Chapter N: Experimenting Interlinear Glossing Text

Yuan-Lu Chen

March 30, 2018

(

Assuming that in the previous chapters the following points are addressed already:

- The nature of glosses has been well-explained (Target audience: CS people without any formal linguistics background):
  - What glosses are: A basic intro of interlinear gloss for non-linguists
  - The golden nature of glosses (encodes NON-LINEAR syntax (i.e. structure parse) and semantics information)
  - The potential of gloss:
    - potential: providing disambiguation, labeling important grammar morphemes in the source language, providing morphological analysis, providing one-to-many and many-to-one relations of source tokens and target tokens.
- A history of machine translation, and a non-mathy description of the methods of doing machine translation. (Target reader: theoretical linguists)

)

1 Introduction

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

In supervised machine learning models, two factors effects the performance of the trained systems (Kotsiantis et al., 2007): a.) the quality of the training data and b.) the choices of the features. The properties of the gloss data as described in *CHAPTERXYZ* make it a better training data than natural language data (Scottish Gaelic in the current case) for the following reasons. First, glosses are more purified that natural language words. The most ideal meaning representation system should be built with one-meaning-to-one-representation mappings; in other words, a meaning is mapped to one and only one representation. Natural languages fail to do so, given that synonyms and ambiguous words/phrases are ubiquitous in natural languages. Glosses provide this one-to-one mapping. Second, the gloss data provides hierarchical (non-linear) syntactic

parsing information to some degree. To determine what the gloss of a word is, linguists have to look for hierarchical (non-linear) context information.

Therefore, theoretically incorporation of the gloss data should improve the translation systems. Specifically, I propose the following hypothesis:

(1) **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 boost up 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, the performance of the systems is improved significantly.

This chapter describes the experiments conducted to test the Gloss-helps hypothesis and the results attest the weak version. The rest of the chapter is organized as follows: Section 2 describes the constant parameter settings across all the experiments, section 3 tests the hypothesis in (1a), section 4 tests the hypothesis in (1b),and section 5 concludes the chapter.

2 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

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

The settings of the hyper-parameters do have effects on the performances of the trained models. A common practice to find the optimal settings of the hyper-parameters is to hold out a subset of the training dataset as the developing dataset, then test the models on the developing data to see what settings are optimal, then merge the developing dataset and training dataset as a new training set, and then train on this new training set using the found optimal hyper-parameters. However, finding the optimal settings of the hyper-parameters is not relevant to our research and causing unnecessary complications. So, the process of optimizing the settings of the hyper-parameters is not implemented, and I simply adopt OpenNMT's default settings. So, the employed settings of the hyper-parameters may be viewed as arbitrarily chosen. 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.

The data and the scripts will be accessible on GitHub[2], so that the results can be reproduced.

3 Gloss Representation Solely Does NOT Outperform Gaelic Sentences

This section tests the strong version of Gloss-helps hypothesis in (1a). 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).

3.1 Procedure of the Experiments

I 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 Gaelic 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.

(2) 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 above is:

(3) <"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).

(4) Definitions of datasets:
Let:

[2] https://github.com/lucien0410/Scottish_Gaelic

a. $\text{Index}_{\text{Train}}$, $\text{Index}_{\text{Validation}}$, and $\text{Index}_{\text{Test}}$ be sets of random indexes from 0 to 8,387.

b. $\text{Index}_{\text{Train}} \cap \text{Index}_{\text{Validation}} \cap \text{Index}_{\text{Test}} = \emptyset$

c. $|\text{Index}_{\text{Train}}| = 6{,}388$; $|\text{Index}_{\text{Validation}}| = 1{,}000$; $|\text{Index}_{\text{Test}}| = 1{,}000$.

The step above just randomly splits the indexes of the 3-tuples into three distinct sets: $\text{Index}_{\text{Train}}$, $\text{Index}_{\text{Validation}}$, and $\text{Index}_{\text{Test}}$. Based on the indexes, we generate the sets of samples. 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. For each pair, the first item is the source sequence, and the second item is the target sequence. The systems learns how to map the source sequence to the target sequence.

(5) 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.”>

(6) 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 (i.e. the model learns how to map the source sequence to the target sequence). Both training set and validation set are known information for the models[3]. Specifically, the neural net system learns how to maps gloss lines to English sentences from samples in (5a) and (5b), and another neural net system learns how to maps Gaelic sentences to English sentences from from samples in (6a) and (6b).

