KEEP IT SIMPLE: TEXT SIMPLIFICATION USING SIMPLE ENGLISH WIKIPEDIA

HTTP://READABILITY.CRYDEE.EU/

(HTTP://READABILITY.CRYDEE.EU/)

Internship defense — Hugo Mougard
July the 10th

OVERVIEW

1. Definition: what is readability?
2. Research statement
3. Related works
4. Approach
5. Future works
6. Conclusion

BUT FIRST, WHAT IT IS NOT

Readability is not legibility: the former is only about the text, not the layout nor the appearance.

MEDIUM
HIGH LEGIBILITY

Right tool for the job

Like all things, it depends. If it's a complex intelligence algorithm that requires high concurrency—sure, something besides Node can be fine... as a service... that I can call from Node. As far as measuring concurrency and speed between Go and Node, it would the equivalent of comparing whether a *for* loop performs better going backwards or forward. These kind of microbenchmarks don't appeal to me. When choosing a programming language or framework or library, always consider the ROI for your product and how effective you'll be between your users, your team, and *yourself*.

The Moral

I don't want this to a very long retrospective—So, similar to TJ's advice—there are lots of awesome solutions out there; pick one you will do well in and will ultimately make your users, company, and yourself *happy*.

CDISCOUNT

LOW LEGIBILITY

SIMPLE ENGLISH WIKIPEDIA, *COMMODORE NUTT*

HIGH READABILITY

Nutt toured the world between 1869 and 1872 with the Thumbs and Lavinia's sister, Minnie Warren. They returned to America rich after performing before royalty. Nutt left Barnum's employ after a disagreement with the showman. He toured with a comic opera company, put together a variety show on the United States West Coast, and operated saloons in Oregon and California. He returned to New York City, and died there of Bright's disease in May 1881.

JAMES JOYCE, *ULYSSES*

LOW READABILITY

It soared, a bird, it held its flight, a swift pure cry, soar silver orb it leaped serene, speeding, sustained, to come, don't spin it out too long long breath he breath long life, soaring high, high resplendent, aflame, crowned, high in the effulgence symbolistic, high, of the ethereal bosom, high, of the high vast irradiation everywhere all soaring all around about the all, the endlessnessnessness...

DEFINITION OF READABILITY

*Readability is the **ease** with which text can be **read** and **understood**.*

— Wikipedia

READABILITY FACTORS

READER RELATED

- **understanding**: background knowledge, language
- **reading**: reading fluency, language

READABILITY FACTORS

TEXT RELATED

- **reading**: syntax, vocabulary
- **understanding**: syntax, vocabulary, idea density, cognitive load

IN THIS WORK

Focus on **vocabulary** aspect of readability.

OVERVIEW

1. Definition: what is readability?
2. Research statement
3. Related works
4. Approach
5. Future works
6. Conclusion

FINE-GRAINED
READABILITY ANALYSIS

Most approaches consider the document as a **whole**.

We want to be more specific.

END GOAL

Be able to find which words or sentences to rewrite, and how.

OVERVIEW

1. Definition: what is readability?
2. Research statement
3. Related works
4. Approach
5. Future works
6. Conclusion

RELATED WORKS

1. Readability formulas
2. Machine learning
3. Simple Wikipedia

EARLY READABILITY RESEARCH

20th century research was centered around formulas to estimate if a text is readable ot not.

AN EFFICIENT SET OF FEATURES

Most formulas use a combination of:

- average number of words per sentence
- average number of syllables per word
- presence of the word in a list of easy words

AN EXAMPLE

Dale–Chall readability formula (Dale and Chall, 1949):

$$0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$$

RELATED WORKS

1. Readability formulas
2. Machine learning
3. Simple Wikipedia

THE TASKS

Classical task	Machine learning task
Score a text	Regression
Sort texts on readability	Regression, classification on pairs of documents
Assign a required grade to a text	Classification with grades as labels
Regroup texts of similar readability	Clustering

LANGUAGE MODEL APPROACH

Schwarm and Ostendorf, 2005:

- bigrams and trigrams LM alone
- combination in a SVM of:
  - LM perplexities
  - readability formulas
  - syntactic features

COMPLEX FEATURES

Pitler and Nenkova, 2008:

- unigram model
- lexical cohesion (cosine similiarity averaged over all sentences)
- syntactic features (as Schwarm and Ostendorf)
- entity coherence (analyse the subjects / objects of consecutive sentences)
- language model over discourse relations

→ proves the superiority of discourse relations over average lengths of sentences and words. **But** discourse relations are not yet easily computable.

