Improving Multi-label Classification of Similar Languages by Semantics-Aware Word Embeddings

The Quyen Ngo[1], Thi Anh Phuong Nguyen[2], My Linh Ha[1], Thi Minh Huyen Nguyen[1], Phuong Le-Hong[1]*
[1]Vietnam National University, Hanoi, Vietnam
[2]Institute of Information Technology,
Vietnam Academy of Science and Technology, Hanoi, Vietnam
Correspondence: *phuonglh@vnu.edu.vn*

Abstract

The VLP team participated in the DSL-ML shared task of the VarDial 2024 workshop which aims to distinguish texts in similar languages. This paper presents our approach to solving the problem and discusses our experimental and official results. We propose to integrate semantics-aware word embeddings which are learned from ConceptNet into a bidirectional long short-term memory network. This approach achieves good performance – our system is ranked in the top two or three of the best performing teams for the task.

1 Introduction

Discriminating between similar languages (e.g., Croatian and Serbian) and language varieties (e.g., Brazilian and European Portuguese) has been a popular research topic related to the study of diatopic language variation from a computational perspective (Aepli et al., 2023). In the DSL-ML shared tasks of The Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 2024 (Chifu et al., 2024), participating teams are expected to provide multi-label annotations for the instances of datasets from five different macro-languages and with different types of multi-label annotations, including BCMS (Bosnian, Croatian, Montenegrin, Serbian) (Rupnik et al., 2023; Miletić and Miletić, 2024), EN (American and British English), ES (Argentinian and Peninsular Spanish), Portuguese (Brazilian and European Portuguese) (Zampieri et al., 2024), and FR (Belgian, Canadian, French and Swiss French) (Găman et al., 2023; Tan et al., 2014; Bernier-Colborne et al., 2023). Participating systems are evaluated on macro-average F1 for each test set, and aggregated over the five test sets.

This paper presents an approach to improving the performance of our participating system in this shared task. The main idea of our approach is the integration of semantic word embeddings which are learned from the ConceptNet knowledge graph into a recurrent neural network model. The ConceptNet word embeddings are readily available for multiple languages that are concerned with this task.

The paper is structured as follows. Section 2 describes our method. Section 3 presents and discusses empirical results on the development datasets and on the private test set as announced by the shared task organizers. Section 4 concludes the paper and outlines several possible directions for future work.

2 Methods

In this shared task, participants are expected to provide multi-label annotations for the test set instances. There are two tracks. In the closed track, systems may only use the labeled training data provided for the task. The use of pre-trained models is allowed as long as they are not specifically pre-trained or fine-tuned on language identification tasks. In the open track, systems may use any data and pre-trained models, except the prohibited datasets listed in the language description. Our system is essentially in the closed track since we do not use any external training data and the ConceptNet embeddings are not specifically pre-trained or fine-tuned on any language identification task.

We aim to develop a method which does not utilize pre-trained models for this task. Thus, we use bidirectional long short-term memory networks (BiLSTMs) for learning text representation.

2.1 Bidirectional LSTM Model

Let $\mathbf{x}_j$ be the embedding of token w_j and $\mathrm{RNN}_\theta(\mathbf{x})$ be an abstraction of a LSTM that processes the sequence of vectors $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, then output for $\mathbf{x}_j$ is defined as $\vec{v}_j := \mathrm{RNN}^l_\theta(\mathbf{x}_j) \oplus \mathrm{RNN}^r_\theta(\mathbf{x}_j)$. We consider multi-layer BiLSTMs where the output $\vec{v}^k_j$ of the k-th layer is fed as input

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 235–240

to the $(k+1)$-th layer. In our experiments, each token embedding $\mathbf{x}_j$ is either *initialized randomly* or is a *static pre-trained word embedding* provided by ConceptNet Numberbatch as presented in the next subsection.

For decoding, we use a fully-connected feed-forward network which is fed the output of the last BiLSTM. The output is simply computed by a softmax layer as common in multiway classification:

$$P(y_j|\vec{v}_j) = \text{softmax}(W\vec{v}_j + \vec{b}),$$

where W and b are parameter matrices. The overall network architecture that we use is as follows:

$$\texttt{EmbeddingLayer}(w) \rightarrow \texttt{stacked BiLSTM}(h) \rightarrow \texttt{Dense}(d, \texttt{ReLu}) \rightarrow \texttt{Dense(softmax)},$$

where the hyperparameters w, h and d are the word embedding size, the recurrent hidden size and the dense hidden size, which are tuned on the development datasets for the best performance.

