Universitat Politècnica de Catalunya (UPC)
Universitat de Barcelona (UB)
Universitat Rovira i Virgili (URV)

Facultat de Informàtica de Barcelona (FIB)
Facultat de Matemàtiques (UB)
Escola Tècnica Superior d'Enginyeria (URV)

Artificial Intelligence Master Thesis

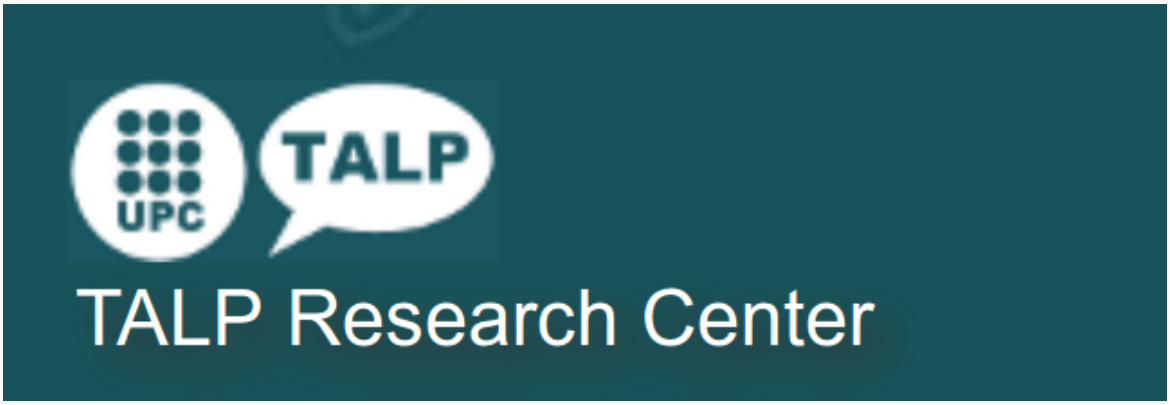
Coverage Model for Character-based Neural Machine Translation

Mohammad Bashir Kazimi

May 2017

Supervised by
Dr. Marta Ruiz Costa-jussà

Co-Supervisor
Dr. Lluís Padró



This master thesis has been carried out under the supervision of Dr. Marta Ruiz Costa-Jussa at the Center for Language and Speech Technologies and Applications (TALP) at Polytechnic University of Catalonia.

I would like to thank Dr. Marta Ruiz Costa-Jussa for giving me the opportunity and guiding me to carry out state of the art research in Neural Machine Translation and providing me with tools and environments for me to be able to do so.

Abstract

In recent years, Neural Machine Translation (NMT) has achieved state-of-the art performance in translating from a language; source language, to another; target language. However, many of the proposed methods use word embedding techniques to represent a sentence in the source or target language. Character embedding techniques for this task has been suggested to represent the words in a sentence better. Moreover, recent NMT models use attention mechanism where the most relevant words in a source sentence are used to generate a target word. The problem with this approach is that while some words are translated multiple times, some other words are not translated. To address this problem, coverage model has been integrated into NMT to keep track of already-translated words and focus on the untranslated ones. In this research, we present a new architecture in which we use character embedding for representing the source and target words, and also use coverage model to make certain that all words are translated. We compared our model with the previous models and our model shows comparable improvements. Our model achieves an improvement of 2.87 BLEU (BiLingual Evaluation Understudy) score over the baseline; attention model, for German-English translation, and 0.34 BLEU score improvement for Catalan-Spanish translation.

Keywords

Machine Learning, Deep Learning, Natural Language Processing, Neural Machine Translation

Contents

List of Figures

List of Tables

1 Introduction

> A satisfactory translation is not always possible, but a good translator is never satisfied with it. It can usually be improved.
>
> Peter Newmark

Machine Translation (MT) is the task of using a software to translate a text from one language to another. Many of the natural languages in the world are quite complex due to the fact that a word could have different meanings based on the context it is used in, and it could also be used in different grammatical categories (e.g. *match* as a *noun* or as a *verb*). Therefore, the main challenge in Machine Translation is the fact that for a correct translation of a word, it is required that many different factors be considered; the grammatical structure, the context, the preceding and succeeding words. Over the years, researchers have developed different methods in order to reduce the amount of manual work and human intervention, and increase the amount of automatic work, and machine dependent translation. The methods in Machine Translation are mainly divided into four categories; *Rule Based Machine Translation (RBMT)*, *Example Based Machine Translation (EBMT)*, *Statistical Machine Translation (SMT)*, and *Neural Machine Translation (NMT)*.

RBMT systems use a a set of language rules developed by linguists in order to translate. The task of translation in a RBMT system involves the analysis of morphological, grammatical, semantic, and syntactic structure of the input words, and generation of syntax and semantics for the target words [1].

In EBMT, translation is performed based on analogy. The machine translates a new text segment by segment based on its similarity with a set of already translated texts. It then combines and puts the sub-parts together as a complete translation of the given text [2]

SMT systems do not need a set of language rules, rather the machine learns from data how to translate using statistical approaches. These systems; however, still need a set of feature functions depending on the input and output words in order to predict a translation [3].

NMT is the most recent approach in Machine Translation which is purely based on a large neural network that is trained to learn and translate text from a source to a target language. Unlike SMT, it does not require pre-designed feature functions and can be trained fully based on training data [4]. NMT has attracted the attention of many researchers in the recent years. The use of neural networks for translation by Baidu [5], the attention from Google's NMT system [6], Facebook's Automatic Text Translation, and many other industries have given the urge for research in NMT a push.

