Towards Automatic Glossing

The 2023 SIGMORPHON Challenge

glose 2023
Paris
Le 28 Juin 2023

Miikka Silfverberg
msilfver@mail.ubc.ca
1. Introduction
2. Machine learning for glossing
3. The SIGMORPHON shared task on interlinear glossing
5. Submissions and results
6. Takeaways
The SIGMORPHON shared task on interlinear glossing

First shared task on automated interlinear glossing

One of many shared tasks on computational morphology that SIGMORPHON has organized since 2016

The 2023 SIGMORPHON Interlinear Glossing Challenge
Interlinear glossing

An interlinear gloss (Arapaho):

| transcription: wohei          | heetne'nee'eestoo3i'          |
| segmentation: wohei           | heet-ne'-nee'eestoo-3i'       |
| gloss: okay                  | FUT-then-do.thus-3PL          |

translation: Well that's what they're going to do.

A semistructured tabular format

Lots of variation in annotation practices: shallow vs. canonical segmentation (e.g. normalizing Eng. PAST markers -d and -t to -ed), tags vary, ...

The Leipzig Glossing Rules (Lehmann, 1982)
Glossing as a supervised learning task

We treat glossing as a supervised learning task.

Competitors receive glossed sentences as training data and learn models which are able to annotate unseen sentences.

We ask the competitors to fill in the missing glosses in a test set which lack annotations.

transcription: Ii nax'ni-diit win dim bakwhl siwetdiit ehl surveyors

segmentation: ii nax'ni-diit win dim bakw-hl si-we-t-diit e-hl surveyors


translation: They heard that what they call surveyors were coming.
Glossing as a supervised learning task

Interlinear glossing is connected to morphological segmentation. If you know the morphemes, it can be straightforward to gloss (chien-s -> dog-PL)

Conversely, if you get the segmentation wrong, there's great risk that you will gloss incorrectly

Major challenges in the glossing task:

- Ambiguity. (Fr: -s might refer to pl. number chien-s or 1st person comprend-s)
- Context can influence the interpretation of lexical and grammatical morphemes
- Unknown lexical and grammatical morphemes
- How to extract information from translations?
Why glossing?

Language preservation and revitalization have become significant areas of focus in policy making and linguistic research.

Both rely on language documentation (often accomplished through glossing).

Language documentation is invariably be a slow process.

Time is of the essence because knowledge about languages is dying as we speak!

The hope is that technological tools can speed up the work.

For many languages, glossed text is the only type of annotated data that is available. It is important that we learn to handle it.
State of the art in glossing

Two main approaches: (1) feature-based and (2) neural models

Feature-based models (CRF, MEMM, etc.) depend on human-engineered features (McMillan-Major, 2020)

These models typically depend on an external morphological segmentation (either human or model generated)

Can train well even on small annotated training sets
State of the art in glossing

Neural sequence-to-sequence models (transformer, LSTM encoder-decoder) independently learn representations (Zhao et al., 2020)

Neural models are extensively used in many tasks like machine translation

Typically, neural models benefit from large annotated training sets
State of the art in glossing

The system by Zhao et al. (2020) elegantly incorporates translations

Neural models can be trained to translate input directly into a gloss. No intermediate segmentation is required (although it can be helpful)
Crosslingual training could boost performance for low-resource languages

Incorporating additional noisy training data from multilingual databases like ODIN (Lewis and Xia, 2010)

Data augmentation techniques (Anastasopoulos and Neubig, 2019) could enhance the training process for glossing models.

Hard attention models (Aharoni and Goldberg, 2017) have delivered strong performance for many morphology tasks.

Pretrained language models like ByT5 (Xue et al. 2022) have demonstrated strong performance in various morphology tasks.
The SIGMORPHON shared task on interlinear glossing

Through spring 2023, teams built glossing systems for 7 languages based on annotated training and development data.

The competition culminated in an evaluation period at the end of May.

We investigated glossing in two different scenarios: the open and closed track.

We got submissions from five teams.

The systems were surprisingly different! Many interesting techniques were included.
A diverse set of languages
A diverse set of languages

For the shared task, we wanted manually annotated high-quality data

Arapaho – polysynthetic (North America)

Gitksan – "analytic to synthetic" morphology (North America)

Lezgi – agglutinating (Caucasus)

Natügu – agglutinating (Austronesia)

Tsez – agglutinating (Caucasus)

Nyangbo – agglutinating (Western Africa)

Uspanteko – "lightly agglutinating" (Central America)
Statistics

We wanted to investigate performance under different training data conditions.

