VarDial in the Wild: Industrial Applications of LID Systems for Closely-Related Language Varieties

Fritz Hohl
Sony Europe B.V., Stuttgart, Germany
fritz.hohl@sony.com

Soh-Eun Shim
Sony Europe B.V., Stuttgart, Germany
soh-eun.shim@sony.com

Abstract

This report describes first an industrial use case for identifying closely related languages, e.g. dialects, namely the detection of languages of movie subtitle documents. We then present a 2-stage architecture that is able to detect macrolanguages in the first stage and language variants in the second. Using our architecture, we participated in the DSL-TL Shared Task of the VarDial 2023 workshop. We describe the results of our experiments. In the first experiment we report an accuracy of 97.8% on a set of 460 subtitle files. In our second experiment we used DSL-TL data and achieve a macro-average F1 of 76% for the binary task, and 54% for the three-way task in the dev set. In the open track, we augment the data with named entities retrieved from Wikidata and achieve minor increases of about 1% for both tracks.

1 Introduction

In the NLP community the problem of identifying languages of documents is often perceived as being solved (Zampieri et al., 2023), also due to the good accuracy of this function in tools like Google Translate. This is especially true for many users as they apply this method in cases where they want to understand text that is not their field of native speaker expertise. However, when applying state of the art language identification tools to applications where an accurate distinction of closely related languages, e.g. dialects is important, it soon becomes clear that these tools often either do not offer variants in their list of covered languages or confuse them regularly. One of these application areas are movie subtitles. As we will see in the next section, although these texts are typically not too small for language identification, they often differ from news domain content, which is the source of the shared task data. Section 3 will describe the architecture of our system. Using this architecture, we participated in the DSL-TL Shared Task of the VarDial 2023 workshop. We describe this Shared Task briefly in Section 4. Our experiments of our system on this Shared Task and other data can be found in Section 5, while Section 6 presents a manual oracle experiment that aims at finding out how much an extended NER-like mechanism can reduce errors. Finally, Section 7 conclude our findings.

Format Title	File Extension
DCTitle format	xml
TTML	xml
Flashplayer TTAF	xml
SMPTE-TT (extension of TTML)	xml
TTML	dxfp
TTML	itt
CAP	cap
STL	stl
Scenarist_SCC V1.0	scc
SRT	srt
WEBVTT	vtt

Table 1: Example subtitle file formats.

2 LID for Subtitles

Subtitle files contain the Closed Captions or Subtitles of movies and similar video content. These files come in a variety of different, partially proprietary formats (see Table 1 for some of them).

The content consists typically of a mix of time stamps, dialogue lines, textual descriptions of visual content, and symbols, e.g. for music (see Fig. 1 for an example extract of such content).

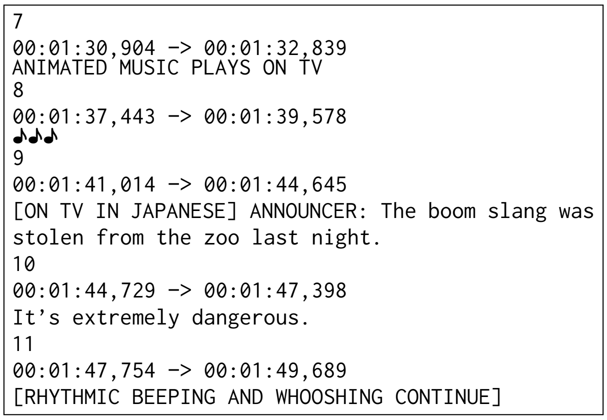
In order to cope with the diversity of formats, and to extract the textual parts in the target language, a preprocessing stage is needed. Afterwards, UTF-8-encoded text can be fed into the Language Identification stage. In our experience the resulting subtitle text documents have a median file size of about 25 kbytes. Subtitles and Closed Captions

