EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LEAD
Graduate School

INTERSECTION 4: LANGUAGE AND LINGUISTICS

Seminar Für Sprachwissenschaft

Text Readability and Automatic Simplification

Connecting Linguistics, Processing and Educational Appllications

Research Questions in the Thesis

1. Can we analyze the the linguistic complexity of texts automatically?
2. Can we apply this analysis in Educational contexts?
3. Will text complexity affect a learner's cognitive processing and performance outcomes?
4. Can we automatically simplify texts to a reading level appropriate to the learner?

Q1: Automatic Assessment of Text Complexity

- **Task:** Predicting the appropriate grade level for a text
- **Methods**:
  1. Texts: Texts written for learners at various grade levels
  2. Linguistic Variables studied:
     - Lexical: e.g., lexical density, type-token ratio
     - Syntactic: e.g., dependent clauses per sentence
     - Morphological: e.g., complexity of a word
     - Psycholinguistic: e.g., age of acquisition of words
  3. Modeling: machine learning (e.g., linear regression, support vector classification)
  4. Evaluation: how correctly can the approach predict the reading level of a text?
- **Results**:
  1. Our readability model is accurate, generalizable across texts and genres. (correlation: 0.9)
  2. It is also the second best model in comparison with 6 other existing academic and commercial systems.

(Vajjala & Meurers, 2012; 2013; 2014a; 2014b)

Q2: Application in Educational Contexts

1. **Reading Demands Project (German)**:
   - **Aim:** Analyze the differences in linguistic complexity of German textbooks between schools, grades.
   - **Results:**

(a) We can identify grade-wise and school-wise differences for certain linguistic features.

(b) For predictive models, there are significant differences between publishers and schools.

(c) Prediction is better at school level than grade level.

(funded by a LEAD Intramural Research Grant)

2. **Proficiency Classification (English)**:
   - **Aim:** Automatically detect the language proficiency of English learners, using linguistic complexity features.
   - **Results:** Performance is comparable with existing research on this topic, on publicly accessible datasets.

Q3: Impact of Text Complexity on Readers

- **Aim:** Understand how text complexity affects a learner's cognitive processing and performance.
- **Methods:** Eye tracking, recall-comprehension questions and generalized additive mixed models.
- **Results:**
  1. Fixation count, second pass reading time and recall are significantly affected by text difficulty.
  2. Subject's language proficiency interacts with text complexity for all the above processes.
  3. Comprehension questions performance is dependent only on subject's language proficiency.

(funded by a LEAD Intramural Research Grant)

Q4: Automatic Text Simplification

- **Aim:** perform text simplification as monolingual machine translation (using Moses toolkit).
- **Method:** use readability models for choosing difficult sentences to simplify. (Vajjala & Meurers, in prep.)
- **Example Simplification from my approach:**
  - *original:* Hyper inflation is a condition where prices **increase rapidly** as a **currency** loses its value.
  - *simplified:* Hyper inflation is a condition where prices increase **very fast** as **money** loses its value.

References

- Sowmya Vajjala and Detmar Meurers, in prep. Readability based sentence ranking for evaluating text simplification.
- Sowmya Vajjala and Detmar Meurers, 2014. On assessing the reading level of individual sentences for text simplification, In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pages 288-297.
- Sowmya Vajjala and Detmar Meurers, 2014a. Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. In: Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics 165:2. (pp. 194-222).
- Sowmya Vajjala and Detmar Meurers, 2014b. Exploring Measures of "Readability"for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In Proceedings of PITR workshop, EACL 2014, pages 21-29.
- Sowmya Vajjala and Detmar Meurers, 2013. On The Applicability of Readability Models to Web Texts. In Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). pages 59-68.
- Sowmya Vajjala and Detmar Meurers, 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7), Association for Computational Linguistics. 163-173.

Sowmya Vajjala
LEAD Graduate School
Gartenstr. 29 · Tübingen · Germany
sowmya@sfs.uni-tuebingen.de

Seminar Für Sprachwissenschaft
Wilhelmstr. 19-23 · Tübingen · Germany
http://sfs.uni-tuebingen.de/~svajjala