Génération de textes
A partir de textes et de graphes

Chaire IA xNLG: Generating from Multiple Sources into Multiple Languages

CNRS / LORIA
Outline

Text → Text

• Document level Simplification
Outline

Text → Text

- Document level Simplification

Graph → Text
Outline

Text → Text

- Document level Simplification

Graph → Text

- Abstract Meaning Representation (AMR) → 21 EU languages
- RDF graph → English
- RDF graph → Low Resource Languages (Breton, Irish, Welsh)
Document Simplification

Cripwell et al. EMNLP Findings 2021, NAACL Findings 2022, EACL 2023, ACL Findings 2023
Example

Complex Input Document

Owls are birds from the order of Strigiformes, comprising over 200 species of mostly solitary and nocturnal birds of prey typified by an upright stance, binocular vision, binaural hearing, and sharp talons. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

Simplified Output Document

Owls are birds. There are over 200 species and are all animals of prey. Most of them are solitary and nocturnal. Owls’ prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

*Avg nb of sentences in Input Document: 39*
Why Simplify?

To aid reader comprehension (Mason, 1978; Williams et al., 2003; Kajiwara et al., 2013)

- Adult vs children
- Native vs non Native
- Reading disability
- Expert vs non-Expert
Simplification Operations

*Owls are birds from the order of Strigiformes, comprising over 200 species* of mostly solitary and nocturnal birds of prey typified by an upright stance, binocular vision, binaural hearing, and sharp talons. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

*Owls are birds. There are over 200 species ...*
Simplification Operations

Owls are birds from the order of Strigiformes, comprising over 200 species of *mostly solitary and nocturnal birds of prey* typified by an upright stance, binocular vision, binaural hearing, and sharp talons. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

Owls are birds. There are over 200 species *and are all animals of prey. Most of them are solitary and nocturnal*.
Simplification Operations

Owls are birds from the order of Strigiformes, comprising over 200 species of mostly solitary and nocturnal birds of prey typified by an upright stance, binocular vision, binaural hearing, and sharp talons. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

Owls are birds. There are over 200 species and are all animals of prey. Most of them are solitary and nocturnal. Owls’ prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).
Previous Work

Mostly on *Sentence Simplification*
Previous Work

Mostly on *Sentence Simplification*

Some work on Sentence-level techniques iteratively applied over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b)
Previous Work

 Mostly on *Sentence Simplification*

Some work on Sentence-level techniques iteratively applied over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b)

*Insufficient for maintaining the discourse coherence of the document (Siddharthan, 2003; Alva-Manchego et al., 2019b).*
Previous Work

Mostly on *Sentence Simplification*

Some work on Sentence-level techniques iteratively applied over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b)

*Insufficient for maintaining the discourse coherence of the document (Siddharthan, 2003; Alva-Manchego et al., 2019b).*

A sentence-level model that uses context information to influence document simplification (Sun et al. 2020)
Previous Work

Mostly on *Sentence Simplification*

Some work on Sentence-level techniques iteratively applied over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b)

*Insufficient for maintaining the discourse coherence of the document (Siddharthan, 2003; Alva-Manchego et al., 2019b).*

A sentence-level model that uses context information to influence document simplification (Sun et al. 2020)

*Unable to outperform the baseline* (Sun et al. 2021)
Our Proposal

Plan-Guided, Context-Aware Document Simplification

PLAN + SIMPLIFY
Our Proposal

*Plan-Guided, Context-Aware Document Simplification*

**PLAN + SIMPLIFY**

PLAN: predict a simplification operation for each sentence in the input document
Our Proposal

Plan-Guided, Context-Aware Document Simplification

PLAN + SIMPLIFY

PLAN: predict a simplification operation for each sentence in the input document

\[ c_1, \ldots, c_n \Rightarrow \hat{o}_1, \ldots, \hat{o}_n \]
Our Proposal

Plan-Guided, Context-Aware Document Simplification

PLAN + SIMPLIFY

PLAN: predict a simplification operation for each sentence in the input document

\[ c_1, \ldots, c_n \Rightarrow \hat{o}_1, \ldots, \hat{o}_n \]

\[ c_1, \ldots, c_n \Rightarrow \text{split, delete, split, rephrase, rephrase, copy} \]
Our Proposal

Plan-Guided, Context-Aware Document Simplification

PLAN + SIMPLIFY

PLAN: predict a simplification operation for each sentence in the input document

\[ c_1, \ldots, c_n \Rightarrow \hat{o}_1, \ldots, \hat{o}_n \]

\[ c_1, \ldots, c_n \Rightarrow \text{split, delete, split, rephrase, rephrase, copy} \]

SIMPLIFY: Controlled Sentence Simplification

\[ c_i, \hat{o}_i \Rightarrow s_i \]
Our Proposal

*Plan-Guided, Context-Aware Document Simplification*

**PLAN + SIMPLIFY**

**PLAN:** predict a simplification operation for each sentence in the input document

\[ c_1, \ldots, c_n \Rightarrow \hat{o}_1, \ldots, \hat{o}_n \]

\[ c_1, \ldots, c_n \Rightarrow \text{split, delete, split, rephrase, rephrase, copy} \]

**SIMPLIFY:** Context-Aware, Controlled Sentence Simplification

\[ Context, c_i, \hat{o}_i \Rightarrow s_i \]
Planning Simplification Operations

\[ c_1, \ldots, c_n \Rightarrow \hat{c}, \ldots, \hat{c}_n \]
Challenges

Simplification Operations have different requirements

Splitting

- mainly depends on the *input sentence’s internal structure*

  *The man who sleeps snores* → *The man sleeps. He snores.*

  *John went shopping after he left work* → *John left work. Afterwards he went shopping.*
Challenges

Simplification Operations have different requirements

Splitting

- mainly depends on the *input sentence’s internal structure*

  *The man who sleeps snores → The man sleeps. He snores.*

  *John went shopping after he left work → John left work. Afterwards he went shopping.*

Deletion, copy and rephrase

- are mostly *context dependent*.

  A sentence can only be omitted if it is either *redundant* with, or of *minor semantic import* relative to, other sentences in the document
Planning Model

RoBERTa classifier with cross-attention over the context

- layers initialised with weights from a context-independent classifier
Planning Model

RoBERTa classifier with cross-attention over the context

- layers initialised with weights from a context-independent classifier

Internal structure

- *Token level* encoder for $c_i$
Planning Model

RoBERTa classifier with cross-attention over the context

- layers initialised with weights from a context-independent classifier

Internal structure

- *Token level* encoder for $c_i$

Context

- fixed window of Sentence level embedding (SBERT) for *surrounding sentences*
Planning Model

RoBERTa classifier with cross-attention over the context

- layers initialised with weights from a context-independent classifier

Internal structure

- **Token level** encoder for $c_i$

Context

- fixed window of Sentence level embedding (SBERT) for *surrounding sentences*
- The left context is *dynamically* updated with previously simplified sentences
Planning Model

RoBERTa classifier with cross-attention over the context

- layers initialised with weights from a context-independent classifier

Internal structure

- **Token level** encoder for $c_i$

Context

- fixed window of Sentence level embedding (SBERT) for *surrounding sentences*
- The left context is **dynamically** updated with previously simplified sentences

Context positional embedding: relative distance of a given sentence from the input sentence $c_i$

Document positional embedding: the document quintile (1-5) that a given sentence falls into
**Dynamic Contextual Classifier**: our model

**Contextual Classifier**: Static left context

**Classifier**: no context

**Tagger**: Sequence tagging on SBERT representations (no internal structure)

**Tagger-Decoder**: Each prediction is conditioned on the input document and on the previously predicted operation tags. SBERT encodings.

