Presentation for INLG 2010 paper on Complex lexico-syntactic reformulation

Advaith Siddharthan

\(^1\)Department of Computer Science
University of Aberdeen
Outline

1 Motivation

2 Reformulation with Transfer Rules
   - Phrasal Parse Trees
   - Minimal Recursion Semantics
   - Typed dependencies

3 Results and Examples

4 Summary
Motivation for Sentence Reformulation

Text Simplification

- Making the same semantic/pragmatic content [Dorr et al.2004] more accessible using linguistic reformulation
  - Reduce lexical or grammatical complexity by splitting long sentences or replacing difficult words
  - Make content more transparent by making discourse relations more explicit.
Text Simplification

- Three Stages (Siddharthan 2003; 2006)
- Analysis
  - Identifies constructs to simplify based on pattern matching on PoS tagged sentences
    - Relative clauses, apposition, coordinate and subordinate clauses
- Transformation
  - 7 Hand-crafted syntactic rules to split sentences
- Generation
  - Address fluency and coherence issues - Information ordering, referring expressions...
An example

Original

- Also contributing to the firmness in copper, the analyst noted, was a report by Chicago purchasing agents, which precedes the full purchasing agents report that is due out today and gives an indication of what the full report might hold.

Simplified by machine

- Also contributing to the firmness in copper, the analyst noted, was a report by Chicago purchasing agents.
- The Chicago report precedes a full purchasing agents report and gives an indication of what the full report might hold.
- The full report is due out today.
Uses of Text Simplification

- **Literacy Aids**
  - My software is used to simplify online texts to linguistic levels suitable for adult second-language learners.
    - Univ. of Washington (Petersen and Ostendorf, 2007)
    - 87% good choice of split point
    - $P=0.59$, $R=0.37$, compared to sentences in manual simplification

- **Enabling Access to Information**
  - Basis for a project on text simplification in Portuguese
    - Univ. of Sao Paolo (Aluísio et al., 2008)
    - Aims to make online text more accessible to Brazilians with low literacy skills
    - Considers adults with 4-8 years of schooling
Uses of Text Simplification

- Education and Literacy
  - Literacy skills improve faster when
    - Reading material is interesting
    - Language is challenging without being intimidating
  - Simplified texts commonly used in language learning
  - Manual simplification helps
    - low ability readers outperform middle ability readers on comprehension tests (L’Allier, 1980; Linderholm et al., 2000)
  - Automatic simplification not quite there yet...

- Enable access to information for disadvantaged groups
  - Language deficits (from aphasia, deafness, aged, etc.)
This research

- Part of a larger study on formulating sentences expressing causation [Siddharthan2010, Siddharthan2011, Siddharthan and Katsos2010]
- 4 constructs
  a. Fructose-induced hypertension is caused by increased salt absorption by the intestine and kidney. (verb)
  b. Fructose-induced hypertension occurs because of increased salt absorption by the intestine and kidney. (preposition)
  c. Fructose-induced hypertension occurs because salt absorption by the intestine and kidney increases. (conjunction)
  d. The cause of fructose-induced hypertension is increased salt absorption by the intestine and kidney. (noun)
- x 2 Information Orders = 8 formulation types
This research

- Part of a larger study on formulating sentences expressing causation [Siddharthan2010, Siddharthan2011, Siddharthan and Katsos2010]

- 4 constructs
  - a’ Increased salt absorption by the intestine and kidney causes fructose-induced hypertension. (verb)
  - b Fructose-induced hypertension occurs because of increased salt absorption by the intestine and kidney. (preposition)
  - c Fructose-induced hypertension occurs because salt absorption by the intestine and kidney increases. (conjunction)
  - d The cause of fructose-induced hypertension is increased salt absorption by the intestine and kidney. (noun)

