Is linguistically-motivated data augmentation worthwhile?

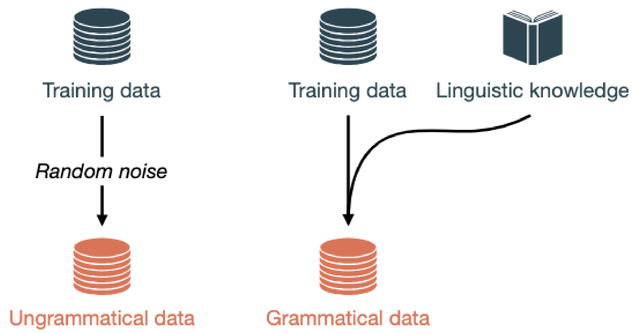
Ray Groshan, Michael Ginn, Alexis Palmer

University of Colorado Boulder

Motivation

Existing NLP methods are not as effective for low-resource languages, as they lack sufficient labeled training data. One way to overcome this is through creating synthetic data using data augmentation. We compared two categories of augmentation strategies, linguistically-motivated and random, through experiments on two low-resource languages, Uspanteko and Arapaho.

Data Augmentation Approaches



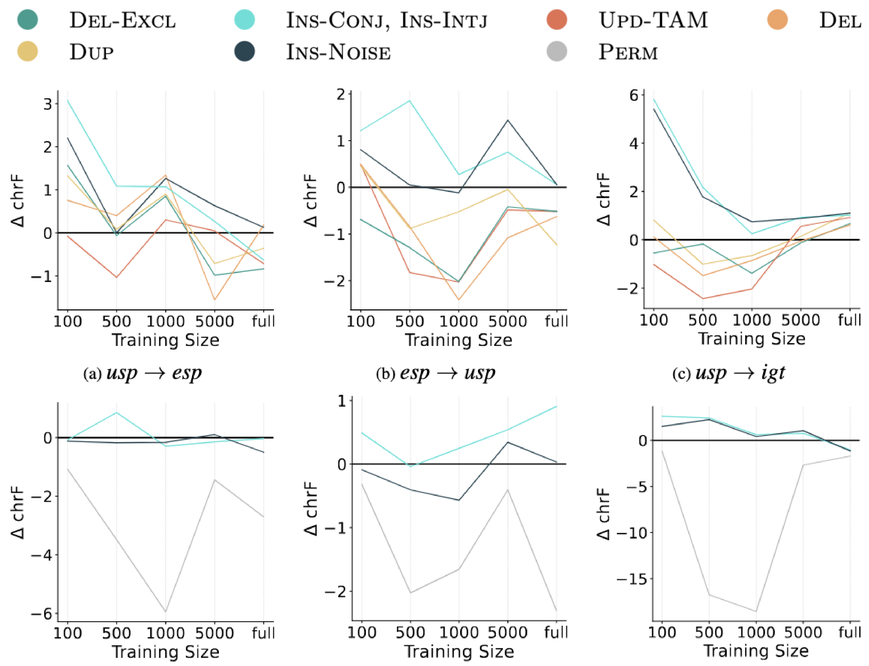
Methods

We used the Ginn et al. (2023) datasets to train google/byt5-small models using curriculum learning for both languages.

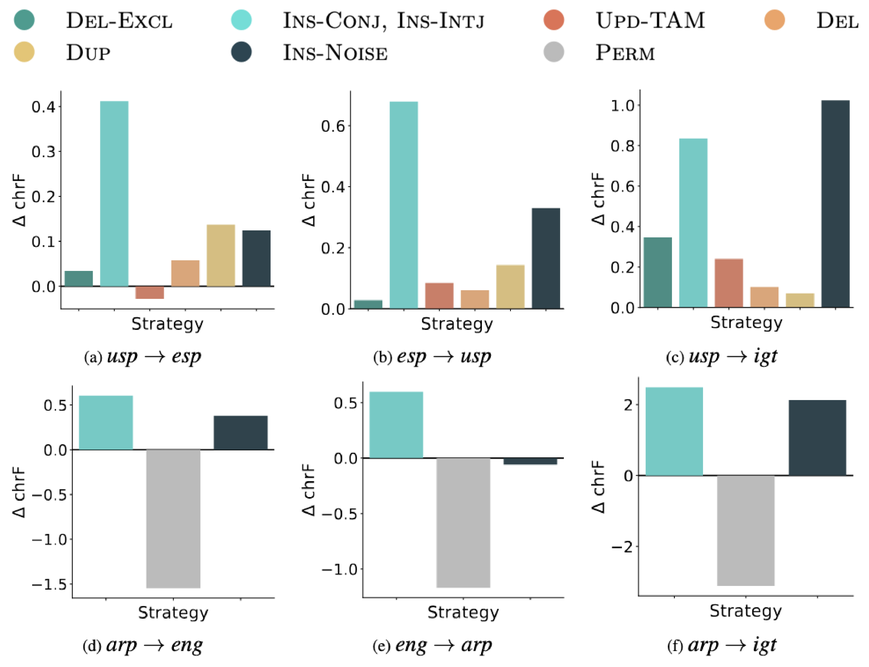
Our augmentation methods were inspired by careful review of grammars and IGT glosses of the languages, as well as easy data augmentation (EDA) as proposed by Wei and Zou (2019). They were evaluated on three tasks: machine translation in both directions and interlinear glossing.

Name	Category	# examples
Uspanteko		
DEL-EXCL	Linguistic	0.2
INS-CONJ	Linguistic	20.0
UPD-TAM	Linguistic	0.3
DEL	Non-linguistic	0.2
DUP	Non-linguistic	0.3
INS-NOISE	Non-linguistic	20.0
Arapaho		
INS-INTJ	Linguistic	20.0
PERM	Linguistic	10.0
INS-NOISE	Non-linguistic	20.0

Results



Difference in (test set) chrF score for various individual augmentation strategies from the baseline for Uspanteko (top) and Arapaho (bottom). Averaged over three runs at each point.



Average difference in (test set) chrF score between combinations including a given strategy and combinations excluding that strategy. Averaged over all runs and training sizes.

Augmentation Strategies

Uspanteko

DEL-EXCL: Randomly deletes a word by index, excluding verbs
INS-CONJ: Inserts a random conjunction or adverb at the start of the sentence
UPD-TAM: Updates the aspect marker on the verb
DEL: Randomly deletes a word by index
DUP: Duplicates the word at a randomly chosen index
INS-NOISE: Inserts a random word at the start of the sentence

Arapaho

INS-INTJ: Inserts an interjection at the start of the sentence
PERM: Produces up to 10 permutations of the original word order
INS-NOISE: Inserts a random word at the start of the sentence

Augmentation Examples

```
Toos cha'
Toos cha'
ADV VI
entonces dice
Entonces Dice
```

INS-CONJ – Uspanteko

```
Yeheihoo beeheeteihini3 hee3eihok
gee.whiz IC.all.powerful-4S said.to.s.o .
Gee whiz the Lord said to him
```

INS-INTJ – Arapaho

Conclusions

We find that linguistically-motivated strategies **can improve** performance, but only if they are not significantly different from the training data.

INS-CONJ and INS-INTJ, the two strategies that created examples most like the training data, offered the most improvement compared to the linguistically-naive strategies. Linguistically-motivated strategies that create grammatically valid, but unlikely examples (i.e. PERM), however, were detrimental to performance.