Demystify languages and varieties similarity: the identification of similar languages and varieties

Sudhir Singh
Rochester Institute of Technology
1 Lomb Memorial Dr, Rochester, NY 14623
ss6842@cs.rit.edu

Abstract

This paper describes the research performed on the DSL (Discriminating Similar Language (Tan et al., 2014a)) dataset to identify similar languages and language varieties. In this research, the naive Bayes and Linear SVM classifiers are used with bags-of-words (counts) and TF-IDF character n-grams as well as word n-grams as features to identify the difference. The models used in this research performed well in comparison to (Tan et al., 2014b) initial results.

1 Introduction:

The identification of a language in a multi-lingual application (natural language processing application) is crucial. The continuous effort of researchers and corporate has pushed the limit of this sub-field. Nowadays, it is very easy to identify languages having different language script with some base models and get really good accuracy. However, in a multi-lingual environment, it is still quite challenging to identify or distinguish between closely related languages or languages varieties.

In recent years, an initiative has been taken to solve this issue. The VarDial evaluation series (Malmasi et al., 2016); (Zampieri et al., 2017); (Zampieri et al., 2018) is running it's fifth iteration in the area of closely related languages and language varieties. The first iteration was compromises of identifying the language in a closely related language group. In this research, I like to identify the similarity or dissimilarity of closely related languages and language varieties.

The paper is divided into sections corresponding to each task. Section 2 has details about the dataset used for this problem. The methods used to address the issue are described in section 3. Section 4 has the experiment results and performance of the methods used. Section 5 carries out the discussion on the topic and research. Section 6 summarizes and concludes the research performed with this paper.

2 Data:

The availability of the dataset for similar languages and language varieties is hard to gather or find. The dataset used to classify similar languages and language varieties is taken from the Discriminating Similar Language (DSL) corpus collection (Tan et al., 2014a). The DSL corpus is collected using a shared task and comprises of news data from different languages. As part of the DSL tasks, the dataset was encoded into "UTF-8" because of the nature of different languages and the character & alphabets used in these varieties of languages.

2.1 Data Representation

As the dataset is a collection of similar and dissimilar languages from different region, to distinguish between them, the language code using ISO 639-1 convention [1] has been used. Also, similar languages and language varieties are kept in groups. The table 1 represent the language and varieties, the language and the code of the language.

The DSL corpus is divided into three categories namely, train, development, and test set. The table 2 describe these categories distribution for each language/variety. The number of samples (each language and varieties) in these categories are equally distributed. All dataset has ".txt" file format. Each line in each file is tab-delimited as follow: `sentence<tab>language code<tab>`

[1] http://www.loc.gov/standards/iso639-2/php/English$_list.php$

Group	Language/Variety	Language Code
A	Bosnian	bs
	Croatian	hr
	Serbian	sr
B	Indonesian	id
	Malay	my
C	Czech	cz
	Slovak	sk
D	Brazilian Portuguese	pt-BR
	European Portuguese	pt-PT
E	Argentine Spanish	es-AR
	Castilian Spanish	es-ES
F	American English	en-US
	British English	en-GB

Table 1: Similar languages and varieties with the language code

Category	sentences	tokens
train.txt	18,000	≥ 20
devel.txt	2,000	≥ 20
test.txt	1,000	≥ 20

Table 2: Data description

3 Methods:

This section describe the process and methodologies used to identify similar languages and language varieties.

3.1 Data Pre-processing:

The data is divided into train, development test and test dataset. The data further processed and below steps are performed to clean it.

- Removed all punctuation
- Removed #NE# tag
- Removed all numeric variables
- Removed quotes (”), new line (`\n`) and white-spaces

3.2 Features selection:

A language is made of characters, symbols and words. Therefore, the best way to identify a language or variety is to find the characters or word n-grams features. I have used character n-grams (1 to 10-grams) and words n-grams (1 to 3-grams) as features. The bag-of-words (count) and term frequency-inverse document frequency (tf-idf) both method works well to identify a language. I have used both feature selection to classify the languages. However, tf-idf worked well for this problem.

3.3 Model selection:

The author (Baldwin and Lui, 2010) used naive Bayes in their experiment for language identification. Additionally, as described in feature selection 3.2, the feature selected for this problem are character n-grams and words n-grams, the naive Bayes model is a good fit.

Naive Bayes assumption:

$$P(w_1, w_2 \dots w_n | c_j) = \prod_{i}^{n} P(w_i | c_j)$$

which gives Naive Bayes classifier:

$$\hat{c} = \textbf{argmax}_{c_j \in C} \hat{P}(c_j) \prod_{w_i = 0 \in d}^{n} \hat{P}(w_i | c_j)$$

The *sklearn* implementation of naive Bayes used in this experiment[2] scikit-learn (Pedregosa et al., 2012). The Laplace smoothing $\alpha = 1$ and prior probability settings has been applied.

