Generalization in Native Language Identification: Learners versus Scientists

Sabrina Stehwien and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{sabrina.stehwien,sebastian.pado@ims.uni-stuttgart.de}

Abstract

English. Native Language Identification (NLI) is the task of recognizing an author's native language from text in another language. In this paper, we consider three English learner corpora and one new, presumably more difficult, scientific corpus. We find that the scientific corpus is only about as hard to model as a less-controlled learner corpus, but cannot profit as much from corpus combination via domain adaptation. We show that this is related to an inherent *topic bias* in the scientific corpus: researchers from different countries tend to work on different topics.

Italiano. *La Native Language Identification (NLI) permette di riconoscere la lingua madre di un autore utilizzando il testo scritto in un' altra lingua. In questo lavoro utilizziamo tre collezioni di testi prodotti da apprendenti di inglese e un nuovo corpus scientifico, presumibilmente più difficile. In realtà, il corpus scientifico risulta essere difficile da modellare quanto un corpus di apprendimento meno controllato; tuttavia, a differenza di questi, esso non beneficia della combinazione di diversi corpora con metodi di domain adaptation. Questo limite è legato ad un'intrinseca specializzazione degli argomenti del corpus scientifico: ricercatori di paesi diversi tendono a trattare argomenti diversi.*

1 Introduction

Native Language Identification (NLI) is the task of recognizing an author's native language (L1) from text written in a second language (L2). NLI is important for applications such as the detection of phishing attacks (Estival et al., 2007) or data collection for the study of L2 acquisition (Odlin, 1989). State-of-the-art methods couch NLI as a classification task, where the classes are the L1 of the author and the features are supposed to model the effects of the author's L1 on L2 (*language transfer*). Such features may be of varying linguistic sophistication, from function words and structural features (Tetreault et al., 2012) on one side to N-grams over characters, words and POS tags (Brooke and Hirst, 2011; Bykh and Meurers, 2012) on the other side.

Like in many NLP tasks, there are few large datasets for NLI. Furthermore, it is often unclear how well the models really capture the desired language transfer properties rather than *topics*. The widely-used International Corpus of Learner English (ICLE, Granger et al. (2009)) has been claimed to suffer from a *topic bias* (Brooke and Hirst, 2011): Authors with the same L1 prefer certain topics, potentially due to the corpus collection strategy (from a small set of language courses). As a result, Brooke and Hirst (2013) question the generalization of NLI models to other corpora and propose the use of domain adaptation. In contrast, Bykh and Meurers (2012) report their ICLE-trained models to perform well on other learner corpora.

This paper extends the focus to a novel corpus type, non-native scientific texts from the ACL Anthology. These are substantially different from learner corpora: (a) most authors have a good working knowledge of English; and (b) due to the conventions of the domain, terminology and structure are highly standardized (Hyland, 2009; Teufel and Moens, 2002). A priori, we would believe that NLI on the ACL corpus is substantially more difficult.

Our results show, however, that the differences between the ACL corpus and the various learner corpora are more subtle: The ACL corpus is about as difficult to model as some learner corpora. However, generalization to the ACL corpus is more difficult, due to its idiosyncratic topic biases.

Corpus	# Docs/L1	Avg # Tokens/Doc	Type
TOEFL11	1100	348	Learner
ICLE	251	612	Learner
Lang-8	176	731	Learner
ACL	54	3850	Science

Table 1: Statistics on datasets

2 Datasets

We consider three learner corpora plus one scientific corpus, described below. We consider the 7 languages that are in the intersection of all datasets (DE, ES, FR, IT, JP, TR, ZH). To obtain a balanced setup comparable across corpora, we determined for each corpus the language with fewest documents, and randomly sampled that number of documents from the other languages (cf. Table 1).

TOEFL11. The TOEFL11 corpus (Blanchard et al., 2013) consists of texts that learners of English with mixed proficiency wrote in response to prompts during TOEFL exams.

ICLE. is the oldest and best-researched NLI corpus, a collection of essays written by students with a high intermediate to advanced level of English.

Lang-8. Lang8, introduced in (Brooke and Hirst, 2011) is a web-scraped version of the Lang-8 website[1] where learners of English post texts for correction by native speakers. Although it counts as a learner corpus, it is much less controlled.

