Journal of Artificial Intelligence Research 61 (2018) 947-974 Submitted 09/17; published 04/18

From Feature to Paradigm: Deep Learning in Machine Translation

Marta R. Costa-jussà MARTA.RUIZ@UPC.EDU
TALP Research Center,
Universitat Politècnica de Catalunya,
08034 Barcelona

Abstract

In the last years, deep learning algorithms have highly revolutionized several areas including speech, image and natural language processing. The specific field of Machine Translation (MT) has not remained invariant. Integration of deep learning in MT varies from re-modeling existing features into standard statistical systems to the development of a new architecture. Among the different neural networks, research works use feed-forward neural networks, recurrent neural networks and the encoder-decoder schema. These architectures are able to tackle challenges as having low-resources or morphology variations.

This manuscript focuses on describing how these neural networks have been integrated to enhance different aspects and models from statistical MT, including language modeling, word alignment, translation, reordering, and rescoring. Then, we report the new neural MT approach together with a description of the foundational related works and recent approaches on using subword, characters and training with multilingual languages, among others. Finally, we include an analysis of the corresponding challenges and future work in using deep learning in MT.

1. Introduction

The information society is continuously evolving towards multilinguality: e.g. different languages other than English are gainining more and more importance in the web; and strong societies, like the European, are and will continue to be multilingual. Different languages, domains, and language styles are combined as potential sources of information. In such a context, Machine Translation (MT), which is the task of automatically translating a text from a source language into a target language, is gaining more and more relevance. Both industry and academy are strongly investigating in the field which is progressing at an incredible speed. This progress may be directly attached to the introduction of deep learning. Basically, deep learning is the evolution of neural networks composed by multiple-layered models, and neural networks are machine learning systems capable of learning a task by training from examples and without requiring being explicitly programmed for that task. MT is just one of the applications where deep learning has succeeded recently. Although neural networks were proposed for MT in late nineties (Forcada & Ñeco, 1997; Castaño & Casacuberta, 1997), and have been integrated in different parts of statistical MT since 2006, it was not until 2013 and 2014 that first competitive neural MT systems were proposed (Kalchbrenner & Blunsom, 2013; Sutskever, Vinyals, & Le, 2014; Cho, van Merrienboer, Gulcehre, Bahdanau, Bougares, Schwenk, & Bengio, 2014b), and in 2015, that neural MT reached the state-of-the-art (Bahdanau, Cho, & Bengio, 2015).

1.1 MT Approaches before Deep Learning

MT has been approached mainly following a rule-based or corpus-based strategy. Rule-based MT systems date back early 70s with the initiatives of Systran (Philipson, 2017) or EUROTRA (Maegaard, 1989). The idea behind rule-based approaches is that transformation from source to target is done by means of performing an analysis of the source text, transfering (with hand-crafted rules) this new source representation to a target representation and generating the final target text.

Corpus-based approaches learn from large amounts of text. One popular and successful approach is the statistical one and, in particular, the phrase-based MT system (Koehn, Och, & Marcu, 2003). This statistical approach benefits from being trained on large datasets. Normally, statistical MT uses parallel texts at the level of sentences, it uses co-occurrences to extract a bilingual dictionary, and finally, it uses monolingual text to compute a language model which estimates the most fluent translation text in the target language.

The main limitations of statistical MT are that it relies on parallel corpora. In rule-based MT, limitations are that it requires many linguistic resources, and a lot of human expert time. There is a considerable amount of research trying to hybridize these two approaches (Costa-jussà, 2015).

Another type of MT approaches, popular in the decade of the 80s, were interlingua-based, which focus on finding a universal representation of all languages. However, these approaches have fallen into disuse because it is very challenging and expensive to manually find a universal representation for all languages.

1.2 MT and Deep Learning

Recent appearence of new training and optimization algorithms for neural networks, i.e. deep learning techniques (Hinton, Osindero, & Teh, 2006; Bengio, 2009; Goodfellow, Bengio, & Courville, 2016), the availability of large quantities of data and the increase of computational power capacity have benefited the introduction of deep learning in MT.

Deep learning is about learning representations with multiple levels of abstraction and complexity (Bengio, 2009). There has been a lot of excitement around deep learning because of the achieved breakthroughs, e.g. the automatic extraction of composition of images from lines to faces (Lee, Grosse, Ranganath, & Ng, 2009), the ImageNet classification (Krizhevsky, Sutskever, & Hinton, 2012) or reducing the error rate in speech recognition by around 10% (Graves, Mohamed, & Hinton, 2013). There has been a lot of recent activity from the scientific community in using deep learning in MT refelected in, for example, an explosion in the number of works in relevant conferences from 2014 up to date.

This manuscript presents an overview from the early stages of how deep learning has started as a feature function in statistical MT (Schwenk, Costa-Jussà, & Fonollosa, 2006) to become an entire new paradigm, which has achieved state-of-the-art results (Jean, Cho, Memisevic, & Bengio, 2015) within one-year of development.

1.3 Manuscript's Contribution and Organization

This manuscript focuses on collecting and describing research done in introducing deep learning in MT. Differently from previous surveys (Zhang & Zong, 2015), we do not detail deep learning techniques, instead we just provide a briefly description to make this manuscript self-contained. We center our attention on:

- Overviewing the integration of deep learning in MT and reporting the MT aspects that have been improved with the different types of neural networks;
- Detailing the new neural MT architecture, citing its foundational works as well as discussing recent advances that face challenging aspects encountered in the neural MT architecture;
- Depicting an analysis of strengths and weaknesses of deep learning in MT.

The rest of this manuscript is organized as follows. Section 2 briefly defines the types of neural networks that have been mostly used to enhance MT including feed-forward, recurrent neural networks and the encoder-decoder schema. Then, section 3 classifies how deep learning has been introduced in MT through enhancements in translation and language modeling, word alignment, reordering and rescoring. Section 4 reviews the emergent deep learning architecture for MT: neural MT together with the description of recent advances in the area. Section 5 underlines the main challenges of using deep learning in MT by depicting the main strengths and weaknesses. Finally, the last section brings discussion about the role and future of deep learning in the MT field and points out further work directions.

2. Brief Description of Neural Networks Types

This section briefly describes the neural networks types used in MT, but for further details on each of these neural network types refer to complete studies (Goodfellow et al., 2016).

Neural networks can be defined as a type of statistical learning algorithms used to estimate functions that can have a large number of inputs. Neural networks are organised in layers, including an input and an output layers and, in between, considering one or several hidden layers. Each layer is composed by neurons which is the elementary unit of the network. Each neuron receives one or several inputs for which the neuron performs a weighted sum of the inputs and pass it through a non-linear function (activation) to produce the output.

Among the main advantages, these algorithms:

- Extract abstractions from data and, currently, they are providing the best performance in multiple domains and applications that learn from data;
- Do not require feature engineering since the algorithms learn them from data;
- Can easily be adapted to new problems, i.e. deep learning architectures which are applied to one particular application can be useful to other applications (Kaiser, Gomez, Shazeer, Vaswani, Parmar, Jones, & Uszkoreit, 2017).

While among the main drawbacks, these algorithms:

- Train complex models which require large amount of data, a huge number of parameters and a high computational cost;
- Require the challenging task of determining the right architecture or topology of the network adequate to each task.

This section introduces the main neural network types that have been employed in MT, either to complement statistical approaches or to be part of the new neural MT approach. Mainly they are feed-forward and recurrent neural networks and the encoder-decoder schema. All of them can be extended with multiple hidden layers of units between the input and the output layers constituting deep neural networks (DNNs). These networks in the MT task are mostly trained under a supervised learning framework, where algorithms learn from a huge collection of examples (i.e. parallel texts at the level of sentences). As a consequence, one of the main big issues in deep learning architectures has been dealing with large vocabularies. Training speed goes down when the vocabulary increases. Main reason for this is that deep learning architectures used for MT (or related tasks) normally require a softmax function to generate the final probabilities of output words. Computing this softmax function involves taking the sum of scores for all words in the vocabulary, which is not feasible for large vocabularies. As we will see in sections 3 and 4, there have been mainly two directions to face this problem: either reducing the vocabulary size by using characters or subwords units instead of words; or use approximations and self-normalization to reduce computation time, but not model size.

