Building Neural Net Machine Translation Systems Using Interlinear Glossed Texts

Abstract

The gloss data that are widely used in theoretical linguistics are hidden treasures for machine translation. The current paper introduces the gloss data to the natural language processing world and demonstrates a practical and effective way to incorporate gloss data into the training data for training neural net machine translation systems.

1 Introduction

Interlinear Glossed Text (IGT) is widely used in linguistic studies. (1) is an example of Scottish Gaelic IGT.

(1) Tha a athair nas sine na a mhàthair.
be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother
'His father is older than his mother.'

In a simple form of IGT, the first line is a sentence of the language of interest, the second line is a word-by-word translation, annotated with relevant grammatical information, and the third line is an English translation (see Bickel et al. (2008) for the complete formats and options of IGT).

The Innovation is to incorporate the gloss information of Interlinear Glossed Text data into neural net machine translation systems.

The properties of the gloss data make it a better training data than natural language data (Scottish Gaelic in the current case) for the following reasons: 1) gloss representations cluster words with different forms into a single representation; 2) gloss representations reserve the semantic difference between homographs; 3) gloss representations are sensitive to hierarchical structures (sentence parsing).

Consider the definite article in the following Gaelic examples.

(2) tha mi a' sireadh **an** leabhair bhig ghuirm.
be-PRES-IND 1S PROG searching-VN **ART** book-G small-G blue-G
'I am looking for the small blue book.' (Lamb, 2001, p. 29)

(3) **am** fear mòr.
ART man big
'a big man.' (Lamb, 2001, p. 31)

(4) thuit **a'** chlach air cas mo mhnà.
fall-PAST **ART** stone on foot 1S-POSS wife-G
'the stone fell on my wife's foot.' (Lamb, 2001, p. 30)

(5) doras **na** sgoile(adh).
door-N **ART** school-G
'the door of the school.' (Lamb, 2001, p. 29)

(6) a chuir air dòigh **nan** àiridhean a-muigh a rubh' Eubhal agus an oidhche seo.
to put-INF on order **ART** sheilings out-LOC to point Eaval and ART night this
'the girls big house.' (Lamb, 2001, p. 100)

(7) fèis **nam** bàrd.
festival **ART** poet.PL.GEN
'festival of the poets.' (Lamb, 2001, p. 107)

The definite article in Scottish Gaelic may be realized as the following forms: as *an*, *am*, *a'*, *na*, *nan* or *nam*. The alternation is determined by the case, gender and number of noun phrase that it modifies, and additionally phonological property of the word following it also changes the form of the definite article (Lamb, 2001). All these different realizations refer to the same concept, the definite article. The gloss representation nicely clusters them together as *ART*. Learning the general distribution of the article and all its different forms is a challenge for the MT system, but the glossing information should make this easier.

Also the glosses distinguish different concepts with the form. Consider the word *a'* in the following examples.

(8) tha mi **a'** sireadh an leabhair bhig ghuirm.
be-PRES-IND 1S **PROG** searching-VN ART book-G small-G blue-G
'I am looking for the small blue book.' (Lamb, 2001, p. 29)

(9) thuit **a'** chlach air cas mo mhnà.
fall-PAST **ART** stone on foot 1S-POSS wife-G
'the stone fell on my wife's foot.' (Lamb, 2001, p. 30)

Critically *a'* in (8) is a progressive aspect marker while in (9) the some form denotes to definite article. Again, the semantic difference is preserved in the gloss representations but not in natural language words. The gloss data also provides hierarchical (non-linear) syntactic parsing information. Consider the Gaelic word, *a'*, in the above examples again, the gloss of which is decided by the hierarchical structure (i.e. constituency) of the sentences instead of the linear order of the words.

In short, glosses are more purified and transparent than natural language words in terms of representing meanings. Therefore, theoretically the incorporation of the gloss data should improve the translation systems. Specifically, we propose the following hypothesis:

(10) **Gloss-helps hypothesis: the translation systems trained with the gloss data incorporated should outperform the systems trained with only Gaelic and English sentences pairs (i.e. without gloss data).**

The hypothesis can have two versions, strong and weak:

a. Strong version: Gloss may replace the source natural language totally, and the system outperforms the system trained with source natural language to target language sentence pairs (i.e. the baseline systems).

b. Weak version: Gloss only increases the performance of the baseline systems, but cannot replace the source language.