ACL. Adapting a method proposed by Lamkiewicz (2014), we extracted a dataset for NLI from the 2012 release of the ACL Anthology Network Corpus (Radev et al., 2013), covering 25,000 papers from the Proceedings of ACL and other ACL-sponsored conferences and workshops. The dataset was extracted according to the e-mail domains of the authors, which were assumed to correspond to their native countries.[2] A document was included if and only if all the e-mail addresses had the same domain. Furthermore, we removed all the headers, acknowledgments and reference sections, since these often contain information on the authors' home country or L1.[3]

3 Models

Our NLI models uses binary features consisting of recurring unigrams and bigrams as proposed by Bykh and Meurers (2012).[4] An n-gram is recurring if it occurs in more than two documents of the same class. As multi-class classifier, we use the LIBLINEAR Support Vector Machine implementation (Chang et al., 2008) with default parameters.

Our standard models are simply trained on one corpus. However, since our focus will be on cross-corpus experimentation, we directly describe the two domain adaptation methods with which we will experiment to improve generalization. We will use the standard terminology of *source* for the (main) training domain and *target* for the testing domain.

Feature Augmentation. Daumé III's (2007) simple but effective domain adaptation method augments the feature space of the problem and can be applied as a preprocessing step to any learning algorithm. It allows feature weights to be learned per domain, by mapping input feature vector onto triplicate version of themselves. The first version of each feature, the "general" version, is identical to the original feature. The second version, the "source" version, is identical to the general version for all instances from the source domain, and zero for all instances from the target domain; vice versa for the third version, the "target" version.

Marginalized Stacked Denoising Autoencoders. Glorot et al. (2011) propose Stacked Denoising Autoencoders (SDAs) for domain adaptation, multilayer networks that reconstruct input data from a corrupted input by learning intermediate (hidden) layers. The intuition is that the intermediate layers model the relevant properties of the input data without overfitting, providing robust features that generalize well across domains. Chen et al. (2012) propose a marginalized SDA (mSDA) which makes the model more efficient while preserving accuracy.

Formally, the input data $\mathbf{x}$ is partially and randomly corrupted into $\tilde{\mathbf{x}}$, e.g., by setting some values to zero. The autoencoder leans a hidden representation from which x is reconstructed: $g(h(\tilde{\mathbf{x}})) \approx \mathbf{x}$. The objective is to minimize the reconstruction error $\ell(\mathbf{x}, g(h(\tilde{\mathbf{x}})))$. We set the corruption probability $p = 0.9$ and the number of layers $l = 1$ in line with previous work. If there are many more features than data points, Chen et al. (2012) use the

[1] http://www.cs.toronto.edu/~jbrooke/Lang8.zip

[2] While this heuristic would fail for countries with a high influx of foreign researchers, like the US, it seems reasonable for countries like Turkey and Japan. Manual evaluation of a small sample showed its precision to be >95%.

[3] The data is available at http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/NLI2015.html

[4] We refrain from using structural features, concentrating on model generalization when using simple lexical features.

Name	Training Data	Test Data
SRC-only	TOEFL11	ICLE Lang-8 ACL
TGT-only	ICLE (2/3) Lang-8 (2/3) ACL (2/3)	ICLE Lang-8 ACL
CONCAT / FA / mSDA -big	TOEFL11 + ICLE (2/3) TOEFL11 + Lang-8 (2/3) TOEFL11 + ACL (2/3)	ICLE Lang-8 ACL
CONCAT / FA / mSDA -small	TOEFL11 + ICLE (1/3) TOEFL11 + Lang-8 (1/3) TOEFL11 + ACL (1/3)	ICLE Lang-8 ACL

Table 2: Model configurations

x most frequent features. The data D is sliced into $\frac{D}{x} = y$ partitions and mSDA is performed on each partition y_i by decoding $g(h(\mathbf{y}_i)) \approx \mathbf{x}$. We set x to 5000 and concatenate the learned intermediate layer units with the original features.

4 Experiments and Results

4.1 Experimental Setup

Table 2 shows all model configurations that we consider. In the SRC-only model, we use the full TOEFL-11 – our largest corpus – as training corpus and test on the other three corpora. The in-domain model (TGT-only), trains and tests always on the same corpora. The next set of models (CONCAT-big, FA-big, mSDA-big) all combine TOEFL-11 as source corpus with two thirds of a target domain corpus, using different combination methods (plain concatenation or the two domain adaptation methods). The final set of models is parallel the previous set, but uses just one third of the target corpora, to assess the influence of the amount of training data.