1.1 Motivation for Neural Machine Translation

There are many different languages spoken around the world, and it is important for the nations to be able to communicate with each other at different stages of life. They need to be able to interact for the sake of advancements in technology, politics, business, and they need to be able to negotiate at times of peace and war. This brings up the concept of *translation*.

Translation from one language to another has had a great global impact in many different areas like education, tourism, religion, business, politics, sports, and many more. Translation is essential for growing one's business in a community of different languages, it is important in making improvements in the economy and exchanging cultures among different nations in the form of media, literature, theatres, movies, and tourism.

Since the early ages, translation in all the fields mentioned above has been done manually and by human translators who could speak more than one language. While we are thankful to the translators who have helped nations communicate and stay together, it is clear that manual translation is costly in terms of time and money. With the aims to reduce the cost in translation, the field of *machine translation* emerged.

Machine Translation aims at automating the task of translation between languages, and researchers have been developing different approaches with the aim to increase the automation, and decrease human intervention. Of the four categories of machine translation discussed in the beginning of this chapter, RBMT, EBMT,and SMT have achieved considerable autonomy in translation. The latest approach in Machine Translation; called Neural Machine Translation has outperformed the former approaches and has enabled the machine to be trained on data and learn how to translate using neural networks.

There have been many advancements in NMT since it first emerged. Researchers have mostly based their research on Recurrent Neural Network (RNN) Encoder-Decoder NMT architecture, which produces remarkable results [7, 8, 9]. One of the most recent advancement is the introduction of *Attention Mechanism* which enables the NMT system to translate sentences of different length by focusing on the most relevant parts of the input sequence [10]. The problem with this model is that while some words in the input sequence are translated multiple times, some others are never translated. To address this problem, *Coverage Model* has been introduced into the system to keep track of already translated words and focus on words that are yet to be translated[11]. One of the main issues in the models mentioned is the fact that word embedding[12] has been used for the source and target words to train the model, and hence, due to computational and memory constraints, we are limited to a finite number of words to train the model. Moreover, the models trained to learn a specific word are not able to understand the word if an affix is added to it. With the aims to address this problem, it has been proposed to use character embedding, rather than word embedding [13]. While both models; Coverage Model, and Character-based NMT Model have shown notable improvements over the *Attention Based NMT Model*, there is still room for more improvements to be made.

1.2 Objective: Coverage for Character Neural Machine Translation

This thesis describes an approach to integrate the coverage model into character-based NMT model with the aims to improve the performance of the existing NMT systems and achieve state-of-the-art results in machine translation.

While there are many possible improvements to make, in the scope of this thesis, we have tried to focus on two things. First, the character embedding has been only used for the source or input words, and target words still use word embedding. Second, the coverage model has been integrated into the character based NMT model which intends to solve two main issues in existing NMT systems; the problem of over-translation and under-translation, and the limitation of vocabulary size and translation of different morphological structure of the same word.

1.3 Outline

This thesis is separated into 5 chapters.

Chapter 2 describes the related work in machine translation. Detailed information have been given on the performances and applications of Rule Based Machine Translation , Example Based Machine Translation, Statistical Machine Translation, and Neural Machine Translation.

Chapter 3 describes contribution of this research in NMT, and explains the proposed coverage for character-based NMT. Detailed information has been given on the architecture of the proposed model.

Chapter 4 provides information on the experiments performed and the results obtained. Information on the types of datasets used have been listed, and the automatic evaluation metric has been explained.

Chapter 5 summarizes the thesis and discusses possible future research on the topic at hand.

2 Related Work in Machine Translation

Machine translation has come a long way. Researchers have developed different methods throughout the years to automate the translation as much as possible. In this chapter, the four main methods of machine translation have been discussed; Rule Based Machine Translation, Example Based Machine Translation, Statistical Machine Translation, and Neural Machine Translation.

2.1 Rule Based Machine Translation

Rule Based Machine Translation (RBMT) is one of the first methods in machine translation. It is mainly based on rules and lexicons for both; the input and output languages, produced by linguists. The rules and lexicons explain the syntactic, semantic and morphological information of the languages [14, 1]. There are three types of RBMT systems; *Dictionary Based Machine Translation*, *Transfer Based Machine Translation*, and *Interlingual Machine Translation*.

2.1.1 Dictionary Based Machine Translation

In Dictionary Based Machine Translation, the words in the input language are mapped to the target language directly based on the dictionary look-up between the two languages and some basic grammar rules and morphological analysis. The structure of the system is depicted in figure 2.1.

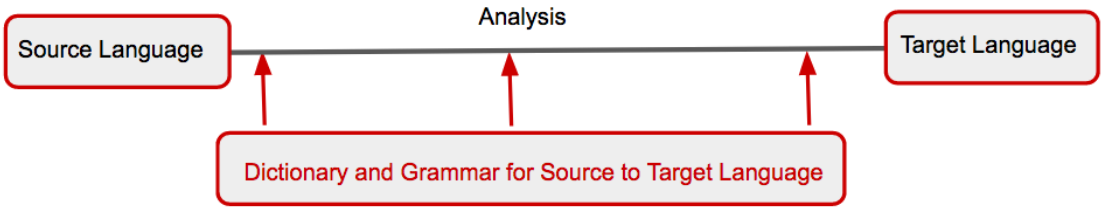


Figure 2.1: Structure of Dictionay Based Machine Translation

2.1.2 Transfer Based Machine Translation

In Transfer Based Machine Translation (TBMT), a morphological and syntactic analyzer is used to produce a representation of the source language from which a target language could be generated. There are three main components in the TBMT systems; *Analysis*, *Transfer*, and *Synthesis* [15].

