based text catogorization. *3rd Symposium on Document Analysis and Information Retrieval (SDAIR-94)*.

Jack Chambers and Peter Trudgill. 1998. *Dialectology (2nd Edition)*. Cambridge University Press.

Michael Clyne. 1992. *Pluricentric Languages: Different Norms in Different Nations*. CRC Press.

Hendrik Petrus Combrinck and Elizabeth Botha. 1994. Text-based automatic language identification. In *Proceedings of the 6th Annual South African Workshop on Pattern Recognition*.

Ted Dunning. 1994. Statistical identification of language. Technical report, Computing Research Lab - New Mexico State University.

Gregory Grefenstette. 1995. Comparing two language identification schemes. In *Proceedings of JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data*, Rome.

Lena Groethe, Ernesto De Luca, and Andreas Nürnberger. 2008. A comparative study on language identification methods. In *Proceedings of LREC*.

Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek, and Pierre Zweigenbaum. 2010. Présentation et résultats du défi fouille de texte deft2010 où et quand un article de presse a-t-il été écrit? *Actes du sixième DÉfi Fouille de Textes*.

Chu-ren Huang and Lung-hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of PACLIC 2008*, pages 404–410.

Norman Ingle. 1980. *A Language Identification Table*. Technical Translation International.

Nikola Ljubešić, Nives Mikelic, and Damir Boras. 2007. Language identification: How to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Meeting of the ACL*.

Bruno Martins and Mario Silva. 2005. Language identification in web pages. *Proceedings of the 20th ACM Symposium on Applied Computing (SAC), Document Engineering Track. Santa Fe, EUA.*, pages 763–768.

Muntsa Padró and Llus Padró. 2004. Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, (33):155–162.

Arjen Poutsma. 2001. Applying monte carlo techniques to language identification. In *Proceedings of Computational Linguistics in the Netherlands*.

Bali Ranaivo-Malancon. 2006. Automatic identification of close languages - case study: Malay and indonesian. *ECTI Transactions on Computer and Information Technology*, 2:126–134.

Radim Rehurek and Milan Kolkus. 2009. Language identification on the web: Extending the dictionary method. In *Proceedings of CICLing. Lecture Notes in Computer Science*, pages 357–368. Springer.

Hidayet Takçı and Tunga Güngör. 2012. A high performance centroid-based classification approach for language identification. *Pattern Recognition Letters*, 3:2077–2084.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of The 7th Workshop on Building and Using Comparable Corpora (BUCC)*.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Erik Tromp and Mykola Pechnizkiy. 2012. Graph-based n-gram language identification on short texts. In *Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning (Benelearn 2011)*, pages 27–34.

Francis Tyers and Murat Serdar Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Proceedings of LREC 2010*, Malta.

John Vogel and David Tresner-Kirsch. 2012. Robust language identification in short, noisy texts: Improvements to liga. In *Third International Workshop on Mining Ubiquitous and Social Environments (MUSE 2012)*.

Peter Vojtek and Maria Belikova. 2007. Comparing language identification methods based on markov processess. In *Slovko, International Seminar on Computer Treatment of Slavic and East European Languages*.

Alexandros Xafopoulos, Constantine Kotropoulos, George Almpanidis, and Ioannis Pitas. 2004. Language identification in web documents using discrete hmms. *Pattern Recognition*, 37:583–594.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012*, pages 233–237, Vienna, Austria.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2012. Classifying pluricentric languages: Extending the monolingual model. In *Proceedings of the Fourth Swedish Language Technlogy Conference (SLTC2012)*, pages 79–80, Lund, Sweden.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN2013*, Sable d'Olonne, France.