The experiments reveal that replacing Gaelic words with glosses doesn't improve the performance of the translation systems. Thus, the strong version (replacing-Gaelic-with gloss) of the Gloss-helps hypothesis is not attested. However, it is found that if the Gaelic data and the gloss data are combined in a specific way as the training data, we term which as Parallel-Partial treatment (see section 5.2), the performance of the systems is improved significantly.

The rest of the paper is organized as follows: Section 2 discusses relevant literature, Section 3 describes the constant parameter settings across all the experiments, section 4 tests the hypothesis in (10a), section 5 tests the hypothesis in (10b) and exemplifies an effective way of incorporating the gloss information, section 6 reports other possible ways of blending glosses and Gaelic sentences, and section 7 is the conclusion.

2 Related Work

Attempts to improve machine translation systems by incorporating explicit linguistic information are reported in the literature. Syntax information is known to be effective in improving statistical machine translation (SMT). The efforts of using syntax information even derive a special type of SMT, termed as syntax-based SMT (Williams et al., 2016). The same trend is also found in neural net machine translation. For example, Sennrich and Haddow (2016) exploit the information of lemmas, part of tags, morphology of words, and dependency parses of sentences to improve MT systems. Nadejde et al. (2017) incorporate the Categorial grammar parse tags of the target sequences.

The Parallel-Partial treatment section 5.2 may be viewed as a form of multi-task Sequence to Sequence Learning (Luong et al., 2015). Specifically, the parallel part of the treatment is very similar to the data manipulation used in building multi-language translation systems (Johnson et al., 2016).

3 Technical Settings of the Machine Translation Experiments

The experiments are conduced by using OpenNMT (Klein et al., 2017), which implements the state-of-the-art neural net machine translation algorithms (Cho et al., 2014a,b; Bahdanau et al., 2014). The following default hyper-parameter

settings of OpenNMT[1] are used across all models so that the only independent variable is the type of the training data:

- Word vector size: 500
- Type of recurrent cell: Long Short Term Memory
- Number of recurrent layers of the encoder and decoder: 2
- Number of epochs: 13
- Size of mini batches: 64

The settings of the hyper-parameters do have effects on the performances of the trained models. However, given that finding the optimal settings of the hyper-parameters is not relevant to our current research and causing unnecessary complications, the process of optimizing the settings of the hyper-parameters is not implemented, and we simply adopt OpenNMT's default settings. The employed settings of the hyper-parameters should be viewed as arbitrarily chosen options, and there are rooms to tune the models for better performance. We will leave the question of what hyper-parameters are optimal for our data for future research. Critically, these settings are viewed as constants, so that we can focus on the effects of different treatments on the source sequences in the translation experiments.

4 Gloss Representation Solely Does NOT Outperform Gaelic Sentences

This section tests the strong version of Gloss-helps hypothesis in (10a). Given the assumption that gloss may be better than any natural language in terms of representing meanings, it is expected that for neural net machine translation systems it is easier to learn how to translate from the glosses of Scottish Gaelic to English than to learn how to translate from Scottish Gaelic to English. However, the results show that there is no significance difference between the two types of data (i.e. GLOSS → English and Gaelic → English).

4.1 Scottish Gaelic Interlinear Glossed Text Data

We use the same Scottish Gaelic IGT corpus[2] for all experiments. The corpus has 8,367 Gaelic sentences, and in term of words, it has 52,778 Gaelic words/glosses. The data of the corpus is from two different sources: linguistics fieldwork and data elicitation.

4.2 Procedure of the Experiments

We use repeated random sub-sampling validation to compare the performances of the two type of models. Totally we have 8,388 indexed 3-tuples of a Gaelic

[1] See their documentation for the complete default hyper-parameter settings: http://opennmt.net/OpenNMT-py/.

[2] Full citation cannot be given without compromising anonymity.

sentence, a gloss line and an English translation. In the interlinear glossed text example below, each line is an argument of a 3-tuple sample.

(11) Tha a athair nas sine na a mhàthair.
be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother
'His father is older than his mother.'

