# Developing Linguistically Informed Neural 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

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: ________________________________

# Dedication

For Eva, the best linguist in my world.

# Acknowledgments

acknowledgment! Many people help me and keep me company in my this journey. Thank you.

# Contents

List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . **8**

List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . **9**

Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . **10**

Chapter 1. Introduction . . . . . . . . . . . . . . . . . . . . . . . . **11**

Chapter 2. What are glosses? Why are they golden representations of meanings? . . . . . . . . . . . . . . . . . . . . . . . . . **12**

- 2.1. Introduction: What are Glosses? . . . . . . . . . . . . . . . . . . 12
- 2.2. The Golden Properties of Glosses . . . . . . . . . . . . . . . . . 14
  - 2.2.1. Glosses Cluster Different Words with the Same Meanings (Synonyms) Into a Single representation . . . . . . . . . . . . . . 15
  - 2.2.2. Glosses Distinguish Homographs' Different Meanings . . . . . 18
  - 2.2.3. Glosses are Sensitive to Hierarchical Structures in Natural Language Sentences . . . . . . . . . . . . . . . . . . . . . . . . 18
- 2.3. Conclusion: What Is a Gloss Line and Why Do They Matter? . . . . 22

Chapter 3. Description of the Scottish Gaelic Interlinear Glossed Text Corpus . . . . . . . . . . . . . . . . . . . . . . . . . . . . **23**

Chapter 4. A Gental Introduction of Machine Learning and Machine Translation . . . . . . . . . . . . . . . . . . . . . . . . . . **24**

- 4.1. What is Machine Learning? . . . . . . . . . . . . . . . . . . . . 24
- 4.2. History of Machine Translation . . . . . . . . . . . . . . . . . . 25
- 4.3. Statistical Machine Translation . . . . . . . . . . . . . . . . . . 25
- 4.4. Neural Net Machine Translation . . . . . . . . . . . . . . . . . . 25
  - 4.4.1. What is neural nets? Why can neural net learn? . . . . . . . . 25

Chapter 5. Building Translation Systems using Interlinear Glossed Text: First Attempt . . . . . . . . . . . . . . . . . . . . . . . . **26**

- 5.1. Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
- 5.2. Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
- 5.3. Technical Settings of the Machine Translation Experiments . . . . . . 28
- 5.4. Gloss Representation Solely Does NOT Outperform Gaelic Sentences 30
  - 5.4.1. Procedure of the Experiments . . . . . . . . . . . . . . . . . . 30
  - 5.4.2. Result . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
  - 5.4.3. Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

Contents—*Continued*

5.5. Discussion and Conclusion . . . 37

Chapter 6. Combining Gaelic Words with Glosses . . . **39**

6.1. Introduction . . . 39

6.1.1. The Underlying Heuristics . . . 40

6.2. The ‘Parallel-Partial’ Treatment Outperforms Any Other Treatments and the Baseline Significantly . . . 40

6.2.1. Related work . . . 40

6.2.2. Data Preprocessing Using the Parallel-Partial Treatment . . . 41

6.2.3. Results of the Parallel-Partial Treatment . . . 42

6.3. Other Possible Treatments . . . 43

6.3.1. The Parallel Treatment . . . 43

6.3.2. Interleaving Gaelic Words and Gloss Items And Concating them 45

6.3.3. Hybrid: Gaelic or Gloss . . . 48

6.4. Summary and Conclusion . . . 50

Chapter 7. Interlinear Glossed Text Data In Other Languages **52**

Chapter 8. Implications on Theoretical Linguistics and Natural Language Processing . . . **53**

Chapter 9. Conclusion . . . **54**

Bibliography . . . **55**

# List of Figures

Figure 5.1. The BLEU score is based on n-gram matches with the reference translation (Koehn, 2009, p. 226-227) . . . . . . . . . . . . . . . . . . . . . 35

# List of Tables

Table 5.1. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{GLOSStoEn}}$ . . . . . . . . . 37

Table 6.1. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{ParaParttoEn}}$ . . . . . . . . 43

Table 6.2. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{ParatoEn}}$ . . . . . . . . . . 44

Table 6.3. BLEU scores of Model$_{\text{GDtoEN}}$, Model$_{\text{ParatoEN}}$ and Model$_{\text{ParaParttoEN}}$ 45

Table 6.4. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{interleavingGdGLOSStoEn}}$ . . . 46

Table 6.5. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{ConcatGLOSSGaelictoEn}}$ . . . 47

Table 6.6. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{HybridDefaultAsGaelictoEn}}$ . . 49

Table 6.7. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{HybridDefaultAsGLOSS}}$ . . . . 50

Table 6.8. BLEU scores of the all treatments . . . . . . . . . . . . . . . . . . 51

Table 7.1. BLEU scores of other languages . . . . . . . . . . . . . . . . . . . 52

# Abstract

Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper.

# Developing Linguistically Informed Neural Machine Translation Systems

Yuan-Lu Chen, Ph.D.
The University of Arizona, 2018

Director: Mike Hammond

Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper. Here is my abstract. It is a summary of the paper.

# 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 they golden representations of meanings?

## 2.1 Introduction: What are Glosses?

Interlinear Glossed Text is widely used in linguistic studies. The following is an example of Interlinear Glossed Text.

