Sentences and Documents in Native Language Identification

Andrea Cimino, Felice Dell'Orletta, Dominique Brunato, Giulia Venturi
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it

Abstract

English. Starting from a wide set of linguistic features, we present the first in depth feature analysis in two different Native Language Identification (NLI) scenarios. We compare the results obtained in a traditional NLI document classification task and in a newly introduced sentence classification task, investigating the different role played by the considered features. Finally, we study the impact of a set of selected features extracted from the sentence classifier in document classification.

Italiano. *Partendo da un ampio insieme di caratteristiche linguistiche, presentiamo la prima analisi approfondita del ruolo delle caratteristiche linguistiche nel compito di identificazione della lingua nativa (NLI) in due differenti scenari. Confrontiamo i risultati ottenuti nel tradizionale task di NLI ed in un nuovo compito di classificazione di frasi, studiando il ruolo differente che svolgono le caratteristiche considerate. Infine, studiamo l'impatto di un insieme di caratteristiche estratte dal classificatore di frasi nel task di classificazione di documenti.*

1 Introduction

Native Language Identification (NLI) is the research topic aimed at identifying the native language (L1) of a speaker or a writer based on his/her language production in a non-native language (L2). The leading assumption of NLI research is that speakers with the same L1 exhibit similar linguistic patterns in their L2 productions which can be viewed as traces of the L1 interference phenomena. Thanks to the availability of large-scale benchmark corpora, such as the TOEFL11 corpus (Blanchard et al., 2013), NLI has been recently gaining attention also in the NLP community where it is mainly addressed as a multi-class supervised classification task. This is the approach followed by the more recent systems taking part to the last editions of the NLI Shared Tasks held in 2013 (Tetreault et al., 2013) and 2017 (Malmasi et al., 2017). Typically, these systems exploit a variety of features encoding the linguistic structure of L2 text in terms of e.g. n-grams of characters, words, POS tags, syntactic constructions. Such features are used as input for machine learning algorithms, mostly based on traditional Support Vector Machine (SVM) models. In addition, rather than using the output of a single classifier, the most effective approach relies on ensemble methods based on multiple classifiers (Malmasi and Dras, 2017).

In this paper we want to further contribute to NLI research by focusing the attention on the role played by different types of linguistic features in predicting the native language of L2 writers. Starting from the approach devised by (Cimino and Dell'Orletta, 2017), which obtained the first position in the essay track of the 2017 NLI Shared Task, we carry out a systematic feature selection analysis to identify which features are more effective to capture traces of the native language in L2 writings at sentence and document level.

Our Contributions *(i)* We introduce for the first time a NLI sentence classification scenario, reporting the classification results; *(ii)* We study which features among a wide set of features contribute more to the sentence and to the document classification task; *(iii)* We investigate the contribution of features extracted from the sentence classifier in a stacked sentence-document system.

2 The Classifier and Features

In this work, we built a classifier based on SVM using LIBLINEAR (Rong-En et al., 2008) as ma-

Raw text features
TOEFL11 essay prompt*
Text length (n. of tokens)
Word length (avg. n. of characters)
Average sentence length and standard deviation*
Character n-grams (up to 8)
Word n-grams (up to 4)
Functional word n-grams (up to 3)
Lemma n-grams (up to 4)
Lexical features
Type/token ratio of the first 100, 200, 300, 400 tokens*
Etymological WordNet features (De Melo, 2014)
etymological n-grams (up-to 4)
Morpho–syntactic features
Coarse Part-Of-Speech n-grams (up to 4)
Coarse Part-Of-Speech+Lemma of the following token n-grams (up to 4)
Syntactic features
Dependency type n-grams (sentence linear order) (up to 4)
Dependency type n-grams (syntactic hierarchical order) (up to 4)
Dependency subtrees (dependency of a word + the dependencies to its siblings in the sentence linear order)

Table 1: Features used for document and sentence classification (* only for document).

chine learning library. The set of documents described in Section 3 was automatically POS tagged by the part–of–speech tagger described in (Cimino and Dell'Orletta, 2016) and dependency–parsed by DeSR (Attardi et al., 2009). A wide set of features was considered in the classification of both sentences and documents. As shown in Table 1, they span across multiple levels of linguistic analysis. These features and the classifier were chosen since they were used by the 1st ranked classification system (Cimino and Dell'Orletta, 2017) in the 2017 NLI shared task.

