Native Language Identification Using Recurring N-grams

Investigating Abstraction and Domain Dependence

Serhiy Bykh Detmar Meurers
Universität Tübingen
Germany

The 24th International Conference on Computational Linguistics (COLING 2012)
Mumbai, India, 8-15 December 2012

Contents

- Introduction
- Related work
- Our approach
- Study 1: Systematically explore recurring n-grams
- Study 2: Investigate domain dependence
- Conclusions
- Outlook

Introduction

Native Language Identification (NLI)

- NLI determines the *native language (L1)* of an author based on a text written in a *second language (L2)*
- Theoretical relevance:
  - advance understanding of L1 Transfer in Second Language Acquisition
- Practical relevance:
  - author profiling, e.g., for systems identifying the native language of writers of phishing emails (Estival et al. 2007)

Related work

- CL research generally approaches NLI by training machine learning classifiers with L1s as classes, and
- surface-based features (Koppel et al. 2005; Tsur & Rappoport 2007; Estival et al. 2007; Wong & Dras 2009; Brooke & Hirst 2011, 2012)
  - uni-/bi-/tri-grams of characters, words and part-of-speech
  - function words
  - ...
- syntactic features (Wong & Dras 2009, 2011; Swanson & Charniak 2012)
  - subject-verb or noun-number disagreement
  - parse-tree based features
  - ...

Our Approach
Overview

- Machine learning approach systematically exploring
  - all recurring n-grams of any length as features
  - linguistic abstraction to different word classes
- Data-driven approach using up to 160 000 features
- Investigate domain dependence of approach by comparing
  I. Single-corpus evaluation: ICLE
    - *International Corpus of Learner English* (Granger et al. 2009)
    - 16 different L1
    - mainly argumentative essays

  II. Cross-corpus evaluation: ICLE vs. NOCE+USE+HKUST
    - independently compiled learner corpora for three L1
    - argumentative essays

Our Approach

Three independently collected learner corpora

- *NOCE: Non-Native Corpus of English* (Díaz Negrillo 2007, 2009)
  - L1 Spanish
- *USE: Uppsala Student English Corpus* (Axelsson 2000, 2003)
  - L1 Swedish
- *HKUST: Hong Kong University of Science and Technology English Examination Corpus* (Milton & Chowdhury 1994)
  - L1 Chinese

Our Approach: Features

Systematic use of all recurring n-grams as features

- Recurring n-grams of all occurring lengths
  - *recurring*:
    - all n-grams occurring in at least 2 texts of training set
    - idea: include all potentially useful information
  - *of all occurring length*:
    - up to the max. length in the training set
    - idea: long n-grams may capture additional cues, e.g., transliterations of idioms (Milton & Chowdhury 1994)

  \+ are efficiently computable using dynamic programming

  cf. Variation n-gram approach to corpus annotation error detection (Dickinson & Meurers 2003, 2005)
- Use all individual lengths n and intervals $[1, n]$:

  n: uni-grams, bi-grams, tri-grams, etc.

  $[1, n]$: uni-grams, uni- & bi-grams, uni- & bi- & tri-grams, etc.

Our Approach: Features

Systematic exploration of abstraction

- Recurring n-grams with three levels of abstraction:
  i. *Word-based n-grams* (word n-grams):
     - strings of words, i.e., the surface forms
  ii. *Open-Class-POS-based* n-grams (OCPOS n-grams):
     - nouns, verbs, adjectives and cardinal numbers are represented by their part-of-speech (POS) tags
  iii. *POS-based* n-grams (POS n-grams):
     - all words are represented by their POS tags

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Our Approach: Features

Example for $max_n(d) = 5$

size of n →

abstraction ↓

	1	2	3	4	5
Word	all	men	are	equal	but
OCPOS	all	NNS	VBP	JJ	but
POS	DT	NNS	VBP	JJ	CC

Word-based n-grams:

$n = 1$: *all, men, are, equal, but*
$n = 2$: *all men, men are, are equal, equal but*
...
$n = 5$: *all men are equal but*

