
Native Language Identification
State of the Art

What's native language identification?

* NLI is the task of identifying the native language (L1) of a writer based solely on a sample of this author writing (1)
* For example, a Spanish journalist writing news in Arabic, a French student writing essays in English or a Brazilian user tweeting in Spanish...

(1) https://sites.google.com/site/nlisharedtask2013/home

Why is NLI important?

* Education: More targeted feedback to language learners about their errors
* Marketing: Better market segmentation based on native idiosyncrasies
* Forensics: Helping in author profiling
* Security: Profiling possible threats
* ...

Some variations / relations

* Native Language Identification: For example, people from different nationalities speaking English
* Language Varieties or Dialects Identification: For example, Portuguese of Portugal vs. Brazil, Spanish of Spain, Argentina, Mexico...

Outline

* Representative Works
* Common Resources / Corpora
* Some Issues
* Research niches
* References

Representative works

* Determining an Author's Native Language by Mining a Text for Errors. Koppel, M., Schler, J., Zigdon, K. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD'05
* A Report on the First Native Language Identification Shared Task. Tetreault, J., Blanchard, D., Cahill, A. In the 8th Workshop on Innovative Use of NLP for Building Educational Applications BEA-8. NAACL-HTL 2013
* Using Other Learner Corpora in the 2013 NLI Shared Task. Brooke, J., Hirst, G. In the 8th Workshop on Innovative Use of NLP for Building Educational Applications BEA-8. NAACL-HTL 2013
* Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. Bykh, S., Meurers, D. The 25th International Conference on Computational Linguistics COLIN 2014
* Author's Native Language Identification from Web-Based Texts. Tofight, P., Köse, C., Rouka, L. International Journal of Computer and Communication Engineering 2012
* Automatic Identification of Language Varieties: The Case of Portuguese. Zampieri, M., Gebrekidan, B. In Proceedings of the Conference on Natural Language Processing 2012
* Automatic Identification of Arabic Language Varieties and Dialects in Social Media. Sadat, F., Kazemi, F., Farzindar, A. In Proceeding of the 1st. International Workshop on Social Media Retrieval and Analysis SoMeRa 2014
* ...

Determining an Author's Native Language by Moining a Text for Errors.
Koppel, M., Schler, J., Zigdon, K.

In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD'05.

* The first published work in this area
* Corpus: International Corpus of Learner English, ICLE
* L1: 258 authors from Russia, Czech Republic, Bulgaria, France, Spain
* L2: English
* Features:
  * Function words (400)
  * Letter n-grams (200)
  * Errors and idiosyncrasies (185 error types + 250 rare POS bigrams)
    * Ortography: repeated letter (remmit/remit), double letter appears only once (comit/commit)...
    * Syntax: run-on sentence, mismatched singular/plural, mismatched tense, that/which confusion...
    * Neologism: e.g fantabolous
    * Rare Parts-of-Speech bigrams in the Brown corpus [http://www.wikiwand.com/en/Brown_Corpus]
* ML Algorithm: Multi-class linear Support Vector Machines
* Evaluation method: 10-fold cross-validation
* Accuracy ~ 80%

A Report on the First Native Language Identification Shared Task.
Tetreault, J., Blanchard, D., Cahill, A.

In the 8th Workshop on Innovative Use of NLP for Building Educational Applications BEA-8. NAACL-HTL 2013

* The first shared task in this area
* Corpus: Test of English as a Foreign Language, TOEFL11
  * 8 prompts (i.e. topics)
  * TOEFL11-train (900), TOEFL11-dev (100), TOEFL11-test (100) per L1
* L1: 1100 essays per language, Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish
* L2: English
* Task:
  * Closed-training: 1 11-way classification task using only TOEFL11-train and optionally TOEFL11-dev
  * Open-training-1: Participants could use any amount of training data excluding TOEFL11
  * Open-training-2: Any kind of training data, even TOEFL11-train and -dev
* 29 teams, 24 papers, up to 5 systems per task
* Features: the most common were word, character and POS n-gram feature
* ML Algorithm: The majority used SVM (13). Also Maximum Entropy (3), Ensemble (3), Discriminant Function Analysis (1) and K-Nearest Neighbors (1)...
* Evaluation method: TOEFL11-test in the three subtasks
* Accuracy:
  * Closed-training: 83.6%
  * Open-training-1: 56.5%
  * Open-training-2: 83.5%

Using Other Learner Corpora in the 2013 NLI Shared Task.
Brooke, J., Hirst, G.

