European master in Language and Communication Technologies

LARGE-SCALE LANGUAGE IDENTIFICATION FOR CLOSELY RELATED LANGUAGES

Master's thesis.

By:

M·MEDVEDEVA, BA

Supervisors:

dr. B·PLANK[†]
prof. dr. D·KLAKOW[‡]

RIJKSUNIVERSITEIT GRONINGEN[†]
&
UNIVERSITÄT DES SAARLANDES[‡]

XXI.VI.MMXVII

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine andere als angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, ________________ ________________________________
(date) (signature)

Abstract

This is the abstract

Acknowledgements

I would like to thank my supervisors, Dr. Barbara Plank and Prof. Dr. Dietrich Klakow, for there enormous help, support, dedication and enthusiasm, during the entire time of my work on this thesis.

I would also like to say special thanks to my LCT coordinators, Dr. Gosse Bouma at University of Groningen and Dr. Jürgen Trouvain at Saarland University, for making this journey between universities smoother, especially when it came to writing the thesis.

And last, but certainly not least, I would like to express my deepest gratitude to my family and friends, especially Martin Kroon, for their love, support and fruitful discussions on linguistics, the universe and everything over dinner table and long-distance calls.

Contents

1. Introduction

- Language Identification in general
- Language identification for similar languages
- Minority languages
- Language influence & loan words
- Conclusion: we want create tools for languages with very few resources (while having very few resources) & make it universal enough to add more languages given enough data

2. Related work

- which language identifiers are there
- which similar languages
- which minority languages
- overview of methods
- some assumptions/conclusions on the way to approach a universal language identifier
- a huge table overview (in appendix?): year/author/model/main features

2.1 Bullet points

- domain
- number of languages
- tested on (document, sentence, word, snippet)
- approach
- representation (byte/char/word ngrams, words, weighted letters, etc)

In this chapter we will discuss the research in language identification that has already been conducted. As we've mentioned before, there were a lot of attempts to solve this problem; from the very start the attempts showed very promising results (**cavnar1994n**; **dunning1994statistical**). Even though many papers show language identification problem as a solved problem due to their near perfect results (CITE CITE), there are certain aspects of it that are still a challenge. The main problem

Brown 2012, 2013, 2014 (USED A 50 TO 1000 LINES FOR TESTING. WEIRD) is trained on up to 1366 languages, but it is predominantly trained and tested on Bible, Wikipedia and Europarl corpus of European parliamentary proceedings (Koehn, 2005FIX CITATION), and even though it has a very high accuracy, it is unclear how well it performs on out-of domain data. Thus we should not only concentrate on increasing the coverage, but also simultaneously improving the robustness by building systems that either trained on more domains or not as domain-dependant, thus more general.

When talking about language identification systems, we need to focus not only on the algorithms, but also on languages, how well the models performs within different domains, ho many languages do they support, and if they can be retrained on more languages, does the performance stay the same. on what size of the text can the predication be accurate enough? is it a whole document? How long is that document? a sentence? a word? what is the representation used for the training? how are the systems evaluated.

A full list of systems discussed can be found in TABLE.

2.2 From DSL

The problem of automatic language identification has been a popular task for at least the last 25 years. From early on, different solutions showed very high results (**cavnar1994n**; **dunning1994statistical**), while the more recent models achieve near-perfect accuracies.

Distinguishing closely-related languages, however, still remains a challenge. The *Discriminating between similar languages* (DSL) shared task (**vardial2017report**) is aimed at solving this problem. For this year's task our team (mm_lct) built a model that discriminates between 14 languages or language varieties across 6 language groups (which had two or three languages or language varieties in them).[1]

The most popular of the more recent systems, such as `langid.py` (**lui2012langid**) and CLD/CLD2[2] produce very good results based on datasets containing fewer than 100 languages, but even a model trained on as many as 131 languages (**kocmi2017lanidenn**) and whatlang (**brown2013selecting**) with trained on 184 and 1100 languages, are not able to distinguish closely-related (and therefore very similar) languages and dialects to a satisfying degree, at least not to the extent of the data available.

