N-gram Fragment Sequence Based Unsupervised Domain-Specific Document Readability

Shoaib Jameel, Xiaojun Qian, Wai Lam

The Chinese University of Hong Kong
Outline

1. Introduction and Motivation
   1. The problem of readability
   2. Why is it important in web search?

2. Related Work
   1. Heuristic readability methods
   2. Supervised readability methods
   3. Unsupervised readability methods

3. Overview of our model
   1. Background
   2. Sequential N-gram Connection Model (SNCM)
      1. SNCM1
      2. SNCM2 - An extended model

4. Empirical Evaluation

5. Conclusions and Future Directions
The Problem of Readability

- Readability is the ease with which humans can understand a piece of textual discourse
- For example, consider the following two text snippets:

Snippet 1: Source → ScienceForKids website
A proton is a tiny particle, smaller than an atom. Protons are too small to see, even with an electron microscope, but we know they must be there because that’s the only way we can explain how atoms behave. To give you an idea how small a proton is, if an atom was the size of a football stadium, then a proton would still be smaller than a marble.

Snippet 2: Source → English Wikipedia
The proton is a subatomic particle with the symbol p or p+ and a positive electric charge of 1 elementary charge. One or more protons are present in the nucleus of each atom. The number of protons in each atom is its atomic number.
Why readability is important in web search?

- Users not only want documents which are a good match to their queries but also want documents which they can comprehend.
- Partially understood in Information Retrieval.
- Current assumption is that all users are alike “one-size-fit-all” scheme.
- For example, for the query *proton*, Google currently ranks a document from the Wikipedia in the top position.
- Users thus have to reformulate query several times.
- Will certainly hurt the user in the end i.e. user will be dissatisfied.
Illustration of the query **proton** in Google

![Google Search for proton](image)

**Proton** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Proton - Cached
The proton is a subatomic particle with the symbol \( p \) or \( \mathbf{p} \) and a positive electric charge of 1 elementary charge. One or more protons are present in the nucleus...

**Proton therapy** - Proton decay - Proton (disambiguation) - Antiproton

**PROTON** - Home
www.proton.com/- Cached
Live the PROTON lifestyle. Visit our online shop for exclusive merchandise. © 2012 Perusahaan Otomobil Nasional Sdn. Bhd. All rights reserved. Follow us:

**PROTON Edar - Home**
proton-edar.com.my/- Cached
... Putrajaya, Sabah, Sarawak, Selangor, Terengganu. Warranty, Service Centres. You are here: Home © 2012 PROTON Edar Sdn. Bhd. All rights reserved.
Special Promotions - Service Centres - Overview - Dealer Locator

**Home | Proton Cars SA**
www.protoncars.co.za/- Cached
Official Proton site; Visit our showroom, vehicle information, specifications, features, pricing, find a dealer, parts and service.

**Proton Radio**
www.protonradio.com/- Cached
Streaming the best in electronic music live and On Demand. Breaks, Minimal, Progressive, House, Techno, Electro, and more on 200+ exclusive shows!

**ProT on Europe Innovation from Public Research - Home - ProT on...**
www.protoneurope.org/- Cached
ProT on Europe, the European Knowledge Transfer Association, created in 2003 by the European Commission and self-supporting since 2007.
An attempt by Google
Related Work

- General heuristic readability methods
  - Readability formulae such as Flesch Kincaid
- Supervised learning methods
  - Language Modeling
  - Support Vector Machines
  - Query log mining and building individual user profile
  - Computational Linguistics
- Unsupervised learning methods
  - Terrain based method
  - Domain-specific readability methods
  - Vector-space based methods
Related Work
General Heuristic Readability Methods

- Very old - existed since 1940’s
- Conjectured that two components play a major role in finding reading difficulty of texts
  - Syntactic component - sentence length, word length, number of sentences etc.
  - Semantic component - number of syllables per word etc.
- Manually tuned parameters
- Simple to apply
- Works very well on general texts [Kevyn and Callan, JASIST - 2005] but fails on web pages and domain-specific documents [Yan et al., CIKM - 2004]
An example of a readability formula
Flesh-Kincaid (F-K) readability method

F-K Formula

\[
206.835 - 1.015 \times \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \times \left( \frac{\text{total syllables}}{\text{total words}} \right)
\]

1. Syntactic component \( \rightarrow \left( \frac{\text{total words}}{\text{total sentences}} \right) \)
2. Semantic component \( \rightarrow \left( \frac{\text{total syllables}}{\text{total words}} \right) \)
3. Numerical values are manually tuned after repeated experiments

