Negation Learning in Transformer-based Neural Machine Translation Models*

Duc M. Trinh
Advisors: Alvin Grissom II and Jane Chandlee

A thesis submitted in partial fulfillment of the requirements for
the degree of Bachelor of Arts in Computer Science & Linguistics

Haverford College
December 10th 2021

I would like to thank my advisors, Alvin Grissom II and Jane Chandlee, for their endless patience, encouragement, and guidance. I would also like to thank Jonathan Washington, Jiangxue Han, Yiying Jiang, and Jeova Neto for the valuable feedback they gave me on earlier drafts of this thesis. I would like to acknowledge Lingxi Hua for providing her linguistic expertise for me and also for lending me her Stat notes. I also would not have made it through college without the friendships of Quoc Anh Ngo, My Nguyen, Ian Davis, Justus Best, and Oleksandr Litus. I am extremely grateful to Ngan Nguyen, my best friend and soulmate, for her neverending support throughout the years. And finally, I would like to express my deepest gratitude for my parents, who have worked tirelessly to get me here. For all they have done for me, I am forever going to be in their debt.

Abstract

Negation is a core component of all human natural languages. Despite this, even the best machine translation models still struggle to deal with negation. In this thesis, we examine why neural machine translation (NMT) models have problems with negation. We specifically focus on Transformer-based models since they are the best performing models when it comes to negation in the literature. In our experiments, we modify testing data to investigate the limit of what the state-of-the-art NMT models learn about negation. We find that removing the negation cue has the highest effect in changing the polarity of a sentence. Furthermore, we discover that NMT models are more likely to translate a sentence as negated when it contains NPI terms than when it does not.

Table of Contents

1 Introduction

Negation is a linguistic phenomenon universal across human natural languages. An example would be: "I am a Haverford student" versus "I am not a Haverford student". All languages must discuss events that do not happen or describe properties that objects do not hold. Since every language supports negation, it comes in all forms, which makes it a complex subject to study.

Machine translation (MT) is subfield of **natural language processing** (NLP), with the primary goal of transforming text or speech from a source language to the text or speech of a target language. The most recent method in MT uses deep neural networks and is called **neural machine translation** (NMT). Given sufficient training data, these models achieve the state-of-the-art results for benchmark problems in MT (Goldberg 2017). NMT models are usually end-to-end models, which mean they take a sequence from a source language as input and ideally produce a sequence in the target language with the same meaning as output.

Since negation is very common, yet has the power to flip the semantic content of a sentence, being able to translate negation correctly is crucial for any MT system. However, research about negation in NLP started late compared to other linguistic phenomena (Jiménez-Zafra et al. 2020). Its application is still limited, with most success being seen only in the field of information retrieval and sentiment analysis (Wilson et al. 2005, Socher et al. 2013, Zhu et al. 2013, Reitan et al. 2015). Previous work shows that negation has been problematic for both statistical machine translation (SMT) and NMT. In some cases, the presence of negation in a sentence can drastically reduce the quality of translation (Hossain et al. 2020). Z-scores are the official ranking criterion in WMT competitions, and Hossain reports many language pairs show a substantially worse Z-scores, up to 60%, for sentences that have negation. A recent study conducted by Tang et al. (2021) demonstrates that even the Transformer (Vaswani et al. 2017), the current best performing NMT architecture, still performs poorly on negation.

Hence, within the current literature, we find that there is not a conclusive answer to the problem of negation and machine translation. In this thesis, we perform experiments to probe what NMT learns about negation (specifically focusing on the Transformer-based model), which might lead to insights on how human languages model negation. The experiments we run involve changing the testing data such as moving the location of negation cue or removing the negation cue altogether to test how well these NMT models respond to such

modifications.

The structure of this paper is as follows: Section 2 covers negation (the basics and some of the complexities that come with negation) as well as MT (its history as well as the common structure of NMT), and there is a subsection towards the end to discuss the problems that negation causes for NMT. Section 3 discusses our experimental set up, which includes the datasets we are using and the specifics of the NMT models we are training. Section 4 goes in-depth about the different experiments that we use to test the limit of the NMT models and their results. Section 5 summarizes the results, limitations of this thesis, and future directions.

2 Literature Review

This section provides an overview of the concepts needed for this thesis as well as some related concepts. It is divided into two subsections: negation and machine translation.

2.1 Negation

This section covers some background notions on negation.

2.1.1 Negation 101

This section goes over our definition of negation and its main components. From a logical standpoint, negation is a simple operator that flips the truth value of a proposition. For example, "the birds are singing" becomes "the birds are not singing". This sentence demonstrates the simplest form of negation where a declarative sentence is negated. In general, negation "relates an expression e to another expression with a meaning that is in some way opposed to the meaning of e" (Horn & Wansing 2020). Despite its seemingly simple nature, negation serves the crucial purposes of facilitating the "uniquely human capacities of denial, contradiction, misrepresentation, lying, and irony" (Horn & Horn 1989). Due to its importance, every natural language has negation in some forms. Since negation varies highly between languages, we now discuss some of the varieties and complexities of negation.

According to Blanco & Moldovan (2011) and Morante & Blanco (2012), negation in a text consists of four main components: cue, event, scope, focus. We can see these four components from the simple Example 1, adapted from Fancellu & Webber (2015):

(1) She is not eating her food.

- **Cue**: the word or multiple-word unit that does the work of negation, the most recognized negation cue in English is the word *not* (e.g. "She is **not** eating her food").
- **Event**: the lexical element that the cue directly refers to (e.g. "She is not **eating** her food").
- **Scope**: the part of the sentence that is negated; if any portion of the scope is false, it will prove the negation false (e.g. "**She is** not **eating her food**").
- **Focus**: the part of the scope that is most explicitly negated. Determining the focus is complex, since it requires interpreting which parts of the scope are supposed to be true or false (e.g. "She is not eating **her food**").

2.1.2 Negation Typology and Related Topics

This section serves to illustrate the various forms negation take on across the world's natural language sand its complexity. **Sentence/sentential negation** is the type of negation that affects the semantic meaning of an entire clause, which has been studied extensively (Dahl 1979, Belletti & Rizzi 1996, Nyberg 2012). We cover the typology of standard negation, the most basic way a language can negate a clause. Then, we talk briefly about **constituent negation**, which is negation of a non-clausal constituent (Hossain et al. 2020). After that, we quickly give an overview of lexical negation. Finally, we talk a bit about negative polarity items.

Standard negation

Standard negation is defined as "the basic means that languages have for negating declarative verbal main clauses" (Miestamo 2007). In English, the most common standard negation method would be to add *not* after the auxiliary verb. There are also non-standard negations that are negation methods for imperatives, existentials, nonverbal clauses, etc, which are outside the scope of this thesis. From a 240-language sample, Dahl (1979) divides standard negation into two types: morphological and syntactic negation. They further divide morphological negation into three main categories: prefixal (Example 2 - Latvian), suffixal (Example 3 - Lezgian), circumfixal (Example 4 - Chukchi). The examples are taken from Miestamo (2007), and explanations of the glossing abbreviations are in the footnote.[1]

[1] nominative (NOM), third person (3), locative case (LOC), negation (NEG), plural (PL), adelative (ADEL),

(2) a. *tēv-s* *strādā* *pļavā*
father-NOM work.3 meadow.LOC
'Father is working in the meadow'

b. *tēv-s* *ne-strādā*
father-NOM NEG-work
'Father is not working'

(3) a. *xürünwi-jri* *ada-waj* *meslät-ar* *q̃aču-zwa*
villager-PL(ERG) he-ADEL advice-PL take-IMPF
'The villagers take advice from him.'

b. *xürünwi-jri* *ada-waj* *meslät-ar* *q̃aču-zwa-č*
villager-PL(ERG) he-ADEL advice-PL take-IMPF-NEG
'The villagers do not take advice from him.'

