Splitting Complex English Sentences

<table>
<thead>
<tr>
<th>John Lee</th>
<th>City University of Hong Kong</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Buddhika K. Pathirage Don</td>
<td>Hong Kong Applied Science and Technology Research Institute</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Previous work
- Approach
- Evaluation
Introduction

• Text simplification
  – Goal: to re-write a sentence to reduce its complexity, but preserve the meaning

• Various types of text simplification
  – Lexical simplification
  – Syntactic simplification
  – Content deletion / insertion
## Introduction

- This paper addresses one kind of syntactic simplification
  - Goal: to split a complex sentence (S) into two simpler sentences (S1, S2)

<table>
<thead>
<tr>
<th>S</th>
<th>The man, carrying numerous books, entered the room.</th>
<th>Input: Complex sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>The man entered the room.</td>
<td>Output: Two simpler sentences</td>
</tr>
<tr>
<td>S2</td>
<td>He was carrying numerous books.</td>
<td></td>
</tr>
</tbody>
</table>
Introduction

• Sentence splitting involves two steps:
  – Detect the text span that should be taken out of the complex sentence
  – Re-generate the two simpler sentences
    • Sentence re-ordering
    • Pronouns or referring expressions
    • Tense changes

• We focus on the first step
Outline

• Introduction
• Previous work
• Approach
• Evaluation
Previous work

• Many previous studies on syntactic simplification (Chandrasekar et al., 2016; Siddharthan, 2002; Inui et al., 2003; Belder & Moens, 2010; Bott et al., 2012; Saggion et al., 2015; etc.)

• Typical approach: parse sentence, then apply rules to transform specific constructs
  – E.g., rules for apposition, relative clauses, subordination, coordination, etc. (Aluisio et al., 2008; Siddharthan and Angrosh, 2014)
Previous work

• Evaluation methodology
  – Task-based, e.g., reading comprehension (Angrosh et al., 2014)
  – Readability metrics, BLEU (Aluisio et al., 2010; Narayan and Gardent, 2014)
  – Human ratings (Stajner et al., 2016)

• Limitation: no clear indication of what goes wrong in the simplification process
Outline

• Introduction
• Previous work
• **Approach**
• Evaluation
Data

• We created a evaluation dataset
  – Derived from aligned sentences from Wikipedia and Simple Wikipedia (Kauchak, 2013)
    • 23,715 one-to-two sentence alignments
    • 1,071 <S,S1,S2> sentence triplets manually annotated to serve as test set
    • Remainder serves as training set

```xml
<S><ref>The man</ref> <split type="participial">, carrying numerous books, </split> entered the room.</S>
<S1><ref1>The man</ref1> entered the room.</S1>
<S2><ref2>He</ref2> was carrying numerous books.</S2>
```

Referring expressions

Split text span
Baseline

• Manually-crafted tree patterns
  – Patterns for six different syntactic constructs
  – Breadth-first search for these patterns in parse tree
Decision Tree

- Baseline yields high recall but low precision
  - Always split a sentence when it has one of the syntactic constructs

- Decision tree approach
  - Determine whether each candidate text span should be split or not
  - Features: POS, child/parent/sibling POS, comma, determiner, text span length
Outline

• Introduction
• Previous work
• Approach
• Evaluation
Result

• Decision tree system outperforms baseline in precision, but has lower recall

<table>
<thead>
<tr>
<th>Construct</th>
<th>Baseline Precision / Recall</th>
<th>Proposed Precision / Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination</td>
<td>0.31 / 0.84</td>
<td>0.61 / 0.80</td>
</tr>
<tr>
<td>Adjectival clause</td>
<td>0.29 / 0.97</td>
<td>0.59 / 0.79</td>
</tr>
<tr>
<td>Participial phrase</td>
<td>0.33 / 0.90</td>
<td>0.56 / 0.58</td>
</tr>
<tr>
<td>Appositive phrase</td>
<td>0.21 / 0.91</td>
<td>0.36 / 0.56</td>
</tr>
<tr>
<td>Subordination</td>
<td>0.39 / 0.84</td>
<td>0.70 / 0.74</td>
</tr>
<tr>
<td>Parataxis</td>
<td>0.78 / 0.99</td>
<td>0.92 / 0.95</td>
</tr>
<tr>
<td>Overall</td>
<td>0.34 / 0.88</td>
<td>0.63 / 0.72</td>
</tr>
</tbody>
</table>
Conclusion

• We investigated the task of automatic complex sentence splitting
  – A subtask of syntactic simplification
    • Decision tree outperforms a baseline based only on matching syntactic patterns
  – We reported the first large-scale, detailed evaluation on this task with a new dataset