2.1 Feed-Forward Neural Networks

Feed-forward Neural Networks (FNNs), see Figure 1 (A), connect the inputs through hidden nodes to the outputs without loops. Basically, these networks can be classified into single and multi-layer perceptrons. The former consist of a function that maps its input to an output value. The latter consist of several fully connected layers in a directed graph. Each layer has several nodes, and each node is a neuron with a non-linear activation function.

2.1.1 Convolutional Neural Networks (CNNs)

A popular type of FNNs are CNNs, whose connectivity pattern between their neurons is inspired by the overlapping of the individual neurons of the animal cortex (LeCun & Bengio, 1998). A convolution operation is the mathematical way to describe this connectivity pattern. Mentioning that there are 3 basic properties of CNNs on top of FNNs which are: local connectivity (only adjacent neurons are connected), parameter sharing (replicated units share the same parameterization) and maxpooling units which is a form on subsampling.

2.2 Recurrent Neural Network

Recurrent Neural Networks (RNNs), see Figure 1 (B), are another class of neural networks. The main characteristic is that connections between units form a directed cycle, which generates an internal state with dynamic temporal behavior. FNNs typically rely on a fixed-size context window making the Markov assumption that a word only depends on n previous words. On the other hand, RNNs are able to use the internal memory to get rid of this Markov assumption and condition on all previous words, which is highly relevant in

language modeling and MT. There are different types of RNNs and, in this manuscript, we focus on the most used in first neural MT systems. A much further detailed explanation on this type of neural networks can be found in the work of Goodfellow et al. (2016).

2.2.1 Long Short Term Memory (LSTM)

The LSTM network (Hochreiter & Schmidhuber, 1997) has the directed cycle structure with a different structure in the repeating cycle. The repeating cycle has three neural network gates (input, memory/forget and output) which allow to discard or keep information solving the problem RNN face on the vanishing gradient (Hochreiter, 1991; Bengio, Simard, & Frasconi, 1994). Intuitevely, the vanishing gradient problem may appear when using gradient-based and backpropagation methods. When training weights with these algorithms, these weights are updated using the gradient of the error function. At this point, and for RNNs, the chain rule is applied for the entire history of the sequence and applying this many times may cause the gradients to tend to zero (specially, when using activation functions as *tanh* or *sigmoid*). Since LSTM where initially proposed, they have successfully been used in a wide range of sequence applications (Graves, 2013).

2.2.2 Gated RNN

An alternative to LSTMs are the Gated RNN (Chung, Gulcehre, Cho, & Bengio, 2015; Cho et al., 2014b), which main difference is that instead of having three gates as LSTMs, GRUs have two gates (reset and update). GRUs have less parameters to train compared to LSTMs which may help in training faster and generalizing better with less data (Chung, Gulcehre, Cho, & Bengio, 2014).

2.2.3 Bi-directional RNN

Bi-directional RNN uses a finite sequence to label each sequence's element using the past and the future context (Schuster & Paliwal, 1997). In this case, predictions are computed by combining the RNN output of processing the sequence from left to right and the RNN output of processing the sequence from right to left.

2.3 Encoder-Decoder

This type of architecture has been inspired in autoencoders which try to predict their own input (Goodfellow et al., 2016). Encoder-decoder architecture generalizes the idea of autoencoders allowing for having different input and output data. The encoder-decoder architecture aims at learning a representation (encoding) of input data, and decodes this representation while minimizing the amount of error for recovering the output data. The main purpose of the internal representation is a dimensionality reduction capable of extracting relevant features from the dataset. A schematic representation is shown in Figure 1 (C).

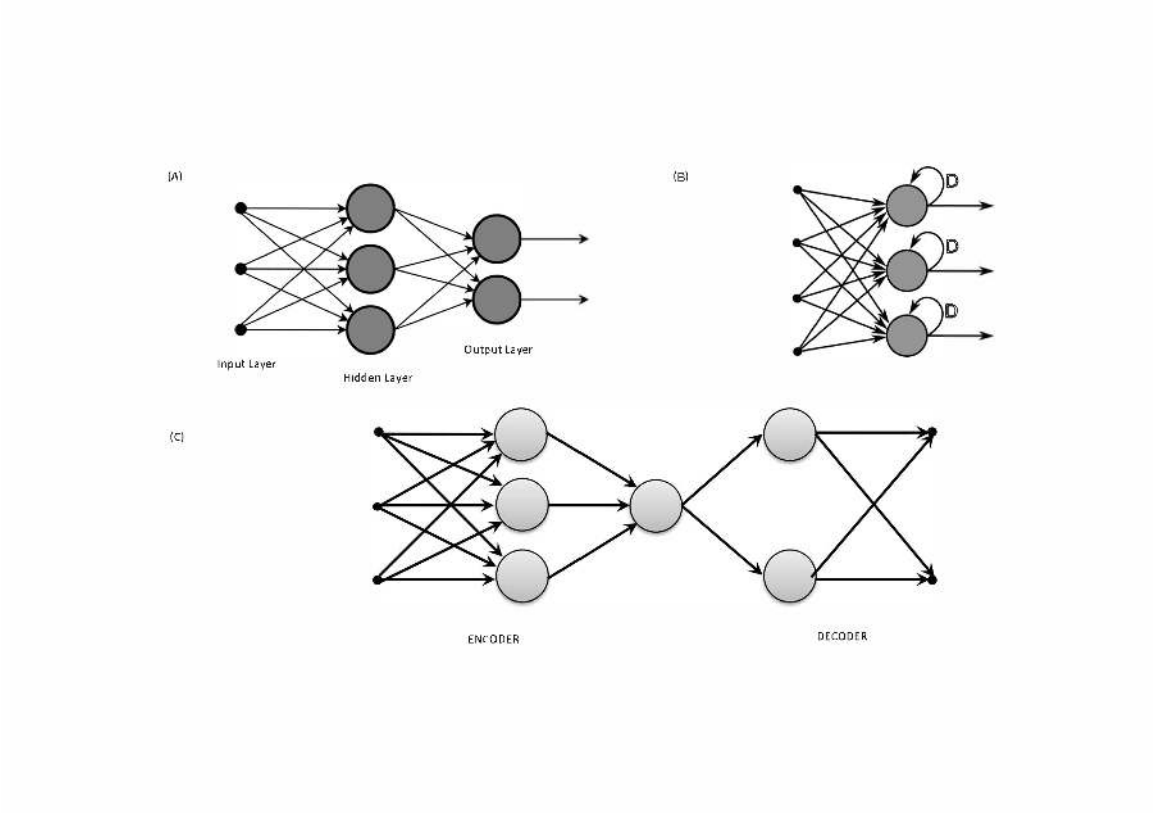

Figure 1: Feed-forward Neural Network (A); Recurrent Neural Network (B) and Encoder-Decoder (C) schemas.

3. Statistical MT Systems with Deep Learning

Statistical MT (Lopez, 2008) focuses on finding the most probable target text given a source text. In training, this approach uses parallel and monolingual corpus to extract several models (i.e. translation, language and reordering) that in inference are combined in a beam-search decoding. The translation model is a bilingual dictionary of words and sequences that are scored based on co-ocurrences extracted from a parallel corpus, which has been automatically aligned at the level of words, usually using IBM models (Brown, Pietra, Pietra, & Mercer, 1993). The aim of this model is to provide accurate translations of source sequences. The language model is a monolingual dictionary of words and sequences and its objective is to model the probability of target sequences. Finally, the reordering model scores changes of order between source and target. These models are assigned different weights through an optimization procedure which maximises the translation quality, measured by an automatic metric, e.g. BLEU (Papineni, Roukos, Ward, & Zhu, 2002). In addition, these systems may benefit from rescoring n-best lists to deal with the decoding search errors.

