CIC-FBK Approach to Native Language Identification

Ilia Markov[1], **Lingzhen Chen**[2], **Carlo Strapparava**[3], **Grigori Sidorov**[1]

[1]Instituto Politécnico Nacional, CIC, Mexico City, Mexico
[2]University of Trento, DISI, Trento, Italy
[3]FBK-irst, Trento, Italy
imarkov@nlp.cic.ipn.mx, lzchen.cs@gmail.com,
strappa@fbk.eu, sidorov@cic.ipn.mx

Abstract

We present the CIC-FBK system, which took part in the Native Language Identification (NLI) Shared Task 2017. Our approach combines features commonly used in previous NLI research, i.e., word n-grams, lemma n-grams, part-of-speech n-grams, and function words, with recently introduced character n-grams from misspelled words, and features that are novel in this task, such as typed character n-grams, and syntactic n-grams of words and of syntactic relation tags. We use log-entropy weighting scheme and perform classification using the Support Vector Machines (SVM) algorithm. Our system achieved 0.8808 macro-averaged F1-score and shared the 1st rank in the NLI Shared Task 2017 scoring.

1 Introduction

Native language identification (NLI) is a natural language processing (NLP) task that aims at automatically identifying the native language (L1) of a language learner based on his/her writing in the second language (L2). Identifying the native language is based on the hypothesis that the L1 of a learner impacts his/her L2 writing due to the language transfer effect. NLI can be used for a variety of purposes, including marketing, security, and educational applications. From the machine-learning perspective, the NLI task is viewed as a multi-class, single-label classification problem, in which automatic methods have to assign class labels (L1s) to objects (texts).

Recent trends in NLI include cross-genre and cross-corpus NLI scenarios (Malmasi and Dras, 2015a), as well as identifying the L1 based on writings in other non-English L2s and cross-lingual NLI research (Malmasi and Dras, 2015b). However, following the practice of the first NLI shared task (Tetreault et al., 2013), this year's task focuses on L2 English data (Malmasi et al., 2017). This can be related to the use of English as *lingua franca* on the Internet and academia, when NLI methods are particularly useful for languages with a large number of foreign speakers. Moreover, following the 2016 Computational Paralinguistics Challenge (Schuller et al., 2016) and the VarDial workshop (Malmasi et al., 2016), this year's competition covers an NLI task based on the spoken response. Overall, this year's task consists of three tracks: NLI on the essay only, NLI on the spoken response only, and NLI on both essay and spoken response. In this paper, we describe the CIC-FBK approach to the essay-only track.

Previous works on identifying the native language from texts explored a large variety of features, including lexical and part-of-speech (POS) features (Koppel et al., 2005a), character n-grams (Ionescu et al., 2014), spelling errors (Koppel et al., 2005b), and syntactic features (Wong and Dras, 2011). Following previous research on the NLI task, we incorporate commonly used word n-grams, lemma n-grams, POS n-grams, and function words. In order to capture the L1 influences at the character level, we use recently introduced character n-grams from misspelled words (Chen et al., 2017), as well as 10 categories of character n-gram features proposed by Sapkota et al. (2015). We also include syntactic features by extracting syntactic dependency-based n-grams of words and of syntactic relation tags (Sidorov et al., 2014) using the algorithm designed by Posadas-Durán et al. (2014, 2017). We describe the features used by the CIC-FBK system in more detail in subsection 3.1.

Our system achieved 0.8808 macro-averaged F1-score and 0.8809 accuracy in the essay-only

track and shared the 1st rank in the NLI Shared Task 2017 scoring, obtaining the 2nd absolute score with the difference of 0.0010 F1-score and 0.0009 accuracy with the 1st place.

2 Data

The dataset used in the NLI Shared Task 2017 is composed of English essays written by non-native learners in a standardized assessment of English proficiency for academic purposes. The corpus consists of 13,200 essays (1,000 essays per L1 for training, 100 for development, and 100 for testing). The essays are sampled from 8 prompts, and score levels (low/medium/high) are provided for each essay. The training, development, and test sets are balanced in terms of the number of essays per L1 group. The 11 L1s covered by the corpus are: Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR). The detailed description of the corpus and its statistics can be found in Malmasi et al. (2017).

