Chapter 1

Introduction

Key information to be included:

1. outline/organization of the dissertation

2. Arguments to be made in the dissertation:

    (a) To understand language, NLP and Linguistics should work together.

    (b) Gloss is the right ‘lingua franca’ for the two fields.

    (c) Linguistics helps NLP.

    (d) NLP helps linguistics.

Chapter 2
What are glosses? Why are them golden representations of meanings?

2.1 Key Points of The Chapter

1. Target Audience: CS people
2. main points:
   (a) summary of the Leipzig Glossing Rules (Bickel et al., 2008).
   (b) glossing is the initial processing of the data guided by some specific syntax theory.
   (c) Glosses contain morphology information (with examples); glosses disambiguate homographs (with many examples); gloss also provide some parsing information because some glosses are determined by structural/constituency context (with examples).

2.2 Introduction: What are Glosses

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.’

2.3 The Golden Properties of Glosses

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. Theoretically, this claim can be tested empirically. Imagine there is a set of special gold meta-linguistic semantic representations, which has the following property: each concept is mapped to one and one representation and each representation is mapped to one and one concept. Given, theses gold representations, it is expected that each gold representation will map to more natural language words than gloss items, and each natural language word will map to more gold representations than gloss items. However, in practice, this is an impossible experiment to conduct, because there are no such gold representation[1]. Glosses provide this one-to-one mapping. Second, the gloss data provides hierarchical (non-linear) syntactic parsing information to some degree.

2.3.1 Glosses cluster different words with the same meanings

Gloss collapses words

2.3.2 Glosses disambiguates ambiguous words

Critically glosses provide 'redundant' information. Here 'redundant' means that

2.3.3 Glosses are sensitive to hierarchical structure of natural language sentences

Critically glosses provide 'redundant' information. Here 'redundant' means that

2.4 What is a Gloss Line

A gloss line is an artificial sentence using the purified 'gloss words'.

[1] It would solve the puzzle of semantics if one should be able to build the set of special gold meta-linguistic semantic representations.

Chapter 3

Description of the Scottish Gaelic Interlinear Glossed Text Corpus

1. Original goal of the corpus: a bank of examples of specific syntactic patterns for syntacticians.

2. Description of UA Celtic Group's Scottish Gaelic documentation project

3. Collection of interlinear glossed text data used in syntax paper/dissertations AND language documentation

4. Auto-glosser and literature on auto-glosser

Chapter 4

A Gental Introduction of Machine Learning and Machine Translation

1. General review of supervised Machine learning (Kotsiantis et al., 2007): The goal is provide a high level of understanding of what machine learning is. Machine Learning is to learn from EXAMPLES/SAMPLES. For example, to define the meaning of 'dog', instead of giving all the definable features of 'dog', we feed the machine with as many as possible of information of entities of dogs that we have access. Montague Semantics is actually a variant of ML, within which 'dog' is defined as 'the set of all the dog that exists in the current world.'. As such, Montague Semantics is Machine Learning, because instead of defining 'dog' with certain arbitrary rules (+/- FEATURE), it says 'all the dog entities in the current world'. Definition by samples/examples not by rules.

2. Literature on machine translation: from statistical machine translation (Koehn, 2009) to neural machine translation (Cho et al., 2014b,a; Bahdanau et al., 2014; Koehn, 2017). (Target audience: linguists)

Chapter 5

Building Translation Systems using Interlinear Glossed Text

(

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)

)

5.1 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 choice 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[1]. 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:

(2) **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).

[1] Theoretically, this claim can be tested empircally. Imagine there is a set of specical gold meta-linguistic semantic representations, which has the following property: each concept is mapped to one and one representation and each representation is mapped to one and one concept. Given, thses gold repretations, it is expected that each gold repretation will map to more natural language words than gloss items, and each natural language word will map to more gold representations than gloss items. However, in practice, this is an impossible experimet to conduct, becuase there are no such gold prepresentations. It would solve the pozzle of semantics if one should be able to build the set of specical gold meta-linguistic semantic representations.

