Admin

Ethics discussion on Wednesday: read papers beforehand

Project status update due Wednesday

Next Monday, Quiz #4 (comprehensive)

Next Wednesday, presentations

Text simplification

Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius and a lot of courage to move in the opposite direction.

- E. F. Schumacher

Goal:

Reduce the reading complexity of a sentence by incorporating more accessible vocabulary and sentence structure while maintaining the content.
Simplify

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

The reverse process, producing electrical energy from mechanical energy, is accomplished by a generator or dynamo.

Text simplification: real examples

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

Alfonso Perez is a former Spanish football player.

Deletion

Text simplification: real examples

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

Alfonso Perez is a former Spanish football player.

Rewording
Endemic types or species are especially likely to develop on islands because of their geographical isolation.

Endemic types are most likely to develop on islands because they are isolated.

The reverse process, producing electrical energy from mechanical energy, is accomplished by a generator or dynamo.

A dynamo or an electric generator does the reverse: it changes mechanical movement into electric energy.
Text simplification: real examples

The reverse process, producing electrical energy from mechanical energy, is accomplished by a generator or dynamo.

A dynamo or an electric generator does the reverse: it changes mechanical movement into electric energy.

Goals today

Introduce the text simplification problem

Highlight why text simplification is important

Show some examples of text simplification approaches

Give one perspective on CS research

Why text simplification?

A lot of text data is available

Problem: much of this content is written above the reading level of many people
Adult literacy

Below Basic: no more than the most simple and concrete literacy skills
Basic: can perform simple and everyday literacy activities
Intermediate: can perform moderately challenging literacy activities
Proficient: can perform complex and challenging literacy activities

Why text simplification?

Broader availability of standard text resources
- language learners
- people with aphasia or other cognitive disabilities
- children

Broader availability of domain-specific text resources
- health and medical documents
  - 90M Americans (over a quarter!) do not have sufficient health literacy to understand currently provided materials
  - Cost of low health literacy is estimated to be hundreds of billions
- academic papers
- legal documents

Goals today

Introduce the text simplification problem ✓
Highlight why text simplification is important ✓
Show some examples of text simplification approaches
Give one perspective on CS research

Spectrum of solutions

Focus on these types of approaches today

writer assist tools/resources
- readability formulas
- simple word lists
- flag difficult text sections
- simplification thesauruses
- rule-based with human verification
- …

manual semi-automated fully automated
**How do we identify difficult words?**

**Quantifying word difficulty**

**Hypothesis:**

The more often a person sees a word, the more familiar they are with that word, and therefore the simpler it is.

**Proxy for “how often you see a word”:**

Frequency on the web!

**Validating the frequency hypothesis**

Google: ~13M unique “words”

Does the frequency of these words relate to people’s knowledge/familiarity with these words?
Validating the frequency hypothesis

Google: ~13M unique “words”

11 bins based on frequency:
1%, 10%, 20%, ..., 100%

- randomly pick 25 words from each bin

- 275 words

Annotate with definition

marmorean:

- a) crimson-and-grey songbird that inhabits town walls and mountain cliffs of southern Eurasia and northern Africa
- b) of or relating to or characteristic of marble
- c) the most common protein in muscle
- d) a color or shade

random definitions from other words in data set

Study participants

50 participants per word * 275 words = 13,750 total annotations!
Frequency correlates with understanding!

How do we identify difficult words?

Medical Text Simplification Tool

Given data (paired sentences) learn a simplification model

Spectrum of solutions

Most modern approaches are data-driven

- readability formulas
- simple word lists
- flag difficult text sections
- simplification thesauruses
- rule-based with human check
- ...

unified
simplified

Given data (paired sentences) learn a simplification model

4/24/23

9
From aligned documents to aligned sentences

E minor (G₄, Min) is a minor scale based on the note E. The E natural minor scale {E, F, G, A, B, C, D} contains the natural 7, D, rather than the flatted 7, C♯ to align with the major dominant chord, E7 (E, B, F). Its key signature has one sharp, F (see below: Scales and keys).

No relative major is G major, and its parallel major is E major.

Most of the electric guitar repertoire is in E minor, as it is a very natural key for the instrument. In standard tuning (E, A, D, G, B, E), four of the instruments six 'top' (unstretched) strings are part of the tonic chord. The key of G minor is also extremely popular in heavy metal music, as its tonic is the lowest note on a standard tuned guitar.

E minor (G₄, Min) is a minor scale based on the note E. Its key signature has one sharp, F if its relative major is G major.