(7) Models:

a. $\text{Model}_{\text{GLOSStoEN}}$ = Model trained with $\text{GLOSStoEN}_{\text{Train}}$ in (5a) and $\text{GLOSStoEN}_{\text{Validation}}$ in (5b)

b. $\text{Model}_{\text{GDtoEN}}$ = Model trained with $\text{GDtoEN}_{\text{Train}}$ in (6a) and $\text{GDtoEN}_{\text{Validation}}$ in (6b)

[3]Technically speaking, the validation set is part of the training data in terms of machine learning. The presence of the validation set is a special requirement of neural net machine learning, which uses the validation set to evaluate the convergence of the training.

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).

(8) Predictions:

a. $\text{Predictions}_{\text{GLOSStoEN}}$ = A list of English sequences that $\text{Model}_{\text{GLOSStoEN}}$ maps to from the gloss sequences in (5c)

b. $\text{Predictions}_{\text{GDtoEN}}$ = A list of English sequences that $\text{Model}_{\text{GDtoEN}}$ maps to from the Gaelic sentences in (6c)

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 (bilingual evaluation understudy)[4] score metric (Papineni et al., 2002) of each prediction is calculated using the multi-bleu.perl[5] script, a public implementation of Moses (Koehn et al., 2007). The BLEU score calculation is an automatic evaluation of how similar two copora are. In the current experiments we are comparing the predicted target sequences with the gold standard. The BLEU score of 100 means the two copora are identical, and the BLEU score of 0 means the two copora are completely distinct from each other.

(9) Gold-Standard = English sentences in (5c) = English sentences in (6c)

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.

(10) 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.

3.2 Result

After ten rounds of repeated random sub-sampling validation, ten pairs of scores of the two models are generated, as shown in the following table. The average score of the $\text{Models}_{\text{GLOSStoEN}}$ is only sightly higher than the average score of the $\text{Models}_{\text{GDtoEN}}$. Also, after doing a paired T-test, the difference between the two types of models is not attested (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]There are other automatic machine translation evaluation algorithms available, such as translation edit rate (Snover et al., 2006) and Damerau-Levenshtein distance (Damerau, 1964; Levenshtein, 1966). BLEU is chosen for the current experiments because it is the most widely used evaluation algorithm, and the correlation between the BLUE score evaluation and human judgment evaluation is also well-acknowledged.

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

Round	Gaelic (Baseline)	GLOSS
0	17.29	18.39
1	16.42	18.00
2	15.29	16.02
3	15.97	20.22
4	17.79	19.02
5	16.73	15.53
6	17.11	18.00
7	16.37	20.08
8	15.93	15.82
9	16.99	15.93
Mean	16.59	17.70

Table 1: BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{GLOSStoEn}}$

3.3 Summary

The ultimate practical goal of the dissertation is to use glossing data to develop better machine translation systems. Here *better* means to be better than a baseline system, which is the machine translation system trained with Gaelic-to-English translation samples. The models in (7b) are the baseline systems, and their scores are in the Gaelic column of table (**??**). These are the target scores that we aims to outperform. The experiment above is the first attempt to improve that scores by using the *gloss treatment*, in which the Gaelic sentences are replaced with gloss lines. However, the result shows that this *gloss treatment* is not effective as the scores of the gloss models are not statistically higher than the baseline Gaelic-to-English models.

3.4 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 (3.2) that the gloss models are not better than the Gaelic models, it is concluded that glosses and natural languages are about the same in terms of representing meanings. The strong version of the Gloss-helps hypothesis does not hold.

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

4 Combining Gaelic words with Glosses

In the previous section, we attempt to build a system 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. This section reports various ways of combining the gloss line data and the Gaelic sentence data, and the

experiments and their results using these different treatments. Critically, a specific way of combining Gloss data and Gaelic date (termed as '*ParaPart*' treatment) boosts the performance significantly. The model trained with this specially arranged training data also significantly outperforms Google's Gaelic-to-English translation system.

In this section, I will first describe the most effective treatment, termed as '*ParaPart*' treatment, and the results, and then I will report other experiments attempting other relevant logical treatments.

4.1 The 'ParaPart' Treatment Outperforms Any Other Treatments and The Baseline Significantly

Round	Gaelic (Baseline)	ParaPart
0	17.29	32.64
1	16.42	32.28
2	15.29	29.94
3	15.97	31.18
4	17.79	32.83
5	16.73	31.11
6	17.11	32.19
7	16.37	33.52
8	15.93	30.93
9	16.99	34.35
Mean	16.59	32.10

Table 2: BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{ParaParttoEn}}$

(M_{GDToEn}=16.59, SD_{GDToEn}=0.74; M_{ParaPart}=32.10, SD_{ParaPart}=1.33,; t(9)=48.95, p=0.000)

4.2 Other Logical Treatments

4.2.1 Para

Round	Gaelic (Baseline)	Para
0	17.29	25.42
1	16.42	25.32
2	15.29	20.72
3	15.97	22.22
4	17.79	24.27
5	16.73	24.55
6	17.11	27.03
7	16.37	25.34
8	15.93	24.24
9	16.99	25.96
Mean	16.59	24.51

