Experiments in Cuneiform Language Identification

Gustavo Henrique Paetzold[1], Marcos Zampieri[2]
[1]Universidade Tecnológica Federal do Paraná, Toledo-PR, Brazil
[2]University of Wolverhampton, Wolverhampton, United Kingdom
ghpaetzold@utfpr.edu.br

Abstract

This paper presents methods to discriminate between languages and dialects written in Cuneiform script, one of the first writing systems in the world. We report the results obtained by the *PZ* team in the Cuneiform Language Identification (CLI) shared task organized within the scope of the VarDial Evaluation Campaign 2019. The task included two languages, Sumerian and Akkadian. The latter is divided into six dialects: Old Babylonian, Middle Babylonian peripheral, Standard Babylonian, Neo Babylonian, Late Babylonian, and Neo Assyrian. We approach the task using a meta-classifier trained on various SVM models and we show the effectiveness of the system for this task. Our submission achieved 0.738 F1 score in discriminating between the seven languages and dialects and it was ranked fourth in the competition among eight teams.

1 Introduction

As discussed in a recent survey (Jauhiainen et al., 2018), discriminating between similar languages, national language varieties, and dialects is an important challenge faced by state-of-the-art language identification systems. The topic has attracted more and more attention from the CL/NLP community in recent years with publications on similar languages of the Iberian peninsula (Zubiaga et al., 2016), and varieties and dialects of several languages such as Greek (Sababa and Stassopoulou, 2018) and Romanian (Ciobanu and Dinu, 2016) to name a few.

As evidenced in Section 2, the focus of most of these studies is the identification of languages and dialects using contemporary data. A few exceptions include the work by Trieschnigg et al. (2012) who applied language identification methods to historical varieties of Dutch and the work by Jauhiainen et al. (2019) on languages written in cuneiform script: Sumerian and Akkadian. Cuneiform is an ancient writing system invented by the Sumerians for more than three millennia.

In this paper we describe computational approaches to language identification on texts written in cuneiform script. For this purpose we use the dataset made available by Jauhiainen et al. (2019) to participants of the Cuneiform Language Identification (CLI) shared task organized at VarDial 2019 (Zampieri et al., 2019). Our submission, under the team name *PZ*, is an adaptation of an n-gram-based meta-classifier system which showed very good performance in previous language identification shared tasks (Malmasi and Zampieri, 2017b,a). Furthermore, we compare the performance of the meta-classifier to the submissions to the CLI shared task and, in particular, to a deep learning approach submitted by the team *ghpaetzold*. It has been shown in previous language identification studies (Medvedeva et al., 2017; Kroon et al., 2018) that deep learning approaches do not outperform n-gram-based methods and we were interested in investigating whether this is also true for the languages and dialects included in CLI.

2 Related Work

Since its first edition in 2014, shared tasks on similar language and dialect identification have been organized together with the VarDial workshop co-located with international conferences such as COLING, EACL, and NAACL. The first and most well-attended of these competitions was the Discrminating between Similar Languages (DSL) shared task which has been organized between 2014 and 2017 (Malmasi et al., 2016b; Zampieri et al., 2014, 2015, 2017). The DSL provided the first benchmark for evaluation of language identification systems developed for similar languages and language varieties using the DSL Corpus Col-

Proceedings of VarDial, pages 209–213
Minneapolis, MN, June 7, 2019 ©2019 Association for Computational Linguistics

Language or Dialect	Code	Texts	Lines	Signs
Late Babylonian	LTB	671	31,893	ca. 260,000
Middle Babylonian peripheral	MPB	365	11,015	ca. 95,000
Neo-Assyrian	NE	3,570	65,932	ca. 490,000
Neo-Babylonian	NEB	1,212	19,414	ca. 200,000
Old Babylonian	OLB	527	7,605	ca. 65,000
Standard Babylonian	STB	1,661	35,633	ca. 390,000
Sumerian	SUX	5,000	107,345	ca. 400,000
Total		13,006	278,837	ca. 1,900,000

Table 1: Number of texts, lines, and signs in each of the seven languages and dialects in the dataset of Jauhiainen et al. (2019), from which the instances of the CLI datasets were taken.

lection (DSLCC) (Tan et al., 2014), a multilingual benchmarked dataset compiled for this purpose. In 2017 and 2018, VarDial featured evaluation campaigns with multiple shared tasks not only on language and dialect identification but also on other NLP tasks related to language and dialect variation (e.g. morphosyntactic tagging, and cross-lingual dependency parsing). With the exception of the DSL, the language and dialect identification competitions organized at VarDial focused on groups of dialects from the same language such as Arabic (ADI shared task) and German (GDI shared task).