In the 8th Workshop on Innovative Use of NLP for Building Educational Applications BEA-8. NAACL-HTL 2013

* Corpus:
  * TOEFL11: 11 L1
  * Lang-8: Website where language learners write journal entries in their L2 to be corrected by native speakers. 11 L1 overlapping with TOEFL11
  * ICLE: 15 L1, 8 overlap with TOEFL11
  * FCE: Small sample of the First Certificate in English. 16 L1, 9 overlap with TOEFL11
  * ICCI: International Corpus of Crosslinguistic Interlanguage. 4 L1 overlap with TOEFL11
  * ICANLE: International Corpus Network of Asian Learners of English. 3 L1 overlap with TOEFL11
* L1: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish
* L2: English
* Features:
  * Function words, Word n-grams (up to bigrams), POS n-grams (up to trigram), character n-grams (up to trigram), dependencies, context-free productions, 'mixed' POS/function n-grams (up to trigram), i.e. n-grams with all lexical words replaced with part of speech.
  * The best model: word n-grams + mixed POS/function n-grams
* ML Algorithm: Support Vector Machines
* Evaluation method: TOEFL11 1-test
* Accuracy:
  * Closed-task: 80.2% (12 / 29)
  * Open-1-task: 56.5% (1 / 3)
  * Open-2-task: 81.6% (2 / 4)

Exploring Syntactic Features for Native Language Identification: A Variatonist Perspective on Feature Encoding and Ensemble Optimization

Bykh, S., Meurers, D.

The 25th International Conference on Computational Linguistics COLIN 2014

* Corpus:
  * TOEFL11 1: 11 L1
  * NT1 1: 5843 texts from ICLE + FCE + BALC + ICNALE + TÜTEL-NLI.
* L1: Arabic (846), Chinese (1048), French (456), German (500), Hindu (400), Italian (467), Japanese (447), Korean (684), Spanish (446), Telugu (200), Turkish (349)
* L2: English
* Features: Context-Free Grammar Rules
  * Only phrasal CFG production rules excluding all terminals (S->NP VP,N -> D NN, ...)
  * Only lexicalized CFG production rules of the type predeterminal -> terminal (JJ -> nice, JJ -> quick, NN -> vacation, ...)
  * The union (combination) of the above two
* ML Algorithm: Logistic Regression
* Evaluation method: Cross-corpus, NT11 for training, TOEFL11-test for testing
* Accuracy: 84.82%

Author's Native Language Identification from Web-Based Texts.
Tofight, P., Köse, C., Rouka, L.

International Journal of Computer and Communication Engineering 2012

* Corpus: 600 publicly available news agencies texts
* L1: 150 texts from each, English, Persian, Turkish, German
* L2: English
* Features:
    * Lexical (64): character n-grams, word-length frequency, vocabulary richness...
    * Syntactic (308): common punctuation signs, function words (e.g. the)...
    * Structural (13): paragraph length, use of greetings...
    * Content-specific (): n-grams with TF > 10
* ML Algorithm: Support Vector Machines
* Evaluation method: 10-fold cross-validation
* Accuracy 70% ~ 80%

Automatic Identification of Language Varieties: The Case of Portuguese.
Zampieri, M., Gebrekidan-Gebre, B.

In Proceedings of the Conference on Natural Language Processing 2012

- Corpus (1000 documents from newsletters):
  - Brazilian corpus: Folha de São Paulo, newspaper 2004
  - Portuguese corpus: Diário de Notícias, newspaper 2007
- Features: word and character n-grams
    - Orthography:
      - Graphical signs: econômico (BP); económico (EP); economic (EN)
      - Mute consonants: ator (BP); actor (EP); actor (EN)
    - Syntax:
      - Pronouns: eu te amo (BP); eu amo-te (EP); I love you (EN)
    - Lexical variation:
      - multa (BP); coima (EP); fine, penalty (EN)
    - Proper nouns
- ML Algorithm: Language probability distributions with log-likelihood function for probability estimation
- Evaluation method: 50/50 split
- Accuracy:
  - Word uni-grams: 99.6%
  - Word bi-grams: 91.2%
  - Character 4-grams: 99.8%

Automatic Identification of Arabic Language Varieties and Dialects in Social Media. Sadat, F., Kazemi, F., Farzindar, A.

In Proceeding of the 1st. International Workshop on Social Media Retrieval and Analysis SoMeRa 2014

* Corpus (blogs and forum documents): 6 regional variations
  * Egyptian: Egypt
  * Iraqui: Iraq
  * Gulf: Bahrein, Emirates, Kuwait, Qatar, Oman, Saudi Arabia
  * Maghrebi: Algeria, Tunisia, Morocco, Libya, Mauritania
  * Levantine: Jordan, Lebanon, Palestine, Syria
  * Others: Sudan
* Features: character n-grams
* ML Algorithm: Markov language model vs. Naïve Bayes
* Evaluation method: 50/50 split
* Accuracy: 98% (78% F-measure)

