Franco Salvador, M.; Rangel, F.; Rosso, P.; Taulé, M.; Martí, MA. (2015). Language variety identification using distributed representations of words and documents. En Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings. Springer International Publishing. 28-40. doi:10.1007/978-3-319-24027-5_3.

Language Variety Identification using Distributed Representations of Words and Documents*

Marc Franco-Salvador[1], Francisco Rangel[1,2], Paolo Rosso[1], Mariona Taulé[3], and M. Antònia Martí[3]

[1] Universitat Politècnica de València, Spain
mfranco@prhlt.upv.es, prosso@dsic.upv.es
[2] Autoritas Consulting, S.A., Spain
francisco.rangel@autoritas.es
[3] Universitat de Barcelona, Spain
{mtaule,amarti}@ub.edu

Abstract. Language variety identification is an author profiling subtask which aims to detect lexical and semantic variations in order to classify different varieties of the same language. In this work we focus on the use of distributed representations of words and documents using the continuous Skip-gram model. We compare this model with three recent approaches: Information Gain Word-Patterns, TF-IDF graphs and Emotion-labeled Graphs, in addition to several baselines. We evaluate the models introducing the Hispablogs dataset, a new collection of Spanish blogs from five different countries: Argentina, Chile, Mexico, Peru and Spain. Experimental results show state-of-the-art performance in language variety identification. In addition, our empirical analysis provides interesting insights on the use of the evaluated approaches.

Keywords: Author profiling, Language variety identification, Distributed representations, Information Gain Word-Patterns, TF-IDF graphs, Emotion-labeled Graphs

1 Introduction

Author profiling aims to identify the linguistic profile of an author on the basis of his writing style. It is used to determine an author's gender, age, personality type and native language, among other traits. In this work we focus on language variety identification. Native language identification aims at identifying the native language of an author on the basis of a text he has written in another language. In contrast, the aim of language variety identification sub-task is to label the texts with its corresponding variant. For example, with a text written in Spanish, the Argentinean, Chilean, Mexican, Peruvian or European Spanish variant. This task has special relevance in text mining in social media. Given that there are millions of user blogs and posts in any given language, it

* This research has been carried out within the framework of the European Commission WIQ-EI IRSES (no. 269180) and DIANA - Finding Hidden Knowledge in Texts (TIN2012-38603-C02) projects. The work of the second author was partially funded by Autoritas Consulting SA and by Spanish the Ministry of Economics by means of a ECOPORTUNITY IPT-2012-1220-430000 grant. We would like to thank Tomas Mikolov for his support and comments about distributed representations.

is important to identify the concrete variety of the language in order to attribute and exploit correctly the information they contain, e.g. opinions about political elections in Mexico do not have the same relevance in Spain, which is at 9,000 kilometers away.

In this work, we are interested in comparing the performance of three recent approaches that we previously applied to other author profiling tasks: Information Gain Word-Patterns, TF-IDF Graphs and Emotion-labeled Graphs. Furthermore, due to the increasing popularity of distributed representations [5], we use the continuous Skip-gram model to generate distributed representations of words, i.e., n-dimensional vectors, applying further refinements in order to be able to use them on documents. In addition, we use the Sentence Vector variation to directly generate representations of documents. We also compare the aforementioned approaches with several baselines: bag-of-words, character 4-grams and TF-IDF 2-grams. In order to evaluate these models, we are presenting the Hispablogs dataset, a new collection of Spanish blogs from five different countries: Argentina, Chile, Mexico, Peru and Spain.

The rest of the paper is structured as follows. Section 2 studies related work in the field of language variety identification. In Section 3 we overview the continuous Skip-gram model, its Sentence Vectors variation, and explain how we generated distributed vectors of documents. Section 4 details the three compared approaches. Finally, in Section 5 we introduce the Hispablogs dataset and evaluate the different approaches to the task of language variety identification.

