

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

July 2016

© Copyright by Minh-Thang Luong 2016
All Rights Reserved

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

1	Introduction	1
1.1	Machine Translation Development	2
1.2	Thesis Outline	5
2	Background	9
2.1	Recurrent Neural Network	10
2.1.1	Recurrent Language Models	12
2.1.2	Better Training RNNs	16
2.2	Neural Machine Translation	19
2.2.1	Training	20
2.2.2	Testing	21
3	Copy Mechanisms	22
4	Attention Mechanisms	23
5	Hybrid Models	24
6	NMT Future	25
7	Conclusion	26
A	Miscellaneous	27

List of Tables

List of Figures

1.1	A general setup of machine translation	2
1.2	Phrase-based machine translation	3
1.3	Source-conditioned neural probabilistic language models	4
1.4	Neural machine translation	5
2.1	Recurrent neural networks	11
2.2	Recurrent language models	12
2.3	Neural machine translation	19
2.4	Neural machine translation	21

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 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 [21] from the original philosophical 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 [57]. Despite several excitement moments that led to hopes that MT will be solved “very soon”, e.g., the 701 translator² developed by scientists at George Town and IBM in the 1950s or a simple vector-space transformation technique³ proposed by Google researchers at the beginning of

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

²http://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/>

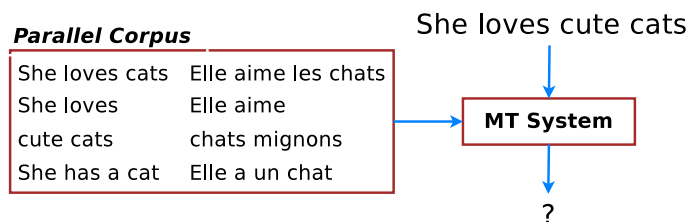


Figure 1.1: **Machine translation** (MT) – a general setup of MT. Systems build translation models from parallel corpora to translate new unseen sentences, e.g., “She loves cute cats”.

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.

1.1 Machine Translation Development

Modern statistical MT started out with a seminal work by IBM scientists [7]. 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 [8, 9, 12, 25, 26, 27, 42], 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.

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

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

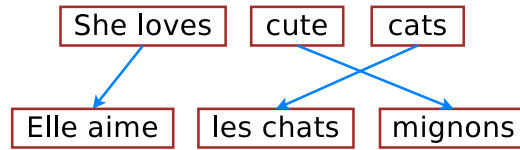


Figure 1.2: **Phrase-based machine translation** (MT) – example of how phrase-based MT systems translate a source sentence “She loves cute cats” into a target sentence “Elle aime les chats mignons”: sentences are split into chunks and phrases are translated.

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. [5] 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 [18], 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, [30, 47, 56], inter alia, start adopting NPLMs alongside with traditional LMs. To make NPLMs even more powerful, recent work [1, 11, 48, 50] propose to condition on source words beside the target context to lower uncertainty in predicting next words (see Figure 1.3).⁵

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 dependencies. For example, in Figure 1.3, 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”.

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

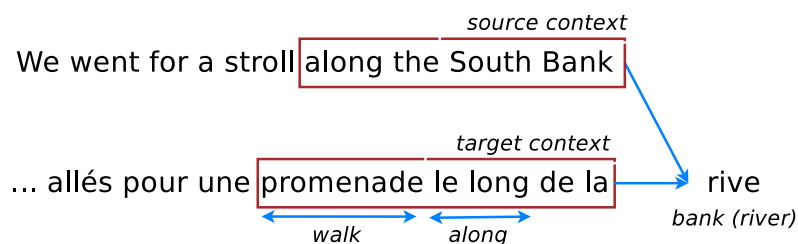


Figure 1.3: **Source-conditioned neural probabilistic language models** (NPLMs) – example of a source-conditioned NPLM proposed by Devlin et al. [11]. To evaluate a how likely a next word “rive” is, the model not only relies on previous target words (context) “promenade le long de la” as in traditional NPLMs [5], but also utilizes source context “along the South Bank” to lower uncertainty in its prediction.