(1) Indonesian (Sneddon et al., 2012, p. 237)

Mereka di Jakarta sekarang. (*sentence of interest*)
they in Jakarta now (*gloss line: word-by-word gloss translation*)

'The are in Jakarta now.' (*English translation*)

A chunk of an interlinear glossed text has three lines. The first line is the sentence of interest. The second line is the gloss line, which is a word-by-word translation of the first line. And the third line a free English translation of the first line.

The conventional way to show the word-by-word translation from the first line to the gloss line is to use vertical alignment. In (1), '*Mereka*' is glossed as '*they*', '*di*' is glossed as '*in*', '*Jakarta*' is glossed as '*Jakarta*', and '*sekarang*' is glossed as '*now*'. These pairs are vertically aligned.

The gloss line also provides morphological information. Consider the following example:

(2) French

aux chevaux
to.ART.PL horse.PL

'to the horses'

The morphemes of a single word is linked by a '.'. The French word 'aux' is actually a combination of three separate morphemes[1]: '*to*', '*ART*', and '*PL*' and '*chevaux*' is decomposed into '*horse*' and '*PL*'. Bickel et al. (2008) compile a set of widely used conventions of IGT called the Leipzig Glossing Rules. Note that they are just guidelines of the formats of Interlinear Glossed Texts, so that Interlinear Glossed Texts can be more standardized.

The underlying intuition of Interlinear Glossed Text is that it provides an access to look into the subparts of a sentence. We may imagine the situation without the gloss line; then all we have is just the sentence and the English translation of that sentence. This will make it really hard to discuss the internal structure of the sentence. On the other hand, with the presence of the gloss line, with which each word is glossed and annotated, we then have a meta-representation in hand to discuss the grammatical properties of the sentence of interest.

An important note of the gloss line is that it is NOT raw linguistic data, and it is already processed. A linguist has already committed to some theory or some analysis on the sentence of interest when he or she transcribes the sentence into a gloss line, even if he or she tries to be as neutral as possible. As such, the question of what the gloss of a word is not trivial at all. Actually, sometimes a whole linguistic paper or thesis is to discuss and argue what the right gloss for a word is.

(3) Mandarin Chinese

| | | |
|---|---|---|
| Zhangsan | **hen** | gao |
| Zhangsan | HEN | tall |

'Zhangsan is tall.'

For example, Grano (2008), Chen (2010), and Liu (2010) discuss the nature of the Mandarin Chinese word '*hen*' in the above example and what the right gloss should

[1] A morpheme is a smallest unit of meaning. For example, '*boys*' has two morphemes in it: '*boy*' and '*-s*', where '*-s*' is a plural marker. Sometimes, the morpheme boundary is not visible. For example '*went*' is composed of '*go*' and '*-ed*'.

be '*hen*'. In cases like this, how one glosses a word is not trivial at all, but determining what the gloss of word is requires a set of evidence and arguments.

## 2.2 The Golden Properties of Glosses

A system of meaning representations is decomposed of three components: a) meanings, b) representations, and c) a mapping between meanings and representations. 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. On the other hand, gloss provides a mapping that is close to this ideal one-to-one mapping. Thus gloss should a better representation in term of representing meanings.

Theoretically, the claim that gloss representation is closer to the ideal one-to-one mapping than natural language representation is can be tested empirically. Let's imagine a set of special golden 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. With this imaginary golden semantic representation system, we may now compare Gaelic words and glosses. First, it is expected that each golden representation token will map to more natural language words than gloss items do.

(4) a. $golden_i \rightarrow \{Gaelic\_word_1, Gaelic\_word_2, \ldots\}_{golden_i}$

b. $golden_i \rightarrow \{gloss_1, gloss_2, \ldots\}_{golden_i}$

c. $|\{Gaelic\_word_1, Gaelic\_word_2, \ldots\}_{gold_i}| \geq |\{gloss_1, gloss_2, \ldots\}_{gold_i}|$

(4a) and (4b) represent a singe golden token may maps to multiple Gaelic words and glosses respectively. If we compare the size of them, it is expected that the set of Gaelic words is bigger than that of glosses, meaning that Gaelic words are more likely to be homographs than glosses are.

For the other direction, we may determine which one, Gaelic words or glosses, is more likely to be ambiguous.

(5) a. $Gaelic\_word_i \rightarrow \{golden_1, golden_2, \ldots\}_{Gaelic\_word_i}$

b. $gloss_i \rightarrow \{golden_1, golden_2, \ldots\}_{gloss_i}$

c. $|\{golden_1, golden_2, \ldots\}_{Gaelic\_word_i}| \geq |\{golden_1, golden_2, \ldots\}_{gloss_i}|$

(5a) and (5b) show the mappings from a Gaelic word to different concepts and the mappings from a gloss to different concepts respectively. (5c) is the expectation that Gaelic words are more likely to be ambiguous than glosses are.

To run statistical experiments to confirm the truth of (4c) and (5c) is the way to empirically support the claim that glosses closer to the golden representations than Gaelic words are. However, in reality, this is an impossible experiment to conduct, because there are no such golden representation[2]. In spite of the impossibility of conducting statistical experiments, we may still use some examples to show the intuition that glosses are better representations than natural languages are. The following sections describes how glosses cluster words with different forms but with the same meaning, and how glosses represent words with same form but with different meanings with different representations.