As part of the DSL 2017 shared task we chose to further explore traditional linear approaches, as well as deep learning methods. In the next Section we shortly discuss previous approaches to the task of discriminating between similar languages. Then in Section **??** we describe our systems and the data, followed by the results in Section **??**, which are discussed in Section **??**. We conclude in Section **??**.

Even though a number of researches in dialect identification have been conducted, (**tiedemann-ljubesic:2012:COLI** **lui2013classifying**; **maier2014language**; **ljubesic2015discriminating**), they mostly deal with particular language groups or language variations. We saw as our goal to create a language identifier that is able to produce comparable results for languages within all provided groups with the same set of features for every language group, so that it can be expanded outside those languages provided by the DSL shared task without any changes other than to the training corpus – as to make the system as language-independent and universal as possible.

Most of the language identifiers that use linear classifiers rely on character n-gram models (**carter2011semi**; **ng2011improving**; **zampieri2012automatic**) and combinations of character and word n-grams (**milne2012study**; **vogel2012robust**; **goldszmidt2013boot**), also including top systems from previous DSL shared tasks (**goutte-leger:2015:LT4VarDial**; **malmasi-dras:2015:LT4VarDial**; **ccoltekin-rama:2016:VarDial3**).

The overviews of the previous DSL shared tasks (**zampieri:2014:VarDial**; **zampieri:2015:LT4VarDial**; **dslrec:2016**) showed that SVMs always produce some of the top results in this task, especially when tested on same-domain datasets (**ccoltekin-rama:2016:VarDial3**). Thus, we chose to put our efforts into improving upon SVM approaches, but still decided to experiment with an neural network to see if we could get comparable results, while using fewer features and reducing the chance of overfitting.

The popularity of using NNs for NLP tasks is growing. A few neural language identifiers already exist as well (**tian2003scalable**; **takcci2012minimal**; **simoes2014language**), however on average traditional systems still seem to outperform them. The results of the DSL 2016 shared task also show the same tendency overall (**bjerva:2016:VarDial3**; **cianflone-kosseim:2016:VarDial3**; **ccoltekin-rama:2016:VarDial3**; **dsl2016**).

[1] The term *language* shall henceforth be used for both 'language' and 'language variety'.

[2] `https://github.com/CLD2Owners/cld2`

2.3 Related Work

Even though a number of researches in dialect identification have been conducted, (**tiedemann-ljubesic:2012:COLING** **lui2013classifying**; **maier2014language**; **ljubesic2015discriminating**), they mostly deal with particular language groups or language variations. We saw as our goal to create a language identifier that is able to produce comparable results for languages within all provided groups with the same set of features for every language group, so that it can be expanded outside those languages provided by the DSL shared task without any changes other than to the training corpus – as to make the system as language-independent and universal as possible.

Most of the language identifiers that use linear classifiers rely on character n-gram models (**carter2011semi**; **ng2011improving**; **zampieri2012automatic**) and combinations of character and word n-grams (**milne2012study**; **vogel2012robust**; **goldszmidt2013boot**), also including top systems from previous DSL shared tasks (**goutte-leger:2015:LT4VarDial**; **malmasi-dras:2015:LT4VarDial**; **ccoltekin-rama:2016:VarDial3**).

The overviews of the previous DSL shared tasks (**zampieri:2014:VarDial**; **zampieri:2015:LT4VarDial**; **dslrec:2016**) showed that SVMs always produce some of the top results in this task, especially when tested on same-domain datasets (**ccoltekin-rama:2016:VarDial3**). Thus, we chose to put our efforts into improving upon SVM approaches, but still decided to experiment with an neural network to see if we could get comparable results, while using fewer features and reducing the chance of overfitting.

The popularity of using NNs for NLP tasks is growing. A few neural language identifiers already exist as well (**tian2003scalable**; **takcci2012minimal**; **simoes2014language**), however on average traditional systems still seem to outperform them. The results of the DSL 2016 shared task also show the same tendency overall (**bjerva:2016:VarDial3**; **cianflone-kosseim:2016:VarDial3**; **ccoltekin-rama:2016:VarDial3**; **dsl2016**).