Statistical MT has been enhanced with neural networks at different levels. This section covers how neural networks have improved language modeling; word alignment; translation

modeling; reordering; and rescoring. Figure 2 shows the proportion of works for each of these models (note the source of this statistic are the cited papers by current manuscript).

As a summary of this section, Table 1 shows the main related works on introducing deep learning on standard statistical MT.

3.1 Language Modeling

Language modeling is the task of scoring sequences of words. Approaches with neural networks to language modeling have a long history (Elman, 1990; Bengio, Ducharme, Vincent, & Janvin, 2003). This subsection covers the way neural networks has enhanced monolingual language modeling in statistical MT systems, while bilingual language modeling is reported in the following section 3.3.

Schwenk et al. (2006) use a continuous space language model inspired in the classical approaches (Bengio et al., 2003) to improve the N-gram-based MT system (Mariño, Banchs, Crego, de Gispert, Lambert, Fonollosa, & Costa-jussà, 2006). The neural network is a multilayer perceptron, trained as a classifier, with input projection, hidden and output layers. The projection layer of input words represents the distributed encoding of input words and uses a linear activation function. The hidden layer uses a hyperbolic tangent as the activation function and the output layer is a softmax layer. In this case, the continuous space language model is used in rescoring for both a phrase-based and N-gram-based MT system.

Following, Mikolov et al. (2012) use recurrent neural network language modeling to rescore n-best lists. Regarding the problem of limited vocabulary (as mentioned in section 2), Hu et al. (2014) use two RNN-based minimum translation unit models. Authors focus on addressing the issues of data sparsity and limited context sizes by leveraging continuous representations and the unbounded history of the recurrent network. They frame the problem as a sequence modeling task over minimal units and, furthermore, model a minimum translation unit as bag-of-words. Results show the approach is complementary to a very strong RNN-based language model based solely on target words.

Niehues et al. (2012) propose a simplified neural language model based on restricted boltzmann machines and integrate it in MT decoding. This fully integration leads to further improvements than n-best rescoring. Also integrated in MT decoding, Vaswani et al. (2013) use a FNN architecture using rectified linear units and noise-contrastive estimation, which does not require repeated summations over the whole vocabulary and enables to train neural networks on larger datasets.

Devlin et al. (2014) introduce a neural network joint language model which augments an n-gram language model (Stolcke, 2002) with an m-word source window. Moreover, the network can be self-normalized, which allows the increase in vocabulary size. Auli et al. (2014) show how the RNN language model can be optimized using BLEU as a criterion. In addition, authors efficiently integrate the RNN in decoding. Given that decoding speed using n-gram LM is still state-of-the-art, some approaches calculate neural probabilities in the n-gram format (Wang, Utiyama, Goto, Sumita, Zhao, & Lu, 2013). In the framework of joint translation and reordering which consists in training sequences that encode translation and word reordering information at the same time, Guta et al. (2015) compare the performance of n-gram, feedforward and recurrent neural networks directly in translation.

In addition, and just recently, there have been studies that investigate the effectiveness of several variations of the neural language model or information. Luong et al. (2015a) investigate whether deep neural language models with three or four layers outperform those with fewer layers in terms of translation quality, and they reach a significant improvement when jointly condition on both the source and target contexts. Costa-jussà and Fonollosa (2016) use character-aware language modeling (Kim, Jernite, Sontag, & Rush, 2016) to rescoring of n-best lists outperforming results on an enhanced phrase-based system.

3.2 Word Alignment

Word alignment is a key task in statistical MT systems since it identifies word-by-word relationships given pairs of sentences that are corresponding translations. IBM models (Brown et al., 1993) are one of the most popular probabilistic formulation of this problem and have been using successfully with the GIZA++ implementation (Och & Ney, 2003).

Recently, there have appeared some approaches that use deep learning to perform this task. Yang et al. (2013) use the methodology of DNN used in speech recognition to learn to extract lexical translation information. The model integrates a multi-layer neural network into a Hidden Markov Model (HMM) framework, from where they extract context dependent lexical translation. The model is trained on a bilingual corpus and use monolingual data to pre-train word embeddings. They improve the quality over classical IBM models. Tamura et al. (2014) improve the previous work by using a RNN which allows for unlimited alignment history.

3.3 Translation Modeling

Given work in the literature, we distinguish studies done on bilingual translation models, on phrase-based models and on syntax-based models. The main difference relies on the fact that the bilingual translation models follow a language model structure with bilingual units (Mariño et al., 2006), while the phrase-based models use bilingual units with no context (Koehn et al., 2003), and, finally, the syntax-based models incorporate explicitly a representation of syntax by parsing the sources and/or target sentences following a type of grammar (Yamada & Knight, 2001).

3.3.1 The Bilingual Language Model

Early approaches in using neural networks in bilingual translation models are normally two-step systems, which means that a n-best list is proposed in a traditionally way, and then the continuous space modeling is used to rescore these lists. Schwenk et al. (2007) propose to project bilingual units onto a continuous space as an extension from previous work on monolingual language modeling (Schwenk et al., 2006). Then, this projection allows to estimate the translation probabilities in this continuous representation. Bilingual units act as neural network inputs. Again, authors face with the problem of computational complexity which is solved by limiting the vocabulary. Zamora et al. (2010) apply the same neural language model to both the bilingual and the monolingual language model and, more relevant, the decoder is extended with neural language modeling during Viterbi, which gives better results than rescoring. Leson et al. (2012) propose a similar architecture, but authors use two vectors in the input layer coming from the source and target language.

The two representations are combined in the hidden layer. Also in order to deal with a larger vocabulary, the output is structured as a clustering tree, where each word belongs to only one class and its associated sub-classes. Wu et al. (2014) tackle the sparsity problem by factorizing bilingual tuples into source and target and using a recurrent model over them.

3.3.2 The Phrase-Based Translation Model

Schwenk et al. (2012) use FNNs to estimate the translation probabilities using a continuous representation, again inspired in previous works (Schwenk et al., 2006, 2007), and they discuss their fully integration in decoding. Gao et al. (2014) report a novel phrase translation model which scores the bilingual phrase as the distance between their feature vectors in a continuous space. This continuous space is learned with a multi-layer neural network and weights are learned on the BLEU score. Sundermeyer et al. (2014) present two word-based and phrase-based approaches to recurrent translation models. The former assumes the one-to-one aligned source and target sentences. The latter models phrasal translation probabilities while avoiding sparsity issues by using single words as input and output units. Furthermore, in addition to the unidirectional formulation, authors experiment with a bidirectional network which can take the full source sentence into account for all predictions.

A particular relevant challenge for the phrase-based translation model is the fact of taking into account larger contexts while producing a translation. While standard translation provide context to the translation, it is generally limited to short contexts. Therefore, in the following we focus on works that aim at successfully employing larger contexts for the phrase-based translation model.

Zou et al. (2013) learn bilingual embeddings from a large unlabeled corpus, while utilizing MT word alignments to constrain translational equivalence. New embeddings are added as a semantic feature in a phrase-based system and significantly outperform the baseline system. Cui et al. (2014) propose to learn topic representation of parallel data using an encoder-decoder and the techniques of pre-training and fine-tuning. Pre-training and fine-tuning allow to partially initialize the error function in a point that it is easier to train. España-i-Bonet et al. (2014) use distributed vector representations of words (Mikolov, Le, & Sutskever, 2013) to handle ambiguous words. Authors identify content words which have different translations. For each of these content words, authors take a window of two previous and two following words and compute their vector representations. They compute a linear combination of these vectors to obtain a context vector. Then, they calculate a score based on the similarity among the vectors of every possible translation option. Costa-jussà et al. (2014) use a deep learning encoder-decoder structure to learn similarity correspondances between training and test sentences and integrate this similarity measure as a new feature in the phrase-based system.

