Admin

- Paper draft due 5pm Wednesday
  - Must be done with all of your experiments
  - "Results" section is required

- 1 hr quiz on Tuesday

Review

- **Corpus analysis**
- **Basic probability**
- **Language modeling**
  - n-gram language models
  - different smoothing techniques
- **Parsing**
  - CFG, PCFGs
  - CKY algorithm
  - Improved models
- **Text and word similarity**

Review

- **Machine translation**
  - MT basics
  - translation models
  - word alignment
- **Machine learning**
  - ML basics
  - NB (multinomial and Bernoulli)
  - smoothing
  - other models (k-NN, SVM)
- **NLP research topics**
  - text modeling
  - text simplification
- **High-level themes**
  - Probabilistic modeling and data-driven modeling
  - Evaluation
Course summary

- Number of assignments: 8 (4 "A" assignments)
- Number of labs: 4
- Pages read: 218
- Number of lines of code: 3,776
- Number of slides: 1,251

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.

Simpler is better.
Text simplification: real examples

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

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

What types of transformations are happening?

Deletion

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.

What types of transformations are happening?

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.

Deletion

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.

What types of transformations are happening?

Rewording
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.

- Deletion and rewording
- Insertion and reordering

Why text simplification?

A lot of text data is available

Problem: much of this content is written above many people’s reading level

Goals today

Introduce the text simplification problem

Understand why it’s important

Examine what makes text difficult/simple

Overview of approaches to text simplification
Adult literacy

<table>
<thead>
<tr>
<th>Below Basic:</th>
<th>no more than the most simple and concrete literacy skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic:</td>
<td>can perform simple and everyday literacy activities</td>
</tr>
<tr>
<td>Intermediate:</td>
<td>can perform moderately challenging literacy activities</td>
</tr>
<tr>
<td>Proficient:</td>
<td>can perform complex and challenging literacy activities</td>
</tr>
</tbody>
</table>

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 (at least a third!) 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

Why text simplification?

Make life easier for computers!

I find forest colored chicken ovum and smoked pork thigh to be dietarily disturbing.

I do not like green eggs and ham.

What makes text difficult/simple?
What makes text difficult/simple?

Lots of previous research going back decades!

Some ideas:
- vocabulary
- sentence structure/grammatical components
  - passive vs. active tense
  - use of relative clauses
  - compound nouns
  - nominalization (turning verbs into nouns)
- ...
- organization/flow

Quantifying text difficulty

- vocabulary
- sentence structure/grammatical components
  - passive vs. active tense
  - use of relative clauses
  - compound nouns
  - nominalization (turning verbs into nouns)
- ...
- organization/flow

How do we measure/quantify these things, particularly with minimal human intervention?

Quantifying word difficulty

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

Proxy for “how often you see a word”:
Frequency on the web!

Validating frequency hypothesis

Google unigrams: ~13M

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

randomly pick 25 words from each bin

275 words

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

Google unigrams: ~13M

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

randomly pick 25 words from each bin

275 words

Annotate with definition

Validating frequency hypothesis

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 woman policeman

random definitions from other words in data set

Study participants

50 participants per word =
- 1,250 annotations/frequency bin
- 13,750 total annotations!
Frequency correlates with understanding!

What does this tell us about simplifying text?

Avoid less frequent words. Use more frequent words.

Quantifying text difficulty

- vocabulary
- sentence structure / grammatical components
- passive vs. active tense
- use of relative clauses
- compound nouns
- nominalization (turning verbs into nouns)
- ...
- organization/flow

Still many, many aspects of language to explore...

Goals today

- Introduce the text simplification problem ✅
- Understand why it’s important ✅
- Examine what makes text difficult/simple ✅
- Overview of approaches to text simplification
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

A semi-automated approach

I disdain green chicken ovum and ham.

Identify difficult words

I disdain green chicken *ovum* and ham.

How can we do this?

A semi-automated approach

I disdain green chicken ovum and ham.

Identify difficult words

I disdain green chicken ovum and ham.

Based on word frequency!
(low-frequency words)

A semi-automated approach

I disdain green chicken ovum and ham.

Identify difficult words

I disdain green chicken ovum and ham.

generate candidate word simplifications from text resources (e.g. thesauruses, dictionaries, etc.)

Human annotator
A semi-automated approach

I disdain green chicken ovum and ham.

dislike
dislike
hate
hate
scorn
scorn
...
...

