Develong Linguistically Imformed Neural Net Machine Translation Systems

by
Yuan-Lu Chen

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF LINGUISTICS
In Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy
In the Graduate College
The University of Arizona

April 4, 2018

Get the official approval page
from the Graduate College
before your final defense.

Statement by Author

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Signed: _____

Chapter 1

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

1.1 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).

1.2 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.'

(summarize and exemplify the Leipzig Glossing Rules)

1.3 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.

1.3.1 Glosses Cluster Different Words with the Same Meanings

Gloss collapses words with different forms with the same meanings into a single gloss. In natural languages, the morphology of a word (i.e. the form of a word) may be sensitive to the phonological environments and changing into different forms. Consider the following English example:

(2) John ate **an** apple.
John eat.past **Det** apple

(3) John ate **a** banana.
John eat.past **Det** banana

In the above example, *an* and *a* have the identical meaning[2].

[1] It would solve the puzzle of semantics if one should be able to build the set of special golden meta-linguistic semantic representations, and the mappings between the golden representations to natural languages.

[2] Semantically, *an* and *a* are existential quantifiers, which declare that a member of a set exists in the world. In formal semantics, *an* and *a* may be defined as follows: $\exists\lambda P[P(x)]$. In the current example, *apple* and *banana* will instantiate P in the formula, and the meanings will be 'an apple exists' and 'a banana exists'. Kratzer and Heim (1998) would be a nice introduction for interested readers to see how linguists, specifically semanticians, define, decompose, and compose meanings of languages formally.

1.3.2 Glosses disambiguates ambiguous words

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

1.3.3 Glosses are sensitive to hierarchical structure of natural language sentences

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

1.4 What is a Gloss Line

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

1.5 Conclusion

Chapter 2

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)

)

2.1 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. Glosses provide this one-to-one mapping. Second, the gloss data provides hierarchical (non-linear) syntactic parsing information. To determine what the gloss of a word is, linguists have to look for hierarchical (non-linear) context information. See chapter 1 for the discussion on the golden properties of glosses.

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

(4) **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 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 2.2 describes the constant parameter settings across all the experiments, section 2.3 tests the hypothesis in (4a), section **??** tests the hypothesis in (4b),and section 5 concludes the chapter.

2.2 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

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

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

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[2], so that the results can be reproduced.

2.3 Gloss Representation Solely Does NOT Outperform Gaelic Sentences

This section tests the strong version of Gloss-helps hypothesis in (4a). 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 $\rightarrow$ English and Gaelic $\rightarrow$ English).

2.3.1 Procedure of the Experiments

I use repeated random sub-sampling validation to compare the performances of the two type of models.

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

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.

(5) 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:

(6) <"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].

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

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.

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

(8) 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.">

(9) 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[4]. Specifically, the neural net system learns how to maps gloss lines to English sentences from samples in (8a) and (8b), and another neural net system learns how to maps Gaelic sentences to English sentences from from samples in (9a) and (9b).

[4]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.

(10) Models:

a. $\text{Model}_{\text{GLOSStoEN}}$ = Model trained with $\text{GLOSStoEN}_{\text{Train}}$ in (8a) and $\text{GLOSStoEN}_{\text{Validation}}$ in (8b)

b. $\text{Model}_{\text{GDtoEN}}$ = Model trained with $\text{GDtoEN}_{\text{Train}}$ in (9a) and $\text{GDtoEN}_{\text{Validation}}$ in (9b)

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).

(11) Predictions:

a. $\text{Predictions}_{\text{GLOSStoEN}}$ = A list of English sequences that $\text{Model}_{\text{GLOSStoEN}}$ maps to from the gloss sequences in (8c)

b. $\text{Predictions}_{\text{GDtoEN}}$ = A list of English sequences that $\text{Model}_{\text{GDtoEN}}$ maps to from the Gaelic sentences in (9c)

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)[5] score metric (Papineni et al., 2002) of each prediction is calculated using the **multi-bleu.perl**[6] 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.

[5]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; 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.

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

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.

(12) Gold-Standard = English sentences in (8c) = English sentences in (9c)

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.

(13) 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.

2.3.2 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)

2.3.3 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,

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 2.1. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{GLOSStoEn}}$

which is the machine translation system trained with Gaelic-to-English translation samples. The models in (10b) are the baseline systems, and their scores are in the Gaelic column of table (2.1). 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.

2.3.4 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 (2.3.2) 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.

There are several remarks that need to make for the current result. First, the result falsifies the point of view about glosses in chapter (1) 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, as the mean BLEU score of the gloss treatment is slightly higher than that of the baseline systems. 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 of 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. The auto-glosser described in chapter (1) may convert the Gaelic sentence to a gloss line, but the conversion is not perfect at all. Given this, even if the gloss treatment should work, it is not practical unless we may convert Gaelic sentence to gloss line perfectly.

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.

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, 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.

Kratzer, Angelika and Irene Heim (1998), *Semantics in generative grammar*. Blackwell Oxford.

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.