Verbalising Graphs into High-, Medium- and Low-Resource Languages

*Chaire IA xNLG: Generating from Multiple Sources into Multiple Languages*

Claire Gardent

CNRS / LORIA

Anja Belz, Thiago Castro-Ferreira, Liam Cripwell, Angela Fan, Albert Gatt, Nikolai Ilinyskh, Chris van der Lee, Simon Mille, Diego Moussalem, Yannick Parmentier, Laura Perez-Beltrachini, Anastasia Shimorina, William Soto-Martinez
Graphs are frequent Data Structures

- Knowledge Graphs
Graphs are frequent Data Structures

- Knowledge Graphs
- RDF Stores
Graphs are frequent Data Structures

- Knowledge Graphs
- RDF Stores
- Meaning Representations
Graphs are frequent Data Structures

- Knowledge Graphs
- RDF Stores
- Meaning Representations
- Tabular Data
Abstract Meaning Representation (AMR)

Ryan describes himself as a genius
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.
Challenges

- Structured input has a different surface form
Challenges

- Structured input has a different surface form
- Structured Input is underspecified
Challenges

- Structured input has a different surface form
- Structured Input is underspecified
- Lack of parallel Graph/text data
Challenges

- Structured input has a different surface form
- Structured Input is underspecified
- Lack of parallel Graph/text data
- Decoding into languages with varied morphology and word order
Outline

Multilingual Models

- AMR ⇒ 21 EU Languages
Outline

Multilingual Models

- AMR $\Rightarrow$ 21 EU Languages

Pre-trained Multilingual Models

- RDF $\Rightarrow$ English, Russian
Outline

Multilingual Models

- AMR $\Rightarrow$ 21 EU Languages

Pre-trained Multilingual Models

- RDF $\Rightarrow$ English, Russian

Parameter Efficient Fine Tuning

- RDF $\Rightarrow$ Breton, Welsh, Irish, Maltese
Outline

Multilingual Models

- AMR $\Rightarrow$ 21 EU Languages

Pre-trained Multilingual Models

- RDF $\Rightarrow$ English, Russian

Parameter Efficient Fine Tuning

- RDF $\Rightarrow$ Breton, Welsh, Irish, Maltese
- AMR $\Rightarrow$ 6 High- and 6 Low-Resource Languages
AMR \Rightarrow 21 EU Languages

Fan and Gardent EMNLP 2020
Ryan describes himself as a genius.
AMR → 21 Languages

Amerikanska 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
The Encoder-Decoder Framework

Two networks

- The encoder produces a continuous representation of the input
- The decoder generates a text from this representation

A Unifying Framework for Text Generation
AMR Encoding

- Transformer encoder
- Linearise (and simplify) AMRs
- Graph structure
  - Node: token + distance from root + branch
- Pretraining (Masked Language Modelling objective)
  - on 30M silver AMRs
Linearising

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
Encoding Structure

Add branch and sibling embeddings (aka positional embedding for graphs)
Multilingual Decoding

- Crosslingual embeddings (XLM Sentence Piece Model and Vocabulary)
- Language Models pretrained on 30M sentences (for each language)
Leveraging Pretraining

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

Multilingual decoding

- Prefix each training instance with a control token
- Trained on multilingual Europarl data
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

- Europarl: 21 Languages
- Input AMR: create AMR structure with JAMR parser
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

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

Multilingual Model

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

[Diagram showing BLEU scores for various languages comparing bilingual baseline to multilingual model]
Results: Europarl

The multilingual model generally outperforms monolingual models
Results: Europarl

The multilingual model generally outperforms monolingual models

The difference is stronger on Mid-Resource Languages
Results: Gold AMR

The difference also holds when generating from gold AMRs
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.

**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
```

fr

```
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

AMR to English

Translation Model

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

Multilingual Model

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

Generate: AMR → English
Comparison: Hybrid vs Multilingual

**Hybrid Translation Model**

Translation Model

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

**Multilingual Model**

Translation Model

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

Generate: AMR → **English**

Translate: **English** → X
Comparison: Hybrid vs Multilingual

Hybrid Translation: En AMR -> En -> Translate to X  Multilingual Model: En AMR -> All

<table>
<thead>
<tr>
<th>Language</th>
<th>BLEU</th>
<th>Hybrid Translation</th>
<th>Multilingual Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>17.5</td>
<td>18.8</td>
<td>18.8</td>
</tr>
<tr>
<td>es</td>
<td>28.1</td>
<td>28.1</td>
<td>28.1</td>
</tr>
<tr>
<td>fr</td>
<td>19.8</td>
<td>20.8</td>
<td>20.8</td>
</tr>
<tr>
<td>it</td>
<td>17.2</td>
<td>17.2</td>
<td>17.2</td>
</tr>
<tr>
<td>nl</td>
<td>18.7</td>
<td>18.7</td>
<td>18.7</td>
</tr>
<tr>
<td>pt</td>
<td>21.1</td>
<td>21.1</td>
<td>21.1</td>
</tr>
<tr>
<td>sv</td>
<td>16.6</td>
<td>16.6</td>
<td>16.6</td>
</tr>
</tbody>
</table>
Comparison: Hybrid vs Multilingual