3 Experiments and Results

We carried out two experiments devoted to classify L2 documents and sentences. The training and development set distributed in the 2017 NLI shared task, i.e. the TOEFL11 corpus (Blanchard et al., 2013), was used as training data. It includes 12,100 documents, corresponding to a total of 198,334 sentences. The experiments were tested on the 2017 test set, including 1,100 documents (18,261 sentences).

The obtained macro average F1-scores were: 0.8747 in the document classification task and 0.4035 in the sentence one. As it was expected, the identification of the L1 of the sentences turned out as a more complex task than L1 document classification. Both document and sentence classification

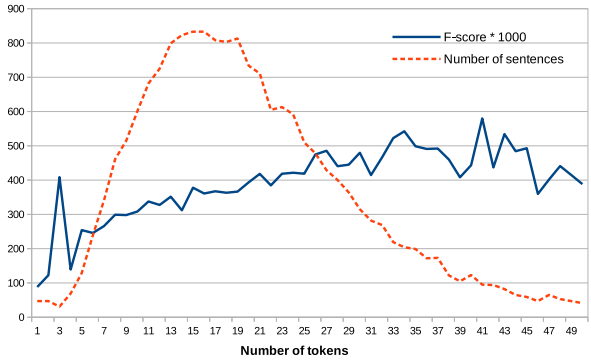


Figure 1: Sentence classification performance across bins of sentences of the same lengths.

are influenced by the number of words but with a different impact. Figure 1 shows that the average performance on sentences is reached for sentences ∼21–token long, which corresponds to the average sentence length for this dataset. As the sentence length increases, the accuracy increases as well. Due to the smaller amount of linguistic evidence, the classification of short sentences is a more complex task. The performance of document classification is more stable: the best f–score is already reached for documents of ∼200–tokens, which corresponds to a very short document compared to the average size of TOEFL11 documents (330 tokens).

Figures 2(a) and 2(b) report the confusion matrices of the two experiments[1]. As it can be seen, both for sentences and documents the best classification performance is obtained for German, Japanese and Chinese, even though with some differences in the relative ranking positions, e.g. German is the top ranked one in the sentence classification scenario and the 2nd ranked one in the document classification one, while Japanese is the best classified L1 in the document experiment and the 4th ranked one in the sentence classification scenario. Conversely, we observe differences with respect to the worst recognized L1s, which are Turkish, Hindi and Korean in the document classification task and Arabic, Spanish and Turkish in the sentence classification one. The two confusion matrices also reveal a peculiar error distribution trend: the confusion matrix of the sentence classification model is much more sparse than the

[1]Since the number of documents and sentences in the two experiments is different, in order to make comparable the values of the two confusion matrices, the sentence classification values were normalized to 100.

document classification one. This means that for each considered L1, the errors made by the sentence classifier are quite similarly distributed over all possible L1s; instead, errors in the document classification scenario are much more prototypical, i.e. the wrong predicted label is assigned to only one or two L1 candidates, which change according to the specific L1. This is shown e.g. by languages belonging to same language family such as Japanese and Korean which belong to the same Altaic family. Specifically, in the document classification scenario Korean is mainly confused with Japanese (10% of errors). This trend holds also in the sentence classification experiment where 17.8% of errors were due to the confusion of Korean with Japanese and vice versa (18.2% of errors). Interestingly enough, the most prototypical errors were also made when contact languages were concerned. This is for example the case of Hindi and Telugu: Hindi documents were mainly confused with Telugu ones (16% of errors) and Telugu documents with Hindi ones (13% of errors). Similarly, in the sentence classification scenario, Hindi sentences were wrongly classified as Telugu sentences in about 20% of cases and vice versa. As previously shown by Cimino et al. (2013), even if these two languages do not belong to the same family, such classification errors might originate from a similar linguistic profile due to language contact phenomena: for instance, both Hindi and Telugu L1 essays are characterized by sentences and words of similar length, or they share similar syntactic structures such e.g. parse trees of similar depth and embedded complement chains governed by a nominal head of similar length.

