Arabic Native Language Identification

Shervin Malmasi
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
mark.dras@mq.edu.au

Abstract

In this paper we present the first application of Native Language Identification (NLI) to Arabic learner data. NLI, the task of predicting a writer's first language from their writing in other languages has been mostly investigated with English data, but is now expanding to other languages. We use L2 texts from the newly released Arabic Learner Corpus and with a combination of three syntactic features (CFG production rules, Arabic function words and Part-of-Speech n-grams), we demonstrate that they are useful for this task. Our system achieves an accuracy of 41% against a baseline of 23%, providing the first evidence for classifier-based detection of language transfer effects in L2 Arabic. Such methods can be useful for studying language transfer, developing teaching materials tailored to students' native language and forensic linguistics. Future directions are discussed.

1 Introduction

Researchers in Second Language Acquisition (SLA) investigate the multiplex of factors that influence our ability to acquire new languages and chief among these factors is the role of the learner's mother tongue. Recently this fundamental factor has been studied in Native Language Identification (NLI), which aims to infer the native language (L1) of an author based on texts written in a second language (L2). Machine Learning methods are usually used to identify language use patterns common to speakers of the same L1.

The motivations for NLI are manifold. The use of such techniques can help SLA researchers identify important L1-specific learning and teaching issues. In turn, the identification of such issues can enable researchers to develop pedagogical material that takes into consideration a learner's L1 and addresses them. It can also be applied in a forensic context, for example, to glean information about the discriminant L1 cues in an anonymous text.

While almost all NLI research to date has focused on English L2 data, there is a growing need to apply the techniques to other language in order to assess the cross-language applicability. This need is partially driven by the increasing number of learners of various other languages.

One such case is the teaching of Arabic as a Foreign Language, which has experienced unparalleled growth in the past two decades. For a long time the teaching of Arabic was not considered a priority, but this view has now changed. Arabic is now perceived as a critical and strategically useful language (Ryding, 2013), with enrolments rising rapidly and already at an all time high (Wahba et al., 2013). This trend is also reflected in the NLP community, evidenced by the continuously increasing research focus on Arabic tools and resources (Habash, 2010).

A key objective of this study is to investigate the efficacy of syntactic features for Arabic, a language which is significantly different to English.

Arabic orthography is very different to English with right-to-left text that uses connective letters. Moreover, this is further complicated due to the presence of word elongation, common ligatures, zero-width diacritics and allographic variants. The morphology of Arabic is also quite rich with many morphemes that can appear as prefixes, suffixes or even circumfixes. These mark grammatical information including case, number, gender, and definiteness amongst others. This leads to a sophisticated morphotactic system.

Given the aforementioned differences with English, the main objective of this study is to determine if NLI techniques can be effective for detecting L1 transfer effects in L2 Arabic.

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 180–186,
October 25, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

2 Background

NLI has drawn the attention of many researchers in recent years. With the influx of new researchers, the most substantive work in this field has come in the last few years, leading to the organization of the inaugural NLI Shared Task in 2013 which was attended by 29 teams from the NLP and SLA areas. A detailed exposition of the shared task results and a review of prior NLI work can be found in Tetreault et al. (2013).

While there exists a large body of literature produced in the last decade, almost all of this work has focused exclusively on L2 English. The most recent work in this field successfully presented the first application of NLI to a large non-English dataset (Malmasi and Dras, 2014a), evidencing the usefulness of syntactic features in distinguishing L2 Chinese texts.

3 Data

Although the majority of currently available learner corpora are based on English L2 (Granger, 2012), data from learners of other languages such as Chinese have also attracted attention in the past several years.

No Arabic learner corpora were available for a long time. This paucity of data has been noted by researchers (Abuhakema et al., 2008; Zaghouani et al., 2014) and is thought to be due to issues such as difficulties with non-Latin script and a lack of linguistic and NLP software to work with the data.