**EncDec \texttt{full}**: Same as Tagger-Decoder but with token encodings
Data

$(C, S)$ pairs with $C$ a complex document and $S$ its simplification (sentences are aligned)

<table>
<thead>
<tr>
<th></th>
<th>Wiki-auto</th>
<th>Newsela-auto</th>
</tr>
</thead>
<tbody>
<tr>
<td># Doc Pairs</td>
<td>85,123</td>
<td>18,319</td>
</tr>
<tr>
<td># Sent Pairs</td>
<td>461,852</td>
<td>707,776</td>
</tr>
<tr>
<td>Avg. $</td>
<td>C</td>
<td>$</td>
</tr>
<tr>
<td>Avg. $</td>
<td>S</td>
<td>$</td>
</tr>
<tr>
<td>Avg. $c_i$</td>
<td>28.64</td>
<td>22.49</td>
</tr>
<tr>
<td>Avg. $s_i$</td>
<td>21.57</td>
<td>15.84</td>
</tr>
<tr>
<td>Avg. $n$</td>
<td>5.43</td>
<td>38.64</td>
</tr>
<tr>
<td>Avg. $k$</td>
<td>4.53</td>
<td>42.60</td>
</tr>
</tbody>
</table>

- $n$: the number of sentences in $C$
- $k$: the number of sentences in $S$
Labeling the data

\[(C, S) \rightarrow (C, S, o)\]

Delete

- \(c_i\) is not aligned to any \(s_j\).
  The complex sentence \(c_i\) is not aligned to any sentence \(s_j\) in the simplified version.
Labeling the data

\((C, S) \rightarrow (C, S, o)\)

Delete

- \(c_i\) is not aligned to any \(s_j\).

Copy

- \(c_i\) is aligned to a single \(s_j\) with a Levenshtein similarity above 0.92.
  
  The complex sentence \(c_i\) is aligned to a similar sentence \(s_j\) in the simplified version.
Labeling the data

\((C, S) \rightarrow (C, S, o)\)

Delete

- \(c_i\) is not aligned to any \(s_j\).

Copy

- \(c_i\) is aligned to a single \(s_j\) with a Levenshtein similarity above 0.92.

Rephrase

- \(c_i\) is aligned to a single \(s_j\) with a Levenshtein similarity below 0.92.
  The complex sentence \(c_i\) is aligned to a sentence \(s_j\) in the simplified version but differs from it.
Labeling the data

\[(C, S) \rightarrow (C, S, o)\]

Delete

- \(c_i\) is not aligned to any \(s_j\).

Copy

- \(c_i\) is aligned to a single \(s_j\) with a Levenshtein similarity above 0.92.

Rephrase

- \(c_i\) is aligned to a single \(s_j\) with a Levenshtein similarity below 0.92.

Split

- \(c_i\) is aligned to multiple \(s_j\)
  - The complex sentence \(c_i\) is aligned to several sentences in the simplified version.
## Planning Accuracy Results

<table>
<thead>
<tr>
<th></th>
<th>Wiki-auto</th>
<th>Newsela-auto</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td><strong>C</strong></td>
<td><strong>R</strong></td>
</tr>
<tr>
<td>EncDec_all</td>
<td>26.9</td>
<td>42.2</td>
</tr>
<tr>
<td>EncDec</td>
<td>29.3</td>
<td>54.5</td>
</tr>
<tr>
<td>Tagger</td>
<td>38.6</td>
<td>54.2</td>
</tr>
<tr>
<td>Classifier</td>
<td>42.1</td>
<td>52.9</td>
</tr>
<tr>
<td>Dyn. Context</td>
<td><strong>44.8</strong></td>
<td><strong>57.9</strong></td>
</tr>
<tr>
<td>+ docpos</td>
<td>43.7</td>
<td>55.4</td>
</tr>
</tbody>
</table>

- Our model (Dyn. Context) shows best results
# Planning Accuracy Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Wiki-auto</th>
<th></th>
<th></th>
<th></th>
<th>Micro</th>
<th>Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td>EncDec_all</td>
<td>26.9</td>
<td>42.2</td>
<td>36.0</td>
<td>51.8</td>
<td>43.2</td>
<td>40.8</td>
</tr>
<tr>
<td>EncDec</td>
<td>29.3</td>
<td>54.5</td>
<td>30.0</td>
<td>51.8</td>
<td>47.7</td>
<td>41.4</td>
</tr>
<tr>
<td>Tagger</td>
<td>38.6</td>
<td>54.2</td>
<td>31.7</td>
<td>58.5</td>
<td>50.6</td>
<td>45.8</td>
</tr>
<tr>
<td>Classifier</td>
<td>42.1</td>
<td>52.9</td>
<td>42.6</td>
<td>49.0</td>
<td>48.4</td>
<td>46.7</td>
</tr>
<tr>
<td>Dyn. Context</td>
<td><strong>44.8</strong></td>
<td><strong>57.9</strong></td>
<td>42.4</td>
<td>54.8</td>
<td><strong>52.8</strong></td>
<td><strong>50.0</strong></td>
</tr>
<tr>
<td>+ docpos</td>
<td>43.7</td>
<td>55.4</td>
<td><strong>43.6</strong></td>
<td>56.7</td>
<td>52.3</td>
<td>49.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newsele-auto</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>26.1</td>
<td>10.8</td>
<td>11.7</td>
<td>9.0</td>
<td>12.2</td>
<td>11.5</td>
</tr>
<tr>
<td></td>
<td>72.2</td>
<td>73.9</td>
<td>75.9</td>
<td>79.7</td>
<td>75.0</td>
<td>75.4</td>
</tr>
<tr>
<td></td>
<td>71.4</td>
<td>72.7</td>
<td>74.1</td>
<td>78.4</td>
<td>73.7</td>
<td>74.1</td>
</tr>
<tr>
<td></td>
<td>77.0</td>
<td>75.6</td>
<td>80.0</td>
<td>78.5</td>
<td>77.4</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td><strong>79.3</strong></td>
<td><strong>77.3</strong></td>
<td>82.8</td>
<td>81.4</td>
<td>79.7</td>
<td>80.2</td>
</tr>
<tr>
<td></td>
<td><strong>80.0</strong></td>
<td><strong>78.1</strong></td>
<td><strong>83.6</strong></td>
<td><strong>82.0</strong></td>
<td><strong>80.3</strong></td>
<td><strong>80.8</strong></td>
</tr>
</tbody>
</table>