### 2.2.1 Glosses Cluster Different Words with the Same Meanings (Synonyms) Into a Single representation

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 the indefinite article in the English examples:

[2] 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.

| | | | | |
|---|---|---|---|---|
| (6) | John | ate | **an** | apple. |
| | John | eat.past | **ART** | apple |

| | | | | |
|---|---|---|---|---|
| (7) | John | ate | **a** | banana. |
| | John | eat.past | **ART** | banana |

In the above example, *an* and *a* have the identical meaning[3]. In English, the same concept is realized as two representations, *a* or *an*, while in the gloss representation the one concept is neatly represented as *ART*.

Critically, synonyms like the English *a* and *an* commonly occur in many other natural languages if not in all languages. The definite article in the language of interest, Scottish Gaelic, is another example to show the noisiness of natural language representations. Consider the definite article in the following Gaelic 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) | **am** | fear | mòr |
| | **ART** | man | big |

'a big man' (Lamb, 2001, p. 31)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (10) | 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)

| | | | |
|---|---|---|---|
| (11) | doras | **na** | sgoile(adh) |
| | door-N | **ART** | school-G |

'the door of the school' (Lamb, 2001, p. 29)

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (12) | a | chuir | air | dòigh | **nan** | àiridhean | a-muigh | a | rubh' | Eubhal | agus | an |
| | to | put-INF | on | order | **ART** | sheilings | out-LOC | to | point | Eaval | and | ART |

[3] 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.

oidhche seo
night this
‘the girls big house’ (Lamb, 2001, p. 100)

(13) 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 the 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. Again, the gloss notation nicely clusters them together as *ART*.

In Mandarin Chinese, similar patterns are found. Consider classifiers in the following examples:

(14) Yani mai-le {**pi**/***tou**} ma , Lulu mai-le {***pi**/**tou**} zhu.
Yani buy-PRF CL/CL horse , Lulu buy-PRF CL/CL pig
‘Yani bought a horse and Lulu bought a pig.’ (Zhang, 2013, p. 136)

In Zhang (2013), the classifier like *pi* and *tou* is a type of *indivual classifier* which co-occurs with countable nouns, like *ma*, ‘horse’, and *zhu*, ‘pig’, and this type of classifier is the head of *UNIT Phrase*. *Pi* and *tou* actually have the same semantics and the syntactic function; however, they are realized in different forms, specifically the form of which has to agree with the noun following it (i.e. *pi* goes with *ma*; *tou* goes with *zhu*). Here the gloss, *CL*, unifies the two forms of the same meaning.

Gloss collapses synonyms in natural languages. 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.

### 2.2.2 Glosses Distinguish Homographs' Different Meanings

In natural languages, there are cases when a single form denotes distinct concepts. Words with this property are termed as homographs. Consider the word *for* in following English examples:

(15) a. I intended **for** Jenny to be present.

b. **For** Jenny, I intended to be present. (Adger, 2003, p.306-307)

*For* in (15a) and (15b) has the same form but different meanings. Specifically, *for* in (15a) is a complementizer with its part of speech being *C*, and it heads the non-finite clause *Jenn to be present*; on the other hand *for* (15b) is a preposition, which takes a Determiner Phrase, *Jenny*, as its benefactive argument.

The Scottish Gaelic word *a'* in the following examples also has different meanings.

(16)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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)

(17)

| | | | | | | |
|---|---|---|---|---|---|---|
| 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 (16) is a progressive aspect marker while in (17) 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.

### 2.2.3 Glosses are Sensitive to Hierarchical Structures in Natural Language Sentences

Before I introduce how gloss information is linked to hierarchical structures, it is necessary to emphasize the importance of hierarchical structures in natural languages.

In this section, I will first review some linguistic arguments for why and how semantics and syntax of languages[4] are all about hierarchical structures instead of linear word orders. Then I will link gloss to hierarchical structures.

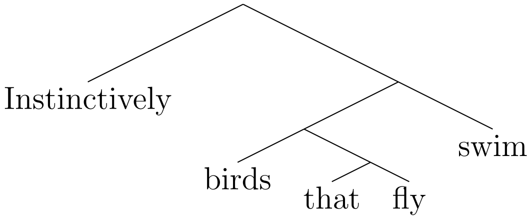
It is well-argued in linguistics that the syntax and semantics of natural languages are determined by hierarchical structures instead of linear orders of words, and essentially it is the sensitivity of hierarchical structures that distinguishes human natural languages from other animal communications (Berwick and Chomsky, 2015).

Semantics is determined by hierarchical structures instead of linear orders. Berwick and Chomsky (2015, p. 117) use the following simple example to demonstrate this property of natural languages:

(18) Instinctively birds that fly swim.

In the example above, *instinctively* is linearly close to *fly* than *swim*; however, it unambiguously modifies *swim* instead of *fly*. The reason for this is the hierarchical structures (Berwick and Chomsky, 2015, p. 117):

(19)



In (19) it is shown that *fly* is more embedded than *swim*, and thus it is hierarchically further away from *instinctively*. So, *instinctively* can only modify *swim* instead of *fly*.

Syntax is also all about hierarchical structures. Consider the following sentence:

(20)
a. Birds that can$_1$ fly can$_2$ swim.

b. *Can$_1$ birds that fly can$_2$ swim?

[4] When it turns to the sound aspect of languages, Phonetics is more about linear order, but Phonology is still sensitive to hierarchical structures just like syntax and semantics.

c. Can$_2$ birds that can$_1$ fly swim?