More recently, the first version of the Arabic Learner Corpus[1] (ALC) was released by Alfaifi and Atwell (2013). The corpus includes texts by Arabic learners studying in Saudi Arabia, mostly timed essays written in class. In total, 66 different L1 backgrounds are represented. While texts by native Arabic speakers studying to improve their writing are also included, we do not utilize these.

We use the more recent second version of the ALC (Alfaifi et al., 2014) as the data for our experiments. While there are 66 different L1s in the corpus, the majority of these have less than 10 texts and cannot reliably be used for NLI. Instead we use a subset of the corpus consisting of the top seven native languages by number of texts. The languages and document counts in each class are shown in Table 1.

Native Language	Texts
Chinese	76
Urdu	64
Malay	46
French	44
Fulani	36
English	35
Yoruba	28
Total	329

Table 1: The L1 classes included in this experiment and the number of texts within each class.

Both plain text and XML versions of the learner texts are provided with the corpus. Here we use text versions and strip the metadata information from the files, leaving only the author's writings.

4 Experimental Methodology

In this study we employ a supervised multi-class classification approach. The learner texts are organized into classes according on the author's L1 and these documents are used for training and testing in our experiments. A diagram conceptualizing our NLI system is shown in Figure 1.

4.1 Word Segmentation

The tokenization and word segmentation of Arabic is an important preprocessing step for addressing the orthographic issues discussed in §1. For this task we utilize the Stanford Word Segmenter[2].

4.2 Parsing and Part-of-Speech Tagging

To extract the syntactic information required for our models, the Arabic texts are POS tagged and parsed using the Stanford Arabic Parser[3].

4.3 Classifier

We use a linear Support Vector Machine to perform multi-class classification in our experiments. In particular, we use the LIBLINEAR[4] package (Fan et al., 2008) which has been shown to be efficient for text classification problems such as this.

4.4 Evaluation Methodology

In the same manner as many previous NLI studies and also the NLI 2013 shared task, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. In recent years this

[1] http://www.arabiclearnercorpus.com/

[2] http://nlp.stanford.edu/software/segmenter.shtml

[3] http://nlp.stanford.edu/projects/arabic.shtml

[4] http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

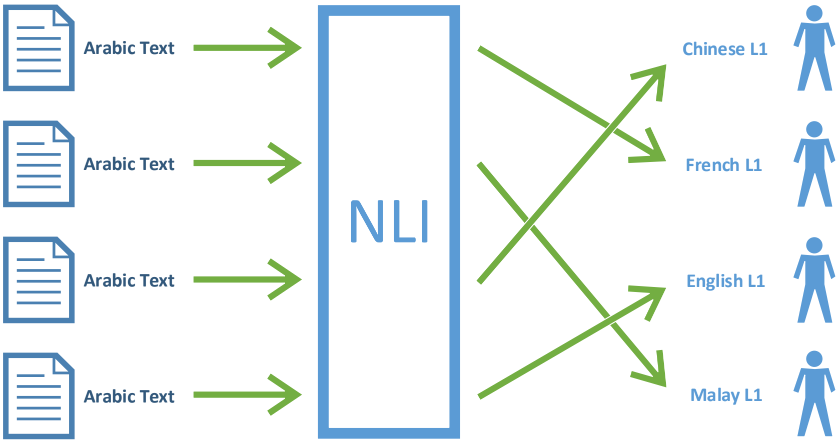


Figure 1: Illustration of our NLI system that identifies the L1 of Arabic learners from their writing.

has become a *de facto* standard for reporting NLI results.

5 Experiments

We experiment using three syntactic feature types described in this section. As the ALC is not balanced for topic, we do not consider the use of lexical features such as word n-grams in this study. Topic bias can occur as a result of the subject matters or topics of the texts to be classified not not evenly distributed across the classes. For example, if in our training data all the texts written by English L1 speakers are on topic A, while all the French L1 authors write about topic B, then we have implicitly trained our classifier on the topics as well. In this case the classifier learns to distinguish our target variable through another confounding variable.

