Coverage for Character Based Neural Machine Translation

Técnicas de Cobertura y Caracteres integrados en la Traducción Automática Basada en Aprendizaje Profundo

M.Bashir Kazimi **Marta R. Costa-jussà**
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
mohammad.bashir.kazimi@est.fib.upc.edu
marta.ruiz@upc.edu

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 languages, and also use coverage model to make certain that all words are translated. Experiments were performed to compare our model with coverage and character model and the results show that our model performs better than the other two models.
Keywords: Machine Learning, Deep Learning, Natural Language Processing, Neural Machine Translation

Resumen: En los últimos años, la traducción automática basada en el aprendizaje profundo ha conseguido resultados estado del arte. Sin embargo, muchos de los métodos propuestos utilizan espacios de palabras embebidos para representar una oración en el idioma de origen y destino y esto genera muchos problemas a nivel de cobertura de vocabulario. Avances recientes en la traducción automática basada en aprendizaje profundo incluyen la utilización de caracteres que permite reducir las palabras fuera de vocabulario. Por otro lado, la mayoría de algoritmos de traducción automática basada en aprendizaje profundo usan mecanismos de atención donde las palabras más relevantes en de la oración fuente se utilizan para generar la traducción destino. El problema con este enfoque es que mientras algunas palabras se traducen varias veces, algunas otras palabras no se traducen. Para abordar este problema, usamos el modelo de cobertura que realiza un seguimiento de las palabras ya traducidas y se centra en las no traducidas. En este trabajo, presentamos una nueva arquitectura en la que utilizamos la incorporación de caracteres para representar el lenguaje origen, y también usamos el modelo de cobertura para asegurarnos que la frase origen se traduce en su totalidad. Presentamos experimentos para comparar nuestro modelo que integra el modelo de cobertura y modelo de caracteres. Los resultados muestran que nuestro modelo se comporta mejor que los otros dos modelos.
Palabras clave: Aprendizaje automático, Aprendizaje profundo, Procesado del Lenguaje Natural, Traducción Automática

1 Introduction

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 MT 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. One of the main methods in MT is Statistical Machine Translation (SMT) which is a data-driven approach and produces translation based on probabilities between the source and target language. The goal is to maximize the conditional probability $p(y|x)$ of a target sentence y given the equivalent source sentence x based on a set of predesigned features (Koehn, 2009).

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 (Luong and Manning, 2015). NMT has attracted the attention of many researchers in the recent years. The use of neural networks for translation by Baidu (Zhongjun, 2015), the attention from Google's NMT system (Wu et al., 2016), Facebook's Automatic Text Translation, and many other industries has given the urge for research in NMT a push.

In this research, we study the state of the art in NMT, and propose a novel approach by combining two of the most recent models in NMT; coverage (Tu et al., 2016) and character model (Costa-jussà and Fonollosa, 2016), in the hopes to achieve state of the art results. The rest of the paper has been organized as follows. Section 2 studies the related work in NMT, section 3 explains the proposed model in this study and points out the contribution of the research, section 4 explains the experiments performed and the results obtained, and finally section 5 summarizes the thesis and points out possible future research.

2 Related Work

NMT has achieved state of the art results in MT, and the first NMT models used the Recurrent Neural Network (RNN) Encoder Decoder architecture (Sutskever, Vinyals, and Le, 2014; Cho et al., 2014). In this approach, the input sentence is encoded by the encoder into a fixed-length vector h_T using a recurrent neural network (RNN), and the fixed-length vector is decoded by the decoder; another RNN, to generate the output sentence. Word-embedding (Mandelbaum and Shalev, 2016) 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 (Bahdanau, Cho, and Bengio, 2014). Attention model uses a bi-directional recurrent neural network to store the information into memory cells instead of a fixed-length vector. Then a small 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 (Yang et al., 2016; Lee, Cho, and Hofmann, 2016). 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 (Chung, Cho, and Bengio, 2016). To address these problems, it has been proposed to use character embedding rather than word embedding, resulting into fully character-level NMT system (Lee, Cho, and Hofmann, 2016), character based NMT models that use character embedding

only for source language (Costa-jussà and Fonollosa, 2016; Kim et al., 2015), and character-level decoders that use character embedding for the target language (Chung, Cho, and Bengio, 2016). 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 (Lee, Cho, and Hofmann, 2016). It has been proved that character NMT models produce improved performance over the attention model (Costa-jussà and Fonollosa, 2016; Yang et al., 2016; Lee, Cho, and Hofmann, 2016; Chung, Cho, and Bengio, 2016).