The multilingual model outperforms the Gen&Translate pipeline
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

![Semantic Accuracy Bar Chart]
Key Takeaways

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

- Pre-training and Multilingual techniques permits bridging the gap between English-Centric AMR graphs 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 AMR graphs 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 AMR graphs 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
Knowledge Graphs ⇒ English, Russian
Gardent et al. ACL 2017, Castro-Ferreira et al. INLG 2020
The WebNLG Challenge
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 ⇒ English

**ALL**: 7.07 - 45.13, **Seen**: 19.87 - 60.54, **Unseen**: 5.13 - 35.7
WebNLG 2017: RDF ⇒ English

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

*Strong differences between models*
WebNLG 2017: RDF ⇒ English

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

Strong differences between models

All models degrades on Unseen Data
WebNLG 2020

Natural Language Generation

- RDF $\Rightarrow$ English
WebNLG 2020

Natural Language Generation

- RDF $\Rightarrow$ 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-UFAI</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>FBCConvAI</td>
<td>Facebook</td>
<td>US</td>
</tr>
</tbody>
</table>

17 teams submitted 48 system runs
WebNLG 2020: Results

Min and Max BLEU Scores per Category of Test Data

- Russian
- Unseen Entity
- English (All)
- Unseen Category
- Seen Category

Min BLEU:
- Russian: 0.25
- Unseen Entity: 0.39
- English (All): 0.51
- Unseen Category: 0.35
- Seen Category: 0.23

Max BLEU:
- Russian: 0.51
- Unseen Entity: 0.54
- English (All): 0.61
- Unseen Category: 0.52
- Seen Category: 0.49
WebNLG 2020: Results

Results are better for English than for Russian
WebNLG 2020: Results

Results are better for English than for Russian

Pre-training improves results: +16 BLEU points for English w.r.t. 2017
WebNLG 2023: Low Resource Languages

<table>
<thead>
<tr>
<th>Language</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>Language</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: 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>

No training Data
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

- English 2020
- Irish
- Maltese
- Welsh
- Breton

Min BLEU

- English 2020: 0.39
- Irish: 0.14
- Maltese: 0.09
- Welsh: 0.12
- Breton: 0.11

Max BLEU

- English 2020: 0.54
- Irish: 0.21
- Maltese: 0.1
- Welsh: 0.2
- Breton: 0.25
WebNLG 2023: Results

Strong degradation overall compared to results on English
WebNLG 2023: Results

Strong degradation overall compared to results on English

Very poor output for Breton
End-to-End RDF $\Rightarrow$ 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 (Prefix Tuning)
  - Structured to capture language relatedness and various tasks
Prefix Tuning

- All parameters of the pre-trained model are frozen
- Only learn the prefix (soft-prompt) parameters

Li and Liang 2021
The soft-prompt is decomposed into Family, Genus, and Language sub-prompts.
Phylogenetic Tree

The soft-prompt is decomposed into Family, Genus, and Language sub-prompts.

**The soft-prompt is decomposed into Family, Genus, and Language sub-prompts.**

**Allows LR languages to benefit from the training data of their related languages** (E.g., The sub-prompt for Goidelic is updated each time an Irish or Gaelic training instance is processed) - Facilitate Transfer

**Prevents the mixture of training data to introduce too much noise to the model.** - Reduce Noise
# 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 West Germanic RDF</td>
<td>Germanic West Germanic RDF</td>
<td>&lt;S&gt; Einstein &lt;P&gt; &lt;mask&gt; &lt;P&gt; Poland</td>
</tr>
<tr>
<td>Prefix LM</td>
<td>Germanic West Germanic English</td>
<td>Germanic West Germanic English</td>
<td>Thank you for &lt;mask&gt; &lt;pad&gt; &lt;pad&gt;</td>
</tr>
<tr>
<td>Suffix LM</td>
<td>Celtic Britonic Welsh Celtic Britonic Welsh</td>
<td>&lt;mask&gt; honno ? &lt;pad&gt; &lt;pad&gt; &lt;pad&gt;</td>
<td></td>
</tr>
<tr>
<td>Deshuffling</td>
<td>Celtic Britonic Breton Celtic Britonic Breton</td>
<td>skuizh ? out Ha &lt;pad&gt; &lt;pad&gt;</td>
<td></td>
</tr>
<tr>
<td>Generate</td>
<td>Celtic Goidelic Irish Celtic Goidelic Irish</td>
<td>Seo &lt;mask&gt; &lt;pad&gt; &lt;pad&gt; &lt;pad&gt; &lt;pad&gt;</td>
<td></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
Results