- x 2 Information Orders = 8 formulation types
This research

- Part of a larger study on formulating sentences expressing causation [Siddharthan2010, Siddharthan2011, Siddharthan and Katsos2010]
- 4 constructs
  - a’ Increased salt absorption by the intestine and kidney causes fructose-induced hypertension. (verb)
  - b’ Because of increased salt absorption by the intestine and kidney, fructose-induced hypertension occurs. (preposition)
  - c Fructose-induced hypertension occurs because salt absorption by the intestine and kidney increases. (conjunction)
  - d The cause of fructose-induced hypertension is increased salt absorption by the intestine and kidney. (noun)
- x 2 Information Orders = 8 formulation types
3 Genres: 
- Pubmed 
- BNC-world 
- BNC- natural science 

144 sentences selected 
- 48 from each genre (pubmed, bnc-world, bnc-natsci) 
- 18 of each type 

1152 sentences in total 
- Each of the 144 originals is rewritten (manually) to generate the other 7 versions
Transfer Rules: Phrasal Parse Trees

The explosion was caused by an incendiary device.
(S (NP (AT The) (NN1 explosion))
  (VP (VBDZ be+ed)
    (VP (VVN cause+ed)
      (PP (II by) (NP (AT1 an) (JJ incendiary) (NN1 device)))))

An incendiary device caused the explosion.
(S (NP (AT1 An) (JJ incendiary) (NN1 device))
  (VP (VVD cause+ed)
    (NP (AT the) (NN1 explosion)))))

Derived Rule:
(S (??X0[NP])
  (VP (VBZ be+s)
    (VP(VVN cause+ed) (PP(II by+) (??X1[NP])))}})
↓
(S (??X1[NP]) (VP (VVZ cause+s) (??X0[NP])))
Transfer Rules: Phrasal Parse Trees

- Similar to Synchronous Tree Substitution Grammars (STSG)
  - Popular in Sentence compression/shortening
  - Machine Translation
- But there are problems for complex reformulations
Problems

- **Lexical reformulations**
  - *was given by* vs *gave*
  - Can be handled by allowing variables over pos tags

- **Agreement**
  - *The car* is pushed by the boys
  - *The boys* pushes the car
  - Number information not easily accessible
    - Very dependent on grammar
    - Head of NP not marked, Coordination, etc
  - Hard to formulate within a single transfer rule

- **Similar strings can be parsed differently in the two aligned sentences**
  - For example, different PP attachment
  - Leads to very complicated rules, often with more than 20 variables
Results

- Split data into
  - development/training (96 instances of passive to active)
  - test sets (48 instances of passive)
- Using the top parse for each sentence
  - 92 rules were obtained - no generalisability
  - coverage of test set < 10%
- Using the top 20 parses for each sentence
  - coverage of test set > 70%
  - BUT more than 4000 rules learnt - just to change voice
- Not the right representation for complex lexico-syntactic reformulation
Outline

1. Motivation

2. Reformulation with Transfer Rules
   - Phrasal Parse Trees
   - Minimal Recursion Semantics
   - Typed dependencies

3. Results and Examples

4. Summary
Tom ate because of his hunger.

MRS:
\[
\text{named}(x5, \text{Tom}), \ _\text{eat}_\text{v}_\text{1}(e2,x5), \\
\ _\text{because}_\text{of}(e2,x11), \ \text{poss}(x11,x16), \\
\text{pron}(x16), \ _\text{hunger}_\text{n}(x11)
\]

Transfer Rule
\[
\ _\text{because}_\text{of}(e,x), \ P(e,y) \iff \ _\text{cause}_\text{v}_\text{1}(e10,x,y,l1), \ l1:P(e,y)
\]

Output MRS:
\[
\text{named}(x5, \text{Tom}), \ l1:\_\text{eat}_\text{v}_\text{1}(e2,x5), \\
\ _\text{cause}_\text{v}_\text{1}(e10,x11,x5,l1), \ \text{poss}(x11,x16), \\
\text{pron}(x16), \ _\text{hunger}_\text{n}(x11), \ x5 \text{ aeq } x16
\]