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 213–221
May 5, 2023 ©2023 Association for Computational Linguistics

are meant to be displayed over the video for a certain amount of time (hence the time stamps). The displayed text for this period can contain either single words, parts of sentences, or multiple sentences. These text portions are reflected by line breaks in the document, i.e. line breaks separate different time periods. Apart from textual descriptions and symbols, texts transcribe mainly the spoken dialogue. DVD and BluRay releases of video content often come with a number of subtitles in different languages and language variants.[1] Also, releases of these media in different regions or countries come with different sets of subtitle languages. As a result, for a single piece of video content, many different subtitle documents exist. If one imagines that e.g. for a single movie, different versions of that movie are needed even in a single language (realizing different ratings, different cuts, or being trailer versions), the number of subtitles documents over an entire catalogue of movies is very large. Ideally, a digital asset management system denotes the language of a single subtitle file. However, in reality, movie studios have a very large catalogue of content that partially predates the widespread availability of low-threshold asset management platforms. This means that there are many subtitle files on many disks in many drawers of which no exact metadata is known. This is where Language Identification really helps as it avoids the need of re-creating subtitle files for existing movies if they are about to be re-released. Also, it helps to verify existing language metadata as these might be incorrect. The user of such a tool will be typically a technician, not a linguist. Also, the user will probably have to face a lot of different languages, some of which will be very foreign to the user. Finally, it will be very difficult for the user to get a gold language label for a file using standard tools. Subtitles are relatively mono-lingual; they might contain a moderate amount of code-switching and the occasional sentence in another language.

3 System Architecture

The requirements for the use case mentioned in Section 2 asked for a system that could not only recognize macrolanguages, but also language variants. An earlier internal language identification system implementation based purely on character n-grams and perplexity proved to be especially



```
7
00:01:30,904 -> 00:01:32,839
ANIMATED MUSIC PLAYS ON TV
8
00:01:37,443 -> 00:01:39,578
♪♪♪
9
00:01:41,014 -> 00:01:44,645
[ON TV IN JAPANESE] ANNOUNCER: The boom slang was
stolen from the zoo last night.
10
00:01:44,729 -> 00:01:47,398
It’s extremely dangerous.
11
00:01:47,754 -> 00:01:49,689
[RHYTHMIC BEEPING AND WHOOSHING CONTINUE]
```

Figure 1: Subtitle file extract.

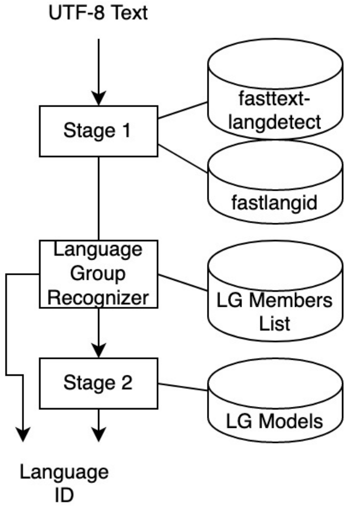


Figure 2: Overall architecture.

[1] For four releases of the movie “Bullet Train” from 2022 alone, we counted 30 different subtitle languages.

weak when it comes to closely-related languages and dialects. Consulting the literature (Goutte et al., 2014; Zampieri et al., 2015) we decided in favor of a two-stage system, where the first stage aims at recognizing a "language group". For us, Language Groups are groups of closely related languages that are difficult to distinguish in the first stage. Language Groups can be macrolanguages that contain e.g. different language variants, or simply a set of languages that are hard to tell apart for a stage 1 algorithm. If the language in stage 1 is marked in a list as being a member of a certain Language Group, a stage 2 takes over and aims at determining the concrete member of the language group. Using this architecture, not all language group member languages need to be recognized in stage 1. Let's now have a closer look at the two stages (see Figure 2).

3.1 Stage 1

For stage 1, our system follows the general approach of Goutte et al. (2014); Zampieri et al. (2015). More concretely, we use the 126 MB-model of fasttext-langdetect (Joulin et al. (2016a,b)) directly. This package uses pretrained fastText embeddings for language identification and provides support for 176 languages. The only difference is that we additionally use the model from fastlangid[2] in order to cure the inexplicable weakness of fasttext-langdetect not to distinguish traditional and simplified Chinese.