2.2 ConceptNet Numberbatch

ConceptNet is a freely-available semantic network, designed to help computers understand the meanings of words that people use (Speer et al., 2017)[1]. Figure 1 illustrates an excerpt of the concept of ConceptNet. It has been used to create word embeddings – representations of word meanings as vectors, similar to word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or fastText (Bojanowski et al., 2016). These word embeddings are free, multilingual, aligned across languages, and designed to avoid representing harmful stereotypes. Their performance at word similarity, within and across languages, was shown to be state of the art at SemEval 2017 (Speer and Lowry-Duda, 2017).

ConceptNet Numberbatch is a set of semantic vectors which are trained on ConceptNet that can be used directly as a representation of word meanings. These embeddings benefit from the fact that they have semi-structured, common sense knowledge from ConceptNet, giving them a way to learn about words that isn't just observing them in context. Unlike most embeddings, ConceptNet Numberbatch is multilingual from the ground up. Words in different languages share a common semantic space, and that semantic space is informed by all of the languages. These appealing properties of ConceptNet embeddings make them suitable for multilingual processing tasks which deal with lexical semantics. Discrete structures of ConceptNet

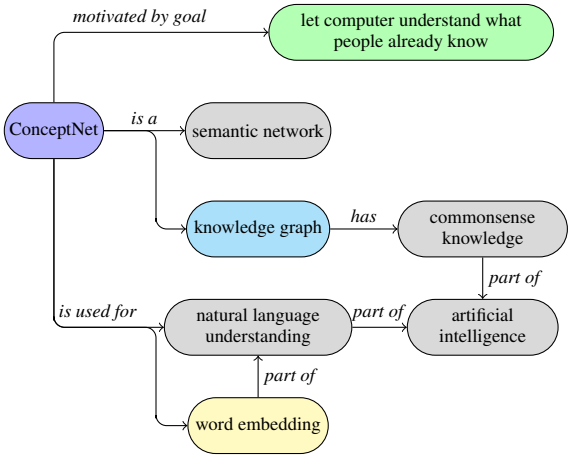

Figure 1: An illustration of ConceptNet in graph.

Language	Training	Dev.	Test
BCMS	368	122	123
EN	2,097	599	300
ES	3,467	989	495
FR	340,363	17,090	12,000
PT	3,467	991	495

Table 1: Statistics of the datasets used in the shared task.

have been recently exploited to improve natural language inference (Le-Hong and Cambria, 2023) and dependency parsing (Le-Hong and Cambria, 2024). In this work, we demonstrate that ConceptNet Numberbatch is also helpful in the problem of similar languages classification.

3 Results

3.1 Datasets

Some statistics of the five datasets of the DSL-ML-2024 shared task are given in Table 1. Some observations about the datasets are as follows.

First, the BCMS dataset contains texts in Bosnian, Croatian, Montenegrin, Serbian. There are no multi-label samples in the training split of this dataset but multi-label samples are present in the development and test splits. The size of this dataset is quite small but its sample text is often very long[2]. These properties make supervised models less accurate. Second, while the English and Portuguese datasets are of the same size, the French training dataset is about 100 times larger. This makes the training of French models much more time consuming.

[1] https://conceptnet.io/

[2] The longest training sample has 159,440 characters.

3.2 Experimental Settings

We carry out two experiments. In the first experiment, the model is applied on randomly initialized word embeddings which are fine-tuned on the training set. This experiment allows us to estimate the performance that a purely supervised learning system can achieve. In the second experiment, the same model is applied on the ConceptNet embeddings. This experiment investigates the advantage of using semantics-aware embeddings in detecting similar languages in a multilingual context. All the models have the same training objective, which is to set the score of the correct language label above the scores of incorrect ones. We use the common cross-entropy loss to minimize the objective function over the training data. This correlates with maximizing the number of correct predictions in the predicted outputs. Note that we consider each target label as atomic; for example, "*EN-GB,EN-US*" is considered a single label instead of two labels "*EN-GB*" and "*EN-US*"[3].

The ConceptNet Numberbatch word embeddings are freely available for download from the ConceptNet open data project[4]; we use the 19.08 version, *numberbatch-en-19.08.txt.gz* for English and *numberbatch-19.08.txt.gz* for multilingual word vectors, each also has 300 dimensions.