The behavior of the two classifiers may suggest that *i)* some features could play a different role in the classification of sentences with respect to documents and *ii)* the document classifier can be improved using features extracted from the output of a sentence classifier in a stacked configuration. To investigate these hypotheses, we carried out an extensive feature selection analysis to study the role of the features in the two classification scenarios.

3.1 Feature Selection

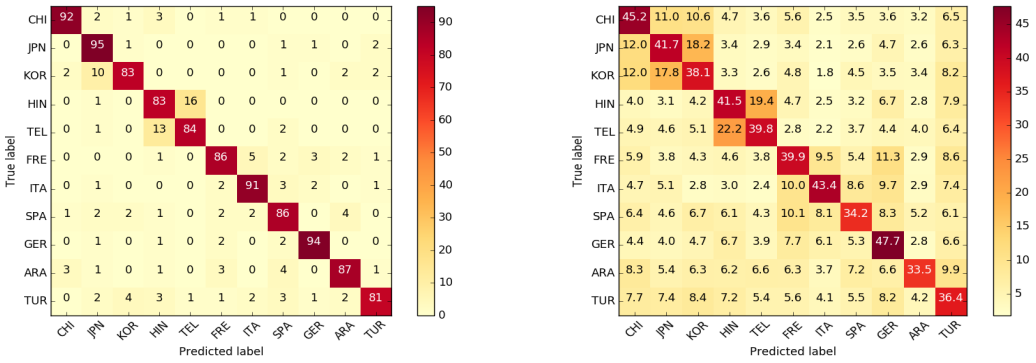
In the first step of the feature selection process, we extracted all the features from the training set and pruned those occurring less than 4 times, obtaining $\sim$ 4,000,000 distinct features both for document and sentence classification. In the second step, we ranked the extracted features through the *Recursive Feature Elimination* (RFE) algorithm implemented in the Scikit-learn library (Pedregosa et al., 2011) using Linear SVM as estimator algorithm. We dropped 1% of features in each iteration. At the end of this step we selected the top ranked features corresponding to $\sim$ 40,000 features both for the sentence and document tasks. These features were further re-ranked using the RFE algorithm (dropping 100 features at each iteration) to allow a more fine grained analysis.

Figure 3(a) compares the percentage of different types of features used in the classification of documents and sentences. As it can be noted, the document classifier uses more words n-grams, especially n-grams characters. Instead, morpho–syntactic and syntactic features are more effective for sentence classification, and the n-grams of lemmas even more than 4 times. Figures 3(b), 3(c) and 3(d) show the variation of relevance of the 40k raw text, morpho-syntactic and syntactic features grouped in bins of 100 features. The lines in the charts correspond to the differences between document and sentence in terms of percentage of a single type of feature in the bin with respect to its total distribution in the whole 40k selected features[2]. Negative values mean that this distribution in the bin is higher for sentence classification.

Among the raw text features (Figure 3(b)), n-grams of words occur more in the 1st bins of document classification, while n-grams of characters and lemma are more relevant in the 1st bins of sentence classification. The n-grams of coarse parts–of-speech are equally distributed in the two rankings, instead both the n-grams of coarse parts–of-speech followed by a lemma and the n-grams of functional words occur more in the 1st bins of sentence classification (Figure 3(c)). This confirms the key role played by lemma in sentence classification.

For what concerns syntactic information (Figure 3(d)), the features that properly capture sentence structure (dependency subtree and the hierarchical syntactic dependencies) are all contained in the first bins of document classification even if their total distribution is lower than in the sentence. This shows that syntactic information is very relevant also when longer texts are classified and that this kind of information is not captured by

[2]Spline interpolation applied for readability purpose.