RELATED WORKS

1. Readability formulas
2. Machine learning
3. Simple Wikipedia

SIMPLE ENGLISH WIKIPEDIA (SEW)

- Wikipedia written in simple english
- goal is:
  - to use only the 1000 most common words in English
  - to keep sentences short
- 100 000 articles

TRANSFORMATION EXAMPLE

From the “Baseball uniform” pages:

- **SEW:** On April 4, 1849, the New York Knickerbockers became the first team to use uniforms.
- **EW:** The New York Knickerbockers were the first baseball team to wear uniforms, taking the field on April 4, 1849 in pants made of blue wool, white flannel shirts and straw hats.

COMPARABLE CORPORA

EW ~|| SEW

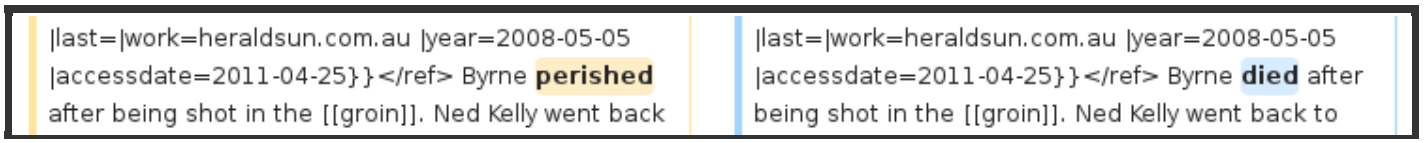
(Zhu et al., 2010)

- align EW and SEW versions of a same article
- gather the differences to compute a non-readable → readable corpus
- 100 000 pairs of sentences

REVISION HISTORY

(Yatskar et al., 2010)

- use EW and SEW revision histories
- gather differences between two consecutive revisions if the modification is about readability



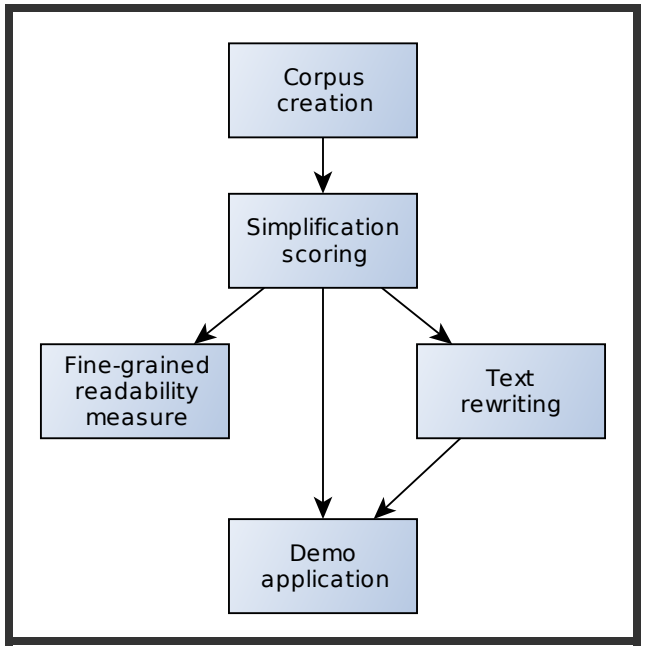
OVERVIEW

1. Definition: what is readability?
2. Research statement
3. Related works
4. Approach
5. Future works
6. Conclusion

APPROACH

1. Overview
2. Readability Lab
3. Corpus creation
4. Simplifications scoring
5. Fine-grained readability measure
6. Text rewriting

OVERVIEW


Corpus creation
Simplification scoring
Fine-grained readability measure
Text rewriting
Demo application

APPROACH

1. Overview
2. Readability Lab
3. Corpus creation
4. Simplifications scoring
5. Fine-grained readability measure
6. Text rewriting

READABILITY LAB

http://readability.crydee.eu/

- experiment with our approach
- publicly available
- source code on Github

APPROACH

1. Overview
2. Readability Lab
3. Corpus creation
4. Simplifications scoring
5. Fine-grained readability measure
6. Text rewriting

CORPUS CREATION

MOTIVATION

Free readability corpora are based on comparable corpora.
We propose a free corpus based on revision history.