More problematically, the entire MT pipeline is already complex with different components needed to be tuned separately, 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 [10, 23, 54]. This is often referred 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.

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 [25]. 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

⁶Forcada and Neco [15] wrote the very first paper on sequence-to-sequence models for translation!

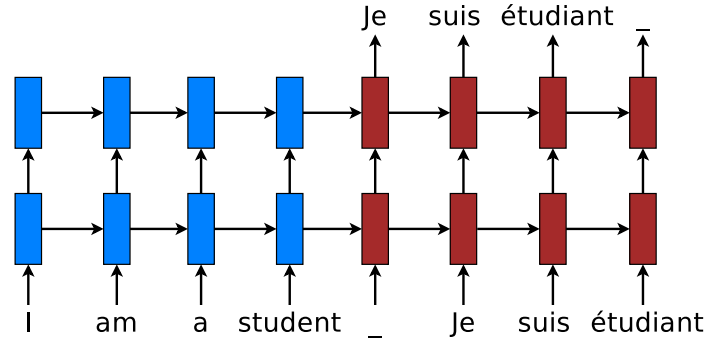


Figure 1.4: **Neural machine translation** – example of a deep recurrent architecture proposed by Sutskever et al. [54] for translating a source sentence “I am a student” into a target sentence “Je suis étudiant”. Here, “_” marks the end of a sentence.

tables or language models as in the case of standard MT.

1.2 Thesis Outline

Despite all the aforementioned advantages and potentials, the early NMT architecture [10, 54] 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 [32], English-German [28, 31], and English-Czech [29]. 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 [33]; and (2) whether we can compress NMT for mobile devices [49]. 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. [2] 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 mechanisms 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

For neural machine translation, it all started from language modeling.

Thang Luong.

Language modeling plays an indispensable role in ensuring that machine translation systems produce fluent target sentences and has always been an active area of research. Despite much effort in improving traditional n -gram language models [14, 18, 19, 44, 45, 51, 55], traditional LMs inherently can only handle short contexts of a few words. Approaches to building neural probabilistic language models (NPLMs) using feed-forward networks such as those initiated by Bengio et al. [5] and enhanced by others [3, 39, 40, 41] have addressed that drawback to model longer contexts. Still, NPLMs can only capture fixed-length contexts and is incapable of handling variable-length sequences, which is the case for sentences. Recurrent neural networks (RNNs) come in handy as a powerful and expressive architecture to handle sequential data and have successfully been applied to the language modeling task [36, 37, 38]. By viewing RNNs as generative models [53] that can produce texts and by pushing another step towards conditioning RNNs on source sentences, recent works [10, 23, 54] have started a new line of research in machine translation, namely Neural Machine Translation (NMT). NMT is technically a source-conditioned NPLM that can be trained end-to-end.

In this chapter, we provide background knowledge on two main topics, RNN and NMT.

We first go through the basics of RNNs, explaining how they can be used to model sentences. Then, we delve into details of one particular type of RNNs, the Long Short-term Memory, that makes training RNNs easier. Given RNNs as a building block, we discuss NMT together with tips and tricks for better training and testing NMT.

2.1 Recurrent Neural Network

Recurrent Neural Network (RNNs) [13] are models that help understand the temporal aspect as well as build up representations for sequential data using a dynamic memory structure. At the surface form, an RNN takes as input a sequence of vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ and processes them one by one. For each new input \mathbf{x}_i , an RNN updates its memory to produce a hidden state \mathbf{h}_i which one can think of as a representation for the partial sequence $\mathbf{x}_{1:i}$. The beauty of RNNs lies in the fact that it can capture the dynamics of an arbitrarily long sequence without having to increase its modeling capacity unlike the case of feedforward network which can only model relationship within a fixed-length sequence. The key secret sauce is in the recurrence formula of an RNN that defines how its hidden state is updated. At its simplest form, a “vanilla” RNN defines its recurrence function as:

$$\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1}) \quad (2.1)$$

In the above formula, f is an abstract function that computes a new hidden state given the current input \mathbf{x}_t and the previous hidden state \mathbf{h}_{t-1} . The starting state \mathbf{h}_0 is often set to $\mathbf{0}$ though it can take any value as we will see later in the context of NMT decoders. A popular choice of f is provided below with σ being a non-linear function such as sigmoid or tanh.¹