(4) a. *čejwə-rkən*
go-DUR
'(S)he goes.'

b. *a-nto-ka* *(itə-rkən)*
NEG-go.out-NEG be-DUR
'(S)he does not go out.'

In syntactic negation, a negation marker is used as the negation method. According to Dahl (1979), the negation marker can be mainly classified into either an uninflected particle (Example 5 - Indonesian, Example 6 - French) or an auxiliary verb (Example 7 - Finnish). The different negation markers can be seen in these examples taken from Miestamo (2007):

(5) a. *mereka menolong kami*
they help us.EXCL
'They helped us.'

b. *mereka tidak menolong kami*
they NEG help us.EXCL
'They did not help us.'

(6) a. *je* *chante*
1SG sing.PRES.1SG
'I sing.'

ergative case (ERG), imperfect (IMPF), durative aspect (DUR), exclusive person (EXCL), first person (1), singular (SG), present tense (PRES), connegative (CNG).

b. *je ne chante pas*
1SG NEG sing.PRES.1SG NEG
'I do not sing.'

(7) a. *koira-t haukku-vat*
dog-PL bark-3PL
'Dogs bark.'

b. *koira-t ei-vät hauku*
dog-PL NEG-3PL bark-CNG
'Dogs do not bark.'

Constituent negation

For this section, we need to define two concepts: sentence polarity and tag questions. If a sentence is affirmative in nature, it is considered to have positive **polarity**. Example 8 is considered to have positive polarity, while Example 9 has negative polarity. However, the polarity of a sentence is not always clear. For example, "I failed my test yesterday" is a truthful description of an event, but the situation it is describing is not a positive one. In cases such as this example, it is not easy to decide the polarity of a sentence (Swart 2010).

(8) She was able to find the perfect gift.

(9) She was not able to find the perfect gift.

A **tag question** is a *yes/no* confirmation question attached to another clause that might be positive or negative in nature (Achiri-Taboh 2015). In English, a positive tag question comes after a sentence with negative polarity and vice versa. We can easily attach tag questions to Examples 8 and 9 as can be seen in Examples 10 and 11:

(10) She was able to find the perfect gift, was she not?

(11) She was not able to find the perfect gift, was she?

These two concepts are going to be important in our discussion of constituent negation. Klima (1964) proposes some tests that can help differentiate between sentential negation and constituent negation in English. Here we present two of his tests: Example 12 tests whether one can add an *either/too* tag to a sentence with multiple negated clauses, and Example 13 tests whether one can assign a positive or negative tag to a question. The examples used for these tests are adapted from Swart (2010):

(12) *either* vs. *too* tags

a. Tom is not affected, and Jerry is not affected either.

b. Tom is unaffected, and Jerry is unaffected {*either/too}.

(13) positive vs. negative tag questions

a. He was not being polite, was he?

b. He was being impolite, {#was he/wasn't he}?

According to Klima (1964), the examples in part (a) of 12 and 13 pass his tests and are considered sentential negation, while the ones in part (b) are considered constituent negation. The reason that sentences with constituent negation fail these tests is that constituent negation does not affect the whole sentence. Hence, it is not clear whether these sentences are positive or negative in nature; we do not know the polarity of the sentences. However, to be able to add a tag question, we need to know the polarity of the sentence, so it gets confusing when we try to add any type of tags question onto a sentence such as Example 13(b). One can argue that both "was he" and "wasn't he" can be used as tag question for this example. On the other hand, examples 12(a) and and 13(a) are clearly negative in nature, so we do not have such trouble. According to Swart (2010), there are more tests in the literature on English that help draw these distinctions, but their results can be conflicting (Horn & Horn 1989).

Lexical negation

Lexical negation uses lexical elements that are inherently negative. We cover two types here: adjectives formed by adding negative affixes and verbs that convey negative meaning by themselves.

Just considering English adjectives (1964), the following is a list of the most notable negative affixes, arranged by their ascending productivity (the frequency that new words are formed with these prefixes):

1. *a-/an*: **a**symptomatic, **a**synchronous
2. *dis-*: **dis**advantageous, **dis**honest.
3. *in-/il-/im-/ir-*: **il**logical, **im**moral.
4. *non-*: **non**intuitive, **non**maternal.
5. *un-*: **un**sullied, **un**enlightened.

Some verbs that inherently have negative meanings are: *fail*, *deny*, *lack*, *refuse*, *reject*, and *doubt* (Hossain et al. 2020, Nyberg 2012). Example 14(a) demonstrates the use of an inherently negative verb in a sentence. The sentence is stating the truth, but the word *fail* comes with a negative meaning. This context makes it hard to judge whether this sentence is positive or negative in nature. On the other hand, a sentence such as Example 14(b), with nearly the same meaning, is clearly negative in nature.

(14) a. I **failed** to get to her in time.

b. I was **not** able to get to her in time.

Negative polarity items

Negative polarity items (NPI) are lexical items that only show up in the scope of negation or semantically related contexts (Horn & Wansing 2020). All the highlighted parts in Example 15 are considered NPIs (taken from Horn & Wansing (2020)):

(15) a. I {have not/*have} **ever** eaten any kumquats **at all**.

b. {Few/*Many} of the assignments have been turned in **yet**.

c. The dean {rarely/*often} **lifts a finger** to help students on probation.

d. I {doubt/*believe} they are **all that** pleased with the proposal.

e. {All/*Many} customers who had **ever** purchased **any** of the affected items were (***ever**) contacted.

The main condition for NPIs is they need to be in **downward entailing** contexts: in other words, contexts where we can make inferences from sets to subsets (but not vice versa) (Horn & Wansing 2020). For example, in Example 15(a), "I have not eaten fruit" entails "I have not eaten kumquats", but not necessarily vice versa. Here, we are narrowing down from a set (fruit) to a subset (kumquats), which indicates that we are in a downward entailing environment.

Since NPI can also show up in negation semantically related context, the existence of an NPI in a sentence does not mean that the sentence has negation. Examples 16, 17, and 18 demonstrate cases where the sentences have NPI but are not under negation. Example 18 demonstrates the phenomenon of *positive anymore* where *anymore* is used in an affirmative sentence (Shields 1997). It can be understood as: "Traffic was not so bad before, but it is now". This usage of *anymore* is rare and might not be grammatical for many English speakers.

(16) That is the best looking car I have **ever** seen.

(17) Do you know **any** of these people?

(18) Traffic is so bad **anymore**.

Even though NPI does tend to show up with negation, the examples above show why we can not use NPI to decide whether a sentence is under negation. We can start imagining some of the difficulties for computing systems to learn these subtle differences and contexts where negation takes place.

2.1.3 Negation and Natural Language Processing

The previous section shows how negation is a common yet complex linguistics phenomenon. This section showcases some works in natural language processing (NLP) that have benefited from making use of negation information.

Work involving negation in NLP started quite late in comparison with other linguistics phenomena (Jiménez-Zafra et al. 2020). Due to its complexity and the late start, computational linguists have not yet grasped how to model negation properly for many NLP systems. Despite our lack of understanding, past literature still indicates that incorporating knowledge about negation can greatly benefit applications such as information retrieval or sentiment analysis (Auerbuch et al. 2004, Socher et al. 2013, Reitan et al. 2015).

Much work on negation started with information retrieval in English, in particular processing clinical records (Chapman et al. 2001, Mutalik et al. 2001, Goldin & Chapman 2003). By nature, medical data contains a lot of explicit negation (Rokach et al. 2008). Hence, being able to process negation content in this field is crucial. Auerbuch et al. (2004) estimate that not utilizing negation content in medical reports can reduce the precision of the overall information retrieval system's by up to 40%. Precision is measured as a ratio between the total number of relevant documents retrieved and the total number of documents retrieved. Zhu et al. (2013) test different models that use discharge summaries to aid web health information queries by patients. They report that the best model for English utilizes NLP-produced information, including negation content.

