NAIST at the NLI 2013 Shared Task

Tomoya Mizumoto, Yuta Hayashibe
Keisuke Sakaguchi, Mamoru Komachi, Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara 630-0192, Japan
{ tomoya-m, yuta-h, keisuke-sa, komachi, matsu }@is.naist.jp

Abstract

This paper describes the Nara Institute of Science and Technology (NAIST) native language identification (NLI) system in the NLI 2013 Shared Task. We apply feature selection using a measure based on frequency for the closed track and try Capping and Sampling data methods for the open tracks. Our system ranked ninth in the closed track, third in open track 1 and fourth in open track 2.

1 Introduction

There have been many studies using English as a second language (ESL) learner corpora. For example, automatic grammatical error detection and correction is one of the most active research areas in this field. More recently, attention has been paid to native language identification (NLI) (Brooke and Hirst, 2012; Bykh and Meurers, 2012; Brooke and Hirst, 2011; Wong and Dras, 2011; Wong et al., 2011). Native language identification is the task of identifying the ESL learner's L1 given a learner's essay.

The NLI Shared Task 2013 (Tetreault et al., 2013) is the first shared task on NLI using the common dataset "TOEFL-11" (Blanchard et al., 2013; Tetreault et al., 2012). TOEFL-11 consists of essays written by learners of 11 native languages (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Koran, Spanish, Telugu, Turkish), and it contains 1,100 essays for each native language. In addition, the essay topics are balanced, and the number of topics is 8.

In the closed track, we tackle feature selection for increasing accuracy. We use a feature selection method based on the frequency of each feature (e.g., document frequency, TF-IDF).

In the open tracks, to address the problem of imbalanced data, we tried two approaches: **Capping** and **Sampling** data in order to balance the size of training data.

In this paper, we describe our system and experimental results. Section 2 describes the features we used in the system for NLI. Section 3 and Section 4 describe the systems for closed track and open track in NLI Shared Task 2013. Section 5 describes the results for NLI Shared Task 2013. Section 6 describes the experimental result for 10-fold cross validation on the data set used by Tetreault et al. (2012).

2 Features used in all tracks

In this section, we describe the features in our systems. We formulate NLI as a multiclass classification task. Following previous work, we use LIBLINEAR [2] for the classification tool and tune the C parameter using grid-search.

We select the features based on previous work (Brooke and Hirst, 2012; Tetreault et al., 2012). All features used are binary. We treated the features as shown in Table 1. The example of features in Table 1 shows the case whose input is "I think not a really difficult question".

We use a special symbol for the beginning and end of sentence (or word) for bigrams and trigrams. For surface forms, we lowercased all words. POS, POS-function and dependency features are extracted

[1] http://www.lextek.com/manuals/onix/stopwords1.html
[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 134–139,
Atlanta, Georgia, June 13 2013.

Name	Description	Example	
Word N-gram (N=1,2)	Surface form of the word.	N=1 N=2	i, think, not BOS i, i think
POS N-gram (N=2,3)	POS tags of the word.	N=2 N=3	BOS PRP, PRP VBP BOS PRP VBP, PRP VBP RB
Character N-gram (N=2,3)		N=2 N=3	^ t, t h, hi, in, nk, k$ ^ t h, t h i
POS-function N-gram (N=2,3)	We use surface form for words in stop word list [1], otherwise we use POS form.	N=2 N=3	RB difficult, difficult NN RB difficult NN
Dependency	the surface and relation name the surface and the dependend token's surface the surface, relation name and the dependend token's surface		(i, nsubj) (think, i) (nsubj, i, think)
Tree substitution grammer	Fragments of TSG		(PRP_UNK-INITC-KNOWNLC) (VB_think) (NP_RB_DT_ADJP_NN) (JJ_UNK-LC)

Table 1: All features for native language identification.

using the Stanford Parser 2.0.2 [3].

We use tree substitution grammars as features. TSGs are generalized context-free grammars (CFGs) that allow nonterminals to re-write to tree fragments. The fragments reflect both syntactic and surface structures of a given sentence more efficiently than using several CFG rules. In practice, efficient Bayesian approaches have been proposed in prior work (Post and Gildea, 2009). In terms of the application of TSG to NLI task, (Swanson and Charniak, 2012) have shown a promising result. Post (2011) also uses TSG to judge grammaticality of a sentence written by language learners. With these previous findings in mind, we also extract TSG rules. We use the training settings and public software from Post (2011)[4], obtaining 21,020 unique TSG fragments from the training dataset of the TOEFL-11 corpus.

3 Closed Track

In this section, we describe our system for the closed track. We use the tools and features described in Section 2.

In our system, feature selection is performed using a measure based on frequency. Although Tsur and Rappoport (2007) used TF-IDF, they use it to decrease the influence of topic bias rather than for increasing accuracy. Brooke and Hirst (2012) used document frequency for feature selection, however it does not affect accuracy.