Where does it fail?
- water \( \rightarrow \) 2 syllables (wa-ter)
- embryology \( \rightarrow \) 5 syllables (em-bry-ol-o-gy)
- star \( \rightarrow \) 1 syllable (which star??)
Related Work

Supervised Learning Methods

Smoothed Unigram Model

1. Deal in American grade levels
2. The basic model is a unigram language model with smoothing
3. Define a generative model for a passage

Unigram Language Model

\[
L(T|G_i) = \sum_{w \in V} C(w) \log P(w|G_i)
\]

where,

- \( T \) is some small passage
- \( L(T|G_i) \) is the log likelihood of a passage belonging to some grade
- \( V \) is the number of words in that passage
- \( w \) is a word in the passage \( T \)
- \( C(w) \) is the number of tokens with type \( w \) in the passage \( T \)
Related Work
Matching queries with users

- Also deal in American grade levels [Liu et al., SIGIR - 2004]
- Used readability features to train the classifier i.e. SVM
- Separate queries based on reading levels
- In the end, they conclude that SVM based method helps better segregate queries based on reading levels

Limitation of supervised methods
Requires extensive amount of training data, which might be expensive and time consuming to obtain
Kate et al., [Kate et al., COLING - 2010] found that language model features play an important role in determining readability of texts.

Pitler and Nenkova, [Pitler and Nenkova, EMNLP - 2008] found that average sentence length and word features are strong features for a classifier.
Related Work

Domain-specific readability methods

- Compute readability in a completely unsupervised fashion
- But they require some external knowledge based for detect domain-specific terms in documents [Yan et al., CIKM - 2006] and [Zhao and Kan, JCDL - 2010]
- Our previous terrain based [Jameel et al., CIKM - 2011] method does not require any ontology or lexicon but considers only unigrams in determining the reading difficulty of texts
The Idea of Cohesion and Scope

- **Document cohesion** is a state or quality that the elements of a text tend to “hang together” [Morris and Hirst, CL - 1991]
- When units of texts are cohesive then the text is readable [Kintsch, Psy. Review - 1988]
- **Document Scope** [Yan et al., CIKM - 2006] refers to the coverage of the concepts (i.e. domain-specific terms)
- Lesser the scope (coverage), more difficult the term.
Our Methodology - An Overview

- Our method is based on automatically finding appropriate n-gram in the Latent Semantic Indexing latent concept space.
- In the latent concept space, n-grams which are central to a document come close to their document vectors and general/common n-grams move far from the document vector.
- We introduce the notion of n-gram specificity.
- We denote the sequence of unigrams in a document \( d \) as \((t_1, t_2, \cdots, t_W)\).
- We form n-grams from this sequence which we denote as \( S = (s_1, s_2, s_3, s_4) \).
- Our motive is two-fold:
  1. Automatic n-gram determination
  2. Compute cost in n-gram formation considering cohesion and specificity (we use specificity in contrast to Document Scope)
Sequential N-gram Connection Model

Notion of n-gram Specificity

We compute specificity by computing cosine similarity between the vectors (NOTE: term and document vectors) in the low dimensional latent concept space

- Central n-grams will come close to their document vectors in the latent concept space
- These central terms in domain-specific documents are mainly domain-specific terms

Computation of n-gram Specificity

Let $s$ be an n-gram fragment. Let $d$ be the document where this n-gram fragment occurs. Let this fragment be represented as a vector in the LSI latent space as $\vec{s}$ and the document vector as $\vec{d}$. We compute the n-gram specificity, $\vartheta(\vec{s}, \vec{d})$ as $\vartheta(\vec{s}, \vec{d}) = \text{cosine_sim}(\vec{s}, \vec{d})$
Sequential N-gram Connection Model
Notion of n-gram Cohesion

We compute cohesion also by computing cosine similarity between (NOTE: two consecutive n-gram vectors) in the latent concept space

▶ If two terms are semantically related to each other i.e. they are cohesive then their vectors will be close to each other in the latent concept space
▶ Their cosine similarities will be high
▶ Other way to look at - they co-occur very often in the collection

Computation of n-gram Cohesion

Suppose $T = (t_1, t_2, \cdots, t_W)$ is the term sequence and $S = (s_1, s_2, \cdots, s_K)$ is one particular n-gram fragmented sequence of $T$. Cohesion is computed as: $\eta(\vec{s}_i, \vec{s}_{i+1}) = \text{cosine}_\text{sim}(\vec{s}_i, \vec{s}_{i+1})$
Our first model: SNCM1