In all cases except SRC-only, we perform 3-fold cross-validation. We report accuracy, and test statistical significance using the Chi-squared test with Yates' continuity correction (Yates, 1984). Due to the balanced nature of our corpora, the frequency (and random) baselines are at 1/7 = 14.3%.

4.2 Main Experimental Results

The main results are shown in Table 3. The SRC-only results show that the only corpus for which an NLI model trained on TOEFL performs reasonably well is, unsurprisingly, its "nearest neighbor" ICLE, while performance on Lang-8 and ACL is poor. However, even on ICLE, performance remains below 80%. In contrast, the TGT-only results show that reasonable NLI results (generally >80%) can be obtained for each domain if there is target data to train on. Notably, the ACL corpus is easier to model than the Lang-8 corpus despite its special status as a scientific corpus and despite its much smaller size. Yet, the results (except for the very easy ICLE) generally leave room for improvement.

Test data Model	ICLE	Lang-8	ACL
SRC-only	79.5	57.7	49.5
TGT-only	96.1	77.1	**85.7**
CONCAT-big	94.4	80.0	75.1
FA-big	97.0***	84.1*	81.2
mSDA-big	**98.9*****	**90.0*****	**88.4****
CONCAT-small	92.5	75.5	68.8
FA-small	96.0***	77.9	74.6
mSDA-small	**98.6*****	86.8***	**86.0*****

Table 3: Classification accuracies. Bold indicates results not significantly different from best result for each test set ($p<0.05$). Significant improvements over results in previous row marked by asterisks (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

SRC plus a lot of TGT data. The next group of results (*-big*) shows what happens when the SRC data and all available TGT data are combined. This experiment establishes an upper bound of performance for when a lot of target domain data is available. Simple CONCATentation does not perform well, with degradation compared to TGT-only for ICLE and, with a notably large slump, ACL. Feature Augmentation works to some extent, but mSDA improves the results much more substantially, to almost 90% accuracy and above, and yielding the largest improvements over TGT-only (ICLE: +2.8%, Lang-8: +12.9%, ACL:+2.7%). We surmise that FA is handicapped by the relatively small sizes of Lang-8, and the very small size of the ACL corpus, which are "overpowered" by the large TOEFL-11 dataset.

SRC plus some TGT data. The final group of results (*-small*) shows the results of combining the SRC data with only half the available TGT data. In comparison to *-big*, the performance drops substantially for CONCAT and FA, but only somewhat for mSDA (ICLE: -0.3%, Lang-8: -3.2%, ACL: -2.4%; difference statistically significant only for Lang-8). This indicates that mSDA can take advantage of relatively small target domain datasets.

Summary. Domain adaptation, in particular mSDA, can construct highly accurate NLI models (85%+) by combining large source datasets with relatively small target datasets. Contrary to

Test Corpus	ICLE	Lang-8	ACL
SRC-only	76.7	54.5	49.5
TGT-only	96.0	74.8	84.9
mSDA-big	98.7	88.2	86.5

Table 4: Accuracies on reduced feature set

expectations, we do not see a clear division between learner and scientific datasets: rather, the less well controlled Lang-8 behaves much more like the ACL dataset than like ICLE, which in turns clusters together with the TOEFL-11 dataset, the other "classical" learner corpus. This explains the good generalization results found by Bykh and Meurers (2012) but indicates that they may be restricted to "classical" learner corpora.

A difference between Lang-8 and ACL, however, is that Lang-8 still profits significantly from domain adaptation while ACL does not. There is a numerical increase, though, so the low number of documents in the ACL dataset may be responsible.

4.3 Topic vs. L1 Transfer at the Feature Level

To better understand the models, we inspect the most highly weighted n-gram features in the NLI models. As expected, in all models we find language and country names which directly indicate the authors' L1 topically (*"I am from China"*), as opposed to language transfer. To test the importance of these features, we construct a stop word list including the relevant language and country names for each language (e.g. *Italian, Italy* for IT), including *Hong, Kong* for ZH. We use this list to filter out all features that include these stop words.

