Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models

Tommi Jauhiainen
Department of Digital Humanities
University of Helsinki
tommi.jauhiainen@helsinki.fi

Heidi Jauhiainen
Department of Digital Humanities
University of Helsinki
heidi.jauhiainen@helsinki.fi

Krister Lindén
Department of Digital Humanities
University of Helsinki
krister.linden@helsinki.fi

Abstract

This paper describes the language identification systems used by the SUKI team in the Discriminating between the Mainland and Taiwan variation of Mandarin Chinese (DMT) and the German Dialect Identification (GDI) shared tasks which were held as part of the third VarDial Evaluation Campaign. The DMT shared task included two separate tracks, one for the simplified Chinese script and one for the traditional Chinese script. We submitted three runs on both tracks of the DMT task as well as on the GDI task. We won the traditional Chinese track using Naive Bayes with language model adaptation, came second on GDI with an adaptive version of the HeLI 2.0 method, and third on the simplified Chinese track using again the adaptive Naive Bayes.

1 Introduction

The third VarDial Evaluation Campaign (Zampieri et al., 2019) included three shared tasks on language, dialect, and language variety identification. The Discriminating between Mainland and Taiwan variation of Mandarin Chinese (DMT) concentrated on finding differences between the varieties of Mandarin Chinese written on mainland China and Taiwan. The task included two tracks, one for the simplified script and another for the traditional one. The German Dialect Identification (GDI) task was already the third of its kind (Zampieri et al., 2017, 2018). In GDI 2019, the task was to distinguish between four Swiss-German dialects. The third task was that of Cuneiform Language Identification (CLI), but we did not participate in that as we were partly responsible for creating its dataset (Jauhiainen et al., 2019a).

We evaluated several language identification methods using the development sets of the DMT and GDI tasks. Our best submissions were created using a similar language model (LM) adaptation technique to the one we used in the second VarDial Evaluation Campaign (Zampieri et al., 2018). In that Evaluation Campaign, we used the HeLI language identification method (Jauhiainen et al., 2016) together with a new LM adaptation approach, winning the Indo-Aryan Language Identification (ILI) and the GDI 2018 shared tasks with a wide margin (Jauhiainen et al., 2018b,c). After the second Evaluation Campaign, we had developed a new version of the HeLI method and further refined the LM adaptation technique (Jauhiainen et al., 2019b). With the HeLI 2.0 method and the refined adaptation technique, we came second in the GDI 2019 shared task using only character 4-grams as features. Furthermore, we had implemented several baseline language identifiers for the CLI shared task (Jauhiainen et al., 2019a). One of them was a Naive Bayes (NB) identifier using variable length character *n*-grams, which fared better than the HeLI method on the CLI dataset. We modified our LM adaptation technique to be used with the NB classifier and this fared better than the adaptive HeLI 2.0 method on both of the Chinese datasets. With the adaptive NB identifier, we won the traditional Chinese track and came third on the simplified one.

In this paper, we first go through some related work in Section 2, after which we introduce the datasets and the evaluation setup used in the DMT and the GDI shared tasks (Section 3). We then use the training and the development sets to evaluate our baseline methods (Sections 4.1 and 4.2) and the HeLI 2.0 method (Section 4.3), after which we evaluate the efficiency of our LM adaptation procedure with the HeLI 2.0 and NB methods in Sections 4.4 and 4.5. Finally we introduce and discuss the results of our official submissions (Section 5) as well as give some conclusions and ideas for future work (Section 6).

Proceedings of VarDial, pages 178–187
Minneapolis, MN, June 7, 2019

2 Related work

In this section, we introduce some background information on previous studies in language identification in general, language identification in the context of Chinese and German languages, as well as LM adaptation.

2.1 Language identification in texts

Language identification (LI) is the task of identifying the language of a text. The same methods which are used for LI are generally also used for dialect and language variety identification. A comprehensive survey of language identification in general has been published in arXiv by Jauhiainen et al. (2018d).

The series of shared tasks in language identification began in 2014 with the Discriminating Between Similar languages (DSL) shared task (Zampieri et al., 2014) and similar tasks have been arranged each year since (Zampieri et al., 2015; Malmasi et al., 2016; Zampieri et al., 2017, 2018).

It is notable that, so far, deep neural networks have not gained an upper hand when compared with the more linear classification methods (Çöltekin and Rama, 2017; Medvedeva et al., 2017; Ali, 2018).