- **Analysis**: In this phase, the syntactic and semantic structures of the input language is extracted using morphological, syntactic, and semantic analyzer and Part of Speech (POS) taggers.
- **Transfer**: The extracted structures of the input language in the previous phase is then transferred to the the same level representation of the target language using lexical and structural rules.
- **Synthesis**: The first phase is applied to the target language in reverse order in order to generate text in the target language using the representation obtained in the *transfer* phase.

2.1.3 Interlingual Machine Translation

In Interlingual Machine Translation, a single intermediary representation for source and target language is created [16] as opposed to two different representations as in the case of Transfer Based Machine Translation explained in section 2.1.2. One of the main advantages of Interlingual Machine Translation is that it can be used to translate to multiple target languages as there is no need to transfer the representation into each target language individually.

In short, the three types of RBMT could be illustrated in the so called Vauquois Triangle [17] in figure 2.2. As observed in the Vauquois Triangle in figure 2.2, going up the triangle increases the quality of translation and reduces the error, but it increases the amount of analysis and synthesis which refer to producing pre-transfer representation for the source words, and generating target translation from the post-transfer representation, respectively [16].

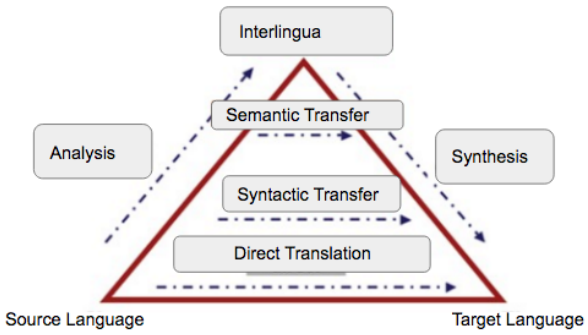


Figure 2.2: Vauquois Traingle for RBMT

2.2 Example Based Machine Translation

Example Based Machine Translation (EBMT) is a method that translates by using a bilingual corpus of texts as a knowledge base [18]. EBMT involves three steps; *Matching*, *Alignment*,

and *Recombination*. Matching refers to the task of finding similar translations for different fragments of the text in a database of existing translations. Alignment is the process of identifying and selecting translation fragments that could be reused, and then comes recombination which involves putting the selected translation fragments together in a meaningful manner [19]. EBMT steps could be integrated into the Vauquois Triangle as depicted in figure 2.3. The EBMT steps are shown in oval shapes. The *analysis* stage for the original Vauquois Triangle has been replaced with *matching*, *alignment* could replace the *transfer* step, and *recombination* is used instead of *synthesis* in the original triangle. Additionally, an *exact match* would mean a *direct translation*.

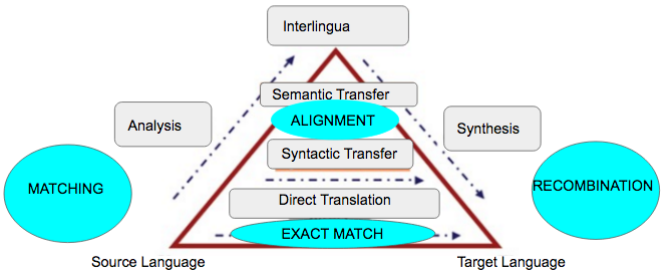


Figure 2.3: Vauquois Traingle for EBMT

2.3 Statistical Machine Translation

In Statistical Machine Translation (SMT), we aim to have the machine translate by learning from data, rather than a set of rules. Given a parallel corpus $D = (x^1, y^1), ..., (x^n, y^n)$ translated by humans, SMT algorithms automatically learn to translate [20]. For a given sentence, there could be many possible translations, therefore the goal is to find the translation with the highest probability, and there are different methods to achieve this. Two main types of SMT models; probabilistic models, and Log-linear models, are explained in the following subsections.

2.3.1 Probabilistic Models

The first type of SMT models is the *probabilistic model* where a translation y' is chosen where the probability $p(y|x)$ is the highest [21].

$$y' = arg \max_{y} p(y)p(x|y) \tag{2.1}$$

where the term $p(x|y)$ is referred to as *translation model*; which shows the confidence of y being a translation for x, and $p(y)$ is called the *language model*; which shows the fluency of the target sentence y. Equation 2.1 also summarizes another main component in SMT, usually referred to as *decoding* or *searching*, which is the process of finding an optimal approach to find y such that the result of the product $p(y)p(x|y)$ is the highest [22]. Each of the aforementioned components are briefly explained as follows.

- **Translation Model**:
  Translation models; the term $p(x|y)$ in equation 2.1, are based on translation probabilities of pairs of words or pairs of phrases. To determine the correspondence of a word or a phrase in one language to that of another, it is important to know the *alignment* between the languages. Alignment model; proposed by Brown et al. [23], indicates which word in the source language is the target word a translation of. Therefore, the main formula for the translation model is as follows.

$$p(x|y) = \sum_{a} p(x, a|y) \tag{2.2}$$

  where a denotes the alignment between the two languages. [21].

- **Language Model**:
  A language model is a probability distribution on a sequence of words. It makes sure the translation generated is fluent. The main formula for the language model is as follows.

$$p(y) = \prod_{i=1}^{n} p(y_i) \tag{2.3}$$

  The probabilities are usually calculated using N-grams. N-grams indicate that the probability that a word comes next in the sequence depends on the probability of previous $N-1$ words [23].

