Text Simplification: A challenge for all Computational Linguists

Advaith Siddharthan

University of Aberdeen
Scotland
Some Background

Text Simplification (Siddharthan, 2014)

- Make information/meaning more accessible through reformulation of text Siddharthan (2006)
  - Reduce lexical or syntactic complexity
Some Background

Text Simplification (Siddharthan, 2014)

- Make information/meaning more accessible through reformulation of text Siddharthan (2006)
  - Reduce lexical or syntactic complexity
  - Reduce text length by deleting peripheral information
Text Simplification (Siddharthan, 2014)

- Make information/meaning more accessible through reformulation of text Siddharthan (2006)
  - Reduce lexical or syntactic complexity
  - Reduce text length by deleting peripheral information
  - Increase text length by adding redundancy (through reformulation, analogy, metaphor, examples...)

- Simplified Language
  - Approaches to ATS
  - Current Status
  - Challenges
Some Background

Text Simplification (Siddharthan, 2014)

- Make information/meaning more accessible through reformulation of text Siddharthan (2006)
  - Reduce lexical or syntactic complexity
  - Reduce text length by deleting peripheral information
  - Increase text length by adding redundancy (through reformulation, analogy, metaphor, examples...)
  - Make content more transparent by making discourse relations explicit
Some Background

Text Simplification (Siddharthan, 2014)

- Make information/meaning more accessible through reformulation of text Siddharthan (2006)
  - Reduce lexical or syntactic complexity
  - Reduce text length by deleting peripheral information
  - Increase text length by adding redundancy (through reformulation, analogy, metaphor, examples...)
  - Make content more transparent by making discourse relations explicit
  - Simplify argumentation
Text Simplification (Siddharthan, 2014)

- Make information/meaning more accessible through reformulation of text Siddharthan (2006)
  - Reduce lexical or syntactic complexity
  - Reduce text length by deleting peripheral information
  - Increase text length by adding redundancy (through reformulation, analogy, metaphor, examples...)
  - Make content more transparent by making discourse relations explicit
  - Simplify argumentation
  - Simplify meaning
Some Background

Text Simplification (Siddharthan, 2014)

- Make information/meaning more accessible through reformulation of text Siddharthan (2006)
  - Reduce lexical or syntactic complexity
  - Reduce text length by deleting peripheral information
  - Increase text length by adding redundancy (through reformulation, analogy, metaphor, examples...)
  - Make content more transparent by making discourse relations explicit
  - Simplify argumentation
  - Simplify meaning

- Readability can also be improved in other ways
  - Highlighting key phrases
  - Changing font, alignment, size, color etc...
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 (Siddharthan, 2010, 2011):

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.
My own work

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.

Lexical Simplification (Siddharthan & Angrosh, 2014):

The first police inquiry led to Mulcaire being jailed in 2007. The 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.
My own work

The first police inquiry led to Mulcaire being jailed in 2007. The 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.

Compression (Angrosh et al., 2014):

The police inquiry led to Mulcaire being jailed in 2007. The enquiry found proof that he has intercepted messages belonging to Rebekah Brooks. Brooks was editor of the Sun. Mulcaire was working for its Sunday stablemate.
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
Examples of Simplified Language

- Motherese: The language adults use to talk to children (Cross, 1977; Papoušek et al., 1987; Gleitman et al., 1984)
Examples of Simplified Language

- Motherese: The language adults use to talk to children (Cross, 1977; Papoušek et al., 1987; Gleitman et al., 1984)

  - Lexical:
    - reduced vocabulary
    - reduction in the number of verb inflections
    - replacement of pronouns with names
Examples of Simplified Language

- Motherese: The language adults use to talk to children (Cross, 1977; Papoušek et al., 1987; Gleitman et al., 1984)

  - Syntactic:
    - reduction of pre-verb length and complexity
    - reduction in the number of embedded clauses and conjunctions
    - shortening of utterance lengths
    - reduction in the number of disfluencies and fragments
Examples of Simplified Language

- Motherese: The language adults use to talk to children (Cross, 1977; Papoušek et al., 1987; Gleitman et al., 1984)
  - Speech: slowing of speech rate
Examples of Simplified Language

- Motherese: The language adults use to talk to children (Cross, 1977; Papoušek et al., 1987; Gleitman et al., 1984)
  - Similar observations for bilingual accommodation (e.g., Giles et al., 1973)
Examples of Simplified Language

- Controlled Language: interest from industries in creating better (less ambiguous and easier to translate) user manuals (O’Brien, 2003)
Examples of Simplified Language

- Controlled Language: interest from industries in creating better (less ambiguous and easier to translate) user manuals (O’Brien, 2003)
  - Lexical:
    - Rule out use of particular acronyms, synonyms, pronouns and ambiguous anaphoric reference, double negations
    - insist on inclusion of relative pronoun
    - standardise format for numbers and dates
    - specify dictionary, rule out ambiguous words.
Examples of Simplified Language

