Readability Assessment for Sentences
Motivation, Methods and Evaluation

Sowmya Vajjala
(with Detmar Meurers)

Center for Language Technology
University of Gothenburg, Sweden
20 November 2014
What is readability analysis?

We want to measure **how difficult** it is to read a text,

- based on **properties of the text**, using criteria which are
  - data-induced: using corpora with graded texts
  - theory-driven: constructs known to reflect complexity

- given a **purpose**, e.g.,
  - humans: to support reading and comprehension
    - read texts at a specific level of language proficiency.
    - carry out specific tasks (e.g., answer questions) etc.,
  - machines: evaluation of generation systems

- sometimes personalized to a user, through information-
  - obtained directly (e.g., questionnaire), or
  - indirectly (e.g., inferred from nature of a search query)
Why do we need readability for sentences?

some application scenarios

- selecting appropriate sentences for language learners in CALL. (Segler 2007; Pilán et al. 2013, 2014)
- understanding the difficulty of survey questions (Lenzner 2013)
- predicting sentence fluency in Machine Translation (Chae & Nenkova 2009)
- for text simplification (Vajjala & Meurers 2014a; Dell’Orletta et al. 2014)
Why do WE need it?

- identifying target sentences for text simplification.
- evaluating text simplification approaches.
Our Approach: Overview

- **Corpus**: publicly accessible, sentence level corpora (texts not prepared by us)
- **Features**: from Vajjala & Meurers (2014b), that work well at a text level.
- **Modeling**:
  1. binary classification (easy vs difficult)
  2. apply document level regression model on sentences.
  3. pair-wise ranking
- **Evaluation**: within and cross corpus evaluations with multiple real-life datasets

- Zhu et al. (2010) created a publicly available, sentence aligned corpus from Wikipedia and Simple Wikipedia.
- ~80,000 pairs of sentences in simplified and unsimplified versions.
- Example pair:
  1. Wiki: *Chinese styles vary greatly from era to era and are traditionally named after the ruling dynasty.*
  2. Simple Wiki: *There are many Chinese artistic styles, which are usually named after the ruling dynasty.*
Corpora-2: OneStopEnglish.com

- OneStopEnglish (OSE) is an English teachers resource website published by the Macmillan Education Group.
- They publish Weekly News Lessons which consist of news articles sourced from *The Guardian*.
- The articles are rewritten by teaching experts for English language learners at three reading levels (elementary, intermediate, advanced)
- We obtained permission to collect articles and compiled a corpus of 76 article triplets (228 in total)
Corpora-2: OneStopEnglish.com
sentence aligned corpus creation

▶ creation process:
1. parse the pdf files and extract text content.
2. split all texts into sentences.
3. compare sentences between versions and match them by their cosine similarity (Nelken & Shieber 2006).

▶ two versions of the corpus:
1. OSE2Corpus: ~ 3000 sentence pairs.
2. OSE3Corpus: ~ 850 sentence triplets.

* contact me if anyone wants to use this corpus.
OSE Corpus: Example

**adv:** In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.

**inter:** In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.

**ele:** In Beijing, people went to the Apple Store with flowers and candles.
Features-1
From Vajjala & Meurers (2014b)

- **Lexical Features**
  - lexical richness features from Second Language Acquisition (SLA) research
    - e.g., Type-Token ratio, noun variation, ...
  - POS density features
    - e.g., # nouns/# words, # adverbs/# words, ...
  - traditional features and formulae
    - e.g., # sentence length in words ...

- **Syntactic Features**
  - syntactic complexity features from SLA research.
    - e.g., # dep. clauses/clause, average clause length, ...
  - other parse tree features
    - e.g., # NPs per sentence, avg. parse tree height, ...
Features -2

- Morphological properties of words
  - e.g., Does the word contain a stem along with an affix? \(\text{abundant} = \text{abound} + \text{ant}\)

- Age of Acquisition (AoA)
  - average age-of-acquisition of words in a text

- Other Psycholinguistic features
  - e.g., word abstractness

- Avg. number of senses per word (obtained from WordNet)
We started with training a sentence-level readability model on Wikipedia corpus:
- Binary classification: simple – hard
  - 65–68% accuracy, depending on training set size.
  - increasing training sample size from 10K to 80K samples did not improve the accuracy much!
- As regression: \( r = 0.4 \)
- Why is it so bad?
What is the problem?

- What happens if we just apply a document level readability model on this corpus?
- Model (Vajjala & Meurers 2014b): outputs readability score on a scale of 1-5, 5 being difficult.
What can we infer?

- There are all sorts of sentences in both versions.
- Wikipedia has more sentences at higher reading levels than Simple Wikipedia.

- Is this the reason binary classification failed?

- one idea: A simple sentence is only simpler than its unsimplified version. It can also still be hard.
  ⇒ Simplification could be relative, not absolute.
Is Simplification Relative?
How can we study this?

-One approach:
   - compute reading levels of normal (N) and simplified (S) sentences using our document level readability model.
   - evaluate simplification classification using the percentages of S<N, S=N and S>N
   - the higher the percentage for S<N, the better the model is, at evaluating sentence level readability.
     - Why?: Simplified versions are expected to be at a lower reading level than Normal versions!!

- How big must |S−N| be to interpret it as a categorical difference in reading level?
  → We call this the $d$-value.
What exactly is *d*-value?

- It is a measure of how fine-grained the model is in identifying reading-level differences between sentences.

