AUTOMATIC LANGUAGE IDENTIFICATION FOR FINNIC LANGUAGES

Denisova D.A. (wamelss@mail.ru)
Saint Petersburg State University, Saint Petersburg, Russia

We present our approach to automatic language identification (LI) for Finnic languages. Most of the languages from this group are low resource, which means that we have to deal not only with LI for similar languages but also with data sparsity.
We trained and tested several classifiers: Neural Network, Support Vector Machine, Logistic Regression, Random Forest and k Nearest Neighbors. The evaluation was performed for sentence-level and word-level LI.
The best accuracy both for sentences and for words was obtained by the Neural Network, 95.62 and 83.34 respectively.

Key words: language identification, closely related languages, low resource languages, neural network

1. Introduction

Language identification (LI) is a task of identifying the natural language the given document or a part thereof is written in. In its essence, LI is a task of text categorization where the documents are mapped to the set of language labels.
LI is not a new problem in NLP, it has been researched since the 1960s. In general, LI has reached almost 100 percent accuracy where there are large documents, high resource languages, and quite different languages involved. Nevertheless, there is still a number of open issues concerning automatic LI.
Problems with LI arise when we deal with closely-related languages, low resource languages, or short texts. Closely-related languages pose a problem because, naturally, it is more difficult to distinguish between those languages due to their many similarities in orthography, word-structure and/or grammar. Low resource or lower density languages are ones that lack sufficient linguistic resources of different kinds, i.e. large corpora, annotated corpora, linguistic tools or even plain texts. Thus, the limited nature of data makes the identification of those languages much more complicated. As for the short texts, most of the research in LI has traditionally been conducted using large documents as data. In recent years, however, significant attention has been focused on the LI of short texts (usually tweets) or even words.
This paper presents the implementation of automatic LI for Finnic languages. This research was a part of a project aimed at creating a morphological tagger for Finnic languages. In our system, automatic LI was a step that preceded the morphological analysis. In the process of creating a LI system, we had to deal with the issues mentioned above which will be further detailed in the following sections.

The rest of the paper is organized as follows. Section 2 provides the relevant work in the field of LI. Section 3 describes the Finnic languages, their similarities, and differences. Section 4 gives an account of the evaluation data and, finally, Section 5 presents our experiments and their results.

2. Related work

Since the first paper on computer LI, published in the 60s, the automatic LI has come a long way. (Jauhiainen et al., 2018) provide a comprehensive outline of all the features and methods used in LI, as well as its open problems.

LI for low resource languages has been investigated by (King, 2015) in his Ph.D. thesis. He focused on data collection and word-level LI in multilingual documents and featured Conditional Random Field model, Hidden Markov Model, and Logistic Regression.

In recent years LI of short texts has often been implemented on Twitter messages. For instance, (Lui and Baldwin, 2014) extended an existing Twitter message to 65 languages and compared several off-the-shelf identifiers on that data. The average accuracy that they report is 86%.

When it comes to LI for similar languages, research usually doesn't discriminate between languages, varieties, and dialects. For example, (Simaki et al., 2017) used several different classifiers (Neural Network, Bagging, Support Vector Machine, etc.) to distinguish between English varieties. They used a large number of different features and obtained an accuracy of 73.86%.

(Ljubešić and Kranjčić, 2015) attempted to discriminate between different Southern Slavic languages on Twitter messages. They used words, character 3 and 6-grams as features and trained several classifiers among which Multinomial Naive Bayes showed the best accuracy. Accuracy on the test data was 99%.

As far as we're concerned, no research into the LI of Finnic languages has been conducted.

3. Finnic languages

LI in this paper is conducted for Finnic languages. Finnic or Baltic Finnic languages are a Finno-Ugric branch of Uralic languages. They are spoken around the Baltic Sea, hence their name. An overview of the Finnic languages can be found in (Viitso, 1998).

Finnic languages are subsequently divided into Northern Finnic and Southern Finnic languages. Northern Finnic languages are Finnish, Karelian, Veps, and Ingrian. Southern Finnic languages are Estonian, Votic and Livonian. The two most prominent languages in this group are Finnish and Estonian that are spoken in their respective states. The rest of the languages are spoken in some regions around the Gulf of Finland, and Lakes Ladoga and Onega. The Livonian language hasn't been examined in our study since there are no more native speakers left. As for the Karelian language that has several widely used dialects, in our LI system, we do not make a distinction between them.

There are many similarities between those languages, especially when it comes to grammar and morphology. They have very similar orthographies, although there are some particularities. For instance, *õ* letter occurs only in Estonian and Votic, while *š* and *ž* letters in Finnish and Estonian, in contrast to the rest of the languages, occur only in loanwords. Finnic languages are agglutinative, and due to their genetic and geographical closeness, many affixes across the languages are very similar.
Table 1 provides some morphological comparisons that point to that similarity.

	Elative case	**Finite verb: 1.PL**
Finnish	sta/stä	mme
Estonian	st	me
Karelian	sta/stä/šta/štä	mma/mmä
Veps	s/š/se/sa/ša	mei
Ingrian	st	mma/mmä
Votic	s/ssa	mma/mmä/mmõ

Table1. Comparison of some affixes in Finnic languages.

Most of the languages of this group, except Finnish and Estonian, are considered to be minority and low-resource languages. Karelian (with all of the dialects combined) has around 30,000[1] speakers, Veps has somewhere from 1,000[2] to 3,000[3] speakers, Ingrian has 120[4] speakers, and Votic has just 25[5]. While Karelian and Veps are used socially, i.e. there are newspapers published in those languages, and they are taught at primary schools in Russia, Ingrian and Votic are considered to be nearly extinct, and there are scarcely any text documents in those languages available.
All of that makes automatic LI for Finnic languages more difficult. We have to take into account the many similarities between those languages, as well as the sparsity of training data. It also means that document LI is unavailable, therefore we have to deal with sentence and word level LI.