(20a) is a declarative sentence. To derive an interrogative sentence from it, the auxiliary needs to be moved; however only *can$_2$* can be moved but not *can$_1$* even *can$_1$* is linearly close to the sentence initial position. Again, it is all because of the hierarchical structures. *Can$_2$* is in the matrix clause while *can$_1$* is in the embedded relative clause.

Glosses, on the other hand, are sensitive to the internal hierarchical structures or constituency of sentences. They provide more clues of the internal hierarchical structures or constituency of sentences than natural language words. Consider the following examples, modified from (15):

(21) For as *complementizer* (glossed as *complementizer*)

a. I intended **for** [Jenny] to be present.

b. I intended **for** [the girl] to be present.

c. I intended **for** [the little girl] to be present.

d. I intended **for** [the little girl who wants to eat some ice scream] to be present.

(22) For as *preposition* (glossed as *preposition*)

a. **For** [Jenny], I intended to be present.

b. **For** [the girl], I intended to be present.

c. **For** [the little girl], I intended to be present.

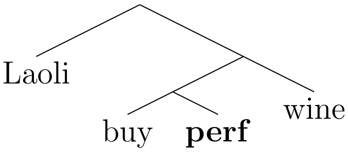
d. **For** [the little girl who wants to eat some ice scream], I intended to be present.

Linear length of the argument of *for* (i.e. the sequences in the square brackets) does not have any effect in determining what the gloss is, and instead it is the hierarchical structures that determines what the gloss is. Then the form of gloss hints to the internal structures of the sentence.

A even more dramatic example comes from Mandarin Chinese. A single sequences of words may have distinctive meanings because of different parses, and the difference of parses is marked by the differences of glosses. In the following examples, the sentence '*Lao3Li3 mai3 hao3 jiu3*'[5] may have two distinct meanings depending on the status of '*hao3*'.

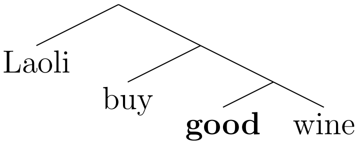
(23) a. Lao3Li3 mai3 hao3 jiu3
Laoli buy **Perf** wine
'Laoli bought a wine'

b.



(24) a. Lao3Li3 mai3 hao3 jiu3
Laoli buy **good** wine
'Laoli buys a good wine'

b.



In sentence (23), '*hao3*' goes with the verb '*mai3*'; as such '*hao3*' is interpreted as a Perfective marker and glossed as '*Perf*'; on the other hand, in (24), '*hao3*' goes with the noun '*jiu3*' and works as an adjective modifying '*jiu3*', and it is glossed as '*good*'.

With all the examples above, we have showed that gloss lines provide more clues of the the internal structures of the sentences are than natural language words do.

[5] These specific examples are extensively discussed in Mandarin Chinese Tone Sandhi literature (e.g. Cheng (1973); Mei (1991); Shih (1997); Wang and Lin (2011)). Critically, the constituency plays a role in Mandarin Chinese Tone Sandhi.

## 2.3 Conclusion: What Is a Gloss Line and Why Do They Matter?

The gloss line is like a linguistic version of ‘word embedding’. A natural language word is first converted to a gloss, which is readable for linguists. Also we may view a gloss line as an artificial sentence using the purified ‘gloss words’, a meaning representation with which one meaning is mapped to one and only one representation. It is a useful and widely used annotation algorithm that requires linguistic knowledge. Given the properties of gloss data, it can be a very useful data for machine translation. Moreover, gloss data is widely used in linguistics literature, so data is already out there and all we need to do to clean the data. A loose end here is that, even if all the arguments should be sound, we still have no statistical evidence to show the usefulness of the gloss data. Chapter 5 and 6 close this loose end, in which I will report machine translation experiments using gloss data.

# 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

   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: fieldwork and data elicitation.

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 dogs 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)

## 4.1 What is Machine Learning?

UG is a type of machine learning. Instead of teaching a set of rules, the children is giving a set of grammatical selectness and then they are able to learn grammar, and produce grammatical sentences.

Experience is the samples. Chomsky (2005) three factors: 1) UG 2) Experience. 3) other cognitive mechanisms/limitations. when the Experience is language X, a child learns language X.

Chomsky (2005) actually is a precise description of machine learning. UG is the learning algorithm. Experience is the training data. Cognitive mechanisms are all the other hardware specification of a machine (the human fresh).

Why UG is better than language-specific rules == why sample-based learning is better than rule-based learning

## 4.2 History of Machine Translation

## 4.3 Statistical Machine Translation

## 4.4 Neural Net Machine Translation

### 4.4.1 What is neural nets? Why can neural net learn?

Introduction of Neural Net Machine Translation

# Chapter 5

# Building Translation Systems using Interlinear Glossed Text: First Attempt

## 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 chapter 2 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 than 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 2 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:

(25) **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 in the current chapter will 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 empirally supported.

This chapter describes the experiments conducted to test the strong version of the Gloss-helps hypothesis. The rest of the chapter is organized as follows: Section 5.2 describes related works in the literature, Section 5.3 describes the constant parameter settings across all the experiments, Section 5.4 tests the hypothesis in (25a), Section 5.5 discusses the results and conclude this chapter.

## 5.2 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.

## 5.3 Technical Settings of the Machine Translation Experiments

The experiments are conducted 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

  In neural net machine translation, a word is represented as a vector. This hyper-parameter means that we are going to use vectors with 500 dimensions to represent words.