Number of training examples

Arapaho – 140000 training tokens, Gitksan – 260 training tokens

The 2023 SIGMORPHON Interlinear Glossing Challenge
Statistics

All of our languages display multi-morphemic words

Morphemes per token

<table>
<thead>
<tr>
<th>Language</th>
<th>Morphemes per token</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arapaho</td>
<td>1.7</td>
</tr>
<tr>
<td>Gitksan</td>
<td>1.6</td>
</tr>
<tr>
<td>Lezgi</td>
<td>1.5</td>
</tr>
<tr>
<td>Natügu</td>
<td>1.5</td>
</tr>
<tr>
<td>Nyangbo</td>
<td>1.5</td>
</tr>
<tr>
<td>Tsez</td>
<td>2.0</td>
</tr>
<tr>
<td>Uspanteko</td>
<td>1.4</td>
</tr>
</tbody>
</table>
Statistics

Distribution of morpheme counts for Natügu tokens

The 2023 SIGMORPHON Interlinear Glossing Challenge
The 2023 SIGMORPHON Interlinear Glossing Challenge

Statistics

Distribution of morpheme counts for Tsez tokens
Tracks

Track 1 – the closed track

\t Esnazał xizaz ixiw raład boqno.
\t \g sister-PL-CONT.ESS behind big sea III-become-PST.UNW
\t \l And a big sea formed behind the sisters.

Track 2 – the open track

\t Esnazał xizaz ixiw raład boqno.
\m esyu-bi-ł xizaz ixiw raład b-oq-n
\g sister-PL-CONT.ESS behind big sea III-become-PST.UNW
\l And a big sea formed behind the sisters.

In the open track, all additional resources, apart from glossed data in the target language, are allowed.

Additional data:
- Glossed third-language data
- Plain text
- Dictionaries
...
We had 10 submissions from a total of 5 teams

All of the teams utilized neural models in some way

Primary glossing model:
- Transformer (2 teams)
- LSTM (1 team)
- CRF (1 team)
- Hard-attentional neural model (1 team)

Submissions were compared against a RoBERTa baseline model provided by the organizers (Ginn, 2023)
Glossing accuracy (open track)

Baseline | Best submission
--- | ---
Arapaho | 87.5 | 90
Gitksan | 22.5 | 25
Lezgi | 45 | 47.5
Natügu | 67.5 | 70
Nyangbo | 80 | 85
Tsez | 62.5 | 65
Uspanteko | 70 | 72.5

The 2023 SIGMORPHON Interlinear Glossing Challenge
Impact of training data size (open track)

Figure 2: Impact of different data characteristics (training data size, out-of-vocabulary rate and type-token-ratio) on average word-level glossing accuracy. In addition to the average performance, we also plot the performance of each individual system. Only complete submissions, for all shared task languages, are included in these plots. Abbreviations refer to languages: Arapaho (arp), Tsez (ddo), Gitksan (git), Lezgi (lez), Natügu (ntu), Nyangbo (nyb) and Uspanteko (usp).

8 Conclusion

The 2023 SIGMORPHON Shared Task on Interlinear Glossing received submissions from five teams which presented a wealth of interesting techniques greatly expanding the field of automated interlinear glossing. The submissions achieved substantial improvements over a baseline RoBERTa system. The winning team TÜ-CL achieved a 23.99%-point improvement over the baseline in the closed track and a 17.42%-point improvement in the open track using a hard attention model.
Submissions

Two of the teams incorporated external data and/or used data augmentation.

Team SigMoreFun used an external dictionary for Gitksan and glossed data from the ODIN database.

The winning team Tü-CL used a hard-attention (HA) model.

The HA system delivered the best performance for all languages in the closed track.

The HA system delivered the best performance for all but two languages in the open track.

The feature-based CRF model by LISNTeam turned out to be the best on the lowest-resourced language Gitksan (and Natügu).
Takeaways

Training data size is one of the most important predictors of performance

Hard attention seems to be a promising approach

In the lowest-resourced settings (like Gitksan in the ST) feature-based systems like CRFs may have an edge
References

Lehman, C. Directions for Interlinear Morphemic Translations. Folia Linguistica (1982).