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}) \quad (2.2)$$

At each timestep t , an RNN can (optionally) emit an output symbol y_t which can either be discrete or real-valued. For the discrete scenario, which is often the case for languages,

¹There could also be an optional bias term in Eq. (2.2).

a probability distribution \mathbf{p} over a set of output classes Y is derived as follows²:

$$\mathbf{s}_t = \mathbf{W}_{hy} \mathbf{h}_t \quad (2.3)$$

$$\mathbf{p}_t = \text{softmax}(\mathbf{s}_t) \quad (2.4)$$

Here, we introduce a new set of weights $\mathbf{W}_{hy} \in \mathbb{R}^{|Y| \times d}$, with d being the dimension of the RNN hidden state, to compute a score vector \mathbf{s}_t , or *logits*, over different individual classes. Often, with a large output set Y , the matrix-vector multiplication in Eq. (2.3) is a major computational bottleneck in RNNs, which results in several challenges for neural language modeling and machine translation that we will address in later chapters. The softmax function transforms the score vector \mathbf{s}_t into a probability vector \mathbf{p}_t , which is defined for each specific element $y \in Y$ as below. For convenience, we overload our notations to use $\mathbf{p}_t(y)$ and $\mathbf{s}_t(y)$ to refer to entries in the vectors \mathbf{p}_t and \mathbf{s}_t that correspond to y .

$$\mathbf{p}_t(y) = \frac{e^{\mathbf{s}_t(y)}}{\sum_{y' \in Y} e^{\mathbf{s}_t(y')}} \quad (2.5)$$

With the above formulae, we have completely defined the RNN weight set θ which consists of *input* connections \mathbf{W}_{xh} , *recurrent* connections \mathbf{W}_{hh} , and *output* connections \mathbf{W}_{hy} . These weights are shared across timesteps as illustrated in Figure 2.1 **Draw a picture on general RNNs**, which enables RNNs to handle arbitrarily long sequences.

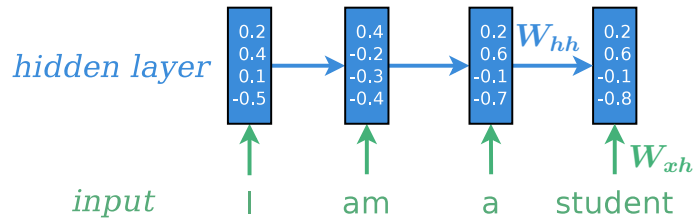


Figure 2.1: **Recurrent neural networks** – example of a recurrent neural network that processes a sequence of input words “I am a student” to build up hidden representations as input symbols are consumed. The recurrent \mathbf{W}_{hh} and feed-forward \mathbf{W}_{xh} weights are shared across timesteps.

²For the real-valued case, we refer readers to mixture density models [6] which have been applied to RNN training, e.g., for hand-writing synthesis [17].

Next, we discuss the training and testing phases of RNNs from a slightly more focused angle, the language learning aspect. For more details on RNNs, we refer readers to the following resources [24, 35, 52].

2.1.1 Recurrent Language Models

To apply RNNs to sentences in languages, or generally sequences of discrete symbols, one can consider one-hot representations $\mathbf{x}_i \in \mathbb{R}^{|V|}$, with V being the vocabulary considered. However, for a large vocabulary V , such a representation choice is problematic as it results in a large weight matrix \mathbf{W}_{xh} and there is no notion of similarity between words. In practice, low-dimensional dense representations for words, or *word embeddings*, are often used to address these problems. Specifically, an embedding matrix $\mathbf{W}_e \in \mathbb{R}^{d_e \times |V|}$ is looked up for each word x_i to retrieve a representation $\mathbf{x}_i \in \mathbb{R}^{d_e}$. As a result, a simple RNN applied to language modeling will generally have $\theta = \{\mathbf{W}_{xh}, \mathbf{W}_{hh}, \mathbf{W}_{hy}, \mathbf{W}_e\}$ as its weights as illustrated in Figure 2.2 **Draw an RNN with embedding**.