- Type of recurrent cell: Long Short Term Memory

  Long Short Term Memory recurrent neural net is a type of neural net that is suitable for sequence to sequence tasks.

- Number of recurrent layers of the encoder and decoder: 2

  This hyper-parameter specifies that we are going to have two recurrent layers of the encoder and decoder.

- Number of epochs: 13

  The training process of a neural net machine translation systems is done epoch by epoch. Each epoch is an iteration of training. Here 13 means that we are going to have 13 iterations of training and thus have 13 epochs.

- Size of mini-batches: 64

  Training a neural net is to let the weights of the connections between the neurons fit the training samples. Theoretically, we may ask the net adjust the weights

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

according all the samples all together at one time. However, in practice, this is not memory efficient, and will cause errors in the process of optimizing the weight parameters. So, instead, the samples are split into smaller mini-batches, and the neural net just updates its weights to make the most accurate predictions for a mini-batch at one time. This hyper-parameter specifies the size of a mini-batch. Actually finding the right mini-batch size is not a trivial but an important question in Deep Learning. See Keskar et al. (2016) and Smith et al. (2017) for the experiments and discussions on the effects of the size of mini-batches.

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. We will leave the question of what hyper-parameters are optimal for our data for future research.

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

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

## 5.4 Gloss Representation Solely Does NOT Outperform Gaelic Sentences

This section tests the strong version of Gloss-helps hypothesis in (25a). 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).

### 5.4.1 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 a Gaelic sentence, a gloss line and an English translation. Each line in the interlinear glossed text example below is an argument of a 3-tuple sample.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (26) | 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:

(27) <“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].

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

(28) 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.

(29) 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.”>

(30) 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 map gloss lines to English sentences from samples in (29a) and (29b), and another neural net system learns how to maps Gaelic sentences to English sentences from from samples in (30a) and (30b).

(31) Models:

a. $\text{Model}_{\text{GLOSStoEN}}$ = Model trained with $\text{GLOSStoEN}_{\text{Train}}$ in (29a) and $\text{GLOSStoEN}_{\text{Validatio}}$ in (29b)

b. $\text{Model}_{\text{GDtoEN}}$ = Model trained with $\text{GDtoEN}_{\text{Train}}$ in (30a) and $\text{GDtoEN}_{\text{Validation}}$ in (30b)

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).

(32) Predictions:

a. $\text{Predictions}_{\text{GLOSStoEN}}$ = A list of English sequences that $\text{Model}_{\text{GLOSStoEN}}$ maps to from the gloss sequences in (29c)

[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.

b. Predictions$_{\text{GDtoEN}}$ = A list of English sequences that Model$_{\text{GDtoEN}}$ maps to from the Gaelic sentences in (30c)

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 assumes that a sentence is a bag of n-grams (n is from 1 to 4). It measures how different the two bags of n-grams (the predicted sentence and the gold standard sentence) are. A bag of words means that the order is not important, and the difference is measured by modified precision. For concreteness, consider the following toy examples:

(33) a. Gold reference: ‘one two three four five’

b. predicted sentence 1: ‘one one two two two’

c. predicted sentence 2: ‘two two two one one’

For simplicity, let's consider unigram precision first. With the bag of words assumption, (33b) and (33c) are identical in terms unigram because they have the same set[7] of unigrams:

(34) a. predicted sentence 1: ‘one one two two two’ =

{‘one’,‘one’,‘two’,‘two’,‘two’} =

{‘two’,‘two’, ‘two’, ‘one’, ‘one’} =

predicted sentence 2: ‘two two two one one’

[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

[7] Here set does not mean the mathematical set but just unordered list.

The unigram bag of word format of the gold-standard of our example is:

(35) gold-standard unigram bag of words

a. {'one','two','three','four','five'}

Now to calculate of the proportional similarity between the predicted bag of words and the gold-standard bags of words, BLEU uses 'modified precision rate'. The technical meaning of 'precision' is whether the predicted items is actually present in the gold-standard. Now the size of the bag of unigrams of the candidate is 5, the denominator of the precision rate is 5, and the nominator is how many items in the candidate set are present in the gold-standard. In the current example, *one* and *two* are both present in the gold-standard, so the nominator of (33b) or (33c) is 5. Now we have a wrongly inflated rate, 5 out of 5, 100% matched, meaning (33b) or (33c) is 100% similar to the gold-standard. To counter the effect of this inflation, BLEU uses 'modified' precision rate. When the item in the gold-standard is matched it is crossed-out, and invisible to the predicted bag of words[8]. With this modified precision measurement, the two *one*s only get one score and the three *two*s only get one score. Now the modified precision rate is 2 out of 5, instead of 5 out of 5.

In terms of bigrams, the same examples will be:

(36) a. Gold reference: {'one_two', 'two_three', 'three_four', 'four_five'}

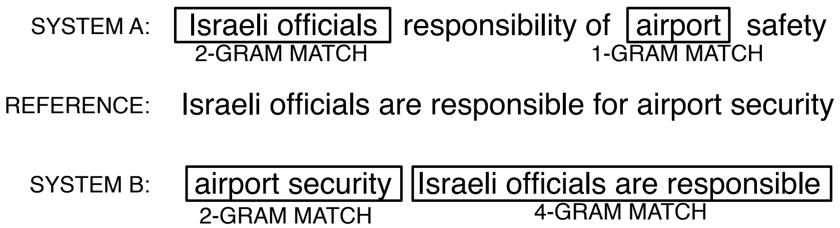
b. predicted sentence 1: {'one_one', 'one_two', 'two_two', 'two_two'}

c. predicted sentence 2: {'two_two', 'two_two', 'two_one', 'one_one'}

The denominator of the precision rate is 4 because the length of the predicted bag of words is 4; predicted sentence 1 in (36b) get 1 score because 'one_two' is matched, yielding a rate of 1 out of 4, while for predicted sentence 2 in (36c) no bigram is matched, yielding a rate of 0 out of 4.