5.1 Context-free Grammar Production Rules

Context-free phrase structure rules (without lexicalizations) are extracted from parse trees of the sentences in each learner text. One such constituent parse tree and extracted rules are shown in Figure 2. These production rules are used as classification features[5]. Linguistically, they capture the global syntactic structures used by writers.

5.2 Arabic Function Words

The distributions of grammatical function words such as determiners and auxiliary verbs have proven to be useful in NLI. This is considered to be a useful syntactic feature as these words indicate the relations between content words and are topic independent. The frequency distributions of a set of 150 function words were extracted from the learner texts and used as features in this model.

السبب فى اختيار الطب هو أننى أحب أن أُدخِلَ السرور في قلوب الناس
وأساعدهم في أزمنة خطيرة.

```
DTNN IN NN DTNN PRP VBD VBP IN VBN DTNN
IN NN DTNN CC NN PRP$ IN NN JJ PUNC
```

Figure 3: An example of a sentence written by a learner and its Part-of-Speech tag sequence. Unigrams, bigrams and trigrams are then extracted from this tag sequence.

5.3 Part-of-Speech n-grams

In this model POS n-grams of size 1–3 were extracted. These n-grams capture small and very local syntactic patterns of language production and were used as classification features.

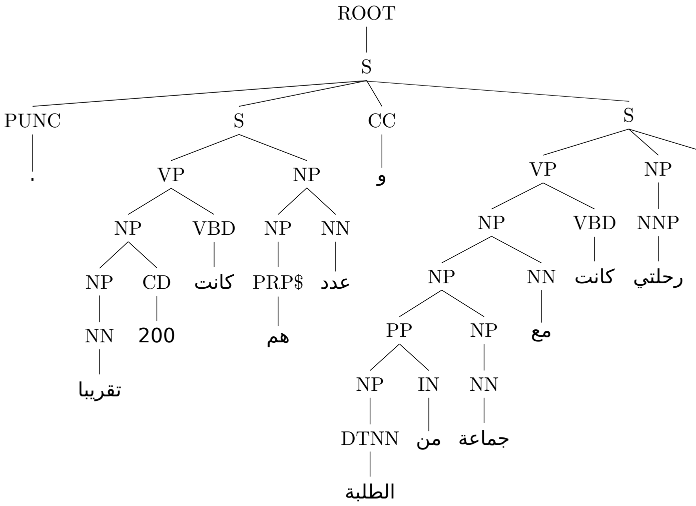
6 Results

The results from all experiments are shown in Table 2. The majority baseline is calculated by using the largest class, in this case Chinese[6], as the default classification. The frequency distributions of the production rules yield 31.7% accuracy, demonstrating their ability to identify structures that are characteristic of L1 groups. Similarly, the distribution of function words is helpful, with 29.2% accuracy.

While all the models provide results well above the baseline, POS tag n-grams are the most useful features, with bigrams providing the highest accuracy for a single feature type with 37.6%. This

[5] All models use relative frequency feature representations

[6] $76/329 = 23.1\%$



```
S  →  S CC S PUNC    VP  →  VBD NP
NP  →  DTNN          PP  →  IN NP
```

Figure 2: A constituent parse tree for a sentence from the corpus along with some of the context-free grammar production rules extracted from it.

Feature	Accuracy (%)
Majority Baseline	23.1
CFG Production Rules	31.7
Function Words	29.2
Part-of-Speech unigrams	36.0
Part-of-Speech bigrams	37.6
Part-of-Speech trigrams	36.5
All features combined	41.0

Table 2: Arabic Native Language Identification accuracy for the three experiments in this study.

seems to suggest that the greatest difference between groups lies in their word category ordering.

Combining all of the models into a single feature space provides the highest accuracy of 41%. This demonstrates that the information captured by the various models is complementary and that the feature types are not redundant.

7 Discussion

The most prominent finding here is that NLI techniques can be successfully applied to Arabic, a morphologically complex language differing significantly from English, which has been the focus of almost all previous research.

This is one of the very first applications of NLI to a language other than English and an important step in the growing field of NLI, particularly with the current drive to investigate other languages. This research, though preliminary, presents an approach to Arabic NLI and can serve as a step towards further research in this area.