2.4 From proposal

Language identification is one of the most challenging, yet incredibly relevant tasks in computational linguistics. Well-performing systems are necessary for both conducting linguistic research and creating various applications that work with texts.

There are however very few systems that can identify any minority languages, while many of them have written traditions and are largely represented on the Internet. These language communities are often under-resourced and without language recognition systems they are slowed down in their progress in developing various instruments both for researching the respective languages and for creating tools that have been available in all major languages for a long time.

A sub-task of language identification is identification of languages within multilingual texts. It is a much more challenging task as the identification has to happen on much smaller character sequences (i.e. word-level). There have been some attempts of solving this problem for texts written by speakers of certain languages (**nguyen2013word**; **maharjan2015developing**), but those are pairs of unrelated languages; today's challenge is to figure out the way to distinguish between related languages and languages that have been in a very long contact. This problem can be viewed as quite similar to dialect distinction, as they are both recognised by particular words(/morphemes/syntax) standing out of overall context. Therefore I would like to propose a system that would be a solution to all of these problems at once.

Being a novelty in computational linguistics, Neural Networks have not been used much for solving the language identification problem before. In my thesis I would like to take advantage of this method.

2.4.1 Data

As this problem has been explored before, quite a lot of corpora have been created already. There are traditional monolingual language corpora, code-switching corpora, monolingual twitter corpora (**vrl2014accurate**), code-switching twitter corpora (**maharjan2015developing**; **vrl2010multilingual**), etc.

As part of this work I would also be building a corpus with larger diversity of phenomena. There are 93 indigenous minority languages spoken just in Russia, according to Ethnologue ((**lewis2009ethnologue**); the number should not be seen as precise because of the language vs. dialect uncertainty). These are languages of different families, with different (and similar) writing systems, some them have been in long contact and some have not. Almost all of them have translations of the Bible, Wikipedia articles, blogs and Twitter accounts written entirely in their language, some of them already have annotated corpora. Putting my background of working with languages of Russia and being a native speaker of Russian to use I could create a languages of Russia database to test how well language identification works on less explored languages.

2.4.2 Method

There are multiple open-source language identification systems available, including `langid.py` (**lui2012langid**), whatlang (**brown2013selecting**), YALI (**majlivs2012yet**), TextCat (**cavnar1994n**), MSR-LID (**goldszmidt2013boot**), etc. They however were trained on different data and therefore their performance is not comparable, as there is a high possibility that they only perform well within their own domains. Moreover, obviously some languages are easier to distinguish between than others and the distribution within those systems might be very different.

As part of the thesis work I will train existing systems on datasets of each other in order to have a proper comparison and to be able to see which are the strong suits of one system against the other (i.e. one works better with sorter sequences, one performs properly only on social media, good to distinguish languages of different families, etc.). By training the systems on all datasets together and in different combinations we'll be able to build a better training set and see if a combination of classifiers might be a good solution.

The focus will be on the closely related languages and languages that have been in a very long contact (and therefore, share a lot of lexicon). Therefore, a finer-grained classification for shorter sequences will be needed. Language identification within multilingual texts will not be a priority, and the system will not be trained on such data. However, languages that borrowed large amounts of lexicon from contact languages can be considered very similar to multilingual texts, and thus I will test the system on code-switching data as well to see how it performs.

The main approach will be trying to use Deep Learning in order to see if it improves the performance. In order for the system to perform well on dialect distinction task and multilingual texts the identification will have to be on n-gram or word-level, or a combination of two.

I am however restricted by computing-power limitations and therefore a scheme of reducing the training data will have to be introduced as well.

Therefore, the main goals of the project are to reduce the training data needed for building a comparable to state-of-the-art language identification system, as well as focusing on large minority languages and closely related data and dialects. The minimum requirement for the data is to be determined: the initial approach is going to be ‘the more the better’, and tested throughout the process (on different simpler classifiers) in order to see whether the results keep improving with the amount of data and whether certain languages require more data due to a lot of borrowings from other languages, etc.