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 bochoiceost 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 5.2 describes the constant parameter settings across all the experiments, section 5.3 tests the hypothesis in (2a), section 5.4 tests the hypothesis in (2b),and section 5 concludes the chapter.

5.2 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[2] 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

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

- Size of mini batches: 64

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, given that finding the optimal settings of the hyper-parameters is not relevant to our research and causing unnecessary complications, the process of optimizing the settings of the hyper-parameters is not implemented, and I simply adopt OpenNMT's default settings. The employed settings of the hyper-parameters should be viewed as arbitrarily chosen, and there are room to tune the models for better performance. 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[3], so that the results can be reproduced.

5.3 Gloss Representation Solely Does NOT Outperform Gaelic Sentences

This section tests the strong version of Gloss-helps hypothesis in (2a). 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

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

no significance difference between the two types of data (i.e. GLOSS → English and Gaelic → English).

5.3.1 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.

(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.’

The 3-tuple representation of the above example is:

(4) <“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].

(5) Definitions of datasets:

Let:

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

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

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.

(6) 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.”>

(7) 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[5]. Specifically, the neural net system learns how to maps gloss lines to English sentences from samples in (6a) and (6b), and another neural net system learns how to maps Gaelic sentences to English sentences from from samples in (7a) and (7b).

(8) Models:

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

b. $\text{Model}_{\text{GDtoEN}}$ = Model trained with $\text{GDtoEN}_{\text{Train}}$ in (7a) and $\text{GDtoEN}_{\text{Validation}}$ in (7b)

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).

(9) Predictions:

a. $\text{Predictions}_{\text{GLOSStoEN}}$ = A list of English sequences that $\text{Model}_{\text{GLOSStoEN}}$ maps to from the gloss sequences in (6c)

b. $\text{Predictions}_{\text{GDtoEN}}$ = A list of English sequences that $\text{Model}_{\text{GDtoEN}}$ maps to from the Gaelic sentences in (7c)

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)[6] score metric

[5] 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.

[6] There are other automatic machine translation evaluation algorithms available, such as translation edit rate (Snover et al., 2006) and Damerau-Levenshtein edit distance (Damerau, 1964; Lev-

(Papineni et al., 2002) of each prediction is calculated using the multi-bleu.perl[7] 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.

(10) Gold-Standard = English sentences in (6c) = English sentences in (7c)

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.

(11) 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.

5.3.2 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

enshtein, 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.

[7] 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 5.1. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{GLOSStoEn}}$

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, $\text{SD}_{\text{GDToEn}}$=0.74; $\text{M}_{\text{GLOSStoEN}}$=17.70, $\text{SD}_{\text{GLOSStoEN}}$=1.78; t(9)=1.97, p=0.080)

5.3.3 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 (8b) are the baseline systems, and their scores are in the Gaelic column of table (5.1). 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.

5.3.4 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 (5.3.2) 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.

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.

5.4 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. In other words, the questions now are:

(12) a. Does adding the gloss data into the Gaelic data will improve the translation system?

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

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 '*Parallel-Partial*' 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 '*Parallel-Partial*' treatment, and the results, and then I will report the experiments done with other relevant logical treatments (i.e. other ways of combining glossing data and Gaelic data).

5.4.1 The 'Parallel-Partial' Treatment Outperforms Any Other Treatments and the Baseline Significantly

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 (7a) and (7b)), and the data of the gloss treatment are pairs of a gloss line and a English sentences (see (6a) and (6b). These two groups of data are combined in a parallel manner in the current treatment. Now the size of training set and validation set is doubled. 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 12,776 samples in the training set and 2,000 samples in the validation set. This is the *parallel* part of the treatment.

Additionally, I 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:

(13) 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:

(14) 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”>

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

(15) 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”>

Results of the Parallel-Partial Treatment With the training and validation data ready, now we can train models and evaluate them. Critically, 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. As shown in the following table, the Parallel-Partial treatment has a tremendous effect in improving the baseline system.