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 (Tu et al., 2016; Mi et al., 2016). The authors claim to have achieved better results as compared to the attention based model.

3 Coverage for Character Based Neural Machine Translation

3.1 Contribution

While researchers have based their models on the RNN Encoder Decoder (Sutskever, Vinyals, and Le, 2014; Cho et al., 2014) and the attention model (Bahdanau, Cho, and Bengio, 2014), to produce character models (Costa-jussà and Fonollosa, 2016; Yang et al., 2016; Kim et al., 2015; Lee, Cho, and Hofmann, 2016) and coverage models (Tu et al., 2016; Mi et al., 2016) 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 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 uses 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. (Bahdanau, Cho, and Bengio, 2014) with the word embedding in the input language replaced by the character embedding as proposed by Costa-jussà and Fonollosa (Costa-jussà and Fonollosa, 2016). 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.

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

$$\overleftarrow{h}_t = \overleftarrow{f}(x_t, \overleftarrow{h}_{t-1}) \quad (2)$$

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 (Costa-jussà and Fonollosa, 2016) and explained 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, n convolution filters H of length w, with w ranging between 1 to 7, are applied to C^k in order to obtain a feature map f^k for the source word k as follows.

$$f^k[i] = tanh(\langle C^k[*, i : i+w-1], H\rangle + b) \quad (3)$$

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.

$$y_H^k = \max_i f^k[i] \tag{4}$$

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 (Kim et al., 2015). A layer of the highway network performs as follows.

$$x_t = \mathbf{t} \odot g(W_H \mathbf{y}^k + b_H) + (1 - \mathbf{t}) \odot \mathbf{y}^k \tag{5}$$

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 1 and 2.

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

$$z_{t'} = f(z_{t'-1}, y_{t'-1}, s_{t'}) \tag{6}$$

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

$$s_{t'} = \sum_{t=1}^{T} \alpha_{t't} h_t \tag{7}$$

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

$$\alpha_{t't} = \frac{\exp(e_{t't})}{\sum_{k=1}^{T} \exp(e_{t'k})} \tag{8}$$

and

$$e_{t't} = a(z_{t'-1}, h_t, C_{t'-1t}) \tag{9}$$

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. (Tu et al., 2016) is calculated as follows.

$$C_{t't} = f(C_{t'-1t}, \alpha_{t't}, h_t, z_{t'-1}) \tag{10}$$

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

$$\log p(y|x) = \sum p(y_{t'}|y_{<t'}, x) \tag{11}$$

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.

$$p(y_{t'}|y_{<t'}, x) = softmax(g(y_{t'-1}, z_{t'}, s_{t'})) \tag{12}$$

where g is a nonlinear function, $z_{t'}$ is the decoding state from equation 6, and $s_{t'}$ is the context vector from equation 7

The overall architecture of the proposed model is illustrated in figures 1, 2, 3. Figure 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 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 which generates a target translation.

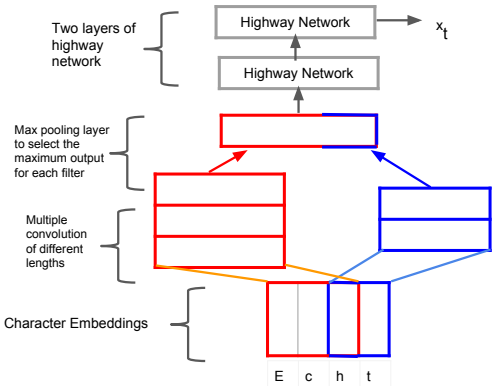

Figure 1: Character based word embedding

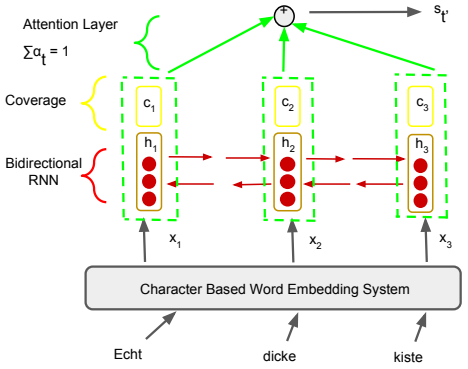

Figure 2: Encoder with coverage & alignment

4 *Experiments*

In order to evaluate the performance of our model, experiments on the same data set has