- **Decoding/Searching**:
  This part of the task in SMT deals with the fact that there are many possible translations, and the goal here is to find the best translation possible. This is done by building partial alignments and translating, and then keeping the alignments with the highest probabilities [23].

2.3.2 Log-linear Models

In Log-linear models, the goal is to find a set of parameters θ that maximizes the following loglikelihood function.

$$\mathcal{L}(\theta, D) = \sum_{i} \log p(y^i|x^i, \theta) \tag{2.4}$$

for $i = 1, ..., n$ where

$$\log p(y|x, \theta) = \sum_{i} \theta_i f_f(x, y) + C(\theta) \tag{2.5}$$

where C is the normalization constant, and θ_i is a set of coefficients to order the set of feature functions f_i that helps in estimating the best translation [3, 24, 25].

The set of coefficients θ_i in equation 2.5 is learnt by algorithms in machine learning, but the set of feature functions f_i in the same equation is what has to be predesigned, and this is where most of the research in SMT has been focused on [26].

2.4 Neural Machine Translation

As mentioned before, in Machine Translation, the goal is to maximize the conditional probability of a target sentence $\mathbf{y}$, given a source sentence $\mathbf{x}$. Research suggests that this probability distribution can be learnt using Neural Networks leading to the evolution of Neural Machine Translation. In this section, NMT and some of the main methods have been explained.

NMT is the most recent approach in Machine Translation which is purely based on a large neural network that is trained to learn and translate text from a source to a target language. Unlike SMT, it does not require pre-designed feature functions and can be trained fully based on training data [4].

NMT has achieved state of the art results in Machine Translation, and the first NMT models used the RNN Encoder Decoder architecture [8, 7]. In this approach, the input sentence is encoded by the encoder into a fixed-length vector h_T using an RNN, and the fixed-length vector is decoded by the decoder; another RNN, to generate the output sentence. Word-embedding [12] has been used for representation of the source and target words. One of the main issues in the simple RNN Encoder Decoder models is that the encoded vector is of a fixed length, and it cannot represent long sentences completely. To address this issue, attention model has been introduced to the simple RNN Encoder Decoder model [10]. Attention model uses a bi-directional RNN to store the information into memory cells instead of a fixed-length vector. Then a neural network called *attention mechanism* uses the input information in the memory cells and the information on the previously translated words by the decoder in order to focus on the most relevant input words for the translation of a specific output word.

In the models mentioned above, word embedding has been used for word representations. While it performs well, it limits the NMT model to a fixed-size vocabulary. Since the models are trained using a large set of vocabularies, and vocabulary is always limited, the models face problems with rare and out-of-vocabulary (OOV) words [13, 27, 28]. Many of the words could have various morphological forms, and could have affixes, and word-embedding models would not be able to distinguish a word it has been trained with if an affix is added to it or a different morphological form of the word is used [29]. To address these problem, it has been proposed to use character embedding rather than word embedding, resulting into fully character-level NMT system [28], character based NMT models that use character embedding only for source language [13, 30], and character-level decoders that use character embedding for the target language [29]. Two additional advantages of character embedding for NMT are its usability for multilingual translation, which is the result of its ability to identify shared morphological structures among languages, and also the fact that as opposed to word embedding models, no text segmentation is required, which enables the system to learn the mapping from a sequence of characters to an overall meaning representation automatically [28]. It has been proved that character NMT models produce improved performance over the attention model [13, 27, 28, 29].

Another issue with the models mentioned earlier; specifically in the case of the attention model, is that they do not track the translation history and hence, some words are translated

many times while some other words are not translated at all or translated falsely. To address this problem, different models of *coverage* have been proposed to track translation history, avoid translating words multiple times and focus on words that are not yet translated [11, 31]. The authors claim to have achieved better results as compared to the attention based model.

3 Coverage for Character Neural Machine Translation

3.1 Contribution

While researchers have based their models on the RNN Encoder Decoder [8, 7] and the attention model [10], to produce character models [13, 27, 30, 28] and coverage models [11, 31] and have achieved state of the art results, both the models address one of the two issues in the earlier models separately. The character model addresses the problem of rare, OOV words, and words with various morphological structures, and uses character embedding rather than word embedding, and the coverage model addresses the problem where some words are translated multiple times while some of the rest are never or falsely translated. In this research, we propose to jointly address the two important problems in the traditional NMT models and introduce *coverage* to *character* model to achieve state of the art results in NMT. The character embedding has only been used for the source words, and the target words still use word embedding.

3.2 Architecture of the Proposed NMT Model

The backbone of the proposed architecture is still the the attention model proposed by Bahdanau et al.[10] with the word embedding in the input language replaced by the character embedding as proposed by Costa-jussà and Fonollosa [13]. Thus, first of all, the encoder computes the input sentence summary $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$ which is the concatenation of $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ for $t = 1, 2, ..., T$. $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the hidden states for the forward and backward RNN encoder reading the information from the input sentence in the forward and reverse order, respectively. The hidden states are calculated as follows.

$$\overrightarrow{h}_t = \overrightarrow{f}(x_t, \overrightarrow{h}_{t-1}) \tag{3.1}$$

$$\overleftarrow{h}_t = \overleftarrow{f}(x_t, \overleftarrow{h}_{t-1}) \tag{3.2}$$

where $\overrightarrow{h}_{t-1}$ and $\overleftarrow{h}_{t-1}$ denote the previous hidden states for the forward and backward RNN, $\overrightarrow{f}$ and $\overleftarrow{f}$ are recurrent activation functions, and x_t is the embedding representation for the t-th input word. In the attention model, x_t is the simple word embedding representation of the word in the source language, but in our case, x_t is the character embedding calculated as proposed by Costa-jussà and Fonollosa [13] as follows.

