Automatic Interlinear Glossing for Under-Resourced Languages Leveraging Translations

Xingyuan Zhao[†], **Satoru Ozaki**[†], **Antonios Anastasopoulos**[‡], **Graham Neubig**[†], **Lori Levin**[†]
[†]Language Technologies Institute, Carnegie Mellon University
[‡]Department of Computer Science, George Mason University
{xingyuanz,sozaki}@andrew.cmu.edu antonis@gmu.edu
{gneubig,lsl}@cs.cmu.edu

Abstract

Interlinear Glossed Text (IGT) is a widely used format for encoding linguistic information in language documentation projects and scholarly papers. Manual production of IGT takes time and requires linguistic expertise. We attempt to address this issue by creating automatic glossing models, using modern multi-source neural models that additionally leverage easy-to-collect translations. We further explore cross-lingual transfer and a simple output length control mechanism, further refining our models. Evaluated on three challenging low-resource scenarios, our approach significantly outperforms a recent, state-of-the-art baseline, particularly improving on overall accuracy as well as lemma and tag recall.

1 Introduction

The under-documentation of endangered languages is an imminent problem for the linguistic community. Of the 7,000 languages in the world, an estimated 50% will face extinction in the coming decades, while around 35–42% still remain substantially undocumented (Austin and Sallabank, 2011; Seifart et al., 2018). Perhaps these numbers are no surprise when one acknowledges the difficulty of language documentation – fieldwork demands time, cultural understanding, linguistic expertise, and financial support among a myriad of other factors. Consequently, an important task for the linguistic community is to facilitate the otherwise daunting documentation process as much as possible. Documentation is not a cure-all for language loss, but it is an important part of language preservation; even in an unfortunate worst-case scenario where a language does disappear, a permanent record of the language is saved for posterity, and could hopefully be useful for revitalization purposes.

Documentation written for a Language A in a Language B necessarily involves careful selection of example data from Language A that illustrate various aspects of Language A's grammar. These data are often provided in the form of a typical semi-structured format known as interlinear glossed text (IGT). IGT consists of three lines: the first line being the original data from Language A, the third line its translation into Language B, and the line in between the other lines being a morpheme-by-morpheme gloss, i.e. annotation of the Language A data. The provision of this second line illustrates the morphological structure of Language A to the reader, who is not necessarily informed on the grammar of Language A. An example is outlined in Figure 1. IGT helps linguists better understand how other languages work without prior knowledge of those languages or their grammars.

However, manual segmentation and IGT annotation is still a time- and labor-consuming work, as it requires linguistic training. Therefore, language documentation projects typically manifest a yawning gap between the amount of material recorded and archived and the amount of data that is thoroughly analyzed with morphological segmentation and gloss (Seifart et al., 2018). This gap can be filled using automatic approaches, which could at least accelerate the annotation process by providing high-quality first-pass annotations. Previous approaches to automatic gloss generation include manual rule crafting and deep rule-based analysis (Bender et al., 2014; Snoek et al., 2014), treating the glossing task as a classification problem focusing only on the morphological tags (Moeller and Hulden, 2018) and requiring a lexicon for

Proceedings of the 28th International Conference on Computational Linguistics, pages 5397–5408
Barcelona, Spain (Online), December 8-13, 2020

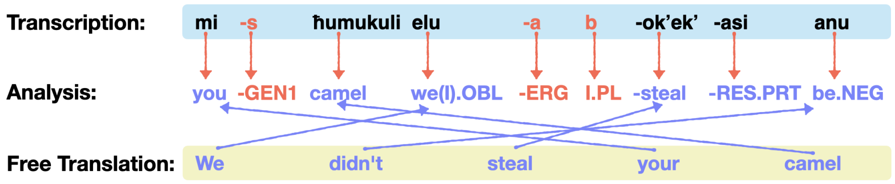
Transcription:
mi -s ħumukuli elu -a b -ok'ek' -asi anu
Analysis:
you -GEN1 camel we(I).OBL -ERG I.PL -steal -RES.PRT be.NEG
Free Translation:
We didn't steal your camel

Figure 1: A Tsez IGT example. Combining information from both the transcription and the translation can aid in deriving the information in the analysis.

stems (Samardžić et al., 2015), and using models based on Conditional Random Fields (CRF) integrated with translation and POS-tagging information (McMillan-Major, 2020). In contrast, our approach is, to our knowledge, the first one to show that modern neural systems are a viable solution for the automatic glossing task, without requiring any additional components or making unrealistic assumptions regarding data or NLP tool availability for low-resource languages.

We rely on the observation that parallel corpora with transcription and translation are likely to be available for many low-resource languages, since the knowledge of the two languages is sufficient for translating the corpus without the need of linguistic training. Documentation approaches relying on parallel audio collection (Bird et al., 2014) are in fact already underway in the Americas (Jimerson and Prud'hommeaux, 2018) and Africa (Rialland et al., 2018; Hamlaoui et al., 2018), among other places. An additional advantage of parallel corpora is that they contain rich information that can be beneficial for gloss generation. As Figure 1 outlines, the stems/lemmas in the analysis are often hiding in the translation, while the grammatical tags could be derived from the segments in the transcription. We hypothesize that the information from the translation can further ground the gloss generation, and especially allow a system that properly takes into account to generalize to produce lemmas or stems unseen during training.