3.3.3 The Syntax-Based Translation Model

Meng et al. (2015) summarize the relevant source information through a convolutional architecture, guided by the target information, and then, this architecture is integrated into a dependency-to-string translation system (Xie, Mi, & Liu, 2011). Zhai et al. (2014) use

a RNN to perform the structure prediction in a bracket transduction grammar MT system (Wu, 1995).

3.4 Reordering

Word reordering responds to the phenomena that words can take different positions in the source and target sentences involved in the translation. This has become one of the most challenging aspects in MT and there is a large body of research works addressing this issue (Bisazza & Federico, 2016). There are mainly two options to apply a reordering model within a statistical MT framework which are integrating the model within the decoder or formulating a preprocessing. The former performs reordering directly in search and in the target language. The latter reorders source words in a way that better matches the target word order (Costa-jussà & Fonollosa, 2006).

3.4.1 Reordering in Decoding

Li, Liu, Sun, Izuha, and Zhang (2014a) propose a recursive autoencoder-based for ITG-based translation (Wu, 1995), which is a type of syntax-based model. Li, Liu, Sun, Izuha, and Zhang (2014b) propose a neural reordering model that conditions reordering probabilities on the words of both the current and previous phrase pairs. (Kanouchi, Sudoh, & Komachi, 2016) use also a recursive autoencoder architecture but using phrase translation and word alignment information and tested in a phrase-based system. Specifically, to alleviate the data sparsity problem, authors build one classifier for all phrase pairs using four recursive autoencoders and a softmax layer. The phrase pairs are represented as continuous space vectors using also a recursive autoencoder. Differently, Setiawan, Huang, Devlin, Lamar, Zbib, Schwartz, and Makhoul (2015) develop new neural network features to model non-local translation phenomena related to the word reordering and improve these features with tensor neural networks. Authors use the hypothesis-enumerating features that estimate the probability of each generated target word and source-enumerating features that estimate the probability for each source word.

3.4.2 Reordering as Preprocessing

Valerio et al. (2015) propose a class of RNN models to exploit source dependency syntax. Yu et al. (2015) propose a RNN-based rule sequence model to capture an arbitrary distance of contextual information in estimating the probability of rule sequences. Cui et al. (2016) present a LSTM-based neural reordering model that directly models word pairs and their alignment.

3.5 Rescoring

Rescoring is the task of re-ranking a list of tentative translations (provided by the decoder) using different knowledge information than the one used by the models in the decoder. Research works in this area are previous and posterior to the neural MT system itself. Previous to it, mostly use neural LM or TM to rescoring statistical-based systems and they have been reported in section of language and translation modeling 3.1 and 3.3, respectively. Posterior to it, refer to the use of a neural MT system (which will be detailed in next section)

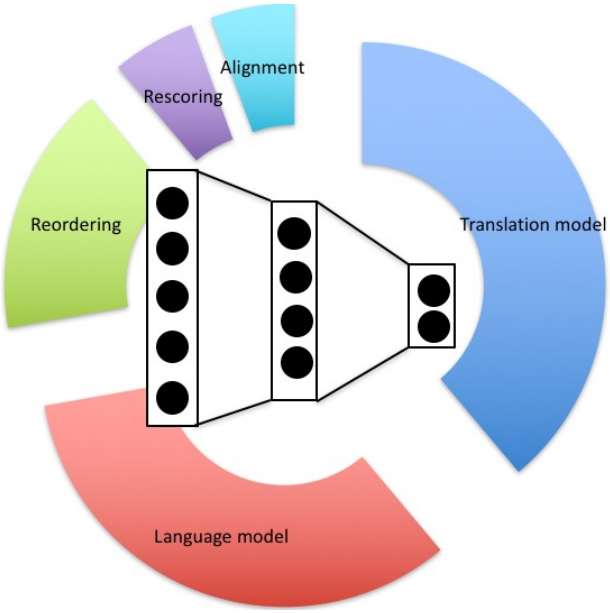

Figure 2: Introduction of neural networks in statistical MT. Most works are applied to main translation and language models (based on the papers cited in this manuscript).

Table 1: Deep learning in Statistical MT: Main Related Works

System	Model	Main related works
Statistical-based	Language Model	(Schwenk et al., 2006; Schwenk, 2010) (Niehues & Waibel, 2012; Vaswani et al., 2013) (Wang et al., 2013; Luong et al., 2015a) (Mikolov, 2012; Devlin et al., 2014) (Auli & Gao, 2014; Hu et al., 2014) (Guta et al., 2015; Costa-jussà et al., 2016)
	Word Alignment	(Yang et al., 2013) (Tamura et al., 2014)
	Translation	(Schwenk et al., 2007; Zamora-Martínez et al., 2010) (Son et al., 2012; Schwenk, 2012) (Gao et al., 2014; Meng et al., 2015) (Sundermeyer et al., 2014) (Zou et al., 2013; Martínez et al., 2014) (Mikolov et al., 2013; Cui et al., 2014) (Costa-jussà et al., 2014) (Wu et al., 2014; Zhai et al., 2014)
	Reordering	(Setiawan et al., 2015) (Miceli Barone & Attardi, 2015) (Yu & Zhu, 2015) (Cui et al., 2016) (Li et al., 2014a, 2014b; Kanouchi et al., 2016) (Setiawan et al., 2015)
	Rescoring	(Neubig et al., 2015; Stahlberg et al., 2016a)

for the same purpose. In this direction, Neubig et al. (2015) rescore n-best lists of a syntax-based system NMT while Stahlberg et al. (2016) improve this approach by using lattices instead.

4. Neural Machine Translation

Neural MT systems are neural networks trained to predict a target sentence given a source sentence. This section inspired by Cho (2015), describes the probabilistic training framework of this new approach, together with a review of the main foundational works and recent advances on neural MT.

As a summary of this section, Table 2 shows the main related works on neural MT.

4.1 The Probabilistic Training Framework

The training framework of the core neural MT aims at maximizing the probability of the target sentence given the source sentence. In particular, the neural MT system maps a source sentence $S = s_1, ..., s_I$ (with I words) into a target sentence $T = t_1, ..., t_J$ (with J words) and parametrizes the conditional distribution $p(T_n|S_n)$ for all training sentences in the corpus set $n = 1...N$. Then, the learning algorithm maximizes the following objective function:

$$argmax_\theta \frac{1}{N} \sum_{n=1}^{N} log\, p_\theta(T_n|S_n) \tag{1}$$

where θ are different model parameters. To deal with variable-length input and output, RNNs (see section 2.2) are able to maintain its internal state while reading a sequence of inputs, which for translation will be a sequence of words, thereby allowing for an input of any length.

4.2 Encoder-Decoder based on RNNs

As mentioned earlier, the neural MT system uses an encoder-decoder schema, which is maximized as shown in Equation 1. This section describes how the explicit encoder and decoder based on RNNs operate. The encoder follows the next steps:

1. Build a word to one-hot vector, which is a binary vector with a single element set to 1 (w_i).
2. Project a one-hot vector into a continuous representation. The encoder projects this vector with a matrix E whose columns are words from the source vocabulary and rows are the number of dimensions chosen ($m_i = Ew_i$). This projection generates a continuous vector for each source word, and each element of the vector is later updated to maximize the log-probability of the correct output sentence.
3. Build the sequence summarization by a RNN, $h_i = \phi_0(h_{i-1}, m_i)$, where ϕ is the activation function of the RNN with θ parameters. If visualizing this vector, it can be seen that similar sentences are close together in the summary-vector space (Sutskever et al., 2014).