Another NLP field that has traditionally made use of negation is sentiment analysis. The objective of sentiment analysis is to classify a string of text into different sentiment classes (usually between positive and negative). We can see how negation is important for this task as a word like *good* usually means positive sentiment, but adding a negation cue can change the overall sentiment (e.g. *not good*). Wilson et al. (2005) suggest that phrases contain negation words can help intensify a sentiment instead of flipping it (e.g. **not only**

good but amazing). Hence, they encode negation as sentiment modification features in their system. By doing this, they allow for a wider range of usage for negation, which can capture intensification such as in "not only good but amazing". Other exemplar works includes Socher et al. (2013)'s neural network, which captures sentiment change when negation is present. An example of how their neural network works is in Figure 1. Their model improves state-of-the-art result for single sentence sentiment classification, going from 80% to 85.4% in accuracy. We also have more recent work, such as a state-of-the-art Twitter sentiment analysis for English that makes use of negation information (Reitan et al. 2015).

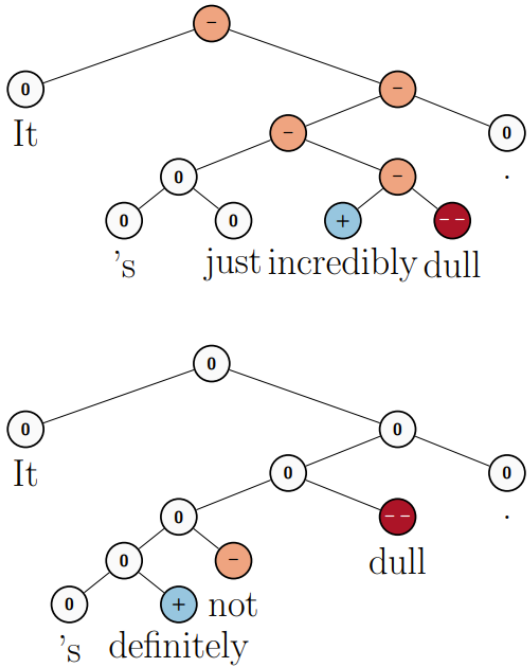

Figure 1: Example of how negation has been used in sentiment analysis. Socher et al. propose a new model called the Recursive Neural Tensor Network (RNTN) that works on a parse tree of a sentence. RNTN is a compositional model that can compose the sentiment information from lower nodes to calculate the overall sentiment at higher nodes in the tree. For example, the lower node *incredibly* conveys a positive sentiment, while the lower node *dull* has a negative sentiment. Socher et al.'s neural network has a composition function that will calculate the sentiment compositionality of these two lower nodes resulting in a sentiment composition of 0 (neutral) for their parent node. This figure was taken from Socher et al. (2013).

However, outside of the tasks mentioned above, researchers have found that negation proves to be problematic for other computational linguistic tasks such as machine translation. Our next section introduces machine translation and gives a brief overview of its evolution.

2.2 Machine Translation

This section is dedicated to give an overview about machine translation (MT). In this section, we will cover MT's objective and its main approaches.

MT is a subfield within NLP that involves using computer software to translate text or speech of one language to another. MT is useful as it is usually faster and can be used without additional human input; it could also be used as a base translation to supplement the human translator. The training objective of corpus-based MT is to maximize the conditional probability of the target sentence y given the source sentence x, which can be written as $argmax_y p(y|x)$ (Bahdanau et al. 2015). The subfield started in the 1950s with rule-based machine translation. This method was not considered robust enough, and it got replaced by statistical machine translation in the 2000s. More recently, the state-of-the-art for MT has moved to employ deep learning neural networks (Luong et al. 2015, Wu et al. 2016).

2.2.1 Rule-Based Machine Translation

This section discusses the oldest approach to MT, rule-based machine translation (RBMT), but is still being used today. RBMT usually operates using grammatical knowledge and lexicon of languages made by linguists to translate sentences (Scott & Barreiro 2009, Khanna et al. 2021). We start to see the drawback with this method as grammar rules are hard for computer scientists to generate. They can not cover all the cases within a language and often conflict with each other. Language also changes over time, which requires updates or replacements of existing rules. However, these models do not require a large language corpus, which makes it suitable for low resource languages (Hurskainen & Tiedemann 2017). Literature has also shown that more complicated RBMT models can match and surpass the performance of MT models that are data-driven (Bayatli et al. 2018).

2.2.2 Statistical Machine Translation

RBMT was eventually largely replaced by statistical machine translation (SMT), a mainstream method for MT from 1990 until recently. This section summarizes some key notations about SMT. SMT's approach differs from RBMT in that it breaks sentences down into

phrases, each of which can be replaced using a word or a phrase of the same meaning (Koehn et al. 2003, Chiang 2007). For example, in a sentence such as "I am going to the student center", the model is going to look to replace the whole phrase "the student center" instead of each individual token. The model is trained on a bilingual corpus, and the transformation is decided via statistics. Given enough data, the training process can be as fast as one day (Oard & Och 2003). However, in cases of low resource language, SMT models run into issues with data sparsity (Lopez 2008). SMT's phrase-based approach allows it to utilize contextual information in the sentence, which greatly improves the translation accuracy compared to previous models that did not use contextual information.

2.2.3 Neural Machine Translation

This section provides some background on neural machine translation (NMT), an approach that has gathered attention in the recent years. With the development of deep learning, neural machine translation (NMT) that makes use of deep neural networks have become the new state-of-the-art for MT. NMT's training objective is to maximize the log-likelihood L with regard to θ (Bahdanau et al. 2015):

$$L_\theta = \sum_{x,y \in C} \log p(y|x;\theta) \tag{1}$$

where C is a parallel corpus with $x = \{x_1, ..., x_n\}$ as an input sentence, $y = \{y_1, ..., y_m\}$ as its translation, and θ as a set of parameters to be learned.

NMT is an end-to-end sequence-to-sequence (seq2seq) model. The model's learning goal is to map from a sequence in the source language to another sequence in the target language. Most mainstream NMT architectures can be broken down into two parts: an *encoder* and a *decoder*. The encoder and decoder are two connected networks, and an example can be seen in Figure 2.

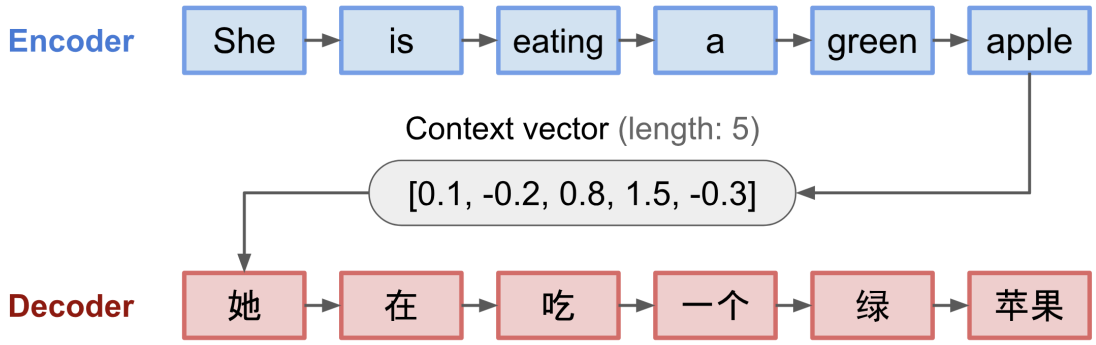

Figure 2: An example of an encoder-decoder network, the most common paradigm for NMT. The encoder converts the information from the input sequence into a context vector, which is then used by the decoder to predict the output sequence. Figure taken from Weng (2018).

A Recurrent Neural Network (RNN) is a type of neural network that is trained to predict the next symbol in a sentence given the previous symbols in the sentence (Cho et al. 2014, Sutskever et al. 2014). In an RNN, the probability of a target sentence given the source sentence is:

$$p(y|x;\theta) = \prod_{j=1}^{m} p(y_j|y_{<j},x;\theta) \tag{2}$$

where m is the number of words in y, y_j is the current generated word, and $y_{<j}$ are the previously generated words.