2 Related Work

Author profiling is a field of growing interest for the research community. In the last years several tasks have been hold on different demographic aspects: i) native language identification at the BEA-8 workshop at NAACL-HT 2013[4]; ii) personality recognition at ICWSM 2013[5] and at ACMMM 2014[6]; and iii) age and gender identification (both in English and Spanish) at PAN 2013[7] and PAN 2014[8] tracks at the CLEF initiative. In PAN 2015[9] the task is concerned with predicting the author's age, gender, and personality. Interest in author profiling was also expressed by industry representatives in the Kaggle platform[10], where companies and research centers share their needs and independent researchers can join the challenge of solving them. A small number of tasks related to author profiling have been organised: i) psychopathy prediction based

[4] https://sites.google.com/site/nlisharedtask2013/
[5] http://mypersonality.org/wiki/doku.php?id=wcpr13
[6] https://sites.google.com/site/wcprst/home/wcpr14
[7] http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html
[8] http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-web/author-profiling.html
[9] http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html
[10] http://www.kaggle.com/

on Twitter usage[11]; ii) personality prediction based on Twitter stream[12]; iii) and gender prediction from handwriting[13].

With respect to previous works on author profiling, in [17] the author divides writing style features in two types: content and style-based features. In [20, 19] participants in the PAN shared task at CLEF approached the task of age and gender identification using combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, etc., together with part-of-speech (PoS) tags and content-based features such as bag-of-words, the TF-IDF of words, dictionary-based words, topic-based words, entropy-based words, and content-based features obtained with Latent Semantic Analysis [3]. Affectivity is explored in [15], showing the relationship between gender and the expression of emotions.

Despite the growing interest in author profiling problems, little attention has been given to language variety identification. In [24] the authors investigated varieties of Portuguese. They collected 1,000 news articles and applied different features such as word and character n-grams to them. Similarly, in [21] the authors differentiate between six different varieties of Arabic in blogs and forums using character n-gram features. Concerning Spanish language varieties, in [9] the authors collected a dataset from Twitter, focusing on varieties from Argentina, Chile, Colombia, Mexico and Spain. They applied four types of features: character n-gram frequency profiles, character n-gram language models, LZW compression and syllable-based language models, all combined with a meta-classifier and evaluated with cross-validation.

In this work we focus on Spanish language variety identification with some differences with regard to the previous works: i) we focus on larger social media texts because we are interested in investigating more complex features which may also model discourse structure; ii) we evaluate the proposed methods both with cross-validation and with an independent test set generated from different authors in order to reduce possible overfitting; iii) the Twitter dataset compiled in the previous work is not publicly available; in contrast, in line with the CLEF initiative we are making our datset available to the research community.

3 Continuous Skip-gram Model

The use of log-linear models has been proposed [11] as an efficient way to generate distributed representations of words, since they reduce the complexity of the hidden layer thereby improving efficiency. The continuous Bag-of-Words model attempts to maximize the classification of a word by using the surrounding words without taking into account the order of the sequence. In contrast, the continuous Skip-gram model uses word ordering to sample distant word that appear less frequently during training time. Compared to traditional approaches such as the Feedforward Neural Net Language model [2] and the Recurrent Neural Net Language model [12], these approaches obtained better performance with a considerably lower training time in semantic and syntactic word relationship tasks. Experimental results also demonstrated that the Skip-gram model offers better performance on average, excelling especially at the semantic

[11] http://www.kaggle.com/c/twitter-psychopathy-prediction

[12] http://www.kaggle.com/c/twitter-personality-prediction

[13] http://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting

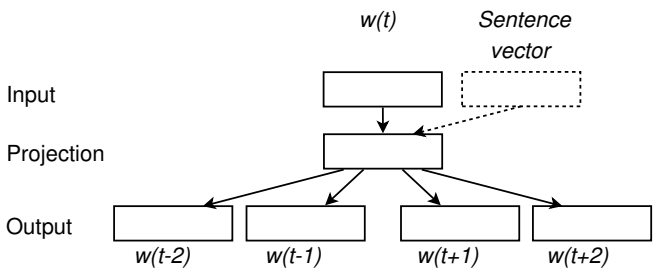

Fig. 1. Skip-gram model architecture. The objective is to predict words within a certain range before and after the current word. Dashed part is used only in place of $w(t)$ when learning sentence vectors.

level. Therefore, in this work we selected that approach to generate our distributed representations.

