Corpora and resources for automatic text simplification: 

*The role of the target audience*

Rémi Cardon
CENTAL – UCLouvain
Automatic Text simplification
- Objectives
- Methods
- Evaluation
- Resources
- Example
- Perspectives
- Conclusion
Outline

- Automatic Text simplification (ATS)
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion
Objectives

- First ATS works: pre-processing step for other natural language processing tasks
  - Part-of-Speech tagging (Chandrasekar et al., 1996)
  - Summarization (Vale et al., 2020)
  - Machine translation (Štajner et al., 2019)
  - Information extraction / retrieval (Evans & Orasan, 2019)
Objectives

• Make information more accessible to humans
  ● Language learners (Tack et al., 2016)
  ● Adults with neurocognitive disorders (Carroll et al., 1999)
  ● Children or adults with reading difficulties (De Belder & Moens, 2010)
  ● General public – specialized texts (Cardon & Grabar, 2020)
Simplification

- Guide for human writers:
  - FALC *facile à lire et à comprendre* – easy to read and understand (Audiau, 2009)
  - OCDE (OCDE, 2015)
  - Haute Autorité de Santé \(^1\)

- Simplification initiatives:
  - UNAPEI (https://www.unapei.org/)
  - Cochrane Foundation (https://france.cochrane.org/)

- Writing recommendations:
  - Short sentences
  - No passive voice
  - Accessible words
  - Layout
  - ... \(^1\)

\(^1\)https://www.has-sante.fr/jcms/c_430286/fr/elaboration-d-un-document-ecrit-d-information-a-
Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion
Methods

- Historically, rule-based systems for syntactic simplification
- Manually design rules for syntactic simplification
  - Delete specific clause types
  - Split complex sentences into several simple sentences
  - Pronoun resolution
  - ...
- Requires efficient syntactic parsers
- Produces a lot of rules
- Costly to make and hard to maintain
Methods

- **Lexical simplification:**
  1. Complex word identification / Lexical complexity prediction: what to simplify
  2. Candidates generation: list of potential substitutes
  3. Candidate selection: keep relevant substitutes (part-of-speech, meaning in context...)
  4. Candidate ranking: identify the simplest adequate substitute
Methods

- Currently, neural methods
- End-to-end: lexical and syntactic simplification in one pass
- Let the machine learn from the data
- Require parallel corpora
  - Corpora with aligned complex / simple sentences
  - Importance of the quality of the data
Automatic Text simplification

- Objectives
- Methods
- Evaluation
- Resources
- Example
- Perspectives
- Conclusion
Three criteria are considered:

- Grammaticality: is the output grammatically correct?
- Meaning preservation: does the output express the same meaning as the input?
- Simplicity: is the output simpler than the input?
Evaluation

- **BLEU**: metric from machine translation (n-gram overlap between output and reference(s))
- **SARI**: metric designed for text simplification (n-gram overlap between input, output and reference(s))
- **Readability measures**
  - Commonly, Flesch Reading Ease and FKGL
- In the past few years, all have been shown to have poor correlation with simplicity
Evaluation

- Manual evaluation
- 5-point Likert scales for the three criteria
  - Grammaticality
  - Meaning preservation
  - Simplicity
- No consistency / consensus from one study to another
- Difficult to interpret
- Performed by researchers or crowdworkers
- Less frequently: eye-tracking, reading comprehension evaluation
• Automatic Text simplification
  • Objectives
  • Methods
  • Evaluation
  • Resources
  • Example
  • Perspectives
  • Conclusion
Resources

- Neural methods are data-hungry
- Building parallel corpora:
  - Manual alignment
  - Automatic alignment
  - Machine translation of existing corpora in other languages
  - Crowdsourcing
## Resources