3 Methodology

Our system incorporates a wide range of features, i.e., word, lemma, and POS n-grams, spelling error character n-grams, typed character n-grams, and syntactic n-grams. We used the tokenized version of essays provided by the organizers. For the evaluation of our approach, we merged the training and development sets, and conducted experiments under 10-fold cross-validation. System performance was measured in terms of both classification accuracy and F1 (macro) score. The former was used as evaluation metric in the majority of previous works on NLI, whilst the later is the official evaluation metric in the NLI Shared Task 2017.

3.1 Features

3.1.1 Word, lemma, and POS *n*-grams

Word and lemma features represent the lexical choice of a writer, while part-of-speech (POS) features capture the morpho-syntactic patterns in a text. Following previous works on the NLI task (Jarvis et al., 2013; Malmasi and Dras, 2017), we use word, lemma, and POS n-grams with n ranging from 1 to 3. We include punctuation marks and split n-grams by a full stop. We lowercase word and lemma n-grams and replace each digit by the same symbol (e.g., 12,345 → 00,000), as proposed in Markov et al. (2017), to capture the format (e.g., 00.000 vs. 00,000), which reflects stylistic choice of a learner and not the value of a number that does not carry stylistic information. Lemmas and POS tags were obtained using the TreeTagger software package (Schmid, 1995).

3.1.2 Function words

Function words are the most common words in a language (e.g., articles, determiners, conjunctions). They are considered one of the most important stylometric features (Kestemont, 2014). Function words can be seen as indicators of the grammatical relations between other words. We use a set of 318 English function words from the scikit-learn package (Pedregosa et al., 2011). Other examined function word lists obtained from the Natural Language Toolkit[1] (127 function words) and the Onix Text Retrieval Toolkit[2] (429 function words), as well as function word skip-grams (Guthrie et al., 2006) did not lead to an improvement in accuracy.

3.1.3 Spelling error character *n*-grams

Spelling errors have been used as features for NLI since Koppel et al. (2005b). They are considered a strong indicator of an author's L1, since they reflect L1 influences, such as sound-to-character mappings in L1. Recently, Chen et al. (2017) introduced the use of character n-grams from misspelled words. The authors showed that adding spelling error character n-grams to other commonly used features (word and lemma n-grams) improves NLI classification accuracy. We extract 39,512 unique misspelled words from the training and development sets using the *spell* shell command. Then we build character n-grams (n = 4) from the extracted misspelled words. Other examined size of spelling error character n-grams (n = 1, 2, 3, and 5), as well as their combinations did not lead to an improvement in system performance.

3.1.4 Typed character *n*-grams

Character level features are sensitive to both the content and the form of a text and able to cap-

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

[2] http://www.lextek.com/manuals/onix/functionwords1.html

ture lexical and syntactic information, punctuation and capitalization information related with the authors' style (Stamatatos, 2013). The effectiveness of character n-gram features for representing the stylistic properties of a text has been demonstrated in previous NLI studies (Ionescu et al., 2014; Chen et al., 2017). Their effectiveness in NLI is hypothesized to be a result of phoneme transfer from the learner's L1, and by their ability to capture orthographic conventions of a language (Tsur and Rappoport, 2007).

Sapkota et al. (2015) defined 10 different character n-gram categories based on affixes, words, and punctuation. In this approach, instances of the same n-gram may refer to different typed n-gram features. For example, in the phrase *less carelessness*, the two instances of the 4-gram *less* are assigned to different character n-gram categories. As an example, consider the following sample sentence:

(1) *Lisa said, "John should repair it tomorrow."*

The character n-grams (n = 4) for the sample sentence (1) for each of the categories proposed by Sapkota et al. (2015) are shown in Table 1. For clarity, spaces are represented by the underscore.

SC	Category	*N*-grams
affix	*prefix*	shou repa tomo
	suffix	ould pair rrow
	space-prefix	_sai _sho _rep _it_ _tom
	space-suffix	isa_ ohn_ uld_ air_
word	*whole-word*	Lisa said John
	mid-word	houl epai omor morr orro
	multi-word *	sa_s hn_s ld_r ir_i it_t
punct	*beg-punct*	"Joh
	mid-punct **	_,_ _"_ _._ _"_
	end-punct	aid, row.