The results for the reduced feature set are shown in Table 4. They do not differ substantially from our previous experiments. Thus, simple country and language mentions do not seem to have a huge impact on NLI. While this does not exclude the possibility of topic effects among less prominent features, many of the features acquired from the learner corpora that received the most weight are actually interpretable in terms of language transfer, thus exposing writing habits that point towards the author's L1. For example, the FR and ES models include misspellings of loanwords (*"exemple"*, *"advertissements"*, *"necesary"*, *"diferent"*) while DE authors are influenced by German punctuation rules for embedded clauses ("*, that*", "*, because*"). We also see lexical transfer expressed as the overuse of words that are more frequent in the L1 ("*concern*" for FR). What is notable in the ICLE corpus are L1-specific register differences that were found to correlate with topics by Brooke and Hirst (2011): JP writers prefer a colloquial style ("*I think*", "*need to*") while FR writers adopt a more formal style ("*may*", "*the contrary*", "*certainly*").

The situation is quite different in the ACL corpus. While we still find mentions of languages (*"of Chinese", "the German"*), many n-grams reflect scientific jargon and preferred research *topics*. For example, TR researchers write about morphology (*"suffixes", "inflectional", "morphological"*) and ES authors discuss Machine Learning (*"stored", "trained", "the system"*). For some languages, the features appear to be a mixture of specific topics and language transfer: for IT, we find *"category", "implement", "availability", "we obtain", "results in", "accounts for"*. Are these indicative of empirical methodology, or merely results of the (over)use of particular collocations? While we cannot answer this at the moment, we believe that the ACL corpus can thus be considered to have an idiosyncratic form of topic bias – but one that is very different from the learner corpora, which explains the difficulty of generalizing to ACL.

5 Conclusion

This study investigated the generalizability of NLI models across learner corpora and a novel corpus of scientific ACL documents. We found that generalizability is directly tied to corpus properties: well-controlled learner corpora (TOEFL-11, ICLE) generalize well to one another (Bykh and Meurers, 2012). Together with the minor effect on performance of removing topic-related features, we conclude that topic bias *within a similar text type* does not greatly affect generalization.

At the same time, "classical" learner corpora do not generalize well to less-controlled learner corpora (Lang-8) or scientific corpora (ACL). Lang-8 and ACL show comparable performance, which seems surprising given the small size of the ACL corpus and its quite different nature. Our analysis shows that the ACL corpus exhibits an idiosyncratic topic bias: scientists from different countries work on different topics, which is reflected in the models. As a result, the improvements that Lang-8 can derive from domain adaptation techniques carry over to the ACL corpus only to a limited extent. Nevertheless, the use of mSDA can halve the amount of ACL data necessary for the same performance, which is a promising result regarding the generalization to other low-resource domains.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Marin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. *ETS Research Report Series*, 2013(2):1–15.

Julian Brooke and Graeme Hirst. 2011. Native Language Detection with 'Cheap' Learner Corpora. In *Proceedings of the 2011 Conference on Learner Corpus Research*, Louvain-la-Neuve, Belgium.

Julian Brooke and Graeme Hirst. 2013. Using Other Learner Corpora in the 2013 Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–196, Atlanta, Georgia.

Serhiy Bykh and Detmar Meurers. 2012. Native Language Identification Using Recurring N-Grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 425–440, Mumbai, India.

Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin, and Soeren Sonnenburg. 2008. Liblinear: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Minmin Chen, Zhixiang (Eddie) Xu, and Kilian Q. Weinberger. 2012. Marginalized Denoising Autoencoders for Domain Adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the 28th International Conference on Machine Learning*, volume 27, pages 97–110.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English, Version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Ken Hyland. 2009. *Academic Discourse*. Continuum, London.

Anna Maria Lamkiewicz. 2014. Automatische Erkennung der Muttersprache von L2-Englisch-Autoren. Magisterarbeit, Institut für Computerlinguistik, Neuphilologische Fakultät, Ruprecht-Karls-Universität Heidelberg.

Terence Odlin. 1989. *Language Transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

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 24th International Conference on Computational Linguistics*, pages 2585–2602, Mumbai, India.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4).

Frank Yates. 1984. Tests of Significance for 2x2 Contigency Tables. *Journal of the Royal Statistical Society Series A*, 147 (3):426–463.