In this work we propose an automated system which creates the hard-to-obtain gloss from an easy-to-obtain parallel corpus. We use deep neural models which have driven recent impressive advances in all facets of modern natural language processing (NLP). Our model for automatic gloss generation uses multi-source transformer models, combining information from the transcription and the translation, significantly outperforming previous state-of-the-art results on three challenging datasets in Lezgian, Tsez, and Arapaho (Arkhangelskiy, 2012; Abdulaev and Abdullaev, 2010; Kazeminejad et al., 2017). Importantly, our approach does not rely on any additional annotations other than plain transcription and translation, also making no assumptions about the gloss tag space. We further extend our training recipes to include necessary improvements that deal with data paucity (utilizing cross-lingual transfer from similar languages) and with the specific characteristics of the glossing task (presenting solutions for output length control).

Our contributions are three-fold:

1. We apply multi-source transformers on the gloss generation task and significantly outperform previous state-of-the-art statistical methods.

2. We propose methods to control the prediction length and extract the alignment of gloss and source token to overcome the drawbacks of neural networks in the gloss generation task.

3. We evaluate our approach in three challenging settings over three languages: Tsez, Lezgian, and Arapaho. We evaluate the generated IGT with various metrics, analyse how they correlate with each other and make informed suggestions for the proper evaluation of the IGT generation task.

2 Interlinear Glossed Text

Interlinear glossed text (IGT) is the name for a format commonly used by linguists in presenting linguistic data. In addition to providing the original sentence in one language (called the "object language") and a free translation into another language (which renders the utterance interpretable by a broader audience), a morpheme-by-morpheme annotation (called the "gloss") of the original sentence is presented between the object language line and the translation line – hence the name “interlinear gloss”. Ex. 1 shows an example of IGT for a sentence from our Tsez dataset.

(1) Tsez

mi-s	ħumukuli	elu-a	b-ok’ek’-asi	anu
you-GEN1	camel	we(I).OBL-ERG	I.PL-steal-RES.PRT	be.NEG

“We didn’t steal your camel.”

Lexical morphemes such as open-class words and their stems are simply glossed as their translations. On the other hand, functional morphemes such as inflectional affixes and closed-class words are glossed with the grammatical categories or function that they encode. Glosses for lexical morphemes are referred to as **STEMs** and those for functional morphemes are referred to as **GRAMs**.

In Ex. 1, the verb *bok’ek’asi* consists of the class I plural prefix *b-*, the stem *ok’ek’* and the resultative participle suffix *-asi*. The stem, being a lexical morpheme, is glossed as its English translation *steal*. On the other hand, the prefix and the suffix are functional morphemes. Thus they are glossed as the collections of grammatical categories they encode. The prefix *b-* is glossed as the combination of I (for class **I** – Tsez has four noun classes I–IV) and PL (for **pl**ural), and the suffix is glossed as RES.PRT (for **res**ultative **part**iciple). When a morpheme is glossed as a collection of multiple GRAMs, they are separated by periods or some other delimiting punctuation.

3 Multi-Source Transformer for Gloss Generation

3.1 Problem Formulation and Model

Our model is built upon the transformer model (Vaswani et al., 2017), a self-attention-based sequence-to-sequence (seq2seq) neural model. Compared to the CRF model used in McMillan-Major (2020), which can only capture local dependencies, a self-attention model can produce context-sensitive hidden representations that take the whole input into account. Moreover, unlike other recurrent (seq2seq) models such as bidirectional LSTM, the Transformer model shows more robust performance in morphology-related tasks under low-resource settings (Ryan and Hulden, 2020). Our architecture choice is also motivated by the promising performance of the model along with its computational efficiency.

The original Transformer is composed of a single encoder and a decoder, each with several layers. Each encoder layer consists of a multi-head self-attention layer and a fully connected feed-forward network, while decoder layers are additionally augmented with multi-head attention over the output of the encoder stack. Our model adds a second encoder to create a multi-source transformer similar to (Zoph and Knight, 2016; Anastasopoulos and Chiang, 2018), in order to incorporate the secondary information from the translation. A visual depiction of our model is outlined in Figure 2.

Let $\mathbf{X^1} = \mathbf{x}^1_1 \ldots \mathbf{x}^1_N$ be a sequence of transcription words, $\mathbf{X}^2 = \mathbf{x}^2_1 \ldots \mathbf{x}^2_M$ a sequence of translation words, and $\mathbf{Y} = \mathbf{y}_1 \ldots \mathbf{y}_K$ be a sequence of the target gloss. A *single-source* gloss generation model attempts to model $P(\mathbf{Y} \mid \mathbf{X}^1)$.

