Using Artificial Neural Networks in Dialect Identification in Less-resourced Languages
The Case of Kurdish Dialects Identification

Hossein Hassani and Oussama H. Hamid
Department of Computer Science and Engineering, University of Kurdistan Hewlêr, Erbil, Kurdistan Region, Iraq

Keywords: Dialect Classification, Natural Language Processing, Artificial Neural Networks, Machine Learning, Kurdish Dialects.

Abstract: Dialect identification/classification is an important step in many language processing activities particularly with regard to multi-dialect languages. Kurdish is a multi-dialect language which is spoken by a large population in different countries. Some of the Kurdish dialects, for example, Kurmanji and Sorani, have significant grammatical differences and are also mutually unintelligible. In addition, Kurdish is considered a less-resourced language. The classification techniques based on machine learning approaches usually require a considerable amount of data. In this research, we are interested in using approaches based on Artificial Neural Network (ANN) in order to be able to identify the dialects of Kurdish texts without the need to have a large amount of data. We will also compare the outcomes of this approach with the previous work on Kurdish dialect identification to compare the performance of these methods. The results showed that the two approaches do not show a significant difference in their accuracy and performance with regard to long documents. However, they showed that the ANN approach performs better than traditional approach for the single sentence classification. The accuracy rate of the ANN sentence classifier was 99% for Kurmanji and 96% for Sorani.

1 INTRODUCTION

Kurdish is an Indo-European multi-dialect language (Hassanpour, 1992). It is mainly spoken in areas touching Iran, Iraq, Turkey, and Syria and also by Kurdish communities in other countries such as Lebanon, Georgia, Armenia, Afghanistan and Kurdish diaspora in Europe and North America (Hassani and Medjedovic, 2016; Foundation Institute Kurde de Paris, 2017a). The population that speaks the language is estimated to be between 36 and 47 million[1].

The most common categorization of Kurdish dialects includes Northern Kurdish (Kurmanji), Central Kurdish (Sorani), Southern Kurdish, Gorani and Zazaki (Haig and Öpengin, 2014; Hassani and Medjedovic, 2016; Malmasi, 2016). Kurdish is written using four different scripts, which are modified Persian/Arabic, Latin, Yekgirtû(unified), and Cyrillic (Hassani and Medjedovic, 2016) The usage of the scripts and their popularity differ depending on the dominance of Persian, Arabic, Turkish, and Cyrillic in the specific regions (Hassani, 2017b).

The dialect diversity of Kurdish implies that automatic dialect identification is an essential task in Kurdish Natural Language Processing (NLP) (Hassani and Medjedovic, 2016; Hassani, 2017a). Language identification is a fundamental task in NLP though a straightforward one (Zaidan and Callison-Burch, 2014). Although dialect identification could be viewed as language identification, the subtle differences that distinguishes one dialect from the other leads to a more complex NLP and computational task (Zaidan and Callison-Burch, 2014). The task of Dialect Identification (DID) is a special case of the more general problem of Language Identification (LID) (Ali et al., 2015; Hassani and Medjedovic, 2016).

Inspired by the models that depict the way that the human brain processes the cognition, Artificial Neural Network (ANN) has been suggested to be used in solving a wide range of problems such as pattern recognition and classification (Jain et al., 1996; Krogh, 2008). scholars (Glüge et al., 2010; Rizwan

[1]The details that appear in (Foundation Institute Kurde de Paris, 2017b) do not show the population of Kurdish diaspora in North America. However, they elsewhere, the same website has estimate this to be about 26,000 in (Foundation Institute Kurde de Paris, 2017a).

Hassani H. and H. Hamid O.
Using Artificial Neural Networks in Dialect Identification in Less-resourced Languages - The Case of Kurdish Dialects Identification.
DOI: 10.5220/0006578004430448

et al., 2016; Sunija et al., 2016; Soorajkumar et al., 2017; Sinha et al., 2017). ANN has been also used in text classification (Ghiassi et al., 2012; Lai et al., 2015; Belinkov and Glass, 2016).