The continuous Skip-gram model [11, 13] is an iterative algorithm which attempts to maximize the classification of the context surrounding a word (see Figure 1). Formally, given a word $w(t)$, and its surrounding words $w(t-c),\ w(t-c+1), ...,\ w(t+c)$ inside a window of size $2c+1$, the training objective is to maximize the average of the log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W}\exp({v'_w}^T v_{w_I})} \tag{2}$$

Although $p(w_{t+j}|w_t)$ can be estimated using the softmax function (Eq. 2) [1], its normalization depends on vocabulary size W which makes its usage impractical for high values of W. For this reason, more computationally efficient alternatives are used instead. Hierarchical softmax has been proposed [16] to approximate the results of the softmax function. This function is based on a binary tree with all $w \in W$ as leaves, each node being the relative probabilities of its child nodes. The algorithm makes it necessary only to process $\log_2(W)$ words for each probability estimation. An alternative introduced in [13] is negative sampling. This function is a simplified version of the Noise Contrastive Estimation (NCE) [4, 14], which is only concerned with preserving vector quality in the context of Skip-gram learning. The basic idea is to use logistic regression to distinguish the target word W_O from a noise distribution $P_n(w)$, having k negative samples for each word. Formally, the negative sampling estimates $p(w_O|w_I)$ as follows:

$$\log\sigma({v'_{w_O}}^T v_{w_I}) + \sum_{i=1}^{k}\mathbb{E}_{w_i}\sim P_n(w)\Big[\log\sigma(-{v'_{w_i}}^T v_{w_I})\Big] \tag{3}$$

where $\sigma(x) = 1/(1+\exp(-x))$. Note that computational complexity is linear with the number of negative samples k. The experimental results in [13] show that this function obtains better results at the semantic level than hierarchical softmax and NCE. Therefore, in this work we will use negative sampling in all our experiments.

3.1 Learning Sentence Vectors

The continuous Skip-gram model can be easily adapted to generate representative vectors of sentences (or documents). Sentence vectors (SenVec) [7] follows Skip-gram architecture to train a special vector sv representing the sentence. Basically, before each context window movement, SenVec uses sv in place of $w(t)$ with the objective of maximizing the classification of the surrounding words (see Figure 1).

3.2 Classification using Distributed Representations

Although SenVec is directly applicable as input to a classifier, we need to combine the word vectors generated with the Skip-gram model to use them when classifying documents. The use of Convolutional Neural Networks with Skip-gram word vectors as input has been proposed [6] with excellent results for sentence classification tasks. However, due to the computational complexity of these networks, we will explore that option in the future and we now employ a simpler solution. Having a list of word vectors[14] $(w_1, w_2, ..., w_n)$ belonging to a document, we generate a vector representation v of its content by estimating the average of their dimensions: $v = n^{-1} \sum_{i=1}^{n} w_i$. This combination is directly named Skip-gram in the evaluation.

4 Alternative methods for language variety identification

We are interested in comparing the performance of distributed representation against three alternative representations successfully used in other author profiling tasks: TF-IDF graphs[15], Information Gain Word-Patterns and Emotion-labeled Graphs. We describe the latter two below. We also compare them against different baselines in author profiling such as Bag-of-Words (BOW), Char. 4-grams and TF-IDF 2-grams[16].

4.1 Information Gain Word-Patterns

Information Gain Word-Patterns (IG-WP) [10] is a bottom-up method for obtaining lexico-syntactic patterns aiming to represent the content of documents. This method is based on the pattern-construction hypothesis, which states that those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions[17]. This method consists of a pipe-line of the following processes. First, the source corpus is morphologically annotated with lemma and PoS tagging using Freeling library[18], and syntactically annotated with dependencies using Treeler[19]. Secondly, a Vector Space Model (VSM) [22] matrix is built, in which contexts are modeled as dependency relations between two lemmas. For each lemma in the

[14] We allow the use of word repetitions.