POS-based n-grams:

$n = 1$: *DT, NNS, VBP, JJ, CC*
$n = 2$: *DT NNS, NNS VBP, VBP JJ, JJ CC*
...
$n = 5$: *DT NNS VBP JJ CC*

OCPOS-based n-grams:

$n = 1$: *all, NNS, VBP, JJ, but*
$n = 2$: *all NNS, NNS VBP, VBP JJ, JJ but*
...
$n = 5$: *all NNS VBP JJ but*

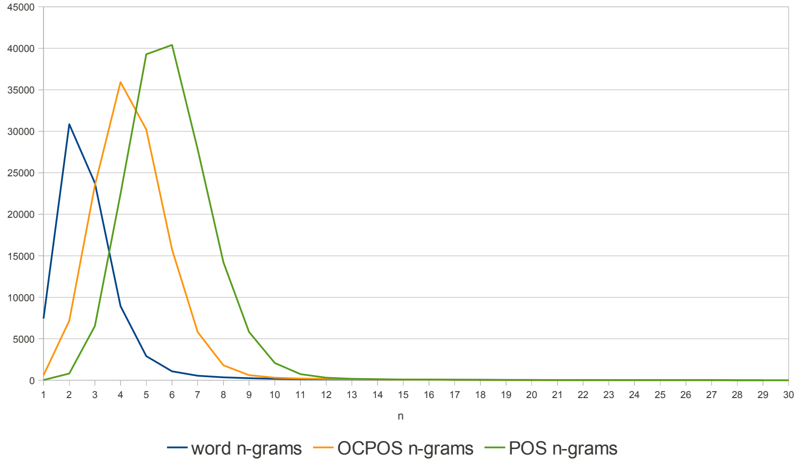
Our Approach: Tools

- POS Tagging:
  - *OpenNLP* POS-tagger (http://opennlp.apache.org)
  - *PennTreebank* tagset (Santorini 1990)
- Machine Learning:
  - *SVM* (*LIBLINEAR*, Fan et al. 2008)
  - Feature representation: Binary vectors
    - $\{1, 0\}$ encoding the presence of a feature in a given text

First Study: Explore recurring n-grams

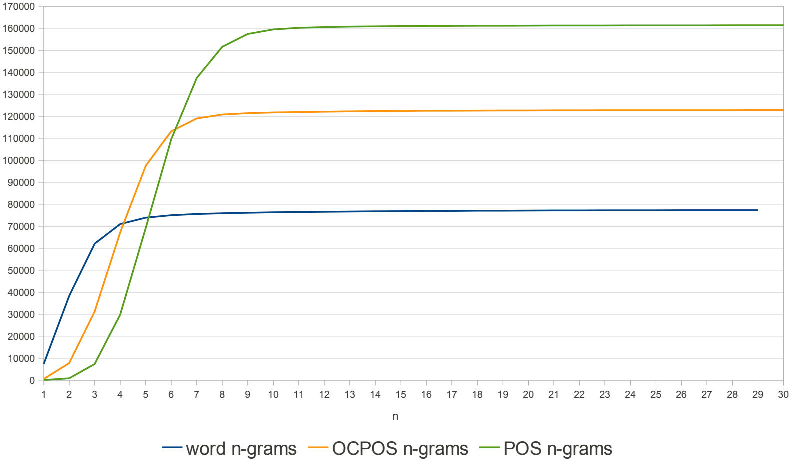
- Setup: just as in Wong & Dras (2009, 2011)
  - Corpus: ICLE, v2
  - Seven L1: Bulgarian, Czech, French, Russian, Spanish, Chinese and Japanese
  - Data split:
    - Training: 7 L1 · 70 essays = 490 essays
    - Testing: 7 L1 · 25 essays = 175 essays
    - Essay Length: 500–1000 words
- Evaluation:
  a) one training and test set, as in Wong & Dras (2009, 2011)
  b) ten randomly selected training and test sets to observe variance

Study 1a: Number of features (single *n*)