We are interested in investigating the performance and accuracy of ANN in the identification of Kurdish dialects in textual formats. An efficient dialect identifier is necessary in Kurdish NLP and CL. Although different approaches have been taken by researchers to address the problem of text classification, we prefer to use a simple approach before embarking into a more complex method. For this we use the perceptron model, which was introduced in the 1960s. We also use a traditional classifier based on Support Vector Machines (SVM) to compare the performance of this method with the previous one. Importantly, we also evaluate our models at the sentence level. That is, we assess the accuracy of the models when they are applied to sentences rather than long documents.

The rest of this article is organized as follows. Section 2 discusses the related work. Section 3 provides the methodology and how the experiments are conducted. Section 4 summarizes the findings and gives the conclusion.

2 RELATED WORK

The Parallel Convolutional Neural Network was suggested (Johnson and Zhang, 2015) for text categorization. It was proposed as an alternative mechanism for effective use of word order by the usage of direct embedding of small text regions. The approach is different from the bag-of-ngram or word-vector Convolutional Neural Network (CNN). Parallel CNN framework allows the learning of several types of embedding which can be combined together. This combination is able to let the parts to complement each and to provide a higher accuracy. According to researchers who suggested this approach, they have been able to achieve a state-of-the-art performance on sentiment classification and topic classification (Johnson and Zhang, 2015).

A Dynamic Artificial Neural Network (DAN2) algorithm was proposed as an alternative approach for text classification (Ghiassi et al., 2012). Like the classical neural networks, DAN2 is also composed of an input layer, several hidden layers and an output layer (Ghiassi et al., 2012). However, unlike classical neural networks, there is no preset number for the hidden layers (Ghiassi et al., 2012). The experiments with DAN2 showed that it outperforms the classical approaches of Machine Learning (ML) which are used for classification such as Key Nearest Neighbors (KNN) and Supervised Vector Machines (SVM) (Ghiassi et al., 2012).

A Recursive Neural Network (RecursiveNN or RNN) (Socher et al., 2011; Socher et al., 2013) was suggested to be used in parsing natural language processing. The experiments with its application showed that it outperforms the state-of-the-art approaches in segmentation, annotation and scene classification (Socher et al., 2011).

A variant of RecursiveNN, Recursive Convolutional Neural Network (RCNN) (Zhu et al., 2015) is used in different tasks of NLP in order through modeling the relations between a dependency tree and distributed representations of a sentence or phrase (Zhu et al., 2015).

Recurrent Convolutional Neural Networks (RecurrentCNN) (Lai et al., 2015) was introduced for text classification. It uses a recurrent structure to capture the contextual information. It then uses a CNN to constructs the representation of text. The results of applying this approach showed that it outperforms CNN and Recursive Neural Networks (RecursiveNN) (Lai et al., 2015).

Character-level Convolutional Networks (ConvNets) (Zhang et al., 2015) was applied to text classification tasks, particularly for discriminating between similar languages and dialects. The experiments with Arabic dialect identification showed promising results (Zhang et al., 2015; Belinkov and Glass, 2016).

The depth and diversity of literature in the context of ANN application with regard to NLP tasks, in addition to the current and ongoing research on this area suggests that to explore the performance of the related algorithms in Kurdish NLP potentially can contribute to ANN applications. However, in this stage we start with basic forms of ANN to investigate how they might work with the dialect diversity and resource paucity in Kurdish.

3 METHODOLOGY

We apply two approaches to identify the dialect of a Kurdish text. The first approach is based on ANN and the second is based on traditional classifiers. For the first approach, we use a Perceptron and for the second we use SVMs. Both methods are explained in the following sections.

3.1 Perceptron

An ANN is a suggested architecture based on the way the human brain works, that is, a network of model

neurons in computer which are able to imitate the process of natural neurons whereby they can be trained to solve different kinds of problems (Krogh, 2008). An ANN includes an input layer, several hidden layers and an output layer.

In a text classifier based on ANN, the input units consists of terms/words, the hidden layers are the computational units and the output layer represents class of the inputs (Sebastiani, 2002). A weight is assigned to each term that act is a parameter in computation (hidden) layers. To classify a text, the words/terms weights are given to the network and the sum of the weights is computed, which leads ot the identification of the category/class of the text (Sebastiani, 2002).