Reasons:

- general process
- easily extensible outside of wikipedia (copy-editing)

CORPUS CREATION

METHODOLOGY

We use previously known methods:

- SEW edit history
- align the sentences with a `diff` program

The corpus itself and the related tooling are freely available

CORPUS METRICS

~36 000 entries. ~25 000 occur only once and ~18 000 / ~21 000 originals have only one readable equivalent.

APPROACH

1. Overview
2. Readability Lab
3. Corpus creation
4. Simplifications scoring
5. Fine-grained readability measure
6. Text rewriting

SIMPLIFICATIONS DICTIONARY CREATION

Objective is to score the translations in the corpus we created: with $\mathcal{P}$ the set of English phrases, go from a corpus $\mathcal{C} \subset \mathcal{P} \times \mathcal{P}$ to a simplification dictionary $\mathcal{D} \subset \mathcal{P} \times \mathcal{P} \times \mathbb{R}$.

$\rightarrow$ need to define score functions $\mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$

PROPERTIES OF GOOD SCORES

We are looking for good lexical simplifications. When we score the simplification (s, t):

- the less common s, the higher $\mathcal{S}(s, t)$
  marvelous → good **>** ok → good
- the more common (s, t) in $\mathcal{C}$, the higher $\mathcal{S}(s, t)$
- the more common t, the higher $\mathcal{S}(s, t)$
  marvelous → good > marvelous → wonderful

HOW RARE IS s ?

Direct language model score is not enough ($P_{LM}\left(s\right)$):

- "exhausted" would likely have a higher LM score than "I am"
- we want to rewrite "exhausted", not "I am"

→ average by s length: $\sqrt[|s|]{P_{LM}\left(s\right)}$

HOW COMMON IS (s, t) IN $\mathcal{C}$?

We can answer with two probabilities:

1. $P_{\mathcal{C}}\big((s,t)\big) = \frac{|\{(s,t) \mid (s,t) \in \mathcal{C}\}|}{|\mathcal{C}|}$
2. $P_{\mathcal{C}}\big((s,t)\big|s\big) = \frac{|\{(s,t) \mid (s,t) \in \mathcal{C}\}|}{|\{(s,t') \mid t' \in \mathcal{P} \wedge (s,t') \in \mathcal{C}\}|}$

Scores using (2.) have c index, for $c$onditional.

HOW COMMON IS t ?

Two worthy definitions:

1. $P_{LM}(t)$ to allow only for short simplifications
2. $\sqrt[|t|]{P_{LM}(t)}$ to allow for paraphrases

Scores using any of (1.) or (2.) have a d index, for $d$ouble language model.

Scores using (2.) have a w index, for weighted

APPROACH

1. Overview
2. Readability Lab
3. Corpus creation
4. Simplifications scoring
5. Fine-grained readability measure
6. Text rewriting

A GENERAL FRAMEWORK

Recursive combination of readability scores with 3 functions:

- π to handle the different scores of a sentence **p**art
- σ to handle the parts of a **s**entence
- θ to handle the sentences of a **t**ext

$$f(t) = \theta\left(\left\{ \sigma\left(\{\pi(s) | x \subset sent \wedge (x, y, s) \in \mathcal{D}\}\right) \middle| sent \text{ is a sentence in } t \right\} \right)$$

AN APPLICATION

We construct f_{max} with:

- $\pi = \max$
- $\sigma = \max$
- $\theta = \text{average}$

in English: f_{max} averages the maximum of the sentences lexical improvement scores to assess the readability of a text.

π SCORING

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

↓

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

σ SCORING

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

↓

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

θ SCORING

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

↓

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

INTEREST

- fine-grained analysis when needed
- still usable as a readability score for the complete text