2.2 Chinese dialect identification

In the DMT shared task, the text material in the dataset is UTF-8 encoded. Before the widespread use of UTF-8, different encodings for different scripts were widely used. Li and Momoi (2001) discussed methods for automatically detecting the encoding of documents for which an encoding was unknown. They present two tables showing distributional results for Chinese characters. In their research they had found that the 4096 most frequent characters in simplified Chinese encoded in GB2312 cover 99.93 percent of all text and they report that earlier results of traditional Chinese in Big5 encoding are very similar with the 4096 most frequent characters covering 99.91% of text.

Huang and Lee (2008) used a bag of words method to distinguish between Mainland, Singapore and Taiwan varieties of Chinese. They reached an accuracy of 0.929.

Brown (2012) displays a confusion matrix of four varieties of the Chinese macrolanguage as part of his LI experiments for 923 languages. The Gan and Wu Chinese were among the languages with the highest error rates of all languages.

Huang et al. (2014) show how light verbs have different distributional tendencies in Mainland and Taiwan varieties of Mandarin Chinese. Using K-Means clustering they show that the varieties can be differentiated.

Xu et al. (2016) describe an approach to distinguish between several varieties of Mandarin Chinese: Mainland, Hong Kong, Taiwan, Macao, Malaysia, and Singapore. In another study (Xu et al., 2018), they used support vector machines (SVM) to distinguish between Gan Chinese dialects.

2.3 German dialect identification

The GDI 2019 task was already the third of its kind (Zampieri et al., 2017, 2018). In 2017, we did not participate in the shared task, which was won using an SVM meta-classifier ensemble with words and character *n*-grams from one to six as features (Malmasi and Zampieri, 2017). We won the 2018 edition using the HeLI method with LM adaptation and character 4-grams as features (Jauhiainen et al., 2018b). We were the only ones employing LM adaptation and won with a wide margin to the second system which was an SVM ensemble using both character and word *n*-grams (Benites et al., 2018).

For a more complete overview of dialect identification for the German language, we refer the reader to our recent paper where we used LM adaptation with the datasets from the GDI 2017 and 2018 shared tasks (Jauhiainen et al., 2019b). Our experiments using a refined LM adaptation scheme with the HeLI 2.0 method produced the best published identification results for both of the datasets.

2.4 Language model adaptation

Language model (LM) adaptation is a technique in which the language models used by a language identifier are modified during the identification process. It is advantageous especially when there is a clear domain (topic, genre, idiolect, time period, *etc.*) difference between the texts used as training data and the text being identified (Jauhiainen et al., 2018a,b,c). If an adaptation technique is successful, the language identifier learns the peculiarities of the new text and is better able to classify it into the given language categories. In the shared tasks of the VarDial Evaluation Campaigns, we are provided with the complete test sentence collection at once. This means, that we can addi-

tionally choose in which order we learn from the test data and even process the same test sentences several times before providing the final language labels.

The LM adaptation technique and the confidence measure we use in the systems described in this article are similar to those used earlier in speech language identification by Chen and Liu (2005) and Zhong et al. (2007). The adaptation technique is an improved version of the one we used in our winning submissions at the second VarDial Evaluation Campaign (Jauhiainen et al., 2018b,c). For a more complete overview of the subject, we refer the reader to our recent article dedicated to language model adaptation for language and dialect identification of text, where we also introduce the improved LM adaptation technique used in this paper (Jauhiainen et al., 2019b).

3 Test setup

In the shared tasks 2019, the participants were provided with separate training and development sets. All the tracks were closed ones, so no external information was to be used in preparing the language identification systems. The training and development sets were released approximately a month before the test set release. When the test sets were released, the participants had two days to submit their predictions on the tracks. The texts in the development portions could be used as an additional training data when processing the test sets and we did so in each case.

The evaluation measure used in both of the shared tasks was the macro F1-score and we used it also when comparing the different methods we used with the development data.