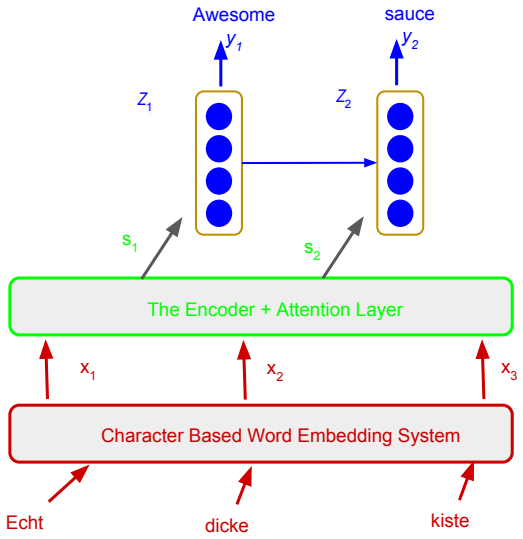

Figure 3: The decoder

been performed using the character model by Costa-jussà and Fonollosa (Costa-jussà and Fonollosa, 2016), the coverage model by Tu et al.(Tu et al., 2016), and finally the proposed model in this study; coverage for character model. This section has been divided into two subsections. Subsection 4.1 explains the data set used and the preprocessing performed on the data, and subsection 4.2 elaborates on the evaluation method and the results obtained.

4.1 Data

The data set used for this experiment is kindly provided by Costa-jussà (Costa-jussà, 2017) and includes a subset of a larger data set which includes 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[1]. As a preprocessing task, the data set has been tokenized and a dictionary of 10 thousand most frequent words have been prepared for training the system. Detailed information about the data set is listed in table 1

Language	Set	# of Sentences	# of Words	# of Vocabs
Ca	Train	83.5k	2.9M	83.5k
	Dev	1k	27.7k	6.9k
	Test	1k	27k	6.7k
Es	Train	100k	2.7M	90k
	Dev	1k	25k	7k
	Test	1k	24.9k	7k

Table 1: Spanish-Catalan Dataset Statistics

[1]http://www.universaldoctor.com/

Model	BLEU Score
Character	53.30
Coverage	53.76
Character+Coverage	54.87

Table 2: BLEU Scores for the NMT Models

4.2 Evaluation and Results

To evaluate the model, the BLEU (BiLingual Evaluation Understudy) evaluation method proposed by Papineni et al.(Papineni et al., 2002) has been used. The main idea behind this evaluation method is that the closer to a human translation the machine translation is, the better the model performs. The result of the experiments performed on the data set mentioned in section 4.1 have been listed in table 2.

As observed in table 2, the proposed model outperforms the other models and achieves state of the art performance. The main motivation for this study is to try to address two main issues in the attention model. First, the attention model uses word embedding for language representation, and thus it suffers from the rare, OOV word problems, and problems with identifying different morphemes added to a word. The second issue is that even though the attention model focuses on most relevant part of the input sentence in order to translate and generate an output sentence, it does not keep track of already-translated words, which leads to multiple translation of some words while the rest are never or falsely translated. The two issues were individually tackled with characters models (Costa-jussà and Fonollosa, 2016; Yang et al., 2016; Lee, Cho, and Hofmann, 2016; Kim et al., 2015), and coverage models (Tu et al., 2016; Mi et al., 2016), respectively. In this research, we tried to improve the state of the art and introduce coverage for character model in NMT. The experiment performed on the data set shown in table 1 clearly shows that our model outperforms earlier models, as shown in table 2. To understand the contribution of our proposed model and see how the combination of character and coverage model compliments the two models and sometimes performs better than both of the models, we list in table 3 some manual analysis on sample translation by the models tested.