A *multi-source* model can jointly model $P(\mathbf{Y} \mid \mathbf{X}^1, \mathbf{X}^2)$, and thus we need two encoders (see Figure 2b). One encoder transforms the input transcription sequence $\mathbf{x}^1_1 \ldots \mathbf{x}^1_N$ into a sequence of input states $\mathbf{h}^1_1 \ldots \mathbf{h}^1_N$:

$$\mathbf{h}^1_n = \text{enc}^1(\mathbf{h}^1_{n-1}, \mathbf{x}^1_n)$$

and the second encoder transforms the translation sequence $\mathbf{x}^2_1 \ldots \mathbf{x}^2_M$ into another sequence of input states $\mathbf{h}^2_1 \ldots \mathbf{h}^2_M$:

$$\mathbf{h}^2_m = \text{enc}^2(\mathbf{h}^2_{m-1}, \mathbf{x}^2_m).$$

An attention mechanism transforms the two sequences of input states into a sequence of *summed context vectors* via two matrices of *attention weights*:

$$\mathbf{c}_k = \sum_n \alpha^1_{kn} \mathbf{h}^1_n + \sum_m \alpha^2_{km} \mathbf{h}^2_m.$$

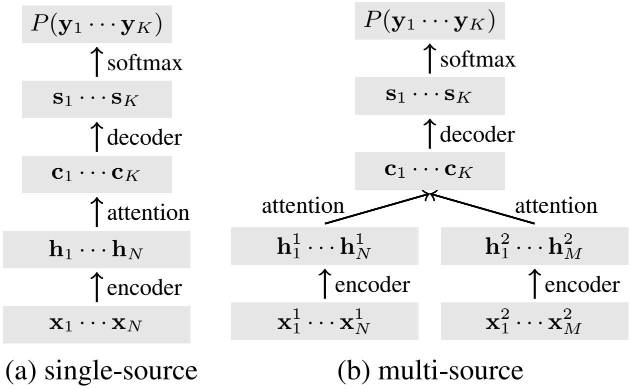
$P(\mathbf{y}_1 \cdots \mathbf{y}_K)$
softmax
$\mathbf{s}_1 \cdots \mathbf{s}_K$
decoder
$\mathbf{c}_1 \cdots \mathbf{c}_K$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$
(a) single-source
$P(\mathbf{y}_1 \cdots \mathbf{y}_K)$
softmax
$\mathbf{s}_1 \cdots \mathbf{s}_K$
decoder
$\mathbf{c}_1 \cdots \mathbf{c}_K$
attention
attention
$\mathbf{h}^1_1 \cdots \mathbf{h}^1_N$
$\mathbf{h}^2_1 \cdots \mathbf{h}^2_M$
encoder
encoder
$\mathbf{x}^1_1 \cdots \mathbf{x}^1_N$
$\mathbf{x}^2_1 \cdots \mathbf{x}^2_M$
(b) multi-source

Figure 2: Simple visualization of the transformer model with a single- and multisource architecture. In the standard *single-source* model, the decoder attends to a single encoder's states. In our *multi-source* setup, we have two input sequences encoded by two different encoders, and attention mechanisms provide two context to the decoder. Note that for clarity's sake there are dependencies not shown.

Finally, the decoder computes a sequence of *output states* from which a probability distribution over output stems/tags can be computed:

$$\mathbf{s}_k = \text{dec}(\mathbf{s}_{k-1}, \mathbf{c}_k, \mathbf{y}_{k-1})$$

$$P(\mathbf{y}_k) = \text{softmax}(\mathbf{s}_k).$$

3.2 Inference

As standard in sequence-to-sequence tasks, we use beam search to search the output space for the most likely output gloss sequence. In addition, though, we incorporate enhancements to handle the particular nature of the gloss generation task.

Length control Unlike other text generation tasks where the generated text can be relatively free in word order, the gloss must map to the transcription morpheme-by-morpheme or word-by-word, dependent on the intended granularity. One drawback of using a seq2seq model for the gloss generation task compared to e.g. a CRF-based approach like (McMillan-Major, 2020) is that a hard constraint of "one output per input" is not enforced by or hard-coded in the model.

Even though structural biases (Cohn et al., 2016) such as hard monotonic attention (Wu and Cotterell, 2019), or source-side coverage mechanisms (Tu et al., 2016; Mi et al., 2016) could remedy this potential issue, we found that there was little need for them, as a simple mechanism to control the final output length during inference was sufficient.[1] The intuition lies in the observation that the length of the output gloss should match that of the input transcription exactly. Hence, during inference we set a minimum desired length of the output sequence, and disallow any candidates shorter than that.[2]

Alignment between gloss and transcription To ensure fair evaluation against the baseline and other models, we need to be able to produce the exact mapping of the output gloss to the input transcription (we discuss the reasoning in Section 4 on evaluation). Luckily, this information lies in the cross-attention weights. Even though neural attention has been reported to behave differently than traditional statistical MT alignments (Ghader and Monz, 2017), we find that our cross-attention between the decoder[3] and the transcription encoder is indeed monotonic and can function as such an alignment. For each output y_k, we align it to a single source word x_i such that $i = \arg\max \alpha^1_{kn}$.