[8] This is very similar to the feature checking mechanism in the Minimalist Program: one interpretable feature normally can only check out one uninterpretable feature.

Figure 5.1. The BLEU score is based on n-gram matches with the reference translation (Koehn, 2009, p. 226-227)



| Metric | System A | System B |
|---|---|---|
| precision (1gram) | 3/6 | 6/6 |
| precision (2gram) | 1/5 | 4/5 |
| precision (3gram) | 0/4 | 2/4 |
| precision (4gram) | 0/3 | 1/3 |
| brevity penalty | 6/7 | 6/7 |
| BLEU | 0% | 52% |

A loose end of the current measurement is that it will wrongly give a shorter predicted sentence a higher precision rate because the shorter the smaller the denominator is. To counter this, the final version of BLUE penalize short predicted sentence by multiplying the ratio between the length of the predicted sequence and the length of the gold-standard sentence. For N from 1 to 4, each N-gram comparison yields a BLEU score; the multi-bleu score is just the average of the 4 BLEU scores (unigram to four-gram).

Put all together, a concise way of describing the calculation of BLUE is the following equation.

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \Big(\prod_{i=1}^{4} \textit{precision}_i\Big)^{\frac{1}{4}} \tag{5.4.1}$$

For a little bit more complicated example of calculating the multi-bleu score, consider the following example in figure 5.1 from Koehn (2009, p. 226-227).

In short, 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.

(37) Gold-Standard = English sentences in (29c) = English sentences in (30c)

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.

(38) Scores:

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

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

This procedure of splitting the data into three sub-sets, training the models, and evaluating the models is executed for ten times.

### 5.4.2 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 slightly 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 ($\text{M}_{\text{GDToEn}}$=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)

| 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 $\text{Model}_{\text{GLOSStoEn}}$

### 5.4.3 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 (31b) 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.

## 5.5 Discussion and Conclusion

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 (5.4.2) 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 (2) 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 (2) 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 chapter.

# Chapter 6
# Combining Gaelic Words with Glosses

## 6.1 Introduction

In the previous chapter, 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:

(39) 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).

### 6.1.1 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:

(40) 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

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

### 6.2.1 Related work

The Parallel-Partial treatment section 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).

### 6.2.2 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 (30a) and (30b) ), and the data of the gloss treatment are pairs of a gloss line and a English sentences (see (29a) and (29b). 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:

(41) 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 three pairs of samples:

(42) 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">

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

(43) 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">

### 6.2.3 Results of the Parallel-Partial Treatment

Critically, the same technical settings and the same test sets in the previous experiments are used, and the same procedures are executed. The same split of the original IGTs is used, so as long as it is the same round, the training, validation and test are the same set of IGTs. The only difference is that now the training and validation IGT data are treated with the Parallel-Partial treatment. The result show that 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 ($M_{\mathrm{GDToEn}}$=16.59, $SD_{\mathrm{GDToEn}}$=0.74; $M_{\mathrm{ParaPart}}$=32.10, $SD_{\mathrm{ParaPart}}$=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.

| 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 6.1. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{ParaParttoEn}}$

*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 6.1.1, and provide strong evidence for the gloss-helps hypothesis in (**??**).

## 6.3 Other Possible Treatments

This section reports other possible ways of blending the Gaelic sentences and gloss lines[1]. 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.

### 6.3.1 The Parallel Treatment

*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

[1] There must be other possible and logical ways to blend in gloss that are beyond my imagination. It seems me to that by simply attempting to incorporate gloss information we open many other doors to the possible ways of improving machine translation systems. This is another merit of combining theoretical linguists to natural language processing.

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 two pairs. For example, the chunk of interlinear glossed text in (44) becomes two samples in (45):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (44) | 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.’

(45) 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”>

*Results of the Parallel Treatment* The experiments had our expected results. The table in (6.2) compares the performances of this treatment and the baseline. Critically, the Parallel treatment is effective in improving the baseline systems ($\text{M}_{\text{GDToEn}}$=16.59, $\text{SD}_{\text{GDToEn}}$=0.74; $\text{M}_{\text{Para}}$=29.56, $\text{SD}_{\text{Para}}$=1.46; t(9)=34.42, $p < 0.01$). However, the best

| Round | Gaelic (Baseline) | Para |
|---|---|---|
| 0 | 17.29 | 30.40 |
| 1 | 16.42 | 30.07 |
| 2 | 15.29 | 26.05 |
| 3 | 15.97 | 29.67 |
| 4 | 17.79 | 28.86 |
| 5 | 16.73 | 29.46 |
| 6 | 17.11 | 30.57 |
| 7 | 16.37 | 30.45 |
| 8 | 15.93 | 28.83 |
| 9 | 16.99 | 31.28 |
| Mean | 16.59 | 29.56 |

Table 6.2. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{ParatoEn}}$

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.