4. Data

Before performing our experiments, we gathered training data. For Finnish, Estonian, Karelian and Veps we used newspaper articles from contemporary newspapers. For Ingrian

[1] https://www.ethnologue.com/language/krl
[2] https://www.ethnologue.com/language/vep
[3] http://www.gks.ru/free_doc/new_site/population/demo/per-itog/tab6.xls
[4] https://www.ethnologue.com/language/izh
[5] https://www.ethnologue.com/language/vot

and Votic no such data was available and therefore we used texts from manuals of those languages and scientific papers that featured fragments of texts written those languages.
The newspapers crawled were "Helsingin Sanomat"[6] for Finnish language, "Postimees"[7] for Estonian, "Oma Mua"[8] for Karelian, and "Kodima"[9] for Veps.
Table 2 sums up the number of words and sentences in the training data.

	Sentences	**Unique words**
Finnish	1,386	7,450
Estonian	1,024	6,562
Karelian	1,847	8,138
Veps	1,821	5,490
Ingrian	155	639
Votic	293	1,051
Total	**6,526**	**29,330**

Table 2. Dataset sizes

All documents were preprocessed before the experiments. All non-alphabetic symbols were removed, and spaces were replaced with underscores for visual clarity. Everything was also converted to lowercase. Documents in each language were combined in one and split into sentences.

5. Experiments

We have chosen a character n-gram model as our feature model. We used overlapping n-grams of length from 2 to 4 characters, underscores were also included in the n-gram to take into account the beginning and end of the word. Out of all the n-grams 200 most frequent ones were chosen for the processing of training data. Since the classes were disproportionate, first, the most frequent 50 n-grams were chosen for each language. All top n-grams were then combined in one list, the repeating n-grams were deleted, and out of that final list, 200 n-grams were chosen for the feature model. Each object in the training data was converted to a vector of counts of those frequent n-grams.
Several classifiers were trained and tested. Parameters that fit the data best were chosen for the final evaluation.
The classifiers are:

[6] https://www.hs.fi
[7] https://www.postimees.ee
[8] http://omamua.ru
[9] http://kodima.rkperiodika.ru

1. Neural Network (NN).

Sequential NN was built using Keras[10] library. The network had 3 layers with a sigmoid activation function and an output layer with softmax activation to produce probabilities of each language. Due to the imbalance of the classes, class weights were calculated and introduced to the model.

2. Support Vector Machine (SVM) with linear kernel.
3. Random Forest (RF) with 250 estimators.
4. Logistic Regression (LR).
5. k Nearest Neighbours (kNN) with k=5 and uniform weights.

All models were evaluated with a 5-fold cross-evaluation.

The results are presented in the Table 3.

	NN	**SVM**	**kNN**	**LR**	**RF**
Sentences	0.9562	0.9249	0.7846	0.9172	0.9124
Words	0.8334	0.6265	0.5781	0.6268	0.6257

Table 3. Evaluation results

As can be seen, the NN has shown the best results in both experiments.

Table 4 present the confusion matrix for NN trained on sentences:

	Predicted language					
	Estonian	**Finnish**	**Ingrian**	**Karelian**	**Veps**	**Votic**
Estonian	**192**	3	1	7	6	1
Finnish	2	**267**	6	6	1	1
Ingrian	0	5	**30**	0	0	1
Karelian	1	3	0	**346**	9	3
Veps	1	2	0	8	**341**	1
Votic	2	3	2	3	2	**50**

Table 4. Confusion matrix for NN (sentences)

We also tried to improve the word level LI by adding morphological features (i.e. affixes that occur in a particular language), but it had the opposite effect. Combination of NN and this affix-based system yielded a 0.7947 accuracy, and therefore, it is not presented in this paper.

[10] https://keras.io

6. Conclusion

In this paper, we described several implementations of the LI for the Finnic languages. The challenges we faced were associated with the close relatedness of languages and scarcity of linguistic data for most of those languages. We chose to test several multi-class classifiers: Neural Network, Support Vector Machine, k Nearest Neighbors, Logistic Regression and Random Forest.

The results obtained in our experiments show that high accuracy LI on a sentence level can be achieved even for similar and low resource languages. Perhaps, NN can be further improved by some parameter modification, but so far the 0.9562 accuracy seems satisfactory. As for the word level LI, additional research needs to be conducted. We suppose that adding some other features may improve the accuracy of identification, though it is yet unclear which features may help since morphological information failed to make the system better. Another improvement may consist of adding more training data, and since it is difficult to find large texts in Ingrian and Votic, a dictionary may be used. Nevertheless, our results may be of use for further research into low resource Finnic languages: our LI system may help with automatic corpora building, code-mixing identification and language preservation.

References

1. *King B.P* (2015) Practical Natural Language Processing for Low-Resource Languages, PhD dissertation. Online version: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/113373/benking_1.pdf
2. *Ljubešić N. and Kranjčić D.* (2015) Discriminating Between Closely Related Languages on Twitter. Informatica, 39, 2015. pp. 1-8.
3. *Lui M. and Baldwin T.* (2014) Accurate Language Identification of Twitter Messages. In Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM), Gothenburg, Sweden, 2014, pp. 17–25.
4. *Simaki V., Simakis P., Paradis C., and Kerren A.* (2017) Identifying the Authors' National Variety of English in Social Media Texts. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017), Varna, Bulgaria, September 2017, pp. 671–678. URL https://doi.org/10.26615/978-954-452-049-6_086.
5. *Viitso T.-R. (1998)* Fennic. The Uralic Languages. Routledge Language Family Descriptions. London – New York: Routledge. pp. 96-114.