The decoder, which is basically the inverse of the encoder, follows the next steps:

1. Compute the internal hidden state of the decoder $z_i = \phi_{\theta'}(h_I, u_{i-1}, z_{i-1})$, h_I represents the summary of the whole source sentence, being u_{i-1} the previous translated word, z_{i-1} the previous hidden state of the decoder.

2. Compute next word probability, by first scoring each word K given a hidden state z_i such that $e(k) = w_k^I z_i + b_k$, then, if simplifying the bias, the score can be normalized to obtain the probability by using softmax.

3. Predict next word. After choosing the i^{th} word, go back to the first step of computing the decoder's internal hidden state, scoring and normalizing the target words and selecting the next $(i+1)^t h$ word, repeating until selecting the end-of-sentence word.

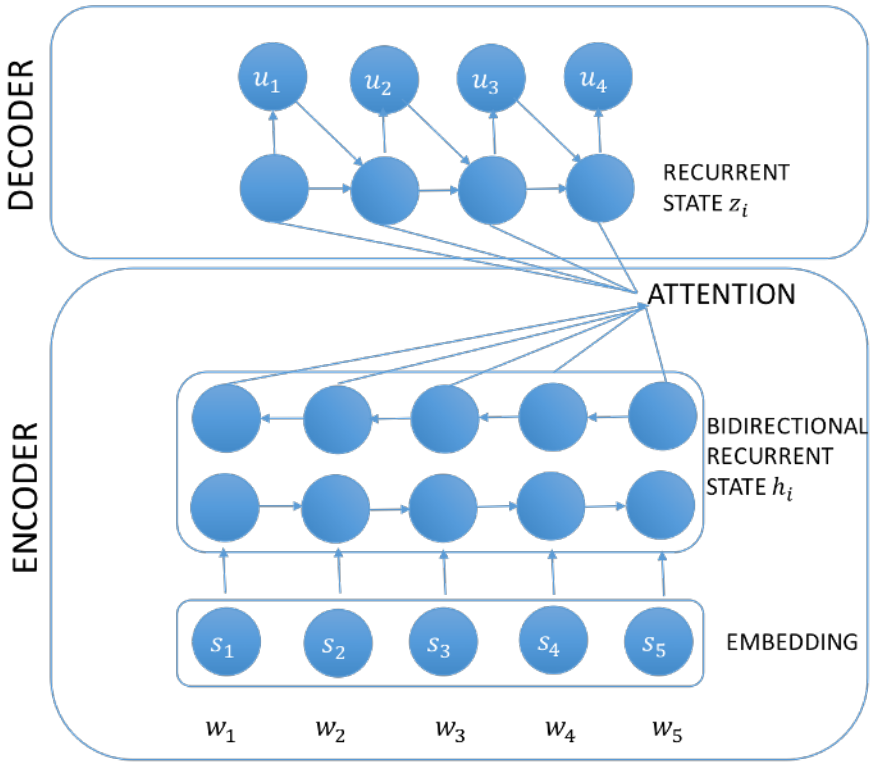

Figure 3: Neural MT architecture.

This simple architecture has led to notable improvements and to achieve state-of-the-art quality translation, as it is explained in the following section.

4.3 Foundational Works

Early research on this neural MT (Forcada & Ñeco, 1997; Castaño & Casacuberta, 1997) were mainly limited by the computational power and short data. The former builds a state-space representation of each input string and unfolds it to obtain the corresponding output string. The latter uses an Elman simple RNN (Elman, 1990) to go from source to target.

First proposed neural MT models mainly use the previous encoder-decoder architecture (Sutskever et al., 2014; Cho et al., 2014b). As explained in previous section 4.2, this architecture allows for encoding the source text into a fixed-length vector and decoding this fixed-length vector into the target text. Both encoding and decoding are trained as a single architecture on a parallel corpus. The main problem with this type of architecture is to compress the source sentence into a fixed-length vector. Cho et al. (2014b) analyse this new approach and show that neural MT performs relatively well on short sentences without unknown words, but its performance degrades rapidly with the increment of sentence length and number of unknown words.

To address the long sentence issues, i.e. mainly caused by encoding the input sentence into a single fixed-lentgh vector, Bahdanau et al. (2015) propose a new mechanism where the decoder decides which parts of the source sentence to pay attention to. This attention mechanism relieves the encoder from having to compress all the source sentence into a fixed-length vector, allowing the neural translation model to deal better with long sentences. See schematic representation of the encoder-decoder with attention in Figure 3.

In the case of Pouget-Abadie et al. (2014), authors propose a way to address the challenge of long sentences by automatically segmenting an input sentence into phrases that can be easily translated by the neural network translation model.

4.4 Recent Advances

Since neural MT is a young paradigm, it still has large room for improvement. It seems that performance on foundational neural MT works largely depended on language pair and quantity of training resources. For example, in WMT 2015, neural MT beated phrase-based systems for only one task (Bojar, Chatterjee, Federmann, Haddow, Huck, Hokamp, Koehn, Logacheva, Monz, Negri, Post, Scarton, Specia, & Turchi, 2015). However, only one year later, with the use of subword units (Sennrich, Haddow, & Birch, 2016b) and enlarging the training data, neural MT systems outperformed phrase-based systems for a large number of tasks (Bojar, Chatterjee, Federmann, Graham, Haddow, Huck, Jimeno Yepes, Koehn, Logacheva, Monz, Negri, Neveol, Neves, Popel, Post, Rubino, Scarton, Specia, Turchi, Verspoor, & Zampieri, 2016).

As follows we describe popular recent advances applied to neural MT that focus on solving big challenges in neural MT such as: covering translation of the entire source sentence; high-inflected languages and large vocabularies; low-resourced languages; and efficiency in training.

4.4.1 Encoder-Decoder Architectures

Encoders and decoders have been recently designed by means of three different successful architectures. First, using RNNs (as mentioned in section 4.2) which use recurrence to process sequences of variable lengths. The main disadvantage of these networks is they process text in a strict order (either left-to-right or right-to-left) and this is computationally expensive since it cannot be parallelized (Cho et al., 2014b). Second approach (convolutional neural networks) overcome this limitation because these networks can process all elements at the same time. These networks allow to compute a vector representation for each sequence of words and their way of dealing with the input sentence allows to learn a hierarchical

structure of the it (Gehring, Auli, Grangier, Yarats, & Dauphin, 2017). Finally, the third approach uses only self-attention mechanisms which allows to model dependencies without limitation to their position (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017).

4.4.2 Attention Mechanism Variations

As mentioned in previous section 4.3 the attention-based mechanism solves the limitation of having to encode in a fixed-length vector the entire input. There have been several proposals to improve the original proposal from Bahdanau et al. (2015). Tu et al. (2016) propose to maintain a coverage vector to keep track of attention history. This helps preventing over-translation (words unnecessarily translated for multiple times) and under-translation (words mistakenly untranslated). As a variation of this proposal, Yang et al. (2016) propose to model the sequences of attention levels for each word with a RNN while checking a fixed window of previous decisions. Given that attention to current time tends to be correlated with previous attentions, Cohn et al. (2016) propose to add information about the last attention when making the next decision. Attention can also be used for the same sentence in which case it is called intra or self-attention. This type of attention has been used for the first in translation in the work from Vaswani et al. (2017) with the new structure of multi-headed attention which allows to focus attention on different parts of the sentence at the same time.