The first and the second columns are BLUE scores of the baseline systems and the systems with the Parallel-Partial treatment respectively. The latter is significantly better than the former ($\mathrm{M}_{\mathrm{GDToEn}}$=16.59, $\mathrm{SD}_{\mathrm{GDToEn}}$=0.74; $\mathrm{M}_{\mathrm{ParaPart}}$=32.10,

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 5.2. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{ParaParttoEn}}$

SD$_{\text{ParaPart}}$=1.33; t(9)=48.95, $p<0.01$). The comparison of the average BLUE scores of the groups of systems shows that the Parallel-Partial treatment improves the performance of the baseline system for 93 percent.

5.4.2 Other Possible Treatments

This section reports 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.

The Parallel Treatment

Method of the Parallel Treatment The Parallel treatment is using the parallel part of the Parallel-Partial treatment. With this treatment, a chunk of interlinear glossed text is split into two pairs. For example, the chunk of interlinear glossed text in (16) becomes two samples in (17):

(16) 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.’

(17) 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”>

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 5.3. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{ParatoEn}}$

Results of the Parallel Treatment The table in (5.3) compares the performances of this treatment and the baseline. Critically, the Parallel treatment is effective in improving the baseline systems (M_{GDToEn}=16.59, $\text{SD}_{\text{GDToEn}}$=0.74; M_{Para}=24.51, SD_{Para}=1.84; t(9)=17.50, $p < 0.01$). However, the best treatment (i.e. the Parallel-Partial treatment) is still far better than this Parallel treatment (M_{Para}=24.51, SD_{Para}=1.84; $\text{M}_{\text{ParaPart}}$=32.10, $\text{SD}_{\text{ParaPart}}$=1.33; t(9)=18.73, $p < 0.01$).

Interleaving Gaelic Words and Gloss Items And Concating them

Round	Gaelic (Baseline)	Para	ParaPart
0	17.29	25.42	32.64
1	16.42	25.32	32.28
2	15.29	20.72	29.94
3	15.97	22.22	31.18
4	17.79	24.27	32.83
5	16.73	24.55	31.11
6	17.11	27.03	32.19
7	16.37	25.34	33.52
8	15.93	24.24	30.93
9	16.99	25.96	34.35
Mean	16.59	24.51	32.10

Table 5.4. BLEU scores of Model$_{\text{GDtoEN}}$, Model$_{\text{ParatoEN}}$ and Model$_{\text{ParaParttoEN}}$

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. Consider the following example:

(18) 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">

Given the chuck of interlinear glossed text data in (18a), the Interleaving treatment generates the sample in (18b). The results are given in the following table. It turns out this treatment has a significant negative effect (M$_{\text{GDToEn}}$=16.59, SD$_{\text{GDToEn}}$=0.74; M$_{\text{interleavingGdGLOSS}}$=12.26, SD$_{\text{interleavingGdGLOSS}}$=0.74,; t(9)=-17.06, p=0.000). This is not the right way of incorporating gloss line data.

Method of Concating Gaelic Words and Gloss Words A quick and close amendment of the Interleaving approach is to concatenate the aligned Gaelic word

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 5.5. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{interleavingGdGLOSStoEn}}$

and gloss item as a single token. Given the same chunk of interlinear glossed text data, this treatment generates the following sample:

(19) 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">

Results of Concating Gaelic Words and Gloss Words The performances of this treatment is given in the following table.
The result shows that this treatment hurts the baseline systems badly instead of improving them (M$_{\text{GDToEn}}$=16.59, SD$_{\text{GDToEn}}$=0.74; M$_{\text{ConcatGLOSSGaelic}}$=15.44, SD$_{\text{ConcatGLOSSGaelic}}$=1.23,; t(9)=-3.64, p=0.010).