(a) (b)

Figure 2: Confusion matrix of document (a) and sentence classification (b).

n-grams of words. Feature types with low number of instances are not reported in these charts. Among these, *etymological n-grams* appears in the first bins both for sentence and document, confirming the relevance of the etymological information already proven for NLI document classification (Nastase and Strapparava, 2017). For sentence classification, it is also relevant *sentence length* and *word length*. Instead, for document, *type/token ratio* plays a very important role. Interestingly, the *average sentence length* does not appear in the 40k features; we found instead the *sentence length standard deviation*, showing that what counts more is the variation in length rather than the average value. Even though not contained in the first bins, also *word and document lengths* and the *TOEFL11 essay prompt* are in the top 40k features.

4 Stacked Classifier

The different role of the features in the two L1 classification tasks suggests that we may improve the traditional NLI document classification by combining sentence and document classifiers. We thus evaluated and extended the stacked sentence-document architecture proposed by (Cimino and Dell'Orletta, 2017). In addition to the linguistic features, they proposed a stacked system using the L1 predictions of a pre–trained sentence classifier to train a document classifier. Thus we run several experiments on the NLI Shared Task 2017 test set to assess i) the importance of the sentence classifier in a stacked sentence-document architecture and ii) which features extracted from the predictions of the L1 sentence classifier maximize the accuracy of the stacked system. The sentence classifier assigned a confidence score for each L1 to each sentence of the documents. Based on the confidence score, we defined the following features: for each L1 i) the mean sentence confidence (*avg*), ii) the standard deviation of confidences (*stddev*), iii) the product of the confidences (*prod*), iv) the top–3 highest and lowest confidence values (*top–3 max-min*). The last two features were introduced to mitigate the effect of spike values that may be introduced by considering the max-min L1 confidences used in (Cimino and Dell'Orletta, 2017). The first row of Table 2 reports the result obtained by (Cimino and Dell'Orletta, 2017) by the stacked classifier on the same test set. The second row reports the results of our document system which does not use features extracted from the sentence classifier. The third row reports the result of a classifier that uses only the features extracted from the predictions of the L1 sentence classifier. The following rows report the contribution of each sentence classifier feature in the stacked architecture showing an improvement (with the exception of the product) with respect to the base classifier. The top–3 highest and lowest confidence values are the most helpful features in a stacked architecture. The best result is obtained when using all the sentence classifier features in the base classifier, which is the state-of-the-art on the 2017 NLI test set.

5 Conclusions

We introduced a new NLI scenario focused on sentence classification. Compared to document classification we obtained different results in terms of accuracy and distribution of errors across the L1s. We showed the different role played by a

