Exploring Measures of 'Readability' for Spoken Language:
Analyzing linguistic features of subtitles to identify age-specific TV programs

Sowmya Vajjala and Detmar Meurers
LEAD Graduate School and Department of Linguistics, University of Tübingen

In a Nutshell

- We investigate if features from readability assessment can be used to identify age-specific TV programs
  - based on a corpus of BBC subtitles,
  - using a text classification approach.
- We show that the authentic material targeting specific age groups exhibits
  - a range of linguistic and psycholinguistic characteristics
  - that are indicative of the complexity of the language used.
- We achieve an accuracy of 95.9% (three-class task).

Motivation

- Reading, listening and watching TV are all ways to obtain information.
  - Some TV programs are created for particular age-groups (similar to graded readers).
- Audio-visual presentation and language are important factors in making age-specific TV programs.
- Is language by itself characteristic of the targeted age-group?
  - We hypothesize that the linguistic complexity of the subtitles is a good predictor.
  - We explore this hypothesis using features from automatic readability assessment.

Corpus

- The BBC started subtitling all scheduled programs on its main channels in 2008.
- Van Heuven et al. (2014) compiled a subtitles corpus from nine BBC TV channels.
- Subtitles of four channels are annotated: CBeebies, CBBC, News and Parliament.
- BBC subtitles corpus in numbers:

Program Category	Age group	# texts	avg. tokens per text	avg. sent. len. (in words)
CBEEBIES	$<$ 6 years	4846	1144	4.9
CBBC	6–12 years	4840	2710	6.7
Adults (News + Parliament)	$>$ 12 years	3776	4182	12.9

- We use a balanced subset consisting of 3776 texts per class.

Features

- Lexical Features
  - lexical richness features from Second Language Acquisition (SLA) research
  - part-of-speech density features
  - traditional features and formulae

  (extracted using Stanford Tagger)
- Syntactic Features
  - syntactic complexity features from SLA research.
  - other parse tree features

  (extracted using Berkeley parser and Tregex pattern matcher)
- Morphological properties of words (from Celex database)
- Age-of-Acquisition (AoA) features from various norms from (Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012)
- abstractness and other word characteristics from the MRC psycholinguistic database
- Avg. number of senses per word (obtained from WordNet)

Setup and Experiments

- We explored several classification algorithms in WEKA: SMO, J48 decision tree, Logistic Regression, etc.
  - SMO marginally outperformed the others (1–1.5%), so all further experiments were performed using SMO.
- **Classification Accuracy for different feature sets:**

Features (#)	Accuracy
Sentence Length baseline (1)	71.4%
All Features (152)	95.9%

- **Accuracy for different feature selection methods:**

CfsSubsetEval (41)	93.9%
Information Gain, top-10 features (10)	84.5%
CfsSubsetEval on top-10 features (6)	84.1%

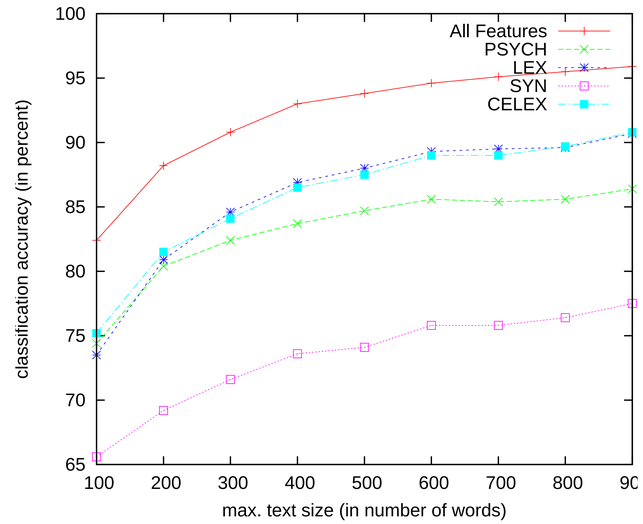
- **Confusion matrix for the model with all 152 features:**

classified as →	CBeebies	CBBC	Adults
CBeebies	3619	156	1
CBBC	214	3526	36
Adults	2	58	3716

Ablation Tests

Features	Acc.	SD
All − AoA_Kup_Lem	95.9%	0.37
All − All AoA Features	95.6%	0.58
All − PSYCH	95.8%	0.31
All − CELEX	94.7%*	0.51
All − CELEX − PSYCH	93.6%*	0.66
SYNTAX only (All − CELEX − PSYCH − LEX)	77.5%*	0.99
LEX	93.1%*	0.70
CELEX	90.0%*	0.79
PSYCH	84.5%*	1.12

Effect of Text Size on Accuracy



Conclusions

- The rich (psycho)linguistic feature set performs very well, achieving a classification accuracy of 95.9%.
- Single most predictive feature: AoA (82.4%), but removing this feature does not affect the accuracy.

⇒ The age-specific nature of authentic material is reflected in a wide range of (psycho)lingusitic properties.

- For practical tasks, accuracies above 90% can also be achieved with feature subsets and relatively short texts.

Outlook:

- Explore the impact of a parser tuned to spoken language.
- Perform more qualitative error analysis to identify where the approach fails and why.