Using Social Networks to Improve Language Variety Identification with Neural Networks

Yasuhide Miura[†,‡]
yasuhide.miura@fujixerox.co.jp

Tomoki Taniguchi[†]
taniguchi.tomoki@fujixerox.co.jp

Motoki Taniguchi[†]
motoki.taniguchi@fujixerox.co.jp

Shotaro Misawa[†]
misawa.shotaro@fujixerox.co.jp

Tomoko Ohkuma[†]
ohkuma.tomoko@fujixerox.co.jp

[†]Fuji Xerox Co., Ltd.
[‡]Tokyo Institute of Technology

Abstract

We propose a hierarchical neural network model for language variety identification that integrates information from a social network. Recently, language variety identification has enjoyed heightened popularity as an advanced task of language identification. The proposed model uses additional texts from a social network to improve language variety identification from two perspectives. First, they are used to introduce the effects of homophily. Secondly, they are used as expanded training data for shared layers of the proposed model. By introducing information from social networks, the model improved its accuracy by 1.67–5.56. Compared to state-of-the-art baselines, these improved performances are better in English and comparable in Spanish. Furthermore, we analyzed the cases of Portuguese and Arabic when the model showed weak performances, and found that the effect of homophily is likely to be weak due to sparsity and noises compared to languages with the strong performances.

1 Introduction

Language identification is a fundamentally important natural language processing (NLP) task that is usually applied before more sophisticated grammatical or semantic analyses. It is especially important in cases when analyzing user generated contents such as social media, which include various languages often, without accurate language information. General purpose language identification tools such as *TextCat* (Cavnar and Trenkle, 1994) and *langid.py* (Lui and Baldwin, 2012) can identify 50–100 languages with accuracy of 86–99%. However, these tools have not considered discrimination between closely related language varieties.

Recently, language identification among similar languages or language varieties has been studied actively to realize more advanced language identification (Goutte et al., 2016). Since 2014, VarDial workshops, which specifically examine linguistic variation, have organized shared tasks of discriminating between similar languages (Zampieri et al., 2014). More recently, language variety analysis has attracted an author profiling community to include it in a PAN shared task that targets social media (Rangel Pardo et al., 2017). A language variety that a person uses often depends on his or her regional and cultural backgrounds. The identification of language variety can enhance a social media analysis by providing such background information.

We tackle this language variety identification in Twitter with a hierarchical neural network model (Lin et al., 2015; Yang et al., 2016b) integrating information from a social network. The use of social network information has shown effectiveness in analyzing various user attributes (Wang et al., 2014; Li et al., 2014, 2015; Rahimi et al., 2015a,b) where homophily (McPherson et al., 2001) exists. Neural networks have recently shown superior performance for solving a variety of problems in NLP. However, for language variety identification, sparse traditional models have shown stronger performance than deep neural models (Medvedeva et al., 2017). Numerous parameters in neural network models make it difficult to apply to language variety identification where the training data are

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 263–270,
Taipei, Taiwan, November 27 – December 1, 2017

limited to a maximum number of several thousands.

We expect to obtain two effects by introducing additional texts from a social network into our model. First, we introduce additional texts that are likely to share the same language variety by homophily. Secondly, we let several layers of our model be trained with more texts by sharing several layers of the model among the processes of a target user and its linked users with a social network. The contributions of this paper are the following:

1. We propose a novel neural network model that uses social network information for language variety identification in Twitter.
2. We show that additional texts of linked users can improve language variety identification.
3. We reveal that a neural network model can be efficiently trained by sharing layers within the processes of a target user and its linked users.