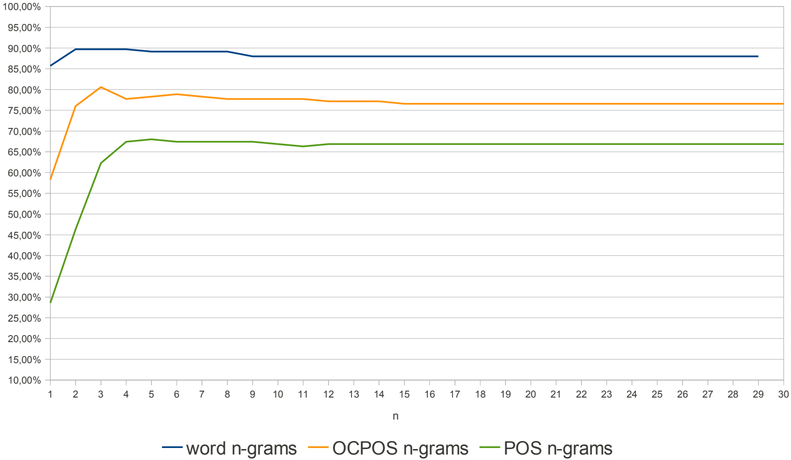
- N-grams with $n \leq 10$ potentially informative, $n > 10$ too sparse
- More abstraction leads to more recurring n-grams
  - e.g. "*all NNS are*" arises from "*all men are*", "*all people are*"

Study 1a: Number of features ([1, n] intervals)



- around 160 000 features for POS [1–10]-grams

Results of Study 1a: Accuracy ([1, n] intervals)



- N-grams with $n \leq 5$ useful (for POS n-grams)
- The more abstraction the lower the accuracy

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

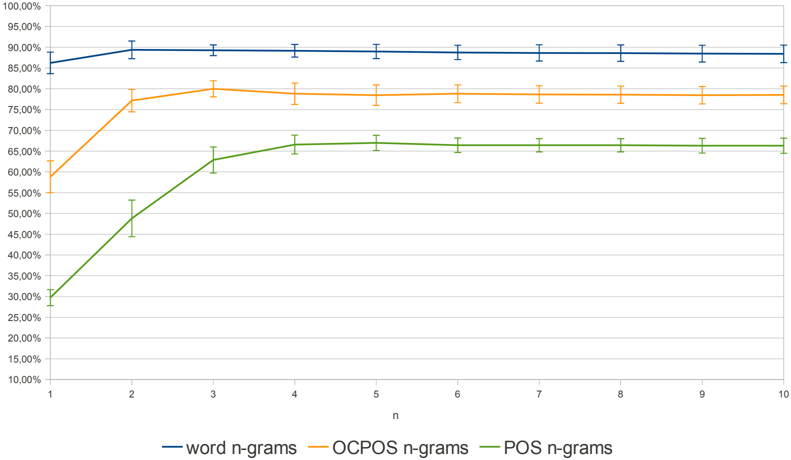
Results of Study 1a: Best Results

- Random baseline, given seven L1 classes: 14.29%
- Best result: 89.71% (using word-based n-grams $[1, 2]$)
- 8% improvement over previous best result on comparable setup (81.71% of Wong & Dras 2011)

	n Intervals			Single *n*		
Features	$[1, n]$	**Accuracy**	**Feature #**	n	**Accuracy**	**Feature #**
word n-grams	2	89.71%	38,300	1	85.71%	7,446
OCPOS n-gr.	3	80.57%	31,263	2	74.86%	7,176
POS n-grams	5	68.00%	69,139	4	65.14%	22,462

- interval results always better than single *n*

Study 1b (ten samples): Mean Accuracy & SSD



- Best *mean* accuracy: 89.37% (word-based n-grams [1, 2]) ≈ best single sample accuracy 89.71%
- Means close to the results on the single sample