3.1 DMT datasets

The scripts commonly used in mainland China and Taiwan are different. In Taiwan, the traditional Chinese script is commonly used whereas in mainland China, the simplified version is the official one (Chen et al., 1996; Huang et al., 2000; McEnery and Xiao, 2003). In order to be able to concentrate on the non-scriptual differences of the two varieties of Mandarin Chinese, Putonghua (Mainland China) and Guoyo (Taiwan), the texts used for the DMT task had been transformed to use the same script. In the simplified track, the Taiwanese texts originally written in the traditional script had been converted into the simplified script and in the traditional track the texts from mainland China originally in the simplified script had been converted to the traditional script. The conversion had been made using a tool called "OpenCC".[1]

The texts used as the source for the datasets were news articles from mainland China and from Taiwan. The participants were provided with training and development sets for both simplified and traditional scripts. Both datasets had been tokenized by inserting whitespace characters between individual words. Furthermore, all punctuation had been removed. The average length of words in all DMT training sets was *c.* 1.7 characters. The training sets contained 9,385 sentences and the development sets consisted of additional 1,000 sentences for each variety. The test sets had 1,000 sentences as well for each variety.

3.2 GDI dataset

The GDI dataset consisted of transcribed speech utterances in four Swiss German dialects. More detailed information about the source of the texts for the GDI datasets, the ArchiMob corpus, are provided by Samardžić et al. (2016). In 2018, the GDI dataset included additional unknown dialects, which were left out in 2019. The sizes of the training and the development sets can be seen in Table 1. The average length of words in the training set was 5.5 characters. The test set contained 4,743 utterances comprising 42,699 words. As of this writing, we are not aware of the distribution of the dialects in the test set.

Variety (code)	Training	Development
Bern (BE)	27,968	7,381
Basel (BS)	26,927	9,462
Lucerne (LU)	28,979	8,650
Zurich (ZH)	28,833	8,086

Table 1: List of the Swiss German varieties used in the datasets distributed for the 2019 GDI shared task. The sizes are in words.

The training, development, and test sets included two files in addition to the speech transcriptions. The first file included normalized forms for each dialectal form in the data. The second file included 400-dimensional iVectors representing the acoustic features of the original speech data, as the text data was transliterated speech. We did not use either of the two additional files in our experiments.

[1] https://github.com/BYVoid/OpenCC

4 Experiments using the development data

We set out to tackle the GDI and the DMT shared tasks with the system based on the HeLI 2.0 method and LM adaptation that we had used for the GDI 2017, GDI 2018 and ILI datasets between the 2018 and 2019 VarDial Evaluation Campaigns (Jauhiainen et al., 2019b).

For the CLI shared task, we had implemented three new baseline identifiers, one of which, a Naive Bayes identifier, managed to overcome the traditional HeLI method when distinguishing between Sumerian and six Akkadian dialects (Jauhiainen et al., 2019a). We were, hence, also interested to see how well these baseline identifiers would perform in the DMT and GDI tasks.

4.1 Simple scoring and the sum of relative frequencies

The first baseline method in the CLI shared task was the simple scoring method. In simple scoring, the frequency information of individual features in the training set is ignored and each time a feature from a language model $dom(O(C_g))$ is encountered in the text M, the language g is given one point. The survey article by Jauhiainen et al. (2018d) gives the following Equation 1 for simple scoring:

$$R_{simple}(g, M) = \sum_{i=1}^{l_{MF}} \begin{cases} 1 & \textbf{, if } f_i \in dom(O(C_g)) \\ 0 & \textbf{, otherwise} \end{cases} \quad (1)$$

where l_{MF} is the number of individual features in the line M and f_i is its ith feature. The language g gaining the highest score R is selected as the predicted language.

The second baseline implementation for the CLI used the sum of relative frequencies of character *n*-grams of varying length. The method is very similar to simple scoring but, instead of simply adding a global constant to the score each time the feature is found in the language model, the observed relative frequency in the respective languages training corpus is added.

In both methods, the only parameter to be decided when using the development data was the range of the character *n*-grams used. These character *n*-grams can span word boundaries and thus long *n*-grams can contain several words. We experimented with a range from 1 to 20 characters and the best attained macro F1-scores on the development sets are listed in Tables 2, 3, and 4.

In the end, we did not submit any results using the two first baseline methods as the third baseline method, the product of relative frequencies, was clearly superior to them.