3.2 Stage 2

Our second stage utilizes an SVM with 1- to 4-character n-gram features, along with word unigram features. SVMs have been shown in prior work to have strong baselines, which have consistently outperformed RNNs in prior experiments (Çöltekin et al., 2018). All n-gram features are weighted with sub-linear tf-idf scaling. The SVM models are trained with scikitlearn (Pedregosa et al., 2011).

4 The DSL-TL Shared Task at VarDial 2023

The Shared Task on "Discriminating Between Similar Languages - True Labels" (DSL-TL) aims at examining the effects of a data set for identifying sentences in similar languages that have been gold-labelled in a new way. Previously, the gold labels for such sentences have been derived from the country (and therefore the language variant) association of the source of a sentence, e.g. a newspaper that is primarily published in a certain country. This method is problematic if e.g these sentences do not contain a variant-specific markers and, thus, do not help an automatic mechanism to determine a language variant. The new way to label sentences is using multiple human annotators to determine a variant label while offering a labeller to also specify that a sentence is not variant-specific. The subject of this labelling campaign has been nearly 13k sentences in a number of language varieties, namely English (American and British), Portuguese (Brazilian and European), and Spanish (Argentinian and Peninsular). The DSL-TL webpage[3] explains the Shared Task in more detail and contains a link to the data used in the Shared Task. (Zampieri et al., 2023) explains the new dataset, the annotation process that led to this dataset, and the performance of baseline algorithms on the dataset. The results of all teams and the shared tasks will be explained in (Aepli et al., 2023).

Language code	# of files	Language code	# of files
bg	2	ms	1
da	3	no	3
de	5	pl	6
el	1	PT-PT	9
en	225	PT-BR	13
es	103	ru	7
fi	3	sr	1
fr	14	sv	3
hu	3	th	4
is	3	tr	4
it	9	zh-hans	6
ja	23	zh-hant	2
mr	7		

Table 2: Subtitle evaluation data.

5 Experiments on Subtitle Files

One of the original use cases for our system is subtitle file language identification.

5.1 Data

Stage 1 was used out of the box; no further training was used. The Language Group models of Stage 2 were trained on prior years of DSLCC data (Tan

[2] https://github.com/currentslab/fastlangid

[3] https://sites.google.com/view/vardial-2023/shared-tasks#h.klf8c6mlh0zk

et al., 2014) for the internal system. The evaluation data set consists of 460 subtitle files that have been converted to UTF-8 text. Table 2 shows the distribution of these files to Gold labels. Gold labels have been raised mainly by a single person taking into account the content of the files, language hints in the filenames that sometimes occur (but which are sometimes also wrong), and existing language identification tools. As most of the labels denote languages which can be easily distinguished, these labels are expected to be correct. One group of labels, though, posed quite a challenge. 22 files belong to the Language Group "Portuguese" with the two members pt-PT and pt-BR. These files were labelled by a native Brazilian Portuguese speaker who expressed doubts on the reliability of his judgments on these files.

5.2 Results

The content in the files was concatenated, then processed by the system as described in Section 4. We also limited the set of language groups and variants to those we expected to be contained in the test data set (because in our experience, the probability to correctly identify the standard variant in a language group is smaller than to identify the macrolanguage). From the 460 files, 451 (or 98.0%) were recognized correctly. The 9 error cases can be divided as follows:

- 7 cases confused pt-PT with pt-BR. In 6 of these cases the file name hints to the possibility that these files might have been indeed created as pt-BR.
- 2 cases were extremely short files (in fact these were the smallest files of the evaluation set).

So, on this evaluation set our system seems to perform quite well as long as the content size is not too small.

6 Experiments on DSL-TL Data

Now we will describe our experiments for the DSL-TL Shared Task. Until mentioned otherwise, all stages have been trained on DSL-TL training data. In Section 6.2 the system will be tested on the DSL-TL dev set, in Section 6.3 on the DSL-TL test set.