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

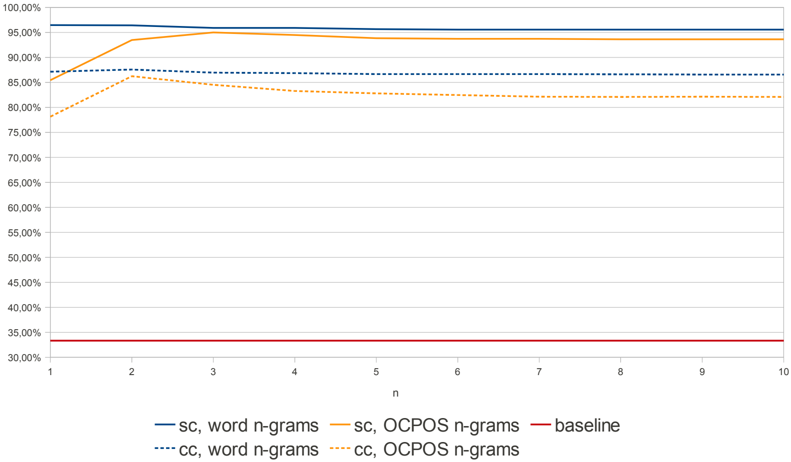
Study 2: Domain Dependence
Motivation

- Very good results for ICLE
- But did we learn something about Native Language Identification – or only about ICLE?
- Brooke & Hirst (2011): ICLE-trained classifier performs poorly for web-data based Lang-8 corpus
  - Lang-8: corpus consisting of short personal narratives, requests for translation of particular phrases, etc.
- Is the drop caused by specific properties of Lang-8 or does it indicate that patterns learned on ICLE do not generalize?

Study 2: Domain Dependence
Setup

- Explore domain dependence with independently collected corpora of argumentative essays on different topics
- Corpora: ICLE vs. USE+NOCE+HKUST
  - L1: Spanish, Swedish and Chinese
- Data split & Evaluation:
  - Training on ICLE:
    - trained ten models on randomly selected essays per L1
    - for each model: 3 L1 · 140 essays = 420 essays
  - Testing (always on data not included in training):
    i. *Single-Corpus (SC)* from ICLE:
    3 L1 · 70 essays = 210 essays
    ii. *Cross-Corpus (CC)* on NOCE+USE+HKUST:
    3 L1 · 70 essays = 210 essays
    (NOCE: 140 essays, pairwise merged to standardize length)

Results for Study 2: Mean accuracy (ten models)



- Best mean accuracy: 96.48% Single Corpus (ICLE)
  87.57% Cross Corpus (ICLE vs. NOCE+USE+HKUST)
- Variance for ten models (SSD): 0.64% Single Corpus
  1.32% Cross Corpus

Study 2: Domain Dependence
Conclusion on Cross-corpus Evaluation

- The ICLE trained classifier in our approach successfully performs NLI on three independently collected corpora.
  - Cross-corpus drop of about 9% when training on ICLE and testing on NOCE+USE+HKUST (baseline: 33.3%)
- Brooke & Hirst (2011)
  - Cross-corpus drop of over 65% when training on ICLE and testing on Lang-8 (baseline: 14.2%)
- Some potential causes for the differences:
  - specific characteristics of Lang-8
  - possibly related to genre differences
    - argumentative essays vs. collated web-data

⇒ Within same genre, surface-based n-gram models seem to provide good cross-corpus performance for NLI.

Summary

- Recurring n-grams:
  - Best result: 89.71% in a task with seven L1 on ICLE
  - N-gram lengths up to 5 were useful
  - N-grams of all lengths together better than single lengths
- Abstraction: word-based n-grams outperformed part-of-speech based n-grams on single-corpus & cross-corpus evaluation
  - Apparently people with different L1 backgrounds make lexical choices indicative across a range of topics.
    - e.g., *might, consider, be able to, make use of*
- Domain Dependence:
  - N-gram patterns learned on ICLE generalized well to three independently collected corpora of the same genre.

