Title: Native-language identification

URL: https://en.wikipedia.org/wiki/Native-language_identification

PageID: 45627703

Categories: Category:Applied linguistics, Category:Bilingualism, Category:Computational linguistics, Category:Machine learning, Category:Natural language processing, Category:Second-language acquisition

---

Native-language identification (NLI) is the task of determining an author's native language (L1) based only on their writings in a second language (L2). [1] NLI works through identifying language-usage patterns that are common to specific L1 groups and then applying this knowledge to predict the native language of previously unseen texts. This is motivated in part by applications in second-language acquisition , language teaching and forensic linguistics , amongst others.

Overview

NLI works under the assumption that an author's L1 will dispose them towards particular language production patterns in their L2, as influenced by their native language. This relates to cross-linguistic influence (CLI), a key topic in the field of second-language acquisition (SLA) that analyzes transfer effects from the L1 on later learned languages.

Using large-scale English data, NLI methods achieve over 80% accuracy in predicting the native language of texts written by authors from 11 different L1 backgrounds. [2] This can be compared to a baseline of 9% for choosing randomly.

Applications

Pedagogy and language transfer

This identification of L1-specific features has been used to study language transfer effects in second-language acquisition. [3] This is useful for developing pedagogical material, teaching methods, L1-specific instructions and generating learner feedback that is tailored to their native language.

Forensic linguistics

NLI methods can also be applied in forensic linguistics as a method of performing authorship profiling in order to infer the attributes of an author, including their linguistic background.

This is particularly useful in situations where a text, e.g. an anonymous letter, is the key piece of evidence in an investigation and clues about the native language of a writer can help investigators in identifying the source.

This has already attracted interest and funding from intelligence agencies. [4]

Methodology

Natural language processing methods are used to extract and identify language usage patterns common to speakers of an L1-group. This is done using language learner data, usually from a learner corpus . Next, machine learning is applied to train classifiers, like support vector machines , for predicting the L1 of unseen texts. [5] A range of ensemble based systems have also been applied to the task and shown to improve performance over single classifier systems. [6] [7]

Various linguistic feature types have been applied for this task. These include syntactic features such as constituent parses, grammatical dependencies and part-of-speech tags.

Surface level lexical features such as character, word and lemma n-grams have also been found to be quite useful for this task. However, it seems that character n-grams [8] [9] are the single best feature for the task.

2013 shared task

The Building Educational Applications (BEA) workshop at NAACL 2013 hosted the inaugural NLI shared task. [10] The competition resulted in 29 entries from teams across the globe, 24 of which also published a paper describing their systems and approaches.

See also

Crosslinguistic influence – Ways bilingual people's languages influence their use of the other

Foreign language writing aid – Assistive technology for non-native language users Pages displaying short descriptions of redirect targets

Computer-assisted language learning – Learning technique

Language education – Process and practice of acquiring a language

Natural language processing – Processing of natural language by a computer

Language transfer – Influence one language has on the acquisition or intelligibility of another

References