6.1 Predicting Macrolanguages for DSL-TL

Regarding stage 1 of our architecture, we wondered with respect to the DSL-TL task whether our existing stage 1 trained on 177 languages (i.e. with data outside the DSL-TL datasets) would perform worse than a stage 1 trained purely on the three language families of the Shared Task using only DSL-TL training data.

In our experiment, applied to the combined DSL-TL dev set data of all languages, the only difference was that our existing stage 1 incorrectly predicted one Argentinian Spanish sentence as Italian (this sentence was "`19. Lucas di Grassi (BRA/Virgin-Cosworth): 1min24s547)`", which, considering the Italian origin of the last name, arguably constitutes a reasonable error. All other stage 1 predictions (also from the DSL-TL-only stage 1) were correct.

6.2 Results on DSL-TL Dev Data

In observance of the influence Named Entities potentially have upon the task, we ran two experiments on the DSL-TL data. First, for the closed task, we varied the number of maximum word n-grams added to observe the difference in performance. Second, we also experimented with adding named entities as retrieved from a linked open database (Vrandečić and Krötzsch, 2014) to the data as our submission to the open task, which allows for the usage of external data. We observe a consistent improvement in both the binary and three-way task by way of this method.

6.2.1 Word n-gram Features

Table 3 shows our results of increasing the maximum number of word n-gram features on the dev data in the binary task. Table 4 shows our results per class. Table 5 and Table 6 show the same results for the three-way classification task. Our results replicate the conclusion made by Çöltekin et al. (2018): increasing the number of word n-gram features becomes useful up to a certain point, after which the effect either levels out or starts hurting performance. We hypothesize that this is due to higher n-gram numbers capturing named entities, but once the granularity exceeds what is typical for a named entity, the features start to lose predictive power.

6.2.2 Adding Named Entities

Our second approach experiments with using additional NER data as retrieved from Wikidata (Vrandečić and Krötzsch, 2014). This NER data consists of 10k person names per country associated with the 6 language variants (i.e. the US, the UK, Spain,

	0	1	2	3
pt	0.66	0.66	**0.68**	0.68
en	0.80	0.81	**0.82**	0.82
es	0.74	**0.76**	0.75	0.75

Table 3: Macro averaged F1 on dev data binary classification task, where the columns indicate the maximum number of word n-gram features used.

	0	1	2	3
PT-BR	0.83	0.82	**0.84**	0.84
PT-PT	0.48	0.5	**0.52**	0.52
ES-ES	0.84	**0.85**	0.85	0.85
ES-AR	0.65	**0.67**	0.66	0.65
EN-GB	0.75	0.77	**0.79**	0.79
EN-US	0.84	0.85	**0.86**	0.86

Table 4: Per class F1 on dev data binary classification task, where the columns indicate the maximum number of word n-gram features used.

Argentina, Portugal, and Brazil). From these lists we selected a number of names randomly per sentence per language variant and added these names as training features to the sentence. Our results for the binary task are listed in Table 7, and the three-way results are noted down in Table 9. Table 8 and Table 10 shows our per class results. We observe that in the binary case, Spanish sees an improvement of 2% while Portuguese and English deteriorate in performance; in the three-way case however, the opposite phenomenon is observed, where considerable improvements are seen for both Portuguese and English, but not Spanish.

6.3 Results on DSL-TL Test Data

The results of our system on Open and Closed Tasks, on Task 1 (three-way labels) and Task 2 (binary labels) as macro averages per language and per single label, as well as the rank of our result among the baselines and the other teams can be found in Table 11. Our results can be found under the team name "SSL" in (Zampieri et al., 2023).

	0	1	2	3
pt	**0.42**	0.42	0.42	0.42
en	0.52	0.54	**0.55**	0.54
es	**0.52**	0.51	0.52	0.52

Table 5: Macro averaged F1 on dev data three-way classification task, where the columns indicate the maximum number of word n-gram features used.