- Controlled Language: interest from industries in creating better (less ambiguous and easier to translate) user manuals (O’Brien, 2003)
  - Syntactic:
    - rule out ellipsis
    - insist on use of article or demonstrative
    - restrict size of noun cluster
    - specify location of prepositions to reduce ambiguity
    - rule out passive voice, insist on indicative mood
    - specify use of punctuation
Examples of Simplified Language

- Controlled Language: interest from industries in creating better (less ambiguous and easier to translate) user manuals (O’Brien, 2003)
  - Textual Structure:
    - specify when lists or tables should be used
    - constrain maximum sentence and paragraph lengths
    - specify keywords to use for coherence
    - restrict use of parentheticals.
Examples of Simplified Language

- Controlled Language: interest from industries in creating better (less ambiguous and easier to translate) user manuals (O’Brien, 2003)
  - Pragmatic:
    - rule out use of metaphor, slang or idiom
    - urge author to be as specific as possible
Texts for target reader populations

- Second language learners
  - Input hypothesis (Krashen, 1985): L2 learners acquire language when the input is comprehensible, but just a little beyond their current level of L2
  - Majority of L2 learning materials make use of simplified texts
Texts for target reader populations

- Deaf readers (Daelemans et al., 2004)
- Aphasic readers (Carroll et al., 1998)
- Children (De Belder & Moens, 2010)
- Low Literacy Adults (Specia, 2010)
Texts for target reader populations

- Deaf readers (Daelemans et al., 2004)
- Aphasic readers (Carroll et al., 1998)
- Children (De Belder & Moens, 2010)
- Low Literacy Adults (Specia, 2010)
- Dyslexic readers (Rello et al., 2013)
  - main difficulties relate to word processing or orthography
Criticisms of manual text simplification

- Can impede language acquisition
- Homogenises vocabulary across the text, and makes important information harder to identify
- Simplified texts are not interesting
- Not really simpler (use of readability formulae as guides to writing)
Questions for automatic text simplification

- Simplify through elaboration?
  - Emphasised by L2 community
  - Application to multidoc summarisation (Siddharthan et al., 2004; Nenkova et al., 2005; Siddharthan et al., 2011)?
  - Connections to paraphrase (Dorr et al., 2004; Siddharthan & McKeown, 2005)?
Questions for automatic text simplification

- Simplify through use of metaphor
  - ubiquitous in science texts for children
  - “The cell as a marketplace”, etc.
  - Also, explanations using examples, analogies, etc.
Be unambiguous and specific (Walker et al., 2011):
- but, simpler words have more senses
- but, simpler words are less specific
Questions for automatic text simplification

- Simplify morphology?
  - Why has no-one tried this!!!
Early Motivations for ATS
Early Motivations for ATS

- Reduce sentence length as a pre-processing step for a parser (Chandrasekar & Srinivas, 1997)
  - defined the task
  - first to learn syntactic rules from aligned parse trees
Early Motivations for ATS

- Extend S-TAG formalism for “Reluctant paraphrase” (Dras, 1999)
  - first to apply synchronous grammars to simplification
  - first to use ILP for controlling output characteristics
Early Motivations for ATS

- Reading aid for aphasics (Devlin & Tait, 1998; Carroll et al., 1998)
  - first to explore lexical simplification
  - first to explore pronoun replacement
Early Motivations for ATS

- Study coherence issues (Siddharthan, 2003)
  - first to detect and fix disfluencies by modelling attentional state and intentional structure
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 & Kauchak, 2011):
    - Do not attempt syntactic simplification
Recent Work

- **Text Simplification as MT (EW to SEW)**
  - PBMT (Specia, 2010; Wubben et al., 2012; Coster & 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 handle morphology, but can reorder, delete or substitute constituents
Recent Work

- **Text Simplification as MT (EW to SEW)**
  - PBMT (Specia, 2010; Wubben et al., 2012; Coster & 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 handle morphology, but can reorder, delete or substitute constituents
  - Quasi-Synchronous TSG (Woodsend & Lapata, 2011):
    - Learns lexical and syntactic 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)
  - PBMT (Specia, 2010; Wubben et al., 2012; Coster & 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 handle morphology, but can reorder, delete or substitute constituents

- Quasi-Synchronous TSG (Woodsend & Lapata, 2011):
  - Learns lexical and syntactic simplifications
  - LEX: No method for modelling lexical context for lexical simplification
  - SYN: Does not do morphology, but can reorder, delete or substitute constituents

  - PBMT for lexical rules
  - Syntactic rules using DRT representations
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**: Hybrid system using Synchronous Dependency Grammars (Siddharthan & Angrosh, 2014):
    - 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
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)
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)