[1] Even in the very low-resource settings, the models produced outputs with the desired length more than 85% of the time without any additional intervention.

[2] This is easily achieved by setting the probability of the end-of-sentence token to negative infinity while the minimum length has not been reached.

[3] We use the cross-attention of the first decoder layer, averaging over all heads.

Language	Translation	Training Examples	Test Examples
LEZGIAN	English	951	119
TSEZ	English/Russian	1584	198
ARAPAHO	English	25208	3151

Table 1: Dataset information for the languages in our experiments.

4 Evaluating Gloss Generation

The characteristics of gloss generation require special care rather than blindly using metrics established for other tasks like machine translation. Previous work uses:

- **Accuracy**: percentage of correct (full) analyses for each token. It is the main metric used in previous work (Samardžić et al., 2015; McMillan-Major, 2020).
- **BLEU** (Papineni et al., 2002): an average of n-gram precision along with a brevity penalty, BLEU is perhaps the most popular reference-based machine translation method. Since our models are inspired from MT, we use it as another indication of quality as it captures accuracy/precision over n-grams, even though the rest of the metrics are more suitable to the automatic glossing task.
- **Precision/Recall**: We further break down the evaluation to focus separately on lemmas and tags. Several previous works prioritize precision over recall, especially by not outputting tags if items are not seen during training, e.g. (Moeller and Hulden, 2018).
- **Error Rate**: A normalized edit distance between the output and the reference, we consider error rate as a good indicator of the overall quality of the generated *sequences*, rather than the more fine-grained metrics such as precision/recall or accuracy. This is also the most reasonable metric if the model is used for computer-assisted glossing where linguists do post-editing.

Comparing all these metrics on our results, we find that unsurprisingly they heavily correlate with each other. We measure correlation with Spearman's rank coefficient, and all correlations end up with coefficients higher than 0.85 ($p < 0.0001$).

5 Experimental Settings

Languages We use three languages as the testbed for evaluating our approach: Arapaho, Lezgian and Tsez. They provide challenging scenarios with different amounts of training data available overall (details are listed in Table 1) as well as the complexity of the produced gloss due to the complex morphosyntax they exhibit.[4] We randomly shuffle and split each dataset into train/validation/test sets with a ratio of 8:1:1.

Arapaho is an Algonquian language spoken in and around Wyoming and western Oklahoma, U.S. With only around 250 fluent speakers, it is very much an endangered language that requires immediate attention and effort for documentation and preservation. Being highly polysynthetic, Arapaho makes heavy use of verb incorporation with morphemes for tense, aspect, modality, evidentiality, and adverbial information, whose (de)incorporation is determined by pragmatic factors such as saliency and emphasis (Cowell et al., 2008). We use the corpus compiled by Kazeminejad et al. (2017), which contains more than 20,000 glossed sentences with English translations.

Lezgian is a member of the Lezgic branch of the Nakho-Daghestanian language family, also known as the (North)east Caucasian family. It is spoken by about 400,000 speakers in southern Dagestan and northern Azerbaijan in the eastern Caucasus. It has a rich consonant inventory with 54 members and an agglutinative morphology. The language has 18 nominal cases, features case-stacking, head-final syntax, ergative-absolutive alignment and no noun-verb agreement (Haspelmath, 1993; Moeller and Hulden, 2018). We use the corpus compiled by Arkhangelskiy (2012), comprised of slightly over 1,000 IGT examples, with free translations in English.

[4] Our modelling approach is not specific to any language, but we only test it on low-resource languages with complex morphosyntax, since they are perhaps the most challenging setting. We leave evaluation on other languages for future work.

Tsez (Dido) is a member of the Tsezic group of the Nakho-Daghestanian language family – it is a close relative of Lezgian.Spoken by 12,467 speakers in Dagestan according to a 2010 Russian census, Tsez also features head-final syntax, ergative-absolutive alignment, rich suffixing morphology, an impressive inventory of cases and a variety of strategies for converb formation (Comrie and Polinsky, forthcoming). We use the Tsez Annotated Corpus compiled by Abdulaev and Abdullaev (2010) which includes almost 1,800 glossed utterances with translations in both Russian and English. We report results using the English translations, since the analysis is also using English stems, and as a result the performance using the Russian translations was worse in preliminary experiments.

Word-level vs. Morpheme-level Settings For each of the language datasets, we have access to gold (hand-created) morpheme-level gloss annotations mapped to morphologically segmented transcriptions. This allows us to evaluate models under two distinct settings, for which we report results separately:

1. **Word-level without gold segmentation**: this setting (referred to as *word-level* hereinafter) closely matches the realistic scenario where we do not have access to gold segmentation of the utterance. This setting is, as a result, more challenging, as the model needs to additionally infer a segmentation. The evaluation in this setting is also performed at the word level (with regards to BLEU and Error Rate) such that the first target for the example in Figure 1 would be “you-GEN1" as a single unit.