The 3-tuple representation of the above example is:

(12) <"Tha a athair nas sine na a mhàthair .", "be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother", "His father is older than his mother .">

First, the samples (i.e. the 3-tuples) are randomly split into three datasets: training set (N=6,388), validation set (N=1,000), and test set (N=1,000)[3].

For each index, the 3-tuple is split into two pairs: <gloss, English>, <Gaelic, English>, so that later we can compare the different effects of gloss lines and Gaelic sentences.

(13) Gloss to English

a. $\text{GLOSStoEN}_{\text{Train}} = \{< gloss_i, En_i > | \ i \in Index_{\text{Train}}\}$

b. $\text{GLOSStoEN}_{\text{Validation}} = \{< gloss_i, En_i > | \ i \in Index_{\text{Validation}}\}$

c. $\text{GLOSStoEN}_{\text{Test}} = \{< gloss_i, En_i > | \ i \in Index_{\text{Test}}\}$

d. Example: <"be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother", "His father is older than his mother .">

(14) Gaelic to English

a. $\text{GDtoEN}_{\text{Train}} = \{< GD_i, En_i > | \ i \in Index_{\text{Train}}\}$

b. $\text{GDtoEN}_{\text{Validation}} = \{< GD_i, En_i > | \ i \in Index_{\text{Validation}}\}$

c. $\text{GDtoEN}_{\text{Test}} = \{< GD_i, En_i > | \ i \in Index_{\text{Test}}\}$

d. Example: <"Tha a athair nas sine na a mhàthair .", "His father is older than his mother .">

The models are trained with the training set and validation set.

(15) Models:

[3]Here the random sampling process is achieved by using the `random.sample(population, k)` function in the standard library of python.

a. $\text{Model}_{\text{GLOSStoEN}}$ =
Model trained with $\text{GLOSStoEN}_{\text{Train}}$ in (13a) and $\text{GLOSStoEN}_{\text{Validation}}$ in (13b)

b. $\text{Model}_{\text{GDtoEN}}$ =
Model trained with $\text{GDtoEN}_{\text{Train}}$ in (14a) and $\text{GDtoEN}_{\text{Validation}}$ in (14b)

The two trained models (gloss-to-English and Gaelic-to-English) then take the right source sequences of the test sets (i.e. glossing lines and Gaelic sentences for $\text{Model}_{\text{GLOSStoEN}}$ and $\text{Model}_{\text{GDoEN}}$ respectively) as inputs and then generate the predicted target sequences (i.e. English sentences).

To evaluate the model, the predicted target sequences are checked against the target sequences of the test set (i.e. the gold standard/human-translated English sentences). Specifically, the BLEU score metric (Papineni et al., 2002) of each prediction is calculated using the multi-bleu.perl[4] script, a public implementation of Moses (Koehn et al., 2007).

(16) Gold-Standard = English sentences in (13c) = English sentences in (14c)

Note that the gold-standard is the same because they are the same English sentences in the 3-tuples samples. Then the two sets of predicted English sentences are evaluated, yielding two BLEU scores.

(17) Scores:

a. $\text{Score}_{\text{GLOSStoEN}} = \text{BLEU}(\text{Gold-Standard}, \text{Predictions}_{\text{GLOSStoEN}})$

b. $\text{Score}_{\text{GDtoEN}} = \text{BLEU}(\text{Gold-Standard}, \text{Predictions}_{\text{GDtoEN}})$

This procedure of splitting the data into three sub-sets, training the models, and evaluating the models is executed for ten times. Recall that all the sets are randomly chosen samples. In this manner, we will be able to reduce the chance of sampling errors and do a t-test to compare the BLEU scores of the treatments.

4.3 Result

After ten rounds of repeated random sub-sampling validation, ten pairs of scores of the two models are generated. See the appendix for the complete BLEU score for each models.

The average score of the $\text{Models}_{\text{GLOSStoEN}}$ is only sightly higher than the average score of the $\text{Models}_{\text{GDtoEN}}$. Also, a paired t-test shows that the difference between the two types of models is NOT statistically significant (M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{\text{GLOSStoEN}}$=17.70, $SD_{\text{GLOSStoEN}}$=1.78; t(9)=1.97, p=0.080).