* If the previous word is more than one character long, two characters are considered; otherwise, only one character is considered.
** We use the tokenized version of essays and set the size of n-grams to 3 for this category. For other categories of typed character n-grams, the size is set to 4.

Table 1: Typed character 4-grams per category for the sample sentence (1) after applying the algorithm proposed by Sapkota et al. (2015).

Typed character n-grams have shown to be predictive features for other classification tasks, such as authorship attribution (Sapkota et al., 2015), author profiling (Markov et al., 2016), and discriminating between similar languages (Gómez-Adorno et al., 2017). In our experiments, typed character n-grams (n = 4) outperformed traditional character n-grams of the same size in most system configurations. In addition, we compared the performance of typed and traditional character n-grams on the 7-way ICLEv2 corpus (Granger et al., 2009), following the corpus splitting as described in Ionescu et al. (2014). In this experiment, typed character n-grams proved to be more indicative than traditional character n-grams when used in combination with features described in this paper.

3.1.5 Syntactic *n*-grams

Syntactic features, including production rules (Wong and Dras, 2011) and Tree Substitution Grammars (TSGs) (Swanson and Charniak, 2012), have been previously explored for NLI. Tetreault et al. (2012) experimented with the Stanford parser (de Marneffe et al., 2006) dependency features and concluded that they are strong indicators of structural differences in L2 writing. We exploit the Stanford dependencies to build syntactic n-gram features by using the algorithm designed and made available by Posadas-Durán et al. (2014, 2017).[3] Consider the following sample sentence:

(2) *I remember this great experience.*

The dependencies generated by the Standard parser for the the sample sentence (2) are the following:

```
root(ROOT, remember),
nsubj(remember, I),
dobj(remember, experience),
det(experience, this),
amod(experience, great).
```

These dependencies, including backoff transformation based on POS, were used as features for NLI in Tetreault et al. (2012). According to the metalanguage proposed in Sidorov (2013a), the syntactic 2-grams of words are the following:

```
remember[I],
remember[experience],
experience[this],
experience[great];
```

when the syntactic 3-grams of words are:

```
remember[I,experience],
remember[experience[this]],
remember[experience[great]],
```

[3]The Python implementation of the algorithm is available on http://www.cic.ipn.mx/ sidorov/MultiSNgrams_3.py

`experience[this,great];`
the syntactic 2-grams of syntactic relation tags are:
`root[nsubj],`
`root[dobj],`
`dobj[det],`
`dobj[amod];`
the syntactic 3-grams of syntactic relation tags are:
`root[nsubj,dobj],`
`root[dobj[det]],`
`root[dobj[amod]],`
`dobj[det,amod].`

Here, the head element is on the left of a square parenthesis and inside there are the dependent elements; the elements separated by a coma refer to non-continuous syntactic n-grams, that is, the elements are at the same level in a syntactic tree.

Syntactic n-grams can be used in any task where traditional n-grams are applied. They allow to introduce syntactic information into machine-learning methods (obviously, at cost of previous syntactic parsing). Syntactic n-grams outperformed traditional n-grams in the task of authorship attribution (Sidorov et al., 2014) and were applied in tasks related with L2, for example, automatic English as L2 grammar correction (Sidorov, 2013b). In our system, we use only continuous syntactic n-grams of words and of syntactic relation tags with n ranging from 2 to 3. The inclusion of non-continuous syntactic n-grams improved 10-fold cross-validation accuracy; however, did not perform well on the test set.

3.2 Frequency threshold

The fine-tuning of feature set size has proved to be a useful strategy for NLI (Jarvis et al., 2013) and other NLP tasks (Stamatatos, 2013; Markov et al., 2017). In our approach, we selected the frequency threshold value that provided the highest 10-fold cross-validation result. We consider only those features that occur in at least two documents in the training corpus and that occur at least 4 times in the entire training corpus. This frequency threshold improves 10-fold cross-validation accuracy by about 1%, compared to the configuration when all the features are considered, and reduces the size of the feature set by approximately 90% of the original. The final size of our feature set is 726,494.

