NEURAL MACHINE TRANSLATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Minh-Thang Luong
September 2016

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Christopher D. Manning) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Dan Jurafsky)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Andrew Ng)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Quoc V. Le)

Approved for the Stanford University Committee on Graduate Studies

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> The Babel fish is small, yellow, leech-like, and probably the oddest thing in the universe. It feeds on brainwave energy ... if you stick a Babel fish in your ear, you can instantly understand anything in any form of language.
>
> *The Hitchhiker's Guide to the Galaxy*. Douglas Adams.

Human languages are diverse with about 6000 to 7000 languages spoken worldwide (Anderson, 2010). As civilization advances, the need for seamless communication and understanding across languages becomes more and more crucial. Machine translation (MT), the task of teaching machines to learn to translate automatically across languages, as a result, is an important research area. MT has a long history (Hutchins, 2007) from the original phiosophical ideas of universal languages in the 17^{th} century to those first practical suggestions in the 1950s, most notably an influential proposal by Weaver (1949) which marked the beginnings of MT research in the United States. In that memorandum, Warren Weaver touched on the idea of using computers to translate, specifically addressing the language ambiguity problem by combining his knowledge of statistics, cryptography, information theory, as well as logical and linguistic universals (Hutchins, 2000). Since then, MT has gone through many periods of great development but also encountered several stagnant phases as illustrated in Figure 1.1. Despite several moments of excitement that led to hopes that MT will be solved "very soon" such as the 701 translator (Sheridan, 1955) developed by scientists at Georgetown and IBM in the 1950s and the popular Google Translate at the

beginning of the 21st century (Brants et al., 2007), MT remains an extremely challenging problem (Kelly, 2014; David, 2016). This motivates my work in the area of machine translation; specifically, in this thesis, the goal is to advance neural machine translation (NMT), a new promising approach for MT developed just recently, over the past two years. The results achieved in this thesis on NMT together with work from other researchers have eventually produced a significant leap in the translation quality as illustrated in Figure 1.1. Before delving into details of the thesis, we now walk the audience through the background and a bit of the development history of machine translation.

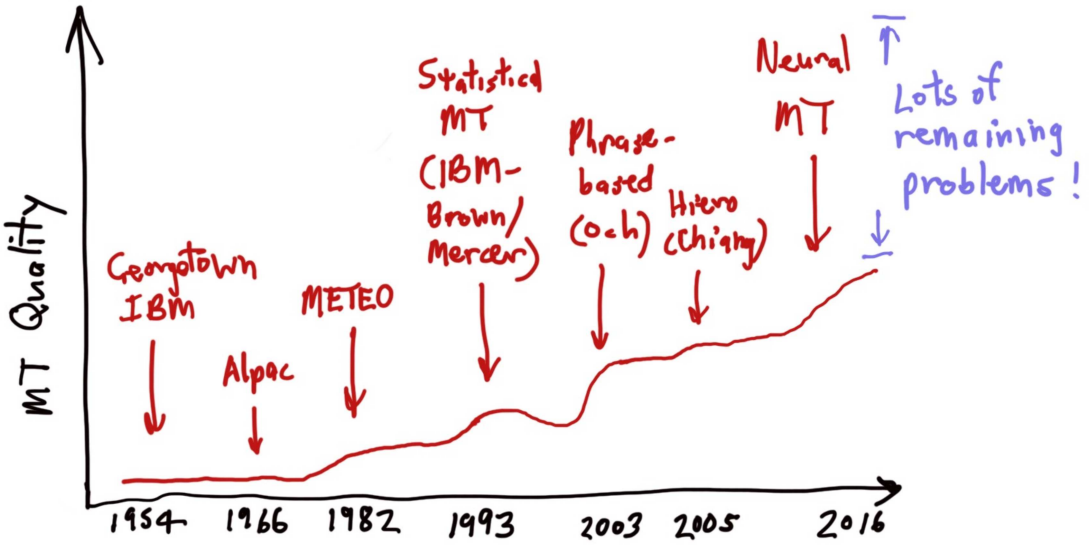

Figure 1.1: **Machine translation progress** – from the 1950s, the starting of modern MT research, until the time of this thesis, 2016, in which neural MT becomes a dominant approach. Image courtesy of Christopher D. Manning.