We use the native language frequency (hereafter we refer to this as NLF). NLF is the number of native languages a feature appears in. Thus, NLF takes values from 1 to 11. Figure 1 shows an example of NLF. The word bigram feature "in Japan" appears only in essays of which the learners' native language is Japanese, therefore the NLF is 1.

The assumption behind using this feature is that a feature which appears in all native languages affects NLI less, while a feature which appears in few native language affects NLI more. The features whose NLFs are 11 include e.g. "there are", "PRP VBP" and "a JJ NN". Table 2 shows some examples of the features appearing in only 1 native language in the TOEFL-11 corpus. The features include place-name or company name such as "tokyo", "korea", "samsung", which are certainly specific for some native language.

[3] http://nlp.stanford.edu/software/lex-parser.shtml

[4] https://github.com/mjpost/post2011judging

Native Language		
Chinese	Japanese	Korean
carry more	this : NN	samsung
i hus become	of tokyo	of korea
JJ whole and	when i worked	debatable whether
striking conclusion	usuful	NN VBG whether
traffic tools	oppotunity for	in thesedays

Table 2: Example of feature appearing in 1 native language for Chinese, Japanese and Korean

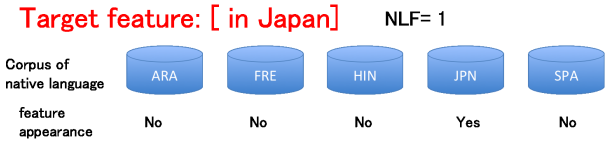

Figure 1: Example of native language frequency

Native Language	# of articles
Japanese	258,320
Mandarin	48,364
Korean	31,188
Spanish	5,106
Italian	2,589
Arabic	1,549
French	1,168
German	832
Turkish	504
Hindi	223
Telugu	19

Table 3: Distribution of native languages in Lang-8 corpus

4 Open tracks

4.1 Lang-8 corpus

For the open tracks, we used Lang-8 as a source to create a learner corpus tagged with the native languages of learners. Lang-8 is a language learning social networking service. [5] Users write articles in their non-native languages and native speakers correct them. We used all English articles written through the end of 2012. We removed all sentences which contain non-ASCII characters. [6]

Almost all users register their native language on the site. We regard users' registered native language as the gold label for each article. We split the learner corpus extracted from Lang-8 into sub-corpora by the native languages. The numbers of articles in all corpora are summarized in Table 3. Unfortunately, some sub-corpora are too small to train the model. For example, the Telugu corpus has only 19 articles.

In order to balance the size of the training data, we tried two approaches: **Capping** and **Sampling**. We confirmed in preliminary experiments that the model with these approaches work better than the model with the original sized data.

Capping

In this approach, we limit the size of a sub-corpus for training to N articles. For a sub-corpus which contains over N articles, we randomly extract articles up to N. We set $N = 5000$ and adapt this approach for Run 1 and Run 3 in the open tracks.

Sampling

In this approach, we equalize the size of all sub-corpora. For corpora which contain less than N articles, we randomly copy articles until their size becomes N. We set $N = 5000$ and adapt this approach for Run 2 and Run 4 in the open tracks.

4.2 Models

We compared two approaches with baseline features and all features.

The models in Run 1 and Run 3 were trained with the data created by the Capping approach, and the models in Run 2 and Run 4[7] were trained by the Sampling approach.

We used only word N-grams ($N = 1, 2$) as baseline features. As extra features we used the following features.

[5] http://lang-8.com/

[6] Some users also add translation in their native languages for correctors' reference.

[7] We did not have time to train the model for Run 4 in the open 1 track.

- POS N-grams ($N = 2, 3$)
- dependency
- character N-grams ($N = 2, 3$)

In open track 2, we also add the TOEFL-11 dataset to the training data for all runs.

5 Result for NLI shared Task 2013

Table 4 shows the results of our systems for NLI Shared Task. Chance accuracy is 0.09. All results outperform random guessing.

5.1 Closed track

In the closed track, we submitted 5 runs. Run 1 is the system using only word 1,2-grams features. Run 2 is the system using all features with NLF feature selection ($1 <$ NLF < 11). Run 3 is the system using word 1,2-grams and POS 2,3-grams features. Run 4 is the system using word 1,2-grams, POS 2,3-grams, character 2,3-grams and dependency features without parameter tuning. Run 5 is the system using word 1,2-grams without parameter tuning. The method using the feature selection method we proposed achieved the best performance of our systems.

5.2 Open tracks

Comparison of the two data balancing approaches

In open track 1, the method of "Sampling" outperforms that of "Capping" (Run 2 $>$ Run 1). This means even duplicated training data can improve the performance.

On the other hand, in open track 2, "Capping" works better than "Sampling" (Run 1 $>$ Run 2 and Run 3 $>$ Run 4). In the first place, the models trained with both Lang-8 data and TOEFL data do not perform better than ones trained with only TOEFL data. This means the less Lang-8 data we use, the better performance we obtain.

Comparison on two feature sets

In open track 1, adding extra features seems to have a bad influence because the result of Run 3 is worse than that of Run 1. This may be because Lang-8 data is out of domain of the test corpus (TOEFL).