Backpropagation is a classic method for training ANNs (Sebastiani, 2002). In this method, a training document (the weight vector) is processed. If the classifier is not able to classify the document properly, an error is raised and "backpropagated" through the network. The network changes the computation parameters in order to correct the decision.

Perceptron is the simplest type of NN classifiers (Sebastiani, 2002) and also it is a kind of Linear Classifier (Gkanogiannis and Kalamboukis, 2009). It begins with an initial model which is refined gradually and iteratively during learning process (Gkanogiannis and Kalamboukis, 2009).

We apply the modified Perceptron learning rule which was suggested for tag recommendations in social bookmarking systems (Gkanogiannis and Kalamboukis, 2009) and adapt it as a dialect identifier. The proposed algorithm is a binary linear classifier and it combines a centroid with a batch Perceptron classifier and a modified perceptron learning rule that does not need any parameter estimation. We use this classifier to detect whether a text is Kurmanji (*Ku*) or Sorani (*So*). In addition, we use a multilayer Perceptron (Pal and Mitra, 1992; Kessler et al., 1997) to identify the text dialect.

For the first case the simple Perceptrons are defined as:

$$ku(\vec{t}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} \vec{W}_i \vec{t}_i + c > 0 \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

$$so(\vec{t}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} \vec{W}_i \vec{t}_i + c > 0 \\ -1 & \text{otherwise} \end{cases} \quad (2)$$

Where:

- $ku(\vec{t})$ determines whether a text has been written in Kurmanji or not;
- $ku(\vec{t})$ determines whether a text has been written in Sorani or not;
- $\vec{W}$ is weights vector;
- c is a constant that is tuned during the training process.

The learning process means the gradual updating of weight vector ($\vec{W}$).

We use a multilayer Perceptron in which "all input units connected to all units of the hidden layer, and all units of the hidden layer connected to all output units" (Kessler et al., 1997) to identify the text dialect. This allows us to add more dialects into our experiments and also makes the model ready to be used in sub-dialect identification in the future.

As we use multilayer Perceptron for multiple class detection purpose, instead of a *sigmoid activation function*, we use a *softmax activation function* for output detection. The *softmax activation function* is defined as below:

$$f(d_i) = \frac{e^{kt_i}}{\sum_{i=1}^{N} e^{kt_j}} \quad (3)$$

where:

- Kt_i shows the inputs to the classifier and
- Kt_j shows the detected dialects.

To minimize the errors in the classifier output a *minimize error function* is used. This function is defined as below:

$$Err = \frac{1}{2} \sum_{i=1}^{nd} \|act - des\|^2 \quad (4)$$

where:

- *act* show the actual outputs;
- *des* show the desired outputs;
- *nd* is the number of dialects.

3.2 Traditional Classifier

We select features for the SVM into two sets of bag-of-words, one set for Kurmanji dialect which we call it *KuTS*, and the other for Sorani dialect which we call *SoTS*. We select 10,000 words from our Kurdish corpora. We select *KuTS* and *SoTS* based on two criteria. The first criteria is the frequency of the words in the related corpus. The second criteria is to have no overlap in the words, which is shown by Equation 5.

$$KuTs \cap SoTs = \varnothing \quad (5)$$

We apply the first condition to restrict the training vectors to the most frequent words. The second condition is applied to investigate whether the common vocabulary plays a role in the efficiency of the classifier.

3.3 Experiment Plan

We use our Kurdish corpora[2] for the experiments. Table 1 gives the general information about this corpora.

Table 1: The number of tokens and word forms inthe Kurdish corpora used in this research.

	Tokens	Word forms
Kurmanji	1,330,443	98,253
Sorani	384,586	67,056

We use 50% of the corpus for training, 10% for development, and the remaining 40% as test data. We have decided to use only 50% of the data for training because as it was mentioned in Section 1, Kurdish is considered a less-resourced language and we are interested in investigating the efficiency of using ANN in the absence of large amount of data.