- For example, let us say \( d = 0.3 \).
  - Now, when \( N = 3.4, S = 3.2 \), \( |S - N| = 0.2, <d \) \( \Rightarrow S = N \).
  - If \( N=3.5, S=3.1 \), \( |S - N| = 0.4, >d \) \( \Rightarrow S < N \).

- What is good for us?: the model should be able to identify as many pairs as possible as \( S < N \).

- \( S = N \) is probably okay, but \( S > N \) is bad.
Influence of $d$

Question 1: Does changing the $d$-value affect our results?

![Graph showing the influence of d-value on percentage of total samples.](image)

desired scenario: percentage of $(S<N) > (S=N) > (S>N)$
Influence of $N$

- Question 2: How does the reading level of the unsimplified sentence (N) affect the results?
Influence of $N$

- Question 2: How does the reading level of the unsimplified sentence ($N$) affect the results?

For harder sentences (when $N \geq 2.5$)

For easier sentences (when $N < 2.5$)
What do we learn from these graphs?

- The accuracy of relative comparison of reading levels of sentences depends on:
  1. minimum $|S-N|$ required to identify a categorical difference ($d$).
  2. reading level of the original, unsimplified sentence ($N$).

- It is difficult to identify simplifications for an already simple sentence.

- But, this approach works well for complex sentences.

*more details about this: Vajjala & Meurers (2014a).
What Next?

- How about modeling this as pair-wise ranking instead?
- Why?
  1. We do not have exact reading level annotations at sentence level.
  2. But we know that the simplified version should have a lower reading level.
- ranking cares only about relative differences, not absolute differences.
  ⇒ perhaps, an ideal learning method for this problem?
Pairwise Ranking: A Primer

- typically used in information retrieval, to rank search results by doing a pair-wise comparison between them.
- learning goal: minimize the number of ordering errors and mis-classified pairs.
- usual purpose: look at a pair of documents and rank them based on their relevance to the query.
- our purpose: learn a binary classifier that can tell which sentence is simpler, given a pair of sentences.
Pairwise Ranking: Evaluation

- errors: percentage of reversed pairs.
  ⇒ accuracy: percentage of correctly ranked pairs.

- if there are two sentences
  (N - unsimplified, S - simplified),
  ▶ rank(S) > rank(N), it is counted as an error.

- if there are three sentences (A, I, B), and our system gives a readability ranking: I,B,A
  ⇒ there are two ranking errors here.
  1. I is ranked higher than A.
  2. B is ranked higher than A.

note: There are other measures of efficiency for ranking, that are tailored to information retrieval applications.
Our Approach with Ranking

Train-Test Data setup

- Training sets:
  2. OSE2-Train: 2000 pairs of sentences from OSE corpus (advanced → beginner, advanced → intermediate, and intermediate → beginner combinations.)
  3. OSE3-Train: 750 sentence triplets from OSE corpus (each triplet has a single sentence in 3 versions).
  4. WikiOSE: mixed training set, consisting of Wiki-Train and OSE2-Train. (size: 4000 pairs)

- Test sets:
  1. Wiki-Test: 78000 pairs.
  2. OSE2-Test: 1000 pairs.
  3. OSE3-Test: 100 triplets.
Results

- algorithm: SVM<sup>rank</sup>

- results:
  - training with 2-level datasets
    | Training   | WIKI-TRAIN | OSE2-TRAIN |
    |------------|------------|------------|
    | WIKI-TEST  | 81.8%      | 77.5%      |
    | OSE2-TEST  | 74.6%      | 81.5%      |
    | OSE3-TEST  | 74.7%      | 79.3%      |
  
  - training with a 3 level corpus and a mixed corpus.
    | Training   | OSE-3LEVEL-750 | WIKI OSE |
    |------------|-----------------|----------|
    | WIKI-TEST-78KPAIRS | 78.6%      | 81.3%    |
    | OSE2-TEST  | 82.4%          | 80.7%    |
    | OSE3-TEST  | 79.6%          | 84.0%    |
How does this compare with the previous approach?

Ranking vs Regression

1. with regression model, depending on $d$-value, we were:
   - able to predict correctly in $\sim 60\%$ of the cases.
   - identified no difference between the sentence versions in $\sim 10\%$ of the cases.
   - identified the order wrongly in $\sim 30\%$ of the cases.

2. ranking approach:
   - We can predict the order correctly with $\sim 80\%$ accuracy.
   - works with multiple datasets and levels of simplification.

$\Rightarrow$ clearly, ranking is working better than regression.
What features work with Ranking?


- performance of feature groups:
  1. Psycholinguistic Features. - 69.1%
  2. Syntactic Complexity Features - 67.1%
  3. Celex Features - 72.2%
  4. Lexical Richness Features - 67%

- around 55-60% accuracy with good single features.
Conclusions so far

- Readability based ranking works well in distinguishing simplified and unsimplified versions of a sentence.
- Features that worked at document level work well on sentences too with good accuracy.
- The approach can also make distinctions between multiple levels without losing on the performance.
- We get good results with cross-corpus and mixed-corpus evaluations too.
  ⇒ It's fairly generalizable to other texts that informational in nature (e.g., news, encyclopaedia articles etc.,)
Current and Future Work

- What feature selection approaches work for ranking?
- What linguistic properties change the most between Advanced to Intermediate vs Intermediate to Beginner?
- How do we eliminate correlated features?
- Using with actual automatic text-simplification approaches to evaluate them in a data-driven manner.
- ...
End of Story!

- Thank you for your patience :-)
- Questions?
- email: sowmya@sfs.uni-tuebingen.de
References