1.1 Machine Translation

Despite much enthusiasm, the beginning period of MT research in the 1950-60s, was mostly about direct word-for-word replacement based on bilingual dictionaries.[1] An MT winner quickly came right after the ALPAC report in 1966 pointing out that "there is no

[1]There are also proposals for "interlingual" and "transfer" approaches but these seemed to be too challenging to achieve, not to mention limitations in hardware at that time(Hutchins, 2007).

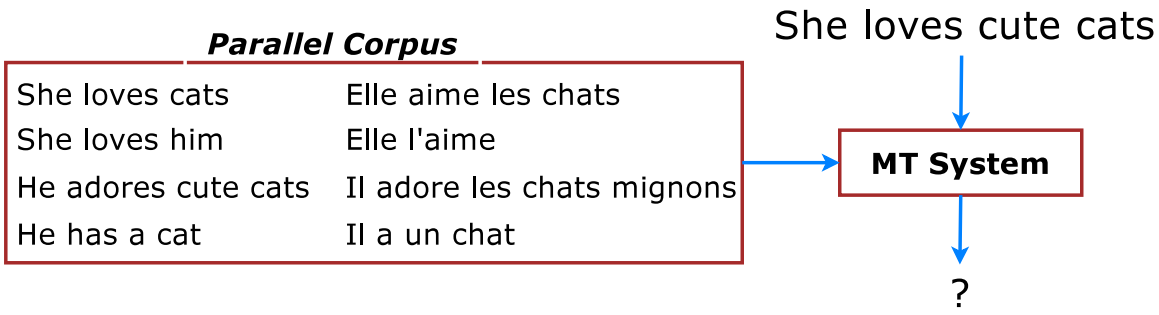

Figure 1.2: **Corpus-based approaches to machine translation** – a general setup in which MT systems are built from parallel corpora of sentence pairs having the same meaning. Once built, systems are used to translate new unseen sentences, e.g., "She loves cute cats".

immediate or predictable prospect of useful machine translation", which hampered MT research for over a decade. Fast-forwarding through the resurgence in the 1980s beginning with Europe, Japan, and gradually the United States, modern statistical MT started out with a seminal work by IBM scientists (Brown et al., 1993). The proposed *corpus-based* approaches require minimal linguistic content and only need a *parallel* dataset of sentence pairs which are translations of one another, to train MT systems. Such a language-independent setup is illustrated in Figure 1.2. In more detail, instead of hand building bilingual dictionaries which can be costly to obtain, Brown and colleagues proposed to learn these dictionaries, or *translation models*, probabilistically from parallel corpora. To accomplish this, they propose a series of 5 algorithms of increasing complexity, often referred as IBM Models 1-5, to learn *word alignment*, a mapping between source and target words in a parallel corpus, as illustrated in Figure 1.3. The idea is simple: the more often two words, e.g., "loves" and "aime", occur together in different sentence pairs, the more likely they are aligned to each other and have equivalent meanings.

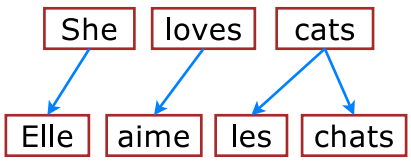

Figure 1.3: **Word-based alignment** – example of an alignment between source and target words. In IBM alignment models, each target word is aligned to at most one source word.

Once a translation model, i.e., a probabilistic bilingual dictionary, has been learned, IBM model 1, the simplest and the most naïve one among the five proposed algorithms, translates a new source sentence as follows. First, it decides on how long the translation is as well as how source words will be mapped to target words as illustrated in Step 1 of Figure 1.4. Then, in Step 2, it produces a translation by selecting for each target position a word that is the best translation for the aligned source word according to the bilingual dictionary. Subsequent IBM models build on top of one another and refine the translation story such as better modeling the reordering structure, i.e., how word positions differ between source and target languages. We refer the audience to the original IBM paper or Chapter 25 of (Jurafsky and Martin, 2009) for more details.