As feature matrix is large (character n-grams and word n-grams) and have distinct vector space, the Linear support vector model has been used for better classification and results along with naive Bayes.

$$w = \sum_{i}^{n} y_i \alpha_i + b$$

The SVM model used in this research is same as used by (Çöltekin and Rama, 2016). The (Çöltekin and Rama, 2016) implementation is similar to (ws-, 2014). In the SVM model, I have used word n-grams and character n-grams separately with the tf-idf feature vector. The SVM model is implemented with *sklearn* scikit-learn (Pedregosa et al., 2012).

3.4 Evaluation methodology:

As language identification is a classification task and we have been provided with standard gold label, the basic confusion matrix makes a good fit to evaluate the predicted labels. Along with confusion matrix, precision, recall and F-1 score will provide a better insight of model accuracy and performance.

[2] https://scikit-learn.org/

4 Experiment Results:

The experiment results with word unigram and bigram using Naive Bayes and Linear SVM.

model	word n-grams	accuracy
Naive Bayes	unigram	0.8765
Linear SVM	unigram	0.8775
Naive Bayes	bigram	0.8745
Linear SVM	bigram	**0.8806**

Table 3: word n-gram accuracy

language	precision	recall	f1-score	support
sk	1.00	1.00	1.00	1000
pt-PT	0.93	0.93	0.93	1000
es-AR	0.92	0.87	0.90	1000
my	0.99	0.99	0.99	1000
pt-BR	0.93	0.93	0.93	1000
cz	1.00	1.00	1.00	1000
id	0.99	0.99	0.99	1000
hr	0.93	0.90	0.91	1000
es-ES	0.88	0.92	0.90	1000
en-US	0.47	0.47	0.47	800
en-GB	0.47	0.47	0.47	800
sr	0.92	0.92	0.92	1000
bs	0.86	0.89	0.88	1000
micro avg	0.88	0.88	0.88	12600
macro avg	0.87	0.87	0.87	12600
weighted avg	0.88	0.88	0.88	12600

Table 4: Classification report all languages with Linear SVM word bigram

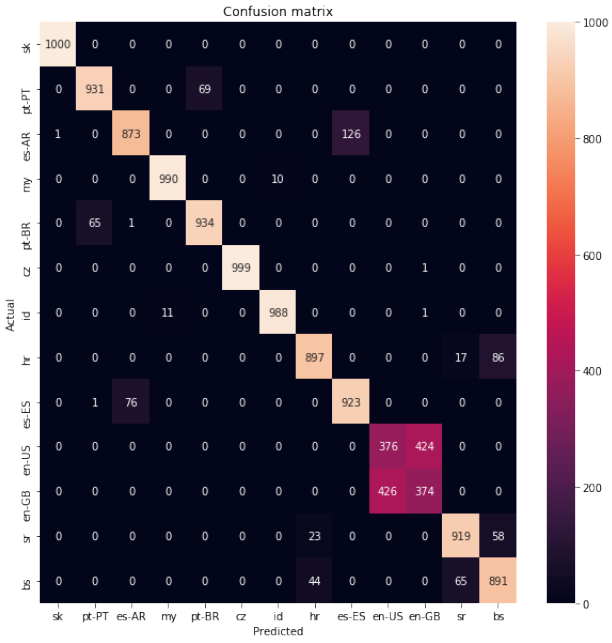

Figure 1: Confusion matrix for all languages with Linear SVM word bigram

The experiment results with character n-grams using Naive Bayes and Linear SVM.

model	char n-grams	accuracy
Naive Bayes	5-gram	0.8629
Linear SVM	5-gram	0.8845
Naive Bayes	6-gram	0.8642
Linear SVM	6-gram	0.8870
Naive Bayes	9-gram	0.8650
Linear SVM	9-gram	**0.8877**

Table 5: Character n-gram accuracy

language	precision	recall	f1-score	support
en-GB	0.46	0.47	0.46	800
hr	0.95	0.93	0.94	1000
my	0.99	0.98	0.99	1000
sr	0.96	0.95	0.95	1000
sk	1.00	1.00	1.00	1000
es-AR	0.91	0.88	0.90	1000
cz	1.00	1.00	1.00	1000
id	0.98	0.99	0.99	1000
en-US	0.45	0.44	0.45	800
pt-BR	0.94	0.95	0.94	1000
es-ES	0.89	0.91	0.90	1000
pt-PT	0.95	0.94	0.94	1000
bs	0.90	0.92	0.91	1000
micro avg	0.89	0.89	0.89	12600
macro avg	0.87	0.87	0.87	12600
weighted avg	0.89	0.89	0.89	12600