4.4.3 Large Vocabularies, Subwords, Character-Based

As mentioned in section 2, computing the final probability of output words is costly for larger vocabularies. Jean et al. (2015) propose a model to reduce the limitation on target vocabulary using sampling to reduce the complexity of computing the normalization constant of the output word probability in neural language models (Bengio & Senecal, 2008). Luong et al. (2015b) address the problem with rare words by using a post-processing step that translates all out-of-vocabulary words using a dictionary. Differently, other approaches, based on the intuition that various word classes are translateable via smaller units than words, use word segmentation techniques and empirically show that subword models improve over a back-off dictionary baseline for unknown words (Sennrich, Haddow, & Birch, 2016a). Furthermore, there are approaches that directly deal with characters (Costa-jussà & Fonollosa, 2016; Lee, Cho, & Hofmann, 2016) or even bytes (Costa-jussà, Escolano, & Fonollosa, 2017b) to reduce vocabulary to the minimum. These works use character-based embeddings trained with convolutional networks and highway networks (Srivastava, Greff, & Schmidhuber, 2015).

4.4.4 Multilinguality and Low-Resources

Having low-resources has always been a limitation to train competitive corpus-based MT systems. Neural MT proposes several ways to tackle this problem. On the one hand, Zoph and Knight (2016) propose to train a high-resource language pair and transfer the learned parameters to the low-resource language pair. On the other hand, there are several proposals to train the systems with multilingual resources. Relevant works in this direction include systems trained from one-to-many languages (Dong, Wu, He, Yu, & Wang, 2015), which

Table 2: Deep learning in Neural MT: Main Related Works

System	Related Progress	Main related works
Neural-based	Foundational	(Forcada & Ñeco, 1997; Castaño & Casacuberta, 1997) (Sutskever et al., 2014; Cho et al., 2014a) (Cho et al., 2014b; Bahdanau et al., 2015) (Pouget-Abadie et al., 2014)
	Enc-dec	(Gehring et al., 2017; Vaswani et al., 2017)
	Attention	(Tu et al., 2016; Yang et al., 2016) (Cohn et al., 2016; Vaswani et al., 2017)
	Subwords	(Jean et al., 2015; Luong et al., 2015b) (Sennrich et al., 2016b; Costa-jussà et al., 2017b) (Costa-jussà & Fonollosa, 2016; Lee et al., 2016)
	Multilinguality	(Zoph et al., 2016; Zoph & Knight, 2016; Firat et al., 2016) (Johnson et al., 2016)
	Linguistics	(Eriguchi et al., 2016; Stahlberg et al., 2016a) (Sennrich & Haddow, 2016)
	Production	(Crego et al., 2016; Wu et al., 2016; Levin et al., 2017)

simultaneously translate sentences from one source language to multiple target languages; many-to-one languages (Zoph & Knight, 2016), in which a standard neural MT model is trained with many sources and one single target language; or many-to-many (Firat, Cho, & Bengio, 2016), in which the neural model is trained with many source and many target languages. Most recent advances include systems which are able to do zero-shot translation (Johnson, Schuster, Le, Krikun, Wu, Chen, Thorat, Viégas, Wattenberg, Corrado, Hughes, & Dean, 2016), meaning that without parallel corpus from language A and B, the system is able to learn translation among these languages. But, in general, phrase-based MT handles low-resource settings better as shown in the work of Koehn et al. (2017) .

4.4.5 Adding Prior/Linguistic Knowledge

Although the idea of neural MT and deep learning in general is adding the minimum prior knowledge possible in the systems, some approaches have shown that adding some kind of linguistic knowledge is useful. Syntactical knowledge is added in the work of Eriguchi et al. (2016) and Stahlberg et al. (2016) while Sennrich and Haddow (2016) train morphological linguistic features for word embeddings in the neural MT model.

4.4.6 Systems In Production

Although efficiency in neural MT is a challenge since training a system may last for weeks (specially when using RNNs), there are already successful neural MT systems in production (Crego, Kim, Klein, Rebollo, Yang, Senellart, Akhanov, & et al., 2016; Wu, Schuster, Chen, Le, Norouzi, & et al., 2016; Levin, Dhanuka, Khalil, Kovalev, & Khalilov, 2017).

5. Neural MT Analysis: Strengths and Weaknesses

Deep learning has been introduced in standard statistical MT systems (see section 3) and as a new MT approach (see section 4). This section makes an analysis of the main strengths and weaknesses of the neural MT approach (see a summary in Figure 5). This analysis helps towards planning the future directions of neural MT.

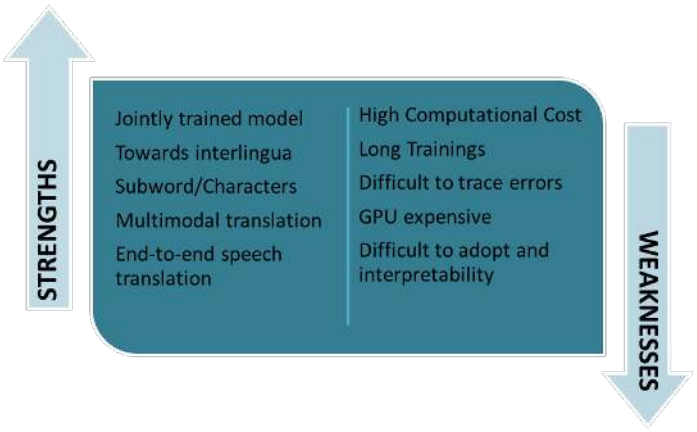

Figure 4: Strengths and Weaknesses analysis for Neural MT.

5.1 Strengths

The main inherent strength of neural MT is that all the model components are jointly trained allowing for an end-to-end optimization.

Another relevant strength is that, given its architecture based on creating an intermediate representation, the neural model could eventually evolve towards a machine-learnt interlingua approach (Johnson et al., 2016). This interlingua representation would be key to outperform MT on low-resourced language pairs as well as to efficiently deal with MT in highly multilingual environments.

In addition, neural MT has shown to be able to learn from different basic unit granularities. Subword-based representations (Sennrich et al., 2016b; Costa-jussà & Fonollosa, 2016; Lee et al., 2016) allow neural MT models with open-vocabulary by translating segmented words. Among the different alternatives to build subword units, the byte pair encoding, which is a data compresion technique, has shown to perform efficiently (Sennrich et al., 2016b). Characters allows to take advantage of intra-word information and they have been implemented only in the source side (Costa-jussà & Fonollosa, 2016) and both in the source and target sides (Lee et al., 2016).

Finally, the new paradigm allows for multimodal machine translation (Elliott, Frank, & Hasler, 2015), allowing to take advantage of image information while translating and

end-to-end speech translation architectures (Weiss, Chorowski, Jaitly, Wu, & Chen, 2017), which reduces concatenating errors.

5.2 Weaknesses

The main inherent weaknesses of neural MT are the difficulty to trace errors, long training time and high computational cost. Other weakness is the high computational cost of training and testing the models. Training can only be faced with GPUs (Graphical Processing Units) which are expensive.

Finally, an added weakness is related to interpretability of the model and the fact that the model works with vectors, matrices and tensors instead of words or phrases. Therefore, the ability to train these neural models from scratch requires background in machine learning and computer science and it is not easy that users/companies are able to comprehend/interpret it. It is difficult to adopt the paradigm. Small companies may prefer other more consolidated paradigms like the phrase and rule-based.

6. Summary, Conclusions and Future Work

Deep learning has been integrated in standard statistical MT systems at different levels (i.e. into the language model, word alignment, translation, reordering and rescoring) from different perspectives and achieving significant improvements in all cases. The field of deep learning is advancing so quickly that it is worth noticing that neural-based techniques that work today may be replaced by new ones in the near future.

In addition, an entire new paradigm has been proposed: neural MT. Curiously, this approach has been proposed almost simultaneaously as the popular phrase-based system (Forcada & Ñeco, 1997; Castaño & Casacuberta, 1997). The proposal was named differently *connectionist MT*, and given that the computational power required was prohibitive at that time and data available was not enough to train such complex systems, the idea was abandoned. Nowadays, thanks to GPUs, the computational power is not such a limitation and the information society is providing large quantities of data which allow to train the large number of parameters that these models have.