- **Our model (Dyn. Context) shows best results**
- **Deletion needs context**
  
  *the context-free classifier (Classifier) under-performs for deletions*
Planning Accuracy Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Wiki-auto</th>
<th>Newsel-a-auto</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>R</td>
</tr>
<tr>
<td>EncDec_all</td>
<td>26.9</td>
<td>42.2</td>
</tr>
<tr>
<td>EncDec</td>
<td>29.3</td>
<td>54.5</td>
</tr>
<tr>
<td>Tagger</td>
<td>38.6</td>
<td>54.2</td>
</tr>
<tr>
<td>Classifier</td>
<td>42.1</td>
<td>52.9</td>
</tr>
<tr>
<td>Dyn. Context + docpos</td>
<td>44.8</td>
<td>57.9</td>
</tr>
</tbody>
</table>

- Our model (Dyn. Context) shows best results

- Deletion needs context

  *the context-free classifier (Classifier) under-performs for deletions*

- A token-level encoding of the complex sentence is important

  *The encoder-decoder and the tagger, which both use a sentence level encoding of the complex sentence, underperform*
Ablations

<table>
<thead>
<tr>
<th>Model</th>
<th>Copy</th>
<th>Rephrase</th>
<th>Split</th>
<th>Delete</th>
<th>Micro</th>
<th>Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>(a) Ablation on Best Model</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyn, $r = 13$, +init, +docpos</td>
<td>80.0</td>
<td>78.1</td>
<td>83.6</td>
<td>82.0</td>
<td>80.3</td>
<td>80.8</td>
</tr>
<tr>
<td>-docpos</td>
<td>79.3</td>
<td>77.3</td>
<td>82.8</td>
<td>81.4</td>
<td>79.7</td>
<td>80.2</td>
</tr>
<tr>
<td>-init</td>
<td>74.9</td>
<td>72.1</td>
<td>77.8</td>
<td>75.2</td>
<td>74.6</td>
<td>75.0</td>
</tr>
<tr>
<td>-init, -docpos</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td><em>(b) Dynamic vs. Static Context</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stat, $r = 9$</td>
<td>71.3</td>
<td>69.5</td>
<td>75.4</td>
<td>73.3</td>
<td>72.0</td>
<td>72.4</td>
</tr>
<tr>
<td>Stat, $r = 13$</td>
<td>72.2</td>
<td>65.3</td>
<td>69.9</td>
<td>68.3</td>
<td>68.5</td>
<td>68.9</td>
</tr>
<tr>
<td>Dyn, $r = 9$</td>
<td>73.1</td>
<td>70.1</td>
<td>75.5</td>
<td>75.9</td>
<td>73.1</td>
<td>73.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td><em>(c) With vs without Initialisation</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyn, $r = 9$</td>
<td>73.1</td>
<td>70.1</td>
<td>75.5</td>
<td>75.9</td>
<td>73.1</td>
<td>73.6</td>
</tr>
<tr>
<td>Dyn, $r = 9$ +init</td>
<td>79.3</td>
<td>78.0</td>
<td>82.7</td>
<td>79.8</td>
<td>79.7</td>
<td>80.0</td>
</tr>
<tr>
<td>Dyn, $r = 13$</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$ +init</td>
<td>79.3</td>
<td>77.3</td>
<td>82.8</td>
<td>81.4</td>
<td>79.7</td>
<td>80.2</td>
</tr>
<tr>
<td><em>(d) Window Size</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stat, $r = 9$</td>
<td>71.3</td>
<td>69.5</td>
<td>75.4</td>
<td>73.3</td>
<td>72.0</td>
<td>72.4</td>
</tr>
<tr>
<td>Stat, $r = 13$</td>
<td>72.2</td>
<td>65.3</td>
<td>69.9</td>
<td>68.3</td>
<td>68.5</td>
<td>68.9</td>
</tr>
<tr>
<td>Dyn, $r = 9$</td>
<td>73.1</td>
<td>70.1</td>
<td>75.5</td>
<td>75.9</td>
<td>73.1</td>
<td>73.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td>Dyn, $r = 9$ +docpos</td>
<td>73.8</td>
<td>72.9</td>
<td>77.2</td>
<td>75.8</td>
<td>74.6</td>
<td>74.9</td>
</tr>
<tr>
<td>Dyn, $r = 13$ +docpos</td>
<td>74.9</td>
<td>72.1</td>
<td>77.8</td>
<td>75.2</td>
<td>74.6</td>
<td>75.0</td>
</tr>
<tr>
<td>Dyn, $r = 9$ +init +docpos</td>
<td>79.4</td>
<td>78.0</td>
<td>83.1</td>
<td>82.0</td>
<td>80.1</td>
<td>80.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$ +init +docpos</td>
<td>80.0</td>
<td>78.1</td>
<td>83.6</td>
<td>82.0</td>
<td>80.3</td>
<td>80.8</td>
</tr>
</tbody>
</table>
Plan-Guided Simplification

$c_1, \ldots, c_n \Rightarrow \hat{o}_1, \ldots, \hat{o}_n$

$c_i, \hat{o}_i \Rightarrow s_i$
Predict simplification operations

\[ c_1, \ldots, c_n \Rightarrow \hat{o}_1, \ldots, \hat{o}_n \]

Simplify each input sentences using controls

\[ c_i, \hat{o}_i \Rightarrow s_i \]
Models

Fine-tuned on sentence pairs and iteratively applied to each input sentence

- Plan-Guided (PG): pipeline

\[ D \rightarrow \hat{o}_1, \ldots, \hat{o}_n \]

\[ c_i, \hat{o}_i \Rightarrow s_i \]
Models

Fine-tuned on *sentence pairs* and iteratively applied to each input sentence

- Plan-Guided (PG): *pipeline*

  \[ D \rightarrow \hat{o}_1, \ldots, \hat{o}_n \]

  \[ c_i, \hat{o}_i \Rightarrow s_i \]

- Sent-BART: *end-to-end*

  \[ c_i \Rightarrow s_i \]
Models

Fine-tuned on sentence pairs and iteratively applied to each input sentence

- Plan-Guided (PG): pipeline

\[ D \rightarrow \hat{c}_1, \ldots, \hat{c}_n \]
\[ c_i, \hat{c}_i \Rightarrow s_i \]

- Sent-BART: end-to-end

\[ c_i \Rightarrow s_i \]

Fine-tuned on full document pairs

- Doc-BART

\[ DOC \Rightarrow SIMPLIFIED \]
Evaluation Metrics

Summarization metrics

- BARTScore (Yuan et al., 2021)
- SMART (Amplayo et al., 2022)

SARI (Xu et al., 2016)

- Most popular simplification metric.
- Computes n-gram edits between input, output, and references.