In an RNN, the encoder network reads in the input sequence token-by-token and compresses that information into vectors. These vectors contain the contextual information that was fed into the encoder and are referred to as **context vectors**. Then, these vectors are directly inputted into the decoder network. The decoder network produces the final output sentence from these vectors and its previous output, token-by-token (Kalchbrenner & Blunsom 2013, Cho et al. 2014, Sutskever et al. 2014, Bahdanau et al. 2015).

Since Vaswani et al. (2017), the **Transformer** has become the current best performing architecture for NMT. The Transformer is a fully attention-based model, its architecture can be seen in Figure 3.

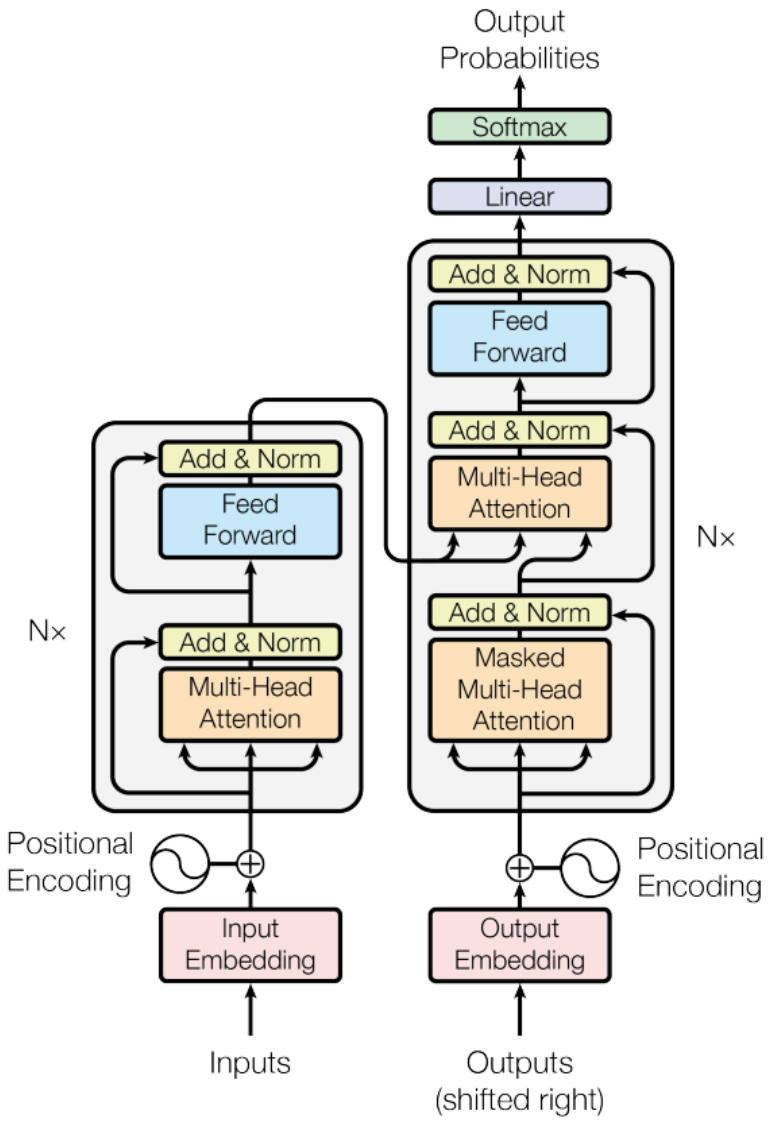

Figure 3: Transformer architecture as proposed in Vaswani et al. (2017). Figure taken from Vaswani et al. (2018)

In an RNN-based model, the most recently seen words are more important to the model, so errors tend to happen when the context happens far before the word we are trying to translate. The attention mechanism gives the encoder access to every word in the sequence at all times. For each word that the decoder predicts, the model can decide which words in the sequence it needs to "attend" to. An example of how this mechanism works can be seen in Figure 4.

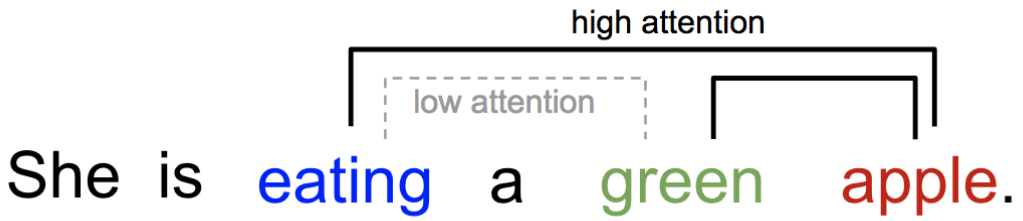

Figure 4: An example of the attention mechanism at work. Since "apple" is a food, the model pays a lot more attention to it when translating "eating". The Figure taken from Weng (2018).

Because "apple" is a food, "eating" is going to pay a higher attention to it than other words in the sequence in the example in Figure 4. This type of information allows the decoder network to more effectively produce a translation. By having access to all of the sequence at all times, the Transformer can learn distant relationships better than RNN models (Vaswani et al. 2018). Since it can change its *attention* for each word in the sequence, the first word in the sequence matters just as much as the last. Not only are the Transformer models more accurate, they can also be parallelized, which makes them faster to train than RNN models.

Traditional machine translation methods usually require hand-crafted features and pre-processing steps based on linguistics knowledge. Compared to these methods, NMT requires minimum domain knowledge and is better at representing data and dealing with data sparsity in training (Yang et al. 2020), yet its performance surpasses RBMT and SMT. However, the traditional problem with using NMT is that while they are excellent at predicting the correct translation, it is difficult to figure out what they "learn" from the data. Hence, it is hard to fix the models when things go wrong. In fact, recent literature shows that negation is still an unsolved problem for NMT (Hossain et al. 2020). The next section discusses recent research involving negation and NMT.

2.3 Neural Machine Translation and Negation

This section examines the different errors that NMT model have when translating negation. Researchers have shown that NMT is superior to SMT when it comes to translating negation (Bentivogli et al. 2016, Beyer et al. 2017). However, NMT's performance still suffers when

negation is involved. Hossain et al. (2020) claim the errors NMT makes when translating negation can be classified into four types (the examples come from the same paper):

- **Negation omission**: the negation cue is not translated in the output sentence; this error is also often referred to as **under-translation**. Example 19 is taken from the Turkish-English WMT18 shared task:

(19) Source Sentence: 'Eğer seçmenlerin eleştirilerini kaldiramiyorsan devlet görevine aday olma.'
Reference Translation: '[...] Don't run for public office, if you can't take heat from voters.'
System Translation: '[...] if you can't take criticism from voters, you're a candidate for state duty,'

- **Negation reversal**: the output sentence has the opposite semantic meaning to the intended meaning. The error can be seen in both Example 19 and 20. Example 20 is taken from the WMT19 Lithuanian-English shared task:

(20) Source Sentence: 'Panaikinti išteisinam ąjį nuosprendį prašo ir žuvusio motociklininko šeimos advokatė.'
Reference Translation: 'The family lawyer of the deceased biker also asks for reversal of the verdict of not guilty.'
System Translation: 'The family lawyer of the dead rider also asks for the conviction to be lifted.'

- **Incorrect negation scope**: the wrong constituent is negated in the sentence, such as Example 21 from the Finnish-English WMT18 shared task:

(21) Source Sentence: 'Viimeinen tuomio ei ole syy.'
Reference Translation: 'The reason is not the Last Judgement.'
System Translation: 'The Last Judgment is not the reason.'