Determine a least cost (which is a readability cost) n-gram connected sequence in the document where at each forward transition sequential n-gram cohesion is minimized. The cost of the n-gram fragment sequence $S$, $C^{(d)}_1(S)$:

$$
C^{(d)}_1(S) = \sum_{k=1}^{K} \left( \frac{1}{\eta(s_{k-1}, s_k)+1} \right)
$$

Our goal is to minimize this cost, $C^{(d)}_1(S)$ and we achieve this using the following optimization scheme:

$$\min_S C^{(d)}_1(S)$$

- We use dynamic programming to find the optimal cost
- The minimized cost obtained at the end of entire document path traversal is a readability cost that a reader expends in order to read the document
We define $C^{(d)}_1(T_i)$ as the optimal cost from the beginning until the term $t_i$ in the document.

$$C^{(d)}_1(T_i) = \text{minimum} \left( C^{(d)}_1(T_{i-1}) + \frac{1}{\eta(S_{X-1}, S_X) + 1}, \right.$$ 

$$\left. C^{(d)}_1(T_{i-2}) + \frac{1}{\eta(S_{Y-1}, S_Y) + 1}, \right.$$ 

$$\ldots,$$ 

$$\ldots,$$ 

$$C^{(d)}_1(T_{i-m}) + \frac{1}{\eta(S_{Z-1}, S_Z) + 1} \right) \tag{1}$$

where,

- $\vec{S}_X$ be a unigram composed of $t_i$
- $\vec{S}_Y$ be a bigram composed of $(t_{i-1}, t_i)$
- $\vec{S}_Z$ be an $m$-gram composed of $(t_{i-m+1}, \ldots, t_i)$
- $S_{X-1}, S_{Y-1} \text{ and } S_{Z-1}$ represent the particular $n$-gram (where $n$ may be from 1 to $m$) in the optimal sequential path that appears just before $\vec{S}_X, S_Y \text{ and } S_Z$ respectively.
Final Readability Cost

We linearly combine specificity values of the n-grams formed during sequential linear n-gram determination scheme

\[ E_1^{(d)} = \alpha C_1^{(d)}(T_W) + (1-\alpha) \sum_{i=1}^{K} s_i(d) \]

where,

- \( \alpha \) (0 \leq \alpha \leq 1) is a parameter controlling the relative contribution of cohesion and specificity
- \( W \) is the total number of terms in the document

Note

- A higher cost indicates that the document is difficult to read and a low cost is indicative of the ease in reading the document
- We shall use the cost values to re-rank the search results obtained from a general purpose IR system
Now, we combine the effect of both cohesion and specificity

\[ C_2^{(d)}(S) = \sum_{k=1}^{K} \left( \beta \psi(\vec{s}_k, \vec{d}) + (1 - \beta) \frac{1}{\eta(s_{k-1}, s_k)+1} \right) \]

where, \( \beta \) (0 \leq \beta \leq 1) is a parameter controlling the relative weights of the two components

Our objective now

\[ \min_S C_2^{(d)}(S) \]
Apply similar dynamic programming

Let the optimal cost for all the terms from \( t_1 \) until position \( t_i \) be \( C_2^{(d)}(T_i) \)

\[
C_2^{(d)}(T_i) = \text{minimum} \left( C_2^{(d)}(T_{i-1}) + \beta \vartheta(\vec{S}_X, \vec{d}) + (1 - \beta) \frac{1}{\eta(S_{X-1}, \vec{S}_X + 1)}, \\
C_2^{(d)}(T_{i-2}) + \beta \vartheta(\vec{S}_Y, \vec{d}) + (1 - \beta) \frac{1}{\eta(S_{Y-1}, \vec{S}_Y + 1)}, \\
\ldots, \\
\ldots, \\
C_2^{(d)}(T_{i-m}) + \beta \vartheta(\vec{S}_Z, \vec{d}) + (1 - \beta) \frac{1}{\eta(S_{Z-1}, \vec{S}_Z + 1)} \right) \tag{2}
\]
and, we rank documents based on

\[ E_2^{(d)} = \frac{C_2^{(d)}(T_W)}{W} \]
Empirical Evaluation

Testbed Data

- We chose two popular domains
  - Science
  - Psychology

- Our test collection contains

  **Psychology**
  - Documents = 170,000
  - n-grams in vocabulary = 154,512

  **Science**
  - Documents = 300,000
  - n-grams in vocabulary = 490,770

- Prepared two sets of data - stopwords kept and removed
Indexing and Retrieval

- Used Zettair\(^1\) to index web pages
- Retrieval using Okapi BM25 ranking function
- Selected top-\(k\) documents (in our case \(k=10\))
- Re-ranked the documents based on the costs obtained from our model and also re-ranked them using scores obtained from other comparative methods
- Topics were created by two humans by following INEX\(^2\) topic creation guidelines