3. Data Collection

Being from Russia and having graduated with a bachelor degree in linguistics in Moscow, I have been a part of multiple projects related to studying minority languages in Russia. My experience showed that there is a lot of effort in Russian linguistic community towards creating resources for those languages.

There are 93 indigenous minority languages spoken in Russia according to Ethnologue ((**lewis2009ethnologue**); the number should not be seen as precise because of the language vs. dialect uncertainty). 78 of these languages are written and over the half of those are prominently represented on the Internet. Some of these languages have Wikipedia articles, blogs and twitter accounts written entirely in their language. The languages belong to different families, but their written orthography, with the exception of a handful of Finnic languages, is based on Cyrillic alphabet. Note, that it does not mean that they all have the same alphabet

maybe move up to the introduction

While many of these languages are largely represented on the Web, they still remain under-resourced. The majority of the languages used in this thesis for language identification have never been tried for this task before.

With the large amount of research in linguistics and NLP being done with the text resources published online, we need to be able to identify the languages of smaller languages in Russia in order to be able to collect substantial amount of data in those languages to conduct linguistic research.

3.1 Data sources

The data collection for this work is largely relying on the data provided by a project by students at the School of Linguistics at National Research University Higher School of Economics that were creating minority languages corpora as part of their master theses.[1] I did not use their ready-to-use corpora, but crawled the links referencing web-pages in particular languages using the crawlers developed within the same project. As they do not have a name I will refer to their project as *Minorlangs*, as the link of their website suggests.

Minorlangs provides a variaty of different resources to collect a corpus, mainly being urls of particular pages, websites of domains that are likely to be written entirely in the minority language (so, not only the text from that url has to be extracted, but all the texts from the links on that page refer to and all the links that those pages refer to on the same website, etc.), as well as the links of webpages from various social media, such as Facebook,[2] Twitter,[3], a question answering website ASKfm[4], Russian biggest social media website Vkontakte[5] and other.

Add a table with distribution of the resources and a few words about it

[1] http://web-corpora.net/wsgi3/minorlangs/about

[2] facebook.com

[3] twitter.com

[4] ask.fm

[5] vk.com

3.2 Data collection methods

To collect the websites *Minorlangs* uses Yandex.XML tool, which allows users to receive the responses to searches in Yandex[6], the most used search engine in Russia.

The authors have manually compiled a list of lexical markers unique to particular minority languages, i.e. most common function words that are unique for this particular language and do not contain diacritic or character that do not appear on the Russian keyboard (as they are often omitted or changed in writing).

Then based on those markers up to 1000 requests per day could be made (the limit set by the service). Certain links were ignored, i.e. known websites with the info about the languages, websites with scientific texts that may include research on minority languages, music and video websites, containing only names of the songs or videos but not full texts, as well as Wikipedia pages. For full description of the system's architecture see (CITE)

cite

I rely on the *Minorlang* project to have used the truly unique markers to identify the websites in particular languages. While I do clear out pages and posts written entirely in Russian, I assume that there is no mix-up between the other languages.

Using the links and a crawler that extracts texts from the pages (provided by the project as well) I extract the data in minority languages from the websites. This data however cannot be used straight away. The crawler collects a lot of meta-information (e.g. mark-up of the pages), such as the dates of the blog posts and navigation links between different parts of the websites. As most of the blogs and news resources are hosted on the Russian websites and do not have that information translated. In order to exclude those uses I exclude the lines that are less than 4 words long from the dataset.