4.2 Product of relative frequencies (NB)

Our third baseline method in CLI was the product of relative frequencies. The method is basically the same as Naive Bayes using the observed relative frequencies of character *n*-grams as probabilities. As with the two previous methods, these character *n*-grams can span word boundaries. Similarly to the sum of relative frequencies method, we calculate the relative frequencies for different *n*-grams from the training corpus, but instead of adding them together, we multiply them as in Equation 2:

$$R_{prod}(g, M) = \prod_{i=1}^{l_{MF}} \frac{c(C_g, f_i)}{l_{C_g^F}} \quad (2)$$

The practical implementation uses the sum of logarithms instead as computers normally cannot handle the extremely small numbers produced by multiplying the observed probabilities of complete sentences. As smoothing, in case $c(C_g, f_i)$ was equal to zero, we used 1 and multiplied the resulting logarithmic value by the penalty modifier p_{mod}. The penalty modifier and the character *n*-gram range used were optimized using the development set. As mentioned earlier, the NB classifier bested the other baseline methods as can be seen in Tables 2, 3, and 4.

4.3 HeLI 2.0

In the HeLI method, we calculate a score for each word using relative frequencies of words or character *n*-grams. The length of the character *n*-grams to use or whether to use the word itself is decided individually for each word encountered in the text to be identified. The whole text gets the average of the scores of the individual words, thus giving equal value to long and short words. No information spanning word boundaries are used.

We have recently introduced a version of HeLI which we decided to call 2.0 as enough changes to the HeLI method had already accumulated (Jauhiainen et al., 2019b). The HeLI 2.0 differs from the HeLI method described by Jauhiainen et al. (2016) in three ways. Firstly, we now always use

Method	*n*-gram range	Smoothing	Splits k	Epochs	*CMmin*	F1 dev
Naive Bayes with LM adaptation	1–15	1.3	*max*	1	0.45	0.9225
Naive Bayes	1–15	1.3	-	-	-	0.9215
Simple scoring	1–15	-	-	-	-	0.8970
HeLI 2.0 with LM adaptation	1–2 + infinite	1.01	*max*	1	-	0.8909
HeLI 2.0	1–2 + infinite	1.01	-	-	-	0.8859
Sum of rel. freq.	5–15	-	-	-	-	0.8204

Table 2: Simplified Chinese. The macro F1-scores attained by different methods on the development set. A *max* in column indicating the number of splits means that k was equal to the number of lines in the evaluation data.

Method	*n*-gram range	Smoothing	Splits k	Epochs	*CMmin*	F1 dev
Naive Bayes with LM adaptation	1–14	1.3	4	1	0	0.9295
Naive Bayes	1–14	1.3	-	-	-	0.9285
HeLI 2.0 with LM adaptation	1–2 + infinite	1.12	*max*	1	0.42	0.9160
HeLI 2.0	1–2 + infinite	1.12	-	-	-	0.9145
Simple scoring	1–6	-	-	-	-	0.9015
Sum of rel. freq.	5–15	-	-	-	-	0.8247

Table 3: Traditional Chinese. The macro F1-scores attained by different methods on the development set. A *max* in column indicating the number of splits means that k was equal to the number of lines in the evaluation data.

Method	*n*-gram range	Smoothing	Splits k	Epochs	*CMmin*	F1 dev
HeLI 2.0 with LM adaptation	4	1.12	9	112	0.15	0.8657
Naive Bayes with LM adaptation	2–6	1.08	40	96	0.16	0.8442
HeLI 2.0	4	1.12	-	-	-	0.6658
Naive Bayes	2–6	1.08	-	-	-	0.6475
Simple scoring	2–7	-	-	-	-	0.5865
Sum of rel. freq.	6–15	-	-	-	-	0.5049

Table 4: GDI 2019. The macro F1-scores attained by different methods on the development set. A *max* in column indicating the number of splits means that k was equal to the number of lines in the evaluation data.

all of the possible training material and use neither rank- nor relative frequency-based cut-off. Secondly, we changed how we calculate the smoothing value. In HeLI, we used a global penalty value for all language models. In HeLI 2.0, we calculate the penalty value relative to the size of each individual language model using a global penalty modifier p_{mod}.[2] Thirdly, when selecting the range of character *n*-grams to use in HeLI 2.0, the minimum size for *n* can be higher than one.

In this description, we define a word as a character *n*-gram of infinite size generated from an individual word. The model sizes used are optimized using a development corpus as is the possible use of *n*-grams of infinite size. Variable n_{max} is the maximum length and the n_{min} the minimum length of the used character *n*-grams.