- **Mistranslation of negated object**: the negated element in the sentence is wrongly translated. An example for this type of error is Example 22 from the WMT18 German-English shared task:

(22) Source Sentence: 'Zu einem Personalienaustausch kam es aber nicht, da der 75-Jährige die Dame auf dem Parkplatz nicht mehr finden konnte.'
Reference Translation: 'No exchange of personal data occurred [...],'
System Translation: 'There was no exchange of personnel [...].'

Using the *polarity* set of *LingEval97*, Sennrich (2017) evaluates different NMT models and finds that negation still creates difficulties for the NMT models, especially when the negation cue is deleted. Ding et al. (2017) assert that neither attention weights nor layer-wise relevance propagation (LRP) can explain the under-translation error for NMT models. Hossain et al. (2020) perform an in-depth study of the impact of negation on translation using 17 translation directions, all involving English. They show that negation is still a challenge to NMT and that there there are fewer translation errors when the target language is more similar to English in regard to the typology of negation. Tang et al. (2021) test the ability to translate negation of different architectures: RNN-, CNN-, and Transformer-based. They claim that Transformer-based models consistently perform better than the other models across different tasks from the *polarity* set of *LingEval97*. The tasks include testing the models translation accuracy when negation is inserted or removed. The most common error for all language directions is found to be under-translation, such as in Example 19, which brings into question whether the NMT model is learning the negation information properly. They perform further experiments such as testing whether the model can classify a negation cue versus other tokens in the sentence. They find out that the Transformer can correctly identify negation cue versus other tokens, but the model still struggles to translate negation nonetheless.

It is clear that the literature still has not figured out the negation problem for MT. Therefore, we endeavor to further contribute to the study of why NMT models struggle with negation. We will create different types of testing datasets that will examine the limits of the NMT approach in an attempt to probe what the NMT model is learning about negation. Ultimately, we aim to contribute to the improvement of these models when it comes to negation. In order to perform our experiment, we need trained NMT models and testing data to work with. The specifics of how we obtain these are described in the next section.

3 Experiment Setup

This section summarizes the steps we undertook before running our experiments. The steps include gathering training data for our models, training our NMT models, and performing

negation detection on the testing set.

3.1 Datasets

This section details the sources of the datasets used for our experiments. The datasets used in these experiments come from the data submissions for the shared translation task of the Workshop on Statistical Machine Translation (WMT) competition. In this competition, participants build different MT systems to solve various MT tasks. The findings are published every year to highlight the newest improvements in the field of MT (Bojar et al. 2017, 2018). The shared task for WMT is a recurring task that requires participants to train MT models for the provided language pairs. We have chosen to focus on the data for English (EN)-Japanese (JA) and English (EN)-Chinese (ZH) from the WMT2021 competition, and to train models for both direction of these language pairs. These two language pairs are chosen due to the fact that they are high resource languages to ensure that we would have enough training data for the NMT models. In addition, they are languages that we had advisors that could provide linguistics expertise needed for the manual evaluation.

The data for English-Japanese comes from JParaCrawl, one of the biggest web-based English-Japanese parallel corpora (Morishita et al. 2020). The dataset contains about 10M (million) sentences formed by crawling the web for English-Japanese bitexts. We use the first 8M sentences as training data, the next 1M sentences as validation data, and the rest of the data as testing data (about 1.1M sentences). We call the model trained using this data the 8M EN-JA model.

For English-Chinese, our data comes from UN Parallel Corpus 1.0 (Ziemski et al. 2016). This dataset contains about 16M sentences formed using official records and other parliamentary documents of the United Nations between 1990 and 2004 that are in the public domain. We use the first 10M sentences as training data, the next 2M sentences as validation data, and the next 2M sentences as testing data. We call the model trained using this data the 10M EN-ZH model.

3.2 Negation Detection

This section explains how we perform negation detection on our testing dataset. Our experiments involve modifying negated sentences and seeing how the models respond to the modification. Because of this, we first need to identify the negated sentences in our testing dataset. Unfortunately, this kind of information is rarely included in the dataset, so we need

to perform the task of negation detection ourselves to get the negated sentences needed for our experiments (Jiménez-Zafra et al. 2020).

For English, we adopt the automatic negation detection script from Hossain et al. (2020) that is available on their GitHub. To detect negation, their system has a cue detector that works for single-token cues (e.g. *not, n't, never, no, nothing, nobody,* etc.) and affixal cues (e.g. *impossible, disagree, fearless,* etc.). Following Hossain et al. experiment's setup, we use their script to detect negation cue in the source sentences when we are translating from English.

In the end, we retrieve about 99,000 negated sentences out of 1.1M testing sentences for English-Japanese (9.0%) and about 98,000 negated sentences out of 2M testing sentences for English-Chinese (4.9%). These sets of negated sentences are the ones being modified for our experiments in Section 4.

In our experiments, it is also necessary to be able to detect negation in the target languages (Japanese and Chinese). To detect negation in Japanese, we use MeCab to parse our sentences (Kudo 2005). Using the part-of-speech information from MeCab, we decide a sentence is negated if it satisfies any of these conditions:

1. The sentence has the token ない tagged with 助動詞 as its part-of-speech.
2. The sentence has the token なかっ tagged with 助動詞 as its part-of-speech.
3. The sentence has the token ませ followed by ん, both tagged with 助動詞 as their part-of-speech.

For Chinese, we are detecting negation using the most five common negation cues in Chinese (Tang et al. 2021):

1. 不 *bu*
2. 没 *mei*
3. 无 *wu*
4. 非 *fei*
5. 别 *bie*

If the sentence contains one of these negation cues, we classify it as a negated sentence.

3.3 NMT Models

This section details how we train the NMT models used in our experiments. We train our NMT models using the *Sockeye* toolkit (Domhan et al. 2020). Our data is tokenized and pre-processed using the same toolkit. Past literature has shown that Transformer-based models have fewer errors caused by negation, so we only focus on this architecture for the thesis. For both EN↔JA and EN↔ZH, we train one character-based Transformer model for each direction. The settings follow Tang et al. (2021), but the mini-batch size is decreased from 4,096 to 2,048 due to limited available computational power. Detailed settings are in Table 1.

Table 1: EN-JA/EN-ZH training settings

Neural network depth	6
Number of Transformer Attention Head	8
Learning rate (initial)	2e-04
Embedding and hidden unit size 512	512
Transformer Feed-forward hidden units	2,048
Mini-batch size (token)	2,048
Dropout	0.1
Optimizer	*Adam*
Checkpoint frequency	4,000
Label smoothing	0.1
Early stopping	32

The models are trained using Haverford College's lab computers equipped with one NVIDIA Quadro P5000 each. Learning rate is reduced by a factor of 0.9 if validation perplexity does not improve after 8 consecutive checkpoints.

For each language pair and direction, the best model is chosen based on its perplexity score on our validation set. Table 2 contains our best NMT models' BLEU score and perplexity tested on our validation set and computed using *Sockeye*.

Table 2: BLEU and perplexity scores evaluated on the validation set for our best NMT models

EN → JA		JA → EN		EN → ZH		ZH → EN	
BLEU	Perpl.	BLEU	Perpl.	BLEU	Perpl.	BLEU	Perpl.
0.321	10.8	0.311	6.28	0.471	5.64	0.264	4.11

BLEU score is a metric that compares the machine produced translation to the reference translation (Papineni et al. 2002). If two translations are the same, the BLEU score will be 1.0. If there is no matched content between the two translations, that is a BLEU score of 0.0. BLEU score ranges between these two extremes. In practice, even a human translator will not achieve a BLEU score of 1.0 since two sentences can convey the same meaning while using different words. In Papineni et al. (2002), given a corpus of 500 sentences, a human translator scored 0.3468 when there are four reference translations provided, and they scored 0.2571 when there are only two reference translations. These reference translations are slightly different but valid translations of the source sentence that are available for each sentence in the corpus. With more references being offered, there is a higher chance the way the human translator phrased their translation is found in one of the references, which increases their BLEU score.