[15] We represent each word as a node and each edge defines a sequence between words.

[16] We tested the value of n iterating for each representation from 1 to 10. The best results were achieved with n equal to 1, 4 and 2 respectively. In all of them the 10,000 most frequent grams were selected.

[17] A construction is a recurrent pattern in language.

[18] `http://nlp.lsi.upc.edu/freeling/`

[19] Treeler is an open-source C++ library of structure prediction methods focussing on tagging and parsing. To get Treeler: `http://devel.cpl.upc.edu/treeler/svn/trunk`

Table 1. Cluster 25 with theirs corresponding lemmas

Cluster: 25	
Lemmas	barba, bigote, cabellera, cabello, cana, ceja, hebra, mecha, mechón, melena, pelo, peluca, pestaña, rizo, trenza (*beard, moustache, head of hair, hair, grey hair, eyebrow, thread, wick, lock of hair, fur, wig, eyelash, curl, braid*)

Table 2. Cluster 643 related to cluster 25

Related_cluster	**Context_set**	**Lemmas**
643	<:* (2.2) <:subj (2.2) <:cd (1.3)	afeitar, ahuecar, alisar, cepillar, encrespar, enmarañar, erizar, mesar, ondear, peinar, rapar, rizar, sombrear, trenzar, tupir (*to shave, to hollow out, to straighten, to brush, to frizz, to tangle, to make stand on end, to pull on, to wave, to comb, to crop, to curl, to tint, to braid, to thicken*)

rows (source lemma) of the matrix, a context is defined by a tuple of three elements: the direction of the dependency, the dependency label and the target lemma:

matrix_context = (dep-dir, dep-lab, lemma-context),

followed by the examples of matrix-context:

$context_1$ = (<, *subj,* ***robar*** *(to steal))*,
$context_2$ = (<, *dobj,* ***peinar*** *(to comb))*,

where, 'subj' and 'dobj' stand for subject and direct object respectively, < indicates the dependency direction, that is, that the lemma in the context is the parent node of the source lemma (in these cases, 'robar' ('to steal') and 'peinar' ('to comb')).

Then, we used the CLUTO toolkit[20] to obtain the clusters of semantically related words that share the same contexts. Next, the relationships between clusters are established using the most descriptive and discriminative contexts of each cluster. Each context consists of a dependency direction, a dependency label, a lemma and a score:

cluster_context = (dep-dir, dep-lab, lemma-context, score)

We obtain as a result a graph of related clusters, exemplified in Table 1 and 2, where cluster 25 is related to cluster 643 by means of the subject and direct object relationships. Table 1 describes the lemmas in cluster 25 (translated lemmas in English appear in italics). Table 2 shows one of the related clusters (i.e., 643) (first column) as a result of the linking cluster process for cluster 25. The second column shows the context_set that relates cluster 25 to cluster 643, and the third column describes the lemmas in the related cluster.

[20] `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

All members (nouns) in cluster 25 are good candidates to be subjects and direct objects of all members (verbs) of cluster 643.

Finally, a set of lexico-syntactic patterns are derived after applying different filters to avoid spurious relationships. The lexico-syntactic patterns are tuples involving two lemmas, related by both a dependency direction and a dependency label:

pattern = *(lemma$_u$, dep-dir, dep-lab, lemma$_v$)*

Considering the examples of cluster 25 and 643, we generated all possible combinations of every lemma from cluster 25 with every lemma in cluster 643. Examples of lexico-syntactic patterns derived from the related clusters 25 and 643 are:

*(**bigote**$_{c_{25}}$ (moustache), <, dobj, **afeitar**$_{c_{643}}$ (to shave))*,
*(**peluca**$_{c_{25}}$ (wig), <, dobj, **peinar**$_{c_{643}}$ (to comb))*,
*(**pelo**$_{c_{25}}$ (hair), <, subj, **encrespar**$_{c_{643}}$ (to curl))*

In the experiments carried out we selected as features the set of 1,000 words from the obtained patterns with the highest information gain. We used the Araknion dataset [10] as input to IG-WP to generate our Spanish patterns.