### Manual approach

<table>
<thead>
<tr>
<th>Language</th>
<th>Textual genre</th>
<th>Target</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG (Pellow and Eskenazi, 2014)</td>
<td>Everyday documents</td>
<td>GP</td>
<td>200 sentence pairs</td>
</tr>
<tr>
<td>ENG (Xu et al., 2015)</td>
<td>Newspapers</td>
<td>CHI</td>
<td>96,037 original sentences</td>
</tr>
<tr>
<td>ENG (Barzilay and Elhadad, 2003)</td>
<td>Encyclopedia Britannica</td>
<td>CHI</td>
<td>2,600 easy-to-read documents</td>
</tr>
<tr>
<td>ENG (Allen, 2009)</td>
<td>Classroom materials</td>
<td>LL</td>
<td>178,967 of simplified words</td>
</tr>
<tr>
<td>ENG (Petersen and Ostendorf, 2007)</td>
<td>Newspapers</td>
<td>LL</td>
<td>2,539 original sentences</td>
</tr>
<tr>
<td>ENG (Xu et al., 2016)</td>
<td>Wikipedia</td>
<td>CS</td>
<td>2,359 original sentences</td>
</tr>
<tr>
<td>ENG (Alva-Manchego et al., 2020a)</td>
<td>Miscellaneous</td>
<td>CS</td>
<td>2,359 original sentences</td>
</tr>
<tr>
<td>Many (Orasan et al., 2013)</td>
<td>Newspapers</td>
<td>PLI</td>
<td>320 original sentences</td>
</tr>
<tr>
<td>SPA (Ett and Saggion, 2014)</td>
<td>Narrative texts</td>
<td>PLI</td>
<td>145 simplified sentences</td>
</tr>
<tr>
<td>SPA (Collados, 2013)</td>
<td>Encyclopedic, scientific, clinical texts</td>
<td>NLP</td>
<td>300 amplified sentences</td>
</tr>
<tr>
<td>FRE (Brouwers et al., 2014)</td>
<td>L1 student materials</td>
<td>L2LL</td>
<td>83 original sentences</td>
</tr>
<tr>
<td>FRE (Grabar and Cardon, 2018)</td>
<td>Newspapers</td>
<td>GP</td>
<td>4,596 sentence pairs</td>
</tr>
<tr>
<td>FRE (Gala et al., 2020)</td>
<td>Newspapers</td>
<td>PLI</td>
<td>52,704 tokens</td>
</tr>
<tr>
<td>DAN (Klerke and Søgaard, 2012)</td>
<td>Newspapers</td>
<td>L2LL</td>
<td>3,701 document pairs</td>
</tr>
<tr>
<td>POR (Caseli et al., 2009)</td>
<td>Newspapers</td>
<td>NLP</td>
<td>2,116 original sentences</td>
</tr>
<tr>
<td>POR (Aluisio et al., 2009)</td>
<td>Popular science articles</td>
<td>PLL</td>
<td>882 original sentences</td>
</tr>
<tr>
<td>GER (Klaper et al., 2013)</td>
<td>Websites</td>
<td>PLI</td>
<td>7,755 original sentences</td>
</tr>
<tr>
<td>GER (Sauberli et al., 2020)</td>
<td>Newspapers</td>
<td>L2LL</td>
<td>3,616 sentence pairs</td>
</tr>
<tr>
<td>JPN Goto et al. (2015)</td>
<td>Newspapers</td>
<td>L2LL</td>
<td>2,885 sentence pairs</td>
</tr>
<tr>
<td>EUS Gonzalez-Dios et al. (2017)</td>
<td>Popular science articles</td>
<td>L2LL</td>
<td>227 original sentences</td>
</tr>
<tr>
<td>RUS Dmitrieva and Tiedemann (2021)</td>
<td>Literary texts</td>
<td>L2LL</td>
<td>69,737 original sentences</td>
</tr>
<tr>
<td>ITA Tonelli et al. (2018)</td>
<td>Administrative texts</td>
<td>GP</td>
<td>157 original sentences</td>
</tr>
<tr>
<td>ITA Brunato et al. (2015)</td>
<td>Children’s literature</td>
<td>PLI</td>
<td>1,060 sentence pairs</td>
</tr>
<tr>
<td>ITA Brunato et al. (2015)</td>
<td>Educational material</td>
<td>L2LL</td>
<td>1,356 original pairs</td>
</tr>
</tbody>
</table>
### (Semi)Automatic Approach