2. **Morpheme-level with gold segmentation**: in this setting (which we will refer to as *morpheme-level*) we have access to the gold segmentation of the transcription, hence the glossing task consists of simply providing the correct stems or tags for each segment. Accordingly, we perform the evaluation at the morpheme level: the first two evaluation units from the Figure 1 example would be “you" and "-GEN1", separated. This setting will provide somewhat of an *oracle* score, that would be achievable if a linguist or the community provide correct segmentations for the transcriptions, or if a morphological segmentation tool is available for that language.

Transliteration Cross-lingual training between typologically related languages has shown promising results in several NLP tasks especially in low-resource settings (McCarthy et al., 2019; Anastasopoulos and Neubig, 2019). Two of our evaluation languages, namely Lezgian and Tsez are fairly similar as they are both members of the Nakho-Daghestanian language family, and as such are ideal for cross-lingual transfer. However, Anastasopoulos and Neubig (2019) pointed out that cross-lingual learning can be inversely impeded if the languages do not use the same script even if they are closely genealogically related languages. Lezgian is written in Cyrillic script while Tsez is written in Latin script. To maximally exploit the power of cross-lingual training, we transliterated Lezgian from Cyrillic script to Latin script, and transliterated Tsez from Latin script to Cyrillic script.[5] With the original and the transliterated versions of the training data at hand, we combine them during training into a single training set for the LANGUAGE TRANSFER Model. The evaluation is of course performed on the original test sets with the original corresponding scripts.

5.1 Implementation

We base our implementation on the Joey-NMT toolkit[6] (Kreutzer et al., 2019), which we extended to support multi-source transformer models.[7] The transcription and translation input sentences can be represented at different granularities: either at the word level or at the more recently popular sub-word level. For simplicity we leave this detail out of the results tables, reporting results with the better-performing option in each case. It is worth noting, though, that for Tsez and Arapaho the sub-word representations (obtained using byte-pair-encoding (BPE)[8]) always lead to better results. For the much smaller Lezgian dataset, we saw no difference between sub-word and word-level models, but this lack of difference can be explained by the overall very small size of the vocabulary for the Lezgian dataset.

[5] We use the transliterator provided by `https://pypi.org/project/transliterate/`.

[6] `https://github.com/joeynmt/joeynmt`

[7] Our code will be open-sourced at https://github.com/yukiyakiZ/Automatic_Glossing.

[8] We use the sentencepiece implementation of the BPE method (Sennrich et al., 2016) with vocab size of 2000 for Lezgian, 2500 for Tsez, and 10000 for Arapaho)

For training all Lezgian and Tsez models and the Arapaho model with the subsampled 2,000 training sentences, we use 2 layers for both encoders and the decoder and 2 attention heads. All the embedding and hidden state dimensions are set to 128. We use a batch size of 20. For training the Arapaho model on the original larger dataset, we use 4 layers for all encoders and decoder, with 4 attention heads. The embedding and hidden state dimension are 256, and batch size is 50. For all models, learning rate is initialized to 0.0005 and optimized through Adam (Kingma and Ba, 2015). We also use dropout (Srivastava et al., 2014) with $p = 0.3$. We set the early stop criterion to a minimum learning rate of 1.0e-6. In the end, Lezgian models trained for 3000 epochs, and Tsez and Arapaho models trained for 1100 epochs without reaching that early stop criterion. For inference, we use beam search with a size of 5. We note that we did not perform any grid-search over the hyperparameter space, which leaves room for further improvements in future work.

5.2 Baseline

We use the model by McMillan-Major (2020) as our baseline, since it is the most recent statistical model for gloss generation. It is additionally the only other, to our knowledge, previously proposed approach that leverages both source and translation information to generate glosses. As required by the baseline model, we first pre-processed the data by converting them to Xigt format (Goodman et al., 2015) and then enriched the data using the INTENT system (Georgi, 2016). This enriching step creates a heuristic alignment between the transcriptions and its translation, which is then used to project Part-of-Speech and morpheme tag information from the translation to the corresponding morpheme using heuristics. This additional information is used through heuristic post-editing: an out-of-vocabulary word, for instance, is assumed to be a stem and the aligned translation word is used as the output gloss stem.

For Lezgian and Tsez, we train and evaluate the baseline model using the exact same training and testing data as for our models.[9] For the much larger Arapaho dataset, we were unable to train the baseline using all 25 thousand training examples; the released code does not support GPU processing and after 7 days of training less than 50% of the training procedure had been completed. Instead, we subsampled the Arapaho training data to only 2,000 examples.

For a fair and complete comparison, we report our system's performance trained on both that subsampled training set and the complete one.