3.3 Filtering Russian out

The long contact with Russian language was bound to influence the minority languages on the territory of Russian Federation. My experience of working with some of these languages (see i.e. **clif**) showed that they are prone to loan a lot of words from Russian. A lot of words do not have an equivalent in the smaller language, and some are just very common code-switches. However, the websites that I am crawling also contain a lot of texts entirely in Russian, sometimes due to the mistakes of the request responses and sometimes just because people write (and comment) in different languages on the same website. Note that the vast majority of speakers of minority languages are bilingual, as well as literate in both languages. It is save to assume that people who speak the minority languages also fluently speak Russian (even with large efforts to save minority languages in Russia, the education is still provided primarily in Russian). Of course keeping the texts in Russian will only reduce the accuracy of the language identification system by mixing up the labels and introducing noise, but filtering out Russian words completely would also be counter-intuitive and would make the data less representative of the real language usage. *Minorlangs* also removes tokens in English (including names of the companies and websites) and emojis. I feel that this information should be kept to preserve the authenticity of the data. I do, however, remove the lines written entirely in Arabic, which is common on the predominantly Muslim territories.

Minorlangs identifies the languages with character trigram model of Russian language and uses it on the word level, thus excluding instances of all Russian words (using not the most reliable

[6] yandex.ru

approach), while as I mentioned before, languages of Russia are largely influenced by Russian and have a significant number of loan words from Russian, which do not always adopt the smaller language morphology (to distinguish with a trigram model) and should definitely be considered part of the language. Thus, I used a much more simplified model. When extracting data, I my system splits it into sentences with an nltk sentence tokenizer (CITE) and then splits those sentences into tokens with an nltk tokenizer, removes the punctuation marks and lowers the register. The words from the sentence are then compared to the unigram list consisting the entire corpus of multi-domain corpus of Russian language, OpenCorpora[7] (164465 unique tokens). If more than the half of the words in the sentence are in the list, the sentence is considered in Russian and is thrown out, otherwise the sentence is treated as belonging to the minority language. This way I can keep longer code-switches without introducing too much noise to the data. The sentences are then added to file, with the label and the url that they came from.[8] Overall, 2145666 sentences have eliminated due to overwhelming amount of Russian words.

cite something

talk about the tokenizers

is this a correct word in English?

3.4 Cross-domain data

Minorlangs also provides the links for downloading the data from Russian social media website Vkontakte for 41 out of 47 languages above. I have crawled that data as well, but have not used it as part of the main data I am training the system on. This data is used to determine the influence of increased training data from a different domain as well as cross-domain testing. We report on the results further in this work.

make a reference to results section

I have not actually downloaded it yet. It does not work, sent an email to the 'project' email, did not get a response yet. I they won't soon, I'll email their supervisor, he used to be my professor. Should be easy to fix, it is just missing a config file, I am sure someone must have it. Add to the table later + write a few sentences in the overview

3.5 Overview of the collected data

A full page report. Talk about how the speakers don't correspond to the presence online

Do not have a full written text for an overview yet. SHould I do it at all. It is not related to the project directly, but it is interesting. The minorlangs website provides their own statistics on the amount of links they collected, but as half of them don't work anymore, I fell like the lines and tokens vs. number of speakers is more interesting.
This is the data after filtering out Russian.
Should I add the vk.com data (when I have it) as 2 additional columns?
The f-scores (word unigram model) will not be in this table. It is only here to see whether the size matters. and the results are quite interesting.
Should I Northern Yukaghir because it is too tiny and only ruins everything?

This work is aimed at developing tools for low-resourced languages, and investigating the (lower) limits of data needed for successful language identification, I did not concentrate as much on collecting as much data as possible. Instead I have only used the data from separate urls (and not the