3.4 Results

Table 2 shows the accuracy of the experiments based on the Perecptron classifier. The table shows the accuracy of the approach for the long texts and single sentences in the test dataset separately.

Table 2: The results of testing the Perceptron classifier.

	Long texts	Sentences
Kurmanji	75%	99%
Sorani	72%	96%

Table 3 shows the accuracy of the experiments based on the traditional classifier. The table shows the accuracy of the approach for the long texts and single sentences in the test dataset separately.

Table 3: The results of testing the traditional classifier.

	Long texts	Sentences
Kurmanji	74.5%	97%
Sorani	71.55%	93%

Table 4 shows the accuracy of the experiments based on the multilayer Perceptron. The table shows the accuracy of the approach for the long texts and single sentences in the test dataset separately.

[2]The corpora consists of variety of texts in Kurmanji and Sorani which is not annotated. It is currently not available for public use.

Table 4: The results of testing the multilayer Perceptron classifier.

	Long texts	Sentences
Kurmanji	88%	58%
Sorani	96%	49%

Table 5 shows the samples of the words and their frequency which was created during the training process.

Table 5: The samples of the words and their frequency from the dataset which was created during the training process. The classes 0 and 1 denote Kurmanji and Sorani dialects, respectively.

Word	Class	Frequency
axa	0	8
axo	1	5
bawerekanî	1	2
bawe	1	1
bawerî	0	13
bawerî	1	10
belł	0	179
belê	1	8
ber	0	2089
ber	1	54
błdenge	1	1
derê	0	69
kar	0	242
kar	1	27
saya	0	33
sayeî	1	3
sêber	1	28
sêberekan	1	49
seravê	0	9
serawê	1	13

In the next section we discuss the presented results.

3.5 Discussion

The results showed that the accuracy of ANN classifier did not present a significant difference against the traditional classifier if the inputs were long texts. However, it showed a considerable difference with regard to sentence classification. On the other hand for Sorani dialect the accuracy in both cases is lower than the rate of classification fo Kurmanji texts/sentence. The reasons for this should be investigated further. However, the preliminary studies suggest that this is the consequence of the smaller dataset that was available for the training process. However, we need to conduct more experiments and to add texts in other Kurdish dialects in order to assess the accuracy of

the model and importantly, the efficiency of the approaches in the absence of a large amount of data.

The multilayer Perceptron showed a different figure. While it performed well for the long texts, it performed quite poor for the sentence classification. This shows that the classifier has not been able to guess the dialect with a high accuracy for short sentences. The preliminary investigations show that this primarily is due to the close relation between Kurmanji and Sorani dialects which makes it difficult to differentiate between the two dialects based on short sentences. However, it requires further study to find out other possible reasons for this outcome.

4 CONCLUSIONS

The article discussed the importance of the task of dialect identification in Kurdish NLP and CL. Through emphasizing the dialect diversity and resource paucity we presented the idea of using ANN to identify the different Kurdish dialects in Kurdish texts. We investigated the efficiency and accuracy of ANN based classifiers in the absence of large amount of texts or corpora. We also compared the outcomes of this approach with the previous work (see (Hassani and Medjedovic, 2016)) on automatic Kurdish dialect identification to compare the accuracy and performance among the two approaches. The results suggested that while the two approaches do not show a significant difference in their accuracy and performance with regard to long documents, the ANN approach performs better than traditional approach for the single sentence classification. However, because we were not able to find any baseline for the sentence classifiers in Kurdish dialect identification studies, we were not able to compare this part of the outcome. Nevertheless, the sentence classifier performed with a high accuracy at 99% for Kurmanji and 96% for Sorani.

The multilayer Perceptron acted differently. It provided quite a poor result for the sentence classification, while showed a reasonable accuracy for the long texts. The early investigations suggest that this behavior could be justified based on the close relation between Kurmanji and Sorani dialects. However, more research is needed to become more certain about this situation and to enhance the classifier to be able to classify short sentences with a higher accuracy.