FKGL (Kincaid et al., 1975)

- Readibility metrics
- Uses surface-level statistics like syllable counts and sentence length.
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>BARTScore $\uparrow$</th>
<th>SMART $\uparrow$</th>
<th>FKGL $\downarrow$</th>
<th>SARI $\uparrow$</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Faith. $(s \rightarrow h)$</td>
<td>P $(r \rightarrow h)$</td>
<td>R $(h \rightarrow r)$</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>Input</td>
<td>-0.93</td>
<td>-2.47</td>
<td>-1.99</td>
<td>-2.23</td>
<td>63.2</td>
</tr>
<tr>
<td>Reference</td>
<td>-1.99</td>
<td>-0.93</td>
<td>-0.93</td>
<td>-0.93</td>
<td>100</td>
</tr>
<tr>
<td>Doc-BART</td>
<td>-2.48</td>
<td>-2.68</td>
<td>-2.76</td>
<td>-2.72</td>
<td>61.9</td>
</tr>
<tr>
<td>Sent-BART</td>
<td>-1.86</td>
<td>-1.63</td>
<td>-1.56</td>
<td>-1.60</td>
<td>78.9</td>
</tr>
<tr>
<td>PG$_{Tag}$</td>
<td>-1.95</td>
<td>-2.22</td>
<td>-2.18</td>
<td>-2.20</td>
<td>5.07</td>
</tr>
<tr>
<td>PG$_{Enc/Dec}$</td>
<td>-1.94</td>
<td>-2.22</td>
<td>-2.18</td>
<td>-2.20</td>
<td>62.2</td>
</tr>
<tr>
<td>PG$_{CH}$</td>
<td>-1.91</td>
<td>-1.68</td>
<td>-1.53</td>
<td>-1.60</td>
<td>77.8</td>
</tr>
<tr>
<td>PG$_{Dyn}$</td>
<td>-1.91</td>
<td>-1.60</td>
<td>-1.54</td>
<td>-1.57</td>
<td>80.2</td>
</tr>
<tr>
<td>PGO$_{Oracle}$</td>
<td>-1.93</td>
<td>-1.39</td>
<td>-1.40</td>
<td>-1.40</td>
<td>85.5</td>
</tr>
</tbody>
</table>

- **Pipeline** (PG Dyn) achieves the highest results of all systems.
Results

<table>
<thead>
<tr>
<th>System</th>
<th>BARTScore ↑</th>
<th>SMART ↑</th>
<th>FKGL ↓</th>
<th>SARI ↑</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Faith. (s → h)</td>
<td>P (r → h)</td>
<td>R (h → r)</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>Input</td>
<td>-0.93</td>
<td>-2.47</td>
<td>-1.99</td>
<td>-2.23</td>
<td>63.2</td>
</tr>
<tr>
<td>Reference</td>
<td>-1.99</td>
<td>-0.93</td>
<td>-0.93</td>
<td>-0.93</td>
<td>100</td>
</tr>
<tr>
<td>Doc-BART</td>
<td>-2.48</td>
<td>-2.68</td>
<td>-2.76</td>
<td>-2.72</td>
<td>61.9</td>
</tr>
<tr>
<td>Sent-BART</td>
<td><strong>-1.86</strong></td>
<td>-1.63</td>
<td>-1.56</td>
<td>-1.60</td>
<td>78.9</td>
</tr>
<tr>
<td><strong>PGTag</strong></td>
<td>-1.95</td>
<td>-2.22</td>
<td>-2.18</td>
<td>-2.20</td>
<td>5.07</td>
</tr>
<tr>
<td><strong>PGEnc/Dec</strong></td>
<td>-1.94</td>
<td>-2.22</td>
<td>-2.18</td>
<td>-2.20</td>
<td>62.2</td>
</tr>
<tr>
<td><strong>PGCIF</strong></td>
<td>-1.91</td>
<td>-1.68</td>
<td><strong>-1.53</strong></td>
<td>-1.60</td>
<td>77.8</td>
</tr>
<tr>
<td><strong>PGDyn</strong></td>
<td>-1.91</td>
<td><strong>-1.60</strong></td>
<td>-1.54</td>
<td><strong>-1.57</strong></td>
<td><strong>80.2</strong></td>
</tr>
<tr>
<td><strong>PGOracle</strong></td>
<td>-1.93</td>
<td><strong>-1.39</strong></td>
<td><strong>-1.40</strong></td>
<td><strong>-1.40</strong></td>
<td><strong>85.5</strong></td>
</tr>
</tbody>
</table>

- **Pipeline** (PG Dyn) achieves the highest results of all systems.

- **Improving planning** (PG Oracle) would substantially increase performance (PG Oracle)
Results

<table>
<thead>
<tr>
<th>System</th>
<th>BARTScore ↑</th>
<th>SMART ↑</th>
<th>FKGL ↓</th>
<th>SARI ↑</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Faith. (s → h)</td>
<td>P (r → h)</td>
<td>R (h → r)</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>Input</td>
<td>-0.93</td>
<td>-2.47</td>
<td>-1.99</td>
<td>-2.23</td>
<td>63.2</td>
</tr>
<tr>
<td>Reference</td>
<td>-1.99</td>
<td>-0.93</td>
<td>-0.93</td>
<td>-0.93</td>
<td>100</td>
</tr>
<tr>
<td>Doc-BART</td>
<td>-2.48</td>
<td>-2.68</td>
<td>-2.76</td>
<td>-2.72</td>
<td>61.9</td>
</tr>
<tr>
<td>Sent-BART</td>
<td>-1.86</td>
<td>-1.63</td>
<td>-1.56</td>
<td>-1.60</td>
<td>78.9</td>
</tr>
<tr>
<td>PGTag</td>
<td>-1.95</td>
<td>-2.22</td>
<td>-2.18</td>
<td>-2.20</td>
<td>5.07</td>
</tr>
<tr>
<td>PGE_enc Dec</td>
<td>-1.94</td>
<td>-2.22</td>
<td>-2.18</td>
<td>-2.20</td>
<td>62.2</td>
</tr>
<tr>
<td>PGCI</td>
<td>-1.91</td>
<td>-1.68</td>
<td>-1.53</td>
<td>-1.60</td>
<td>77.8</td>
</tr>
<tr>
<td>PGDyn</td>
<td>-1.91</td>
<td>-1.60</td>
<td>-1.54</td>
<td>-1.57</td>
<td>80.2</td>
</tr>
<tr>
<td>PGOOracle</td>
<td>-1.93</td>
<td>-1.39</td>
<td>-1.40</td>
<td>-1.40</td>
<td>85.5</td>
</tr>
</tbody>
</table>

- **Pipeline** (PG Dyn) achieves the highest results of all systems.

- **Improving planning** (PG Oracle) would substantially increase performance (PG Oracle)

- **E2E simplification of full document** (Doc-BART) yields poor results
Context-Aware, Plan-Guided Document Simplification

\[ c_1, \ldots, c_n \Rightarrow \hat{o}_1, \ldots, \hat{o}_n \]

\[ \text{Context}, c_i, \hat{o}_i \Rightarrow s_i \]
Context-Aware, Plan-Guided Simplification

PG (plan-guided) pipeline

First PLAN,
Input $D \Rightarrow$ Simplification Plan
$c_1, \ldots, c_n \Rightarrow \hat{o}, \ldots, \hat{o}_n$

then SIMPLIFY
Input $S +$ Simplification Operation $\Rightarrow$ Simplified $S$
$c_i, \hat{o}_i \Rightarrow s_i$

... but SIMPLIFICATION is not
ConBART: Context-Aware Sentence Simplification

- Modification of the BART architecture
- Generation is conditioned on both an input sentence $c_i$ and a representation of the document context $Z_i$ of that sentence
- Same context modeling as for planner (SBERT encoding of the neighbouring sentences)
Models

Text-Only Models (BART, LED)

- input = sentence, paragraph or document
- Model: LongFormer (LED)
Models

Text-Only Models (BART, LED)

- input = sentence, paragraph or document
- Model: LongFormer (LED)

Contextual Model (ConBART)

- Input: sentence + context window of \( n \) sentences
- Model: Context-aware, controlled sentence simplification (ConBART)
Models

Text-Only Models (BART, LED)

- input = sentence, paragraph or document
- Model: LongFormer (LED)

Contextual Model (ConBART)

- Input: sentence + context window of \( n \) sentences
- Model: Context-aware, controlled sentence simplification (ConBART)

Plan-Guided Pipelines (\( \hat{O} \rightarrow M \))

- \( \hat{O} \), a predicted simplification plan
- \( M \), a simplification model (BART, LED, ConBART)
Which context helps most?

