Data linguistica 31

Exploring Natural Language Processing for single-word and multi-word lexical complexity from a Second Language Learner perspective

av David Alfter

Akademisk avhandling för filosofie doktorsexamen
i språkvetenskaplig databehandling,
som enligt beslut av humanistiska fakultetsnämnden
vid Göteborgs universitet kommer att försvaras offentligt

GÖTEBORGS UNIVERSITET

Göteborg 2021

TITLE: Exploring Natural Language Processing for single-word and multi-word lexical complexity from a Second Language Learner perspective
LANGUAGE: English
AUTHOR: David Alfter

Abstract

In this thesis, we investigate how natural language processing (NLP) tools and techniques can be applied to vocabulary aimed at second language learners of Swedish in order to classify vocabulary items into different proficiency levels suitable for learners of different levels.

In the first part, we use feature-engineering to represent words as vectors and feed these vectors into machine learning algorithms in order to (1) learn CEFR labels from the input data and (2) predict the CEFR level of unseen words. Our experiments corroborate the finding that feature-based classification models using 'traditional' machine learning still outperform deep learning architectures in the task of deciding how complex a word is.

In the second part, we use crowdsourcing as a technique to generate ranked lists of multi-word expressions using both experts and non-experts (i.e. language learners). Our experiment shows that non-expert and expert rankings are highly correlated, suggesting that non-expert *intuition* can be seen as on-par with expert *knowledge*, at least in the chosen experimental configuration.

The main practical output of this research comes in two forms: prototypes and resources. We have implemented various prototype applications for (1) the automatic prediction of words based on the feature-engineering machine learning method, (2) practical implementations of language learning applications using graded word lists, and (3) an annotation tool for the manual annotation of expressions across a variety of linguistic factors. As for the resource side, we have started the creation of a sense-based graded vocabulary list, further enriched with data linked from various other sources as well as manual linguistic annotation across multiple linguistic features.

KEYWORDS: natural language processing, lexical complexity, vocabulary, second language learners, automatic complexity prediction, CEFR, linguistic resources, machine learning.

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

Data linguistica 31
ISSN 0347-948X
ISBN 978-91-87850-79-0
GUPEA <http://hdl.handle.net/2077/66861>

PRINTED in Sweden by Stema Specialtryck AB, Borås 2021