Phylogenetic prefix-tuning outperforms full fine-tuning and a SoTA approach for KG-to-Text generation.
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- Performance degrades on out of domain data (unseen)
Key Takeaways

- Pretraining (2017 vs 2020) improves performance
- 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 degrades on out of domain data (unseen)
- Performance is very poor for Low Resource Languages (2023)
- PEFT techniques help improve performance for these languages

BLEU for Breton: 10 (NLG+MT) → 18.15 (PEFT E2E Model)
AMR Graph --> High- and Low-Resource Languages

Soto-Martinez et al. 2024, In Submission
Hierarchical Fine-Tuning

- Iterative fine-tuning of a multilingual model (12 languages) into 12 monolingual models

Phylogenetic Knowledge

- At each iteration, the training languages are chosen using phylogenetic knowledge
LoRA (Low Rank Matrices) Adaptation

During training:

\[ h = W x + B A x \]

\[ h = (W + BA) x \]

\[ W_{merged} \in \mathbb{R}^{d \times d} \]

Pretrained Weights:

\[ W \in \mathbb{R}^{d \times d} \]

\[ A \sim \mathcal{N}(0, \sigma^2) \]

\[ B = 0 \]

\[ r \]

After training:

\[ h \]

\[ x \]

Merged Weights:

\[ W_{merged} \in \mathbb{R}^{d \times d} \]

A new model is created. No overhead during inference.
HQL outperforms or is on par with multi- and monolingual approaches.
Comparison with Baselines

HQL outperforms the Gen&Trans approach for LR Languages
Comparison with Baselines

Figure 4: Average BLEU score of all languages against total number of seen training instances. HQL models include results on the intermediary levels of the hierarchy.
Comparison with Baselines

![Graph showing comparison between different models](image)

Figure 4: Average BLEU score of all languages against total number of seen training instances. HQL models include results on the intermediary levels of the hierarchy.

*HQL optimises faster than the 3 baselines.*
Comparison with Baselines

On average, HQL outperforms all 3 baselines.

Figure 4: Average BLEU score of all languages against total number of seen training instances. HQL models include results on the intermediary levels of the hierarchy.
Comparison with Previous Work

<table>
<thead>
<tr>
<th>Model</th>
<th>DEU</th>
<th>ENG</th>
<th>SPA</th>
<th>ITA</th>
</tr>
</thead>
<tbody>
<tr>
<td>F&amp;G</td>
<td>15.3</td>
<td>24.9</td>
<td>21.7</td>
<td>19.8</td>
</tr>
<tr>
<td>Ribeiro</td>
<td>20.6</td>
<td>—</td>
<td>30.7</td>
<td>26.4</td>
</tr>
<tr>
<td>Xu</td>
<td>25.7</td>
<td>—</td>
<td>31.4</td>
<td>28.4</td>
</tr>
<tr>
<td>Martinez</td>
<td>23.2</td>
<td>44.8</td>
<td>34.6</td>
<td>29.0</td>
</tr>
<tr>
<td>MonoQL</td>
<td>18.2</td>
<td>49.2</td>
<td>38.6</td>
<td>22.7</td>
</tr>
<tr>
<td>MultiQL</td>
<td>19.8</td>
<td>42.9</td>
<td>34.1</td>
<td>27.2</td>
</tr>
<tr>
<td>Gen&amp;Trans*</td>
<td>28.0</td>
<td>49.2</td>
<td>39.6</td>
<td>33.8</td>
</tr>
<tr>
<td>DLHQL</td>
<td>21.2</td>
<td>44.2</td>
<td>37.4</td>
<td>29.2</td>
</tr>
<tr>
<td>PTHQL</td>
<td>22.8</td>
<td>43.4</td>
<td>37.2</td>
<td>29.7</td>
</tr>
</tbody>
</table>

Table 4: BLEU score on AMR3.0 test data.

English Gen&Trans is simply the result of MonoQL.

*HQL performs on par with previous work on HRL while using fewer data.*
Cross-Modal, Multilingual Graph/Text similarity metrics are needed to
Cross-Modal, Multilingual Graph/Text similarity metrics are needed to

- filter noisy training data
Cross-Modal, Multilingual Graph/Text similarity metrics are needed to

- filter noisy training data
- guide generation
Cross-Modal, Multilingual Graph/Text similarity metrics are needed to

- filter noisy training data
- guide generation
- generalise Graph-to-Text Models to other languages and other domains
Questions ?