The focus of the aforementioned language and dialect identification competitions was diatopic variation and thus the data made available in these competitions was synchronic contemporary corpora. In the 2019 edition of the workshop, for the first time, a task including historical languages was organized. The CLI shared task provided participants with a dataset containing languages and dialects written in cuneiform script: Sumerian and Akkadian. Akkadian is divided into six dialects in the dataset: Old Babylonian, Middle Babylonian peripheral, Standard Babylonian, Neo Babylonian, Late Babylonian, and Neo Assyrian (Jauhiainen et al., 2019).

The CLI shared task is an innovative initiative that opens new perspectives in the computational processing of languages written in cuneiform script. There have been a number of studies applying computational methods to process these languages (e.g. Sumerian (Chiarcos et al., 2018)), but with the exception of Jauhiainen et al. (2019), to the best of our knowledge, no language identification studies have been published. CLI is the first competition organized on cuneiform script texts in particular and in historical language identification in general.

3 Methodology and Data

The dataset used in the CLI shared task is described in detail in Jauhiainen et al. (2019). All of the data included in the dataset was collected from the Open Richly Annotated Cuneiform Corpus (Oracc)[1] which contains transliterated texts. Jauhiainen et al. (2019) created a tool to transform the texts back to the cuneiform script. The dataset features texts from seven languages and dialects amounting to a little over 13,000 texts. The list of languages and dialects is presented in Table 1.

3.1 System Description

Our submission to the CLI shared task is a system based on a meta-classifier trained on several SVM models. Meta-classifiers (Giraud-Carrier et al., 2004) and ensemble learning methods have proved to deliver competitive performance not only in language identification (Malmasi and Zampieri, 2017b,a) but also in many other text classification tasks (Malmasi et al., 2016a; Sulea et al., 2017).

The meta-classifier is an adaptation of previous submissions to VarDial shared tasks described in (Malmasi and Zampieri, 2017a). It is essentially a bagging ensemble trained on the outputs of linear SVM classifiers. As features, the system uses the following character n-gram and character skip-gram features:

- character n-grams of order 1–5;
- 1-skip character bigrams and trigrams;
- 2-skip character bigrams and trigrams;
- 3-skip character bigrams and trigrams.

Each feature class is used to train a single linear SVM classifier using LIBLINEAR (Fan et al.,

[1] http://oracc.museum.upenn.edu/

2008). The outputs of these SVM classifiers on the training data are then used to train the meta-classifier.

4 Results

Table 2 showcases the results obtained by our team (*PZ* in bold) and the best submission by each of the eight teams which participating in the CLI shared task. Even though the competition allowed the use of other datasets (open submission), we have used only the dataset provided by the shared task organizers to train our model.

Our submission was ranked 4th in the shared task, only a few percentage points below the top-3 systems: *NRC-CNRC, tearsofjoy*, and *Twist Bytes*. The meta-classifier achieved much higher performance at distinguishing between these Mesopotamian languages and dialects than the neural model by *ghpaetzold*, which ranked 6th in the competition. We present this neural model in more detail comparing its performance to our meta-classifier in Section 4.1.

System	F1 (macro)
NRC-CNRC	0.769
tearsofjoy	0.763
Twist Bytes	0.743
PZ	**0.738**
ghmerti	0.721
ghpaetzold	0.556
ekh	0.550
situx	0.128

Table 2: Results for the CLI task obtained by the team *PZ* (in bold) in comparison to the the best entries of each of the eight teams in the shared task. Results reported in terms of F1 (macro).