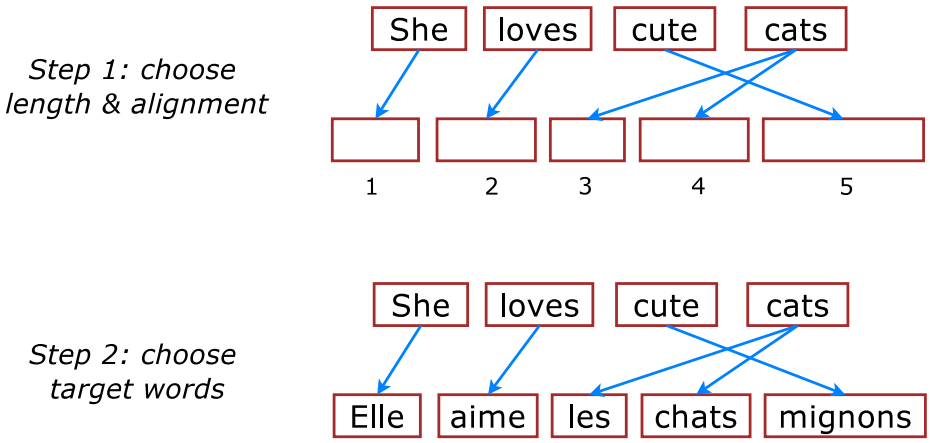

Figure 1.4: **A simple translation story** – example of the generative story in IBM Model 1 to produce a target translation given a source sentence and a learned translation model.

There are, however, two important details that we left out in the above translation story, the *search* process and the *language modeling* component. In Step 1, one might wonder among the exponentially many choices, how do we know what the right translation length is and how source words should be mapped to target words? The search procedure informally helps us "browse" through a manageable set of candidates which are likely to include a good translation; whereas, the language model will help us select the best translation among these candidates. We will defer details of the search process to later since it is dependable on the exact translation model being used. Language modeling, on the other hand, is an important concept which has been studied earlier in speech recognition (Katz, 1987). In

a nutshell, a language model (LM) learns from a corpus of monolingual text in the target language and collect statistics on which sequence of words are likely to go with one another. When applying to machine translation, an LM will assign high scores for coherent and natural-sounding translations and low scores for bad ones. For our example in the above figure, if the model happens to choose a wrong alignment, e.g., "cute" goes to position 3 while "cats" goes to positions 4 and 5, an LM will alert us with a lower score given to that incorrect translation "Elle aime mignons les chats" compared to the translation "Elle aime les chats mignons" with a correct word ordering structure.[2]

While the IBM work had a huge impact on the field of statistical MT, researchers quickly realized that word-based MT is insufficient as words require context to properly translate, e.g., "bank" has two totally different meanings when preceded by "financial" and "river". As a result, *phrase-based models*, (Marcu and Wong, 2002; Zens et al., 2002; Koehn et al., 2003), inter alia, became the de facto standard in MT research and are still being the dominant approach in existing commercial systems such as Google Translate until now. Much credit went to Och's work on *alignment templates*, starting with his thesis in 1998 and later in (Och and Ney, 2003, 2004). The idea of alignment templates is to enable phrase-based MT by first symmetrizing[3] the alignment to obtain many-to-many correspondences between source and target words; in contrast, the original IBM models only produce one-to-many alignments. From the symmetrized alignment, several heuristics have been proposed to extract phrase pairs; the general idea is that phrase pairs need to be "consistent" with their alignments: each word in a phrase should not be aligned to a word outside of the other phrase. These pairs are stored in what called a *phrase table*

[2] For completeness, translation and language models are integrated together in an MT system through the *Bayesian noisy channel* framework as follows:

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, P(t|s) \approx \underset{t}{\operatorname{argmax}}\, P(s|t)P(t) \tag{1.1}$$

Here, we have a source sentence s in which we ask our *decoder*, an algorithm that implements the aforementioned search process, to find the best translation, the argmax part. $P(s|t)$ represents the *translation* model, the faithfulness of the translation in terms of meaning preservation between the source and the target sentences; whereas $P(t)$ represents the *language* model, the fluency of the translated text.