It is difficult to quantify how much does MT improve with the neural approach. It varies from language pair and task. For example, results on the WMT 2016 evaluation (Bojar et al., 2016) show that neural MT achieved best results (in terms of human evaluation) in some language directions such as German-English, English Romanian, English-German, Czech-English, English-Czech; but not in others like Romanian-English, Russian-English, English-Russian, English-Finnish. Neural MT may be more affected by large language differences, low resources and variations in training versus test domain (Aldón, 2016; Costa-jussà, Aldón, & Fonollosa, 2017a; Costa-jussà, 2017; Koehn & Knowles, 2017). Interpreting MT systems has never before been more difficult. In the evolution of MT, we have first lost rules (in the transition from the rule to the statistical-based approach) and recently, we have lost translation units (in the transition from the statistical to the neural-based approach). Nowadays, the new neural-based approaches to MT are opening new questions, e.g. is it a machine-learnt interlingua something attainable? which are the minimal units to be translated?

This manuscript recompiles and systematizes the foundational works in using deep learning in MT which is progressing incredibly fast. Deep learning is influencing many areas in natural language processing and the expectations on the use of these techniques are controversial.

It is adventurous to envisage how neural algorithms are going to impact MT in the future but it seems that they are here to stay as proven by recent news on big companies adopting the neural MT approach e.g. Google (Johnson et al., 2016) and Systran (Crego et al., 2016). Furthermore, deep learning is already taking the field dramatically further as shown by the appearence of first end-to-end speech-to-text translation (Weiss et al., 2017) and multimodal MT (Elliott et al., 2015), interlingua-based representations (Firat, Cho, Sankaran, Vural, & Bengio, 2017) and unsupervised MT (Artetxe, Labaka, Agirre, & Cho, 2017; Lample, Denoyer, & Ranzato, 2017).

Acknowledgements

This work is supported by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, the contract TEC2015-69266-P (MINECO/ FEDER, EU) and the contract PCIN-2017-079 (AEI/MINECO).

References

Aldón, D. (2016). *Sistema de Traducción Neuronal Usando Bitmaps*. B.s. thesis, Universitat Politècnica de Catalunya.

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). Unsupervised neural machine translation. *CoRR*, *abs/1710.11041*.

Auli, M., & Gao, J. (2014). Decoder integration and expected bleu training for recurrent neural network language models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 136–142, Baltimore, Maryland. Association for Computational Linguistics.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, *abs/1409.0473*.

Bengio, Y., & Senecal, J. S. (2008). Adaptive importance sampling to accelerate training of a neural probabilistic language model. *Trans. Neur. Netw.*, *19*(4), 713–722.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *Trans. Neur. Netw.*, *5*(2), 157–166.

Bengio, Y. (2009). Learning deep architectures for ai. *Found. Trends Mach. Learn.*, *2*(1), 1–127.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, *3*, 1137–1155.

Bisazza, A., & Federico, M. (2016). A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Comput. Linguist.*, *42*(2), 163–205.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., & Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pp. 131–198, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., & Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, *19*(2), 263–311.

Castaño, M. A., & Casacuberta, F. (1997). A connectionist approach to mt.. In *Proc. of the EUROSPEECH Conference*.

Cho, K. (2015). Natural language understanding with distributed representation. *CoRR*, *abs/1511.07916*.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014a). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar. Association for Computational Linguistics.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014b). 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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1724–1734.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, *abs/1412.3555*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated feedback recurrent neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 2067–2075. JMLR.org.

Cohn, T., Hoang, C. D. V., Vymolova, E., Yao, K., Dyer, C., & Haffari, G. (2016). Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 876–885, San Diego, California. Association for Computational Linguistics.

Costa-jussà, M. R. (2015). How much hybridization does machine translation need?. *Journal of the Association for Information Science and Technology*, *66*(10), 2160–2165.

Costa-jussà, M. R. (2017). Why catalan-spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. In *Proceedings of the EACL Workshop: Vardial*, Valencia.

Costa-jussà, M. R., Aldón, D., & Fonollosa, J. A. R. (2017a). Chinese-spanish neural machine translation enhanced with character and word bitmap fonts. *Machine Translation*, Accepted for publication.

Costa-jussà, M. R., Escolano, C., & Fonollosa, J. A. (2017b). Byte-based neural machine translation. In *Proc. of the 1st Workshop on Subword and Character Level Models in NLP*, pp. 154–158.

Costa-jussà, M. R., & Fonollosa, J. A. R. (2006). Statistical machine reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pp. 70–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

Costa-jussà, M. R., & Fonollosa, J. A. R. (2016). Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 357–361, Berlin, Germany. Association for Computational Linguistics.

Costa-jussà, M. R., Gupta, P., Rosso, P., & Banchs, R. E. (2014). English-to-hindi system description for wmt 2014: Deep source-context features for moses. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 79–83, Baltimore, Maryland, USA. Association for Computational Linguistics.

Costa-jussà, M., España, C., Madhyastha, P., Escolano, C., & Fonollosa, J. (2016). The talpupc spanishenglish wmt biomedical task: Bilingual embeddings and char-based neural language model rescoring in a phrase-based system. In *Proceedings of the WMT*, Berlin.

Crego, J. M., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., & et al., P. B. (2016). Systran's pure neural machine translation systems. *CoRR*, *abs/1610.05540*.

Cui, L., Zhang, D., Liu, S., Chen, Q., Li, M., Zhou, M., & Yang, M. (2014). Learning topic representation for smt with neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 133–143, Baltimore, Maryland. Association for Computational Linguistics.

Cui, Y., Wang, S., & Li, J. (2016). Lstm neural reordering feature for statistical machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 977–982, San Diego, California. Association for Computational Linguistics.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., & Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-task learning for multiple language translation. In *ACL*.

Elliott, D., Frank, S., & Hasler, E. (2015). Multi-language image description with neural sequence models. *CoRR*, *abs/1510.04709*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179 – 211.

Eriguchi, A., Hashimoto, K., & Tsuruoka, Y. (2016). Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 823–833, Berlin, Germany.

Firat, O., Cho, K., & Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 866–875, San Diego, California.

Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., & Bengio, Y. (2017). Multi-Way, Multilingual Neural Machine Translation. *Accepted for publication in Computer Speech and Language, Special Issue in Deep learning for Machine Translation*.

Forcada, M. L., & Ñeco, R. P. (1997). Recursive hetero-associative memories for translation. In *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks: Biological and Artificial Computation: From Neuroscience to Technology*, IWANN '97, pp. 453–462, London, UK, UK. Springer-Verlag.

Gao, J., He, X., Yih, W., & Deng, L. (2014). Learning continuous phrase representations for translation modeling. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 699–709.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *CoRR*, *abs/1705.03122*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Graves, A. (2013). Generating sequences with recurrent neural networks. *CoRR*, *abs/1308.0850*.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649. IEEE.

Guta, A., Alkhouli, T., Peter, J.-T., Wuebker, J., & Ney, H. (2015). A comparison between count and neural network models based on joint translation and reordering sequences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1401–1411, Lisbon, Portugal. Association for Computational Linguistics.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.*, *18*(7), 1527–1554.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, *91*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780.

Hu, Y., Auli, M., Gao, Q., & Gao, J. (2014). Minimum translation modeling with recurrent neural networks. In *Proceedings of the 14th Conference of the European Chapter*

of the Association for Computational Linguistics, pp. 20–29, Gothenburg, Sweden. Association for Computational Linguistics.

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1–10, Beijing, China. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, *abs/1611.04558*.

Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., & Uszkoreit, J. (2017). One model to learn them all. *CoRR*, *abs/1706.05137*.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Kanouchi, S., Sudoh, K., & Komachi, M. (2016). Neural reordering model considering phrase translation and word alignment for phrase-based translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pp. 94–103, Osaka, Japan. The COLING 2016 Organizing Committee.