First of all, each source word k is represented with a matrix C^k which is a sequence of vectors representing the character embedding for each character in the source word k. Then a number n of convolution filters H of length w, with w ranging between 1 to 7, is applied to C^k in order to obtain a feature map f^k for the source word k as follows.

$$f^k[i] = tanh(\langle C^k[*, i : i + w - 1], H\rangle + b) \tag{3.3}$$

where b is the bias and i is the i-th element in the feature map. For each convolution filter H, the output with the maximum value is selected by a max pooling layer in order to capture the most important feature.

$$y_H^k = \max_i f^k[i] \tag{3.4}$$

The concatenation of these output values for the n convolution filters H; $\mathbf{y}^k = [y_{H1}^k, y_{H2}^k, ..., y_{Hn}^k]$, is the representation for the source word k. Addition of two highway network layers has been proved to give a better representation of the source words [30]. A layer of the highway network performs as follows.

$$x_t = \mathbf{t} \odot g(W_H \mathbf{y}^k + b_H) + (1 - \mathbf{t}) \odot \mathbf{y}^k \tag{3.5}$$

where g is a nonlinear function, $\mathbf{t} = \sigma(W_T \mathbf{y}^k + b_T)$ is the *transform gate*, (1 - $\mathbf{t}$) is the *carry gate*, and x_t is the character embedding that is used in equations 3.1 and 3.2.

The decoder then generates a summary $z_{T'}$of the target sentence as follows.

$$z_{t'} = f(z_{t'-1}, y_{t'-1}, s_{t'}) \tag{3.6}$$

where $s_{t'}$ is the representation for the source words calculated as follows.

$$s_{t'} = \sum_{t=1}^{T} \alpha_{t't} h_t \tag{3.7}$$

where h_t is calculated by the encoder as explained earlier, and $\alpha_{t't}$ is computed as follows.

$$\alpha_{t't} = \frac{\exp(e_{t't})}{\sum_{k=1}^{T} \exp(e_{t'k})} \tag{3.8}$$

and

$$e_{t't} = a(z_{t'-1}, h_t, C_{t'-1t}) \tag{3.9}$$

is called the attention mechanism or the *alignment model* which scores how relevant the input word at position t is to the output word at position t', $C_{t'-1t}$ is the previous coverage and coverage model proposed by Tu et al. [11] is calculated as follows.

$$C_{t't} = f(C_{t'-1t}, \alpha_{t't}, h_t, z_{t'-1}) \tag{3.10}$$

Then, the output sentence is generated by computing the conditional distribution over all possible translation.

$$\log p(y|x) = \sum p(y_{t'}|y_{<t'}, x) \tag{3.11}$$

where y and x are the output and input sentences, respectively, and $y_{t'}$ is the t'-th word in the sentence y. Each conditional probability term $p(y_{t'}|y_{<t'}, x)$ is computed using a feed forward neural network as follows.

$$p(y_{t'}|y_{<t'}, x) = softmax(g(y_{t'-1}, z_{t'}, s_{t'})) \tag{3.12}$$

where g is a nonlinear function, $z_{t'}$ is the decoding state from equation 3.6, and $s_{t'}$ is the context vector from equation 3.7.

The overall architecture of the proposed model is illustrated in figures 3.1, 3.2, 3.3. Figure 3.1 illustrates the character based word embedding model which takes as input the embeddings for each character in the source word x_t, and outputs a final word level representation of it. The output is then fed to the encoder; depicted in figure 3.2 which outputs a context vector s'_t based on the attention mechanism and coverage model. The context vector s'_t is then fed to the decoder illustrated in figure 3.3 which generates a target translation.