[4]The script can be downloaded from: https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

4.4 Discussion

It is assumed that the performances of the machine translation systems are correlated with the quality of the representation of meanings in the source sequences. Better representations of meanings yield better machine translation systems. Given the results in (4.3) that the gloss models are not better than the Gaelic models, it is concluded that glosses and natural languages are equally good in terms of representing meanings. The strong version of the Gloss-helps hypothesis does not hold.

There are several remarks that need to make for the current result. First, the result falsifies the point of view about glosses that the gloss line is a golden semantic representation hand-crafted by linguists. It turns that this artificial language, the gloss lines, is only marginal better than Gaelic statistically. This can be viewed as an evidence of language evolution. The written form of a natural language is actually already optimized for representing semantics to the same degree as the gloss line representations. Second, if we want to actually apply the gloss treatment to translate a Gaelic sentence to English, we encounter an immediate problem. The actual source sequence is a Gaelic sentence, while the required source sequence for the gloss treatment is a gloss line. Given this, even if the gloss treatment should work, it is not practical unless we may convert Gaelic sentences to gloss lines perfectly.

We may now combine Gaelic sentences and gloss lines as the training data to test the weak version of the Gloss-helps hypothesis. The experiments and results are reported in the next section.

5 The Right Way of Combining Gaelic Words with Glosses: Parallel-Partial Treatment

In the previous section, we attempt to build systems by using the *gloss treatment* to outperform the baseline system. It turns that using gloss line solely is not effective enough to improve the system. However, this result does not falsify the gloss-helps hypothesis; instead, it indicates that combination of the gloss line data and the Gaelic sentence data is necessary. In other words, the questions now are:

(18) a. Will adding the gloss data into the Gaelic data improve the translation system?

b. If yes, what are the right ways of blending these two types of meaning representations together?

This section reports that a specific way of combining Gloss data and Gaelic date (termed as '*Parallel-Partial*' treatment) boosts the performance significantly. The models trained with this specially arranged training data also significantly outperform Google's Gaelic-to-English translation system[5] (see the appendix for the BLEU scores).

[5] We used a free Google translation API (Han, 2018) to translate the Gaelic sentences in our test set. Then we calculate the BLEU scores of with the target sequences of our test set as the gold standard.

5.1 The Underlying Heuristics

At a high level, neural net sequence to sequence learning algorithm is to learn how to map a high-dimension space to another high-dimension space. In the settings of machine translation, each dot in the high-dimension space is a meaning representation. Linking one dot to another dot is converting one meaning representation to another, yielding the effect of translation. Given this heuristics, we may just feed the machine with all the available meaning mappings. Given the assumption that the gloss lines are linguistically guided meaning representations, they are suitable training data for building machine translation systems. Specially, with the gloss data, we let the machine to learn the following mappings:

(19) Mappings Learned in the ParaPart treatment
   a. Gaelic sentences → English sentences
   b. Gloss lines → English sentences
   c. Gloss lines → Gaelic sentences
   d. Gaelic words → Gloss items

5.2 The 'Parallel-Partial' Treatment Outperforms the Baseline Significantly

5.2.1 Data Preprocessing Using the Parallel-Partial Treatment

The Parallel-Partial treatment uses the training and validation data of the baseline system and that of the gloss treatment system. The training and validation data of the baseline system are pairs of a Gaelic sentence and a English sentences (see (14a) and (14b)), and the data of the gloss treatment are pairs of a gloss line and a English sentences (see (13a) and (13b). These two groups of data are combined in a parallel manner in the current treatment. Additionally, we also fed the machine with mappings from glossing lines to Gaelic sentences. Now the size of training set and validation set is tripled. In the baseline system and the gloss treatment system, we have 6,388 samples in the training set and 1,000 samples in the validation set. The current treatment has 19,164 (6,388*3) samples in the training set and 3,000 (1,000*3) samples in the validation set. This is the *parallel* part of the treatment.

Additionally, we utilize the alignment property between the Gaelic word and the gloss to further build pairs of a Gaelic word and a gloss. These pairs are also included into the training set and validation set of the current treatment. This is the *partial* part of the treatment.

For concreteness, consider the following interlinear glossed text:

(20) Tha a athair nas sine na a mhàthair.
   be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother
   'His father is older than his mother.'