	0	1	2	3
PT	0.04	**0.05**	0.03	0.03
PT-BR	0.78	0.77	**0.79**	0.79
PT-PT	0.43	**0.45**	0.44	0.45
ES	**0.40**	0.39	0.39	0.39
ES-ES	**0.69**	0.67	0.68	0.68
ES-AR	0.46	**0.48**	0.48	0.48
EN	0.07	**0.13**	0.12	0.10
EN-GB	0.70	0.71	**0.74**	0.73
EN-US	0.78	**0.79**	0.79	0.79

Table 6: Per class F1 on dev data three-way classification task, where the columns indicate the maximum number of word n-gram features used.

7 Country-Informed NER Oracle Performance

We were wondering how much a slightly different approach would fare using an NER-like system that could tell one the country a Named Entity is typically associated with. One (probably inefficient) way to implement this system would be to google a maximum-length Named Entity candidate (the English DSL-TL data is very consistently capitalized in terms of Named Entities) and to take the first country reference that is found in the results (maybe normalizing the result, e.g. from “English” to “UK”). When a Knowledge Panel appears in the results, this information should first be taken into account. As we cared only about oracle performance, we did this process manually.

In order to reduce the English dev set sentences to manually label this way, we only looked at the 92 sentences (from 599) our automatic method from above incorrectly predicted.

For each of these sentences, we marked the capitalized Proper Nouns and then started a Google search. We also looked at mentioned currencies. An example of this data can be found in Table 12.

When comparing the labels from the Oracle mechanism to the DSL-TL labels, we found four groups:

- In 32 cases (34.8%) the Oracle mechanism found the TL label.

- In 25 cases (27.2%) the Oracle mechanism did not come to a conclusion as either no usable Named Entities were found or there was no majority for a country of the found Named Entity (e.g. there were two names associated with the US and two with the UK).

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
pt	**0.68**	0.67	0.64	0.59	0.57	0.58	0.52	0.51	0.47	0.48
en	**0.82**	0.81	0.82	0.81	0.78	0.77	0.75	0.75	0.74	0.69
es	0.75	**0.77**	0.77	0.77	0.75	0.77	0.77	0.77	0.76	0.77

Table 7: Macro averaged F1 on dev data binary classification task, where the columns indicate the number of person names appended to each training instance.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
pt-pt	**0.53**	0.5	0.45	0.34	0.3	0.33	0.22	0.19	0.13	0.14
pt-br	**0.84**	0.83	0.83	0.83	0.83	0.83	0.82	0.82	0.82	0.82
en-gb	**0.79**	0.78	0.79	0.76	0.73	0.71	0.67	0.67	0.65	0.58
en-us	0.85	0.85	**0.86**	0.85	0.84	0.83	0.83	0.83	0.82	0.81
es-es	**0.85**	0.85	0.85	0.85	0.84	0.84	0.84	0.84	0.82	0.83
es-ar	0.66	0.68	0.69	0.69	0.67	0.7	0.7	0.7	0.7	**0.71**

Table 8: Per class F1 on dev data binary classification task, where the columns indicate the number of person names appended to each training instance.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
pt	0.42	**0.46**	0.46	0.38	0.30	0.24	0.19	0.15	0.13	0.10
en	0.55	0.58	**0.62**	0.57	0.53	0.45	0.36	0.30	0.24	0.20
es	**0.53**	0.49	0.40	0.33	0.28	0.22	0.20	0.19	0.18	0.18

Table 9: Macro averaged F1 on dev data three-way classification task, where the columns indicate the number of person names appended to each training instance.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
pt	0.03	0.16	0.28	**0.28**	0.26	0.26	0.25	0.24	0.24	0.24
pt-pt	**0.45**	0.43	0.34	0.17	0.09	0.06	0.03	0.01	0.02	0.01
pt-br	**0.78**	0.78	0.76	0.68	0.54	0.41	0.28	0.2	0.13	0.07
en	0.14	0.2	**0.33**	0.31	0.32	0.29	0.26	0.25	0.24	0.24
en-gb	**0.73**	0.73	0.73	0.66	0.55	0.38	0.24	0.12	0.05	0.04
en-us	**0.8**	0.8	0.79	0.75	0.71	0.67	0.58	0.52	0.42	0.34
es	0.0	0.0	0.0	0.0	0.0	0.43	0.48	**0.5**	0.5	0.5
es-es	0.71	0.71	**0.72**	0.72	0.71	0.69	0.6	0.46	0.33	0.21
es-ar	0.5	**0.52**	0.52	0.51	0.49	0.47	0.39	0.25	0.16	0.13

Table 10: Per class F1 on dev data three-way classification task, where the columns indicate the number of person names appended to each training instance.