NLI technology has practical applications in various fields. One potential application of NLI is in the field of forensic linguistics (Gibbons, 2003; Coulthard and Johnson, 2007), a juncture where the legal system and linguistic stylistics intersect (Gibbons and Prakasam, 2004; McMenamin, 2002). In this context NLI can be used as a tool for Authorship Profiling (Grant, 2007) in order to provide evidence about the linguistic background of an author.

There are a number of situations where a text, such as an anonymous letter, is the central piece of evidence in an investigation. The ability to extract additional information from an anonymous text can enable the authorities and intelligence agencies to learn more about threats and those responsible for them. Clues about the native language of a writer can help investigators in determining the source of anonymous text and the importance of this analysis is often bolstered by the fact that in such scenarios, the only data available to users and investigators is the text itself. One recently studied example is the analysis of extremist related activity on the web (Abbasi and Chen, 2005).

Accordingly, we can see that from a forensic point of view, NLI can be a useful tool for intelligence and law enforcement agencies. In fact, recent NLI research such as that related to the work presented by (Perkins, 2014) has already attracted interest and funding from intelligence agencies (Perkins, 2014, p. 17).

In addition to applications in forensic linguistics, Arabic NLI can aid the development of research tools for SLA researchers investigating language transfer and cross-linguistic effects. Similar data-driven methods have been recently applied to generate potential language transfer hypotheses from the writings of English learners (Malmasi and Dras, 2014c). With the use of an error annotated corpus, which was not the case in this study, the annotations could be used in conjunction with similar linguistic features to study the syntactic contexts in which different error types occur (Malmasi and Dras, 2014b).

Results from such approaches could be used to create teaching material that is customized for the learner's L1. This approach has been previously shown to yield learning improvements (Laufer and Girsai, 2008). The need for such SLA tools is particularly salient for a complex language such as Arabic which has several learning stages (Mansouri, 2005), such as phrasal and interphrasal agreement morphology, which are hierarchical and generally acquired in a specific order (Nielsen, 1997).

The key shortcoming of this study, albeit beyond our control, is the limited amount of data available for the experiments. To the best of our knowledge, this is the smallest dataset used for this task in terms of document count and length. In this regard, we are surprised by relatively high classification accuracy of our system, given the restricted amount of training data available.

While it is hard to make comparisons with most other experiments due to differing number of classes, one comparable study is that of Wong and Dras (2009) which used some similar features on 7-class English dataset. Despite their use of a much larger dataset[7], our individual models are only around 10% lower in accuracy.

We believe that this is a good result, given our limited data. In their study of NLI corpora, Brooke and Hirst (2011) showed that increasing the amount of training data makes a very significant difference in NLI accuracy for both syntactic and lexical features. This was verified by Tetreault et al. (2012) who showed that there is a very steep rise in accuracy as the corpus size is increased towards 11,000 texts[8]. Based on this, we are confident that given similarly sized training data, an Arabic NLI system can achieve similar accuracies. On a broader level, this highlights the need for more large-scale L2 Arabic corpora.

Future work includes the application of our methods to large-scale Arabic learner data as it becomes available. With the ongoing development of the Arabic Learner Corpus and other projects like the Qatar Arabic Language Bank (Mohit, 2013), this may happen in the very near future.

The application of more linguistically sophisticated features also merits further investigation, but this is limited by the availability of Arabic NLP tools and resources. From a machine learning perspective, classifier ensembles have been recently used for this task and shown to improve classification accuracy (Malmasi et al., 2013; Tetreault et al., 2012). Their application here could also increase system accuracy.

We also leave the task of interpreting the linguistic features that differentiate and characterize L1s to future work. This seems to be the next logical phase in NLI research and some methods to automate the detection of language transfer features have been recently proposed (Swanson and Charniak, 2014; Malmasi and Dras, 2014c). This research, however, is still at an early stage and could benefit from the addition of more sophisticated machine learning techniques.