\(^1\)http://www.seg.rmit.edu.au/zettair/
\(^2\)http://www.inex.otago.ac.nz/tracks/adhoc/gtd.asp
Annotations and Metrics

- Asked two human annotators to annotate documents based on the following

Annotation Guidelines

- 0 → very low domain-specific readability
- 1 → reasonably low domain-specific readability
- 2 → average domain-specific readability
- 3 → reasonably high domain-specific readability
- 4 → very high domain-specific readability

- Cohen’s kappa ≈ 0.8

NDCG: Normalized Discounted Cumulative Gain

\[ W(q_s) = \frac{1}{Z_n} \sum_{i=1}^{n} \frac{2r(i) - 1}{\log(1+i)} \]
Results ($\alpha = \beta = 0.5$), $m = 3$, SVD factors=200

(a) Psychology

<table>
<thead>
<tr>
<th></th>
<th>NDCG@3</th>
<th>NDCG@5</th>
<th>NDCG@7</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>0.515</td>
<td>0.548</td>
<td>0.582</td>
<td>0.618</td>
</tr>
<tr>
<td>C-L</td>
<td>0.525</td>
<td>0.553</td>
<td>0.584</td>
<td>0.612</td>
</tr>
<tr>
<td>Flesch</td>
<td>0.449</td>
<td>0.490</td>
<td>0.537</td>
<td>0.579</td>
</tr>
<tr>
<td>Fog</td>
<td>0.513</td>
<td>0.547</td>
<td>0.577</td>
<td>0.612</td>
</tr>
<tr>
<td>LIX</td>
<td>0.516</td>
<td>0.550</td>
<td>0.584</td>
<td>0.619</td>
</tr>
<tr>
<td>SMOG</td>
<td>0.517</td>
<td>0.550</td>
<td>0.579</td>
<td>0.616</td>
</tr>
<tr>
<td>CHM</td>
<td>0.465</td>
<td>0.456</td>
<td>0.473</td>
<td>0.482</td>
</tr>
<tr>
<td>Counts</td>
<td>0.551</td>
<td>0.575</td>
<td>0.603</td>
<td>0.649</td>
</tr>
<tr>
<td>MLF</td>
<td>0.530</td>
<td>0.554</td>
<td>0.581</td>
<td>0.631</td>
</tr>
<tr>
<td>%UNK</td>
<td>0.558</td>
<td>0.585</td>
<td>0.611</td>
<td>0.653</td>
</tr>
<tr>
<td>SNCM1</td>
<td>0.537</td>
<td>0.571</td>
<td>0.602</td>
<td>0.651</td>
</tr>
<tr>
<td>SNCM2</td>
<td>0.581*</td>
<td>0.607*</td>
<td>0.635*</td>
<td>0.680*</td>
</tr>
</tbody>
</table>

(b) Science

<table>
<thead>
<tr>
<th></th>
<th>NDCG@3</th>
<th>NDCG@5</th>
<th>NDCG@7</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>0.524</td>
<td>0.547</td>
<td>0.562</td>
<td>0.564</td>
</tr>
<tr>
<td>C-L</td>
<td>0.541</td>
<td>0.551</td>
<td>0.572</td>
<td>0.576</td>
</tr>
<tr>
<td>Flesch</td>
<td>0.554</td>
<td>0.560</td>
<td>0.566</td>
<td>0.574</td>
</tr>
<tr>
<td>Fog</td>
<td>0.593</td>
<td>0.508</td>
<td>0.538</td>
<td>0.640</td>
</tr>
<tr>
<td>LIX</td>
<td>0.541</td>
<td>0.562</td>
<td>0.583</td>
<td>0.585</td>
</tr>
<tr>
<td>SMOG</td>
<td>0.584</td>
<td>0.538</td>
<td>0.500</td>
<td>0.523</td>
</tr>
<tr>
<td>CHM</td>
<td>0.400</td>
<td>0.406</td>
<td>0.407</td>
<td>0.412</td>
</tr>
<tr>
<td>Counts</td>
<td>0.595</td>
<td>0.563</td>
<td>0.564</td>
<td>0.627</td>
</tr>
<tr>
<td>MLF</td>
<td>0.557</td>
<td>0.584</td>
<td>0.611</td>
<td>0.657</td>
</tr>
<tr>
<td>%UNK</td>
<td>0.562</td>
<td>0.590</td>
<td>0.619</td>
<td>0.660</td>
</tr>
<tr>
<td>SNCM1</td>
<td>0.617*</td>
<td>0.645*</td>
<td>0.672*</td>
<td>0.713*</td>
</tr>
<tr>
<td>SNCM2</td>
<td>0.602*</td>
<td>0.625*</td>
<td>0.650*</td>
<td>0.702*</td>
</tr>
</tbody>
</table>