<table>
<thead>
<tr>
<th>(Semi)Automatic Approach</th>
<th>Data Source</th>
<th>Approach</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG Woodsend and Lapata (2011)</td>
<td>Wikipedia</td>
<td>GP</td>
<td>14,831</td>
</tr>
<tr>
<td>ENG Botha et al. (2018)</td>
<td>Wikipedia</td>
<td>GP</td>
<td>1,004,944</td>
</tr>
<tr>
<td>ENG Pavlick and Callison-Burch (2016)</td>
<td>Miscellanea</td>
<td>CS</td>
<td>4.5 million</td>
</tr>
<tr>
<td>ITA Tonelli et al. (2016)</td>
<td>Wikipedia</td>
<td>GP</td>
<td>530</td>
</tr>
<tr>
<td>FRE Brouwers et al. (2014)</td>
<td>Wikipedia</td>
<td>L2LL</td>
<td>72</td>
</tr>
<tr>
<td>FRE Cardon and Grabar (2020)</td>
<td>Wikipedia</td>
<td>GP</td>
<td>297,494</td>
</tr>
<tr>
<td>ITA Brunato et al. (2016)</td>
<td>Web corpus</td>
<td>GP</td>
<td>63,000</td>
</tr>
</tbody>
</table>

Resources

- Training corpora: scrapped from the Internet (mostly Wikipedia)
- Evaluation corpora:
  - Subsets of training corpora
  - Crowdsourced
    - End-to-end: ask annotators to manually simplify a relatively small number of sentences
    - Lexical simplification: ask annotators to propose substitutes
• Existing but not yet taken into account: typologies of transformations
• Two types of transformations:
  • Linguistic transformations
  • String edits
<table>
<thead>
<tr>
<th></th>
<th>Lexical Simplification</th>
<th>Syntactic Simplification</th>
<th>Discourse-level Simplification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dyslexia</strong></td>
<td>Long words</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Less frequent words</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Homophones</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Words that are orthographically similar</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>New words</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-words</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Aphasia</strong></td>
<td>Information density</td>
<td>Long sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noun compounds</td>
<td>Long sequences of adjectives</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive voice</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Object relative clauses</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison of word meaning</td>
<td></td>
</tr>
<tr>
<td><strong>Intellectual Disability (ID)</strong></td>
<td>Limited vocabulary</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Deaf and Hard-of-Hearing</strong></td>
<td>Limited vocabulary</td>
<td>Complex sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Morphology</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Syntax</td>
<td></td>
</tr>
<tr>
<td><strong>Autism Spectrum Disorder</strong></td>
<td>Words related to emotions</td>
<td></td>
<td>Figurative language</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Texts that require little social knowledge</td>
</tr>
<tr>
<td><strong>Second Language Learners</strong></td>
<td>Limited vocabulary</td>
<td></td>
<td>Tight text structure</td>
</tr>
<tr>
<td><strong>Children</strong></td>
<td>New words</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Limited vocabulary</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion
French biomedical text simplification
Medical corpus (CLEAR) – 4,500 sentence pairs:
  - Wikipedia/Vikidia
  - Drug leaflets (for medical practitioners and patients)
  - Medical literature reviews, and manually simplified versions
Additional data (WikiLarge) – 300,000 sentence pairs
Paraphrases – 4,516 medical terms paraphrased
Questions:
  - Do we really need large amounts of data?
  - Do we need specialized data for specialized text simplification?
  - Can a resource of a different linguistic nature (paraphrases) help?
Experiment

- Algorithm
- OpenNMT-py
- Encoder-decoder with attention ²:
  - Two bi-LSTM layers of 500 cells for encoder and decoder
  - ADAM optimizer
  - learning rate 0.001
  - dropout probability 0.3
  - attention dropout probability 0.2
- Use of the paraphrase table:
  - SL and LE: –replace_unk
  - LS: –phrase_table with the paraphrase table