A lot of classical guitar music is in E minor, because this key is very suited for this instrument. When it is tuned normally, four of the instrument's six strings are part of the tonic chord. The key is also very popular in heavy metal music, because the lowest note on a guitar, E, can be used a lot.

E minor was one of the most often used keys by Felix Mendelssohn.
From aligned documents to aligned sentences

How could you do this?

Simplification approaches

Many different data-driven approaches

- Lexical (change a word at a time)
- Phrasal (change phrases)
- Syntactic (use grammatical structure)
- Neural networks

Lexical simplification

The ACL was established in 1962.

The ACL was started in 1962.

Simplification is accomplished by changing one word (or phrase) at a time.
Preprocessing

The first school was established in 1857

The first school was started in 1857

The district was established in 1993 by merging …

The district was made in 1993 by joining …

Automatically word-align sentences

Extract candidate simplifications

The first school was established in 1857

The first school was started in 1857

The district was established in 1993 by merging …

The district was made in 1993 by joining …

extract aligned candidate word pairs:
- different words
- same part of speech
- not in a list of common words (stoplist)

Simplification rules learned

<table>
<thead>
<tr>
<th>word</th>
<th>candidate simplifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>abolish</td>
<td>remove, replace, stop</td>
</tr>
<tr>
<td>established</td>
<td>began, made, settled, started</td>
</tr>
<tr>
<td>merging</td>
<td>becoming, joining</td>
</tr>
</tbody>
</table>

... 

Learned simplification rules for 14,478 words

On average 2.25 candidate simplifications

Not all rules apply in all contexts

The ACL was established in 1962.

The ACL was began in 1962. ✗

The ACL was made in 1962. ?

The ACL was settled in 1962. ✗

The ACL was started in 1962. ✓
Data for learning context

Enter a simpler word that could be substituted for the red, bold word in the sentence. A simpler word is one that would be understood by more people or people with a lower reading level (e.g. children).

Food is procured with its suckers and then crushed using its tough “beak” of chitin.

Collected data for 500 words

<table>
<thead>
<tr>
<th>simplification</th>
<th># of people that suggested simplification (out of 50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>obtained</td>
<td>17</td>
</tr>
<tr>
<td>gathered</td>
<td>9</td>
</tr>
<tr>
<td>gotten</td>
<td>8</td>
</tr>
<tr>
<td>grabbed</td>
<td>4</td>
</tr>
<tr>
<td>acquired</td>
<td>2</td>
</tr>
<tr>
<td>made</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Learning to apply rules

500 examples

Food is procured with its suckers and then crushed using its tough “beak” of chitin.

1. Food is obtained with its suckers and then crushed using its tough “beak” of chitin.
2. Food is gathered with its suckers and then crushed using its tough “beak” of chitin.
3. Food is gotten with its suckers and then crushed using its tough “beak” of chitin.
4. Food is grabbed with its suckers and then crushed using its tough “beak” of chitin.
...
Learning a ranker

1. Food is obtained with its suckers and then crushed using its tough "beak" of chitin.
2. Food is gathered with its suckers and then crushed using its tough "beak" of chitin.
3. Food is gotten with its suckers and then crushed using its tough "beak" of chitin.
4. Food is grabbed with its suckers and then crushed using its tough "beak" of chitin.

Applying the ranker

The ACL was started in 1962.
The ACL was made in 1962.
The ACL was settled in 1962.
The ACL was started in 1962.

Results

Previous approach:
- Coverage: 85% (of the words that could be changed)
- Accuracy: 54% (of the suggestions are correct)

Our approach:
- Coverage: 86%
- Accuracy: 76%

Simplification approaches

Many different data-driven approaches
- Lexical (change a word at a time)
- Phrasal (change phrases)
- Syntactic (use grammatical structure)
- Neural networks
Phrase-based sentence simplification

I disdain green ham with green eggs

Unsimplified sentence is probabilistically broken into phrases

“phrase” is a sequence of words

Phrase-based sentence simplification

Each phrase is probabilistically simplified
Learned phrase examples

<table>
<thead>
<tr>
<th>original</th>
<th>simple</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ham</td>
<td>ham</td>
<td>0.7</td>
</tr>
<tr>
<td>ham</td>
<td>pork</td>
<td>0.2</td>
</tr>
<tr>
<td>ham</td>
<td>meat</td>
<td>0.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>like to eat a variety</td>
<td>like to eat a variety</td>
<td>0.5</td>
</tr>
<tr>
<td>like to eat a variety</td>
<td>like to eat lots</td>
<td>0.3</td>
</tr>
<tr>
<td>like to eat a variety</td>
<td>like to eat many</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Learn these aligned phrases and probabilities from the aligned sentences

Phrase-based sentence simplification

Problem: does not account for phrasal deletion

63

64

I disdain green ham with green eggs

I do not like green eggs and ham
I do not like ham and green eggs
I do not like green eggs and green ham
I do not like green eggs with ham
I do not like eggs with ham

... Model is probabilistic and considers many, many variations!