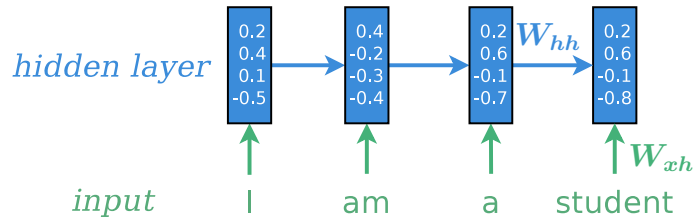


Figure 2.2: **Recurrent language models** – example of a recurrent neural network that processes a sequence of input words “I am a student” to build up hidden representations as input symbols are consumed. The recurrent \mathbf{W}_{hh} and feed-forward \mathbf{W}_{xh} weights are shared across timesteps.

In language modeling (LM), the task is to specify a probability distribution over sequences of symbols (often, words) so that one can judge if a sequence of words is more likely or “fluent” than another. To accomplish that, an LM decomposes the probability of a word sequence $y = y_1, \dots, y_m$ as:

$$p(y) = \prod_{i=1}^m p(y_i | y_{<i}) \quad (2.6)$$

In the above formula, each of the individual term $p(y_i|y_{<i})$ is the conditional probability of the current word y_i given previous words $y_{<i}$, also referred as the *context* or the *history*. To model these conditional probabilities, traditional n -gram as well as feedforward-based neural language models have to resort to the Markovian assumption to model only a fixed window of context, i.e., $p(y_i|y_{i-n+1}, \dots, y_{i-1})$. An RNN-based language model naturally lends itself to model the full history as we shall see now.

An RNN-based language model (RNNLM) is a special case of RNNs in which: (a) the input and output are sequences of discrete words, (b) the output sequence ends with a special symbol $\langle \text{eos} \rangle$ that marks the boundary, e.g., $y = \{ \text{"I"}, \text{"am"}, \text{"a"}, \text{"student"}, \langle \text{eos} \rangle \}$, and (c) the input sequence is a shift-by-1 version of the output sequence with $\langle \text{sos} \rangle$ as a starting symbol, e.g., $x = \{ \langle \text{sos} \rangle, \text{"I"}, \text{"am"}, \text{"a"}, \text{"student"} \}$. We illustrate this in Figure 2.2.

Training Given a training dataset of N discrete output sequences $y^{(1)}, \dots, y^{(N)}$ with lengths m_1, \dots, m_N accordingly. The learning objective is to minimize the negative log-likelihood, or the *cross-entropy* loss, of these training examples:

$$J(\theta) = \sum_{i=1}^N -\log p(y^{(i)}) \quad (2.7)$$

$$= \sum_{i=1}^N \sum_{t=1}^{m_i} -\log p(y_t^{(i)} | y_{<t}^{(i)}) \quad (2.8)$$

RNN learning is often done using mini-batch stochastic gradient descent (SGD) algorithms in which a small set of training examples, a *mini-batch*, is used to compute the gradients and update weights one at a time. Using mini-batches has several advantages: (a) the gradients are more reliable and consistent than the “online” setting which updates per example, (b) less computation is required to update the weights unlike the case of full-batch learning which has to process all examples before updating, and (c) with multiple examples in a mini-batch, one can turn matrix-vector multiplications such as those in Eq. (2.2) and Eq. (2.3) into matrix-matrix multiplications which can be deployed efficiently on GPUs.