Preparing trained NMT models sets the stage for performing experiments. For all of our experiments, we will start from the set of negated sentences from the test set and perform some modifications on it to create the desired testing set for our specific experiment. Then, we run the experiment and report our findings. The specific experiments performed are laid out in the next section.

4 Experiments and Results

All of our experiments in this section manipulate the negation cue in the sentences to some degree. The first experiment tests how the model translates when the negation cue is moved around in the sentence. The second experiment examines how the MT output change when the negation cue is removed completely from the sentence. The third experiment is a follow up of the second experiment, with a focus on sentences with NPI.

4.1 Experiment 1: Negation Cue Shuffle

In our first experiment, we aim to test whether the NMT model can still translate negation if the negation cue is not in the proximity of the negated event. Our hypothesis is that the further the negation cue is moved from its original location, the higher the chance that the machine's output is going to be different from the original translation. Hence, we create distance between the negation cue and its original location by moving the negation cue to every possible location in the sentence.

From our set of negated English sentences described in Section 3, we extract a set of about 3,000 sentences that contain the negation cue *not* from our dataset. Then, we create a test set of about 100,000 sentences by shuffling the location of the negation cue around. This test set contains the original sentences as well as their modified versions, most of the modified sentences are not grammatical in English. A set of sentences from our dataset would look like Example 23:

(23) He is **not** answering the phone.

- **not** He is answering the phone.
- He **not** is answering the phone.
- He is answering **not** the phone.
- He is answering the **not** phone.
- He is answering the phone. **not**

For our experiment, we run this test set through the EN-JA and EN-ZH models. Then, we compare the MT translations of the shuffled sentences with the reference translations of the unshuffled sentences to get our results. We try two different ways to measure our results. First, we define a sentence translation as affected by the modification if the unshuffled sentence's translation differs from the shuffled sentence's translation. Then, our first metric is the proportion of translations affected, and the result of our experiment can be seen in Figure 5. The results confirm our hypothesis; the further the negation cue is moved, the more likely the machine output will be changed.

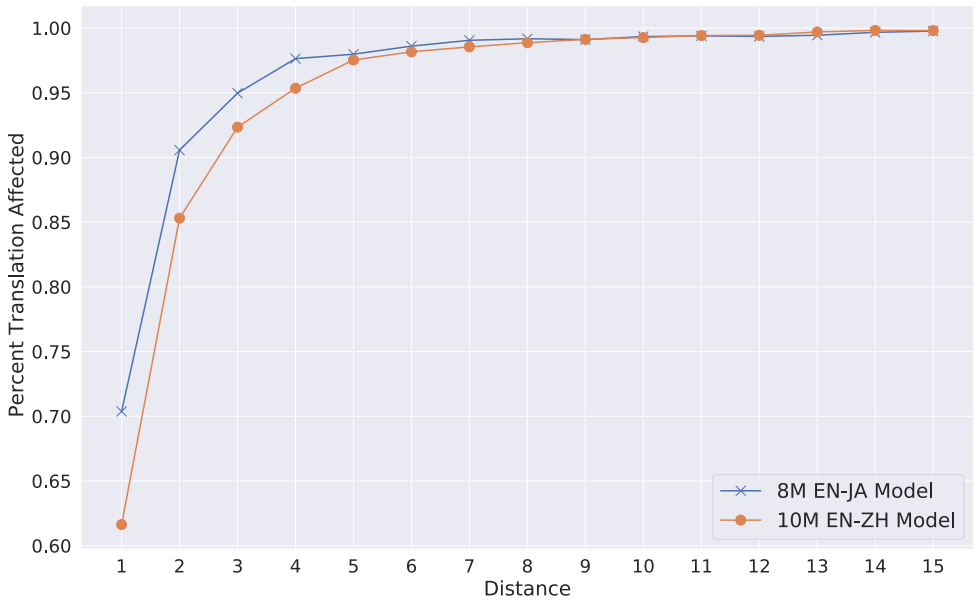

Figure 5: Proportion of translations affected versus distance that the negation cue has moved measured in words. The further the negation cue is moved from its original location, the more likely the MT output will be changed.

However, a change in the translation does not necessarily mean that the negated content has been affected by the modification. Hence, we have a second way to decide if a sentence has been affected by the modification. We assume that if there is a negation cue in the source sentence, then the reference translation will also include a negation cue. If the shuffled sentence's translation no longer contains a negation cue, we count that sentence as having its polarity changed, which is the second metric for our experiments: the proportion of translations with polarity changed. The result using this metric is shown in Figure 6. Once again, we see a confirmation of our hypothesis that the MT output is more likely to change as the negation cue is moved further away from its original location.

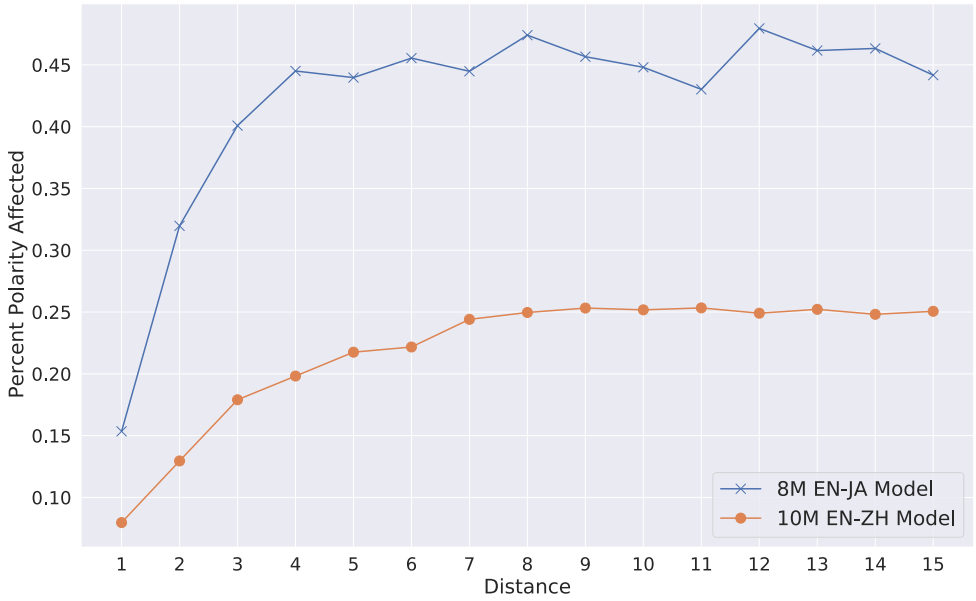

Figure 6: Proportion of polarity changed versus distance that the negation cue has moved measured in words. The further the negation cue is moved away from its original location, the higher the chance that the MT output does not contain a negation cue

4.2 Experiment 2: Negation Cue Removal

In our second experiment, we test whether the machine translation will be affected if we remove the negation cue from the sentence entirely. Our hypothesis is that the MT output will change if we remove the negation cue. In fact, we expect there is a high chance the output sentence will no longer be negated.

To do this, we start from the English negated sentences described in Section 3, we randomly select 50,000 sentences that contain the negation cue *not*, then create a modified test set by removing the negation cue from the sentence. Hence, a sentence such as "I do **not** want to go out today" becomes "I do want to go out today".

Similar to the previous experiment, we run this test set through the EN-JA and the EN-ZH models and compare the MT translations with the reference translations. We use the same two metrics that we described in the previous experiment: proportion of translations

affected and proportion of translations with polarity changes. Using the first metric, our results are as followed:

- For EN-JA, out of a dataset of 50,000 sentence pairs, removing a negation cue affects the translation of 47,566 sentences (95.13%).
- For EN-ZH, out of a dataset of 50,000 sentence pairs, removing a negation cue affects the translation of 49,497 sentences (98.99%).

These numbers confirm our hypothesis that removing a negation cue has a really high change to affect the MT output.