3.3 Weighting scheme

We use log-entropy weighting scheme, which showed good results in previous studies on NLI (Jarvis et al., 2013; Chen et al., 2017). Log-entropy weighting scheme consists of local weighting (denoted as $L_{log}(i,j)$) and global weighting (denoted as $G_{ent}(i)$). The local weighting is calculated by taking the logarithm value of adding-one smoothed term frequency:

$$L_{log}(i,j) = \log(frequency(i,j)+1), \quad (1)$$

where $frequency(i,j)$ is the frequency of term i with regard to document j. The global entropy weighting is calculated by the following formula:

$$G_{ent}(i) = 1 + \frac{\sum_{j=1}^{J} p_{ij} \log p_{ij}}{\log(J+1)}, \quad (2)$$

where J is the total number of documents in the corpus. $\sum_{j=1}^{J} p_{ij} \log p_{ij}$ is the additive inverse of entropy of the conditional distribution given i and

$$p_{ij} = \frac{frequency(i,j)}{\sum_{j} frequency(i,j)}. \quad (3)$$

The final weighting W is calculated as follows:

$$W = L_{log}(i,j) \times G_{ent}(i). \quad (4)$$

Other examined feature representations, i.e., binary feature representation, tf, $tf\text{-}idf$, and normalized feature representation did not enhance system performance. Using log-entropy weighting scheme outperforms $tf\text{-}idf$, the second best scheme in our experiments, by 2.6% in 10-fold cross-validation accuracy.

3.4 Classifier

Support Vector Machines (SVM) is considered among the best performing classification algorithms for text categorization tasks; moreover, it was the classifier of choice for the majority of the teams in the previous edition of the NLI shared task. We use the liblinear scikit-learn (Pedregosa et al., 2011) implementation of SVM with 'ovr' multi-class strategy. We set the penalty hyper-parameter C to 100 based on our model selection result.

4 Results

We present the results of our experiments in two phases. First, we show the performance of each type of features in isolation under 10-fold cross validation on the merged training and development sets. Then, we compare the performance

obtained on the test set with other participating teams. We present the 10-fold cross-validation results in terms of classification accuracy. For each experiment, the difference between accuracy and F1 (macro) score was less than 0.0003.

The individual performance of the features used in our system with the configurations described in the previous section, as well as the number of features (N) of each type are shown in Table 2.

Features	Accuracy	N
words n-grams (n = 1–3)	0.8463	230,714
lemma n-grams (n = 1–3)	0.8454	228,229
POS n-grams (n = 1–3)	0.4930	14,510
function words	0.5004	302
spelling error character 4-grams	0.3779	12,322
typed character 4-grams	0.7779	35,480
syntactic n-grams of words (n = 2–3)	0.7064	148,728
syntactic n-grams of SR tags (n = 2–3)	0.2361	5,344
combination of the above	0.8640	726,494

Table 2: 10-fold cross-validation accuracy of each feature type individually on the merged training and development sets.

In line with the previous works on the NLI task (Tetreault et al., 2013; Jarvis et al., 2013; Chen et al., 2017), in our configurations word and lemma n-grams are the most predictive features. They showed 0.8463 and 0.8454 10-fold cross-validation accuracy, respectively, when evaluated in isolation. Typed character n-grams also performed well with a much smaller feature size, achieving 0.7779 accuracy. Syntactic n-grams of syntactic relation tags showed the lower accuracy when evaluated in isolation; however, when used in combination with other features, they improve 10-fold cross-validation accuracy by 0.2%. The combination of all the features showed 0.8640 10-fold cross-validation accuracy on the merged training and development sets.

The NLI Shared Task 2017 organizers reported several 1st ranked teams based on McNemar's statistical significance test with an alpha value of 0.05. The official results for the essay-only track in terms of F1 (macro) score and classification accuracy for the 1st ranked teams, as well as the baseline results are shown in Table 3.

The CIC-FBK best run differs 0.0009 in terms of classification accuracy from the highest result achieved by the ItaliaNLP Lab system, which corresponds to one correctly predicted label. All the 17 participating teams in the NLI Shared Task 2017 achieved higher level of F1 (macro) score than the official baseline of 0.7104.