4.2 Emotion-labeled Graphs

The Emotion-labeled Graphs (EmoGraphs) model [18] obtains morphosyntactic categories from the Freeling library for each word in the text. Each PoS is modeled as a node in the graph and each edge defines a PoS sequence in the text. The graph obtained is enriched with semantic and affective information. Adjectives are annotated with their polarity and the Spanish Emotion Lexicon [23] is used to identify their associated emotions. WordNet Domains[21] is used to obtain the topics of nouns. On the basis of what was investigated in [8], verbs are annotated with one of the following semantic categories: i) perception (see, listen, smell...); ii) understanding (know, understand, think...); iii) doubt (doubt, ignore...); iv) language (tell, say, declare, speak...); v) emotion (feel, want, love...); vi) and will (must, forbid, allow...). We can see an example in Figure 2.

Once the graph is built, the objective is to use a machine learning approach to classify texts into its corresponding language variety. We obtain two kind of features on the basis of graph analysis: i) general properties of the graph describing the overall style of the modelled texts, such as nodes-edges ratio, average degree, weighted average degree, diameter, density, modularity, cluster coefficient or average path length; ii) and specific properties of its nodes and how they are related to each other, such us eigenvector and betweenness values.

EmoGraphs aims at modelling the way the authors express their emotions in the discourse structure and offers a competent representation in age and gender author profiling tasks [18].

[21] `http://wndomains.fbk.eu/`

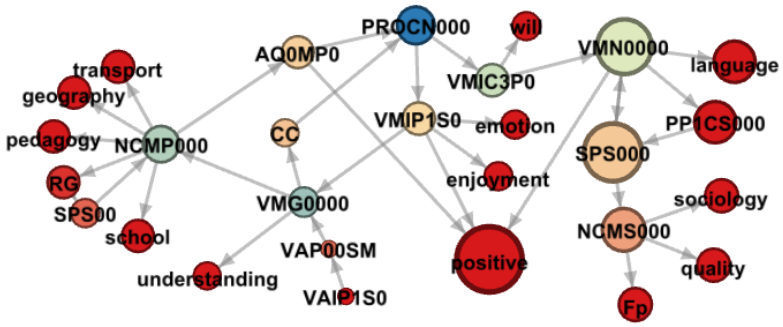

Fig. 2. EmoGraph of "He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público" (*"I have been taking online courses about valuable subjects that I enjoy studying and might help me to speak in public"*). Node sizes are proportional to its eigenvector and node colors depend on its betweenness.

5 Evaluation

In this section we evaluate the performance of the aforementioned models for the language variety identification task. Given a document d and a corpus D_{tr} with documents in C different language varieties, a system has to classify d into one of the categories of C using the labeled collection D_{tr}.

Dataset and Methodology To perform this task we created and used the Hispablogs dataset,[22] a new collection of Spanish blogs from five different countries: Argentina, Chile, Mexico, Peru and Spain. There are 450 training and 200 testing blogs respectively for each language variety, with a total of 2,250 and 1,000 blogs. Each user blog is represented by a set of user posts, with 10 posts per user/blog. We measured the quality of the models by evaluating the accuracy of the classification of the test set using a model trained with the training set. We observed that during the prototyping step, sentence vectors and word vector averages offered better results when they were estimated from a reduced number of words. Taking advantage of the dataset, we treated each post as an independent instance[23], and determined the language variety of the blog in function of the probabilities of classification of its posts: $class = \text{argmax}_{c \in C} \sum_{i=1}^{n} P(c|po_i)$, where C is the total number of classes and $(po_1, ..., po_n)$ is the list of posts in a concrete blog. Following state-of-the-art approach [7], in the evaluation we used a logistic classifier[24] for both SenVec and Skip-gram approaches[25].

[22] The Hispablogs dataset can be downloaded at: `https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs`

[23] Although our method ensures that all contexts are kept together, a sliding window could be used as an alternative.

[24] Similar results with higher training time were obtained with other classifiers such as Support Vector Machines.