6 Results and Analysis

A summary of our main results on the three languages in the realistic word-level scenario (without gold segmentation) is outlined in Table 2, reporting all metrics discussed in Section 4. To determine the gloss corresponding to each morpheme when calculating accuracy, precision and recall, we use the attention-based alignment discussed in Section 3.2.

Our models significantly outperform the baseline in every evaluation metric in all datasets. The results in the subsampled Arapaho dataset are less conclusive (the baseline achieves slightly higher BLEU and lower WER) but both models in this very challenging setting are quite bad to begin with; with an order of magnitude more data, our model's outputs are exceptional, while the baseline was unable to be trained.

In both Tsez and Arapaho, the multi-source approach yields significantly better performance on both lemma and grammatical tag generation than a single-source model. In Lezgian, however, we don't observe any substantial differences between the two models, which we suspect is because the very limited training data may not be sufficient for the model to learn a good gloss-translation cross-attention.[10] Nevertheless, the improvements in Tsez and Arapaho indicate that our model is indeed able to leverage the information provided by the translation to further aid in automatic glossing.

Results with Gold Segmentation In Table 3 we present the same set of results, but this time using the morpheme-level transcription with the gold segmentation. The first thing to note is that the outputs are now better across the board for all metrics and for all models. This indicates that proper segmentation remains a challenge and that the creation of segmentation tools is a valuable endeavor. Nevertheless,

[9] We use the publicly released code: `https://github.com/mcmillanmajora/IGTautoglossing`.

[10] This hypothesis is based on a manual inspection of visualizations of the attention weights.

Dataset	Model	Evaluation				
		BLEU	Accuracy (morpheme)	Prec./Rec. (lemma)	Prec./Rec. (tags)	Error Rate (WER)
Lezgian	Baseline	3.8	0.13	0.12/0.08	0.20/0.05	0.72
	Ours (single-source)	16.6	0.35	0.34/0.38	0.34/0.30	0.54
	Ours (multi-source)	15.5	0.34	0.33/0.36	0.33/0.26	0.55
	+Cross-Lingual Transfer	**26.8**	0.52	0.56/0.57	**0.41/0.41**	**0.40**
	+Length Control	**26.8**	**0.53**	**0.56/0.58**	**0.41/0.41**	**0.40**
Tsez	Baseline	9.4	0.25	0.42/0.24	0.47/0.17	0.64
	Ours (single-source)	21.4	0.35	0.43/0.42	0.49/0.46	0.51
	Ours (multi-source)	25.6	0.37	0.43/0.42	0.52/0.48	0.46
	+Cross-Lingual Transfer	27.5	0.39	0.45/0.45	**0.53/0.49**	**0.43**
	+Length Control	**27.7**	**0.40**	**0.46/0.45**	**0.53/0.49**	**0.43**
Arapaho	Baseline (2000)	8.7	0.18	0.23/0.12	0.15/0.06	0.45
	Ours (multi-source)(2000)	4.2	0.22	0.22/0.21	0.32/0.30	0.50
	Ours (single-source)	39.7	**0.58**	0.70/0.64	0.80/0.74	**0.23**
	Ours (multi-source)	**40.1**	**0.58**	**0.70/0.65**	**0.80/0.75**	**0.23**
	+Length Control	37.1	0.56	0.68/0.65	0.78/0.76	0.26

Table 2: Results on the realistic scenario where no source-side segmentation information is available (word-level transcription). Our approach outperfoms the baselines in all datasets. We **highlight** the best results in each dataset and metric.

even in cases where the baseline is particularly strong when provided with the oracle segmentation, as in the Tsez dataset, its recall in particular lags behind the recall of our approach, for both lemmas and tags.

Cross-Lingual Transfer Cross-lingual transfer significantly improves performance, especially for Lezgian, which only has 951 training sentences before data augmentation using transliteration and cross-lingual training, leading to a 25% reduction in morpheme accuracy error and 27% error reduction in WER (cf. Table 2). Upon manual inspection of the outputs, we find that cross-lingual transfer helps the model make fewer mistakes when predicting stems. As an example, one test sentence with the gold gloss *kazakh nation was* is glossed correctly with cross-lingual transfer, but without it it is glossed as *1pl.abs is-*PST *was*. There may be overlaps between the Tsez and Lezgian vocabularies, which may be contributing to the improved stem predictions.

Cross-lingual transfer also helps in the higher-resource Tsez setting too, with relatively smaller improvements. Unlike the case with Lezgian, we do not observe much improvements in stem prediction after employing cross-lingual transfer.

Overall, the improvements we obtain does show that cross-lingual transfer between related languages is quite a promising direction for future research.

Error Analysis For any model to produce correct IGT given input in some language, it is essential that the model know the morphology of that language. With data-driven machine learning models, it must rely on the segmentation in the training data as its only clue to understanding the morphological structure of the object language. This implies that the model is highly subject to any bias presented in the training data annotation. There is considerable variation in glossing style depending on the language, the linguist, and what the gloss is intended to illustrate. It is not always possible to follow suggested standards for glossing such as the Leipzig Glossing Rules (Bickel et al., 2008).