Rank	Team	F1 (macro)	Accuracy
1	ItaliaNLP Lab	0.8818	0.8818
1	**CIC-FBK**	**0.8808**	**0.8809**
1	Groningen	0.8756	0.8755
1	NRC	0.8740	0.8736
1	taraka_rama	0.8716	0.8718
1	UnibucKernel	0.8695	0.8691
1	WLZ	0.8654	0.8655
-	Official baseline	0.7104	0.7109
-	Random baseline	0.0909	0.0909

Table 3: Results for the essay-only track for the 1st ranked teams. The results for our team are highlighted in bold typeface.

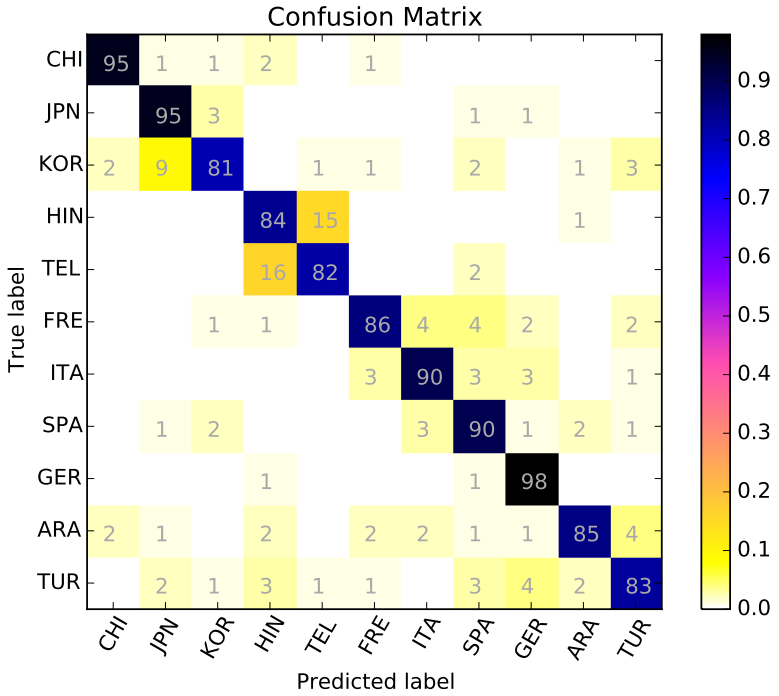
The CIC-FBK system showed 0.8639 F1 (macro) score and 0.8640 accuracy under 10-fold cross-validation on the merged training and development sets. Our other runs in the NLI Shared Task 2017 included small modifications in system configurations, such as variations in frequency threshold values and different strategy for dealing with digits (e.g., 12,345 $\rightarrow$ 0,0). However, since these modifications showed only marginal accuracy variations and did not improve system performance on the test set, the results for these runs are omitted in this paper.

The confusion matrix for our best run is shown in Figure 1. The highest level of confusion is between Hindi and Telugu classes. Korean and Japanese is another problematic language pair, in which Korean native speakers are often classified as Japanese. The highest accuracy of 0.9800 was achieved for German native speakers. These results are in line with the ones reported in the previous edition of the NLI share task (Tetreault et al., 2013), where the teams achieved low levels of accuracy for the Hindi/Telugu (none of the systems was able to reach 0.8000 accuracy for Hidni) and the Korean/Japanese pairs. In future work, we intend to tackle these two language pairs in isolation in order to improve the overall system performance.

5 Conclusions

We presented the description of the best submission of the CIC-FBK team to the NLI Shared Task 2017. Our approach combines features commonly used in the NLI task with recently introduced spelling error character n-grams, as well as with typed character n-grams, and syntactic n-grams of words and of syntactic relation tags.

Figure 1: Confusion matrix for the best CIC-FBK run.



Typed character n-grams and syntactic n-grams are new types of features that are introduced in the NLI task for the first time. It was found during the preliminary experiments on the training and development sets that these features improve the classification accuracy when used in combination with other types of features, such as word n-grams, lemma n-grams, part-of-speech n-grams, spelling error character n-grams, and function words. The CIC-FBK system achieved 0.8808 F1 (macro) score and 0.8809 accuracy and shared the 1st rank in the competition.

Acknowledgements

This work was partially supported by the Mexican Government (CONACYT project 240844, SNI, COFAA-IPN, and SIP-IPN 20171813, 20171344, 20172008). We would like to thank Vivi Nastase for the discussion about spelling errors.