The corpus derived from the training data containing only the word internal[3] *n*-grams of the size n for the language g is called C_g^n. The values $v_{C_g^n}(f)$ for the *n*-gram f are calculated for each language g, as shown in Equation 3:

$$v_{C_g^n}(f) = \begin{cases} -\log_{10}\left(\frac{c(C_g^n, f)}{l_{C_g^n}}\right) & \text{, if } c(C_g^n, f) > 0 \\ -\log_{10}\left(\frac{1}{l_{C_g^n}}\right) p_{mod} & \text{, if } c(C_g^n, f) = 0 \end{cases} \quad (3)$$

where $c(C_g^n, f)$ is the number of *n*-grams f in C_g^n.

The domain $dom(O(C^n))$ is the set of all character *n*-grams of length n found in the models of any of the languages $g \in G$. Separately for each individual word t on the line M to be identified, we determine the length n of the character *n*-grams we use. The word t is divided into overlapping character *n*-grams of the length n. The length n is the highest where at least one of the character *n*-grams generated from the word t is found in $dom(O(C^n))$. However, if an individual *n*-gram f generated from the word t is not found in $dom(O(C^n))$, it is discarded at this point. The number of retained *n*-grams is l_{t^F}. The score for individual words t on the line M is then calculated as in Equation 4.

[2] This is the same smoothing method which we used with the product of relative frequencies.

[3] The beginning and the end of a word are marked using whitespaces.

$$v_g(t) = \sum_{i=1}^{l_{tF}} v_{C_g^n}(f) \quad (4)$$

The whole line M is then scored as in Equation 5:

$$R_g(M) = \frac{\sum_{i=1}^{l_{MT}} v_g(t_i)}{l_{MT}} \quad (5)$$

where l_{MT} is the number of words in the line M. The predicted language g of the line M is the one having the lowest score. We optimized the penalty modifier p_{mod} as well as the minimum and maximum size of the *n*-grams, i.e. n_{min} and n_{max}, using the developments set. Whether or not to use the charcter *n*-grams of infinite size (words), was also decided with the development set. The best results attained by the HeLI 2.0 method on each of the development sets can be seen in Tables 2, 3, and 4.

4.4 HeLI 2.0 with adaptive language models

The fifth method we used in the experiments with the development data was the domain-adaptive version of HeLI 2.0. We used a similar LM adaptation method in the shared tasks of VarDial 2018, clearly winning the GDI and ILI tasks. For Jauhiainen et al. (2019b), we devised an improved version of the adaptation method, which is used here. In order to select the best material to be used in LM adaptation, we need a confidence measure which indicates the best identified lines. In Jauhiainen et al. (2019b), we evaluated three confidence measures and the score difference between the best and the second best scoring languages, CM_{BS}, proved to be the best performing one. The confidence measure is calculated as in Equation 6:

$$CM_{BS}(M) = R_h(M) - R_g(M) \quad (6)$$

where g is the best and h the second best scoring language.

For the shared tasks, we get a complete set of lines to be identified as one collection. We denote an individual line M, as before, and the set of lines is denoted MC. In the adaptation algorithm, we first perform a preliminary identification using the HeLI 2.0 method for each line M of the development or the test set MC. The number of lines a to process simultaneously in adaptation is the number of lines in MC divided by the number of splits k. The number of splits is optimized using the development set. For each line, we also calculate the confidence measure CM_{BS}. We remove a most confident lines from MC and mark them as finally identified with the given language labels. Then we add the information from the finally identified lines to their respective language models. Then we use the new language models to re-identify the lines remaining in MC, again using the a most confident lines to augment the language models. This process is repeated until all the lines in MC have been removed. In the iterative version of the adaptation method, the whole adaptation process is repeated several times (epochs).

For all submissions but one,[4] we used a confidence threshold when deciding whether the information from a line was added to the language models, that is not all lines were always used for adaptation. The confidence threshold, $CMmin$, was also optimized for the development set.[5] The number of splits, k, and the number of epochs were also optimized using the development set.

The best results using HeLI 2.0 with LM adaptation on each development set can be seen in Tables 2, 3, and 4. This was the best performing method with the GDI development set but behind NB with the traditional Chinese and even behind simple scoring with the simplified Chinese.

4.5 Naive Bayes with adaptive language models

As our NB implementation seemed to outperform the HeLI 2.0 method in some experiments, we implemented a method using it together with the same LM adaptation scheme we used with HeLI 2.0 in the previous section.

The best results using the Naive Bayes with LM adaptation on each development set can be seen in Tables 2, 3, and 4.