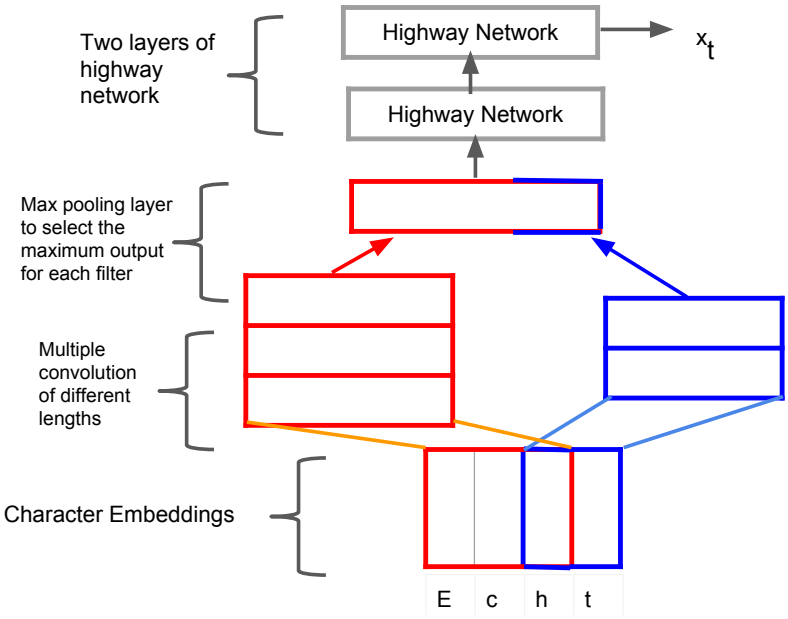


Figure 3.1: Character based word embedding

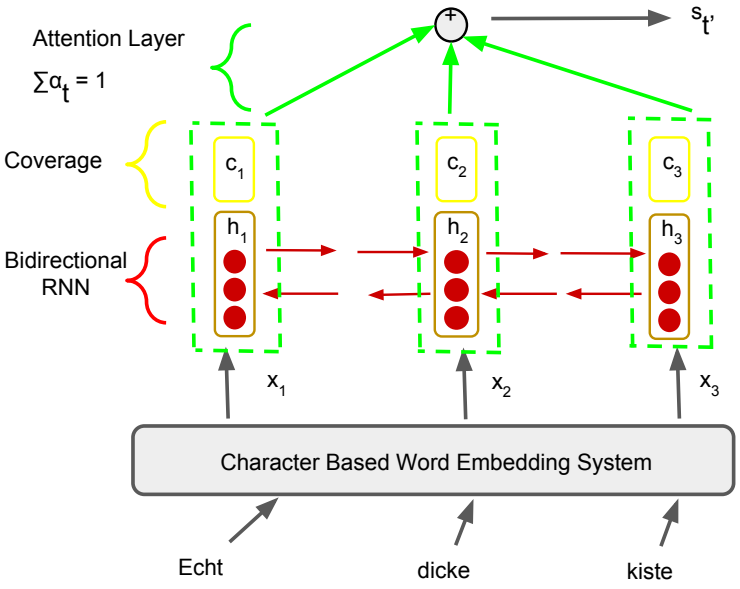


Figure 3.2: Encoder with coverage & alignment

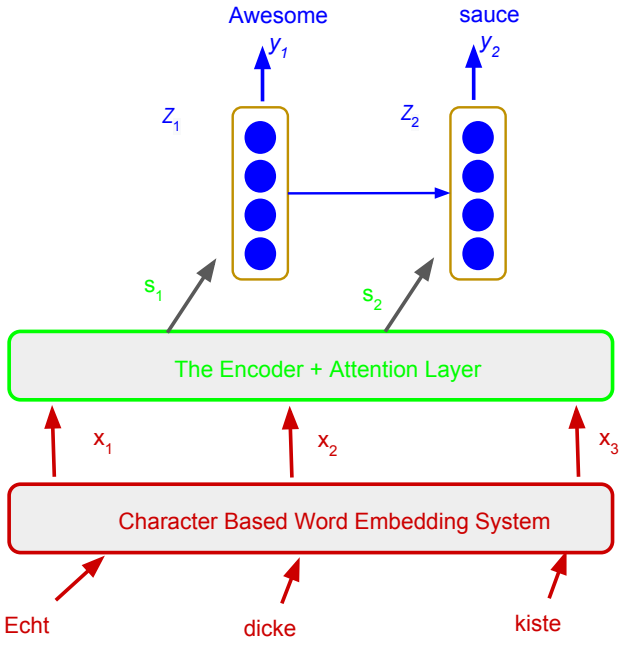


Figure 3.3: The decoder

4 Experiments

In order to evaluate the performance of our model, experiments on the same data sets have been performed using the baseline; attention model by Bahdanau et al.[10], the coverage model by Tu et al.[11], and the proposed model in this study; coverage for character model. This section has been divided into two subsections. Subsection 4.1 explains the data sets used and the preprocessing performed on the data, and subsection 4.2 elaborates on the evaluation method and the results obtained.

4.1 Data Set

The NMT model proposed in this research has been trained and tested on two data sets. The first one is a parallel corpus of German-English taken from Workshop for Machine Translation (WMT'17)[1], and the second one is that of Catalan-Spanish. The Catalan-Spanish data set has been kindly provided by Costa-jussà [32] and contains a set of paper edition over 10 years of a bilingual Catalan newspaper , El Periodico, in addition to a corpus of medical domain provided by Universal-Doctor project[2]. As a preprocessing task, the data sets have been tokenized and a dictionary of 90 thousand most frequent words have been prepared for training the system. The information on both of the data sets are summarized in tables 4.1, and 4.2.

Language	Set	# of sentences	# of Words	# of Vocabs
De	Train	5.6M	121.3M	2M
	Dev	3k	73k	14k
	Test	3k	63	13k
En	Train	5.6M	129.4M	1M
	Dev	3k	73k	10k
	Test	3k	65k	9k

Table 4.1: German-English Dataset

4.2 Evaluation and Results

To evaluate the quality of the translation by our model and compare it to the baseline model by Bahdanau et al.[10] , and the coverage model by Tu et al.[11] based on the experiments performed, we use the BiLingual Evaluation Understudy (BLEU) evaluation method proposed by Papineni et al.[33]. The evaluation method; BLEU, has been explained in subsection 4.2.1, and the results have been listed in subsection 4.2.2.

[1] http://www.statmt.org/wmt17/translation-task.html
[2] http://www.universaldoctor.com/

Language	Set	# of Sentences	# of Words	# of Vocabs
Ca	Train	6.5M	179.9M	713k
	Dev	2.2k	60k	11k
	Test	2.2k	60k	12k
Es	Train	6.5M	165.2M	737k
	Dev	2.2k	55k	8k
	Test	2.2k	56k	8k

Table 4.2: Spanish-Catalan Dataset

4.2.1 BLEU Score Evaluation

To evaluate the quality of translation manually is time consuming and expensive. Therefore, among many automatic evaluation metrics, BLEU metric is used [33, 3]. The BLEU evaluation method scores a translation on a scale of 0 to 1, but in the research society, it is usually reported as percentage. The closer to 1 the BLEU score is, the more similar the translation is to the actual translation, hence the better the quality of the automatic translation. In other words, the higher number of overlaps between an automatic translation and an actual translation; usually referred to as reference translation, the higher the BLEU score. To compute the similarity between two translations, the precision and recall are calculated [3]. *Precision* is calculated as the ratio of matching words and the total number of words in the translation, while *recall* is the ratio of matching words and the total number of words in the *reference sentence*. For example for the two given automatic translations and the corresponding reference sentence below, the precision and recall are listed in table 4.3.

