Text simplification using synchronous dependency grammars: Generalising automatically harvested rules

Angrosh Mandya & Advaith Siddharthan

Department of Computing Science
University of Aberdeen
Some Background

Text Simplification

- Making the same semantic/pragmatic content more accessible through reformulation
  - Reduce lexical or grammatical complexity by replacing difficult words or splitting long sentences
Some Background

Text Simplification

- Making the same semantic/pragmatic content more accessible through reformulation
  - Reduce lexical or grammatical complexity by replacing difficult words or splitting long sentences
  - Make content more transparent by making discourse relations explicit
Some Background

Text Simplification

- Making the same semantic/pragmatic content more accessible through reformulation
  - Reduce lexical or grammatical complexity by replacing difficult words or splitting long sentences
  - Make content more transparent by making discourse relations explicit
- Reduce text length by deleting peripheral information
- Conceptual simplification
The original police inquiry, which led to Mulcaire being jailed in 2007, also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks, who was editor of the Sun when Mulcaire was working exclusively for its Sunday stablemate.
Text Simplification

The original police inquiry, which led to Mulcaire being jailed in 2007, also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks, who was editor of the Sun when Mulcaire was working exclusively for its Sunday stablemate.

Syntactic Simplification (Siddharthan, 2010):

The original police inquiry led to Mulcaire being jailed in 2007. The inquiry also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks. Rebekah Brooks was editor of the Sun. Mulcaire was working exclusively for its Sunday stablemate then.
The original police inquiry, which led to Mulcaire being jailed in 2007, also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks, who was editor of the Sun when Mulcaire was working exclusively for its Sunday stablemate.

This work:

The first police inquiry led to Mulcaire being jailed in 2007. This police enquiry also found proof that he has successfully intercepted voicemail messages belonging to Rebekah Brooks. Rebekah Brooks was editor of the Sun. Mulcaire was working only for its Sunday stablemate then.
RegenT is a framework for Text Simplification that can:
- Can handle complex lexico-syntactic simplification
- Can get agreement and other morphology correct
- Can use hand written rules for most common syntactic cases (Siddharthan 2010)
- Can use automatically harvested rules for lexicalised paraphrase (Siddharthan & Angrosh, 2014-EACL)
- Goal is to generalise lexical context in automatically harvested rules to make them generalise better
Recent Work

- Text Simplification as MT (EW to SEW)
Recent Work

- Text Simplification as MT (EW to SEW)
  - PBMT (Specia 2010; Wubben et al., 2012; Coster and Kauchak, 2011):
    - Do not attempt syntactic simplification
Recent Work

- Text Simplification as MT (EW to SEW)
  - PBMT (Specia 2010; Wubben et al., 2012; Coster and Kauchak, 2011):
    - Do not attempt syntactic simplification
  - Syntax based SMT (Zhu et al., 2010)
    - Does not do lexical simplification
    - Motivation for learning syntactic rules from corpora unclear
    - Does not do morphology, but can reorder, delete or substitute constituents
Text Simplification as MT (EW to SEW)

- PBMT (Specia 2010; Wubben et al., 2012; Coster and Kauchak, 2011):
  - Do not attempt syntactic simplification

Syntax based SMT (Zhu et al., 2010)

- Does not do lexical simplification
- Motivation for learning syntactic rules from corpora uncleaned
- Does not do morphology, but can reorder, delete or substitute constituents

QTSG (Woodsend & Lapata, 2011)

- Uses different framework for lexical (leaf nodes) and syntactic (internal nodes) simplifications
- LEX: No method for modelling lexical context for lexical simplification
- SYN: Does not do morphology, but can reorder, delete or substitute constituents
Recent Work

Text Simplification as MT (EW to SEW)

- None of these MT systems can correctly transform passive to active voice:
  - Apples are liked by John
  - John likes apples

- Automatically acquired syntactic rules are troublesome in many other ways, e.g.,
  - John, who likes apples, eats pie.
  - John likes apples. He eats pie.
Recent Work

- **Text Simplification as MT (EW to SEW)**
  - None of these MT systems can correctly transform passive to active voice:
    - Apples are liked by John
    - John likes apples
  - Automatically acquired syntactic rules are troublesome in many other ways, e.g.,
    - John, who likes apples, eats pie.
    - John likes apples. He eats pie.