The simplest weight update formula with η as a learning rate is given below:

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} - \eta \nabla J(\boldsymbol{\theta}) \quad (2.9)$$

Single-timestep Backpropagation To compute the gradients for the loss $J(\boldsymbol{\theta})$, we first need to be able to derive the gradients of the per-timestep loss $l_t = \log \mathbf{p}_t(y_t)$ with respect to both the RNN weights $\{\mathbf{W}_{xh}, \mathbf{W}_{hh}, \mathbf{W}_{hy}\}$ and the inputs $\{\mathbf{x}_t, \mathbf{h}_{t-1}\}$. We denote these gradients as $\{d\mathbf{W}_{xh}, d\mathbf{W}_{hh}, d\mathbf{W}_{hy}, d\mathbf{x}_t, d\mathbf{h}_{t-1}\}$ respectively and define intermediate gradients $d\mathbf{s}_t, d\mathbf{h}_t$ similarly. Starting with the loss l_t , we employ backpropagation through structures [16] to derive each gradient one by one in the following order: $l_t \rightarrow \mathbf{s}_t \rightarrow \{\mathbf{h}_t, \mathbf{W}_{hy}\} \rightarrow \{\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{W}_{xh}, \mathbf{W}_{hh}\}$. First, from Eq. (2.5), we have:

$$d\mathbf{s}_t = \frac{\partial l_t}{\partial \mathbf{s}_t} = \frac{\partial}{\partial \mathbf{s}_t} \left(\mathbf{s}_t(y_t) - \log \sum_{y'} e^{\mathbf{s}_t(y')} \right) \quad (2.10)$$

Computing per-coordinate gradient $\mathbf{s}_t(y)$ gives:

$$\frac{\partial}{\partial \mathbf{s}_t(y)} \left(\mathbf{s}_t(y_t) - \log \sum_{y'} e^{\mathbf{s}_t(y')} \right) = \begin{cases} 1 - \mathbf{p}_t(y_t) & y = y_t \\ -\mathbf{p}_t(y) & y \neq y_t \end{cases} \quad (2.11)$$

The above gradients can be concisely written in vector form as:

$$d\mathbf{s}_t = \mathbf{1}_{y_t} - \mathbf{p}_t \quad (2.12)$$

Here, \mathbf{p}_t is the probability distribution defined in Eq. (2.4) and has been calculated in the forward pass, so we simply reuse it. $\mathbf{1}_{y_t}$ is a one-hot vector with 1 at position y_t . Applying Corollary 1, noting that $\mathbf{s}_t = \mathbf{W}_{hy}\mathbf{h}_t$ in Eq. (2.3), we arrive at:

$$d\mathbf{h}_t = \mathbf{W}_{hy}^\top \cdot d\mathbf{s}_t \quad (2.13)$$

$$d\mathbf{W}_{hy} = d\mathbf{s}_t \cdot \mathbf{h}_t^\top \quad (2.14)$$

At this point, we have derived part of the backpropagation procedure which can be applied to any hidden unit type, e.g., the aforementioned vanilla RNN or the LSTM unit that we

will describe shortly in the next section.

Vanilla RNN Backpropagation First of all, we can simplify the notation to have $\mathbf{T}_{\text{rnn}} = [\mathbf{W}_{xh} \mathbf{W}_{hh}]$ and $\mathbf{z}_t = [\mathbf{x}_t; \mathbf{h}_{t-1}]$, so the RNN formulation in Eq. (2.2) becomes:

$$\mathbf{h}_t = \sigma(\mathbf{T}_{\text{rnn}} \mathbf{z}_t) \quad (2.15)$$

Applying Lemma 2, we have:

$$d\mathbf{z}_t = \mathbf{T}_{\text{rnn}}^\top \cdot (\sigma'(\mathbf{T}_{\text{rnn}} \mathbf{z}_t) \odot d\mathbf{h}_t) \quad (2.16)$$

$$d\mathbf{T}_{\text{rnn}} = (\sigma'(\mathbf{T}_{\text{rnn}} \mathbf{z}_t) \odot d\mathbf{h}_t) \cdot \mathbf{z}_t^\top \quad (2.17)$$

