hand-written and automatically harvested rules

Presentation for EACL 2014 paper on Hybrid text simplification

Advaith Siddharthan & Angrosh Mandya

Department of Computing Science
University of Aberdeen
Text Simplification

- Making the same semantic/pragmatic [Dorr et al.2004] content more accessible through reformulation
  - Reduce lexical or grammatical complexity by replacing difficult words or splitting long sentences [Siddharthan and Angrosh 2014]
  - Make content more transparent by making discourse relations explicit.
- Conceptual simplification.
- Remove redundant information [Nenkova et al.2005, Siddharthanan et al.2011]
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

[Siddharthan2010, Siddharthan2011]:

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.

Siddharthan & Angrosh (2014):

The original police inquiry led to Mulcaire being jailed in 2007. The police 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.
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
But, for the moment...

- We want 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 cases (syntactic)
  - Can use automatically harvested rules for lexicalised paraphrase.
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)
  - 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
Recent Work

- Text Simplification as MT (EW to SEW)
  - QTSG (Woodsend & Lapata, 2011)
    - Uses different framework for lexical (leaf nodes) and syntactic (internal nodes) simplifications
    - 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.
**Text Simplification Framework**

- **Analysis:** Dependency Parse using the Stanford Parser
- **Transformation:** Rules applied to Dependencies
  - Handwritten Rules (~250, fairly mature)
  - Automatically acquired Rules (~3000 in this work, around 20K in current system (more papers on the way))
- **Generation:**
  - Gen-light (Siddharthan 2010): Generate sentences from dependency graphs.
The/DT cat/NN is/VBZ chased/VBN by/IN the/DT dogs/NNS ./.

det(cat-2, The-1)
nsubjpass(chased-4, cat-2)
auxpass(chased-4, is-3)
det(dogs-7, the-6)
agent(chased-4, dogs-7)
punct(chased-4, .-8)
Generation from Typed Dependencies

The/DT cat/NN is/VBZ chased/VBN by/IN the/DT dogs/NNS ./.

det(cat-2, The-1)
nsubjpass(chased-4, cat-2)
auxpass(chased-4, is-3)
det(dogs-7, the-6)
agent(chased-4, dogs-7)
punct(chased-4, .-8)
Generation from Typed Dependencies

Dependency Trees
- Generation is inorder: left subtrees, root, right subtrees
- Order to process involves typical generation decisions guided by type and statistical preferences for word/phrase order
- But we can use word order from original sentence

The/DT cat/NN is/VBZ chased/VBN by/IN the/DT dogs/NNS ./.

det(cat-2, The-1)
nsubjpass(chased-4, cat-2)
auxpass(chased-4, is-3)
det(dogs-7, the-6)
agent(chased-4, dogs-7)
punct(chased-4, .-8)
Simplification Rule: Passive to Active

1. Delete:
   1. nsubjpass(X0, X1)
   2. auxpass(X0, X2)
   3. agent(X0, X3)

2. Insert:
   1. nsubj(X0, X3)
   2. dobj(X0, X1)

\[
\text{chased:4} \quad \rightarrow \quad \text{chased:4}
\]

\[
\begin{align*}
\text{cat:2} & \quad \text{is:3} & \quad \text{dogs:7} \\
\text{det} & \quad \text{det} & \quad \text{det}
\end{align*}
\]

\[
\begin{align*}
\text{The:1} & \quad \text{the:6} \\
\text{The:1} & \quad \text{the:6}
\end{align*}
\]
Simplification Rule: Passive to Active

1. Delete:
   1. nsubjpass(X0, X1)
   2. auxpass(X0, X2)
   3. agent(X0, X3)

2. Insert:
   1. nsubj(X0, X3)
   2. dobj(X0, X1)

3. Traversal Order Specifications:
   1. Node X0: [X3, X0, X1]
Simplification Rule: Passive to Active

1. Delete:
   1. nsubjpass(X0, X1)
   2. auxpass(X0, X2)
   3. agent(X0, X3)

2. Insert:
   1. nsubj(X0, X3)
   2. dobj(X0, X1)

3. Traversal Order Specifications:
   1. Node X0: [X3, X0, X1]

4. Morphological Specifications:
   1. Node X0: Tense from X2, Number from X3
Summary: Gen-light

- Transformation rules need 4 lists:
  1. DELETE: List of GRs to delete from input.
  2. INSERT: List of GRs to insert into input.
  3. ORDERING: List of nodes with subtree order specified.
  4. NODE-OPERATIONS: List of lexical substitutions, morphological changes and deletion operations on nodes.

- Some of 3,4 could be implemented in a generator.
Automatic Rule Acquisition

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

  DELETE:
  - amod(X0, X1[extensive, large, massive, sizable...])

  INSERT:
  - amod(X0, X2[big])
Automatic Rule Acquisition

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

    DELETE:
    amod(X0, X1[extensive, large, massive, sizable...])