**Grammar:**

\(<\text{RULE}_1, \text{DELETE}_1, \text{INSERT}_1>\)
\(<\text{RULE}_2, \text{DELETE}_2, \text{INSERT}_2>\)
...
\(<\text{RULE}_N, \text{DELETE}_N, \text{INSERT}_N>\)
**Advantages of RegenT 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 of RegenT 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 in RegenT

- 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 in RegenT

- 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)
```
Lexical Simplification
Lexical Simplification

- Lexical Substitution
  - Identify easier synonyms
  - Filter with context vector or WSD to disambiguate senses
### Lexical Simplification

- Technical terms often do not have easier synonyms
  - Explanations can be constructed using ontology relations (e.g., MESH)
    - “Pulmonary atresia” is simplified as “Pulmonary atresia (a_type_of birth defect)”
  - Definition can be obtained with “Google define:”
    - “Pulmonary atresia” is defined as “Pulmonary atresia is a form of heart disease that occurs from birth (congenital heart disease), in which the pulmonary valve does not form properly.”
ATS for different languages

- Basque (Aranzabe et al., 2012)
- Bulgarian (Lozanova et al., 2013)
- Danish (Klerke & Søgaard, 2013)
- Dutch (Daelemans et al., 2004)
- English (De Belder & Moens, 2010; Zhu et al., 2010; Coster & Kauchak, 2011; Woodsend & Lapata, 2011; Wubben et al., 2012; Siddharthan & Angrosh, 2014; Narayan & Gardent, 2014)
- French (Seretan, 2012; Brouwers et al., 2014)
- Korean (Chung et al., 2013)
- Italian (Barlacchi & Tonelli, 2013)
- Japanese (Inui et al., 2003)
- Portuguese (Aluísio et al., 2008; Watanabe et al., 2009)
- Spanish (Bott et al., 2012)
- Swedish (Smith & Jönsson, 2011; Abrahamsson et al., 2014)
Open questions and challenges

- **Lack of corpora for most languages?**
  - Can we use resources from MT?
  - PPDB generates paraphrases by pivoting over translations
    - Gigabytes of paraphrases available for different languages
    - But, noisy
    - And, no context available for lexical simplification
Open questions and challenges

- **Errorful language processing?**
  - Parser error is major cause of incorrect simplification
    - Relative clause attachment
    - Comma separated lists
Open questions and challenges

- *How good are simplified language resources?*
  - We need to better understand the quality of resources such as Simple English Wikipedia
  - Can we make use of other simplified language resources?
Open questions and challenges

- *How good does automatic text simplification need to be?*
  - The typical target reader of a text simplification system has poor reading skills.
  - Bad system output might be unusable, even when it could be understood by a fluent reader.
Open questions and challenges

- *How should text simplification systems be evaluated?*
  - few ATS studies to date with target reader populations
  - evaluations of fluency and correctness have been on a small scale
  - not clear how good text simplification systems really are.
Open questions and challenges

- How should text simplification systems be evaluated?
  - few ATS studies to date with target reader populations
  - evaluations of fluency and correctness have been on a small scale
  - not clear how good text simplification systems really are.

- Automated evaluations:
  - BLEU/NIST (Coster & Kauchak, 2011)
  - Readability scores (Louis & Nenkova, 2013)
Open questions and challenges

- How should text simplification systems be evaluated?
  - few ATS studies to date with target reader populations
  - evaluations of fluency and correctness have been on a small scale
  - not clear how good text simplification systems really are.

- Ratings by fluent readers:
  - Fluency, Simplicity, Meaning (Siddharthan, 2006; Wubben et al., 2012; Woodsend & Lapata, 2011)
Open questions and challenges

How should text simplification systems be evaluated?

- few ATS studies to date with target reader populations
- evaluations of fluency and correctness have been on a small scale
- not clear how good text simplification systems really are.

Online Methods:
- Eye-tracking (Bott et al., 2012)
Open questions and challenges

*How should text simplification systems be evaluated?*

- few ATS studies to date with target reader populations
- evaluations of fluency and correctness have been on a small scale
- not clear how good text simplification systems really are.

**Offline Methods:**

- Cloze tests (Jonnalagadda et al., 2009)
- Sentence Recall (Siddharthan & Katsos, 2012)
- MCQs (Canning, 2002; Angrosh et al., 2014)
Open questions and challenges

**How easy are text simplification systems to adapt for particular users?**

- some systems have been developed with particular target populations in mind
- other systems described are intended as general purpose
- what are the costs associated with modifying systems?
Open questions and challenges

*Can we attempt other simplification operations?*
- focus so far on lexical or syntactic simplification
- simplifying meaning or argument?
- making discourse relations explicit?
- introducing redundancy through paraphrase?
References


Seretan, V. 2012. Acquisition of Syntactic Simplification Rules for French. LREC.


