# LIMSI's Participation in the 2013 Shared Task on Native Language Identification

**Thomas Lavergne, Gabriel Illouz, Aurélien Max**
LIMSI-CNRS
Univ. Paris Sud
Orsay, France
{firstname.lastname}@limsi.fr

**Ryo Nagata**
LIMSI-CNRS & Konan University
8-9-1 Okamoto
Kobe 658-0072 Japan
rnagata@konan-u.ac.jp

## Abstract

This paper describes LIMSI's participation to the first shared task on Native Language Identification. Our submission uses a Maximum Entropy classifier, using as features character and chunk $n$-grams, spelling and grammatical mistakes, and lexical preferences. Performance was slightly improved by using a two-step classifier to better distinguish otherwise easily confused native languages.

## 1 Introduction

This paper describes the submission from LIMSI to the 2013 shared task on Native Language Identification (Tetreault et al., 2013). The creation of this new challenge provided us with a dataset (12,100 TOEFL essays by learners of English of eleven native languages (Blanchard et al., 2013)) that was necessary to us to develop an initial framework for studying Native Language Identification in text. We expect that this challenge will draw conclusions that will provide the community with new insights into the impact of native language in foreign language writing. We believe that such a research domain is crucial, not only for improving our understanding of language learning and language production processes, but also for developing Natural Language Processing applications to support text improvement.

This article is organized as follows. We first describe in Section 2 our maximum entropy system used for the classification of a given text in English into the native languages of the shared task. We then introduce the various sets of features that we have included in our submission, comprising basic $n$-gram features (3.1) and features to capture spelling mistakes (3.2), grammatical mistakes (3.3), and lexical preference (3.4). We next report the performance of each of our sets of features (4.1) and our attempt to perform a two-step classification to reduce frequent misclassifications (4.2). We finally conclude with a short discussion (section 5).

## 2 A Maximum Entropy model

Our system is based on a classical maximum entropy model (Berger et al., 1996):

$$p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp(\theta^\top F(x, y))$$

where $F$ is a vector of feature functions, $\theta$ a vector of associated parameter values, and $Z_\theta(x)$ the partition function.

Given $N$ independent samples $(x^i, y^i)$, the model is trained by minimizing, with respect to $\theta$, the negative conditional log-likelihood of the observations:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p(y^i|x^i).$$

This term is complemented with an additional regularization term so as to avoid overfitting. In our case, an $\ell_1$ regularization is used, with the additional effect to produce a sparse model.

The model is trained with a gradient descent algorithm (L-BFGS) using the Wapiti toolkit (Lavergne et al., 2010). Convergence is determined either by error rate stability on an held-out dataset or when limits of numerical precision are reached.



*Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 260–265,
Atlanta, Georgia, June 13 2013. 

## 3 Features

Our submission makes use of basic features, including $n$-grams of characters and part-of-speech tags. We further experimented with several sets of features that will be described and compared in the following sections.

### 3.1 Basic features

We used $n$-grams of characters up to length 4 as features. In order to reduce the size of the feature space and the sparsity of these features, we used a hash kernel (Shi et al., 2009) of size $2^{16}$ with a hash family of size 4. This allowed us to significantly reduce the training time with no noticeable impact on the model's performance.

Our set of basic features also includes $n$-grams of part-of-speech (POS) tags and chunks up to length 3. Both were computed using an in-house CRF-based tagger trained on PennTreeBank (Marcus et al., 1993). The POS tags sequences were post-processed so that word tokens were used in lieu of their corresponding POS tags for the following: coordinating conjunctions, determiners, prepositions, modals, predeterminers, possessives, pronouns, and question adverbs (Nagata, 2013).

For instance, from this sentence excerpt:

```
[NP Some/DT people/NNS] [VP
might/MD think/VB] [SBAR that/IN]
[VP traveling/VBG] [PP in/IN]...
```

we extract $n$-grams from the pseudo POS-tag sequence:

```
Some NNS MD VB that VBG in...
```

and $n$-grams from the chunk sequence:

```
NP VP SBAR VP PP...
```

The length of chunks is encoded as separate features that correspond to mean length of each type of chunks. As shown in (Nagata, 2013), length of noun sequences is also informative and thus was encoded as a feature.

### 3.2 Capturing spelling mistakes

We added a set of features to capture information about spelling mistakes in the model, following the intuition that some spelling mistakes may be attributed to the influence of the writer's native language.