With the interlinear glossed text, the parallel treatment will generate two pairs of samples:

(21) a. Gaelic to English:
   <"Tha a athair nas sine na a mhàthair", "His father is older than his mother.">

b. Gloss to English:
<“be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother”, “His father is older than his mother”>

c. Glosses to Gaelic:
<“be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother”, “Tha a athair nas sine na a mhàthair”>

The partial treatment then generates pairs of a Gaelic word and a gloss token:

(22) a. <“Tha”, “be.pres”>
b. <“a”, “3sm.poss”>
c. <“athair”, “father”>
d. <“nas”, “comp”>
e. <“sine”, “old.cmpr”>
f. <“na”, “comp”>
g. <“a”, “3sm.poss”>
h. <“mhàthair”, “mother”>

The samples like (21) and (22) are the training data for the Parallel-Partial treatment.

5.2.2 Results of the Parallel-Partial Treatment

The same technical settings and the same test sets in the previous experiments are used, and the same procedures are executed. The only difference is the training and validation data. The result show that the Parallel-Partial treatment has a tremendous effect in improving the baseline system (M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{ParaPart}$=32.10, $SD_{ParaPart}$=1.33; $t(9)=48.95$, $p<0.01$).

5.2.3 Discussion

With the ParaPart treatment, the baseline systems are improved for more than 93 percent. This result suggest the validity of our heuristics in section 5.1, and provide strong evidence for the gloss-helps hypothesis in (10).

6 Other Possible Treatments

There other possible ways of blending the Gaelic sentences and gloss lines. However, all of these treatments are not as effective as the Parallel-Partial treatment. Again, the same procedure and the same test datasets are used across all the experiments.

6.1 The Parallel Treatment

6.1.1 Method of the Parallel Treatment

The Parallel treatment is using the parallel part of the Parallel-Partial treatment without exploiting the alignment properties of gloss lines. It is expected that

this treatment will improve the baseline systems but will not be as effective as the Parallel-Partial treatment.

With this treatment, a chunk of interlinear glossed text is split into three pairs. For example, the chunk of interlinear glossed text in (23) becomes three samples in (24):

(23) Tha a athair nas sine na a mhàthair.
be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother
'His father is older than his mother.'

(24) a. Gaelic to English:
<"Tha a athair nas sine na a mhàthair", "His father is older than his mother.">

b. Gloss to English:
<"be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother", "His father is older than his mother">

c. Gloss to Gaelic:
<"be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother", "Tha a athair nas sine na a mhàthair">

6.1.2 Results of the Parallel Treatment

The experiments had our expected results. The Parallel treatment is effective in improving the baseline systems (M_{GDToEn}=16.59, SD_{GDToEn}=0.74; M_{Para}=29.56, SD_{Para}=1.46; t(9)=34.42, $p < 0.01$). However, the best treatment (i.e. the Parallel-Partial treatment) is still far better than this Parallel treatment (M_{Para}=29.56, SD_{Para}=1.46; $M_{ParaPart}$=32.10, $SD_{ParaPart}$=1.33; t(9)=8.76, $p < 0.01$).

Critically, the comparison between the Parallel-Partial treatment and current Parallel-Only treatment shows the effectiveness of the word-gloss alignments. Our conjecture on the effectiveness is that with the pairs of a gloss item and a Gaelic word present in the training data, the burden of the attention algorithm (Bahdanau et al., 2014) is largely alleviated. In other words, instead of asking the attention algorithm to estimate what to attend to, we explicitly teach the machine the alignment between the Gaelic word and the corresponding gloss.

6.2 Interleaving Gaelic Words and Gloss Items And Concating them

6.2.1 Method of the Interleaving Treatment

Instead of putting the pairs of a Gaelic sentence and a English sentences and the pairs of a gloss line and a English sentence in a parallel manner, we may just literally blend a Gaelic sentence and a gloss line by interleaving them[6]. Consider the following example:

(25) a. Tha a athair nas sine na a mhàthair.
be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother
'His father is older than his mother.'

[6] Nadejde et al. (2017) incorporate the Categorial grammar parse tags into natural sentences by interleaving the tags and the words.

b. <“Tha be.pres a 3sm.poss athair father nas comp sine old.cmpr na comp a 3sm.poss mhàthair mother”, “His father is older than his mother”>

Given the chuck of interlinear glossed text data in (25a), the Interleaving treatment generates the sample in (25b).