Since the model is trained in an end-to-end fashion, the gradients of the entire network, including the embedding matrices for tokens with respect to the sum of the losses are computed using the backpropagation algorithm. We perform multiple training epochs, using early stopping – the training process is stopped when the accuracy does not increase after three consecutive epochs on the development dataset. The maximal sequence length of each sentence is set to 40 tokens[5]. The models are all trained by the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5×10^{-5}. The batch size is set to 32[6]. On each dataset, we run a set of experiments with a different number of hidden units in each recurrent layer or in the dense layer (cf. subsection 3.4). Each experiment is run five times, its results are averaged for reporting.

[3]We have not tried any multi-label classification method in this task; the problem is considered multi-class classification.

[4]ConceptNet Numberbatch: https://github.com/commonsense/conceptnet-numberbatch

[5]This threshold is validated on the training split of the English dataset where 84.07% of samples are ≤ 40 tokens.

[6]All models are implemented in the Scala programming language using the BigDL library.

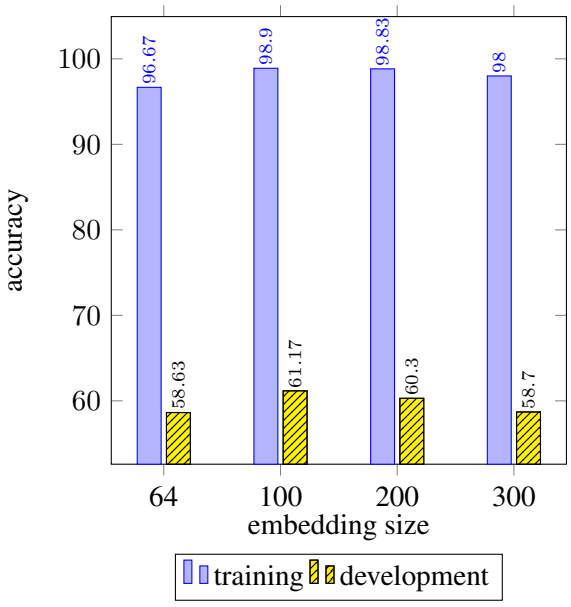

Figure 2: Maximal performance of the LSTM-r model with respect to the word embedding size and the recurrent size on the English training and development sets.

3.3 Evaluation Metrics

The organizers of this task provide an evaluation script and a baseline system which uses tf-idf-weighted character-level and word-level n-gram features in a linear SVM classifier[7]. The official scoring script provides per-class F1-scores, weighted and macro-averaged F1-scores. However, during the development stage, we did not use this script to evaluate our models; we used the common accuracy score on the training set and development set when validating the models. Despite of not being a good score for evaluating imbalanced datasets, this metric is found to be effective in model tuning.

3.4 Results

In this subsection, we first present the performance of our models on the development datasets. We then report the performance on the test datasets. We carried out the same experiments for all the languages. For brevity, we report only the process on the English dataset.

In the simple LSTM-r model where the word embeddings are initialized randomly, we vary the word embedding size w in the range $[64, 100, 200, 300]$ and the recurrent size h in the range $[100, 200, 300]$. The dense hidden size is fixed at 32 heuristically. Figure 2 shows the accuracy of this model on the

[7]https://github.com/yvesscherrer/DSL-ML-2024/

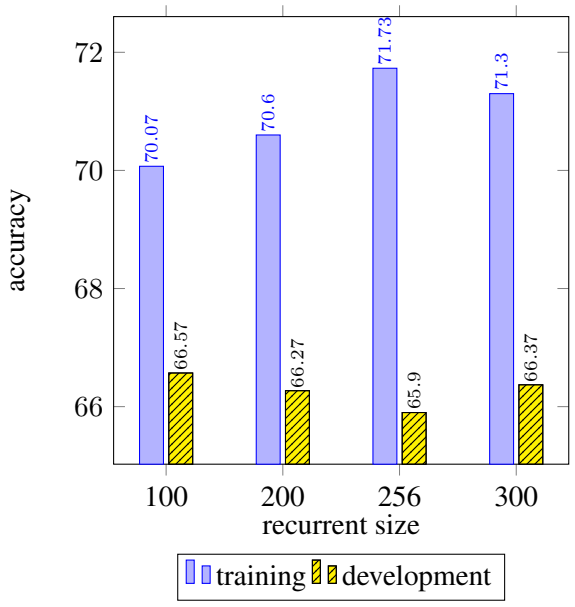

Figure 3: Performance of the LSTM-c model with respect to the recurrent size on the English training and development sets. The ConceptNet word embedding size is 300.