To extract these features, each document is processed using the `ispell`[1] spell checker. This results in a list of incorrectly written word forms and a set of potential corrections. For each word, the best correction is next selected using a set of rules, which were built manually after a careful study of the training dataset.

When a corrected word is found, the incorrect fragment of the word is isolated by striping from the original and corrected words common prefix and suffix, keeping only the inner-most substring difference. For example, given the following mistake and correction:

$$appartment \rightarrow apartment$$

this procedure generates the following feature:

$$pp \rightarrow p$$

Such a feature may for instance help to identify native languages (using latin scripts) where doubling of letters is frequent.

### 3.3 Capturing grammatical mistakes

Errors at the grammatical level are captured using the "language tool" toolkit (Milkowski, 2010), a rule-based grammar and style checker. Each rule firing in a document is mapped to an individual feature.

This triggers features such as BEEN_PART_AGREEMENT, corresponding to cases where the auxiliary *be* is not followed by a past participle, or EN_A_VS_AN, corresponding to confusions between the correct form the articles *a* and *an*.

### 3.4 Capturing lexical preferences

Learners of a foreign language may have some preference for lexical choice given some semantic content that they want to convey[2]. We made the following assumption: the lexical variant chosen for each word may correspond to the less ambiguous choice if mapping from the native language to English[3].

[1] `http://www.gnu.org/software/ispell/`

[2] We assumed that we should not expect thematic differences in the contents of the essays across original languages, as the prompts for the essays were evenly distributed.

[3] This assumption of course could not hold for advanced learners of English, who should make their lexical choices independently of their native language.

Thus, for each word in an English essay, if we knew a corresponding word (or *sense*) that a writer may have thought of in her native language, we would like to consider the most likely translation into English, according to some reliable probabilistic model of lexical translation into English, as the lexical choice most likely to be made by a learner of this native language.

As we obviously do not have access to the word in the native language of the writer, we approximate this information by searching for the word that maximizes the translation probability of translating back from the native language after translating from the original English word. This in fact corresponds to a widely used way of computing paraphrase probabilities from bilingual translation distributions (Bannard and Callison-Burch, 2005):

$$\hat{e}_l \approx \underset{e}{\operatorname{argmax}} \sum_f p_l(f|e).p_l(e|f)$$

where $f$ ranges over all possible translations of English word $e$ in a given native language $l$.

Preferably, we would like to obtain candidate translations into the native language in context, that is, by translating complete sentences and using *a posteriori* translation probabilities. We could not do this for a number of reasons, the main one being that we did not have the possibility of using or building Statistical Machine Translation systems for all the language pairs involving English and the native languages of the shared task. We therefore resorted to simply finding, for each English word, the most likely back-translation into English *via* a given native language. Using the Google Translation online Statistical Machine Translation service[4], which proposed translations from and to English and all the native languages of the shared task, a further approximation had to be made as, in practice, we were only able to access the most likely translations for words in isolation: we considered only the best translation of the original English word in the native language, and then kept its best back-translation into English. We here note some common intuitions with the use of roundtrip translation as a Machine Translation evaluation metrics (Rapp, 2009).

[4] http://translate.google.com

Table 1 provides various examples of back-translations for English adjectives obtained *via* each native language. The samples from the Table show that our procedure produces a significant number of non identical back-translations. They also illustrate some types of undesirable results obtained, which led us to only consider as features for our classifier the proportion of words in essays for which the above-defined back-translation yielded the same word, considering all possible native languages. We only considered content words, as out-of-context back-translation for function words would be too unreliable. Table 2 shows values for some documents of the training set. As can be seen, there are important differences across languages, some languages obtaining high scores on average (e.g. French and Japanese) and others obtaining low scores on average (e.g. Korean, Turkish). Furthermore, the highest score is only rarely obtained for the actual native language of each document, showing that keeping the most probable language according to this value alone would not allow to obtain a good classification performance.

## 4 Experiments

### 4.1 Results per set of features

For all our experiments reported here, we used the full training data provided using cross-validation to tune the regularization parameter. Our results are presented in the top part of Table 3. Using our complete set of features yields our best performance on accuracy, corresponding to a 0.75% absolute improvement over using our basic $n$-gram features only. No type of features allows a significant improvement over the $n$-gram features when added individually.

### 4.2 Two-step classification

Table 4 contains the confusion matrix for our system across languages. It clearly stands out that two language pairs were harder to distinguish: Hindi (hin) and Telugu (tel) on the one hand, and Korean (kor) and Japanese (jpn) on the other.

