Data linguistica 29

Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning

av Ildikó Pilán

Akademisk avhandling för filosofie doktorsexamen
i språkvetenskaplig databehandling,
som enligt beslut av humanistiska fakultetsnämnden
vid Göteborgs universitet kommer att försvaras offentligt torsdagen
den 14 juni 2018 kl. 13.15 i Stora hörsalen (2150), Eklandagatan 86.

GÖTEBORGS UNIVERSITET

Göteborg 2018

TITLE: Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning
LANGUAGE: English
AUTHOR: Ildikó Pilán

Abstract

With the ever-growing presence of electronic devices in our everyday lives, it is compelling to investigate how technology can contribute to make our language learning process more efficient and enjoyable. A fundamental piece in this puzzle is the ability to measure the complexity of the language that learners are able to deal with and produce at different stages of their progress.

In this thesis work, we explore automatic approaches for modeling linguistic complexity at different levels of learning Swedish as a second and foreign language (L2). For these purposes, we employ natural language processing techniques to extract linguistic features and combine them with machine learning methods. We study linguistic complexity in two types of L2 texts: those written by experts for learners and those produced by learners themselves. Moreover, we investigate this type of data-driven analysis for the smaller unit of sentences.

Automatic proficiency level prediction has a number of application potentials for the field of Intelligent Computer-Assisted Language Learning, out of which we investigate two directions. Firstly, this can facilitate locating learning materials suitable for L2 learners from corpora, which are valuable and easily accessible examples of authentic language use. We propose a framework for selecting sentences suitable as exercise items which, besides linguistic complexity, encompasses a number of additional criteria such as well-formedness and independence from a larger textual context. An empirical evaluation of the system implemented using these criteria indicated its usefulness in an L2 instructional setting. Secondly, linguistic complexity analysis enables the automatic evaluation of L2 texts which, besides being helpful for preparing learning materials, can also be employed for assessing learners' writing. We show that models trained partly or entirely on reading texts can effectively predict the proficiency level of learner essays, especially if some learner errors are automatically corrected in a pre-processing step. Both the sentence selection and the L2 text evaluation systems have been made freely available on an online learning platform.

KEYWORDS: natural language processing, linguistic complexity, readability, CEFR, second language learning, corpus examples, text classification, machine learning, domain adaptation.

DISTRIBUTION:
Department of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

Data linguistica 29
ISSN 0347-948X
ISBN 978-91-87850-68-4
GUPEA <http://hdl.handle.net/2077/55895>

PRINTED in Sweden by Repro Lorensberg Göteborg 2018