Outlook

- Systematically explore more types of linguistic abstractions as features, especially
  - their usefulness across genres, and
  - the insights they provide for understanding L1 Transfer in Second Language Acquisition.
- Explore target languages other than English
  - to ensure generalizability of results
  - to enhance study of morphological or word order transfer

Thank you for your attention!

Questions?

References

Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning* 21(4), 323–338. URL http://purl.org/dm/papers/amaral-meurers-call08.html.

Axelsson, M. W. (2000). USE – The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal* 24, 155–157. URL http://icame.uib.no/ij24/use.pdf.

Axelsson, M. W. (2003). *Manual: The Uppsala Student English Corpus (USE)*. Uppsala University, Department of English, Sweden. Available at http://www.engelska.uu.se/Research/English_Language/Research_Areas/Electronic_Resource_Projects/USE-Corpus.

Brooke, J. & G. Hirst (2011). Native Language Detection with 'Cheap' Learner Corpora. In *Learner Corpus Research 2011 (LCR 2011)*. Louvain-la-Neuve.

Brooke, J. & G. Hirst (2012). Measuring Interlanguage: Native Language Identification with L1-influence Metrics. In *Proceedings of the 8th ELRA Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, pp. 779–784. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/129_Paper.pdf.

Chang, C.-C. & C.-J. Lin (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Daelemans, W., J. Zavrel, K. van der Sloot & A. van den Bosch (2007). *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, Tilburg, The Netherlands. URL http://ilk.uvt.nl/downloads/pub/papers/ilk.0703.pdf. Version 6.0.

Dickinson, M. & W. D. Meurers (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, pp. 107–114. URL http://purl.org/dm/papers/dickinson-meurers-03.html.

Dickinson, M. & W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. pp. 322–329. URL http://aclweb.org/anthology/P05-1040.

Díaz Negrillo, A. (2007). A Fine-Grained Error Tagger for Learner Corpora. Ph.D. thesis, University of Jaén, Spain.

Díaz Negrillo, A. (2009). *EARS: A User's Manual*. Munich, Germany: LINCOM Academic Reference Books.

Estival, D., T. Gaustad, S. Pham, W. Radford & B. Hutchinson (2007). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. pp. 263–272.

Fan, R., K. Chang, C. Hsieh, X. Wang & C. Lin (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research* 9, 1871–1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (2009). *International Corpus of Learner English, Version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten (2009). The WEKA Data Mining Software: An Update. In *The SIGKDD Explorations*. vol. 11, pp. 10–18.

Koppel, M., J. Schler & K. Zigdon (2005). Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05)*. New York, pp. 624–628. URL http://portal.acm.org/ft_gateway.cfm?id=1081947&type=pdf&CFID=9343231&CFTOKEN=78617578.

Milton, J. C. P. & N. Chowdhury (1994). Tagging the interlanguage of Chinese learners of English. In *Proceedings joint seminar on corpus linguistics and lexicology, Guangzhou and Hong Kong, 19-22 June, 1993, Language Centre, HKUST*. Hong Kong, pp. 127–143. URL http://hdl.handle.net/1783.1/1087.

Platt, J. C. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Tech. Rep. MSR-TR-98-14, Microsoft Research.

Santorini, B. (1990). *Part-of-speech Tagging Guidelines for the Penn Treebank, 3rd Revision, 2nd Printing*. Tech. rep., Department of Computer Science, University of Pennsylvania. URL http://www.cs.bgu.ac.il/~nlpproj/nlp02/papers/treebank-tagset.pdf.

Swanson, B. & E. Charniak (2012). Native Language Detection with Tree Substitution Grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 193–197. URL http://www.aclweb.org/anthology/P12-2038.

Tsur, O. & A. Rappoport (2007). Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA '07)*. Stroudsburg, pp. 9–16. URL http://leibniz.cs.huji.ac.il/tr/952.pdf.

Wong, S.-M. J. & M. Dras (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*. pp. 53–61.

Wong, S.-M. J. & M. Dras (2011). Exploiting Parse Structures for Native Language Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK., pp. 1600–1610. URL http://aclweb.org/anthology/D11-1148.