[3] Symmetrization is achieved by training IBM models in both directions, source to target and vice versa, then intersecting the alignments. There are subsequent techniques that jointly train alignments in both directions such as (Liang et al., 2006).

together with various scores to evaluate phrase pairs in different aspects, e.g., how equivalent the meaning is, how good the alignment is, etc. Figure 1.5 gives an example of how a phrase-based system translates.

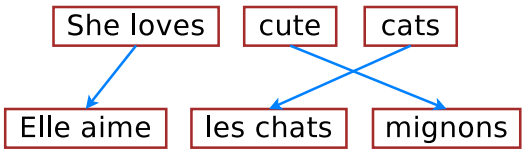

Figure 1.5: **Phrase-based machine translation** (MT) – example of how phrase-based MT systems translate a source sentence "She loves cute cats" into a target sentence "Elle aime les chats mignons": sentences are split into chunks and phrases are translated.

State-of-the-art MT systems, in fact, contain more components than just the two basic translation and language models. There are many knowledge sources that can be useful to the translation task, e.g., language model, translation model, reversed translation model, reordering model, word penalty, phrase penalty, unknown-word penalty, etc. To incorporate all of these features, modern MT systems use a popular approach in natural language processing, called the *maximum-entropy* or *log-linear* models (Berger et al., 1996; Och and Ney, 2002), which has as its special case the Bayesian noisy channel model that we briefly mentioned through Eq. (1.1). Training log-linear MT models can be done using the standard *maximum likelihood estimation* approach. However, in practice, these models are learned by directly optimizing translation quality metrics such as BLEU (Papineni et al., 2002) in a technique known as *minimum error rate training* or *MERT* (Och, 2003). Here, BLEU is an inexpensive automatic way of evaluating the translation quality; the idea is to count words and phrases that overlap between machine and human outputs.

1.2 Neural Machine Translation

The aforementioned approach, while has been successfully deployed in many commercial systems, does not work very well and suffers from the following two major drawbacks. First, translation decisions are *locally determined* as we translate phrase-by-phrase and long-distance dependencies are often ignored. Second, it is slightly "strange" that language

models (LMs), despite being a key component in the MT pipeline, utilize context information that is both short, consisting of only a handful of previous words, and target-only, never looking at the source words. These shortcomings in LMs gives rise to a new wave of *hybrid* systems which aim to empower phrase-based MT with neural network components, most notably neural probabilistic language models (NPLMs).

NPLMs were first proposed by Bengio et al. (2003) as a way to combat the "curse" of dimensionality suffered by traditional LMs. In traditional LMs, one has to explicitly store and handle all possible n-grams occurred in a training corpus, the number of which quickly becomes enormous. As a result, existing MT systems often limit themselves to use only short, e.g., 5-gram, LMs (Heafield, 2011), which capture little context and cannot generalize well to unseen n-grams. NPLMs address these concerns by using distributed representations of words and not having to explicitly store all enumerations of words. As a result, many MT systems, (Schwenk, 2007; Vaswani et al., 2013; Luong et al., 2015a), inter alia, start adopting NPLMs alongside with traditional LMs. To make NPLMs even more powerful, recent work (Schwenk, 2012; Son et al., 2012; Auli et al., 2013; Devlin et al., 2014) propose to condition on source words beside the target context to lower uncertainty in predicting next words (see Figure 1.6).[4]

