Automatic CEFR Level Prediction for Estonian Learner Text

Sowmya Vajjala and Kaidi Lõo

The 3rd workshop on NLP for computer-assisted language learning
Swedish Language Technology Conference
Uppsala, Sweden, 13-14 November 2014

What is the paper about?

- We developed an approach to predict the CEFR level of texts written by language learners in Estonian.
- It is a data-driven, machine learning approach
  - ... informed by linguistic knowledge (morphology, parts-of-speech etc.,)
  - ... uses publicly accessible data and tools.

Research Questions:

1. **Prediction**: How accurately can we predict the CEFR level for a learner text?
2. **Understanding**: What linguistic properties are more prominent between proficiency levels?

Why do we want to do this?

1. It’s useful in real-life.
    - for placement tests at a language teaching institute
    - as a feedback aid to students learning a new language
    - in exams like GRE, TOEFL etc.,

Why do we want to do this?

1. It's useful in real-life.
   - for placement tests at a language teaching institute
   - as a feedback aid to students learning a new language
   - in exams like GRE, TOEFL etc.,
2. We can learn more about how people learn a second language. E.g.,
   - Do learners struggle with morphology in the beginning?
   - As proficiency increases, does lexical proficiency increase or decrease?

Why do we want to do this?

1. It's useful in real-life.
   - for placement tests at a language teaching institute
   - as a feedback aid to students learning a new language
   - in exams like GRE, TOEFL etc.,
2. We can learn more about how people learn a second language. E.g.,
   - Do learners struggle with morphology in the beginning?
   - As proficiency increases, does lexical proficiency increase or decrease?
3. ...and of course, its fun!

Estonian primer

- Estonian is agglutinative. Word forms can be formed by joining the morphemes together.
  - e.g., *jalgades –>jalga+de+s* (stem for foot +plural marker+inessive case marker)
- It is fusional i.e., word forms can be formed by changing the stem.
  - e.g., *jalg* (foot, nominative), *jala* (genitive), *jalga* (partitive)
- It has 14 productive cases (grammatical and semantic cases).
  - Cases express relations between words and are sometimes used instead of postpositions (*jalal* and *jala peal* have the same meaning: *on the foot*)
- Cases have different alternative case endings.
  - e.g., Valid allative plural forms for *jalg* (foot) are: *jalgadele, jalule, jalgele*

our features rely on these properties of the language.

Results - Summary

- We get a classification accuracy of 79%, with a feature set consisting of 78 features.
- We reach almost the same accuracy, with a smaller subset of 27 features.
- There seems to be a lot of correlation between the most predictive features though.
- Comparing classification and regression, we find classification better.
- Morphological features are more prominent between A2,B1 and B2,C1 but not B1,B2.

... to know more, visit our poster!