Table 6: Classification report for all languages with Linear SVM character 9-grams

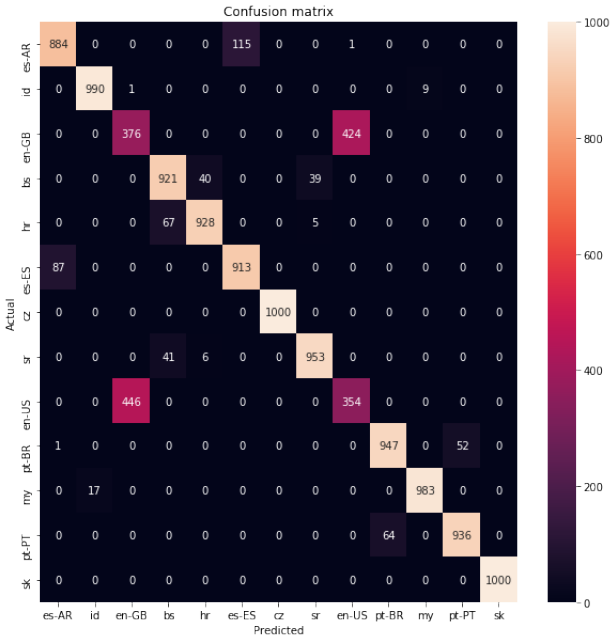

Figure 2: Confusion matrix for all languages with Linear SVM character 9-grams

4.1 Result analysis:

The research performed is in the final phase. Both models have performed well for the identified configuration and produced good accuracy. However, the models are confused for many sentences. The confusion matrix gives a visual representation of Linear SVM confusion matrix for word bigram and character 9-grams. Though the models performed sufficiently suitable for each language group, there are few errors in prediction for different group label. The Naive Bayes model confused for many different groups; however, the Linear SVM with character 9-gram is just confused for few different category groups which are pt-BR & es-ER, en-GB & id, en-US & es-AR. The error has been to further analyzed, and another configuration has to be tested in order to improve the model. The Linear SVM with character 6-gram works well for the situation; however, has little less accuracy compared to character 9-grams.

Below is the comparison of model accuracy used by authors:-

Author	model	n-grams	%
Zampieri	NaiveBayes	5-gram	0.8740
Sudhir	SVM	9-gram	0.8877
Mathur	RNN	SLIDE	0.9512

Table 7: Comparison of model accuracy used by authors

5 Discussion

The accuracy received using Linear SVM with character 6-gram and 9-grams slightly higher than (Tan et al., 2014b) initial results. However, the research needs more improvement concerning different language group identification and improving the overall accuracy using neural network architecture. After the implementation of neural network architecture and comparing results will benefit the research and will be beneficial in identifying language similarities.

6 Conclusion

In this term project paper, I described the implementation and results of my experiment in identifying similar languages and varieties. The naive Bayes and linear SVM model have been used with bags-of-words (counts) and tf-idf of character and words n-grams to identify the language and their similarity. The score of both the models are around 88% which is slightly higher than (Tan et al., 2014b) initial results. Based on the experiment, one can infer that the selection of feature is crucial to improve the system or model accuracy. Additionally, the character n-gram and word n-gram may vary in each language group for better results. Although the results of my experiments are slightly higher, however the models confused between different group languages in few instances as seen in the confusion matrix. There is undoubtedly scope of identifying this behavior and improving the model.

7 Future work

As mentioned in conclusion 6, though the accuracy of models is slightly higher than (Tan et al., 2014b), however, the experiment indicates character or word level similarity in different group languages in few instances. With a strong straightforward point, I believe, this behavior has to be analyzed further. To improve upon this behaviour, I wish to use neural network architecture in future work of my research to identify the languages. I strongly believe that the character n-gram with CNN (Ali, 2018) or RNN (Mathur et al., 2017) would produce better result as experimented by researchers.

References

2014. *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland.

Mohamed Ali. 2018. Character level convolutional neural network for german dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 172–177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California. Association for Computational Linguistics.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear svms and neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan. The COLING 2016 Organizing Committee.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Priyank Mathur, Arkajyoti Misra, and Emrah Budur. 2017. LIDE: language identification from text documents. *CoRR*, abs/1701.03682.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2012. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014a. 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)*, pages 11–15, Reykjavik, Iceland.

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

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.

A Appendices