Type	Track	Results for	Recall	Precisio	F1-score	Rank
Closed	Track 1	Macro Average	0.4978	0.4734	0.4817	12 of 13
Closed	Track 1	“en” label	0	0	0	14 of 14
Closed	Track 1	“en-GB” label	0.7807	0.7063	0.7417	9 of 14
Closed	Track 1	“en-US” label	0.8462	0.763	0.8024	6 of 14
Closed	Track 1	“es” label	0.3205	0.3623	0.3401	13 of 14
Closed	Track 1	“es-AR” label	0.4135	0.5046	0.4545	10 of 14
Closed	Track 1	“es-ES” label	0.767	0.6371	0.696	5 of 14
Closed	Track 1	“pt” label	0	0	0	11 of 14
Closed	Track 1	“pt-PT” label	0.8997	0.7079	0.7923	1 of 14
Closed	Track 1	“pt-BR” label	0.4526	0.5794	0.5082	7 of 14
Closed	Track 2	Macro Average	0.7521	0.7885	0.7604	8 of 15
Closed	Track 2	“en-GB” label	0.7895	0.7895	0.7895	10 of 15
Closed	Track 2	“en-US” label	0.8526	0.8471	0.8498	10 of 15
Closed	Track 2	“es-AR” label	0.5789	0.828	0.6814	10 of 15
Closed	Track 2	“es-ES” label	0.9223	0.7724	0.8407	5 of 15
Closed	Track 2	“pt-PT” label	0.9097	0.7861	0.8434	1 of 15
Closed	Track 2	“pt-BR” label	0.4599	0.7079	0.5575	11 of 15
Open	Track 1	Macro Average	0.4937	0.5068	0.4889	1/1
Open	Track 1	“en” label	0.1333	0.1481	0.1404	1/1
Open	Track 1	“en-GB” label	0.693	0.7248	0.7085	1/1
Open	Track 1	“en-US” label	0.8205	0.7711	0.795	1/1
Open	Track 1	“es” label	0.4038	0.3772	0.3901	1/1
Open	Track 1	“es-AR” label	0.3609	0.4948	0.4174	1/1
Open	Track 1	“es-ES” label	0.7379	0.658	0.6957	1/1
Open	Track 1	“pt” label	0.322	0.1473	0.2021	1/1
Open	Track 1	“pt-PT” label	0.7525	0.7401	0.7463	1/1
Open	Track 1	“pt-BR” label	0.219	0.5	0.3046	1/1
Open	Track 2	Macro Average	0.7647	0.7951	0.7729	2 of 2
Open	Track 2	“en-GB” label	0.7544	0.8037	0.7783	2 of 2
Open	Track 2	“en-US” label	0.8718	0.8293	0.85	2 of 2
Open	Track 2	“es-AR” label	0.6917	0.8288	0.7541	2 of 2
Open	Track 2	“es-ES” label	0.9078	0.8202	0.8618	2 of 2
Open	Track 2	“pt-PT” label	0.9097	0.7839	0.8421	1 of 2
Open	Track 2	“pt-BR” label	0.4526	0.7045	0.5511	2 of 2

Table 11: Results on DSL-TL Test data.

sentence	TL label	Oracle label
The **Grenfell Tower** fire shifted the tectonic plates of **British** society, triggering a wave of investigations and renewing a national conversation about social housing. One year on, **Jack Hardy** reviews the major episodes from a traumatic year.	EN-GB	EN-GB
EASTLEIGH'S rapidly rising star **Luke Coulson** is putting club before country. The **Spitfires**' 22-year-old league top scorer will sacrifice his place in the **England** C squad in order to play in Tuesday's **FA Cup** first round replay at **Swindon Town**.	EN-US	EN-GB
GOOD Samaritans cornered three loose ponies which galloped through oncoming traffic on the A31 on Monday night.	EN-GB	None

Table 12: Example oracle data.