- The best two models use a medium size context (ConBART, LED_{para})
Which context helps most?

- The best two models use a medium size context (ConBART, LED<sub>para</sub>)
- Full Document context does not work well (BART<sub>doc</sub>, LED<sub>doc</sub>)
Which context helps most?

- The best two models use a medium size context (ConBART, LED\textsubscript{para})
- Full Document context does not work well (BART\textsubscript{doc}, LED\textsubscript{doc})
- LongFormers improve results (BART\textsuperscript{X} vs. LED\textsuperscript{X})
Does planning help?

- Planning systematically improves performance
Does planning help?

- Planning systematically improves performance
- Planning needs improving

*Models simplifying based on the oracle plan have much higher performance*
Human Evaluation

On 198 paragraphs, Binary question, Score = Proportion of yes for each model
Comparison: $PG_{Dyn}$, $LED_{para}$, $\hat{O} \rightarrow LED_{para}$, $\hat{O} \rightarrow ConBART$

Fluency

All systems achieve high fluency (92.9% - 95.5%).

Semantic Adequacy

$\hat{O} \rightarrow LED_{para}$ is best (81.1%).

Simplificity

$\hat{O} \rightarrow LED_{para}$ and $\hat{O} \rightarrow ConBART$ achieve highest simplicity (89.4%).

Planning helps
## Performance on OOD Data

<table>
<thead>
<tr>
<th>System</th>
<th>Flu</th>
<th>Ad</th>
<th>Simp</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LED_{para}</td>
<td>0.932</td>
<td>0.632</td>
<td>0.664</td>
<td>0.743</td>
</tr>
<tr>
<td>\hat{O} \rightarrow LED_{para}</td>
<td>0.890</td>
<td>0.684</td>
<td>0.760</td>
<td>0.778</td>
</tr>
<tr>
<td>\hat{O} \rightarrow ConBART</td>
<td>0.890</td>
<td>0.760</td>
<td>0.764</td>
<td>0.805</td>
</tr>
</tbody>
</table>

Human Evaluation, training on Newsela, testing on Wikipedia

- Plan-guidance helps
Performance on OOD Data

<table>
<thead>
<tr>
<th>System</th>
<th>Flu</th>
<th>Ad</th>
<th>Simp</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LED&lt;sub&gt;para&lt;/sub&gt;</td>
<td>0.932</td>
<td>0.632</td>
<td>0.664</td>
<td>0.743</td>
</tr>
<tr>
<td>$\hat{O} \rightarrow$ LED&lt;sub&gt;para&lt;/sub&gt;</td>
<td>0.890</td>
<td>0.684</td>
<td>0.760</td>
<td>0.778</td>
</tr>
<tr>
<td>$\hat{O} \rightarrow$ ConBART</td>
<td>0.890</td>
<td>0.760</td>
<td>0.764</td>
<td>0.805</td>
</tr>
</tbody>
</table>

Human Evaluation, training on Newsela, testing on Wikipedia

- Plan-guidance helps

- Paragraph-level models likely less adaptable to unseen domains because biased towards paragraph length of training data.
Example Output

**Input**

Glenn Edward Greenwald (born March 6, 1967) is an American journalist and author.

He is best known for a series of reports published from June 2013 by "The Guardian" newspaper detailing the United States and British global surveillance programs, and based on classified documents disclosed by Edward Snowden. Greenwald and the team he worked with won both a George Polk Award and a Pulitzer Prize for those reports.

He has written several best-selling books, including "No Place to Hide". Before the Snowden file disclosures, Greenwald was considered one of the most influential opinion columnists in the United States. After working as a constitutional attorney for ten years, he began blogging on national security issues before becoming a "Salon" contributor in 2007 and then for "The Guardian" in 2012. He now writes for (and has co-edited) "The Intercept", which he founded in 2013 with Laura Poitras and Jeremy Scahill.

Greenwald's work on the Snowden story was featured in the documentary "Citizenfour", which won the 2014 Academy Award for Best Documentary Feature. Greenwald appeared on-stage with director Laura Poitras and Snowden's girlfriend, Lindsay Mills, when the Oscar was given. In the 2016 Oliver Stone feature film "Snowden", Greenwald was played by actor Zachary Quinto.

**Output**

Glenn Greenwald is an American journalist and author.

He is best known for a series of reports published from June 2013 by the Guardian newspaper. They are based on documents leaked by Edward Snowden.

He has written several best-selling books, including "No Place to Hide." Before the Snowden file leaks, Greenwald was one of the most respected opinion columnists in the United States. He began writing about national security issues before becoming a "Salon" writer in 2007 and a writer for "The Guardian" in 2012. He now writes for "The Guardian."

Greenwald's work on the Snowden story was featured in the documentary "Citizenfour." The movie won an Academy Award. Greenwald worked with director Laura Poitras and Snowden's girlfriend, Lindsay Mills, to make the documentary. The 2016 Oliver Stone feature, "Snowden," was played by Zachary Quinto.
Key Takeaways and Perspectives

- Controlled and context-aware sentence simplification can effectively be used to simplify documents
Key Takeaways and Perspectives

- Controlled and context-aware sentence simplification can effectively be used to simplify documents
- Document simplification is under-studied
Key Takeaways and Perspectives

- Controlled and context-aware sentence simplification can effectively be used to simplify documents

- Document simplification is under-studied
  - Few training and test data for languages other than English
Key Takeaways and Perspectives

- Controlled and context-aware sentence simplification can effectively be used to simplify documents

- Document simplification is under-studied
  - Few training and test data for languages other than English
  - Few work on User adaptation

  Simplification models for different types of users (native vs. non-native, adult vs. children, expert vs. layman, people with reading disability)
Graph $\Rightarrow$ Text
Graphs are frequent data structures

- Knowledge Graphs
- RDF Stores
- Meaning Representations
- Tabular Data
Graph ⇒ Text

AMR ⇒ Text

- Generating into 21 EU Languages

RDF ⇒ Text

- WebNLG Challenge

- Generating into Low Resource Languages (Breton, Welsh, Irish)
AMR ⇒ Text

Multilingual Generation

Fan and Gardent EMNLP 2020
Challenges of Multilingual Graph-to-Text

- Structured input has a different surface form
Challenges of Multilingual Graph-to-Text

- Structured input has a different surface form
- Structured Input is often very underspecified
Challenges of Multilingual Graph-to-Text

- Structured input has a different surface form
- Structured Input is often very underspecified
- Decoding into languages with varied morphology and word order
Abstract Meaning Representation (AMR)

Ryan describes himself as a genius
AMR → 21 Languages

Amerikanská tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.

