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
June 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

IMAGINE A READER.
MAYBE ABI IN FALL 2015.

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.

what does this mean?

Human languages are diverse and rich in categories with about 6000 to 7000 languages spoken worldwide.[1] 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 [11] from the original phiosophical ideas of universal languages in the seventeen century to the first practical instances of MT in the twentieth century, e.g., one proposal by Weaver [28]. Despite several excitement moments that led to hopes that MT will be solved "very soon", e.g., the 701 translator[2] developed by scientists at GeorgeTown and IBM in the 1950s or a simple vector-space transformation technique[3] proposed by Google researchers at the beginning of

Important proposal but surely can't be an example of a "practical instance"

either: "exciting moments" or "moments of excitement"

[1] http://www.linguisticsociety.org/content/how-many-languages-are-there-world

[2] http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

[3] https://www.technologyreview.com/s/519581/how-google-converted-language-translation-into-a-problem-of-vector-space-mathematics/

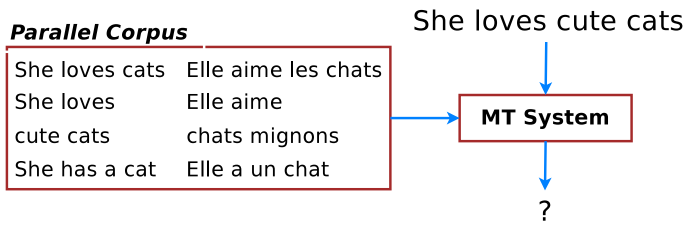

Figure 1.1: **Machine translation** (MT) – a general setup of MT. Systems build translation models from parallel corpora to translate new unseen sentences, e.g., "She loves cute cats".

the twenty-first century, MT remains ~~to be~~ an extremely challenging problem.[4] To understand why MT is difficult, let us trace through one "evolution" path of MT which crosses through techniques that are used extensively in commercial MT systems.

Modern statistical MT started out with a seminal work by IBM scientists [3]. The proposed technique requires minimal linguistic content and only needs a *parallel corpus*, i.e., a set of pairs of sentences that are translations of one another, to train machine learning algorithms to tackle the translation problem. Such a language-independent setup is illustrated in Figure 1.1 and remains ~~to be~~ the general approach for nowadays MT systems. For over twenty years since the IBM seminal paper, approaches in MT such as [4, 5, 8, 13, 14, 15, 22], are, by and large, similar according to the following two-stage process (see Figure 1.2). First, source sentences are broken into chunks which can be translated in isolation by looking up a "dictionary", or more formally a *translation model*. Translated target words and phrases are then put together to form coherent and natural-sounding sentences by consulting a *language model* (LM) on which sequences of words, i.e., *n-grams*, are likely to go with one another.

I think this history is too loose for a thesis on MT. There was a huge change from the original purely word-based IBM models of the 1990s and the phrase-based models introduced by Och et al. in the early 2000s.

That is, this isn't true of IBM work.

I think this is too short and unclear for a reader who doesn't already know it. It doesn't express that the phrase table contains a whole bunch of possible translations, with scores, nor does it explain what a language model is.

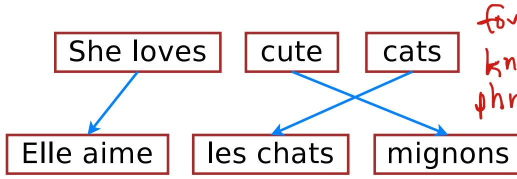

Figure 1.2: **Phrase-based machine translation** (MT) – example of how phrase-based MT systems translate a source sentence "She loves cute cats" into a target sentence "Elle aime les chats mignons": sentences are split into chunks and phrases are translated.

[4] http://www.huffingtonpost.com/nataly-kelly/why-machines-alone-cannot-translation_b_4570018.html

can't have null subject in finite clause in English

The aforementioned approach, while has been successfully deployed in many commercial systems, it 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 language models (NLMs).

Scare quotes are bad style! Maybe it's more "unfortunate". You should explain why they were like this, even if it's unfortunate.

NLMs were first proposed by Bengio et al. [2] 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 [10], which capture little context and cannot generalize well to unseen n-grams. NLMs 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, [18, 23, 27], inter alia, start ed adopting NLMs alongside with traditional LMs. To make NLMs even more powerful, recent work [7, 24] propose s to condition on source words ~~beside~~ as well as the target context to lower uncertainty in predicting next words (see Figure 1.3).[5]

I guess you say why here but different order better?

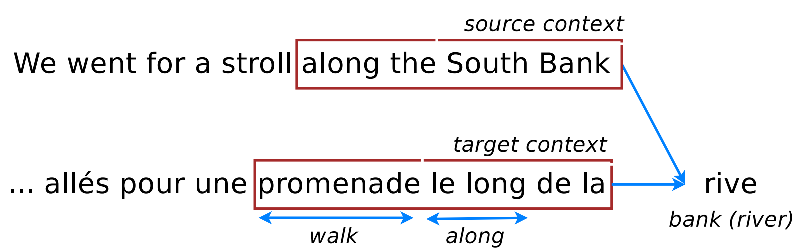

Figure 1.3: **Source-conditioned neural language model** (NLM) – example of a source-conditioned NLM proposed by Devlin et al. [7]. To evaluate ~~s~~ how likely a next word "rive" is, the model not only relies on previous target words (context) "promenade le long de la" as in traditional NLMs [2], but also utilizes source context "along the South Bank" to lower uncertainty in its prediction.