2 Related Works

2.1 Language Variety Identification

The increase of web documents in various languages has raised interest in identifying language varieties automatically. Inspired by some early works in Malay and Indonesian (Ranaivo-Malançon, 2006), south Slavic languages (Ljubešić et al., 2007), and Chinese varieties (Huang and Lee, 2008), studies of language varieties, similar languages, or dialects have expanded to examine numerous languages. The recent expansion of language variety identification has been well surveyed in works by Goutte et al. (2016) and Zampieri et al. (2017). As in other NLP tasks, various neural network models have been applied recently to language variety identification (Belinkov and Glass, 2016; Bjerva, 2016; Cianflone and Kosseim, 2016; Criscuolo and Aluisio, 2017; Medvedeva et al., 2017). However, these neural network models have shown inferior performance compared to sparse traditional models in comparisons (Malmasi et al., 2016; Zampieri et al., 2017).

2.2 NLP with Social Network Information

Social media have attracted numerous NLP studies to analyze its texts. Social media contain interactions among users such as follow, hashtag, mention, reply, and retweet. Many studies have exploited such social network information to enhance NLP models. The use of social networks has shown effectiveness for strengthening NLP tasks such as sentiment analysis (Speriosu et al., 2011; Tan et al., 2011; Vanzo et al., 2014; Ren et al., 2016; Yang and Eisenstein, 2017), skill inference (Wang et al., 2014), user attribute extraction (Li et al., 2014, 2015), geolocation prediction (Rahimi et al., 2015a,b), and entity linking (Yang et al., 2016a).

Integration of social network information into neural network models is accomplished in these studies through joint training (Li et al., 2015), context-based sub-networks (Ren et al., 2016), embedding of social network components (Yang et al., 2016a), and social attention (Yang and Eisenstein, 2017). These are effective approaches in terms of accuracy but they make models more difficult to train with additional parameters. We designed our model to share layers among different processes to facilitate training of the neural network model.

3 Models

3.1 NN-HIER

We prepare a basic neural network model NN-HIER, which is a variant of known hierarchical models (Lin et al., 2015; Yang et al., 2016b). NN-HIER in Figure 1 portays the architecture of this model. For each user, the model accepts the words of user tweets. The words are embedded with a word embedding layer and are processed with a recurrent neural network (RNN) layer, a max-pooling layer, an attention mechanism (Bahdanau et al., 2014) layer, and fully connected (FC) layers. As an implementation of RNN, we used Gated Recurrent Unit (GRU) (Cho et al., 2014) with a bi-directional setting.

The bi-directional GRU outputs $\overrightarrow{\boldsymbol{h}}$ and $\overleftarrow{\boldsymbol{h}}$ are concatenated to form $\boldsymbol{g}$ $where$ $\boldsymbol{g}_t = \overrightarrow{\boldsymbol{h}_t} \| \overleftarrow{\boldsymbol{h}_t}$. $\boldsymbol{g}$ is further processed with a max-over time process in $\mathrm{MaxPooling}$ to obtain a tweet representation $\boldsymbol{m}$. $\mathrm{Attention_U}$ computes a user representation $\boldsymbol{o}$ as a weighted sum of $\boldsymbol{m}_n$ with weight α_n:

$$\boldsymbol{o} = \sum_n \alpha_n \boldsymbol{m}_n$$

$$\alpha_n = \frac{\exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_n\right)}{\sum_l \exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_l\right)} \quad (1)$$

$$\boldsymbol{u}_n = \tanh\left(\boldsymbol{W}_\alpha \boldsymbol{m}_n + \boldsymbol{b}_\alpha\right) \quad (2)$$