Phrase-based sentence simplification

Problem: does not account for phrasal deletion

65

66
Phrase-based sentence simplification

We add phrasal deletion

I disdain the food green ham with green eggs
I do not like green eggs and ham

Each phrase is probabilistically simplified
Phrases can also be probabilistically deleted

Delete phrases

0.5% of learned phrases are deletions

<table>
<thead>
<tr>
<th>Phrase-table entry</th>
<th>probability of deletion</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>0.037</td>
</tr>
<tr>
<td>the</td>
<td>0.033</td>
</tr>
<tr>
<td>of the</td>
<td>0.0015</td>
</tr>
<tr>
<td>or</td>
<td>0.0014</td>
</tr>
<tr>
<td>however ,</td>
<td>0.00095</td>
</tr>
<tr>
<td>the city of</td>
<td>0.00034</td>
</tr>
<tr>
<td>generally</td>
<td>0.00033</td>
</tr>
<tr>
<td>approximately</td>
<td>0.00025</td>
</tr>
<tr>
<td>however ,</td>
<td>0.00022</td>
</tr>
<tr>
<td>etc</td>
<td>0.00013</td>
</tr>
</tbody>
</table>

Qualitatively: Phrase-based

Critical reception for The Wild has been negative.

Reviews for The Wild has been negative.

rewording

Qualitatively: Phrase-based

Bauska is a town in Bauska county, in the Zemgale region of southern Latvia.

Bauska is a town in Bauska county, in the region of Zemgale.

rewording/reordering, deletion
Nicolas Anelka is a French football player. He plays for Chelsea.

rewording, deletion, sentence splitting

Each edge of a tesseract is of the same length.

Same edge of the same length.

He often recuperated at Menton, near Nice, France, where he eventually died on 1892 January 31.

He died.

Compared to three previous systems:

**Pros:**
- phrase-based approach tends to be more similar to human simplifications than other approaches
- deletion improves the quality
- model is fairly easy to understand

**Cons:**
- tends to only make minor changes to the sentences
- some disfluencies due to long distance dependencies
Simplification approaches

Many different data-driven approaches
- Lexical (change a word at a time)
- Phrasal (change phrases)
- Syntactic (use grammatical structure)
- Neural networks

Syntax-based approach

Rather than operating on phrases, operate on grammar trees

Learn probabilistic, syntax-based rules

They may occasionally eat  

sometimes, they eat

Learn probabilistic, syntax-based rules

The scary cats from the park  
may occasionally walk around on two legs

sometimes, the scary cats from the park walk around on two legs
An aside

sometimes, the scary cats from the park walk around on two legs.

A syntax-based approach

Our life is frittered away by detail. Simplify, simplify.
- H.D. Thoreau

Our life is frittered away.
- Lab Machine 227-31

Qualitatively: syntax-based

Overall Bamberga is the tenth brightest main belt asteroid after, in order, Vesta, Pallas, Ceres, Iris, Hebe, Juno, Melpomene, Eunomia and Flora.

Syntax:
Overall Bamberga is the tenth brightest main belt asteroid.

Phrase-based: (same as input)

Quantitatively

Compared to phrase-based:

Pros:
- Much more significant simplifications
- More grammatical
- Simpler

Cons:
- Was sometimes aggressive about removing content
Goals today

Introduce the text simplification problem ✓

Highlight why text simplification is important ✓

Show some examples of text simplification approaches ✓

Give one perspective on CS research ✓

Future thoughts/challenges

What is simple?
- different domains may have different notion
- how do we measure/evaluate simplicity

Collaborators!

Will Coster (Pomona)
Dan Feblowitz (Pomona)
Melissa Grueter (Pomona)
Colby Horn (Middlebury)
Katie Manduca (Middlebury)
Max Schwarzer (Pomona)
Mui Tanprasert (Pomona)
Gondy Leroy (University of Arizona)

Course recap

Corpus analysis

Regex

Language modeling (smoothing!)

Linguistics basics

Learning grammars PCFGs

Parsing
Course recap

Text similarity
Word similarity
Machine translation
Word alignment
Machine learning (Naïve Bayes, SVMs)
Neural nets (basics, Word2Vec, large language models)

Course recap

HashMaps/dictionaries
Java
How to count words! 😊

Course recap

How many lines of code?

How many slides?