Table 3: BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{ParatoEn}}$

(M_{GDToEn}=16.59, SD_{GDToEn}=0.74; M_{Para}=24.51, SD_{Para}=1.84,; t(9)=17.50, p=0.000)

4.2.2 Interleaving Gaelic Words and Gloss Words

Round	Gaelic (Baseline)	interleavingGdGLOSS
0	17.29	13.67
1	16.42	12.49
2	15.29	11.01
3	15.97	12.33
4	17.79	12.56
5	16.73	12.13
6	17.11	11.55
7	16.37	12.78
8	15.93	12.43
9	16.99	11.65
Mean	16.59	12.26

Table 4: BLEU scores of $Model_{GDtoEN}$ and $Model_{interleavingGdGLOSStoEn}$

(M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{interleavingGdGLOSS}$=12.26, $SD_{interleavingGdGLOSS}$=0.74,; t(9)=-17.06, p=0.000)

4.2.3 Concating Gaelic Words and Gloss Words

Round	Gaelic (Baseline)	ConcatGLOSSGaelic
0	17.29	15.42
1	16.42	14.31
2	15.29	15.38
3	15.97	14.18
4	17.79	18.63
5	16.73	14.89
6	17.11	15.16
7	16.37	15.20
8	15.93	15.50
9	16.99	15.72
Mean	16.59	15.44

Table 5: BLEU scores of $Model_{GDtoEN}$Model and textsubscriptConcatGLOSS-GaelictoEn

(M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{ConcatGLOSSGaelic}$=15.44, $SD_{ConcatGLOSSGaelic}$=1.23,; t(9)=-3.64, p=0.010)

4.2.4 Replacing ambiguous Gaelic Words

(M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{ReplacingGaelic}$=9.24, $SD_{ReplacingGaelic}$=0.89,; t(9)=-21.03, p=0.000)

Round	Gaelic (Baseline)	HybridDefaultAsGaelic
0	17.29	9.44
1	16.42	9.07
2	15.29	7.69
3	15.97	9.12
4	17.79	9.08
5	16.73	10.45
6	17.11	8.62
7	16.37	10.00
8	15.93	10.52
9	16.99	8.46
Mean	16.59	9.24

Table 6: BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{HybridDefaultAsGaelictoEn}}$

4.2.5 Replacing ambiguous GLOSS Items

Round	Gaelic (Baseline)	HybridDefaultAsGLOSS
0	17.29	15.95
1	16.42	15.60
2	15.29	14.15
3	15.97	14.72
4	17.79	15.74
5	16.73	14.88
6	17.11	14.45
7	16.37	16.41
8	15.93	15.15
9	16.99	17.61
Mean	16.59	15.47

Table 7: BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{HybridDefaultAsGLOSS}}$

(M_{GDToEn}=16.59, SD_{GDToEn}=0.74; $M_{\text{ReplacingGLOSS}}$=15.47, $SD_{\text{ReplacingGLOSS}}$=1.03,; t(9)=-3.67, p=0.005)

4.2.6 Complete Tables

4.3 literature

XXXwhat about 4 **??** Linguistics-informed MT: (Sennrich and Haddow, 2016)

Multi-task Sequence to Sequence Learning: (Luong et al., 2015)
what is Multi-task learning: (Ruder, 2017)
add ccc to target seq: (Nadejde et al., 2017)
google zero shot: (Johnson et al., 2016)

References

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

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

Cho, Kyunghyun, 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*.

Damerau, Fred J (1964), "A technique for computer detection and correction of spelling errors." *Communications of the ACM*, 7, 171–176.

Johnson, Melvin, 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*.

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

Koehn, Philipp, 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*, 177–180, Association for Computational Linguistics.

Kotsiantis, Sotiris B, I Zaharakis, and P Pintelas (2007), "Supervised machine learning: A review of classification techniques." *Emerging artificial intelligence applications in computer engineering*, 160, 3–24.

Levenshtein, Vladimir I (1966), "Binary codes capable of correcting deletions, insertions, and reversals." In *Soviet physics doklady*, volume 10, 707–710.

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

Nadejde, Maria, 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*, 68–79.

Papineni, Kishore, 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*, 311–318, Association for Computational Linguistics.

Ruder, Sebastian (2017), "An overview of multi-task learning in deep neural networks." *CoRR*, abs/1706.05098, URL http://arxiv.org/abs/1706.05098.

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

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006), "A study of translation edit rate with targeted human annotation." In *In Proceedings of Association for Machine Translation in the Americas*, 223–231.