    INSERT:
    amod(X0, X2[big])

    MOVE:
    X1 → X2
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.
Synchronous Dependency Grammar

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.
RegenT: Regenerating Text

Meta grammar that maps patterns in English dependency graphs to Simple English dependency graphs

- ~ 140 Hand-written Rules for relative clauses (11), apposition (8), coordination (33), subordination (47), cleft constructions (40) and voice (8).
- ~ 100 more hand-written stylistic rules compiled from old unix tool “diction”.

Automatically acquired rules for paraphrase and lexical simplification

- In this paper, 3180 rules from SEW revision histories
- (work in progress, currently over ~10K rules)
Applying the grammar (Decoding)

- rules are applied iteratively and exhaustively

The cat was chased by a dog that was barking

\[
\begin{align*}
\text{det}(cat-2, \ The-1) \\
\text{nsubjpass}(chased-4, \ cat-2) \\
\text{auxpass}(chased-4, \ was-3) \\
\text{det}(dog-7, \ a-6) \\
\text{agent}(chased-4, \ dog-7) \\
\text{nsubj}(barking-10, \ dog-7) \\
\text{aux}(barking-10, \ was-9) \\
\text{rcmod}(dog-7, \ barking-10)
\end{align*}
\]

- Apply rule for relative clauses:
  1. Delete: 2. Insert:
     (a) \text{rcmod}(X0, \ X1)  (a) \text{nsubj}(X1, \ X0)
     (b) \text{nsubj}(X1, \ X0)
Applying the grammar (Decoding)

- rules are applied iteratively and exhaustively

The cat was chased by a dog that was barking

\[
\text{det(cat-2, The-1)} \\
\text{nsubjpass(chased-4, cat-2)} \\
\text{auxpass(chased-4, was-3)} \\
\text{det(dog-7, a-6)} \\
\text{agent(chased-4, dog-7)} \\
\text{nsubj(barking-10, dog-7)} \\
\text{aux(barking-10, was-9)}
\]

- Apply rule for voice change:
  1. Delete: 
     (a) nsubjpass(X0, X1) 
     (b) auxpass(X0, X2) 
     (c) agent(X0, X3)
  2. Insert: 
     (a) nsubj(X0, X3) 
     (b) dobj(X0, X1)
Applying the grammar (Decoding)

- rules are applied iteratively and exhaustively

The cat was chased by a dog that was barking

\[\text{det}(\text{cat-2, The-1})\]
\[\text{dobj}(\text{chased-4, cat-2})\]
\[\text{det}(\text{dog-7, a-6})\]
\[\text{nsubj}(\text{chased-4, dog-7})\]
\[\text{aux}(\text{barking-10, was-9})\]
\[\text{nsubj}(\text{barking-10, dog-7})\]
Applying the grammar (Decoding)

- rules are applied iteratively and exhaustively

The cat was chased by a dog that was barking

A dog chased the cat. The dog was barking.
## Results

<table>
<thead>
<tr>
<th></th>
<th>FLUENCY</th>
<th></th>
<th>SIMPLICITY</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEW</td>
<td>QTSG</td>
<td>HYB</td>
<td>SEW</td>
</tr>
<tr>
<td>Mean</td>
<td>4.15</td>
<td>2.48</td>
<td>4.04</td>
<td>4.05</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MEANING</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEW</td>
<td>QTSG</td>
</tr>
<tr>
<td>Mean</td>
<td>3.96</td>
<td>2.92</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

- EW: English Wikipedia; SEW: Simple English Wikipedia; QTSG system (Woodsend & Lapata, 2011); HYB: Our hybrid system.
- 5 participants (computational linguists); 25 sentences
- All differences in means are significant, except Fluency between HYB and SEW.
### Results

<table>
<thead>
<tr>
<th></th>
<th>FLUENCY</th>
<th></th>
<th>SIMPLICITY</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EW</td>
<td>SEW</td>
<td>QTSG</td>
<td>HYB</td>
</tr>
<tr>
<td>Mean</td>
<td>4.15</td>
<td>2.48</td>
<td>4.04</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3.99</td>
<td>4.06</td>
<td>1.97</td>
<td>3.52</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MEANING</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EW</td>
<td>SEW</td>
</tr>
<tr>
<td>Mean</td>
<td>-</td>
<td>3.96</td>
</tr>
<tr>
<td>Median</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Mean</td>
<td>-</td>
<td>4.03</td>
</tr>
<tr>
<td>Median</td>
<td>-</td>
<td>4</td>
</tr>
</tbody>
</table>

- 5 participants; 25 sentences (this paper)
- MTurk Eval; 72 participants; 60 sentences
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

More expressive and robust than a similar system based on QTSG

- outperform QTSG for fluency, simplicity and meaning preservation.
- allows linguistically sound rules to be written for complex lexico-syntactic transformations, including passive to active voice.
References

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.
2011.
Information status distinctions and referring expressions: An empirical study of references to people in news summaries.

Advaith Siddharthan.
2010.
Complex lexico-syntactic reformulation of sentences using typed dependency representations.

Advaith Siddharthan.
2011.
Text simplification using typed dependencies: A comparison of the robustness of different generation strategies.
In Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011), Nancy, France.