His hunger caused Tom to eat.
Minimal Recursion Semantics: Transfer Rules

- Writing transfer rules is intuitive and easy in MRS.
- Bi-directional grammar for generation ensures grammaticality.
  - No need to worry about lexicalisation or agreement, etc.
- But there are problems...
Problems

- Bidirectional grammars such as the ERG fail to parse ill-formed input
  - Also fail to analyse some well-formed input because of limitations in coverage
- The ERG gives an acceptable parse ‘out of the box’ for only around 50-60% of sentences from scientific papers
- The generator can get slow and memory intensive for long sentences
- Much of this processing effort during generation is redundant
  - The input sentence can be used to narrow down generation choices
  - Infrastructure does not exist to support this
Outline

1. Motivation

2. Reformulation with Transfer Rules
   - Phrasal Parse Trees
   - Minimal Recursion Semantics
   - Typed dependencies

3. Results and Examples

4. Summary
Typed Dependencies

The explosion was caused by an incendiary device.

Typed Dependencies

(|ncsubj| |cause+ed:4_VVN| |explosion:2_NN1| _)
(|aux| |cause+ed:4_VVN| |be+ed:3_VBDZ|)
(|passive| |cause+ed:4_VVN|)
(|iobj| |cause+ed:4_VVN| |by:5_II|)
(|dobj| |by:5_II| |device:8_NN1|)
(|det| |device:8_NN1| |an:6_AT1|)
(|ncmod| _ |device:8_NN1| |incendiary:7_JJ|)
(|det| |explosion:2_NN1| |the:1_AT|)
The explosion was caused by an incendiary device.

Typed Dependencies

(\textit{ncsubj} \textit{cause+ed:4_VVN} | \textit{explosion:2_NN1} | _)
(\textit{aux} \textit{cause+ed:4_VVN} | \textit{be+ed:3_VBDZ})
(\textit{passive} \textit{cause+ed:4_VVN})
(\textit{iobj} \textit{cause+ed:4_VVN} | \textit{by:5_{II}})
(\textit{dobj} \textit{by:5_{II}} | \textit{device:8_NN1})
(\textit{det} \textit{device:8_NN1} | \textit{an:6_AT1})
(\textit{ncmod} _ \textit{device:8_NN1} | \textit{incendiary:7_JJ})
(\textit{det} \textit{explosion:2_NN1} | \textit{the:1_AT})
Generation from Typed Dependencies

- **Phrasal Parse Trees**
  - Generation is depth first l-r search for nodes

- **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

```
S
  NP
    the:1 explosion:2 be+ed:3 VP
      by:5 NP
        the:1 device:8
          an:6 incendiary:7
  VP
    cause+ed PP
      by:5 NP
        an:6 incendiary:7 device:8
```
Typed Dependencies: Transfer Rules

1. Match and Delete:
   1. (possible) | ??X0 |
   2. (iobj | ??X0 | ??X1(by_II))
   3. (dobj | ??X1 | ??X2)
   4. (ncsubj | ??X0 | ??X3 | _)
   5. (aux | ??X0 | ??X4)

2. Insert:
   1. (ncsubj | ??X0 | ??X2 | _)
   2. (dobj | ??X0 | ??X3 | )

cause+ed:4
device:8 explosion:2
an:6 incendiary:7 the:1
Typed Dependencies: Transfer Rules

1 Match and Delete:
   1 (|passive| |??X0|)
   2 (|iobj| |??X0| |??X1(by_II)|)
   3 (|dobj| |??X1| |??X2|)
   4 (|ncsubj| |??X0| |??X3| _)
   5 (|aux| |??X0| |??X4|)