References

Lingzhen Chen, Carlo Strapparava, and Vivi Nastase. 2017. Improving native language identification by using spelling errors. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (ACL 2017)*. ACL, Vancouver, Canada.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. ELRA, Genoa, Italy, pages 449–454.

Helena Gómez-Adorno, Ilia Markov, Jorge Baptista, Grigori Sidorov, and David Pinto. 2017. Discriminating between similar languages using a combination of typed and untyped character n-grams and words. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*. ACL, Valencia, Spain, pages 137–145.

Sylviane Granger, Estelle Dagneaux, and Fanny Meunier. 2009. *The International Corpus of Learner English. Version 2*. Presses Universitaires de Louvain.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A close look at skip-gram modelling. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. ELRA, Genoa, Italy, pages 1222–1225.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. ACL, Doha, Qatar, pages 1363–1373.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth*

Workshop on Innovative Use of NLP for Building Educational Applications. ACL, Atlanta, Georgia, USA, pages 111–118.

Mike Kestemont. 2014. Function words in authorship attribution. From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (EACL 2014)*. ACL, Gothenburg, Sweden, pages 59–66.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005a. Automatically determining an anonymous author's native language. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2005)*. Springer, Atlanta, Georgia, USA, pages 209–217.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005b. Determining an author's native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD 2005)*. ACM, Chicago, Illinois, USA, pages 624–628.

Shervin Malmasi and Mark Dras. 2015a. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL-HLT 2015)*. ACL, Denver, Colorado, USA, pages 1403–1409.

Shervin Malmasi and Mark Dras. 2015b. Multilingual native language identification. *Natural Language Engineering* 1:1–56.

Shervin Malmasi and Mark Dras. 2017. Native language identification using stacked generalization. *arXiv preprint arXiv:1703.06541* .

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*. ACL, Copenhagen, Denmark.

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 VarDial Workshop*. Osaka, Japan.

Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander Gelbukh. 2016. Adapting cross-genre author profiling to language and corpus. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*. CLEF and CEUR-WS.org, Évora, Portugal, volume 1609, pages 947–955.

Ilia Markov, Efstathios Stamatatos, and Grigori Sidorov. 2017. Improving cross-topic authorship attribution: The role of pre-processing. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*. Springer, Budapest, Hungary, in press.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Juan-Pablo Posadas-Durán, Grigori Sidorov, and Ildar Batyrshin. 2014. Complete syntactic n-grams as style markers for authorship attribution. In *Proceedings of the 13th Mexican International Conference on Artificial Intelligence (MICAI 2014)*. Springer, Tuxtla Gutiérrez, Mexico, pages 9–17.

Juan-Pablo Posadas-Durán, Grigori Sidorov, Helena Gómez-Adorno, Ildar Batyrshin, Elibeth Mirasol-Mélendez, Gabriela Posadas-Durán, and Liliana Chanona-Hernández. 2017. Algorithm for extraction of subtrees of a sentence dependency parse tree. *Acta Polytechnica Hungarica* in press.

Upendra Sapkota, Steven Bethard, Manuel Montes-y-Gómez, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL-HLT 2015)*. ACL, Denver, Colorado, USA, pages 93–102.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. ACL, Dublin, Ireland, pages 47–50.

Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Interspeech 2016*. pages 2001–2005.

Grigori Sidorov. 2013a. *Non-linear Construction of N-grams in Computational Linguistics*. SMIA.

Grigori Sidorov. 2013b. Syntactic dependency based n-grams in rule based automatic English as second language grammar correction. *International Journal of Computational Linguistics and Applications* 4(2):169–188.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications* 41(3):653–860.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law & Policy* 21(2):427–439.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers – Volume 2 (ACL 2012)*. ACL, Jeju Island, Korea, pages 193–197.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*. ACL, Atlanta, GA, USA.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24st International Conference on Computational Linguistics (COLING 2012)*. The COLING 2012 Organizing Committee, Mumbai, India, pages 2585–2602.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA 2007)*. ACL, Prague, Czech Republic, pages 9–16.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*. ACL, Edinburgh, Scotland, UK, pages 1600–1610.