[25] We used 300-dimensional vectors, context windows of size 10, and 20 negative words for each sample. We preprocessed the text with word lowercase, tokenization, removing the words of length one, and with phrase detection using word2vec tools: `https://code.google.com/p/word2vec/`

Method	Accuracy
Skip-gram	**0.722**
SenVec	**0.708**
BOW	0.527
IG-WP	0.520
Char. 4-grams	0.515
EmoGraphs	0.393
TF-IDF 2-grams	0.322
Random baseline	0.200
TF-IDF graphs	0.181

Table 3. Accuracy results in language variety identification.

We compared Skip-gram and SenVec approaches with the BOW, Char. 4-grams, TF-IDF 2-grams, TF-IDF graphs, EmoGraphs and IG-WP models (cfr. Section 4). As we can see in Table 3,[26] TF-IDF graphs obtained the lowest results, even lower than the random baseline (0.2 accuracy), followed by EmoGraphs. Looking at the results of TF-IDF 2-grams, we think that in this concrete task TF-IDF-based models are not able to capture differences between language varieties. However, EmoGraphs took advantage of additional information (topics, verbs, sentiments, emotions) to achieve a better performance. The two baselines, BOW and Char. 4-grams, were competitive despite their simplicity. Character n-gram models proved able to extract syntactic variations (differences in vocabulary, verbal inflections) between speakers of different language varieties. The IG-WP approach does not seem to outperform BOW, but demonstrated the potentiality of word-patterns and has the advantage of reducing considerably dimensionality by taking into account linguistic information. We note that in this task of language variety identification, content-based features such as BOW or IG-WP obtained better results than style-based ones such as EmoGraphs or TF-IDF graphs. This may be due to the fact that language variety relies more on the use of words than on discourse structure. Finally, both Skip-gram and SenVec models based on distributed representations significantly outperformed the others. Using the average of the word vectors, the Skip-gram model performs slightly better than SenVec, which infers a unique representation of documents, and proves to be a good alternative to more complex approaches. We think that the use of user blog posts as representations, instead of complete blogs, may have helped to reduce the noise in the vectors in both approaches.

In Figure 3, we highlight the capability of distributed representations to model the semantic properties of language. Comparing how all the models learn their features over the complete training partition, and evaluating their classifiers with cross-validation on the same dataset, we appreciate a very low improvement compared to training features on a different dataset (test set setting). Other models seem to experience some kind of over-fitting, obtaining much higher results in the cross-validation setting.

We can observe in Table 4 the difference in difficulty in the classification of Spanish language varieties using the Skip-gram model. The Spain-Spanish variety is the easiest one to detect compared to the Argentinian variety, which has the lowest results. In

[26] In this work, statistically significant results according to a χ^2 test are highlighted in bold.

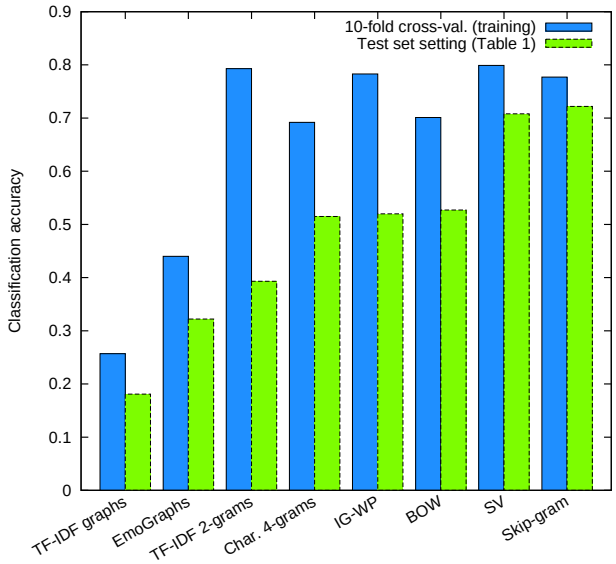

Fig. 3. Models over-fitting analysis.