Americkí predstavitelia usporiadali stretnutie expertnej skupiny v januári 2002 v New Yorku.

US officials held an expert group meeting in January 2002 in New York.

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Funcionarios estadounidenses celebraron una reunión de un grupo de expertos en enero de 2002 en Nueva York.

Romance, Germanic, Slavic, Uralic
Encoding AMRs

Add branch and sibling embeddings (aka positional embedding for graphs)
AMR Encoding

- Transformer encoder
- Linearise (and simplify) AMRs
- Graph structure
  - Node: token + distance from root + branch
- Pretraining
  - on 30M silver AMRs
Multilingual Decoding

- Crosslingual embeddings (XLM Sentence Piece Model and Vocabulary)
- Language Models pretrained on 30M sentences (for each language)
- Multilingual decoding
  - Prefix each training instance with a control token
  - Trained on multilingual Europarl data
Multilingual Decoding

Decoding into Slovak

hold
:ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York

Amerikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.

Decoding into French

hold
:ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.
# Training Data

**hold**

:ARG0 person : ARG0-of have-org-role : ARG1 : op1 United : op2 States : ARG2 official

:ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group

:time date-entity : year 2002 : month 1

:location city : op1 New : op2 York

<table>
<thead>
<tr>
<th>French</th>
<th>Spanish</th>
<th>Slovak</th>
<th>Bulgarian</th>
<th>Swedish</th>
</tr>
</thead>
</table>

- Europarl: 21 Languages

- Input AMR: create AMR structure with JAMR parser
Multilingual AMR-to-NL Model

- Encoder: pretraining on Silver AMRs
- Decoder: language model pretraining
Test Data

- Silver AMR: 21 languages, Europarl
- Gold AMR: 4 languages
Comparison: Bilingual vs Multilingual

Bilingual Baseline

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1
United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.
Comparison: Bilingual vs Multilingual

Bilingual Baseline

hold
:ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.

Multilingual Model

hold
: ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.
Comparison: Bilingual vs Multilingual

**Bilingual Baseline**

hold
:ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.

**Multilingual Model**

hold
: ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York

Amersikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.
Results: Europarl
Results: Europarl

Bilingual Baseline: En AMR -> X
Multilingual Model: En AMR -> All

BLEU

High Resource | Mid Resource
Results: Gold AMR

Bilingual Baseline: En AMR -> X

Multilingual Model: En AMR -> All

<table>
<thead>
<tr>
<th>Language</th>
<th>Bilingual Baseline</th>
<th>Multilingual Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>es</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>it</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>de</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>
Comparison: Hybrid vs Multilingual

Hybrid Translation Model

AMR to English

US officials held an expert group meeting in January 2002 in New York.

Multilingual Model

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.
Comparison: Hybrid vs Multilingual

**Hybrid Translation Model**

```
hold
:ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York
```

AMR to English

```
US officials held an expert group meeting in January 2002 in New York.
```

Translation Model

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.

**Multilingual Model**

```
hold
: ARG0 person : ARG0-of have-org-role : ARG1 : op1
United : op2 States : ARG2 official
: ARG1 meet : ARG0 person : ARG1-of expert : ARG2-of group
: time date-entity : year 2002 : month 1
: location city : op1 New : op2 York
```

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.
Comparison: Hybrid vs Multilingual

Hybrid Translation Model

Hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1
United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity year 2002 :month 1
:location city :op1 New :op2 York

AMR to English

US officials held an expert group meeting in January 2002 in New York.

Translation Model

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.

Multilingual Model

Hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1
United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d’un groupe d’experts en janvier 2002 à New York.
Comparison: Hybrid vs Multilingual

Hybrid Translation: En AMR -> En -> Translate to X

Multilingual Model: En AMR -> All

BLEU
Human Evaluation

- Evaluators: colleagues from NLP mailing lists
- 50 sentences per language
  - Half low BLEU
  - Half high BLEU
Human Evaluation

• Semantic Accuracy:
  Does the generated text correctly paraphrase the reference?

• Morphology:
  Is the morphology correct? Are agreement constraints e.g., verb/subject, noun/adjective respected?

• Word Order:
  Is the word order natural sounding?
Human Evaluation: Semantic Accuracy

![Bar chart showing semantic accuracy for different languages. The accuracy values range from 1.6 to 2.4.]
Key Takeaways

- Pre-training and Multilingual techniques permits bridging the gap between English-Centric AMRs and target languages with varied syntax and morphology
Key Takeaways

- Pre-training and Multilingual techniques permits bridging the gap between English-Centric AMRs and target languages with varied syntax and morphology
  - Pretrained LMs and AMRs, Crosslingual embeddings, Multilingual training
Key Takeaways

- Pre-training and Multilingual techniques permits bridging the gap between English-Centric AMRs and target languages with varied syntax and morphology
  - Pretrained LMs and AMRs, Crosslingual embeddings, Multilingual training
- Multilingual models benefits from increased training data and perform better on average than bilingual
Key Takeaways

- Pre-training and Multilingual techniques permits bridging the gap between English-Centric AMRs and target languages with varied syntax and morphology
  - Pretrained LMs and AMRs, Crosslingual embeddings, Multilingual training

- Multilingual models benefits from increased training data and perform better on average than bilingual

- Multilingual End-to-End models outperform NLG+MT models
RDF → Text

The WebNLG Challenge

Gardent et al. ACL 2017, Castro-Ferreira et al. 2020, Cripwell et al. 2023
The WebNLG Challenge

RDF Graph
Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred Worden and was commanded by David Scott.
WebNLG 2017: RDF ⇒ English

<table>
<thead>
<tr>
<th></th>
<th>Train+Dev</th>
<th>Test (Seen Category)</th>
<th>Test (Unseen Category)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td># (Graph,Text)</td>
<td>20,370</td>
<td>2,495</td>
<td>2,413</td>
<td>25,298</td>
</tr>
<tr>
<td># Graphs</td>
<td>7,812</td>
<td>971</td>
<td>891</td>
<td>9,674</td>
</tr>
</tbody>
</table>

- DBPedia graphs with root entity of various categories.
- English texts are crowdsourced
WebNLG 2017: RDF ⇒ English

<table>
<thead>
<tr>
<th></th>
<th>Train+Dev</th>
<th>Test (Seen Category)</th>
<th>Test (Unseen Category)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td># (Graph,Text)</td>
<td>20,370</td>
<td>2,495</td>
<td>2,413</td>
<td>25,298</td>
</tr>
<tr>
<td># Graphs</td>
<td>7,812</td>
<td>971</td>
<td>891</td>
<td>9,674</td>
</tr>
</tbody>
</table>

10 seen categories:

- Astronaut, University, Monument, Building, Comics Character, Food, Airport, SportsTeam, City and WrittenWork

5 unseen categories:

- Athlete, Artist, MeanOfTransportation, CelestialBody, Politician
WebNLG 2017: RDF ⇒ English