²Hyperparameters inspired by Abdul Rauf et al., (2020)
No paraphrase table (SL)
Table used during Simplification (LS)
Table used during training (LE)
<table>
<thead>
<tr>
<th>Modèle</th>
<th>Test sur WikiLarge FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Le 14 octobre 1960, le candidat à la présidence John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l’Union du Michigan.</td>
</tr>
<tr>
<td>Référence</td>
<td>John F. Kennedy, un candidat à la présidence, a proposé l’idée de ce qui devint le Peace Corps sur les marches de l’Union du Michigan le 14 octobre 1960.</td>
</tr>
<tr>
<td>WikiLarge</td>
<td><em>No change</em></td>
</tr>
<tr>
<td>CLEAR</td>
<td>le cancer est de la médecine</td>
</tr>
<tr>
<td>SL 1:75</td>
<td><em>Aucun changement</em></td>
</tr>
<tr>
<td>LE 1:25</td>
<td>Le 14 octobre 1960, le candidat <em>au</em> président John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l’Union du Michigan.</td>
</tr>
<tr>
<td>LE 1:50</td>
<td>En 1960, le candidat à la présidence John F. Kennedy a proposé l’idée de ce qui est devenu le Peace Corps sur les marches de l’Union du Michigan.</td>
</tr>
<tr>
<td>LE 1:75</td>
<td>John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l’Union du Michigan.</td>
</tr>
<tr>
<td>Model</td>
<td>Test on CLEAR</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
</tr>
<tr>
<td>Source</td>
<td>une hypotension artérielle peut être observée en cas d’administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 4.2)</td>
</tr>
<tr>
<td>Référence</td>
<td>une hypotension artérielle peut être observée en cas d’administration intraveineuse trop rapide, inférieure à 60 min</td>
</tr>
<tr>
<td>WikiLarge</td>
<td>Une artérielle artérielle peut être observée en cas de crise intraveineuse trop rapide et inférieure à 60 minutes</td>
</tr>
<tr>
<td>CLEAR</td>
<td>le traitement de la naissance de la naissance de l’repos [...] de la médecine [...] de la médecine de la peau [...] de la peau de la</td>
</tr>
<tr>
<td>SL 1:50 &amp; SL 1 : 75</td>
<td>une hypotension artérielle peut être observée en cas d’administration intraveineuse trop rapide, inférieure à 60 minutes</td>
</tr>
<tr>
<td>LS 1:75</td>
<td>une tension inférieure à la normale artérielle peut être observée en cas d’administration intraveineuse trop rapide, inférieure à 60 minutes</td>
</tr>
<tr>
<td>LE 1:50</td>
<td>une hypotension artérielle peut être observée en cas d’administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 3)</td>
</tr>
<tr>
<td>LE 1:75</td>
<td>une diminution de la tension artérielle peut être observée en cas d’administration intraveineuse trop rapide, inférieure à 60 minutes</td>
</tr>
</tbody>
</table>
Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion
Perspectives – Objectives

- Go beyond the sentence level
  - Some works have begun exploring paragraph simplification
- Personalized text simplification
Perspectives – Methods

- Design less data-hungry methods (unsupervised)
- Add constraints to the text generation step
  - Sentence length
  - Vocabulary used
  - Syntactic constraints
  - ...

Perspectives

- Design better evaluation metrics
  - Adequacy with the target audience
  - Adequacy with the text genre
- Involve members of the target audience in the process
Perspectives

- Build resources with members of the target audiences
- Build resources with experts (writers) of accessible language for a given audience
Outline

• Automatic Text simplification
  • Objectives
  • Methods
  • Evaluation
  • Resources
  • Example
  • Perspectives
  • Conclusion
Conclusion

- Current methods begin to be able to produce good paraphrases according to some surface criteria...
  - ... but which criteria for which audience?
  - ... but evaluating simplicity is still an open question

- Need for more inclusion of members of target audiences, starting at the design step
- Need for input from other fields (psycholinguistics, sociolinguistics...)