1	Src	dos regidors es presenten als comicis.
	Tgt	dos concejales se pre-sentan alos comicios.
	Ch	dos ediles se presentan en los comicios.
	Cov	dos concejales sepresentan a los comicios.
	Ch+Cov	dos concejales se presentan a los comicios.
2	Src	la falta de públic l ' ha condemnat a mort en una zona clau de l ' oci barceloní que , pel que es veu , té més poder de convocatòria.
	Tgt	la falta de público lo ha condenado a muerte en una zona clave del ocio barcelonés que , por lo que se ve , tiene más poder de convocatoria.
	Ch	Palma alguna de público le ha condenado a muerte en una zona clave del ocio barcelonés que , por lo que se ve , tiene más poder de convocatoria.
	Cov	a falta de público al ha condenado a muerte en una zona clave del ocio barcelonés que , por lo que se ve , tiene mejor de convocatoria.
	Ch+Cov	la falta de público le ha condenado a muerte en una zona clave del ocio barcelonés que , por el que se ve , tiene además poder de convocatoria.a.
3	Src	Una firma austríaca va voler vendre sang amb sida a l ' Àsia..
	Tgt	Una firma austriaca quiso vender sangre con sida en Asia.
	Ch	Una firma UNK quiso vender sangre con sida en Asia.
	Cov	Una seguidores UNK quiso vender sangre con sida en Asia.
	Ch+Cov	~~Unafirma~~ UNK quiso vender sangre consida en Asia
4	Src	com a conseqüència de la progressiva reducció dels marges.
	Tgt	como consecuencia de la progresiva reducción de los márgenes.
	Ch	a consecuencia de la UNK reducción de los márgenes.
	Cov	a consecuencia de la UNK reducción de los márgenes.
	Ch+Cov	como consecuencia de la progresiva reducción de los márgenes.
5	Src	... requereix un esforç que involucri " departaments de Turisme , Joventut i Educació , i també de coordinació en l ' àmbit europeu " ...
	Tgt	... requiere un esfuerzo que involucre " a departamentos de Turismo , Juventud y Educación , y también de coordinación a nivel europeo ...
	Ch	... requiere un esfuerzo que UNK " departamentos de Turismo , Joventut y Educación , y que tiene que UNK en el ámbito europeo ...
	Cov	... requiere un esfuerzo que UNK " departamentos de Turismo , Joventut y Educación , y que tiene que UNK en el ámbito europeo ...
	Ch+Cov	... requiere un esfuerzo que UNK " departamentos de Turismo , Juventud y Educación , y también de coordinación en el ámbito europeo " ...

Table 3: Manual Analysis. Src and Tgt represent Source and Target sentences, Ch, Cov, and Ch+Cov represent translation by Character, Coverage, and the proposed ~~moedl~~, respectively. In example 1 and 2, the proposed model behaves like the coverage model, in example 3, it behaves like the character model, and examples 4 and 5, it performs better than both of the other models.

5 Summary

The recent model; attention, proposed by Bahdanau et al.(Bahdanau, Cho, and Bengio, 2014) tackles the problem of fixed-length encoding vector in the RNN Encoder Decoder model used by Sutskever et al.(Sutskever, Vinyals, and Le, 2014) and Cho et al. (Cho et al., 2014). It gives NMT the ability to be able 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 ~~(Costa-jussà and Fonollosa, 2016; Kim et al., 2015; Yang et al., 2016; Lee, Cho, and Hofmann, 2016)~~ and the coverage models, which keep track of translation history ~~(Tu et al., 2016; Mi et al., 2016)~~ 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 corpus shown in table 1 has

been experimented and the results have been listed in table 2. It is clearly observed that the model in this study outperforms the previous models and achieves state of the art performance in NMT.

As in the case of character model, the character embedding has been used only for the source language, and the target language is still limited to word embedding. further research is required in order to study how character embedding added for the target language impacts the performance of the model, and it is left to investigate more factors affecting the performance of NMT systems.

Acknowledgements

This work is supported by Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional, through contract TEC2015-69266-P (MINECO/FEDER, UE) and the postdoctoral senior grant Ramón y Cajal.

References

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chung, J., K. Cho, and Y. Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *CoRR*, abs/1603.06147.

Costa-jussà, M. R. 2017. 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.

Costa-jussà, M. R. and J. A. R. Fonollosa. 2016. Character-based neural machine translation. *CoRR*, abs/1603.00810.

Kim, Y., Y. Jernite, D. Sontag, and A. M. Rush. 2015. Character-aware neural language models. *arXiv preprint arXiv:1508.06615*.

Koehn, P. 2009. *Statistical machine translation*. Cambridge University Press.

Lee, J., K. Cho, and T. Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Luong, M.-T. and C. D. Manning. 2015. Stanford neural machine translation systems for spoken language domains.

Mandelbaum, A. and A. Shalev. 2016. Word embeddings and their use in sentence classification tasks. *CoRR*, abs/1610.08229.

Mi, H., B. Sankaran, Z. Wang, and A. Ittycheriah. 2016. A coverage embedding model for neural machine translation. *CoRR*, abs/1605.03148.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. 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.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tu, Z., Z. Lu, Y. Liu, X. Liu, and H. Li. 2016. Coverage-based neural machine translation. *CoRR*, abs/1601.04811.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, Z., W. Chen, F. Wang, and B. Xu. 2016. A character-aware encoder for neural machine translation. In *COLING*.

Zhongjun, H. 2015. Baidu translate: research and products. *ACL-IJCNLP 2015*, page 61.