More broadly, additional NLI experiments with different languages are needed. Comparative studies using equivalent syntactic features but with distinct L1-L2 pairs can help us better understand Cross-Linguistic Influence and its manifestations. Such a framework could also help us better understand the differences between different L1-L2 language pairs.

8 Conclusion

In this work we identified the appropriate data and tools to perform Arabic NLI and demonstrated that syntactic features can be successfully applied, despite a scarcity of available L2 Arabic data. Such techniques can be used to generate cross-linguistic hypotheses and build research tools for Arabic SLA. As the first machine learning based investigation of language transfer effects in L2 Arabic, this work contributes important additional evidence to the growing body of NLI work.

[7] Wong and Dras (2009) had 110 texts per class, with average text lengths of more than 600 words.

[8] Equivalent to 1000 texts per L1 class.

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Ghazi Abuhakema, Reem Faraj, Anna Feldman, and Eileen Fitzpatrick. 2008. Annotating an Arabic Learner Corpus for Error. In *LREC*.

Abdullah Alfaifi and Eric Atwell. 2013. Arabic Learner Corpus v1: A New Resource for Arabic Language Research.

Abdullah Alfaifi, Eric Atwell, and I Hedaya. 2014. Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners. In *Proceedings of the Learner Corpus Studies in Asia and the World (LCSAW)*, Kobe, Japan.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. In *Conference of Learner Corpus Research (LCR2011)*, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain.

Malcolm Coulthard and Alison Johnson. 2007. *An introduction to Forensic Linguistics: Language in evidence*. Routledge.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

John Gibbons and Venn Prakasam. 2004. *Language in the Law*. Orient Blackswan.

John Gibbons. 2003. Forensic Linguistics: An Introduction To Language In The Justice System.

Sylviane Granger. 2012. Learner corpora. *The Encyclopedia of Applied Linguistics*.

Tim Grant. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law*, 14(1):1–25.

Nizar Y Habash. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Batia Laufer and Nany Girsai. 2008. Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, 29(4):694–716.

Shervin Malmasi and Mark Dras. 2014a. Chinese Native Language Identification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Shervin Malmasi and Mark Dras. 2014b. From Visualisation to Hypothesis Construction for Second Language Acquisition. In *Graph-Based Methods for Natural Language Processing*, Doha, Qatar, October. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014c. Language Transfer Hypotheses with Linear SVM Weights. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. Nli shared task 2013: Mq submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June. Association for Computational Linguistics.

Fethi Mansouri. 2005. Agreement morphology in Arabic as a second language. *Cross-linguistic aspects of Processability Theory*, pages 117–253.

Gerald R McMenamin. 2002. *Forensic linguistics: Advances in Forensic Stylistics*. CRC press.

Behrang Mohit. 2013. QALB: Qatar Arabic language bank. In *Qatar Foundation Annual Research Conference*, number 2013.

Helle Lykke Nielsen. 1997. On acquisition order of agreement procedures in Arabic learner language. *Al-Arabiyya*, 30:49–93.

Ria Perkins. 2014. *Linguistic identifiers of L1 Persian speakers writing in English: NLID for authorship analysis*. Ph.D. thesis, Aston University.

Karin C. Ryding. 2013. Teaching Arabic in the United States. In Kassem M Wahba, Zeinab A Taha, and Liz England, editors, *Handbook for Arabic language teaching professionals in the 21st century*. Routledge.

Ben Swanson and Eugene Charniak. 2014. Data Driven Language Transfer Hypotheses. *EACL 2014*, page 169.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, Beata Beigman-Klebanov, and Martin Chodorow. 2012. Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. In *Proc. Internat. Conf. on Computat. Linguistics (COLING)*.

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*, pages 48–57, Atlanta, Georgia, June. Association for Computational Linguistics.

Kassem M Wahba, Zeinab A Taha, and Liz England. 2013. *Handbook for Arabic language teaching professionals in the 21st century*. Routledge.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proc. Australasian Language Technology Workshop (ALTA)*, pages 53–61.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large Scale Arabic Error Annotation: Guidelines and Framework. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).