I do not like green eggs and ham.

Evaluation/experimentation

I disdain green chicken ovum and ham.

How do we tell if our system is useful?

An experiment

Examine if people’s learning and understanding improve with the simplified article

An experiment

Page 1:

Page 2:

Page 3:

answer the same questions again!

answer some different questions with the text

read one version of the article and

answer some questions related to

the article topic

original

original

Q1

Q4, Q5, Q6, ...

Q1

Q2

Q3

Q4

Q5

Q6

...
Results with the text: understanding (questions Q3, Q4, Q5, …)

Results without the text: learning (questions Q1, Q2, Q3, …)

Spectrum of solutions:
- readability formulas
- simple word lists
- flag difficult text sections
- simplification thesauruses
- rule-based with human check
- …

Data-driven approach:
- Given training data (paired sentences)
- learn a simplification model
Collecting simplification data

I took a speed reading course and read War and Peace in twenty minutes. It involves Russia.
— Woody Allen

Wikipedia for text simplification

“We use Simple English words and grammar here. The Simple English Wikipedia is for everyone! That includes children and adults who are learning English.”

Wikipedia for text simplification

“Simple does not mean little. Writing in Simple English means that simple words are used. It does not mean readers want simple information. Articles do not have to be short to be simple; expand articles, include a lot of information, but use basic vocabulary.”

Wikipedia for text simplification

4.4M articles

97K articles

unsimplified

simplified

I do not like green eggs and ham.
I find forest colored chicken ovum and pork rump to be dietarily disturbing.

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

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.
From aligned documents to aligned sentences

E minor (Em, Min) is a minor scale based on the note E. The E natural minor scale (E G A B C D E) consists of the pitches E, F, G, A, B, C, and D. The E harmonic minor scale (E F G A B C D E) contains the natural F, D, rather than the flat A. D – to align with the major dominant chord, DT (E G A B C D E).

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish football player. Alfonso Perez Munoz 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.

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

I do not like green eggs and ham. I find forest colored chicken eggs to be dietarily disturbing.

Wikipedia for text simplification

167K aligned sentence pairs

Simplification approaches
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

I disdain green ham with green eggs

Each phrase is probabilistically simplified (translation model)

Phrase-based sentence simplification

I disdain green ham with green eggs

I do not like ham and green eggs

Phrases are probabilistically reordered (language model)
I disdain the food green ham with green eggs
I do not like green eggs and ham

Why is that a problem here?

Problem: does not allow for phrasal deletion

We add phrasal deletion

Each phrase is probabilistically simplified (translation model)
\[ p(\text{NULL} | \text{the food}) \]
Phrase-based performance

Experiments

5 approaches
- **none** – output the unsimplified sentence
- **K&M** – noisy channel sentence compression with PCFGs
  - Only allows for deletion
  - Uses syntactic information
- **T3** – Cohn and Lapata (2009)
  - All transformation operations
  - Uses syntactic information
  - Only been previously employed for sentence compression
- **Moses** – noisy channel, phrase-based without deletion
- **Moses+Del** – with deletion

Evaluation

3 measures
- **BLEU** (0-1.0)
  - weighted mean of n-gram precisions
  - brevity penalty to avoid overly short results
- **word-F1** (0-1.0)
  - F1 measure of system word occurrences
  - F1 combines precision and recall into one measure
- **Simple String Accuracy - SSA** (0-1.0)
  - length normalized edit distance

Results

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>word-F1</th>
<th>SSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>0.5937</td>
<td>0.5967</td>
<td>0.6179</td>
</tr>
<tr>
<td>K&amp;M</td>
<td>0.4352</td>
<td>0.4352</td>
<td>0.4871</td>
</tr>
<tr>
<td>T3*</td>
<td>0.2437</td>
<td>0.2190</td>
<td>0.3651</td>
</tr>
<tr>
<td>Moses</td>
<td>0.5987</td>
<td>0.6076</td>
<td>0.6224</td>
</tr>
<tr>
<td>Moses+Del</td>
<td>0.6046</td>
<td>0.6149</td>
<td>0.6259</td>
</tr>
</tbody>
</table>

All results are significantly different at the p=0.01 level

* T3 was only trained on 30K sentence pairs
Results: phrasal systems

If we remove those sentence pairs from the test set that are identical:

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>0.4560</td>
</tr>
<tr>
<td>Moses</td>
<td>0.4723</td>
</tr>
<tr>
<td>Moses+Del</td>
<td>0.4752</td>
</tr>
</tbody>
</table>

Moses+Del results

In 8.5% of the test sentences deletion was used

<table>
<thead>
<tr>
<th>Case</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>output</td>
</tr>
<tr>
<td>Moses+DEL</td>
<td>correct change</td>
</tr>
<tr>
<td></td>
<td>incorrect change</td>
</tr>
</tbody>
</table>

Results separated by sentence pairs that were different ("correct change") and those that were the same and did not require any simplification ("incorrect change")

Qualitatively: Phrase-based

Critical reception for The Wild has been negative.
Reviews for The Wild has been negative.

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
Qualitatively: Phrase-based

Nicolas Anelka is a French footballer who currently plays as a striker for Chelsea in the English premier league.

Nicolas Anelka is a French football player. He plays for Chelsea.

rewording, deletion, sentence splitting

Qualitatively: Phrase-based

Each edge of a tesseract is of the same length.

Same edge of the same length.

Qualitatively: Previous approach

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

He died.

Phrase-based limitations

Phrasal reordering is only motivated by the resulting words, not the input sentence

- tends not to reorder much

In general, tends not to change much when simplifying

<table>
<thead>
<tr>
<th>System</th>
<th>Length ratio</th>
<th>% unmodified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses+Del (phrase-based)</td>
<td>0.9907</td>
<td>56.9%</td>
</tr>
<tr>
<td>In-corpus average</td>
<td>0.85</td>
<td>26.7%</td>
</tr>
</tbody>
</table>
Syntax-based approach

Rather than operating on phrases, operate on grammar trees

Learn probabilistic, syntax-based rules

Learn probabilistic, syntax-based rules

An aside

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
The hard part

Results

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>oracle</th>
<th>length ratio</th>
<th>% unmodified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>0.5640</td>
<td>0.6627</td>
<td>0.8487</td>
<td>57.3%</td>
</tr>
<tr>
<td>Moses+Del</td>
<td>0.6046</td>
<td>0.6421</td>
<td>0.9907</td>
<td>56.9%</td>
</tr>
<tr>
<td>Baseline (no change)</td>
<td>0.5937</td>
<td>-*</td>
<td>1.0</td>
<td>100%</td>
</tr>
<tr>
<td>In-corpus average</td>
<td>-</td>
<td>-</td>
<td>0.85</td>
<td>26.7%</td>
</tr>
</tbody>
</table>

Human Evaluation

Human annotators were asked to rate outputs from simplify, Moses+Del, and the gold standard for grammaticality, meaning preservation, and overall simplification quality.

<table>
<thead>
<tr>
<th></th>
<th>Grammar</th>
<th>Meaning</th>
<th>Simplicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>4.7</td>
<td>4.1</td>
<td>2.9</td>
</tr>
<tr>
<td>Moses+Del</td>
<td>4.5</td>
<td>4.2</td>
<td>2.0</td>
</tr>
<tr>
<td>Gold standard</td>
<td>4.5</td>
<td>3.7</td>
<td>2.7</td>
</tr>
</tbody>
</table>

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

Our life is frittered away.
- Lab Machine 227-31
Qualitatively: syntax-based

After Anton Szandor Lavey's death, his position as head of the church of satan passed on to Blanche Barton.

Syntax:
After Anton Szandor Lavey's death, his position passed on to Blanche Barton.

Phrase-based:
(same as input)

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)

Future thoughts/challenges

How do people do it?

What is simple?
- different domains may have different notion

How do domain constraints affect approaches
- medical and legal
  - deletion is frowned upon
  - insertions are much more common (e.g. definitions)
- can our algorithms vary the simplicity?

Future work

More/better data

Word-level changes seem to be very effective. Can we automate the semi-automated approaches?
- some work here already with Katie Manduca and Colby Horn!

Incorporate more syntactic information

Discourse modeling (between sentence)
Questions?

References

- **Word difficulty analysis:**
  Gondy Leroy and David Kauchak (2013). The Effect of Word Familiarity on Actual and Perceived Text Difficulty. In JAMIA.

- **Semi-supervised approach:**

- **Data generation:**

- **Phrase-based approach:**

- **Syntax-based approach:**