- 6 participants, 10 systems
- Models: 3 rule-based, 1 SMT, 5 neural
WebNLG 2017: RDF $\Rightarrow$ English

ALL: 7.07 - 45.13, Seen: 19.87 - 60.54, Unseen: 5.13 - 35.7
WebNLG 2020

Natural Language Generation

- RDF $\Rightarrow$ English
WebNLG 2020

Natural Language Generation

- RDF ⇒ English
WebNLG 2020

Natural Language Generation

- RDF ⇒ English
- RDF ⇒ Russian
WebNLG 2020

Natural Language Generation

- RDF ⇒ English
- RDF ⇒ Russian

Semantic Parsing

- English ⇒ RDF
- Russian ⇒ RDF
WebNLG 2020: RDF ⇒ English

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test NLG/SP</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td># (Graph,Text)</td>
<td>35,426</td>
<td>4,664</td>
<td>5,150</td>
<td>47,395</td>
</tr>
<tr>
<td># Graphs</td>
<td>13,211</td>
<td>1,667</td>
<td>1,779</td>
<td>17,409</td>
</tr>
</tbody>
</table>

16 seen categories

Astronaut, University, Monument, Building, Comics Character, Food, Airport, SportsTeam, City, WrittenWork, Athlete, Artist, CelestialBody, MeanOfTransportation, Politician, Company

3 unseen categories:

Film, Scientist, and MusicalWork

Unseen entities: graphs from seen categories, but unseen root entity

E.g., Nie Haisheng in category Astronaut
# WebNLG 2020: Participation

<table>
<thead>
<tr>
<th>System</th>
<th>Affiliation</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>MED</td>
<td>Sber AI Lab</td>
<td>Russia</td>
</tr>
<tr>
<td>RALI-UMontréal</td>
<td>Université de Montréal</td>
<td>Canada</td>
</tr>
<tr>
<td>ORANGE-NLG</td>
<td>Orange Labs</td>
<td>France</td>
</tr>
<tr>
<td>CUNI-UFAL</td>
<td>Charles University</td>
<td>Czechia</td>
</tr>
<tr>
<td>TGen</td>
<td>AIST</td>
<td>Japan</td>
</tr>
<tr>
<td>BT5</td>
<td>Google</td>
<td>US</td>
</tr>
<tr>
<td>UPC-POE</td>
<td>Universitat Politècnica de Catalunya</td>
<td>Spain</td>
</tr>
<tr>
<td>DANGNT-SGU</td>
<td>Saigon University</td>
<td>Vietnam</td>
</tr>
<tr>
<td>Huawei</td>
<td>Huawei Noah’s Ark Lab</td>
<td>UK</td>
</tr>
<tr>
<td>AmazonAI</td>
<td>Amazon AI (Shanghai)</td>
<td>China</td>
</tr>
<tr>
<td>NILC</td>
<td>University of São Paulo</td>
<td>Brazil</td>
</tr>
<tr>
<td>NUIG-DSI</td>
<td>National University of Ireland</td>
<td>Ireland</td>
</tr>
<tr>
<td>CycleGT</td>
<td>Amazon</td>
<td>China</td>
</tr>
<tr>
<td>OSU NeurAL NLG</td>
<td>The Ohio State University</td>
<td>US</td>
</tr>
<tr>
<td>FBConvAI</td>
<td>Facebook</td>
<td>US</td>
</tr>
</tbody>
</table>

17 teams submitted 48 system runs
WebNLG 2020: Results
WebNLG 2023: Low Resource Languages

### Data

<table>
<thead>
<tr>
<th></th>
<th>Silver Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breton</td>
<td>13,211</td>
<td>1,399</td>
<td>1,778</td>
</tr>
<tr>
<td>Welsh</td>
<td>13,211</td>
<td>1,665</td>
<td>1,778</td>
</tr>
<tr>
<td>Irish</td>
<td>13,211</td>
<td>1,665</td>
<td>1,778</td>
</tr>
<tr>
<td>Maltese</td>
<td>13,211</td>
<td>1,665</td>
<td>1,778</td>
</tr>
</tbody>
</table>
# WebNLG 2023: Low Resource Languages

## Data

<table>
<thead>
<tr>
<th></th>
<th>Silver Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breton</td>
<td>13,211</td>
<td>1,399</td>
<td>1,778</td>
</tr>
<tr>
<td>Welsh</td>
<td>13,211</td>
<td>1,665</td>
<td>1,778</td>
</tr>
<tr>
<td>Irish</td>
<td>13,211</td>
<td>1,665</td>
<td>1,778</td>
</tr>
<tr>
<td>Maltese</td>
<td>13,211</td>
<td>1,665</td>
<td>1,778</td>
</tr>
</tbody>
</table>

## Participants

<table>
<thead>
<tr>
<th>Team</th>
<th>Affiliation</th>
<th>Country</th>
<th>Breton</th>
<th>Welsh</th>
<th>Irish</th>
<th>Maltese</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUNI-Wue</td>
<td>Charles University</td>
<td>Czechia</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DCU/TCD-FORGe</td>
<td>ADAPT/DCU/Trinity College</td>
<td>Ireland</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Interno</td>
<td>Pulkovo Observatory</td>
<td>Russia</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>IREL</td>
<td>IIT Hyderabad</td>
<td>India</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DCU-NLG-PBN</td>
<td>ADAPT/DCU</td>
<td>Ireland</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>
WebNLG 2023: Pipeline NLG+MT Models

Participants

<table>
<thead>
<tr>
<th>Team</th>
<th>Affiliation</th>
<th>Country</th>
<th>Breton</th>
<th>Welsh</th>
<th>Irish</th>
<th>Maltese</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUNI-Wue</td>
<td>Charles University</td>
<td>Czechia</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DCU/TCD-FORGe</td>
<td>ADAPT/DCU/Trinity College</td>
<td>Ireland</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Interno</td>
<td>Pulkovo Observatory</td>
<td>Russia</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>IREL</td>
<td>IIT Hyderabad</td>
<td>India</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DCU-NLG-PBN</td>
<td>ADAPT/DCU</td>
<td>Ireland</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>

RDF ⇒ English

- T5 or mT5 fine-tuned on English WebNLG data
- GPT3-5 in context learning, no fine-tuning

English ⇒ LR Language

- Machine Translation: NLLB or Google Translate
WebNLG 2023: Results

Min and Max BLEU Scores per Language

<table>
<thead>
<tr>
<th>Language</th>
<th>Min BLEU</th>
<th>Max BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>English 2020</td>
<td>0.14</td>
<td>0.54</td>
</tr>
<tr>
<td>Irish</td>
<td>0.12</td>
<td>0.21</td>
</tr>
<tr>
<td>Maltese</td>
<td>0.11</td>
<td>0.2</td>
</tr>
<tr>
<td>Welsh</td>
<td>0.09</td>
<td>0.1</td>
</tr>
<tr>
<td>Breton</td>
<td>0.39</td>
<td>0.25</td>
</tr>
</tbody>
</table>
End-to-End RDF → Celtic Language