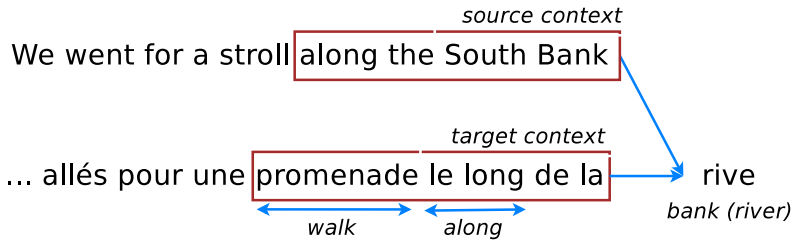

Figure 1.6: **Source-conditioned neural probabilistic language models** (NPLMs) – example of a source-conditioned NPLM proposed by Devlin et al. (2014). To evaluate a how likely a next word "rive" is, the model not only relies on previous target words (context) "promenade le long de la" as in traditional NPLMs (Bengio et al., 2003), but also utilizes source context "along the South Bank" to lower uncertainty in its prediction.

These hybrid MT systems with NPLM components, while having addressed shortcomings of traditional phrase-based MT, still translate locally and fail to capture long-range

[4]In (Devlin et al., 2014), the authors have constructed a model that conditions on 3 target words and 11 source words, effectively building a 15-gram LM.

dependencies. For example, in Figure 1.6, the source-conditioned NPLM does not see the word "stroll", or any other words outside of its fixed context windows, which can be useful in deciding that the next word should be "bank" as in "river bank" rather "financial bank". More problematically, the entire MT pipeline is already complex with different components needed to be tuned separatedly, e.g., translation models, language models, reordering models, etc.; now, it becomes even worse as different neural components are incorporated. Neural Machine Translation to the rescue!

Neural Machine Translation (NMT) is a new approach to translating text from one language into another that captures long-range dependencies in sentences and generalizes better to unseen texts. The core of NMT is a single deep neural network with hundreds of millions of neurons that learn to directly map source sentences to target sentences (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014). This is often referred as the sequence-to-sequence or encoder-decoder approach.[5] NMT is appealing since it is conceptually simple and can be trained end-to-end. NMT translates as follows: an *encoder* reads through the given source words one by one until the end, and then, a *decoder* starts emitting one target word at a time until a special end-of-sentence symbol is produced. We illustrate this process in Figure 1.7.

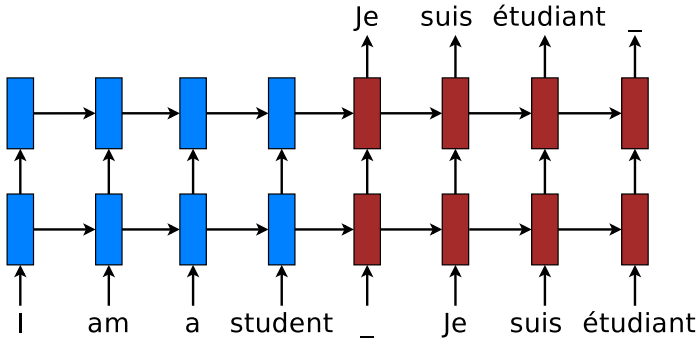

Figure 1.7: **Neural machine translation** – example of a deep recurrent architecture proposed by Sutskever et al. (2014) for translating a source sentence "I am a student" into a target sentence "Je suis étudiant". Here, "_" marks the end of a sentence.

Such simplicity leads to several advantages. NMT requires minimal domain knowledge: it only assumes access to sequences of source and target words as training data and

[5]Forcada and Neco (1997) wrote the very first paper on sequence-to-sequence models for translation!

learns to directly map one into another. NMT beam-search decoders that generate words from left to right can be easily implemented, unlike the highly intricate decoders in standard MT (Koehn et al., 2003). Lastly, the use of recurrent neural networks (RNNs) allow NMT to generalize well to very long word sequences while not having to explicitly store any gigantic phrase tables or language models as in the case of standard MT.