4.1 Comparison to a Neural Model

We take the opportunity to compare the performance of our system with an entirely different type of model submitted by team *ghpaetzold*. This comparison was motivated by the lower performance obtained by the neural models in comparison to traditional machine learning models in previous VarDial shared tasks (Zampieri et al., 2018). It was made possible due to the collaboration between the *ghpaetzold* team and ours.[2]

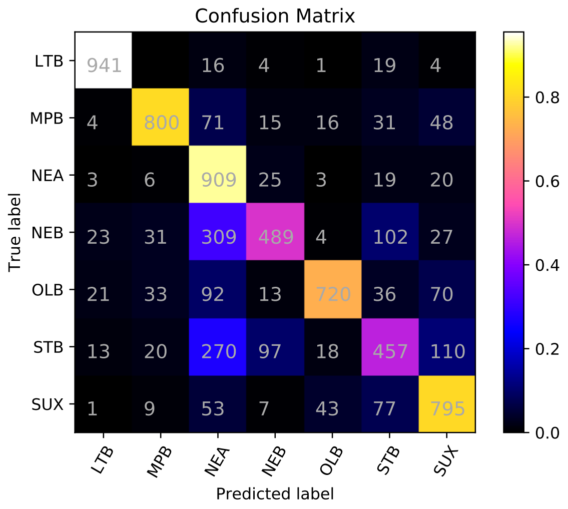
As demonstrated by Ling et al. (2015), compositional recurrent neural networks can offer very reliable performance on a variety of NLP tasks. Previous language identification and dialect studies (Medvedeva et al., 2017; Kroon et al., 2018; Butnaru and Ionescu, 2019) and the results of the previous shared tasks organized at VarDial (Zampieri et al., 2017, 2018), however, showed that deep learning approaches do not outperform more linear n-gram-based methods so we were interested in comparing the performance of a neural model to the meta-classifier for this dataset.

A compositional network is commonly described as a model that builds numerical representations of words based on the sequence of characters that compose them. They are inherently more time-consuming to train than typical neural models that use traditional word vectors because of the added parameters, but they compensate by being able to handle any conceivable word passed as input with very impressive robustness (Paetzold, 2018, 2019).

The model takes as input a sentence and produces a corresponding label as output. First, the model vectorizes each character of each word in the sentence using a typical character embedding layer. It then passes the sequence of vectors through a set of 2 layers of Gated Recurrent Units (GRUs) and produces a numerical representation for each word as a whole. This set of representations is then passed through another 2-layer set of GRUs to produce a final vector for the sentence as a whole, and then a dense layer is used to produce a softmax distribution over the label set. The model uses 25 dimensions for character embeddings, 30 nodes for each GRU layer and 50% dropout. A version of each model was saved after each epoch so that the team could choose the one with the lowest error on the development set as their submission.

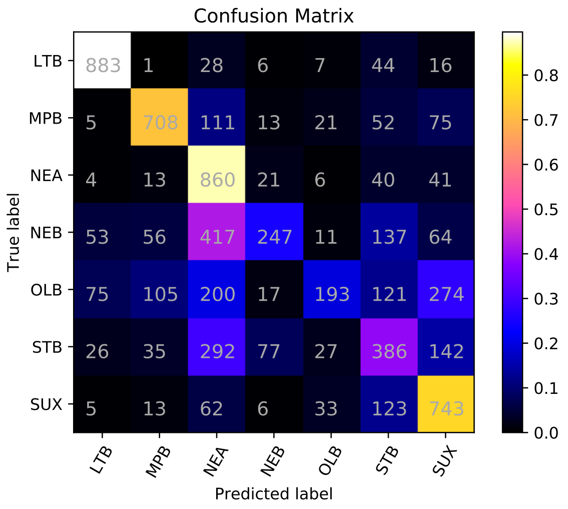
Inspecting the two confusion matrices depicted in Figures 1 and 2, we found that the neural model did not do very well at differentiating between Standard Babylonian and Neo Assyrian, as well as between Neo Babylonian and Neo Assyrian, leading to many misclassifications. These two language pairs were also the most challenging for the meta-classifier, however, the number of missclassified instances by the meta-classifier was much lower.

[2] One of the *ghpaetzold* team members was also a member of the *PZ* team.



	LTB	MPB	NEA	NEB	OLB	STB	SUX
LTB	941		16	4	1	19	4
MPB	4	800	71	15	16	31	48
NEA	3	6	909	25	3	19	20
NEB	23	31	309	489	4	102	27
OLB	21	33	92	13	720	36	70
STB	13	20	270	97	18	457	110
SUX	1	9	53	7	43	77	795

Figure 1: Confusion matrix for the meta-classifier.



	LTB	MPB	NEA	NEB	OLB	STB	SUX
LTB	883	1	28	6	7	44	16
MPB	5	708	111	13	21	52	75
NEA	4	13	860	21	6	40	41
NEB	53	56	417	247	11	137	64
OLB	75	105	200	17	193	121	274
STB	26	35	292	77	27	386	142
SUX	5	13	62	6	33	123	743

Figure 2: Confusion matrix for the neural model.