Translation 1: Israeli officials responsibility of airport safety
Translation 2: airport security Israeli officials are responsible
Reference: Israeli officials are responsible for airport security

Metric	**Translation 1**	**Translation 2**
Precision	3/6 (50%)	6/6 (100%)
Recall	3/7 (43%)	6/7 (85.7 %)

Table 4.3: Precision and Recall for two sample translations

As observed in the results of precision and recall and recall in table 4.3, even though the automatic translation 1 gives a correctly ordered translation, the words translated do not exactly match. The automatic translation 2 gets a higher precision and recall because the translated words match exactly. The problem with evaluating only based on precision and recall is that the word orderings are not taken into account. BLEU method alleviates this problem by using N-gram overlap between the translation and the reference. The precision is calculated using N-grams of size $1-4$ which results in scoring sequential matching of words higher than those of unordered. It also penalizes for brevity to a translation of smaller length than the reference [3]. The BLEU method is formulated as follows.

$$BLEU = \min(1, \frac{translation - length}{reference - length})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} \tag{4.1}$$

Calculating the BLEU scores for the two translations are listed in table 4.4. The matching N-grams are shown in colors as follows.

Translation 1: Israeli officials responsibility of airport safety
Translation 2: airport security Israeli officials are responsible
Reference: Israeli officials are responsible for airport security

Metric	**Translation 1**	**Translation 2**
Precision (1-gram)	3/6	6/6
Precision (2-gram)	1/5	4/5
Precision (3-gram)	0/4	2/4
Precision (4-gram)	0/3	1/3
Brevity Penalty	6/7	6/7
BLEU	0 %	52 %

Table 4.4: BLEU Score for two sample translations

4.2.2 Results

The result of the BLEU evaluation metric for the translation of the test data sets produced by the models are listed in tables 4.5 and 4.6.

Model	**BLEU score**
Baseline (Attention)	80.20
Coverage	80.54
Our model (Coverage+Char)	**81.45**

Table 4.5: BLEU score for the Catalan-Spanish Data Set

Model	**BLEU score**
Baseline (Attention)	18.81
Coverage	21.68
Our model (Coverage+Char)	**22.41**

Table 4.6: BLEU score for the German-English Data Set

As it can be observed from tables 4.5 and 4.6, using coverage combined with character embedding improves the performance of the NMT system for both of the data sets. There is an improvement of 0.34 BLEU score over the baseline for the Catalan-Spanish data set and an improvement of 2.87 over the baseline for the German-English data set. Since Catalan and Spanish are more similar languages while German and English are not as similar, it is yet to be investigated why the improvement over the baseline is not as much for similar languages as it is for a more different language pair.
In order to address the two main issues with the baseline model, we have listed some example translations as follows.

1. **Catalan-Spanish Example**:
   Source: l ' equip de Tabárez va començar com un remolí , però es va anar apaivagant amb el pas dels minuts .
   Baseline: empezó como un torbellino , pero se fue UNK con el paso de los minutos .
   Coverage: el equipo de UNK empezó como un remolino , pero fue UNK con el paso de los minutos .
   Character+Coverage: el equipo de UNK comenzó como un torbellino , pero fue UNK con el paso de los minutos .
   Target: el equipo de Tabárez comenzó como un torbellino , pero fue secándose con el paso de los minutos .

2. **German-English Example**:
   Source: als Teil des Gipfeltreffens trafen sich auch im Vorfeld die Verteidigungsminister des Blocks zur Verabschiedung des Maßnahmenplans für das Jahr 2013 , worin der Dialog und Konsens in Bezug auf die Verteidigung der Region gestärkt werden soll .
   Baseline: as part of the competition , the defence minister of the block met in the run-up to the adoption of the plan for the year 2013 , which aims to strengthen dialogue and consensus in relation to the defense of the region .
   Coverage: as part of the summit , the Defence Minister met previously in the run-up to the adoption of the five-year plan for the year 2013 , which aims to strengthen dialogue and consensus in relation to the defence of the region .
   Character+Coverage: as part of the summit , the Defence Ministers were previously in the run-up to the adoption of the five-year plan for the year 2013 , which aims to strengthen dialogue and consensus on defence in the region .
   Target: also , as part of the summit , the bloc 's foreign defence ministers met in advance to approve the 2013 Action Plan , which seeks to strengthen dialogue and consensus on defence in the region .

3. **German-English Example**:
   Source: terroristische Angriffe gab es auf beiden Seiten .
   Baseline:there were armed attacks on both sides .
   Coverage: there were armed attacks on both sides .
   Character+Coverage: there were terrorist attacks on both sides .
   Target: terrorist attacks occurred on both sides .