Soto-Martinez et al. AACL-IJCNLP 2023
Pipeline vs. End-to-End

For Breton, there is no (good) MT system
Pipeline vs. End-to-End

For Breton, there is no (good) MT system

× NLG+MT pipeline
Pipeline vs. End-to-End

For Breton, there is no (good) MT system

- NLG+MT pipeline
- Full-fine tuning (BLEU: 0.10)
Pipeline vs. End-to-End

For Breton, there is no (good) MT system

❌ NLG+MT pipeline

❌ Full-fine tuning (BLEU: 0.10)

✅ Parameter Efficient Fine Tuning (PEFT)
Pipeline vs. End-to-End

For Breton, there is no (good) MT system

- NLG+MT pipeline
- Full-fine tuning (BLEU: 0.10)

- Parameter Efficient Fine Tuning (PEFT)
  - Soft-Prompt
  - Structured to capture language relatedness and various tasks
Phylogenetic Tree

Celtic

Goidelic
- Irish
- Scottish Gaelic

Brythonic
- Breton
- Welsh

Germanic

West Germanic
- English
- English RDF

Soft Prompt

<table>
<thead>
<tr>
<th>50 Tokens Task</th>
<th>15 Tokens Source Family</th>
<th>15 Tokens Source Genus</th>
<th>15 Tokens Source Language</th>
<th>15 Tokens Target Family</th>
<th>15 Tokens Target Genus</th>
<th>15 Tokens Target Language</th>
<th>n Tokens Input Sequence</th>
</tr>
</thead>
</table>
Training and Testing

**Step 1: Self-supervised Training (Language Models)**

*Trains the Soft Prompt on unsupervised, monolingual tasks*

<table>
<thead>
<tr>
<th>Task</th>
<th>Source</th>
<th>Target</th>
<th>Original Input Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masked LM</td>
<td>Germanic</td>
<td>West Germanic</td>
<td>RDF</td>
</tr>
<tr>
<td>Prefix LM</td>
<td>Germanic</td>
<td>West Germanic</td>
<td>English</td>
</tr>
<tr>
<td>Suffix LM</td>
<td>Celtic</td>
<td>Britonic</td>
<td>Welsh</td>
</tr>
<tr>
<td>Deshuffling</td>
<td>Celtic</td>
<td>Britonic</td>
<td>Breton</td>
</tr>
<tr>
<td>Generate</td>
<td>Celtic</td>
<td>Goidelic</td>
<td>Irish</td>
</tr>
</tbody>
</table>
Training and Testing

**Step 1: Self-supervised Training (Language Models)**
*Trains the Soft Prompt on unsupervised, monolingual tasks*

<table>
<thead>
<tr>
<th>Task</th>
<th>Source</th>
<th>Target</th>
<th>Original Input Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masked LM</td>
<td>Germanic</td>
<td>West Germanic</td>
<td>RDF</td>
</tr>
<tr>
<td>Prefix LM</td>
<td>Germanic</td>
<td>West Germanic</td>
<td>English</td>
</tr>
<tr>
<td>Suffix LM</td>
<td>Celtic</td>
<td>Britonic</td>
<td>Welsh</td>
</tr>
<tr>
<td>Deshuffling</td>
<td>Celtic</td>
<td>Britonic</td>
<td>Breton</td>
</tr>
<tr>
<td>Generate</td>
<td>Celtic</td>
<td>Goidelic</td>
<td>Irish</td>
</tr>
</tbody>
</table>

**Step 2: Fine-Tuning on Dev RDF-to-Text data (RDF-to-Text Models)**
*Trains the RDF-to-Text Task sub-prompt for each target language*
Training and Testing

**Step 1: Self-supervised Training (Language Models)**
*Trains the Soft Prompt on unsupervised, monolingual tasks*

<table>
<thead>
<tr>
<th>Task</th>
<th>Source</th>
<th>Target</th>
<th>Original Input Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masked LM</td>
<td>Germanic</td>
<td>West Germanic</td>
<td>RDF</td>
</tr>
<tr>
<td>Prefix LM</td>
<td>Germanic</td>
<td>West Germanic</td>
<td>English</td>
</tr>
<tr>
<td>Suffix LM</td>
<td>Celtic</td>
<td>Britonic</td>
<td>Welsh</td>
</tr>
<tr>
<td>Deshuffling</td>
<td>Celtic</td>
<td>Britonic</td>
<td>Breton</td>
</tr>
<tr>
<td>Generate</td>
<td>Celtic</td>
<td>Goidelic</td>
<td>Irish</td>
</tr>
</tbody>
</table>

**Step 2: Fine-Tuning on Dev RDF-to-Text data (RDF-to-Text Models)**
*Trains the RDF-to-Text Task sub-prompt for each target language*

**Inference**
*The Language sub-prompt is set to the target language.*
Results

![Graph showing BLEU scores for Breton, Welsh, and Irish with different prefixes and fine-tuning methods]
Impact of Data Size

![Graph showing the impact of data size on BLEU score across different languages (BR, CY, EN, GA). The graph plots BLEU score on the y-axis and training samples per language on the x-axis. Different languages show varying trends with respect to data size.]
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- Performance is lower for Russian than for English
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- Performance is lower for Russian than for English
- Performance degrades on out of domain data (unseen)
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- Performance is lower for Russian than for English
- Performance degrades on out of domain data (unseen)
- Performance is very poor for Low Resource Languages (2023)
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- Performance is lower for Russian than for English
- Performance degrades on out of domain data (unseen)
- Performance is very poor for Low Resource Languages (2023)
- Most LRL Models are NLG+MT pipelines
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- Performance is lower for Russian than for English
- Performance degrades on out of domain data (unseen)
- Performance is very poor for Low Resource Languages (2023)
- Most LRL Models are NLG+MT pipelines
- LRL for which no MT models exist (Breton) show very poor results for these models
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- Performance is lower for Russian than for English
- Performance degrades on out of domain data (unseen)
- Performance is very poor for Low Resource Languages (2023)
- Most LRL Models are NLG+MT pipelines
- LRL for which no MT models exist (Breton) show very poor results for these models
- PEFT techniques help improve performance for these languages

  BLEU for Breton: 10 (NLG+MT) → 18.15 (PEFT E2E Model)
- LLMs for simplification, graph-to-text
- LLMs for simplification, graph-to-text
  - Prompt engineering, chain-of-thought decomposition
• LLMs for simplification, graph-to-text
  ○ Prompt engineering, chain-of-thought decomposition
  ○ LLMs for evaluation
• LLMs for simplification, graph-to-text
  ◦ Prompt engineering, chain-of-thought decomposition
  ◦ LLMs for evaluation
• Generalising RDF-to-Text Models
  ◦ to other languages and other domains
• LLMs for simplification, graph-to-text
  ◦ Prompt engineering, chain-of-thought decomposition
  ◦ LLMs for evaluation
• Generalising RDF-to-Text Models
  ◦ to other languages and other domains
• Evaluation
• LLMs for simplification, graph-to-text
  ▪ Prompt engineering, chain-of-thought decomposition
  ▪ LLMs for evaluation
• Generalising RDF-to-Text Models
  ▪ to other languages and other domains
• Evaluation
  ▪ Metrics for simplificity and Meaning Preservation

Le Scao and Gardent AACL-IJCNLP Findings 2023, Cripwell et al. EMNLP 2023
Questions ?