- In 31 cases (33.7%) the Oracle mechanism found a label different from the TL label, but we would have labeled the sentence differently. Our opinion was mainly informed by the subject matter and the country of the Named Entities as we are of the opinion that it would be only of local interest and therefore could have been published only by a local newspaper.

  Sometimes we could also trace the sentence to a newspaper from a certain country. This opinion is obviously based on our belief of the intention-based criteria for documents of a language variant. Therefore, the reader might either count this group as errors or as correct cases. We discussed our suspicion that the labelers are heavily using their knowledge of the country of Named Entities in order to come to a label. For this group, we basically claim that the labelers did not follow that suspicion.

- In 4 cases (4.3%) the Oracle mechanism found the wrong label and we agree that the TL label is correct.

8 Conclusion

We reported on our use case for Language Identification, namely movie subtitles. For movie subtitles there is a need to also recognize close languages and variants.

We presented the 2-stage architecture of our Language Identification system that uses a second stage if a language is identified in the first stage that is marked as being a member of a "language group". We reported on the results of our experiments. In the first experiment we reported an accuracy of 97.8% on a set of 460 subtitle files. In our second experiment we used DSL-TL data and achieved for the dev set a macro-averaged F1 of 54% in the three-way classification task, and 76% in the binary classification task, where we see an increase in performance by adding Named Entities retrieved from a knowledge base. On the DSL-TL test set for the closed task we achieved a macro-averaged F1 of 48% in the three-way classification task, and 76% in the binary classification task. On the open task, we achieved 49% for the three-way, and 77% on the binary task.

We reported on a small experiment using a manually executed "country-informed" NER on those sentences of the English DSL-TL dev set that were incorrectly predicted by our system. We did this in order to see how much head room remains in the NER-based approach to identify DSL-TL data as some sort of oracle data. As it turns out, this mechanism can reduce the number of errors by at least a third. As future work we intend to examine the question whether it is better to keep language group languages separated for training stage 1 or whether the data for a language group should be mixed together before training in order to achieve a better stage 1 performance.

There are two aspects that are not clear to us when it comes to the DSL-TL dataset, and that might lead to future research. First, how strongly influence named entities the human labellers and is this detrimental to the label accuracy? To quote (Goutte et al., 2016)

> [...] named entities [...] can influence [...] also the performance of human annotators.

Second, again (Goutte et al., 2016) mentions that a

> general tendency we observed is that it is easier to identify an instance that is not from the speaker's own language than the opposite. Our results indicate that humans are better in telling what is not a text written in their own language or variety than telling what it is.

We understand from the description of the new annotation process that a labeller could annotate sentences in his/her own, or another variant. As the ratio of native speakers for the corresponding variants seems not to be mentioned, it is hard to assess the influence of this aspect on the results. Maybe it would be worthwhile to trying to evaluate the accuracy of the new labels. This is, of course, not easy as then another, more accurate process would be needed to come to "platinum" labels.

Limitations

The NER-based extensions to our base 2-stage algorithm increase the accuracy only for documents that contain enough Named Entities. Such documents can be found in news domains, but not in all other domains. Depending on the implementation, the extensions might additionally rely on a correct capitalization in the document. Basically, all Language Identification systems assume that a document is using the written version of a variant. If a document is transcribing the spoken variant, it will have problems to be processed.

Acknowledgements

We would like to thank the reviewers who provided us with valuable comments. We also want to thank Sony's Culver City Laboratory (CCL) for the collaboration in this area, and to Sony Pictures Entertainment and Sony Music Publishing for the provided data.

References

Noëmi Aepli, Çağrı Çöltekin, Rob van der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. Association for Computational Linguistics.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 2018. Tübingen-Oslo team at the VarDial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 55–65, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 139–145, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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 BUCC Workshop*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels. *arXiv preprint arXiv:2303.01490*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.