One bias in the glosses for the so-called "non-distal local cases" in the Tsez dataset is of particular interest. Tsez features 28 non-distal local cases, which are combinatorially formed by selecting one out of the 7 locational series (in-, cont-, super-, sub-, ad-, apud- and poss-) and one out of the 4 directional series (-essive, -lative, -ablative and -versative). Each locational component is associated with a morpheme; so is each directional component. The local case marking is formed by concatenating the morphemes corresponding to each component. For example, the superablative case marking *-ƛ’aaj* is formed by

Dataset	Model	Evaluation BLEU	Accuracy (morpheme)	Prec./Rec. (lemma)	Prec./Rec. (tags)	Error Rate (MER)
Lezgian	Baseline	12.5	0.29	0.22/0.21	0.30/0.26	0.61
	Ours (single-source)	35.8	0.72	0.77/0.80	0.75/0.76	0.24
	Ours (multi-source)	32.1	0.70	0.77/0.79	0.71/0.73	0.27
	+Cross-Lingual Transfer	37.0	0.75	**0.82/0.82**	0.75/0.77	0.24
	+Length Control	**37.2**	**0.76**	**0.82/0.82**	**0.75/0.78**	**0.23**
Tsez	Baseline	43.4	0.84	0.85/0.82	0.84/0.81	0.20
	Ours (single-source)	58.8	0.84	0.89/0.89	0.87/0.87	0.12
	Ours (multi-source)	60.9	**0.87**	**0.91/0.91**	**0.91/0.90**	0.11
	+Cross-Lingual Transfer	62.8	**0.87**	**0.91/0.91**	0.90/0.90	**0.10**
	+Length Control	**62.9**	**0.87**	**0.91/0.91**	0.90/0.90	**0.10**
Arapaho	Baseline(2000)	14.1	0.43	0.40/0.35	0.52/0.44	0.43
	Ours (multi-source)(2000)	18.7	0.55	0.62/0.58	0.69/0.65	0.25
	Ours (single-source)	**64.0**	**0.84**	**0.84/0.77**	0.88/0.88	**0.09**
	Ours (multi-source)	62.6	**0.84**	0.81/0.77	**0.91/0.89**	**0.09**
	+Length Control	62.0	**0.84**	0.81/0.77	0.90/0.89	**0.09**

Table 3: Results assuming gold-level source-side segmentation (morpheme-level transcription). Our approach outperfoms the baselines in all datasets. We **highlight** the best results.

	Model	Predicted Gloss			
	Tsez Transcription	ci-n	nesi-ƛ’	Musa-ƛin	zow-n
Analysis	Gold	name-TOP	DEM1.ISG.OBL-SUPER.ESS	Musa-QUOT	be.NPRS-PST.UNW
	Baseline	name TOP	DEM1.ISG.OBL **CONT**.ESS	Musa QUOT	be.NPRS PST.UNW
	Ours (word)	name-TOP **name-TOP**	DEM1.ISG.OBL-SUPER.ESS-**QUOT**	Musa	be.NPRS-PST.UNW
	Ours (morph)	name-TOP	DEM1.ISG.OBL-SUPER.ESS	Musa-QUOT	be.NPRS-PST.UNW
	English Translation	‘And his name was Musa.’			

Table 4: Four glosses for the same Tsez sentence: gold gloss, baseline prediction with morpheme-level (gold) segmentation, and outputs by our best model without and with gold segmentation.

the “super-” morpheme *-ƛ’* and the “-ablative” morpheme *-aaj* (Comrie and Polinsky, forthcoming). This means the superablative marking can be glossed with the two morphemes separate, as in *-ƛ’-aaj* ‘-SUPER-ABL’ or concatenated, as in *-ƛ’aaj* ‘-SUPER.ABL’. Our Tsez dataset follows the latter convention. However, this makes the compositionality of the local case morphemes opaque to a naive language model – the model does not know that *-ƛ’aaj* can be further broken apart into *-ƛ’* and *-aaj*.

To examine this, we present a Tsez example containing a superessive (‘SUPER.ESS’) marking by comparison four predicted glosses: gold, baseline, our best model (multi-source with cross-lingual transfer and length control) with and without the gold segmentation. The glosses, along with the original Tsez sentence and its English translation are provided in Table 4. The baseline model is provided with morphologically segmented input, yet it incorrectly glosses the superessive as a contessive (‘CONT.ESS’). This is reasonable, because morpheme-level segmentation in this dataset does not elucidate the compositionality of the local case morphemes. This problem can be potentially remedied by providing subword-level input, which can approximate finer morphological segmentation. As such, the superessive is correctly glossed in our model even when provided input without gold segmentation but with, crucially, BPE.