To see whether the polarity of the sentence has changed, we detect the negation cue of the translation in the target language and compare the results between the original sentence's translation and that of the modified sentence. Our results are:

- For EN-JA, out of a dataset of 50,000 sentence pairs, removing a negation cue changes the polarity of 33,637 sentences (67.27%)
- For EN-ZH, out of a dataset of 50,000 sentence pairs, removing a negation cue changes the polarity of 29,291 sentences (58.58%)

From these numbers, we conclude that negation is likely to go away in the machine output if we remove the negation cue from the input, which supports our hypothesis. However, the numbers are lower than we expected. We hypothesize that there are other elements in the negated sentences besides than the negation cue which are informing the model that the sentence is negated. An informal qualitative evaluation was performed to identify shared elements in the sentences whose polarity were not changed. In our evaluation, we found that a high proportion of the sentences whose polarity were not changed in our experiment contain NPI terms. This finding led us to conduct our third experiment where we replicate the second experiment with a focus on negated sentences with NPI.

4.3 Experiment 3: Negation Cue Removal with NPI

In our third experiment, we aim to find out whether the MT output will be affected if we remove the negation cue from a sentence with an NPI term. Our hypothesis is that the translation will be affected, but to a lesser degree than in our second experiment.

Starting from the negated English sentences described in Section 3, we select all sentences that contain the negation cue *not*, then we look for the set of sentences that have

an NPI term from these sentences. The NPI terms we try to look for are: *neither, either, nor, yet, ever, anything, anymore, at all*. Hence, a sentence will only be selected if it satisfies both of these requirements. An example would be "We have **not** seen Eric **at all** today". Finally, our test set is achieved by removing the negation cue from these sentences.

The two metrics that we discussed before are used once again in this experiment: proportion of translations affected and proportion of translations with polarity changes. Using the first metric, our results are as follow:

- For EN-JA, out of a dataset of 4,715 sentence pairs, removing a negation cue affects the translation of 4,645 sentences (98.52%).
- For EN-ZH, out of a dataset of 6,727 sentence pairs, removing a negation cue affects the translation of 6,228 sentences (92.58%).

We see the same results as in our second experiment: the chance that the translation will change after removing a negation cue is really high. However, we see that the EN-JA model is more affected here than in the second experiment, and the EN-ZH model is less affected here than in the second experiment. It could be due to the linguistic difference between how the two languages translate English NPIs, but that is out of the scope of this thesis.

Next, we look at the number of MT output that have their polarity changed after the negation removal is applied. The results are as follow:

- For EN-JA, out of a dataset of 4,715 sentence pairs, removing a negation cue changes the polarity of 2,633 sentences (55.84%).
- For EN-ZH, out of 6,727 sentence pairs, removing a negation cue changes the polarity of only 1,460 sentences (21.70%).

For both languages, we see a drop in the percent of translations with polarity changes from our previous experiment. The percent that the polarity changes drops from 67.27% to 55.84% for English-Japanese, and it drops almost 40% from 58.58% to 21.70% for English-Chinese. These changes are significant since this experiment is almost a replica of the previous experiment; the only thing we change in our setup is we are now looking at a smaller subset of the testing set from the previous setup. Hence, this result suggests that the model is assigning some negative meaning to the NPI terms, and the NPI terms themselves are enough to get negation in the translation. If that is indeed the case, it is likely that the model might introduce negation in sentences due to existence of NPI terms even if that

source sentence is not under negation. We have seen examples of sentences with NPI that are not under negation in Section 2. We suggest that future NMT researchers can incorporate more sentences with NPI to their training dataset (both negated and non-negated) to make sure the model can "learn" a wider usage of NPI.

5 Conclusion

In this thesis, we examine the problem of negation in machine translation. We aimed to gain a better understanding of what NMT models are learning about negation to aid future research in developing better NMT models that can properly make use of and deal with negation information. We pushed the limits of NMT models by presenting them with testing sets where the negation cue has been shuffled around or removed completely. We found that these modifications can greatly affect the output translations and their polarity. In particular, removing the negation cue has the highest effect in changing the polarity of a sentence. We also learned that NMT models seem to be able to assign some negative meaning to NPI terms, which may confuse the model when a NPI is not used under negation.

There were many limits and assumptions made in our research. Our methods of negation detection for English and Chinese are simple, so there are potentially some misclassified negated sentences. We also assume that given negation in the source language, there will be negation in the corresponding reference translation even though this might not necessarily be true. Given more time, we would have replicated the same tests we did with more language pairs. We would also have done more manual evaluation of the machine output. There are many more interesting experiments that we could have done as we are only limiting ourselves to a few experiments using negation cue in this thesis. For example, we could have improved the precision of our evaluation if we were able to systematically classify the sentences that were affected by our experiments (perhaps by parsing them). We could have also used models with target language as English as all experiments are on models with English as the source language. Future work can use the knowledge we learn about NPI terms to augment new training data that could aid NMT models in learning difficult examples (such as NPI terms in positive sentences). We leave behind a testing framework that is scalable and expandable to other tasks for English. We believe that the set up we established in this thesis should be able to support the research directions that we have suggested here as well as future work of other researchers.

References

1964. Affixal negation in english. *WORD* 20(sup1). 21–45. doi:10.1080/00437956.1964.11659838. https://doi.org/10.1080/00437956.1964.11659838.

Achiri-Taboh, Blasius. 2015. A generalized question tag in English: Are English tag questions collapsing? *English Today* 31(1). 48–54. doi:10.1017/S0266078414000546. https://www.cambridge.org/core/journals/english-today/article/generalized-question-tag-in-english/7973B4089691CFE43E12151F6D57B321.

Auerbuch, Mordechai, Tom H. Karson, Benjamin Ben-Ami, Oded Maimon & Lior Rokach. 2004. Context-sensitive medical information retrieval. *Studies in Health Technology and Informatics* 107(Pt 1). 282–286.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR* .

Bayatli, S, S Kurnaz, I Salimzianov, Jonathan North Washington & F M Tyers. 2018. Rule-Based Machine Translation From Kazakh To Turkish 13.

Belletti, Adriana & Luigi Rizzi. 1996. *Parameters and Functional Heads: Essays in Comparative Syntax*. Oxford University Press.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: A Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 257–267. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1025. https://aclanthology.org/D16-1025.

Beyer, Anne, Vivien Macketanz, Aljoscha Burchardt & Philip Williams. 2017. Can Out-of-the-box NMT Beat a Domain-trained Moses on Technical Data? *Proceedings for EAMT 2017 User Studies and Project/Product Descriptions* 41–46.

Blanco, Eduardo & Dan Moldovan. 2011. Some Issues on Detecting Negation from Text. In *Twenty-Fourth International FLAIRS Conference*, Palm Beach, FL, United States.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia & Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, 169–214. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/W17-4717. https://aclanthology.org/W17-4717.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn & Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 272–303. Belgium, Brussels: Association for Computational Linguistics. doi:10.18653/v1/W18-6401. https://aclanthology.org/W18-6401.

Chapman, Wendy W., Will Bridewell, Paul Hanbury, Gregory F. Cooper & Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* 34(5). 301–310. doi: 10.1006/jbin.2001.1029. https://www.sciencedirect.com/science/article/pii/S1532046401910299.

Chiang, David. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics* 33(2). 201–228. doi:10.1162/coli.2007.33.2.201. https://doi.org/10.1162/coli.2007.33.2.201.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. https://aclanthology.org/D14-1179.

Dahl, Östen. 1979. Typology of sentence negation. *Linguistics* 17(1-2). doi: 10.1515/ling.1979.17.1-2.79. https://www.degruyter.com/document/doi/10.1515/ling.1979.17.1-2.79/html.

Ding, Yanzhuo, Yang Liu, Huanbo Luan & Maosong Sun. 2017. Visualizing and Understanding Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1150–1159. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-1106. https://aclanthology.org/P17-1106.