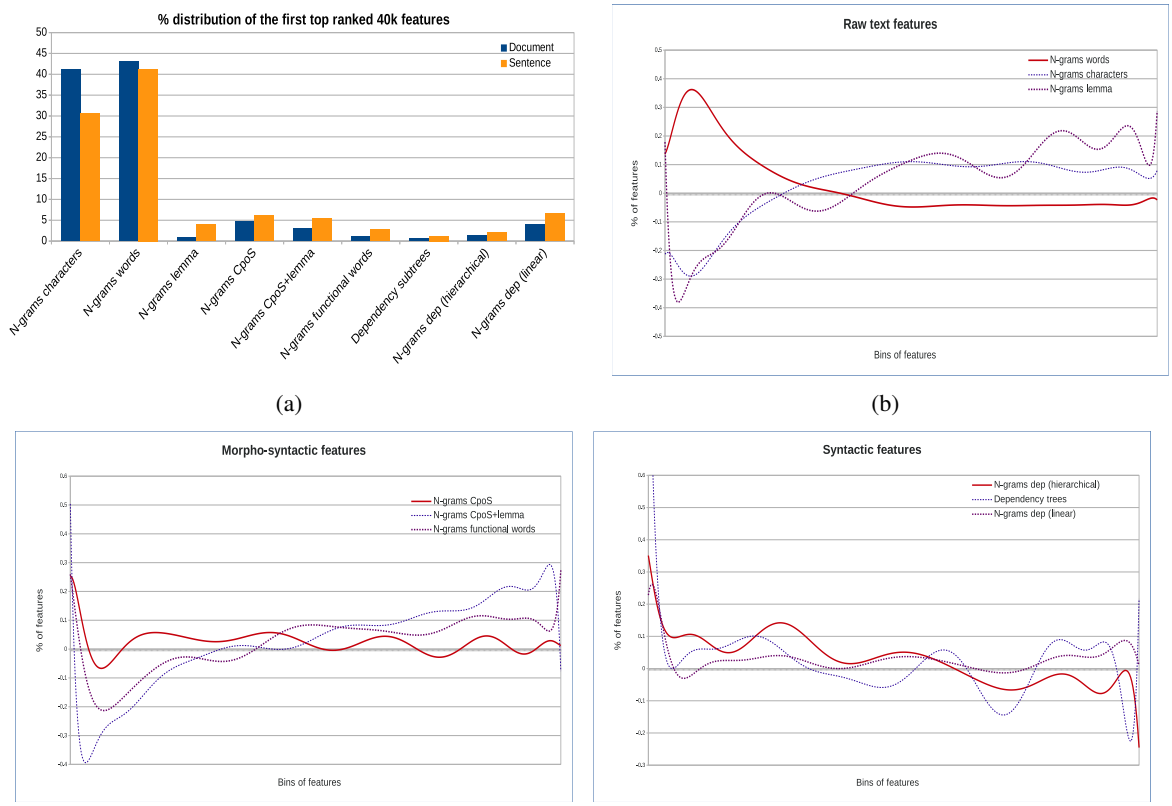


Figure 3: Distribution of the first top ranked 40k features in the document and sentence classification.

Model	F1-Score
Cimino and Dell'Orletta (2017)	0.8818
Base classifier	0.8747
Sentence features	0.8363
Base class. + avg	0.8773
Base class. + stddev	0.8773
Base class. + prod	0.8747
Base class. + top–3 max-min	0.8800
Base class. + all sentence feat.	**0.8828**

Table 2: Results of the stacked system.

wide set of linguistic features in the two NLI scenarios. These differences may justify the performance boost we achieved with a stacked sentence-document system. We also assessed which features extracted from the sentence classifier maximizes NLI document classification.

6 Aknowledgments

The work presented in this paper was partially supported by the 2–year project (2018-2020) SchoolChain – Soluzioni innovative per la creazione, la certificazione, il riuso e la condivisione di unità didattiche digitali all'interno del sistema Scuola, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, and Joseph Turian. 2009. *Accurate dependency parsing with a stacked multilayer perceptron.* In Proceedings of the 2nd Workshop of Evalita 2009. December, Reggio Emilia, Italy.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English.* Technical report, Educational Testing Service.

Andrea Cimino and Felice Dell'Orletta. 2016. *Building the state-of-the-art in POS tagging of italian tweets.* In Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA), December 5-7.

Andrea Cimino and Felice Dell'Orletta. 2017. *Stacked sentence–document classifier approach for improving native language identification.* In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EMNLP 2017, September 8, pages 430-437.

Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. *Linguistic Profiling based on General-purpose Features*

and Native Language Identification. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2013, June 13, 2013, Atlanta, Georgia, USA, pages 207–215.

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. BEA@EMNLP 2017, September 8.

Gerard de Melo. 2014. *Etymological Wordnet: Tracing the history of words.* In Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014), Paris, France. ELRA.

Vivi Nastase and Carlo Strapparava. 2017. *Word etymology as native language interference.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2702–2707.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12:28252830.

Fan Rong-En, Chang Kai-Wei, Hsieh Cho-Jui, Wang Xiang-Rui, and Lin Chih-Jen. 2008. LIBLINEAR: A library for large linear classification. 2008. *LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, 9:18711874.

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 Innovative Use of NLP for Building Educational Applications, Atlanta, GA, USA. Association for Computational Linguistics.