- **RegenT**
  - Uses handwritten linguistically sound rules for syntactic simplification
  - Uses automatically acquired rules for lexicalised constructs
Ding et. al. (2005) introduce Synchronous Dependency Insertion Grammar (SDIG)

- Elementary trees (ETs) are sub-sentential dependency structures containing one or more lexical items.
- SDIG is a meta grammar that rewrites ETs
- Our simplification rules are one such meta grammar
TS with Synchronous Dependency Grammars

**RULE:** PRODUCING \textsubscript{2} \textsc{by} PRODUCING

DELETE:

1. xcomp(X0[reproduce], X1[producing])
2. dobj(X1[producing], X2[spores])

INSERT:

1. amod(X2, X1)
2. prep\textsubscript{by}(X0, X2)

\begin{tikzpicture}
  \node at (0,0) {reproduce} child {node {producing} child {node {dobj} child {node {spores}}}} !-! child {node {spores} child {node {amod} child {node {producing}}}}
  \node at (2,0) {reproduce} child {node {producing} child {node {xcomp} child {node {spores}}}} !-! child {node {spores} child {node {prep\textsubscript{by}} child {node {producing}}}};
\end{tikzpicture}
Advantages over PBMT

RULE: described_as2called

DELETE:
- prep_as(X0[described], X1)

INSERT:
- dobj(X2[called], X1)

- Coulter was described as a polemicist.
- Coulter was called a polemicist.
Advantages over PBMT

**RULE:** \texttt{described\_as}2\texttt{called}

**DELETE:**

- \texttt{prep\_as}(X0[described], X1)

**INSERT:**

- \texttt{dobj}(X2[called], X1)

- Coulter has \texttt{described} herself as a polemicist.
- Coulter has \texttt{called} herself a polemicist.
Automatic Rule Acquisition

- Automatically acquired rules for paraphrase and lexical simplification
  - Compare GRs of aligned sentences: create Delete, Insert lists

DELETE:
  acomp(X0[considered], X1[antiquated])

INSERT:
  acomp(X0, X2[old])
Automatic Rule Acquisition

- Automatically acquired rules for paraphrase and lexical simplification
  - Compare GRs of aligned sentences: create Delete, Insert lists
  - Generalise rule from multiple instances:
     - **DELETE:**
       - amod(X0, X1[extensive, large, massive, sizable, giant...])
     - **INSERT:**
       - amod(X0, X2[big])
Automatic Rule Acquisition

- Automatically acquired rules for paraphrase and lexical simplification
  - Compare GRs of aligned sentences: create Delete, Insert lists
  - Reduce context size:

**DELETE:**
- \( \text{aux}(X0[\text{amend}], X1[\text{to}]) \)
- \( \text{infmod}(X2[\text{proposals}], X0[\text{amend}]) \)
- \( \text{dobj}(X0[\text{amend}], X3[\text{constitution}]) \)

**INSERT:**
- \( \text{aux}(X0[\text{change}], X1[\text{to}]) \)
- \( \text{infmod}(X2[\text{proposals}], X0[\text{change}]) \)
- \( \text{dobj}(X0[\text{change}], X3[\text{constitution}]) \)
Automatic Rule Acquisition

- Automatically acquired rules for paraphrase and lexical simplification
  - Compare GRs of aligned sentences: create Delete, Insert lists
  - Reduce context size:

  **DELETE:**
  - dobj(X0[amend], X3[constitution])

  **INSERT:**
  - dobj(X0[change], X3[constitution])
Automatic Rule Acquisition

- Automatically acquired rules for paraphrase and lexical simplification
  - Compare GRs of aligned sentences: create Delete, Insert lists
  - Extract elementary trees:

DELETE:
  \[ \text{amod}(X_0[\text{rule}], X_1[\text{general}]) \]
  \[ \text{prep_in}(X_2[\text{increase}], X_3[\text{precipitation}]) \]