This is one of the *tricks* that we use to better utilize GPUs by creating larger matrices and vectors, i.e., \mathbf{T}_{rnn} and \mathbf{z}_t . From Eq. (2.16) and Eq. (2.17), one can easily extract the following gradients: (a) $d\mathbf{x}_t$ – embedding gradients which we use to sparsely update the embedding weights \mathbf{W}_e , (b) $d\mathbf{h}_{t-1}$ – gradients of the previous hidden state, which is needed by the backpropagation-through-time algorithm that we will discuss next, and (c) $d\mathbf{W}_{xh}$ as well as $d\mathbf{W}_{hh}$ – the RNN input and recurrent connections.³

Backpropagation Through Time (BPTT) Having defined a single-timestep backpropagation procedure, we are now ready to go through the BPTT algorithm [46, 58]. Inspired by Sutskever [52], we summarize the BPTT algorithm for RNNs below with the following remarks: (a) Lines 3, 5, 6, 7 accumulate the gradients of RNN weights $\{\mathbf{W}_{hy}, \mathbf{W}_{xh}, \mathbf{W}_{hh}, \mathbf{W}_e\}$ over time; (b) In line 7, $d\mathbf{x}_t$ refers to gradients of words participating in the current mini-batch which we use to sparsely update \mathbf{W}_e ;⁴ and (c) Line 4 accumulates gradients for the current hidden state \mathbf{h}_t by considering two paths, a “vertical” one from the current loss at

³One can also separately derive these gradients as follows:

$$\begin{aligned} d\mathbf{x}_t &= \mathbf{W}_{xh}^\top \cdot (\sigma'(\mathbf{T}_{\text{rnn}} \mathbf{z}_t) \odot d\mathbf{h}_t) \\ d\mathbf{h}_{t-1} &= \mathbf{W}_{hh}^\top \cdot (\sigma'(\mathbf{T}_{\text{rnn}} \mathbf{z}_t) \odot d\mathbf{h}_t) \\ d\mathbf{W}_{xh} &= (\sigma'(\mathbf{T}_{\text{rnn}} \mathbf{z}_t) \odot d\mathbf{h}_t) \cdot \mathbf{x}_t^\top \\ d\mathbf{W}_{hh} &= (\sigma'(\mathbf{T}_{\text{rnn}} \mathbf{z}_t) \odot d\mathbf{h}_t) \cdot \mathbf{h}_{t-1}^\top \end{aligned}$$

⁴In multi-layer RNNs, $d\mathbf{x}_t$ is used to send gradients down to the below layers.

time t and a “recurrent” one from the timestep $t + 1$ which was set in Line 8 earlier.

Algorithm 1: BPTT algorithm for “vanilla” RNNs

```

1 for  $t = T \rightarrow 1$  do
    // Output backprop
2    $ds_t \leftarrow \mathbf{1}_{y_t} - \mathbf{p}_t$ 
3    $d\mathbf{W}_{hy} \leftarrow d\mathbf{W}_{hy} + ds_t \cdot \mathbf{h}_t^\top$ 
4    $d\mathbf{h}_t \leftarrow d\mathbf{h}_t + \mathbf{W}_{hy}^\top \cdot ds_t$ 
    // RNN backprop
5    $d\mathbf{W}_{xh} \leftarrow d\mathbf{W}_{xh} + (\sigma'(\mathbf{T}_{\text{rnn}}\mathbf{z}_t) \odot d\mathbf{h}_t) \cdot \mathbf{x}_t^\top$ 
6    $d\mathbf{W}_{hh} \leftarrow d\mathbf{W}_{hh} + (\sigma'(\mathbf{T}_{\text{rnn}}\mathbf{z}_t) \odot d\mathbf{h}_t) \cdot \mathbf{h}_{t-1}^\top$ 
    // Input backprop
7    $d\mathbf{x}_t \leftarrow \mathbf{W}_{xh}^\top \cdot (\sigma'(\mathbf{T}_{\text{rnn}}\mathbf{z}_t) \odot d\mathbf{h}_t)$ 
8    $d\mathbf{h}_{t-1} \leftarrow \mathbf{W}_{hh}^\top \cdot (\sigma'(\mathbf{T}_{\text{rnn}}\mathbf{z}_t) \odot d\mathbf{h}_t)$ 
9 end
```