| Round | Gaelic (Baseline) | Para | ParaPart |
|---|---|---|---|
| 0 | 17.29 | 30.40 | 32.64 |
| 1 | 16.42 | 30.07 | 32.28 |
| 2 | 15.29 | 26.05 | 29.94 |
| 3 | 15.97 | 29.67 | 31.18 |
| 4 | 17.79 | 28.86 | 32.83 |
| 5 | 16.73 | 29.46 | 31.11 |
| 6 | 17.11 | 30.57 | 32.19 |
| 7 | 16.37 | 30.45 | 33.52 |
| 8 | 15.93 | 28.83 | 30.93 |
| 9 | 16.99 | 31.28 | 34.35 |
| Mean | 16.59 | 29.56 | 32.10 |

TABLE 6.3. BLEU scores of $Model_{GDtoEN}$, $Model_{ParatoEN}$ and $Model_{ParaParttoEN}$

### 6.3.2 Interleaving Gaelic Words and Gloss Items And Concating them

*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[2]. Consider the following example:

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

(46) 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 (46a), the Interleaving treatment generates the sample in (46b).

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.

The results are given in the following table.

| 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 6.4. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{interleavingGdGLOSStoEn}}$

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

(47) 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.

*Results of Concating Gaelic Words and Gloss Words* The performances of this treatment is given in the following table.

| 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 6.5. BLEU scores of Model$_{\text{GDtoEN}}$ and Model$_{\text{ConcatGLOSSGaelictoEn}}$

The result shows that this treatment hurts the baseline systems 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).

### 6.3.3 Hybrid: Gaelic or Gloss

*Method of Hybrid* The Hybrid treatment aims to reduce the potential lexical ambiguity. A Gaelic word may maps to multiple glosses, and a glosses may maps to multiple Gaelic words. Let's assume a toy chunk of interlinear glossed text data (a one-word sentence):

(48) 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 >. To decied, we need to know 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 words, then <Gloss_item, English translation> is chosen; on 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.

The following is an example of the hybrid treatment:

(49) tha nathairichean a'chuir an t-eagal orm
be.pres snake.vn put det fear on.1s
'Snakes frighten me'

The length of the Gaelic sentence and the gloss line in the above IGT is 6. This means 6 ambiguity comparisons need to be made to decide which one, Gaelic word or gloss, should take the position. The final production of this hybrid treatment on the above IGT is shown as follows:

(50) <“be.pres nathairichean a’chuir det t-eagal on.1s” ,“Snakes frighten me” >

This treatment has the same potential downside as the Concat treatment: sparsity. In this treatment, the size of the lexicon is the size of the lexicon of Gaelic word plus that of gloss, but what are really visible to the neural net is only about half size of the whole potential lexicon, because for each position it is either a Gaelic word or a gloss.

| 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 6.6. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{HybridDefaultAsGaelictoEn}}$

*Result of Hybrid* When the default setting is the Gaelic word, the performances are significantly worse than than the baseline systems ($\text{M}_{\text{GDToEn}}$=16.59, $\text{SD}_{\text{GDToEn}}$=0.74; $\text{M}_{\text{ReplacingGaelic}}$=9.24, $\text{SD}_{\text{ReplacingGaelic}}$=0.89,; t(9)=-21.03, $p < 0.01$), as shown in table (6.6).

When the default setting is the Gaelic word, the performances are sightly worse than than the baseline systems ($\text{M}_{\text{GDToEn}}$=16.59, $\text{SD}_{\text{GDToEn}}$=0.74; $\text{M}_{\text{ReplacingGaelic}}$=15.47, $\text{SD}_{\text{ReplacingGaelic}}$=1.03,; t(9)=-3.67, $p < 0.01$ ), as shown in table (6.7).

The results show that this hybrid treatment indeed suffers from the sparsity problem.

| 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 6.7. BLEU scores of $\text{Model}_{\text{GDtoEN}}$ and $\text{Model}_{\text{HybridDefaultAsGLOSS}}$

## 6.4 Summary and Conclusion

Chapter 2 argues that the gloss representation is a golden representation of meanings, and thus theoretically with the gloss information incorporated, the performance of machine translation systems should improve. The experiments reported in this chapter reveal an effective way of combining gloss data and Gaelic sentences. It is found that the Parallel-Partial is highly effective, and the Gloss-helps hypothesis is empirically supported by the results. The complete BLEU scores of various treatments are given in the following table.

In the next chapter, I will discuss the implications the experiments have for theoretical linguistics and natural language processing.

| Round | Baseline | GLOSS | ParaPart | Para | Interleaving | Concat | HybrGaelic | HybrGLOSS |
|---|---|---|---|---|---|---|---|---|
| 0 | 17.29 | 18.39 | 32.64 | 30.40 | 13.67 | 15.42 | 9.44 | 15.95 |
| 1 | 16.42 | 18.00 | 32.28 | 30.07 | 12.49 | 14.31 | 9.07 | 15.60 |
| 2 | 15.29 | 16.02 | 29.94 | 26.05 | 11.01 | 15.38 | 7.69 | 14.15 |
| 3 | 15.97 | 20.22 | 31.18 | 29.67 | 12.33 | 14.18 | 9.12 | 14.72 |
| 4 | 17.79 | 19.02 | 32.83 | 28.86 | 12.56 | 18.63 | 9.08 | 15.74 |
| 5 | 16.73 | 15.53 | 31.11 | 29.46 | 12.13 | 14.89 | 10.45 | 14.88 |
| 6 | 17.11 | 18.00 | 32.19 | 30.57 | 11.55 | 15.16 | 8.62 | 14.45 |
| 7 | 16.37 | 20.08 | 33.52 | 30.45 | 12.78 | 15.20 | 10.00 | 16.41 |
| 8 | 15.93 | 15.82 | 30.93 | 28.83 | 12.43 | 15.50 | 10.52 | 15.15 |
| 9 | 16.99 | 15.93 | 34.35 | 31.28 | 11.65 | 15.72 | 8.46 | 17.61 |
| Mean | 16.59 | 17.70 | 32.10 | 29.56 | 12.26 | 15.44 | 9.24 | 15.47 |