INSERT:
  \[ \text{amod}(X_0, X_4[\text{normal}]) \]
  \[ \text{prep_in}(X_2, X_5[\text{rain}]) \]
Automatic Rule Acquisition

- Automatically acquired rules for paraphrase and lexical simplification
  - Compare GRs of aligned sentences: create Delete, Insert lists
  - Removing lexical context from longer rules:

  **DELETE:**
  - amod(X0[satellites], X1[irregular])
  - prep_of(X2, X0[satellites])

  **INSERT:**
  - amod(X3[moons], X4[non-spherical])
  - prep_of(X2, X3)
Automatic Rule Acquisition

- Automatically acquired rules for paraphrase and lexical simplification
  - Compare GRs of aligned sentences: create Delete, Insert lists
  - Expand context using WordNet

**DELETE:**

\[
\text{prep\_with}(X0[\text{striking}], X1[\text{accelerator,bar,bludgeon, bough,bow,branch, cane,club,crutch,gas,gun, handspike,implement,joint,joystick,lever, limb,margarin,margarine,marge,oleo,oleomargarine, pedal,peg,pin,reefer,spliff,staff,stalk,stem, stick,stick,throttle,treadle,trigger}])
\]

**INSERT:**

\[
\text{prep\_with}(X2[\text{hitting}], X1)
\]
IC(c) = −log p(c), p(c) is likelihood of seeing a concept or any of its hyponyms in a corpus

sim_{resnik}(c_1, c_2) = \max_{c \in S(c_1, c_2)} IC(c)

sim_{lin}(c_1, c_2) = \frac{2 \cdot sim_{res}(c_1, c_2)}{IC(c_1) + IC(c_2)}

We use Lin Similarity
## Results: Generalisable?

11,000 sentences

<table>
<thead>
<tr>
<th>Rule Version</th>
<th>Rule Applications</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikicontext</td>
<td>7610</td>
<td>3.41</td>
</tr>
<tr>
<td>WordNet context (Lin=0.8)</td>
<td>7870</td>
<td>11.85</td>
</tr>
<tr>
<td>WordNet context (0.4)</td>
<td>8488</td>
<td>11.85</td>
</tr>
<tr>
<td>WordNet context (0.1)</td>
<td>10715</td>
<td>40.80</td>
</tr>
<tr>
<td>Nocontext</td>
<td>31589</td>
<td>315.09</td>
</tr>
</tbody>
</table>
Results: Generalisable and Accurate?

90 sentences; 28 Raters; Mechanical Turk

<table>
<thead>
<tr>
<th>Rater</th>
<th>FLUENCY</th>
<th>SIMPLICITY</th>
<th>MEANING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WN</td>
<td>WI</td>
<td>NC</td>
</tr>
<tr>
<td>Mean</td>
<td>3.28</td>
<td>3.59</td>
<td>2.49</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
Results: Comparison to other systems

60 sentences; 72 participants; Mechanical Turk

<table>
<thead>
<tr>
<th></th>
<th>FLUENCY</th>
<th></th>
<th>SIMPLICITY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EW</td>
<td>SEW</td>
<td>QTSG</td>
</tr>
<tr>
<td>Mean</td>
<td>3.99</td>
<td>4.06</td>
<td>1.97</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MEANING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EW</td>
</tr>
<tr>
<td>Mean</td>
<td>4.03</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
</tr>
</tbody>
</table>
We have described the first HYBRID text simplification system:

- hand-written rules for common syntactic simplifications
- automatically harvested rules for a much larger set of lexicalised simplifications
- lexical context expanded to facilitate rule application in new contexts

Modelling of lexical context is not perfect

But, more expressive and robust than a similar system based on QTSG
Simplification in the real world

We are twelve billion light years from the edge,
That's a guess,
No-one can ever say it’s true,
But I know that I will always be with you.

- Katie Melua
Simplification in the real world

We are twelve billion light years from the edge,
That's a guess,
No-one can ever say it's true,
But I know that I will always be with you.

- Katie Melua

We are 13.7 billion light-years from the edge of the observable universe,
That's a good estimate,
With well-defined error bars,
And with the available information, I predict that I will always be with you.

- Simon Singh