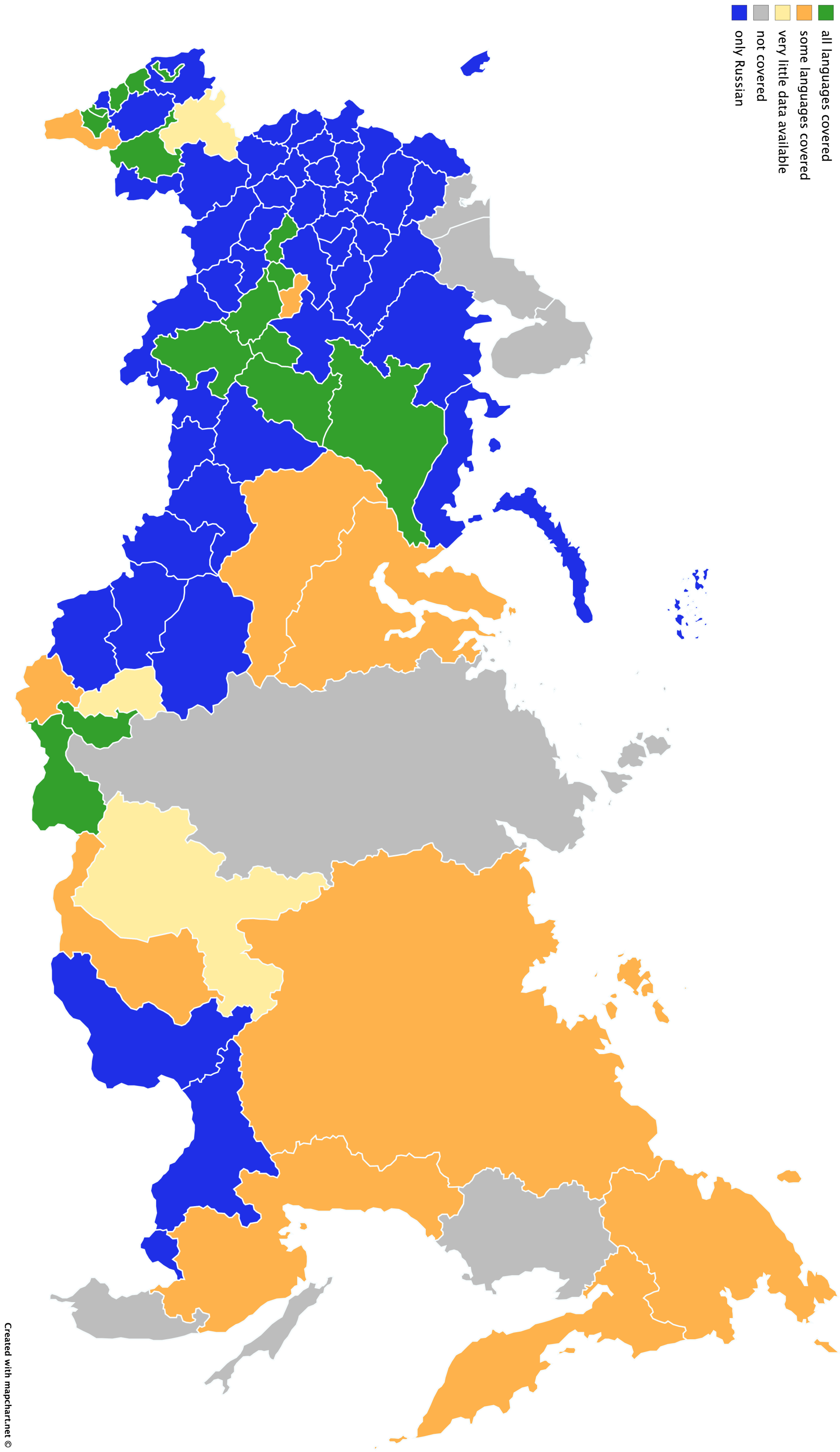
[7] http://opencorpora.org/

[8] The data is available MAKE A REPO WITH DATA ASK ABOUT COPYRIGHTS

entire domains) and Vkontakte data. You can see the overview of the collected data in the Table (REFERENCE). The table is sorted by the number of sentences collected per language.

reference. Keep here or as an appendix?