Hybrid: Gaelic or Gloss

Method of Hybrid The Hybrid treatment aims to reduce the potential lexical ambiguity. A Gaelic word may maps to multiple gloss, and a glosses may maps to

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 5.6. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{ConcatGLOSSGaelictoEn}}$

multiple Gaelic words. Let's assume a toy chunk of interlinear glossed text data (a one-word sentence):

(20) Gaelic_word
Gloss_item
English translation

Now we aim to build a single sample that is either <Gaelic_word, English translation > or <Gloss_item, English translation >. The criterion is which one, the Gaelic word or the gloss item, is less ambiguous. The less ambiguous one is the winner. For example, if the Gaelic word is potentially mapped to 10 glosses and if the gloss item is potentially mapped 2 Gaelic word, then <Gloss_item, English translation> is chosen; other the other hand if the ambiguity situation is reverted, then <Gaelic_word, English translation > is chosen. However, when the situation is tight (i.e. both the Gaelic word and gloss item are equally ambiguous), a default setting is needed to be chosen. The choices of the default setting split this single treatment into two treatments: default as Gaelic or default as gloss.

Result of Hybrid When the default setting is the Gaelic word, the performances are significantly worse than than the baseline systems, as shown in table (5.7).

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 5.7. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{HybridDefaultAsGaelictoEn}}$

(M_{GDToEn}=16.59, $\text{SD}_{\text{GDToEn}}$=0.74; $\text{M}_{\text{ReplacingGaelic}}$=9.24, $\text{SD}_{\text{ReplacingGaelic}}$=0.89,; t(9)=-21.03, $p < 0.01$). When the default setting is the Gaelic word, the performances are sightly worse than than the baseline systems, as shown in table (5.8). (M_{GDToEn}=16.59, $\text{SD}_{\text{GDToEn}}$=0.74; $\text{M}_{\text{ReplacingGaelic}}$=15.47, $\text{SD}_{\text{ReplacingGaelic}}$=1.03,; t(9)=-3.67, $p < 0.01$).

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 5.8. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{HybridDefaultAsGLOSS}}$

5.5 Summary and Conclusion

The chapter reports machine translation experiments that aims to find how the gloss line information can improve the performance of the baseline Gaelic-to-English translation systems. It is found that the Parallel-Partial is highly effective. The complete BLEU scores of various treatments are given in the following table. The aim of chapter is to report and document how the experiments are done and what the results are. This is merely reporting the linguist and non-linguistic facts. The implications and relevant works in the literature will be discussed in the next chapter.

Round	Baseline	GLOSS	ParaPart	Para	Interleaving	Concat	HybrGaelic	HybrGLOSS
0	17.29	18.39	32.64	25.42	13.67	15.42	9.44	15.95
1	16.42	18.00	32.28	25.32	12.49	14.31	9.07	15.60
2	15.29	16.02	29.94	20.72	11.01	15.38	7.69	14.15
3	15.97	20.22	31.18	22.22	12.33	14.18	9.12	14.72
4	17.79	19.02	32.83	24.27	12.56	18.63	9.08	15.74
5	16.73	15.53	31.11	24.55	12.13	14.89	10.45	14.88
6	17.11	18.00	32.19	27.03	11.55	15.16	8.62	14.45
7	16.37	20.08	33.52	25.34	12.78	15.20	10.00	16.41
8	15.93	15.82	30.93	24.24	12.43	15.50	10.52	15.15
9	16.99	15.93	34.35	25.96	11.65	15.72	8.46	17.61
Mean	16.59	17.70	32.10	24.51	12.26	15.44	9.24	15.47

Table 5.9. BLEU scores of the treatments

(Hi Mike: The current chapter reports the what are done and how (i.e. the fact); the next chapter I will discuss the why questions, and discuss similar works in the literature.

)

Bibliography

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

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

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.

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 (2009), *Statistical machine translation*. Cambridge University Press.

Koehn, Philipp (2017), "Neural machine translation." *CoRR*, abs/1709.07809, URL http://arxiv.org/abs/1709.07809.

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.

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.

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.