APPROACH

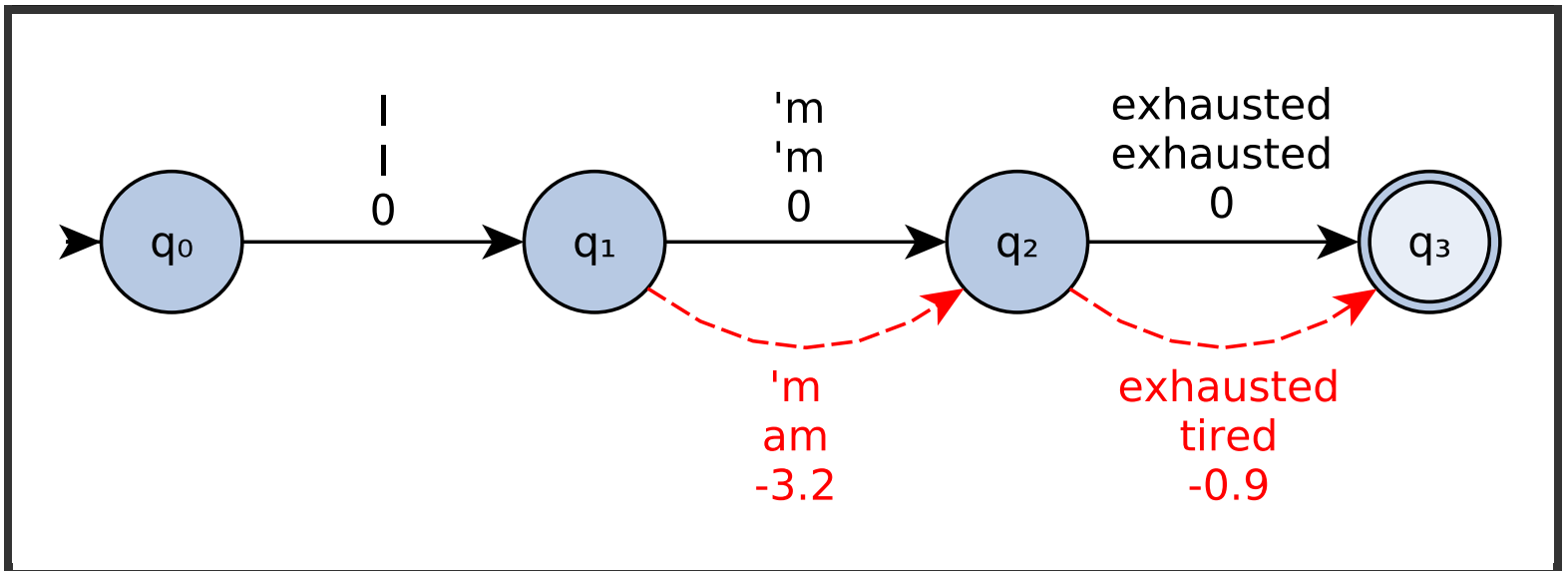
1. Overview
2. Readability Lab
3. Corpus creation
4. Simplifications scoring
5. Fine-grained readability measure
6. Text rewriting

TEXT REWRITING

We can use our lexicon to rewrite text: with **weighted transducers**.

WEIGHTED TRANSDUCERS

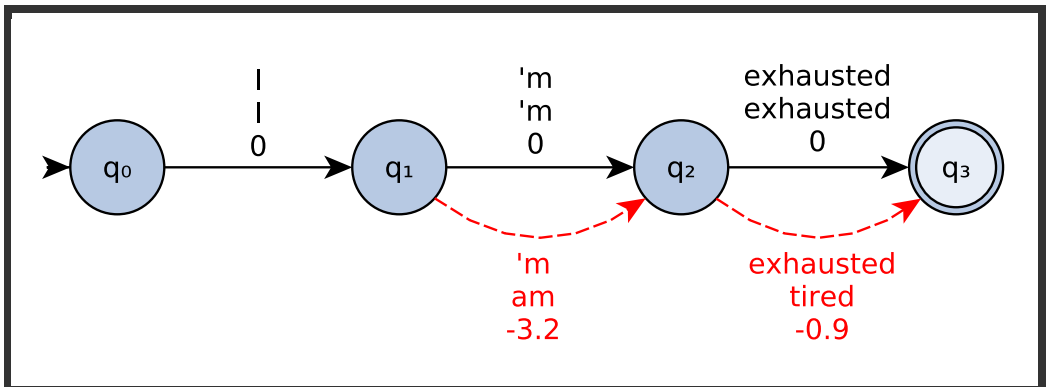
EXAMPLE



WEIGHTED TRANSDUCERS

COMBINING WEIGHTS

score with +	score with min	output
0	0	I'm exhausted
-0.9	-0.9	I'm tired
-3.2	-3.2	I am exhausted
-4.1	-3.2	I am tired



OVERVIEW

1. Definition: what is readability?
2. Research statement
3. Related works
4. Approach
5. Future works
6. Conclusion

AUTOMATIC EVALUATION

- evaluate simplification scoring by comparing agreement of top simplifications with a gold
- compare new readability measures with ML approaches and readability formulas:
  - compute their correlation
  - compare them on a gold corpus

MANUAL EVALUATION

Randomize output of our top simplifications with a SotA system and ask human judges to decide which make the most sense.