	Classified as...				
Lang.	AR	CL	ES	MX	PE
AR	58.5	8	8.5	11	14
CL	5	73.5	5	6	10.5
ES	3	3.5	85.5	4	4
MX	8	4.5	5	70	12.5
PE	6.5	6	4	10	73.5

Table 4. Test set confusion matrix (in %) of Skip-gram model in language variety identification.

general, Latin American varieties are closer to each other and it is more difficult to differentiate between them.

6 Conclusions

In the task of Spanish language variety identification, we introduced Hispablogs -a new collection of Spanish blogs from five different countries-, and evaluated two continuous Skip-gram-based approaches: vectors of words and documents. Compared to the alternative approaches that we previously used in other author profiling tasks (e.g. EmoGraphs), the results obtained using Skip-gram are significantly superior, especially when evaluated with an independent dataset. This may be due to their ability to model semantics. In this particular task, features that model contents perform better than features which model the discourse structure. This suggest that language varieties differ more in the use of words at the lexical level than in the discourse structure, that is, what is said is more important than the way it is said. Future work will investigate further how to apply distributed representations to other author profiling tasks. We are also interested in comparing our approaches to [9] when they release their dataset.

References

1. Barto, A.G.: Reinforcement learning: An introduction. MIT press (1998)
2. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. The Journal of Machine Learning Research 3, 1137–1155 (2003)
3. Dumais, S.T.: Latent semantic analysis. Annual review of information science and technology 38(1), 188–230 (2004)
4. Gutmann, M.U., Hyvärinen, A.: Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. The Journal of Machine Learning Research 13(1), 307–361 (2012)
5. Hinton, G.E., McClelland, J.L., Rumelhart, D.E.: Distributed representations. Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press (1986)
6. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the International Conference on Empirical Methods in Natural Language Processing (2014)
7. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (2014)
8. Levin, B.: English verb classes and alternations. University of Chicago Press, Chicago (1993)
9. Maier, W., Gómez-Rodríguez, C.: Language variety identification in spanish tweets. In: Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants. pp. 25–35. Association for Computational Linguistics, Doha, Qatar (October 2014), http://emnlp2014.org/workshops/LT4CloseLang/call.html
10. Martí, M.A., Bertran, M., Taulé, M., Salamó, M.: Distributional approach based on syntactic dependencies for discovering constructions. Computational Linguistics (under review) (2015)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at International Conference on Learning Representations (2013)
12. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010. pp. 1045–1048 (2010)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26. pp. 3111–3119 (2013)
14. Mnih, A., Teh, Y.W.: A fast and simple algorithm for training neural probabilistic language models. arXiv preprint arXiv:1206.6426 (2012)
15. Mohammad, S.M., Yang, T.: Tracking sentiment in mail: how gender differ on emotional axes. In: In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (2011)
16. Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In: Proceedings of the international workshop on artificial intelligence and statistics. pp. 246–252. Citeseer (2005)
17. Pennebaker, J.W.: The secret life of pronouns: What our words say about us. Bloomsbury Press (2011)
18. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information Processing & Management, Special Issue on Emotion and Sentiment in Social and Expressive Media (In Press) (2015)

19. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: In: Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 1180 (2014)
20. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: In: Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops. CEUR-WS.org, vol. 1179 (2013)
21. Sadat, F., Kazemi, F., Farzindar, A.: Automatic identification of arabic language varieties and dialects in social media. In: In Proceeding of the 1st. International Workshop on Social Media Retrieval and Analysis SoMeRa (2014)
22. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
23. Sidorov, G., Miranda-Jimnez, S., Viveros-Jimnez, F., Gelbukh, F., Castro-Snchez, N., Velsquez, F., Daz-Rangel, I., Surez-Guerra, S., Trevio, A., Gordon-Miranda, J.: Empirical study of opinion mining in spanish tweets. In: 11th Mexican International Conference on Artificial Intelligence, MICAI. pp. 1–4 (2012)
24. Zampieri, M., Gebrekidan-Gebre, B.: Automatic identification of language varieties: The case of portuguese. In: In Proceedings of the Conference on Natural Language Processing (2012)