Language	M. F1	W. F1	EM	VLP
BCMS	27.22	36.97	00.00	1
EN	76.98	77.64	16.67	2
	75.88	76.30	26.67	1
ES	75.39	76.06	45.51	1
	74.14	74.36	42.31	2
FR	25.96	25.96	–	2
	25.74	25.74	–	1
PT	66.36	69.13	13.56	1
	56.58	62.01	00.00	2

Table 2: Official results of our systems on the test datasets as announced by the organizers. M. F1, W. F1 and EM is the macro F1, the weighted F1 and the exact match score, respectively. The VLP column is the submission index where number 1 indicates the LSTM-r model and number 2 indicates the LSTM-c model.

training and development splits. This model has a peak performance when $w = 100$ (and $h = 300$, not shown in the figure), having an accuracy of 61.17% on the development split. It seems that the model overfits the training data, which has a relatively small size.

In the enriched LSTM-c model where the words embeddings are ConceptNet embeddings, we vary the recurrent size h in the range $[100, 200, 256, 300]$. As above, the dense hidden size is 32. Figure 3 shows its accuracy. This model is not able to achieve a high accuracy on the training set but its development accuracy is significantly better than the LSTM-c model. This is maybe due to our choice of freezing the embedding layer, that is, the ConceptNet embeddings are not fine-tuned during training. The best accuracy of LSTM-c on the development dataset is 66.57%, which is 5.4% of absolute points better than that of LSTM-r.

Table 2 presents the official results of our LSTM-r and LSTM-c models on the test datasets of all the languages (Chifu et al., 2024). The ConceptNet embeddings are not available for BCMS languages, there is thus only one submission for this dataset.

As shown in these results, the ConceptNet embeddings help improve the accuracy of the English and French datasets by about 1.2% for English and 0.2% for French. However, they are not helpful for the Spanish and Portuguese datasets. It is surprising that the LSTM-c model is significantly worse than the LSTM-r model on the Portuguese datasets with a gap of about 10% of macro F1. It is possible due to a technical problem of our system during the training stage for this model. We plan to investigate further on this problem once the gold labels of the test datasets are available for additional analysis.

4 Conclusion

In this paper, we have presented a recurrent neural network model for tackling the problem of distinguishing similar languages. Our method utilizes semantics-aware ConceptNet embeddings for four languages. Despite its simplicity, the proposed model achieves relatively good results.

We are currently using the simple multi-class classification approach for this task. We plan to apply specific methods of multi-label classification for the task in a future work.

Recent works have shown that learning to classify texts can be beneficial by unsupervised representation learning methods such as contrastive learning (Su et al., 2022). The goal of contrastive learning is to learn a representation of text such that similar instances are close together in the representation space, while dissimilar instances are far apart. A combination of similarity embeddings learned by contrastive learning and semantics embeddings learned from knowledge graphs such as WordNet and ConceptNet can be helpful for this task.

Finally, in the last few years, pre-trained large language models such as XLM-R (Conneau et al.,

2020), GPT (Brown et al., 2020) and LLaMa (Touvron et al., 2023) are making new waves in the field of natural language processing due to their emergent ability and generalizability. We have investigated using a pre-trained XLM-R model for this shared task but initial results are mediocre compared to our proposed approach. However, a more throughout inquiry of using large language models is necessary before a firm conclusion about their usefulness can be drawn.

Acknowledgments

This study is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14.

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 *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Adrian Chifu, Goran Glavaš, Radu Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. FreCDo: A large corpus for French cross-domain dialect identification. *Procedia Computer Science*, 225:366–373.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, pages 1–15, San Diego, CA, USA.

Phuong Le-Hong and Erik Cambria. 2023. A semantics-aware approach for multilingual natural language inference. *Language Resources and Evaluation*, 57(2):611–639.

Phuong Le-Hong and Erik Cambria. 2024. Integrating graph embedding and neural models for improving transition-based dependency parsing. *Neural Computing and Applications*, 36(6):2999–3016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, pages 1–12, Scottsdale, Arizona, USA.

Aleksandra Miletić and Filip Miletić. 2024. A gold standard with silver linings: Scaling up annotation for distinguishing Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. European Language Resources Association.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, Doha, Qatar.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, pages 4444–4451.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland. Association for Computational Linguistics.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2024. Language variety identification with true labels. In *Proceedings of LREC-COLING*.