1.3 Thesis Outline

Despite all the aforementioned advantages and potentials, the early NMT architecture (Sutskever et al., 2014; Cho et al., 2014) still has many drawbacks. In this thesis, I will highlight three problems pertaining to the existing NMT model, namely the *vocabulary size*, the *sentence length*, and the *language complexity* issues. Each chapter is devoted to solving each of these problems in which I will describe how I have pushed the limits of NMT, making it applicable to a wide variety of languages with state-of-the-art performance such as English-French (Luong et al., 2015c), English-German (Luong et al., 2015b; Luong and Manning, 2015), and English-Czech (Luong and Manning, 2016). Towards the *future* of NMT, I answer two questions: (1) whether we can improve translation by jointly learning from a wide variety of sequence-to-sequence tasks such as parsing, image caption generation, and auto-encoders or skip-thought vectors (Luong et al., 2016); and (2) whether we can compress NMT for mobile devices (See et al., 2016). In brief, this thesis is organized as follows. I start off by providing background knowledge on RNN and NMT in Chapter 2. The aforementioned three problems and approaches for NMT future are detailed in Chapters 3, 4, 5, and 6 respectively, which we will go through one by one next. Chapter 7 wraps up and discusses remaining challenges in NMT research.

Copy Mechanisms

A significant weakness in conventional NMT systems is their inability to correctly translate very rare words: end-to-end NMTs tend to have relatively small vocabularies with a single <unk> symbol that represents every possible out-of-vocabulary (OOV) word. In Chapter 3, we propose simple and effective techniques to address this *vocabulary size* problem

through teaching NMT to "copy" words from source to target. Specifically, we train an NMT system on data that is augmented by the output of a word alignment algorithm, allowing the NMT system to emit, for each OOV word in the target sentence, the position of its corresponding word in the source sentence. This information is later utilized in a post-processing step that translates every OOV word using a dictionary. Our experiments on the WMT'14 English to French translation task show that this method provides a substantial improvement of up to 2.8 BLEU points over an equivalent NMT system that does not use this technique. With 37.5 BLEU points, our NMT system is the first to surpass the best result achieved on a WMT'14 contest task.

Attention Mechanisms

While NMT can translate well for short- and medium-length sentences, it has a hard time dealing with long sentences. An attentional mechanism was proposed by Bahdanau et al. (2015) to address that *sentence length* problem by selectively focusing on parts of the source sentence during translation. However, there has been little work exploring useful architectures for attention-based NMT. Chapter 4 examines two simple and effective classes of attentional mechanism: a *global* approach which always attends to all source words and a *local* one that only looks at a subset of source words at a time. We demonstrate the effectiveness of both approaches on the WMT translation tasks between English and German in both directions. With local attention, we achieve a significant gain of 5.0 BLEU points over non-attentional systems that already incorporate known techniques such as dropout. Our ensemble model using different attention architectures yields a new state-of-the-art result in the WMT'15 English to German translation task with 25.9 BLEU points, an improvement of 1.0 BLEU points over the existing best system backed by NMT and an n-gram reranker.

Hybrid Models

Nearly all previous NMT work has used quite restricted vocabularies, perhaps with a subsequent method to patch in unknown words such as the copy mechanisms mentioned earlier. While effective, the copy mechanims cannot deal with all the complexity of human languages such as rich morphology, neologisms, and informal spellings. Chapter 5 presents

a novel word-character solution to that *language complexity* problem towards achieving open vocabulary NMT. We build hybrid systems that translate mostly at the *word* level and consult the *character* components for rare words. Our character-level recurrent neural networks compute source word representations and recover unknown target words when needed. The twofold advantage of such a hybrid approach is that it is much faster and easier to train than character-based ones; at the same time, it never produces unknown words as in the case of word-based models. On the WMT'15 English to Czech translation task, this hybrid approach offers an addition boost of $+2.1-11.4$ BLEU points over models that already handle unknown words. Our best system achieves a new state-of-the-art result with 20.7 BLEU score. We demonstrate that our character models can successfully learn to not only generate well-formed words for Czech, a highly-inflected language with a very complex vocabulary, but also build correct representations for English source words.

NMT Future

Chapter 6 answers the two aforementioned questions for the future of NMT: whether we can utilize other tasks to improve translation and whether we can compress NMT models.