5 Conclusion and Future Work

In this paper we presented a meta-classifier system submitted by the team *PZ* to the Cuneiform Language Identification shared task organized at VarDial 2019. Our submission is an adaptation of a sophisticated meta-classifier which achieved high performance in previous language and dialect identification shared tasks at VarDial (Malmasi and Zampieri, 2017a). The meta-classifier combines the output of multiple SVM classifers trained on character-based features. The meta-classifier ranked 4th in the competition among eight teams only a few percentage points below the top-3 systems in the competition.

Finally, we compared the performance of the meta-classifier with a compositional RNN model that uses only the text from the instance as input trained on the same dataset. The comparison shows that, while the neural model does offer competitive performance against some of the systems submitted to the shared task, the more elaborate features used by the meta-classifier allows it to much more proficiently distinguish between very similar language pairs, such as Neo Babylonian and Neo Assyrian, leading to a performance gain of 18.2% F-score and 2 positions in the shared task rankings. The results obtained by the meta-classifier in comparison to the neural model corroborate the findings of previous studies (Medvedeva et al., 2017) in the last two VarDial evaluation campaigns (Zampieri et al., 2017, 2018).

In the future we would like to analyze the results obtained by the highest performing teams in the CLI shared task. The top team achieved the best performance in the competition using a neural-based method. This is, to the best of our knowledge, the first time in which a deep learning approach outperforms traditional machine learning methods in one of the VarDial shared tasks. The great performance obtained by the NRC-CNRC team might be explained by the use of more suitable deep learning methods such as BERT (Devlin et al., 2018).

Acknowledgements

We would like to thank Shervin Malmasi for his valuable suggestions and feedback. We further thank the CLI shared task organizer, Tommi Jauhiainen, for organizing this interesting shared task.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

Andrei Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus. *arXiv preprint arXiv:1901.06543*.

Christian Chiarcos, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling. 2018. Towards a Linked Open Data Edition of Sumerian Corpora. In *Proceedings of LREC*.

Alina Maria Ciobanu and Liviu P Dinu. 2016. A computational perspective on the romanian dialects. In *Proceedings of LREC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874.

Christophe Giraud-Carrier, Ricardo Vilalta, and Pavel Brazdil. 2004. Introduction to the Special Issue on Meta-learning. *Machine learning*, 54(3):187–193.

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019. Language and Dialect Identification of Cuneiform Texts. In *Proceedings of VarDial*.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic language identification in texts: A survey. *arXiv preprint arXiv:1804.08186*.

Martin Kroon, Masha Medvedeva, and Barbara Plank. 2018. When Simple N-gram Models Outperform Syntactic Approaches: Discriminating between Dutch and Flemish. In *Proceedings of VarDial*.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.

Shervin Malmasi and Marcos Zampieri. 2017a. Arabic Dialect Identification Using iVectors and ASR Transcripts. In *Proceedings of VarDial*.

Shervin Malmasi and Marcos Zampieri. 2017b. German Dialect Identification in Interview Transcriptions. In *Proceedings of VarDial*.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016a. Predicting Post Severity in Mental Health Forums. In *Proceedings of CLPsych*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016b. Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In *Proceedings of VarDial*.

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 VarDial*.

Gustavo Paetzold. 2018. UTFPR at IEST 2018: Exploring Character-to-Word Composition for Emotion Analysis. In *Proceedings of WASSA*.

Gustavo Paetzold. 2019. UTFPR at SemEval-2019 Task 6: Relying on Compositionality to Find Offense. In *Proceedings of SemEval*.

Hanna Sababa and Athena Stassopoulou. 2018. A Classifier to Distinguish Between Cypriot Greek and Standard Modern Greek. In *Proceedings of SNAMS*.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017. Exploring the use of Text Classification in the Legal Domain. *Proceedings of ASAIL*.

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 BUCC*.

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska Jong, and Theo Meder. 2012. An Exploration of Language Identification Techniques in the Dutch Folktale Database. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage*.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of VarDial*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shuon, 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 VarDial*.

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 VarDial*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of VarDial*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the dsl shared task 2015. In *Proceedings of LT4VarDial*.

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2016. Tweetlid: A Benchmark for Tweet Language Identification. *Language Resources and Evaluation*, 50(4):729–766.