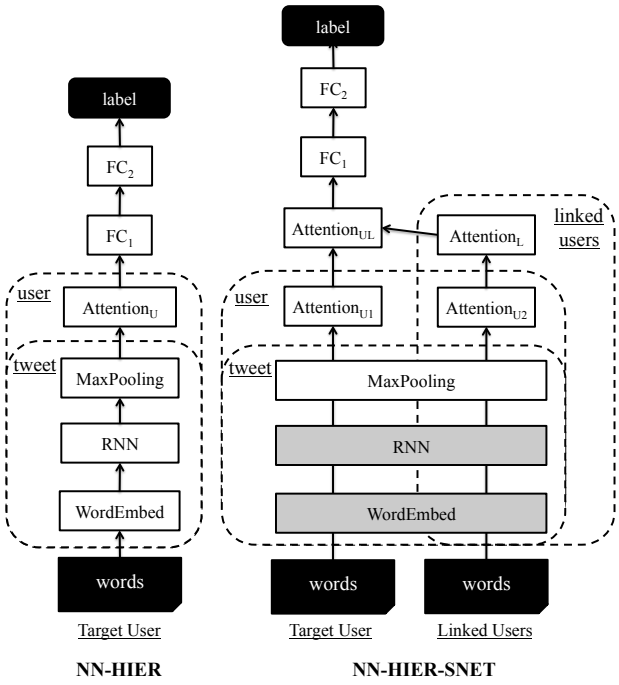

Figure 1: Architectures of NN-HIER and NN-HIER-SNET. tweet represents tweet-level processes, user represents user-level processes, and linked users represents linked-users-level processes. Shaded layers are layers with shared weights over the target user process and the linked users process.

where $\boldsymbol{v}_\alpha$ is a weight vector, $\boldsymbol{W}_\alpha$ is a weight matrix, and $\boldsymbol{b}_\alpha$ a bias vector. $\boldsymbol{u}_n$ is an attention context vector calculated from $\boldsymbol{m}_n$ with a single FC layer (Eq. 2). $\boldsymbol{u}_n$ is normalized with softmax to obtain α_n as a probability (Eq. 1). Finally, the user representation is passed respectively to FC_1 and FC_2.

3.2 NN-HIER-SNET

We extend NN-HIER by adding an additional level of hierarchy to process linked users of a target user. This extension is intended to introduce effects of homophily into our model. NN-HIER-SNET in Figure 1 presents this extended model. NN-HIER-SNET includes additional attention layers $\mathrm{Attention_L}$ and $\mathrm{Attention_{UL}}$ to process linked users. $\mathrm{Attention_L}$ accepts multiple user representations and combined them as in $\mathrm{Attention_U}$ to form a linked users representation. $\mathrm{Attention_{UL}}$ further merges a target user representation (an output of $\mathrm{Attention_{U1}}$) and $\mathrm{Attention_L}$ to obtain an updated target user representation.

An important characteristic of NN-HIER-SNET is that the weights of WordEmbed and RNN are shared across the target user process and the linked users process. This sharing allows the weights to be trained with more texts than those of NN-HIER. Attention processes over tweets are separated ($\mathrm{Attention_{U1}}$ and $\mathrm{Attention_{U2}}$) so that the target user process and the linked users process can pick tweets differently between the two kinds of user processes.

	en	es	pt	ar
$\#train_8$	2,880	3,360	960	1,920
$\#dev_1$	360	420	120	240
$\#test_1$	360	420	120	240
#total	3,600	4,200	1,200	2,400
#langvar	6	7	2	4
#mention	73,897	59,685	11,541	20,287

Table 1: Numbers of training data, development data, test data, entire data (total), language varieties (langvar), and mentioned users (mention).

	en	es	pt	ar
#node	77,497	63,885	12,741	22,687
avg degree	3.12	3.45	2.26	2.59
isolated nodes	3.92%	5.83%	6.33%	28.88%

Table 2: Characteristics of nodes in mention networks for each language. Avg degree is an average node degree and isolated nodes are the percentage of labeled nodes that are not connected to other labeled nodes.

4 Data

We used PAN@CLEF 2017 Author Profiling Training Corpus[1] to train the proposed models. The dataset consists of 11,400 Twitter users labeled with language variety of English (en), Spanish (es), Portuguese (pt), and Arabic (ar). Because the proposed model of Section 3.2 integrates texts of linked users, we additionally collected timelines of mentioned users in this dataset as linked users using Twitter REST APIs. #mention in Table 1 are the numbers of users mentioned for each language.