5 Results and discussion

The participants were allowed to submit three separate runs to each of the two tracks of the DMT shared task, as well as to the one track of the GDI shared task. For each track, we submitted results using the HeLI 2.0 with LM adaptation, the Naive

[4] We did not use the confidence threshold with HeLI 2.0 using LM adaptation for the simplified Chinese script.

[5] The confidence threshold was used with NB and LM adaptation for the traditional Chinese script, but it got optimized to zero.

Bayes, and the NB with LM adaptation thus using all submissions available to us. The parameters we used with each method were the same as with the respective development sets.

A total of seven teams provided language identification results for the DMT shared task and six teams for the GDI shared task. Tables 5, 6, and 7 show the macro F1-scores of our submitted runs on the test set. Additionally, the tables show the methods and features used by the other teams together with their F1 scores. We submitted the runs using a team name "SUKI", which is the same we have used in the previous years. The results of the other participating teams were collected from the results packages provided by the organizers after the competition. The system description papers of the other teams were not available at the time of writing. The identity of other participants was also unknown. We were, however, provided with a short description of each system[6] which we used to provide these results.

The simplified Chinese track was won by a team called "hezhou" by a clear margin to the second and the third submissions. According to their system description, the "hezhou" team used a variety of features learned from outside sources, such as a pretrained BERT model for Chinese and word-embeddings trained on People's Daily News. Their results are interesting, but they cannot be directly compared with the ones provided by other teams as the track was supposed to be a closed one.

If we discount the results of the "hezhou" team, the two top places in all three tracks of the DMT and GDI shared tasks were divided between our "SUKI" team and the team "tearsofjoy". "tearsofjoy" used a two stage SVM approach in all of their top runs. After the first stage, the most confidently identified sentences were added to the training data, this step thus functioning as LM adaptation scheme similar to ours. They also submitted results using an SVM ensemble without adaptation and their respective score differences are similar to our systems with and without LM adaptation.

Interestingly, the HeLI 2.0 with adaptation is better than naive Bayes with adaptation on GDI and vice-versa in the DMT. In DMT, the optimal character n-gram range for NB was up to 15 characters, which spans several words. In the experiments with the simplified Chinese development data, even the simple scoring method performed better than the HeLI 2.0 method with LM adaptation. The optimal maximum length of character n-grams was 15 characters also when using the simple scoring method. In the Chinese data, 15 characters span on average five words. From these results we could surmise, that the poor performance of the HeLI method in the DMT shared task was at least partly due to the lack of capturing features spanning several words.

There is a notable inconsistency in our test results with the simplified Chinese. The HeLI 2.0 with LM adaptation performs almost as well as the NB with LM adaptation. This might be due to the fact that we had forgotten to use the confidence threshold with the simplified Chinese development set for the HeLI method and therefore we did not use one with the test set either. It could very well be that the use of a confidence threshold was disadvantageous with both of the Chinese test sets. Our winning submission for the traditional Chinese track used the NB with LM adaptation and with a confidence threshold of 0.[7]

The fact that we did not come in the first place in the GDI 2019 shared task is undoubtedly partly due to the fact that we did not use the provided iVector-files at all in the classification task unlike the other top three teams. Using the information from the iVectors together with our language identifier implementations would not have been trivial.

At the time of writing this article, the participants do not have access to the correct language labels of the test sets, which hinders a detailed error analysis.

6 Conclusions and future work

The two varieties of Chinese used in the DMT shared task seem to be distinguishable from each other quite well. Whether it is due to more structural or more functional differences is left to be determined by experts in Chinese.

We were happy to see that some of the other teams had taken notice of the success of our LM adaptation scheme at the ILI and the GDI 2018 shared tasks. In the GDI 2019 shared task, the use of some sort of an LM adaptation procedure was of paramount importance; the macro F1 scores rose

[6] As part of the submissions, the participants were asked to provide a short description.txt file.

[7] Using a confidence threshold of 0 was the result of optimization with the development set.