Common resources / corpora

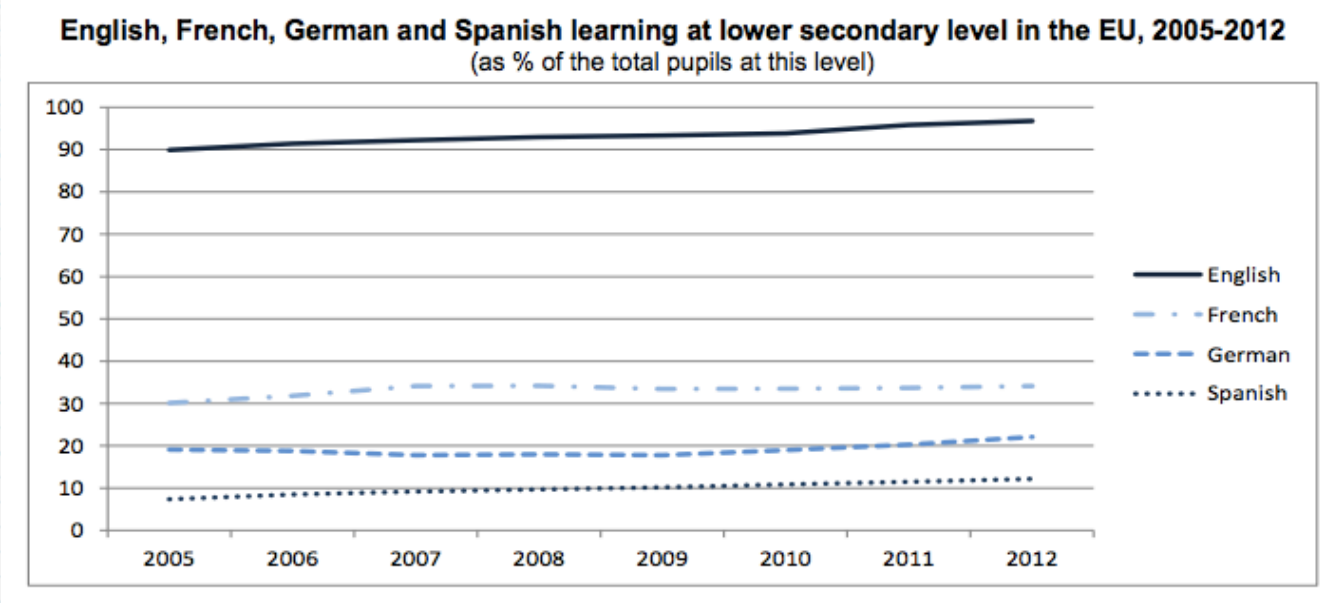
- ICLE: International Corpus of Learner English (Granger et al., 2009)
  - Essays written by college-level English language learners
  - Issues: Quite small, topic bias
  - L1 (11) except Arabic, Hindi and Telugu; L2 English
- Lang-8: http://www.lang8.com
  - Social networking service where users write in the language they are learning and get corrections from native speakers
- FCE: First Certificate in English (Yannakoudakis et al., 2011)
  - Essays written for an English assessment exam
  - L1 (16) except Arabic, Hindi and Telugu; L2 English
- ICCI: International Corpus of Crosslinguistic Interlanguage (Tono et al., 2012)
  - Descriptive and argumentative essays written by young learnets, i.e. those in grade school
  - L1 (4); L2 English
- TOEFL11: Test of English as a Foreign Language (Blanchard et al., 2013)
  - Essays written during high-stakes college-entrance test
  - L1 (11) Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish; L2 English
- ICANLE: International Corpus Network of Asian Learners of English (Ishikawa, 2011)
  - Essays from college students
  - L1 (10): Asian background (Chinese, Japanese, Korean); L2 English
- BALC (Randall and Groom, 2009); TÜTEL-NLI (Bykh et al., 2013)

Some Issues

* Most of the corpora were built from formal media, based on essays of proficient students
* Different corpora have essays with different proficiency level. Even high differences inside the same corpus
* Very few research works from social media, were people express themselves in other languages without taking care about their errors
* All the works are focused on English as a second language

Research niches

* Spanish as a second language (http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-25092014-AP/EN/3-25092014-AP-EN.PDF)



* Spanish variations (Hispablogs)

Lang	Blogs	Words	Max_W	Min_W	Avg_W	Std_W
AR	450	1408103	11117	502	3126.90	2183.75
CL	450	1081068	10336	384	2402.37	2378.07
ES	450	1376478	11141	336	3058.84	2234.14
MX	450	1697091	11946	725	3771.31	2514.51
PA	450	950076	13090	120	2111.28	2264.03
PE	450	1602195	13205	620	356043	2515.71

References

* Blanchard, D, Treteault, J., Higgins, D., Cahill, A., Chodorow, M. TOEFL11: A Corpus of Non-Native English. Technical Report, Educational Testing Service. 2013
* Bykh, S., Vajjala, S., Krivanek, J., Meurers, D. Combining Shallow and Linguistically Motivated Features in Natural Language Identification. In Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications BEA-8 at NAACL-HLT 2013
* Granger, S., Dagneaus, E., Meunier, F. The International Corpus of Learner English: Handbook and CD-ROM, version 2. Presses Universitaries de Louvain. 2009
* Ishikawa, S. A New Horizon in Learner Corpus Studies: The Aim of the ICNALE Projects. In Corpora and Language Tecnologies in Teaching. Learning and Research. University of Strathclyde Publishing. 2011
* Randall, M., Groom, N. The BUiD Arab Learner Corpus: A Resource for Studying the Acquisition of L2 English Spelling. In Proceedings of the Corpus Linguistic Conference CL 2009
* Yannakoudakis, H., Briscoe, T., Medlock, B. A New Dataset and Method for Automatically Grading ESOL Texts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011