As for future work, we are interested in expanding the research to cover the texts written in other scripts for example, Persian/Arabic. We are also interested in including other Kurdish dialects such as Hawrami in the classification process. In addition, we believe that the multilayer Perceptron requires further studies, particularly on the error minimization process. We are planning to work on the mentioned areas in an extended paper that follows the current work.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their constructive suggestions and recommendations which have improved the content of the paper.

REFERENCES

Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., Bell, P., and Renals, S. (2015). Automatic Dialect Detection in Arabic Broadcast Speech. *arXiv preprint arXiv:1509.06928*.

Belinkov, Y. and Glass, J. (2016). A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects. *arXiv preprint arXiv:1609.07568*.

Foundation Institute Kurde de Paris (2017a). The Kurdish Diaspora.

Foundation Institute Kurde de Paris (2017b). The Kurdish Population.

Ghiassi, M., Olschimke, M., Moon, B., and Arnaudo, P. (2012). Automated text classification using a dynamic artificial neural network model. *Expert Systems with Applications*, 39(12):10967–10976.

Gkanogiannis, A. and Kalamboukis, T. (2009). A Modified and Fast Perceptron Learning Rule and its Use for Tag Recommendations in Social Bookmarking Systems. *ECML PKDD Discovery Challenge 2009 (DC09)*, page 71.

Glüge, S., Hamid, O. H., and Wendemuth, A. (2010). A Simple Recurrent Network for Implicit Learning of Temporal Sequences. *Cognitive Computation*, 2(4):265–271.

Haig, G. and Öpengin, E. (2014). Introduction to Special Issue-Kurdish: A critical research overview. *Kurdish Studies*, 2(2):99–122.

Hassani, H. (2017a). BLARK for multi-dialect languages: towards the Kurdish BLARK. *Language Resources and Evaluation*, pages 1–20.

Hassani, H. (2017b). Kurdish Interdialect Machine Translation. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 63–72. Association for Computational Linguistics.

Hassani, H. and Medjedovic, D. (2016). Automatic Kurdish Dialects Identification. *Computer Science & Information Technology*, 6(2):61–78.

Hassanpour, A. (1992). *Nationalism and language in Kurdistan, 1918-1985*. Edwin Mellen Pr.

Jain, A. K., Mao, J., and Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, 29(3):31–44.

Johnson, R. and Zhang, T. (2015). Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112. Association for Computational Linguistics.

Kessler, B., Numberg, G., and Schütze, H. (1997). Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics.

Krogh, A. (2008). What are artificial neural networks? *Nature biotechnology*, 26(2):195.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, volume 333, pages 2267–2273. Association for the Advancement of Artificial Intelligence.

Malmasi, S. (2016). Subdialectal Differences in Sorani Kurdish. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 89–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Pal, S. K. and Mitra, S. (1992). Multilayer Perceptron, Fuzzy Sets, and Classification. *IEEE Transactions on neural networks*, 3(5):683–697.

Rizwan, M., Odelowo, B. O., and Anderson, D. V. (2016). Word Based Dialect Classification Using Extreme Learning Machines. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2625–2629. IEEE.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Sinha, S., Jain, A., and Agrawal, S. S. (2017). Empirical analysis of linguistic and paralinguistic information for automatic dialect classification. *Artificial Intelligence Review*, pages 1–26.

Socher, R., Lin, C. C., Manning, C., and Ng, A. Y. (2011). Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 129–136.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive Deep Models For Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Soorajkumar, R., Girish, G., Ramteke, P. B., Joshi, S. S., and Koolagudi, S. G. (2017). Text-Independent Automatic Accent Identification System for Kannada Language. In *Proceedings of the International Conference on Data Engineering and Communication Technology*, pages 411–418. Springer.

Sunija, A., Rajisha, T., and Riyas, K. (2016). Comparative Study of Different Classifiers for Malayalam Dialect Recognition System. *Procedia Technology*, 24:1080–1088.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In *Advances in neural information processing systems*, pages 649–657.

Zhu, C., Qiu, X., Chen, X., and Huang, X. (2015). A Re-ranking Model for Dependency Parser with Recursive Convolutional Neural Network. *arXiv preprint arXiv:1505.05667*.