Team/run	Method	Features	F1 dev	F1 test
hezhou	Ensemble of BERT, LSTM, SVM, ...	word-embeddings, word *n*-grams, ...		0.8929
tearsofjoy, run 1	SVM with LM adaptation	ch. *n*-grams 1–4, word *n*-grams 1–2		0.8738
SUKI, run 1	**Naive Bayes with LM adaptation**	**ch. *n*-grams 1–15**	**0.9225**	**0.8735**
SUKI, run 3	**HeLI 2.0 with LM adaptation**	**ch. *n*-grams 1–2, words**	**0.8909**	**0.8710**
SUKI, run 2	**Naive Bayes**	**ch. *n*-grams 1–15**	**0.9215**	**0.8685**
itsalexyang	Ensemble of Naive Bayes and BiLSTM	ch. *n*-grams 2–3, word embeddings		0.8530
tearsofjoy, run 2	SVM ensemble	ch. *n*-grams 2–5, words		0.8445
Adaptcenter	Ensemble of CNN and ?	? and words		0.8124
ghpaetzold	RNN	characters		0.7934
gretelliz92	NN with 4 dense layers	TF-IDF vectors		0.7496

Table 5: Simplified Chinese. The macro F1-scores attained by submitted methods on the test set. The results of our submissions are bolded.

Team/run	Method	Features	F1 dev	F1 test
SUKI, run 1	**Naive Bayes with LM adaptation**	**ch. *n*-grams 1–14**	**0.9295**	**0.9085**
hezhou	Ensemble of BERT, LSTM, SVM, ...	word-embeddings, word *n*-grams, ...		0.9009
tearsofjoy, run 1	SVM with LM adaptation	ch. *n*-grams 1–4, words		0.8844
SUKI, run 2	**Naive Bayes**	**ch. *n*-grams 1–14**	**0.9285**	**0.8815**
SUKI, run 3	**HeLI 2.0 with LM adaptation**	**ch. *n*-grams 1–2, words**	**0.9160**	**0.8712**
itsalexyang	Ensemble of Naive Bayes and BiLSTM	ch. *n*-grams 2–3, word embeddings		0.8687
tearsofjoy, run 2	SVM ensemble	ch. *n*-grams 2–5, words		0.8643
Adaptcenter	Ensemble of CNN and ?	? and words		0.8317
ghpaetzold	RNN	characters		0.7959
gretelliz92	NN with 4 dense layers	TF-IDF vectors		0.7484

Table 6: Traditional Chinese. The macro F1-scores attained by different methods on the test set. The results of our submissions are bolded.

Team/run	Method	Features	F1 dev	F1 test
tearsofjoy, run 2	SVM with LM adapt.	ch. *n*-grams 1–5, word *n*-grams 1–2, iVect.		0.7593
SUKI, run 1	**HeLI 2.0 with LM adapt.**	**ch. 4-grams**	**0.8657**	**0.7541**
benf	SVM ens. with LM adapt.	various ch. and word level TF-IDF, iVect.		0.7455
SUKI, run 3	**Naive Bayes with LM adapt.**	**ch. *n*-grams 2–6**	**0.8442**	**0.7451**
tearsofjoy, run 3	SVM ens.	ch. *n*-grams 2–5, words, iVect.		0.6517
SUKI, run 2	**Naive Bayes**	**ch. *n*-grams 2–6**	**0.6475**	**0.6460**
BAM	Ens. of CNN, LSTM, and KRR	?		0.6255
dkosmajac	Ens. of QDA and RF	textual + iVect.		0.5616
ghpaetzold	RNN	characters		0.5575

Table 7: GDI 2019. The macro F1-scores attained by different methods on the test set. The results of our submissions are bolded.

to a completely different level when this was used. The use of LM adaptation did not have such a high importance in the DMT shared task as it did with the GDI 2019, but it still always improved the results.

If we discount the "hezhou" submission, the results of both of the shared tasks once more indicate that deep neural networks do not reach the same accuracy in language identification as SVM, NB, or the HeLI methods.

We think that the poor results of the HeLI 2.0 method on the Chinese data were partly due to the shortness of the words and the importance of information spanning word boundaries. We would like to experiment with giving the HeLI method access to a larger context to verify that this indeed is the case. We will also seek to find a way to incorporate external information, such as that provided by the iVector-files in GDI 2019, to the task of language identification of text.

Acknowledgments

This work has been partly funded by the Kone Foundation and FIN-CLARIN. We also thank the anonymous reviewer of an earlier paper for the idea to use character *n*-grams of infinite size in the formulaic description of the HeLI 2.0 method instead of words.

References

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)*, pages 172–177, Santa Fe, USA.