We divided this dataset into $train_8$, dev_1, and $test_1$ using a stratified sampling with a ratio of 8:1:1. Table 1 presents statistics of these divisions and Table 2 shows characteristics of nodes in the mention networks of each language. $test_1$ differs from the test data of PAN@CLEF 2017 Author Profiling Task. We chose to use a subset of the training data as test data because the true test data can not be accessed publicly (Potthast et al., 2014).

[1] http://pan.webis.de/clef17/pan17-web/author-profiling.html

Model	en	es	pt	ar
SVM-W2	83.06	95.74	98.61	80.00
SVM-W2C6	85.56	95.71	98.99	82.08
SVM-W2C6-SNET	82.69	92.86	99.17	80.42
SVM-W2C6-SNET-R	86.11	93.33	99.17	82.71
NN-HIER	85.83	93.57	99.17	78.75
NN-HIER-SNET	91.39	95.48	93.33	80.42

Table 3: Accuracies of the proposed models and the baselines. Underlined values represent the best values for each language.

5 Experiment

5.1 Baselines

We prepared four support vector machine based baselines: SVM-W2, SVM-W2C6, SVM-W2C6-SNET, and SVM-W2C6-SNET-R.

SVM-W2
A support vector machine model with tf-idf weighted word 1–2 grams. We prepared SVM-W2 with a soft margin setting and configured parameter $C \in \{0.1, 0.5, 1.0, 5.0, 1e^2, 5e^2, 1e^3\}$ using the development sets. For multi-class classification, we used an one-vs.-the-rest scheme.

SVM-W2C6
An extended model of SVM-W2 which additionally uses tf-idf weighted character 1–6 grams. This setting is simple, but similar models have shown state-of-the-art performance in past VarDial tasks (Malmasi et al., 2016; Zampieri et al., 2017).

SVM-W2C6-SNET
An extension of SVM-W2C6 with features of linked users. tf-idf weighted word 1–2 grams and tf-idf weighted character 1–6 grams of linked users were added to SVM-W2C6 with a feature space separated from other SVM-W2C6 features.

SVM-W2C6-SNET-R
A variant of SVM-W2C6-SNET with a restricted feature space for linked users. The size of feature space for linked users are restricted to be equal to the size of feature space for target users in this model. Since, in general, there are more texts of linked users than that of target users, this restriction suppresses the effect of social network information.

5.2 Model Configurations

We trained our model by stochastic gradient descent over shuffled mini-batches with cross-entropy loss as an objective function. Word embeddings were pre-training using streaming tweets by fastText (Bojanowski et al., 2016) using the skip-gram algorithm. The details of the model configurations including a text processor and layer unit sizes are described in Appendix A.

Model	en	es	pt	ar
Majority Baseline	16.67	14.29	50.00	25.00
NN-HIER	85.83	93.57	99.17	78.75
Label Propagation	75.28	78.81	83.33	60.42

Table 4: Accuracies of the label propagation approach and the comparison approaches.

5.3 Result

We evaluated NN-HIER, NN-HIER-SNET, and the baseline models using $train_8$, dev_1, and $test_1$ for each language. Table 3 presents the model accuracies. By introducing additional texts of linked users, the accuracy of the proposed model improved by 1.67–5.56, except for Portuguese. Even compared to the best performing baselines, the model performed better in English and comparably in Spanish. Compared to the improvements obtained in NN-HIER-SNET, the expansion texts of linked users in the baseline models has shown only slight improvements in English, Portuguese, and Arabic with SVM-W2C6-SNET-R.

6 Discussions

6.1 Effects of Homophily

The experiment revealed the effectiveness of combining texts of a target user with social network information for language variety identification. For comparison, we additionally performed a label propagation experiment to observe performances of language variety identification without texts. Following the approaches by Rahimi et al. (2015a) and Rahimi et al. (2015b), we extracted an undirected graph of the social network from target users and their linked users. The labels of training users were propagated to test users using the algorithm of Zhou et al. (2004) with $\alpha = 0.99$.