As BPE is only an approximation of morphological segmentation, it is not the perfect solution. The same model incorrectly attaches QUOT, i.e. the quotative particle to the end of *nesi-ƛ’* ‘DEM1.ISG.OBL-SUPER.ESS’, one word before the correct destination *Musa* ‘Musa’. Upon inspection, we found that the input word *Musaƛin* is segmented into *Mus-aƛin* after BPE, and the occurrences of *aƛin* in the training set are overwhelmingly combinations of *-a* and *ƛin* ‘QUOT’, where *-a* is a verbal suffix. In contrast, *Musa* ‘Musa’ is a proper name. The model cannot deduce that *ƛin* is its own morpheme and may appear after nouns, when it has only seen it in tandem with a verbal suffix which, obviously, appears only after verbs.

7 Related Work

Several works have studied the automated IGT generation task (Palmer et al., 2009; Baldridge and Palmer, 2009; Samardžić et al., 2015; Moeller and Hulden, 2018; McMillan-Major, 2020). They mainly used machine learning methods such as CRF and SVM to generate gloss and proposed a series of heuristic post-editing algorithms to improve the performance. Among them, Palmer et al. (2009), Baldridge and Palmer (2009) combined machine labeling and active learning for creating IGT. Moeller and Hulden (2018) tested LSTMs to predict the morphological labels within glosses, but underperformed against CRF models in that task. McMillan-Major (2020) exploited parallel information in gloss generation. These models either depended on an assumption of only a finite set of possible morphological tags, or required additional linguistic information and feature engineering such as POS tags and word alignment.

8 Conclusion

In this work, we make the initial attempts to leverage neural-based models with dual sources – source transcription and its translation – for the automatic gloss generation task. We further extend our model with cross-lingual transfer to overcome data paucity, and with simple techniques to adapt to the characteristics of the glossing task. Our multi-source transformer-based model significantly outperforms previous work on various evaluation metrics in three low-resource languages. We also hold both qualitative and quantitative error analysis. Future research directions include (i) combining a multi-source model and a multi-task model together by taking both word-level transcription and translation as input, to generate morphologically segmented transcription and morpheme-level gloss, (ii) adding additional structural biases for attention, and (iii) data augmentation algorithms and heuristic post-editing methods.

Acknowledgements

This work was supported by the National Science Foundation under grant 1761548.

References

Arsen K Abdulaev and Isa K Abdullaev. 2010. Cezyas folklor [tsez folklore]. *Leipzig/Makhachkala: Lotos.*

Antonios Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. In *Proc. INTERSPEECH.*

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China, November. Association for Computational Linguistics.

TA Arkhangelskiy. 2012. Electronic corpora of the albanian, kalmyk, lezgian, and ossetic languages. *Automatic Documentation and Mathematical Linguistics*, 46(2):118–123.

Peter K Austin and Julia Sallabank. 2011. *The Cambridge handbook of endangered languages.* Cambridge University Press.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore, August. Association for Computational Linguistics.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. Learning grammar specifications from IGT: A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

B. Bickel, B. Comrie, and M. Haspelmath. 2008. The leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Steven Bird, Lauren Gawne, Katie Gelbart, and Isaac McAlister. 2014. Collecting bilingual audio in remote indigenous communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1015–1024, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885.

Bernard Comrie and Maria Polinsky. forthcoming. Tsez. In Yuri Koryakov, Yury Lander, and Timur Maisak, editors, *The Caucasian Languages. An International Handbook*. De Gruyter Mouton.

Andrew Cowell, Alonzo Moss Sr., and Alonzo Moss. 2008. *The Arapaho Language*. University Press of Colorado.

Ryan Alden Georgi. 2016. *From Aari to Zulu: massively multilingual creation of language tools using interlinear glossed text*. Ph.D. thesis.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: extensible interlinear glossed text for natural language processing. *Language Resources and Evaluation*, 49:455–485.

Fatima Hamlaoui, Emmanuel-Moselly Makasso, Markus Müller, Jonas Engelmann, Gilles Adda, Alex Waibel, and Sebastian Stüker. 2018. BULBasaa: A bilingual basaa-French speech corpus for the evaluation of language documentation tools. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Martin Haspelmath. 1993. *A Grammar of Lezgian*. De Gruyter Mouton.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages—the case of Arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18, Honolulu, March. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China, November. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy, August. Association for Computational Linguistics.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. *Proceedings of the Society for Computation in Linguistics*, 3(1):338–349.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, Colorado, June. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Annie Rialland, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Elodie Gauthier, Pierre Godard, and Jamison Cooper-Leavitt. 2018. Parallel corpora in mboshi (Bantu c25, congo-brazzaville). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Zach Ryan and Mans Hulden. 2020. Data augmentation for transformer-based G2P. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 184–188, Online, July. Association for Computational Linguistics.

Tanja Samardžić, Robert Schikowski, and Sabine Stoll. 2015. Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 68–72, Beijing, China, July. Association for Computational Linguistics.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen Levinson. 2018. Language documentation twenty-five years on. *Language*, 94:e324–e345, 12.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of plains cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34.