Fernando Benites, Ralf Grubenmann, Pius von Däniken, Dirk von Grünigen, Jan Deriu, and Mark Cieliebak. 2018. Twist Bytes German dialect identification with data mining optimization. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 218–227, Santa Fe, USA.

Ralf D. Brown. 2012. Finding and Identifying Text in 900+ Languages. *Digital Investigation*, 9:S34–S43.

Çağr Çöltekin and Taraka Rama. 2017. Tübingen system in vardial 2017 shared task: experiments with language identification and cross-lingual parsing. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 146–155, Valencia, Spain.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. SINICA CORPUS : Design Methodology for Balanced Corpora. In *Language, Information and Computation : Selected Papers from the 11th Pacific Asia Conference on Language, Information and Computation : 20-22 December 1996, Seoul*, pages 167–176, Seoul, Korea. Kyung Hee University.

Yingna Chen and Jia Liu. 2005. Language Model Adaptation and Confidence Measure for Robust Language Identification. In *Proceedings of International Symposium on Communications and Information Technologies 2005 (ISCIT 2005)*, volume 1, pages 270–273, Beijing, China.

Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao, and Kuang-Yu Chen. 2000. Sinica treebank: Design criteria, annotation guidelines, and on-line interface. In *Proceedings of the Second Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 12*, CLPW '00, pages 29–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chu-Ren Huang and Lung-Hao Lee. 2008. Contrastive Approach towards Text Source Classification based on Top-Bag-of-Word Similarity. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 404–410, Cebu City, Philippines.

Chu-Ren Huang, Jingxia Lin, Menghan Jiang, and Hongzhi Xu. 2014. Corpus-based study and identification of mandarin chinese light verb variations. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 1–10, Dublin, Ireland.

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019a. Language and Dialect Identification of Cuneiform Texts. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Minneapolis, Minnesota, USA.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018a. HeLI-based Experiments in Discriminating Between Dutch and Flemish Subtitles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 137–144, Santa Fe, NM.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018b. HeLI-based Experiments in Swiss German Dialect Identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 254–262, Santa Fe, NM.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018c. Iterative Language Model Adaptation for Indo-Aryan Language Identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 66–75, Santa Fe, NM.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2016. HeLI, a Word-Based Backoff Method for Language Identification. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 153–162, Osaka, Japan.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019b. Language Model Adaptation for Language and Dialect Identification of Text. *arXiv preprint*, arXiv:1903.10915.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018d. Automatic Language Identification in Texts: A Survey. *arXiv preprint arXiv:1804.08186*.

Shanjian Li and Katsuhiko Momoi. 2001. A Composite Approach to Language/Encoding Detection. In *Nineteenth International Unicode Conference (IUC19)*, San Jose, California, USA.

Shervin Malmasi and Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 164–169, Valencia, Spain.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jrg 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*, pages 1–14, Osaka, Japan.

A. M. McEnery and R. Z. Xiao. 2003. The lancaster corpus of mandarin chinese.

Maria Medvedeva, Martin Kroon, and Barbara Plank. 2017. When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 156–163, Valencia, Spain.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob–A corpus of spoken Swiss German. In *Proceedings of the Language Resources and Evaluation (LREC)*, pages 4061–4066, Portoroz, Slovenia).

Fan Xu, Mingwen Wang, and Maoxi Li. 2016. Sentence-level Dialects Identification in the Greater China Region. *International Journal on Natural Language Computing (IJNLC)*, 5(6).

Fan Xu, Mingwen Wang, and Maoxi Li. 2018. Building Parallel Monolingual Gan Chinese Dialect Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 244–249, Miyazaki, Japan. European Language Resources Association (ELRA).

Marcos Zampieri, Shervin Malmasi, Nikola Ljubeic, Preslav Nakov, Ahmed Ali, Jrg Tiedemann, Yves Scherrer, and Nomi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–15, Valencia, Spain.

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.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei Butnaru, and Tommi Jauhiainen. 2019. A Report on the Third VarDial Evaluation Campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A Report on the DSL Shared Task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland.

Marcos Zampieri, Liling Tan, Nikola Ljubeić, Jrg 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 (LT4VarDial)*, pages 1–9, Hissar, Bulgaria.

Shan Zhong, Yingna Chen, Chunyi Zhu, and Jia Liu. 2007. Confidence measure based incremental adaptation for online language identification. In *Proceedings of International Conference on Human-Computer Interaction (HCI 2007)*, pages 535–543, Beijing, China.