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*.

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. In *ACL Workshop on Neural Machine translation*.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pp. 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pp. 1097–1105, USA. Curran Associates Inc.

Lample, G., Denoyer, L., & Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *CoRR*, *abs/1711.00043*.

LeCun, Y., & Bengio, Y. (1998). The handbook of brain theory and neural networks.. chap. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258. MIT Press, Cambridge, MA, USA.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the*

26th Annual International Conference on Machine Learning, ICML '09, pp. 609–616, New York, NY, USA. ACM.

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

Levin, P., Dhanuka, N., Khalil, T., Kovalev, F., & Khalilov, M. (2017). Toward a full-scale neural machine translation in production: the booking.com use case. *CoRR*, *abs/1709.05820*.

Li, P., Liu, Y., Sun, M., Izuha, T., & Zhang, D. (2014a). A neural reordering model for phrase-based translation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1897–1907, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Li, P., Liu, Y., Sun, M., Izuha, T., & Zhang, D. (2014b). A neural reordering model for phrase-based translation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1897–1907, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Lopez, A. (2008). Statistical machine translation. *ACM Comput. Surv.*, *40*(3), 8:1–8:49.

Luong, M.-T., Kayser, M., & Manning, C. D. (2015a). Deep neural language models for machine translation. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, Beijing, China.

Luong, T., Sutskever, I., Le, Q., Vinyals, O., & Zaremba, W. (2015b). Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 11–19, Beijing, China. Association for Computational Linguistics.

Maegaard, B. (1989). Eurotra: the machine translation project of the european communities. *Perspectives in artificial intelligence*, *II*.

Mariño, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A. R., & Costa-jussà, M. R. (2006). N-gram-based machine translation. *Comput. Linguist.*, *32*(4), 527–549.

Martínez, E., España-Bonet, C., Tiedemann, J., & Màrquez, L. (2014). Word's vectore representations meet machine translation. In *Proc. of the 8th Workshop on Syntax, Semantics and Structure*, Doha, Qatar.

Meng, F., Lu, Z., Wang, M., Li, H., Jiang, W., & Liu, Q. (2015). Encoding source language with convolutional neural network for machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 20–30, Beijing, China. Association for Computational Linguistics.

Miceli Barone, A. V., & Attardi, G. (2015). Non-projective dependency-based pre-reordering with recurrent neural network for machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 846–856, Beijing, China. Association for Computational Linguistics.

Mikolov, T. (2012). Statistical language models based on neural networks..

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *CoRR*, *abs/1309.4168*.

Neubig, G., Morishita, M., & Nakamura, S. (2015). Neural reranking improves subjective quality of machine translation: NAIST at WAT2015. In *Proceedings of the 2nd Workshop on Asian Translation, WAT 2015, Kyoto, Japan, October 16, 2015*, pp. 35–41.

Niehues, J., & Waibel, A. (2012). Continuous space language models using restricted boltzmann machines..

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipson, J. (2014; accessed September 2017). Systran: A brief history of machine translation..

Pouget-Abadie, J., Bahdanau, D., van Merrienboer, B., Cho, K., & Bengio, Y. (2014). Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 78–85, Doha, Qatar. Association for Computational Linguistics.

Schuster, M., & Paliwal, K. (1997). Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, *45*(11), 2673–2681.

Schwenk, H. (2010). Continuous-space language models for statistical machine translation. *Prague Bull. Math. Linguistics*, *93*, 137–146.

Schwenk, H. (2012). Continuous space translation models for phrase-based statistical machine translation. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pp. 1071–1080.

Schwenk, H., Costa-Jussà, M. R., & Fonollosa, J. A. R. (2006). Continuous space language models for the IWSLT 2006 task. In *2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006*, pp. 166–173.

Schwenk, H., Costa-Jussà, M. R., & Fonollosa, J. A. R. (2007). Smooth bilingual n-gram translation. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pp. 430–438.

Sennrich, R., & Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pp. 83–91, Berlin, Germany.

Sennrich, R., Haddow, B., & Birch, A. (2016a). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pp. 371–376, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Setiawan, H., Huang, Z., Devlin, J., Lamar, T., Zbib, R., Schwartz, R., & Makhoul, J. (2015). Statistical machine translation features with multitask tensor networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 31–41, Beijing, China. Association for Computational Linguistics.

Son, L. H., Allauzen, A., & Yvon, F. (2012). Continuous space translation models with neural networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pp. 39–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway networks. *CoRR*, *abs/1505.00387*.

Stahlberg, F., Hasler, E., & Byrne, B. (2016a). The edit distance transducer in action: The university of cambridge english-german system at wmt16. In *Proceedings of the First Conference on Machine Translation*, pp. 377–384, Berlin, Germany. Association for Computational Linguistics.

Stahlberg, F., Hasler, E., Waite, A., & Byrne, B. (2016b). Syntactically guided neural machine translation. *CoRR*, *abs/1605.04569*.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pp. 257–286.

Sundermeyer, M., Alkhouli, T., Wuebker, J., & Ney, H. (2014). Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 14–25, Doha, Qatar. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112.

Tamura, A., Watanabe, T., & Sumita, E. (2014). Recurrent neural networks for word alignment model. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1470–1480, Baltimore, Maryland. Association for Computational Linguistics.

Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Compu-*

tational Linguistics (Volume 1: Long Papers), pp. 76–85, Berlin, Germany. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*.

Vaswani, A., Zhao, Y., Fossum, V., & Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1387–1392.

Wang, R., Utiyama, M., Goto, I., Sumita, E., Zhao, H., & Lu, B.-L. (2013). Converting continuous-space language models into n-gram language models for statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 845–850, Seattle, Washington, USA. Association for Computational Linguistics.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., & Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. *CoRR*, *abs/1703.08581*.

Wu, D. (1995). Trainable coarse bilingual grammars for parallel text bracketing. In *Proceedings of the Third Workshop on Very Large Corpora (VLC)*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., & et al., W. M. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, *abs/1609.08144*.

Wu, Y., Watanabe, T., & Hori, C. (2014). Recurrent neural network-based tuple sequence model for machine translation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1908–1917, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Xie, J., Mi, H., & Liu, Q. (2011). A novel dependency-to-string model for statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 216–226, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pp. 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang, N., Liu, S., Li, M., Zhou, M., & Yu, N. (2013). Word alignment modeling with context dependent deep neural network. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 166–175, Sofia, Bulgaria. Association for Computational Linguistics.

Yang, Z., Hu, Z., Deng, Y., Dyer, C., & Smola, A. J. (2016). Neural machine translation with recurrent attention modeling. *CoRR*, *abs/1607.05108*.

Yu, H., & Zhu, X. (2015). Recurrent neural network based rule sequence model for statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural*

Language Processing (Volume 2: Short Papers), pp. 132–138, Beijing, China. Association for Computational Linguistics.

Zamora-Martínez, F., Bleda, M. J. C., & Schwenk, H. (2010). N-gram-based machine translation enhanced with neural networks for the french-english btec-iwslt'10 task. In *2010 International Workshop on Spoken Language Translation, IWSLT 2010, Paris, France, December 2-3, 2010*, pp. 45–52.

Zhai, F., Zhang, J., Zhou, Y., & Zong, C. (2014). Rnn-based derivation structure prediction for smt. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 779–784, Baltimore, Maryland. Association for Computational Linguistics.

Zhang, J., & Zong, C. (2015). Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, 1541–1672.

Zoph, B., & Knight, K. (2016). Multi-source neural translation. *CoRR*, *abs/1601.00710*.

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. *CoRR*, *abs/1604.02201*.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1393–1398, Seattle, Washington, USA. Association for Computational Linguistics.