Presentation for INLG 2014 paper on text simplification using synchronous dependency grammars

Angrosh Mandya & Advaith Siddharthan

Department of Computing Science
University of Aberdeen
Text Simplification

- Making the same semantic/pragmatic content [Dorr et al.2004] 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 [Dorr et al. 2004] 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
Text Simplification

- Making the same semantic/pragmatic content [Dorr et al.2004] 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 [Angrosh et al.2014]
- 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.
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
[Siddharthan2010, Siddharthan2011, Angrosh et al.2014]:

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 [Siddharthan2010, Siddharthan2011]
- Can use automatically harvested rules for lexicalised paraphrase [Siddharthan and Angrosh2014]

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)
    - Do not attempt syntactic simplification
Recent Work

- Text Simplification as MT (EW to SEW)
    - Do not attempt syntactic simplification
  - Syntax based SMT [Zhu et al.2010]
    - Does not do lexical simplification
    - Motivation for learning syntactic rules from corpora unclesad
    - Does not do morphology, but can reorder, delete or substitute constituents
Recent Work

- **Text Simplification as MT (EW to SEW)**
    - Do not attempt syntactic simplification
  - Syntax based SMT [Zhu et al.2010]
    - Does not do lexical simplification
    - Motivation for learning syntactic rules from corpora unclead
    - Does not do morphology, but can reorder, delete or substitute constituents

- **QTSG [Woodsend and Lapata2011]**
  - 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
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:** PRODUCING2BY_PRODUCING

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

**INSERT:**
1. `amod(X2, X1)`
2. `prep_by(X0, X2)`
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:** described_as2called

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

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

- Coulter has described herself as a polemicist.
- Coulter has 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:
- aux(X0[amend], X1[to])
- infmod(X2[proposals], X0[amend])
- dobj(X0[amend], X3[constitution])

INSERT:
- aux(X0[change], X1[to])
- infmod(X2[proposals], X0[change])
- 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
  - Reduce context size:

```plaintext
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:**
- amod(X0[rule], X1[general])
- prep_in(X2[increase], X3[precipitation])

**INSERT:**
- amod(X0, X4[normal])
- prep_in(X2, X5[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:

prep_with(X0[striking], X1[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:

prep_with(X2[hitting], X1)
WordNet Expansion

\[ IC(c) = -\log p(c), \]  
\[ p(c) \text{ is likelihood of seeing a concept or any of its hyponyms in a corpus} \]

\[ \text{sim}_{\text{resnik}}(c_1, c_2) = \max_{c \in S(c_1, c_2)} IC(c) \]

\[ \text{sim}_{\text{lin}}(c_1, c_2) = \frac{2 \cdot \text{sim}_{\text{res}}(c_1, c_2)}{IC(c_1) + IC(c_2)} \]

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></td>
</tr>
<tr>
<td>WordNet context (Lin=0.8)</td>
<td>7870</td>
<td>3.41</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>SIMPLICITY</th>
<th>MEANING</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>
<tr>
<td>Mean</td>
<td>-</td>
<td>4.03</td>
<td>2.23</td>
</tr>
<tr>
<td>Median</td>
<td>-</td>
<td>4</td>
<td>2</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
## References

Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan.
2014.
Lexico-syntactic text simplification and compression with typed dependencies.

William Coster and David Kauchak.
2011.
Learning to simplify sentences using wikipedia.

Bonnie Dorr, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith Miller, Teruko Mitamura, Owen Rambow, et al.
2004.
Semantic Annotation and lexico-syntactic paraphrase.
LREC.

Ani Nenkova, Advaith Siddharthan, and Kathleen McKeown.
2005.
Automatically learning cognitive status for multi-document summarization of newswire.

Advaith Siddharthan and Mandya Angrosh.
2014.
Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules.
In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL’14), Gothenburg, Sweden.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown.

Advaith Siddharthan.


Advaith Siddharthan.


Lucia Specia.


Kristian Woodsend and Mirella Lapata.


Sander Wubben, Antal van den Bosch, and Emiel Krahmer.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych.
2010.
A monolingual tree-based translation model for sentence simplification.
Association for Computational Linguistics.