Example 1 shows that the baseline does not *cover* all the words and *l ' equip de Tabárez* has not been translated, and it could be observed the coverage model, and consequently our model handles this problem. Moreover, while *remolino* is a fair translation, *torbellino* is a more adequate translation in this context, which is captured well through character embedding. Example 2 points out the fact that the baseline and the coverage model alone translates the word *Verteidigungsminister* as *Defence Minister* while our model; since it is using character embedding, handles it well, resolves the semantic ambiguity, and understands that the usage is in plural. Finally, example 3 also depicts the fact that using character embedding, you get a better representation and hence our model translates the sentence closer to the actual translation, even though the other two models also translate adequately.

5 Conclusions

5.1 Summary

The recent model; attention, proposed by Bahdanau et al.[10] tackles the problem of fixed-length encoding vector in the RNN Encoder Decoder model used by Sutskever et al.[8] and Cho et al. [7]. It gives NMT the ability to translate sentences of any length. It faces two main problems; the rare, and OOV words problem along with problems with different possible morphemes for a single word, *and* the problem of over-translation and under-translation. The character models which use character embedding [13, 30, 27, 28] and the coverage models, which keep track of translation history [11, 31] have individually addressed both the issues, respectively.

In this research, *coverage* has been introduced to the *character* model which aims to address the main issues mentioned earlier altogether, and improve the state of the art in NMT. The data sets in tables 4.1 and 4.2 have been experimented and the results have been listed in tables 4.5 and 4.6. It is clearly observed that the model in this study outperforms the previous models and achieves state of the art performance in NMT.

5.2 Future Work

In this research, character embedding has been used only for the source language, and the words in the target language are still represented by word-embedding. In addition, we have used the character embedding proposed by Costa-jussà and Fonollosa[13] since first of all, the character embedding models by Yang et. al[27] and Lee et. al[28] were under research while this research was being carried out. Secondly, in [13], words have been used for the attention model while [27] and [28] use characters, and it only makes sense to use *coverage* for words and not characters. Further research is needed to check how the model performs if the target language also uses character embedding. Moreover, there is still more than enough room for improvements in Neural Machine Translation, and further research needs to be done to find out more factors that could affect the performance of the systems.

List of Acronyms

NMT	Neural Machine Translation
SMT	Statistical Machine Translation
RNN	Recurrent Neural Networks
TALP	Center for Language and Speech Technologies and Applications
MT	Machine Translation
RBMT	Rule Based Machine Translation
EBMT	Example Based Machine Translation
BLEU	BiLingual Evaluation Understudy
TBMT	Transfer Based Machine Translation
POS	Parts Of Speech
OOV	Out Of Vocabulary
WMT	Workshop for Machine Translation

Bibliography

[1] Marta R, Mireia Farrús, José B Marino, and José AR Fonollosa. Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems. *Computing and informatics*, 31(2):245–270, 2012.

[2] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle.

[3] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[4] Minh-Thang Luong and Christopher D Manning. Stanford neural machine translation systems for spoken language domains. 2015.

[5] HE Zhongjun. Baidu translate: research and products. *ACL-IJCNLP 2015*, page 61, 2015.

[6] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[9] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *EMNLP*, volume 3, page 413, 2013.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[11] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Coverage-based neural machine translation. *CoRR*, abs/1601.04811, 2016.

[12] Amit Mandelbaum and Adi Shalev. Word embeddings and their use in sentence classification tasks. *CoRR*, abs/1610.08229, 2016.

[13] Marta R. Costa-Jussà and José A. R. Fonollosa. Character-based neural machine translation. *CoRR*, abs/1603.00810, 2016.

[14] A-L Lagarda, Vicente Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Diaz-de Liano. Statistical post-editing of a rule-based machine translation system. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 217–220. Association for Computational Linguistics, 2009.

[15] Jayashree Nair, K Amrutha Krishnan, and R Deetha. An efficient english to hindi machine translation system using hybrid mechanism. In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, pages 2109–2113. IEEE, 2016.

[16] Dorr Bonnie J. Interlingual machine translation: A parameterized approach. *Artificial Intelligence*, 63(1 & 2):429–492, 1993.

[17] Bernard Vauquois. A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *IFIP Congress-68*, pages 254–260, Edinburgh, 1968.

[18] Davide Turcato and Fred Popowich. What is example-based machine translation? In *Recent advances in example-based machine translation*, pages 59–81. Springer, 2003.

[19] Harold Somers. Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157, 1999.

[20] Adam Lopez. Statistical machine translation. *ACM Comput. Surv.*, 40(3):8:1–8:49, August 2008.

[21] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[22] Lambert Mathias. *Statistical machine translation and automatic speech recognition under uncertainty*. PhD thesis, Citeseer, 2007.

[23] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85, June 1990.

[24] Christophe Servan and Simon Petitrenaud. Calculation of phrase probabilities for statistical machine translation by using belief functions. In *The 24th International Conference on Computational Linguistics (COLING 2012)*, 2012.

[25] Holger Schwenk. Continuous space language models. *Computer Speech & Language*, 21(3):492–518, 2007.

[26] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander M Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, et al. A smorgasbord of features for statistical machine translation.

[27] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. A character-aware encoder for neural machine translation. In *COLING*, 2016.

[28] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017, 2016.

[29] Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. A character-level decoder without explicit segmentation for neural machine translation. *CoRR*, abs/1603.06147, 2016.

[30] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. *arXiv preprint arXiv:1508.06615*, 2015.

[31] Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. A coverage embedding model for neural machine translation. *CoRR*, abs/1605.03148, 2016.

[32] Marta R. Costa-jussà. Why catalan-spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. In *A: Workshop on NLP for Similar Languages, Varieties and Dialects. "Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)"*, pages 55–62, 2017.

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.