For the Sake of Simplicity:

Unsupervised Extraction of Lexical Simplifications from Wikipedia

Mark Yatskar, Bo Pang*, Cristian Danescu-Niculescu-Mizil, Lillian Lee

Cornell University and Yahoo! Research*

I SPENT ALL NIGHT READING SIMPLE.WIKIPEDIA.ORG; AND NOW I CAN'T STOP TALKING LIKE THIS.

Example lexical simplifications

The GDP is computed **[annually]**

→ The GDP is computed **[every year]**

[Unsupervised extraction of lexical simplifications] from Wikipedia

→**[Finding ways to make simple phrases]** from Wikipedia

Longer than complex version

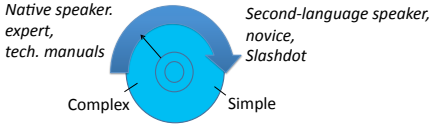
These are *not* syntactic transformations.

(cf. Chandrasekar, Srinivas '97; Siddharthan, Nenkova, McKeown '04; Vickrey, Koller '08)

Applications

Add hyperlinks from phrases to simplifications.

Eventually, the "Style Dial":



Where do we find simple language?

Adults speaking to children? Children's books? Speech to people with a low level of fluency? Textbooks for second language acquisition? Experts teaching novices about a subject?...?

Simple English Wikipedia
(simple.wikipedia.org)

Use SimpleWiki edits to learn to simplify…

BUT, it's not that simple

Look at the words used in SimpleWiki?

- BUT wikis evolve; an article might not yet be fully simple.

Look at the way people edit SimpleWiki?

- BUT, maybe they're correcting spelling or fixing inaccuracies.

(cf. Nelken , Yamangil '08; Shnarch, Barak , Dagan '09)

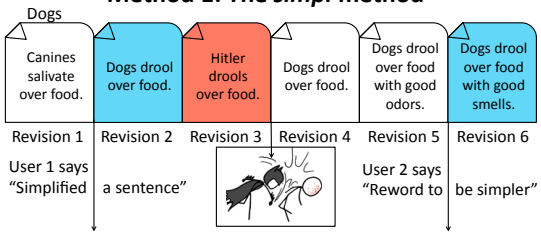
Use EnglishWiki as a filtering mechanism for SimpleWiki edits

Get help from English Wikipedia
(en.wikipedia.org)

Use MT with (EnglishWiki,SimpleWiki) pairs?

- BUT you need to align revisions in the two wikis
- MAYBE some Simple Wiki articles are copied from English Wiki and then simplified?
- BUT which SimpleWiki revision should be aligned?

Method 1: The *simpl* method



1. Find all SimpleWiki revision pairs where the comment contains "simpl".

2. Align sentences between the elements of the pair using TF-IDF
3. Extract possible substitutions from the alignment
4. Rank by PMI(complicated phrase,simple phrase)

Method 2: Edit Mixture Model

P of seeing 'A' rewritten into 'a' $P(a|A) = \sum_{\forall o_i} P(o_i|A)P(a|A,o_i)$ P of 'A' being rewritten into 'a' by an operation; P of 'A' being rewritten by an operation

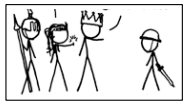
An edit operation:

Fix (o_1) – A correction either in language or content

Simplify(o_2) – Lexical content is simplified

No-op(o_3) – Lexical content is unchanged

Spam(o_4) – Inappropriate text



Estimation Details

Assumptions:

1. EnglishWiki only has fixes
2. The probability of fix being performed in SimpleWiki is proportional to the same probability in EnglishWiki

Let f(A) be the fraction of docs in a collection that had revision that rewrote A

$P(o_1|A) = \alpha f_{English}(A)$.

$P(a \mid A, o_1) = \frac{\#(a, A)}{\#(*, A)}$ in English Wiki

$P(o_2|A) = \max(0, f_{Simple}(A) - \alpha f_{English}(A))$

$P(o_4|A) = 0$. (Assume : no spam.)

$P(o_3|a,A) = 0$. if $a \neq A$

$P(a \mid A) = \frac{\#(a, A)}{\#(*, A)}$ in Simple Wiki, because it is the collection with all edit types

$$P(a|A,o_2) = \frac{P(a|A) - P(o_1|A)P(a|A,o_1)}{P(o_2|A)}$$

- RHS can be estimated using statistics from Simple and English Wikipedia.
- Rank by either $P(a|A,o_2)$ or $p(o_2|A)$. Latter more robust to infrequent edits.

Data

- 38k Simple and English articles
- Number of revisions in English: 5.5 million
- Number of revisions in Simple: 150k
- Number of commented revisions in Simple: 85k
- Articles can be copied from English to Simple

get data at www.cs.cornell.edu/home/llee/data/simple

Results

Method	Prec@100	# of pairs
Human	86%	2000
Edit Model	77%	1079
Simpl Method	66%	2970
Frequent	17%	-
Random	17%	-

Top 100 pairs from each method were manually annotated

Manually assembled dictionary: SpList (by a SimpleWiki author)

Edit and Simpl produce correct pairs not found in SpList (71% and 62%)

Correct Simplifications

Simpl Method

voyage → trip	inducted → added	frequently → often
legend → story	associated → linked	obligatory → required
prevent → stop	founded → started	passed away → died
lies → is	region → area	disbanded → broke up

Edit Model

located → found	virtually → almost	indigenous → native
originated → started	depicted → shown	numerous → many
components → parts	generate → make	discussed → talked about
delivered → gave	classified as → called	collapsed → fell down

Incorrect Simplifications

Simpl Method – trusts revisions

large → big	could → can	the → a
designed → made	and → or	is → was

Edit Model – misses fixes

fail → eat	counting → recounting
cloud → mist	mistakes → members
juice → trees	will become → became

Attempt: bootstrapping on the *simpl* method

1. Find new revisions by getting revisions that contain high ranking substitutions
   - Converges quickly without finding many new revisions
2. Find new comments from revisions that contain high ranking substitutions
   - Example: "reword." Not many comments to find

Related Work

Simplifying medical text for non-doctors: Deléger, Zweigenbaum '09;Elhadad,Sutaria '07

Wordnet + frequency simplification: Devlin, Tait '98;

Tense and coreference simplification: Beigman Klebanov, Knight, Marcu '04

Identifying simple versus non-simple text: Napoles, Dredze '10

Future Work

- Estimate probabilities using EM
- Account for word complexity with inherent model of complexity
- Train a model for rewriting into Simple