Table 6.8. BLEU scores of the all treatments

# Chapter 7

# Interlinear Glossed Text Data In Other Languages

To run the same experiments on Online Database of Interlinear Text (Lewis and Xia, 2010; Xia et al., 2016) to show that incorporating gloss information works effectively across different languages.

| Language Names | Language (Baseline) | GLOSS | ParaPart | Sample size |
|---|---|---|---|---|
| Mandarin | 1.13 | 0.00 | 10.25 | 3719 |
| German | 0.00 | 3.13 | 18.08 | 4037 |
| Greek | 0.00 | 0.00 | 0.00 | 1188 |
| Finnish | 0.00 | 0.00 | 2.56 | 1968 |
| French | 0.96 | 0.00 | 9.49 | 3099 |
| Hausa | 0.00 | 0.00 | 4.41 | 1490 |
| Hungarian | 0.00 | 0.00 | 0.00 | 1197 |
| Indonesian | 0.00 | 0.00 | 2.34 | 1211 |
| Icelandic | 0.00 | 0.00 | 9.55 | 1719 |
| Italian | 0.00 | 0.00 | 0.00 | 1366 |
| Korean | 2.70 | 2.57 | 11.90 | 3630 |
| Dutch | 0.00 | 0.00 | 9.50 | 1454 |
| Norwegian | 0.00 | 0.00 | 3.16 | 1403 |
| Polish | 0.00 | 0.00 | 5.25 | 1713 |
| Passamaquoddy | 0.00 | 0.00 | 7.41 | 1166 |
| Russian | 0.00 | 0.00 | 4.64 | 2266 |
| Spanish | 0.00 | 0.00 | 4.42 | 1733 |
| Turkish | 0.00 | 0.00 | 2.62 | 1420 |

Table 7.1. BLEU scores of other languages

# Chapter 8

# Implications on Theoretical Linguistics and Natural Language Processing

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

# Chapter 9

# Conclusion

Conclusion

# Bibliography

Adger, David (2003), *Core syntax: A minimalist approach*, volume 33. Oxford University Press Oxford.

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

Berwick, Robert C and Noam Chomsky (2015), *Why only us: Language and evolution*. MIT press.

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

Chen, Yuan-Lu (2010), *Degree Modification and Time Anchoring in Mandarin*. Ph.D. thesis.

Cheng, Chin-Chuan (1973), *A synchronic phonology of Mandarin Chinese*, volume 4. Walter de Gruyter.

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*.

Chomsky, Noam (2005), "Three factors in language design." *Linguistic Inquiry*, 36, 1–22, URL https://doi.org/10.1162/0024389052993655.

Chomsky, Noam (2006), *Language and mind*. Cambridge University Press.

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

Grano, Thomas (2008), "Mandarin hen and the syntax of declarative clause typing." *Unpublished manuscript. Accessed online:< http://home. uchicago. edu/˜ tgrano/grano_hen. pdf>. First accessed*, 4.

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*.

Keskar, Nitish Shirish, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang (2016), "On large-batch training for deep learning: Generalization gap and sharp minima." *CoRR*, abs/1609.04836, URL http://arxiv.org/abs/1609.04836.

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.

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

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

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

Lewis, William D. and Fei Xia (2010), "Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages." *Literary and Linguistic Computing*, 25, 303–319, URL +http://dx.doi.org/10.1093/llc/fqq006.

Liu, Chen-Sheng Luther (2010), "The positive morpheme in chinese and the adjectival structure." *Lingua*, 120, 1010–1056.

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*.

Mei, Tsu-Lin (1991), "Tone sandhi and morphological relics." *Journal of Chinese Linguistics Monograph Series*, 454–471.

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.

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

Shih, Chilin (1997), "Mandarin third tone sandhi and prosodic structure." *Linguistic Models*, 20, 81–124.

Smith, Samuel L., Pieter-Jan Kindermans, and Quoc V. Le (2017), "Don't decay the learning rate, increase the batch size." *CoRR*, abs/1711.00489, URL http://arxiv.org/abs/1711.00489.

Sneddon, James Neil, K Alexander Adelaar, Dwi N Djenar, and Michael Ewing (2012), *Indonesian: A comprehensive grammar*. Routledge.

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.

Wang, Chiung-Yao and Yen-Hwei Lin (2011), "Variation in tone 3 sandhi: The case of prepositions and pronouns." In *Proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)*, 138–155.

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

Xia, Fei, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily M. Bender (2016), "Enriching a massively multilingual database of interlinear glossed text." *Language Resources and Evaluation*, 50, 321–349, URL https://doi.org/10.1007/s10579-015-9325-4.

Zhang, Niina Ning (2013), *Classifier Structures in Mandarin Chinese*, volume 263. Walter de Gruyter.