In order to improve the performance of our model, we performed a two-step classification focused on these difficult pairs. For this, we built additional classifiers for each difficult pairs. Both are built

| eng | abrupt | affirmative | amazing | ambiguous | anarchic | atrocious | attentive | awkward |
|---|---|---|---|---|---|---|---|---|
| ara | **sudden** | **positive** | amazing | **mysterious** | **messy** | **terrible** | **heedful** | **inappropriate** |
| chi | **sudden** | **sure** | amazing | ambiguous | anarchic | atrocious | **careful** | awkward |
| fre | **sudden** | affirmative | amazing | ambiguous | anarchic | atrocious | **careful** | awkward |
| ger | abrupt | affirmative | **incredible** | ambiguous | **anarchical** | **gruesome** | **attentively** | awkward |
| hin | **suddenly** | **positive** | amazing | **vague** | **chaotic** | **brutal** | **observant** | **clumsy** |
| ita | abrupt | affirmative | amazing | ambiguous | **anarchist** | atrocious | **careful** | **uncomfortable** |
| jap | **sudden** | **positive** | **surprising** | ambiguous | **anarchy** | **heinous** | **cautious** | awkward |
| kor | **fortuitous** | **positive** | amazing | ambiguous | anarchic | **severe** | **kind** | awkward |
| spa | abrupt | affirmative | **surprising** | ambiguous | anarchic | atrocious | attentive | **clumsy** |
| tel | abrupt | affirmative | amazing | ambiguous | anarchic | **formidable** | attentive | awkward |
| tur | **sudden** | **positive** | amazing | **uncertain** | anarchic | **brutal** | attentive | **strange** |

Table 1: Examples of back translations for English adjectives from the training set *via* each of the eleven native languages of the shared task. Back-translations that differ from the original word are indicated using a bold face.

| Doc id. | Native l. | ARA | CHI | FRE | GER | HIN | ITA | JPN | KOR | SPA | TEL | TUR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 976 | ARA | 0.80 | 0.88 | 0.91 | 0.95 | 0.75 | 0.91 | 0.87 | 0.73 | 0.89 | 0.79 | 0.71 |
| 29905 | CHI | 0.84 | 0.81 | 0.93 | 0.87 | 0.79 | 0.89 | 0.89 | 0.56 | 0.93 | 0.62 | 0.75 |
| 61765 | FRE | 0.73 | 0.84 | 0.90 | 0.71 | 0.73 | 0.83 | 0.86 | 0.50 | 0.91 | 0.58 | 0.66 |
| 100416 | GER | 0.78 | 0.80 | 0.86 | 0.83 | 0.72 | 0.89 | 0.86 | 0.70 | 0.90 | 0.67 | 0.67 |
| 26649 | HIN | 0.68 | 0.75 | 0.88 | 0.89 | 0.67 | 0.85 | 0.86 | 0.69 | 0.86 | 0.75 | 0.77 |
| 39189 | ITA | 0.68 | 0.85 | 0.92 | 0.94 | 0.74 | 0.93 | 0.89 | 0.69 | 0.92 | 0.72 | 0.72 |
| 3044 | JPN | 0.83 | 0.81 | 0.89 | 0.83 | 0.68 | 0.94 | 0.91 | 0.71 | 0.94 | 0.83 | 0.70 |
| 3150 | KOR | 0.75 | 0.86 | 0.91 | 0.84 | 0.76 | 0.88 | 0.87 | 0.55 | 0.88 | 0.67 | 0.73 |
| 6614 | SPA | 0.79 | 0.90 | 0.86 | 0.85 | 0.78 | 0.85 | 0.92 | 0.67 | 0.90 | 0.70 | 0.68 |
| 12600 | TEL | 0.65 | 0.74 | 0.84 | 0.73 | 0.71 | 0.92 | 0.90 | 0.76 | 0.95 | 0.82 | 0.58 |
| 5565 | TUR | 0.70 | 0.77 | 0.88 | 0.78 | 0.70 | 0.84 | 0.86 | 0.72 | 0.84 | 0.74 | 0.71 |

Table 2: Values corresponding to the proportion of content words in a random essay for each native language for which back-translation yielded the same word.