lang	lang family	native speakers	id	lines	tokens		f1
Tatar	Turkic	4280000	tat	416428	4208681		0.92
Chechen	Nakho-Dagestanian	1354705	che	201007	1863447		0.91
Bashkir	Turkic	1150000	bak	125134	1170574		0.92
Kabardian	Abkhazo-Adyghean	515672	kbd	102305	885434		0.88
Buryat	Mongolic	283000	bxr	101286	987638		0.88
Chuvash	Turkic	1152404	chv	86599	637411		0.83
Eastern & Meadow Mari	Uralic	365316	mhr	82981	700941		0.89
Yakut	Turkic	450140	sah	82836	727122		0.82
Ingush	Nakho-Dagestanian	305868	inh	55123	512757		0.71
Erzya	Uralic	431692	myv	52984	372254		0.82
Komi-Zyrian	Uralic	160599	kpv	48947	352996		0.73
Udmurt	Uralic	324338	udm	47574	404315		0.81
Tuvinian	Turkic	253673	tyv	46509	343603		0.83
Lezghian	Nakho-Dagestanian	402173	lez	43710	343389		0.91
Southern Altai	Turkic	55720	alt	32298	239846		0.94
Karachay-Balkar	Turkic	305364	krc	27705	186895		0.85
Nogai	Turkic	88328	nog	26805	344323		0.85
Avar	Nakho-Dagestanian	715297	ava	26478	153654		0.86
Adyghe	Abkhazo-Adyghean	117489	ady	24426	204491		0.71
Evenki	Tungusic	4800	evn	23999	381639		0.86
Gilyak	isolate	198	niv	22642	244880		0.66
Lak	Nakho-Dagestanian	145895	lbe	17214	108980		0.93
Moksha	Uralic	2025	mdf	15163	165356		0.69
Khakas	Turkic	42604	kjh	13545	158186		0.93
Kalmyk	Mongolic	80546	xal	12549	94406		0.83
Komi-Permyak	Uralic	63106	koi	9972	53919		0.65
Nanai	Tungusic	1347	gld	9058	57540		0.75
Chukot	Chukotko-Kamchatkan	5095	ckt	6292	48756		0.66
Nenets	Uralic	21926	yrk	5698	42505		0.77
Shor	Turkic	2839	cjs	3237	20962		0.81
Tabassaran	Nakho-Dagestanian	126136	tab	3182	16766		0.52
Kumyk	Turkic	426212	kum	2879	19522		0.52
Udi	Nakho-Dagestanian	2266	udi	2798	19823		0.85
Abaza	Abkhazo-Adyghean	37831	abq	1893	13527		0.68
Muslim Tat	Indo-European	26000*	ttt	1531	25404		0.91
Dargwa	Nakho-Dagestanian	42000	dar	1183	7928		0.74
Tsakhur	Nakho-Dagestanian	10596	tkr	1111	7126		0.53
Southern Yukaghir	Uralic	50	yux	1101	7531		0.82
Koryak	Chukotko-Kamchatkan	1665	kpy	1002	5819		0.72
Kubachi	Nakho-Dagestanian	3400	dar2	721	9825		0.52
Karagas	Turkic	93	kim	651	4173		0.84
Mansi	Uralic	938	mns	584	3415		0.45
Archi	Nakho-Dagestanian	970	aqc	152	1769		0.22
Rutul	Nakho-Dagestanian	30360	rut	139	1618		0.82
Khanty	Uralic	9584	kca	113	984		0.78
Even	Tungusic	5656	eve	86	809		0.69
Northern Yukaghir	Uralic	30-150	ykg	34	245		0.22



add map as an appendix

4. Methods and design

- split languages into different categories based on the amount of data we have and into domains
- describe the kind of experiments we run: feature extraction, setups
  - baseline: ?SVM with unigrams
  - using all data:
Naive Bayes (baseline?)
SVMs: different features, different parameters, preprocessing
NNs: different features, different parameters, preprocessing
  - the same with different sizes of datasets (test set or cross-val?)
  - the same with cross-domain training/testing (in various combinations)

mention that I excluded short phrases, but It does not mean that there are no short sentences, because I split lines into sentences

5. Results and discussion

- compare to TextCat
- ?compare to best DSL 2017 (DSL setup/Russian data)
- ?compare to PAN 2017 (PAN setup/Russian data)
- ?compare to state-of-the-art re-trainable language identification systems (their setup/Russian data)
- try retraining our best system on other datasets, for instance DSL data/our setup, etc to see if it is universally better or is just tuned to Russian languages specifically.

Discussion of why things worked and did not work.

6. Conclusion

Are there any complex features that contain some sort of language description that helps identifying it? Or does everything hurt?