Domhan, Tobias, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber & Kenneth Heafield. 2020. The Sockeye 2 Neural Machine Translation Toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 110–115. Virtual: Association for Machine Translation in the Americas. https://aclanthology.org/2020.amta-research.10.

Fancellu, Federico & Bonnie Webber. 2015. Translating Negation: A Manual Error Analysis.

In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, 2–11. Denver, Colorado: Association for Computational Linguistics. doi:10.3115/v1/W15-1301. https://aclanthology.org/W15-1301.

Goldberg, Yoav. 2017. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies* 10(1). 1–309. doi:10.2200/S00762ED1V01Y201703HLT037. https://www.morganclaypool.com/doi/abs/10.2200/S00762ED1V01Y201703HLT037.

Goldin, Ilya & Wendy Chapman. 2003. Learning to Detect Negation with 'Not' in Medical Texts. *Proc ACM-SIGIR 2003* .

Horn, Laurence R. & Professor and Director of Undergraduate Studies Department of Linguistics Laurence R. Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Horn, Laurence R. & Heinrich Wansing. 2020. Negation. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University spring 2020 edn. https://plato.stanford.edu/archives/spr2020/entries/negation/.

Hossain, Md Mosharaf, Antonios Anastasopoulos, Eduardo Blanco & Alexis Palmer. 2020. It's not a Non-Issue: Negation as a Source of Error in Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3869–3885. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.345. https://aclanthology.org/2020.findings-emnlp.345.

Hurskainen, Arvi & Jörg Tiedemann. 2017. Rule-based Machine translation from English to Finnish. In *Proceedings of the Second Conference on Machine Translation*, 323–329. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/W17-4731. https://aclanthology.org/W17-4731.

Jiménez-Zafra, Salud María, Roser Morante, María Teresa Martín-Valdivia & L. Alfonso Ureña-López. 2020. Corpora Annotated with Negation: An Overview. *Computational Linguistics* 46(1). 1–52. doi:10.1162/coli_a_00371. https://doi.org/10.1162/coli_a_00371.

Kalchbrenner, Nal & Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709. Seattle, Washington, USA: Association for Computational Linguistics. https://aclanthology.org/D13-1176.

Khanna, Tanmai, Jonathan N. Washington, Francis M. Tyers, Sevilay Bayatlı, Daniel G. Swanson, Tommi A. Pirinen, Irene Tang & Hèctor Alòs i Font. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation* doi:10.1007/s10590-021-09260-6. https://doi.org/10.1007/s10590-021-09260-6.

Klima, Edward S. 1964. Negation in English. In *The Structure of Language*, Englewood Cliffs: Prentice Hall.

Koehn, Philipp, Franz J. Och & Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133. https://aclanthology.org/N03-1017.

Kudo, Takumitsu. 2005. MeCab : Yet Another Part-of-Speech and Morphological Analyzer. *undefined* https://www.semanticscholar.org/paper/MeCab-%3A-Yet-Another-Part-of-Speech-and-Analyzer-Kudo/70b849773678010942a0975f2887e527c17cda76.

Lopez, Adam. 2008. Statistical machine translation. *ACM Computing Surveys* 40(3). 8:1–8:49. doi:10.1145/1380584.1380586. https://doi.org/10.1145/1380584.1380586.

Luong, Thang, Hieu Pham & Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal: Association for Computational Linguistics. doi:10.18653/v1/D15-1166. https://aclanthology.org/D15-1166.

Miestamo, Matti. 2007. Negation - An Overview of Typological Research: Negation - An Overview of Typological Research. *Language and Linguistics Compass* 1(5). 552–570. doi:10.1111/j.1749-818X.2007.00026.x. https://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2007.00026.x.

Morante, Roser & Eduardo Blanco. 2012. *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 265–274. Montréal, Canada: Association for Computational Linguistics. https://aclanthology.org/S12-1035.

Morishita, Makoto, Jun Suzuki & Masaaki Nagata. 2020. JParaCrawl: A Large Scale

Web-Based English-Japanese Parallel Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 3603–3609. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.443.

Mutalik, Pradeep G., Aniruddha Deshpande & Prakash M. Nadkarni. 2001. Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *Journal of the American Medical Informatics Association* 8(6). 598–609. doi:10.1136/jamia.2001.0080598. https://doi.org/10.1136/jamia.2001.0080598.

Nyberg, Joacim. 2012. *Negation in Japanese*. http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-78395.

Oard, Douglas W. & Franz Josef Och. 2003. Rapid-Response Machine Translation for Unexpected Languages. In *PROCEEDINGS OF THE MT SUMMIT IX*, https://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=79EEB095701EB5F9047C5EC55F16D1CA?doi=10.1.1.580.3538.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. https://aclanthology.org/P02-1040.

Reitan, Johan, Jørgen Faret, Björn Gambäck & Lars Bungum. 2015. Negation Scope Detection for Twitter Sentiment Analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 99–108. Lisboa, Portugal: Association for Computational Linguistics. doi:10.18653/v1/W15-2914. https://aclanthology.org/W15-2914.

Rokach, Lior, Roni Romano & Oded Maimon. 2008. Negation recognition in medical narrative reports. *Information Retrieval* 11(6). 499–538. doi:10.1007/s10791-008-9061-0. https://doi.org/10.1007/s10791-008-9061-0.

Scott, Bernard & Anabela Barreiro. 2009. OpenLogos MT and the SAL representation language. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, 19–26. Alacant, Spain. https://aclanthology.org/2009.freeopmt-1.5.

Sennrich, Rico. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:*

Volume 2, Short Papers, 376–382. Valencia, Spain: Association for Computational Linguistics. https://aclanthology.org/E17-2060.

Shields, Kenneth. 1997. Positive Anymore in Southeastern Pennsylvania. *American Speech* 72(2). 217–220. doi:10.2307/455794. http://www.jstor.org/stable/455794.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng & Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics. https://aclanthology.org/D13-1170.

Sutskever, Ilya, Oriol Vinyals & Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Swart, Henriëtte. 2010. Negation in a Cross-Linguistic Perspective. In Henriëtte Swart (ed.), *Expression and Interpretation of Negation: An OT Typology* Studies in Natural Language and Linguistic Theory, 1–53. Dordrecht: Springer Netherlands. doi:10.1007/978-90-481-3162-4_1. https://doi.org/10.1007/978-90-481-3162-4_1.

Tang, Gongbo, Philipp Rönchen, Rico Sennrich & Joakim Nivre. 2021. Revisiting Negation in Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 9. 740–755. doi:10.1162/tacl_a_00395. https://doi.org/10.1162/tacl_a_00395.

Vaswani, Ashish, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer & Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. *arXiv:1803.07416 [cs, stat]* http://arxiv.org/abs/1803.07416. Comment: arXiv admin note: text overlap with arXiv:1706.03762.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Weng, Lilian. 2018. Attention? Attention! https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html.

Wilson, Theresa, Janyce Wiebe & Paul Hoffmann. 2005. Recognizing contextual polarity in

phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* HLT '05, 347–354. USA: Association for Computational Linguistics. doi:10.3115/1220575.1220619. https://doi.org/10.3115/1220575.1220619.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes & Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144. http://arxiv.org/abs/1609.08144.

Yang, Shuoheng, Yuxin Wang & Xiaowen Chu. 2020. A Survey of Deep Learning Techniques for Neural Machine Translation. *arXiv:2002.07526 [cs]* http://arxiv.org/abs/2002.07526.

Zhu, Dongqing, Wu Stephen, Masanz James, Ben Carterette & Hongfang Liu. 2013. Using discharge summaries to improve information retrieval in clinical domain: 2013 Cross Language Evaluation Forum Conference, CLEF 2013. *CEUR Workshop Proceedings* 1179. http://www.scopus.com/inward/record.url?scp=84922021333&partnerID=8YFLogxK.

Ziemski, Michal, Marcin Junczys-Dowmunt & Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3530–3534. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1561.