| | FRE | GER | ITA | SPA | TUR | ARA | HIN | TEL | KOR | JPN | CHI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FRE | 79 | 4 | 4 | 3 | 2 | 3 | 0 | 0 | 2 | 2 | 1 |
| GER | 0 | 89 | 2 | 4 | 1 | 0 | 1 | 0 | 2 | 1 | 0 |
| ITA | 6 | 1 | 83 | 6 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| SPA | 4 | 4 | 5 | 72 | 2 | 3 | 3 | 2 | 1 | 1 | 3 |
| TUR | 3 | 2 | 1 | 3 | 81 | 1 | 3 | 2 | 0 | 3 | 1 |
| ARA | 3 | 0 | 1 | 3 | 3 | 81 | 5 | 2 | 1 | 0 | 1 |
| HIN | 1 | 1 | 1 | 3 | 2 | 1 | 64 | 26 | 1 | 0 | 0 |
| TEL | 0 | 0 | 1 | 0 | 0 | 1 | 17 | 81 | 0 | 0 | 0 |
| KOR | 1 | 1 | 0 | 0 | 3 | 1 | 0 | 0 | 80 | 12 | 2 |
| JPN | 1 | 0 | 2 | 2 | 0 | 3 | 0 | 1 | 13 | 73 | 5 |
| CHI | 0 | 1 | 0 | 0 | 2 | 2 | 0 | 2 | 3 | 3 | 87 |

Table 4: Confusion matrix on the Test set.

| Features | X-Val | Test |
|---|---|---|
| ngm | 74.83% | 75.27% |
| ngm+ort | 74.98% | 75.29% |
| ngm+grm | 75.18% | 75.63% |
| ngm+lex | 74.85% | 75.47% |
| all | 75.57% | 75.81% |
| 2-step (a) | 75.46% | 75.69% |
| 2-step (b) | 75.89% | 75.98% |

Table 3: Accuracy results obtained by cross-validation and using the provided Test set for various combinations of features and our two 2-step strategies. The feature sets are: character and part-of-speech $n$-grams features (ngm), spelling features (ort), grammatical features (grm), and lexical preference features (lex).

from the same feature sets as for the first-step model but with only three labels: one for each language of the pair and one for any other language.

The training data used for these new models include all documents from both languages as well as document misclassified as one of them by the first-step classifier (using cross-validation to label the full training set). The formers keep their original labels while the later are relabeled as *other*.

Document classified in one of the difficult pairs by the first-step classifier were post-processed with these new models. When the new label predicted is *other*, the second best choice of the first step is used.

We investigated two setups for the first classifier: (a) using the original 11 native languages classifier, and (b) using a new classifier with languages of the difficult pairs merged, resulting in 9 native "languages".

Our results, shown in Figure 3 for easy comparison, improve over our system using all features only when the first-pass classifier uses the set of 9 merged pseudo-languages (b). We obtain a moderate 0.32% absolute improvement in accuracy over one-step classification on cross-validation, and 0.17% improvement on the Test set.

## 5 Discussion and conclusion

We have submitted on maximum entropy system to the shared task on Native Language Identification, for which our basic set of $n$-gram features already obtained a level of performance, around 75% in accuracy, close to the best performance reported in our submission. The additional feature sets that we have included in our system, while improving the model, did not allow us to capture a deeper influence of the native language.

A first analysis reveals that the model fails to fully use the additional feature sets due to lack of context. Future experiments will need to link more closely these features to the documents for which they provide useful information.

Due to time constraints and engineering issues, the two-pass system was not ready by the time of submission. The results that we have included in this report show that it is a promising approach that we should continue to explore. We also plan to conduct experiments that exploit the information about the level of English available in the essays, something that we did not consider for this submission. While this information is not directly available, it may be infered from the data as a first-step classification. We believe that studying its influence on the mistakes make learners of different native language is a promising direction.

The approach that we have described in this submission, as most of previously published approaches for this task, attempts to find mistakes in the text of the documents. The most typical mistakes are then used by the classifier to detect the native language. This does not take into consideration the fact that native English writers also make errors. It would be interesting to explore the divergence between various sets of writers/learners, not from the mean of non-native writers, but from the mean of native writers.

## References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan.

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1), March.

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

Thomas Lavergne, Olivier Cappé, and François Yvon.

2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Marcin Milkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software - Practice and Experience*, 40(7):543–566.

Ryo Nagata. 2013. Generating a language family tree from indo-european non-native english texts (to appear). In *Proceedings the 51th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Reinhard Rapp. 2009. The backtranslation score: Automatic mt evalution at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 133–136, Suntec, Singapore.

Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alex Smola, and S.V.N. Vishwanathan. 2009. Hash kernels for structured data. *Journal of Machine Learning Research*, 10:2615–2637, December.

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*, Atlanta, GA, USA, June. Association for Computational Linguistics.