This way of blending gloss lines and Gaelic sentences may add useful information into the training data; however, the downside of this method is to increase the length our samples on the source sequence side. In neural net machine learning, the longer the sequences are, the harder it is to preserve all the information. So, this treatment may not be effective.

It turns out this treatment has a significant negative effect (M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{interleavingGdGLOSS}$=12.26, $SD_{interleavingGdGLOSS}$=0.74,; t(9)=-17.06, p=0.000). This is not the right way of incorporating gloss line data.

6.2.2 Method of Concating Gaelic Words and Gloss Words

A quick and close amendment of the Interleaving approach is to concatenate the aligned Gaelic word and gloss item as a single token. Given the same chunk of interlinear glossed text data, this treatment generates the following sample:

(26) a. Tha a athair nas sine na a mhàthair.
be.pres 3sm.poss father comp old.cmpr comp 3sm.poss mother
‘His father is older than his mother.’

b. <“Tha_be.pres a_3sm.poss athair_father nas_comp sine_old.cmpr na_comp a_3sm.poss mhàthair_mother”, “His father is older than his mother”>

Concating words and glosses does solve the long sequence problem; however, it causes the sparse data problem. In this arrangement, the number of the types of tokens is increased; the number of tokens of each type is decreased. Thus all the samples are put in a larger space. So, the treatment may not be effective either.

6.2.3 Results of Concating Gaelic Words and Gloss Words

The result shows that this treatment hurts the baseline systems instead of improving them (M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{ConcatGLOSSGaelic}$=15.44, $SD_{ConcatGLOSSGaelic}$=1.23,; t(9)=-3.64, p=0.010).

7 Conclusion

In the current paper, we introduce an very effective way of incorporating the gloss data into neural net machine translation systems. The immediate merit is that it works. Additional, how theoretical linguistics may work hand in hand with natural language processing, and how neural net machine learning may exploit linguistics are important questions in both fields (see Pater (2017) for a nice discussion on this topic). In addition to practically building better MT systems, the current work also exemplifies how theoretical linguistics may work hand in hand with natural language processing successfully.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Balthasar Bickel, Bernard Comrie, and Martin Haspelmath. 2008. The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses. *Revised version of February*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

SuHun Han. 2018. Googletrans. https://github.com/ssut/py-googletrans.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

William Lamb. 2001. *Scottish Gaelic*, volume 401. Lincom Europa.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting target language ccg supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Joe Pater. 2017. Generative linguistics and neural networks at 60: foundation, friction, and fusion.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892*.

Philip Williams, Rico Sennrich, Matt Post, and Philipp Koehn. 2016. Syntax-based statistical machine translation. *Synthesis Lectures on Human Language Technologies*, 9(4):1–208.

A Supplemental Material

Round	Baseline	GLOSS	ParaPart	Para	Interleaving	Concat	Google Translation
0	17.29	18.39	32.64	25.42	13.67	15.42	22.09
1	16.42	18.00	32.28	25.32	12.49	14.31	25.38
2	15.29	16.02	29.94	20.72	11.01	15.38	23.72
3	15.97	20.22	31.18	22.22	12.33	14.18	23.21
4	17.79	19.02	32.83	24.27	12.56	18.63	22.31
5	16.73	15.53	31.11	24.55	12.13	14.89	23.41
6	17.11	18.00	32.19	27.03	11.55	15.16	24.53
7	16.37	20.08	33.52	25.34	12.78	15.20	22.78
8	15.93	15.82	30.93	24.24	12.43	15.50	25.67
9	16.99	15.93	34.35	25.96	11.65	15.72	23.42
Mean	16.59	17.70	32.10	24.51	12.26	15.44	23.65

Table 1: BLEU scores of the treatments: Ten rounds of repeated random sub-sampling validation are conducted. For each round, the same sets of IGTs are used. Each column is a treatment, and each row is a single round of repeated random sub-sampling validation. The last column is the scores of Google Translation. We used a free Google translation API (Han, 2018) to translate the same set of test Gaelic sentences into English.