For the first question, we examine three multi-task learning (MTL) settings for sequence to sequence models: (a) the *one-to-many* setting – where the encoder is shared between several tasks such as machine translation and syntactic parsing, (b) the *many-to-one* setting – useful when only the decoder can be shared, as in the case of translation and image caption generation, and (c) the *many-to-many* setting – where multiple encoders and decoders are shared, which is the case with unsupervised objectives and translation. Our results show that training on a small amount of parsing and image caption data can improve the translation quality between English and German by up to 1.5 BLEU points over strong single-task baselines on the WMT benchmarks. Rather surprisingly, we have established a new *state-of-the-art* result in constituent parsing with 93.0 F_1 by utilizing translation data. Lastly, we reveal interesting properties of the two unsupervised learning objectives, autoencoder and skip-thought, in the MTL context: autoencoder helps less in terms of perplexities but more on BLEU scores compared to skip-thought.

For the second question, we examine three simple magnitude-based pruning schemes to compress NMT models, namely *class-blind*, *class-uniform*, and *class-distribution*, which

differ in terms of how pruning thresholds are computed for the different classes of weights in the NMT architecture. We demonstrate the efficacy of weight pruning as a compression technique for a state-of-the-art NMT system. We show that an NMT model with over 200 million parameters can be pruned by 40% with very little performance loss as measured on the WMT'14 English-German translation task. This sheds light on the distribution of redundancy in the NMT architecture. Our main result is that with *retraining*, we can recover and even surpass the original performance with an 80%-pruned model.

Chapter 2

Background

Chapter 3

Copy Mechanisms

Chapter 4

Attention Mechanisms

Chapter 5

Hybrid Models

Chapter 6

NMT Future

Chapter 7

Conclusion

Bibliography

Stephen R. Anderson. 2010. How many languages are there in the world? `http://www.linguisticsociety.org/content/how-many-languages-are-there-world`. Accessed: 2016-09-10.

Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. 2013. Joint language and translation modeling with recurrent neural networks. In *ACL*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *JMLR* 3:1137–1155.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–71.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *EMNLP*.

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

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.

Ernest David. 2016. Winograd schemas and machine translation. *arXiv preprint arXiv:1608.01884* .

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *ACL*.

Mikel L. Forcada and Ramón Neco. 1997. Recursive hetero-associative memories for translation. *Biological and Artificial Computation: From Neuroscience to Technology* pages 453–462.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *WMT*.

W. John Hutchins. 2000. Warren Weaver and the launching of MT: brief biographical note. In *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, John Benjamins, pages 17–20.

W. John Hutchins. 2007. Machine translation: A concise history. In Chan Sin Wai, editor, *Computer Aided Translation: Theory and Practice*, Chinese University of Hong Kong.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35(3):400–401.

Nataly Kelly. 2014. Why machines alone cannot solve the worlds translation problem. `http://www.huffingtonpost.com/nataly-kelly/why-machines-alone-cannot-translation_b_4570018.html`. Accessed: 2016-09-10.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *NAACL*.

Minh-Thang Luong, Michael Kayser, and Christopher D. Manning. 2015a. Deep neural language models for machine translation. In *CoNLL*.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *ICLR*.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *IWSLT*.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *ACL*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015b. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015c. Addressing the rare word problem in neural machine translation. In *ACL*.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *EMNLP*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics* 30(4):417–449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech and Languages* 21(3):492–518.

Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. In *COLING*.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. 2016. Compression of neural machine translation models via pruning. In *CoNLL*.

Peter Sheridan. 1955. Research in language translation on the IBM type 701. In *IBM Technical Newsletter*, 9.

Le Hai Son, Alexandre Allauzen, and Franois Yvon. 2012. Continuous space translation models with neural networks. In *NAACL-HLT*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *EMNLP*.

Warren Weaver. 1949. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, MIT Press, Cambridge, MA, pages 15–23. Reprinted from a memorandum written by Weaver in 1949.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. *Phrase-Based Statistical Machine Translation*, Springer Berlin Heidelberg, pages 18–32.