	Closed	Open 1	Open 2
Run	**Accuracy**	**Accuracy**	**Accuracy**
1	0.811	0.337	0.699
2	*0.817	0.356	0.661
3	0.808	0.285	0.703
4	0.771	-	0.665
5	0.783	-	-

Table 4: Result for systems which submitted in NLI 2013 *We re-evaluated the Run2 because we submitted the Run1 with the same output as Run2.

In open track 2, adding extra features makes the performance better (Run 3 $>$ Run 1, Run 4 $>$ Run 2). In-domain TOEFL data seem to be effective for training with extra features. In order to improve the result with extra features in open track 2, domain adaptation may be effective.

6 Experiment and Result for 10 fold Cross-Validation

We conducted an experiment using 10-fold cross validation on the data set used by Tetreault et al. (2012). Table 5 shows the results for different feature set. The table consists of 3 blocks; the first block is results of the system using 1 feature, the second block is the result of the system using word 1,2-grams feature and another feature, and the third block is the result of the system using word 1,2-grams and more features.

In the first block results, the system using the word 1,2-grams feature achieved 0.8075. It is the highest accuracy in the first block, and third highest accuracy in the results of Table 5. From the second block of results, adding an extra feature does not improve accuracy, however in the third block the systems in (14) and (15) outperform the system using only word 1,2-grams.

Table 6 shows the results of using feature selection by NLF. The table consists of 3 blocks; the first block is the results of the system using features whose NLF is smaller than N (N = 11, 10, 9, 8), the second block is the results of the system using features whose NLF is greater than N (N = 1, 2, 3, 4), and the third block is the results of the system using features whose NLF is smaller than 11 and greater than N (N = 1, 2, 3, 4).

The best accuracy is achieved by excluding fea-

	Feature	Accuracy
(1)	Word 1,2-gram	0.8075
(2)	POS 2,3-gram	0.5555
(3)	POS,Function 2,3-gram	0.7080
(4)	Chracter 2,3-gram	0.6678
(5)	Dependency	0.7236
(6)	Tree substitution grammar	0.6455
(7)	1 + 2	0.7825
(8)	1 + 3	0.7913
(9)	1 + 4	0.7953
(10)	1 + 5	0.8020
(11)	1 + 6	0.7999
(12)	1 + 2 + 3	0.7849
(13)	1 + 2 + 3 + 4	0.8000
(14)	1 + 2 + 3 + 4 + 5	**0.8097**
(15)	ALL	0.8088

Table 5: 10-fold cross validation results for each feature

	Accuracy
NLF < 11	0.8176
NLF < 10	0.8157
NLF < 9	0.8123
NLF < 8	0.8098
$1 <$ NLF	0.8062
$2 <$ NLF	0.8062
$3 <$ NLF	0.8057
$4 <$ NLF	0.8053
$1 <$ NLF < 11	**0.8209**
$2 <$ NLF < 11	0.8206
$3 <$ NLF < 11	0.8201
$4 <$ NLF < 11	0.8195

Table 6: 10-fold cross validation results using feature selection by NLF. (feature selection is not applied to word N-grams features.)

tures whose NLF is 1 or 11. While the results of the first block and the second block are intuitive, the results of the third block are not (looking at the second block of Table 6, excluding features whose NLF is greater than N (1, 2, 3, 4) reduces accuracy). One possible explanation is that features whose NLF is 1 includes features that rarely appear in the training corpus.

7 Conclusion

In this paper, we described our systems for the NLI Shared Task 2013. We tried feature selection using native language frequency for the closed track and Capping and the Sampling data to balance the size of training data for the open tracks. The feature selection we proposed improves the performance for NLI. The system using our feature selection achieved 0.817 on the test data of NLI Shared Task and 0.821 using 10-fold cross validation. While the Sampling system outperformed Capping system for open track 1, the Capping system outperformed Sampling system in open track 2 (because it reduced the amount of out of domain data).

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. Technical report, Educational Testing Service.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. In *Proceedings of LCR 2011*.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *Proceedings of COLING 2012*, pages 391–408.

Serhiy Bykh and Detmar Meurers. 2012. Native language identification using recurring n-grams – investigating abstraction and domain dependence. In *Proceedings of COLING 2012*, pages 425–440.

Matt Post and Daniel Gildea. 2009. Bayesian Learning of a Tree Substitution Grammar. In *Proceedings of the ACL-IJCNLP 2009*, pages 45–48.

Matt Post. 2011. Judging Grammaticality with Tree Substitution Grammar Derivations. In *Proceedings of ACL 2011*, pages 217–222.

Ben Swanson and Eugene Charniak. 2012. Native Language Detection with Tree Substitution Grammars. In *Proceedings of ACL 2012*, pages 193–197.

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 COLING 2012*, pages 2585–2602.

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*.

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 CACLA*, pages 9–16.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of EMNLP 2011*, pages 1600–1610.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2011. Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124.