X

[5] In [7], the authors ~~have~~ constructed a model that conditions on 3 target words and 11 source words, effectively building a 15-gram LM.

X

These hybrid MT systems with NLM components, while having addressed shortcomings of traditional phrase-based MT, still translate locally and fail to capture long-range dependencies. For example, in Figure 1.3, the source-conditioned NLM 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 need~~ed~~ing 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!~~ Too informal. Rewrite or omit.

You haven't mentioned these. You should, if you are now using the term as if the reader should understand.

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 artificial neurons that learn to directly map source sentences to target sentences [6, 12, 26]. This is often referred to as the sequence-to-sequence or encoder-decoder approach.[6] 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 1.4.

Don't most current models have way less "neurons" than this?

What is this? Surely work like Kalchbrenner's original proposal is also NMT but not sequence to sequence? It is an encoder-decoder approach though. These are different things.

Say why this is an advantage

Again, is this a definition or current practice? If I built a system that generated a character or a phrase at a time, wouldn't it still be an NMT system?!?

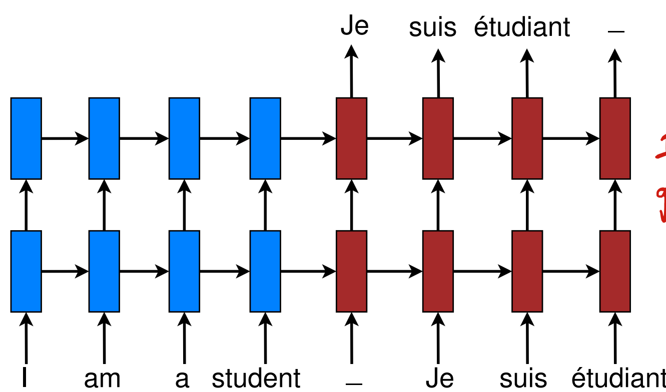

Figure 1.4: **Neural machine translation** – example of a deep recurrent architecture proposed by Sutskever et al. [26] for translating a source sentence "I am a student" into a target sentence "Je suis étudiant". Here, "_" marks the end of a sentence.

X

[6]Forcada and Neco [9] wrote the very first paper on sequence-to-sequence models for translation.

You also haven't explained what an RNN is up until here, even in an informal "the rough idea is" kind of a way.

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 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 [13]. Lastly, the sugg use of recurrent neural networks (RNNs) ~~allow~~ allows 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.

Agree

They need to be parallel!

Agree

Also, the data need is no different to statistical phrase-based MT, so not an advantage

maybe have some sections and start second one here?

Despite all these advantages and potentials, the early NMT architecture [6, 26] 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~~ one of these problems. ~~in which~~ In each chapter 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 [20], English-German [16, 19], and English-Czech [17]. 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 [21]; and (2) whether we can compress NMT for mobile devices [25]. 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~~ to the future of NMT 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~~ the first 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~~ I 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

X

X

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~~ My 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~~ this NMT system is the first to surpass the best result achieved on a WMT'14 contest task. This chapter is based on work with ____ first published as ____ .

Do this for each section in Chapter 1. It's required.

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. [1] 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

not mentioned before

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

Why are these important questions?

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.

Throughout, dissertation should basically be 1st person, unless contextually specific reference to others.

For the first question, ~~we~~ I 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: an autoencoder helps less in terms of perplexit~~ies~~y 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.

Some conclusion on what has been learned and what lies ahead

See also notes on bibliography!

Chapter 2

Background

Chapter 3

Copy Mechanisms

Chapter 4

Attention Mechanisms

Chapter 5

Hybrid Models

Chapter 6

NMT Future

Chapter 7

Conclusion

Bibliography

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

Consistently use initials or full first names. Full first names is better

[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. 3:1137–1155, 2003.

[3] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. 19(2):263–311, 06 1993.

[4] D. Cer, M. Galley, D. Jurafsky, and C. D. Manning. Phrasal: A statistical machine translation toolkit for exploring new model features. In *ACL, Demonstration Session*, 2010.

[5] David Chiang. Hierarchical phrase-based translation. 33(2):201–228, 2007.

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

[7] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul. Fast and robust neural network joint models for statistical machine translation. In *ACL*, 2014.

[8] Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACL, Demonstration Session*, 2010.

[9] Mikel L. Forcada and Ramón Neco. Recursive hetero-associative memories for translation. pages 453–462, 1997.

[10] Kenneth Heafield. KenLM: faster and smaller language model queries. In *WMT*, 2011.

[11] W. John Hutchins. Machine translation: A concise history, 2007.

[12] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *EMNLP*, 2013.

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

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *ACL, Demonstration Session*, 2007.

[15] Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. An end-to-end discriminative approach to machine translation. In *ACL*, 2006.

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

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

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

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

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

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

[22] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. 29(1):19–51, 2003.

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

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

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

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

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

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