Table 4 presents the performance of this label propagation approach. The performance are better than the majority baseline but are substantially lower than those from our text model (NN-HIER). Especially, the performance of Arabic is weak compared to other languages since the percentage of isolated nodes is high in Arabic (Table 2). The result suggests that social network information without texts is ineffective for language variety identification, at least in a dataset of several thousand users.

Model	en	es	pt	ar
NN-HIER-SNET	91.39	95.48	93.33	80.42
NN-HIER-SNET-NS	75.56	84.76	93.33	80.00

Table 5: Accuracies of NN-HIER-SNET with non-shared (NS) layers.

6.2 Effects of Shared Layers

NN-HIER-SNET includes shared layers to suppress the increase of neural networks parameters. To ascertain the effects of these shared layers, we additionally evaluated NN-HIER-SNET with non-shared layers. NN-HIER-SNET-NS in Table 5 presents performances of this setting. As expected, the performances were fundamentally inferior to the shared layers architecture. They performed especially badly in English and Spanish, for which the numbers of mentioned users were high (Table 1). The result shows that the shared layer architecture is effective for language variety identification.

6.3 Social Network Characteristics and Performances of Proposed Model

NN-HIER-SNET showed improvements over the baseline models in English and Spanish. These two languages are more dense than Portuguese and Arabic in terms of average node degree (Table 2), and are likely to obtain richer information from social networks. We further investigated the languages of tweets in linked users to capture additional characteristics of social networks for each language. Table 6 shows the summary of this investigation. In all four languages, the top linked language is same as a target language. However, their percentages vary from 78.08–94.26%, indicating differences in the amount of texts in different languages. These texts with different languages will likely to be noises in the texts of linked users for language variety identification. As in average node degree, English and Spanish are in better conditions compared to Portuguese and Arabic with smaller noises. Sparsity and noises in social networks will likely to weaken the effect of homophily, resulting to small or negative improvements in performances.

7 Conclusion

We proposed a neural network model that integrates information from a social network for language variety identification. The model showed 1.67–5.56 improvements in accuracy from introducing additional texts with shared layers. Furthermore, compared to the performance of a state-of-the-art baseline model, the model performed better in English and comparably well in Spanish. In Portuguese and Arabic, the model performed weakly compared to the baseline models. We analyzed characteristic of social network in these languages and found that their sparsity and noises have possibly weakened the effect of homophily. The result underscores the promising future of applying neural network models to language variety identification.

Target Language	Linked Language 1st	2nd	3rd
en	en: 94.26%	und: 3.50%	fr: 0.39%
es	es: 87.09%	en: 6.67%	und: 4.22%
pt	pt: 78.08%	en: 9.48%	und: 7.14%
ar	ar: 82.07%	en: 9.39%	und: 6.77%

Table 6: Top 3 languages and their percentages in tweets of linked users. Language und is given in a case when the automatic language detection of a tweet has failed.

As future works of this study, we plan to expand the use of the proposed models for application to other user attributes. We expect that a user attribute having a tendency for homophily is likely to benefit from the proposed model as in language variety identification. Additionally, we plan to perform a comparison of the model against an alternative approach to introduce social network information. Recently, neural network models like Graph Convolutional Networks (Kipf and Welling, 2016) are proposed to process graph data. We would like to observe the differences between a hierarchal approach and a graph process approach in a utilization of social network information.

Acknowledgments

We would like to thank the members of Okumura–Takamura Group at Tokyo Institute of Technology for having fruitful discussions about social media analysis. We would also like to thank the anonymous reviewer for their comments to improve this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yonatan Belinkov and James Glass. 2016. A character-level convolutional neural network for distinguishing similar languages and dialects. In *Proceedings*

of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), pages 145–152.

Johannes Bjerva. 2016. Byte-based language identification with deep convolutional networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 119–125.