SYNTACTIC REWRITING

Go from string → string transducing to tree → tree transducing to handle syntactic rewritings.

OVERVIEW

1. Definition: what is readability?
2. Research statement
3. Related works
4. Approach
5. Future works
6. Conclusion

CONCLUSION

Contributions:

- a new readability corpus and related tools
- a scoring method for simplifications
- a way to derive a fine-grained readability measure from it
- how to also apply it to text rewriting
- a tool to conquer them all, Readability Lab

SLIGHTLY OFF-TOPIC ♥

THANK YOU VERY MUCH FOR YOUR ATTENTION

DO YOU HAVE ANY QUESTION?

HTTP://READABILITY.CRYDEE.EU/
(HTTP://READABILITY.CRYDEE.EU/)

WEIGHTED TRANSDUCERS

DEFINITION

- Σ input alphabet
- Δ output alphabet
- Q set of states
- $I \subseteq Q$ set of initial states
- $F \subseteq Q$ set of final states
- $E \subseteq Q \times (\Sigma \cup \{\varepsilon\}) \times (\Delta \cup \{\varepsilon\}) \times \mathbb{K} \times Q$ set of transitions
- $\lambda : I \to \mathbb{K}$ the initial weight function
- $\rho : F' \to \mathbb{K}$ the final weight function mapping F' to $\mathbb{K}$

WEIGHTED TRANSDUCERS
CONSTRUCTION

For a text of length n and its tokens $x_1, \ldots, x_n$:

- $\Sigma = \mathcal{T}$
- $\Delta = \mathcal{T}$
- $Q = \{q_i | 0 \leq i \leq n\}$
- $I = \{q_0\}$
- $F = \{q_n\}$
- $E = \{(q_{i-1}, x_i, y, s, q_i) | (x_i, y, s) \in \mathcal{D} \wedge 1 \leq i \leq n\}$
  $\cup \{(q_{i-1}, x_i, x_i, 1, q_i) | 1 \leq i \leq n\}$
- $\lambda : x \mapsto 1$
- $\rho : x \mapsto 1$

SCORE $\mathcal{S}$

$$\mathcal{S}(s,t) = \log \frac{P_{\mathcal{D}}\big((s,t)\big)^{\lambda_1}}{\sqrt[|s|]{P_{LM}(s)}^{\lambda_2}}$$

$$= \lambda_1 \log P_{\mathcal{D}}\big((s,t)\big) - \lambda_2 \log \sqrt[|s|]{P_{LM}(s)}$$

SCORE $\mathcal{S}_c$

c FOR **C**ONDITIONAL

$$\mathcal{S}(s,t) = \log \frac{P_{\mathcal{D}}\Big((s,t) \,\Big|\, s\Big)^{\lambda_1}}{\sqrt[|s|]{P_{LM}\Big(s\Big)}^{\lambda_2}}$$

SCORE $\mathcal{S}_d$

d FOR **D**OUBLE LANGUAGE MODEL

$$\mathcal{S}(s,t) = \log \frac{P_{\mathcal{D}}\Big((s,t)\Big)^{\lambda_1} P_{LM}\Big(t\Big)^{\lambda_3}}{\sqrt[|s|]{P_{LM}\Big(s\Big)}^{\lambda_2}}$$

SCORE $\boldsymbol{S}_{dc}$

$$\mathcal{S}(s,t) = \log \frac{P_{\mathcal{D}}\Big((s,t) \,\Big|\, s\Big)^{\lambda_1} P_{LM}\Big(t\Big)^{\lambda_3}}{\sqrt[|s|]{P_{LM}\Big(s\Big)}^{\lambda_2}}$$

SCORE $\boldsymbol{S}_{wd}$

w FOR **W**EIGHTED REPLACEMENT

$$\mathcal{S}(s,t) = \log \frac{P_{\mathcal{D}}\Big((s,t)\Big)^{\lambda_1} \sqrt[|t|]{P_{LM}\Big(t\Big)}^{\lambda_3}}{\sqrt[|s|]{P_{LM}\Big(s\Big)}^{\lambda_2}}$$

SCORE $\boldsymbol{S}_{wdc}$

$$\mathcal{S}(s,t) = \log \frac{P_{\mathcal{D}}\Big((s,t) \,\Big|\, s\Big)^{\lambda_1} \sqrt[|t|]{P_{LM}\Big(t\Big)}^{\lambda_3}}{\sqrt[|s|]{P_{LM}\Big(s\Big)}^{\lambda_2}}$$