Table: Comparison of SNCM variants when $\alpha = \beta = 0.5$ against the comparative methods in both domains. * denotes statistically significant results for all comparisons according to paired t-test ($p < 0.05$). Stopwords are kept in these results.
Query-wise improvements

(a) Psychology

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Queries Improved SNCM1</th>
<th>Queries Improved SNCM2</th>
<th>Average Improvement SNCM1</th>
<th>Average Improvement SNCM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>53</td>
<td>59</td>
<td>17.56%</td>
<td>18.06%</td>
</tr>
<tr>
<td>C-L</td>
<td>61</td>
<td>61</td>
<td>22.84%</td>
<td>22.86%</td>
</tr>
<tr>
<td>Flesch</td>
<td>65</td>
<td>65</td>
<td>25.66%</td>
<td>25.66%</td>
</tr>
<tr>
<td>Fog</td>
<td>68</td>
<td>65</td>
<td>20.02%</td>
<td>17.12%</td>
</tr>
<tr>
<td>LIX</td>
<td>60</td>
<td>62</td>
<td>22.05%</td>
<td>24.03%</td>
</tr>
<tr>
<td>SMOG</td>
<td>58</td>
<td>60</td>
<td>23%</td>
<td>23.08%</td>
</tr>
<tr>
<td>CHM</td>
<td>86</td>
<td>88</td>
<td>36%</td>
<td>38%</td>
</tr>
<tr>
<td>Counts</td>
<td>29</td>
<td>40</td>
<td>1.02%</td>
<td>12.05%</td>
</tr>
<tr>
<td>MLF</td>
<td>49</td>
<td>60</td>
<td>2.01%</td>
<td>20.76%</td>
</tr>
<tr>
<td>%UNK</td>
<td>3</td>
<td>32</td>
<td>0</td>
<td>9.34%</td>
</tr>
</tbody>
</table>

(b) Science

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Queries Improved SNCM1</th>
<th>Queries Improved SNCM2</th>
<th>Average Improvement SNCM1</th>
<th>Average Improvement SNCM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>95</td>
<td>95</td>
<td>22.34%</td>
<td>22.01%</td>
</tr>
<tr>
<td>C-L</td>
<td>90</td>
<td>91</td>
<td>20.12%</td>
<td>20.36%</td>
</tr>
<tr>
<td>Flesch</td>
<td>92</td>
<td>92</td>
<td>21.56%</td>
<td>21.50%</td>
</tr>
<tr>
<td>Fog</td>
<td>80</td>
<td>80</td>
<td>17.90%</td>
<td>17.90%</td>
</tr>
<tr>
<td>LIX</td>
<td>90</td>
<td>90</td>
<td>20.19%</td>
<td>20.13%</td>
</tr>
<tr>
<td>SMOG</td>
<td>92</td>
<td>92</td>
<td>25.56%</td>
<td>26%</td>
</tr>
<tr>
<td>CHM</td>
<td>121</td>
<td>119</td>
<td>32%</td>
<td>29.99%</td>
</tr>
<tr>
<td>Counts</td>
<td>82</td>
<td>79</td>
<td>19.76%</td>
<td>17.55%</td>
</tr>
<tr>
<td>MLF</td>
<td>83</td>
<td>75</td>
<td>21.45%</td>
<td>19.23%</td>
</tr>
<tr>
<td>%UNK</td>
<td>77</td>
<td>69</td>
<td>17.55%</td>
<td>16.53%</td>
</tr>
</tbody>
</table>

Table: Performance comparison based on queries for SNCM1 and SNCM2.
Conclusions and Future Work

- Our unsupervised domain-specific readability ranking model that does not require any external knowledge-base.
- We find n-grams in documents based on an optimization scheme.
- Our results indicate an improvement over the state-of-the-art.

In the future....

- How hyperlink structure of the web can aid in readability?
- Can other observable features such web page fonts, layout etc. help in determining readability of documents?
References

References