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

William B. Cavnar and John M. Trenkle. 1994. Ngram-based text categorization. In *Proceedings of the Third Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Andre Cianflone and Leila Kosseim. 2016. N-gram and neural language models for discriminating similar languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 243–250.

Marcelo Criscuolo and Sandra Maria Aluisio. 2017. Discriminating between similar languages with word-level convolutional neural networks. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 124–130.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

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.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 165–174.

Jiwei Li, Alan Ritter, and Dan Jurafsky. 2015. Learning multi-faceted representations of individuals from heterogeneous evidence using neural networks. *arXiv preprint arXiv:1510.05198*, abs/1510.05198.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 899–907.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language identification: how to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg 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 (VarDial3)*, pages 1–14.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

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.

Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the reproducibility of PAN's shared tasks: Plagiarism detection, author identification, and author profiling. In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015a. Twitter user geolocation using a unified text and network prediction model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 630–636.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015b. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367.

Bali Ranaivo-Malançon. 2006. Automatic identification of close languages – case study: Malay and indonesian. *ECTI Transaction on Computer and Information Technology*, 2(2):126–133.

Francisco Manuel Rangel Pardo, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In *Working Notes Papers of the CLEF 2017 Evaluation Labs*, volume 1866.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive Twitter sentiment classification using neural network. In *Proceedings of Thirtieth AAAI Conference on Artificial Intelligence*.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, pages 53–63.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1397–1405.

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for sentiment analysis in Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 2345–2354.

Zhongqing Wang, Shoushan Li, Hanxiao Shi, and Guodong Zhou. 2014. Skill inference with personal and skill connections. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 520–529.

Yi Yang, Ming-Wei Chang, and Jacob Eisenstein. 2016a. Toward socially-infused information extraction: Embedding authors, mentions, and entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1461.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016b. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

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 the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15.

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.

Denny Zhou, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328.

A Supplemental Materials

Text Processor We applied unicode normalization, Twitter user name normalization, and URL normalization for text pre-processing. Pre-processed texts were tokenized with Twokenize[2] for English and NLTK[3] WordPunctTokenizer for three other languages. Words are converted to lower case form, with ignored capitalization.

Pre-training of Embeddings

We collected tweets using Twitter Streaming APIs to pre-train the word embedding matrix of the models. Neural network models are known to perform better when word embeddings are pre-trained by a large-scale dataset. The following steps describe details of the collection process.

1. Tweets with lang metadata of en, es, pt, and ar were collected via Twitter Streaming APIs during March–May 2017.
2. Retweets are removed from the collected tweets.
3. Tweets posted by bots[4] are deleted from the collected tweets.

Table 7 presents the number of resulting tweets. We pre-trained word embeddings with these tweets by fastText using the skip-gram algorithm. The pre-training parameters are dimension=100, learning rate=0.025, window size=5, negative sample size=5, and epoch=5.

Layer Unit Sizes & Maximum Linked Users

We set the following unit size of $\mathrm{RNN} = 100$, unit size of $FC_1 = 100$, and the unit size of FC_2 to the number of labels. The context vector sizes of attention layers were set to $\mathrm{Attention_U} = 200$, $\mathrm{Attention_{U1}} = 200$, $\mathrm{Attention_{U2}} = 200$, and $\mathrm{Attention_{UL}} = 200$. To make our model tractable, we limited the maximum number of linked users to 3.

Optimization Strategy

We used cross-entropy loss as an objective function of the proposed models. The objective function was minimized through stochastic gradient descent over shuffled mini-batches with a learning rate of 0.01, momentum of 0.9, and gradient clipping of 3.0. The model parameters were set to the best performing parameters in terms of loss in the development data.

	en	es	pt	ar
#tweet	12.39M	3.71M	3.16M	2.87M

Table 7: Number of tweets collected for each language with Twitter Streaming APIs.

[2] https://github.com/myleott/ark-twokenize-py

[3] http://www.nltk.org/

[4] We assembled a Twitter client list consisting of 80 clients that are used for manual postings.