2 Insert:
   1 (|ncsubj| |??X0| |??X2| _)
   2 (|dobj| |??X0| |??X3|)

3 Traversal Order Specifications:
   1 Node ??X0: [??X2, ??X0, ??X3]

   cause+ed:4 [8,4,2]
   device:8   explosion:2
   an:6       incendiary:7    the:1
Evaluation

- Evaluation Criterion
  - how easy it is to write transformation rules
  - how many rules are required for intuitive lexico-syntactic reformulations
  - how robust the transformation is to parsing errors

- Methodology
  - hand-written transformation rules
  - developed looking at one third of the corpus (48 sentence)
  - tested on the remaining two thirds (96 sentence)

- Metrics
  - Recall: The proportion of sentences in the test set for which a transform was performed
  - Precision: The proportion of transformed sentence that were correct
Results

<table>
<thead>
<tr>
<th>Handcrafted rules</th>
<th>n</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-CAUSEBY-A → A-CAUSE-B</td>
<td>7</td>
<td>.82 (.00)</td>
<td>.71 (.75)</td>
<td>.76 (.86)</td>
</tr>
<tr>
<td>CAUSEOF → A-CAUSE-B</td>
<td>6</td>
<td>.97 (.90)</td>
<td>.78 (1.00)</td>
<td>.86 (.95)</td>
</tr>
<tr>
<td>BEC-A-B→ B-BEC-A</td>
<td>8</td>
<td>.85 (.92)</td>
<td>.83 (.87)</td>
<td>.84 (.89)</td>
</tr>
<tr>
<td>BEC-OF-A-B→ A-CAUSE-B</td>
<td>9</td>
<td>.75 (.92)</td>
<td>.70 (1.00)</td>
<td>.72 (.97)</td>
</tr>
</tbody>
</table>

- **Number of Rules (n)**
- **Precision, Recall and F-Measure** for lexico-syntactic reformulation using hand-crafted rules over GRs. Numbers in brackets are over the subset of the corpus that contains only the original sentences from PubMed and the BNC.
Examples

- **BEC-OF-A-B → A-CAUSE-B**
  - Because of transvection, the expression of a gene can be sensitive to the proximity of a homolog.
  - Transvection can cause the expression of a gene to be sensitive to the proximity of a homolog.

- **CAUSEOF-B-A → A-CAUSE-B**
  - Almost certainly, however, the underlying cause of the war was the problem of Aquitaine.
  - Almost certainly, however, the underlying problem of Aquitaine caused the war.
A small set of around 20 rules sufficient for simplification by clause-splitting

The List of typed dependencies can be transformed repeatedly to perform multiple simplifications

Order of application of rules does not seem to matter

Mr Powell was greeted on his arrival by Prince Saud, the Saudi foreign minister, who expressed his sorrow and vowed to cooperate with the United States in fighting terrorism.

1. Prince Saud greeted Mr Powell on his arrival
2. Prince Saud is the Saudi foreign minister
3. Prince Saud expressed his sorrow
4. And Prince Saud vowed to cooperate with the United States in fighting terrorism

4 transforms: 1 apposition, 1 relative clause, 1 conjoined VP, 1 passive voice
Conclusions

- We find typed dependency structures to be the most suited for complex reformulation tasks
  - Phrasal trees do not provide adequate level of abstraction
  - Semantic representations do not provide wide coverage
- The framework we have described is adequate for a range of regeneration tasks focused on text simplification
- In future
  - Combine with summarisation task to generate abridged texts
  - Evaluate with groups known to have reading difficulties
  - Look at Lexical Simplification
Semantic Annotation and lexico-syntactic paraphrase. LREC.

Advaith Siddharthan and Napoleon Katsos. 2010. 
Reformulating discourse connectives for non-expert readers. 
In Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010), Los Angeles, CA.

Advaith Siddharthan. 2010. 
Complex lexico-syntactic reformulation of sentences using typed dependency representations.
Text simplification using typed dependencies: A comparison of the robustness of different generation strategies.

Advaith Siddharthan.

2011.


In Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011), Nancy, France.