2.1.2 Better Training RNNs

Even though computing RNN gradients is straightforward once the BPTT algorithm has been plotted out, training is inherently difficult due to the nonlinear iterative nature of RNNs. Among all reasons, the two classic problems of RNNs that often arise when dealing with very long sequences are the *exploding* and *vanishing* gradients as described by Bengio et al. [4]. In short, exploding gradients refer to the phenomenon that the gradients become exponentially large as we backpropagate over time, making learning unstable. Vanishing gradients, on the other hand, is the opposite problem when the gradients go exponentially fast towards zero, turning BPTT into truncated BPTT that is unable to capture long-range dependencies in sequences.

Let us try to explain the aforementioned problems informally and refer readers to more rigorous and in-depth analyses in [4, 20, 34, 43]. The main cause of these two problems all lies in Line 8 of the BPTT algorithm, in which we can rewrite the assignment as $d\mathbf{h}_{t-1} = \mathbf{W}_{hh}^\top \cdot \text{diag}(\sigma'(\mathbf{T}_{\text{rnn}}\mathbf{z}_t)) \cdot d\mathbf{h}_t$. We can try to understand the behavior of RNNs over time by assuming no contribution of the intermediate losses in Line 4 for a moment. Given such an assumption, a signal sent from the current hidden state over K steps will become

$d\mathbf{h}_{t-K} = \prod_{i=1}^K (\mathbf{W}_{hh}^\top \cdot \text{diag}(\sigma'(\mathbf{T}_{\text{rnn}} \mathbf{z}_{t-i+1}))) \cdot d\mathbf{h}_t$. Assuming that the non-linear function σ is bounded, e.g., sigm and tanh , and behaves “nicely”, what we need to deal with now is the multiplication of the recurrent matrix over time. This leads to the fact that the behavior of RNNs is often governed by the characteristics of the recurrent matrix \mathbf{W}_{hh} , which most analysis examines in terms of the largest eigen value of \mathbf{W}_{hh} and norms of the signals. Roughly speaking, if the largest eigen value is large enough, exploding gradients will be likely to happen. On the contrary, if the largest eigen value is below a certain threshold, vanishing gradients will occur as clearly explained by Pascanu et al. [43].

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}) \quad (2.18)$$

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} = \text{diag}(\sigma'(\dots)) \mathbf{W}_{hh}^\top \quad (2.19)$$

$$\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \right\| \leq \|\text{diag}(\sigma'(\dots))\| \|\mathbf{W}_{hh}^\top\| \quad (2.20)$$

$$\leq \gamma \lambda_1 \quad (2.21)$$

$$\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-k}} \right\| \leq (\gamma \lambda_1)^k \rightarrow 0 \quad \text{if } \lambda_1 < \frac{1}{\gamma} \quad (2.22)$$

$$\frac{\partial \mathbf{c}_t}{\partial \mathbf{c}_{t-1}} = \mathbf{I} \quad (2.23)$$

Long Short-Term Memory We use the formulation of [59]. For a single LSTM block at layer l and time t , the new hidden state \mathbf{h}_t^l and memory cell \mathbf{c}_t^l are calculated from \mathbf{h}_t^{l-1} ,

h_{t-1}^l and c_{t-1}^l like so:

$$\begin{pmatrix} i_t^l \\ f_t^l \\ o_t^l \\ \hat{h}_t^l \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} T_{\text{lstm}} \begin{bmatrix} h_t^{l-1} \\ h_{t-1}^l \end{bmatrix} \quad (2.24)$$

$$c_t^l = f_t^l \circ c_{t-1}^l + i_t^l \circ \hat{h}_t^l \quad (2.25)$$

$$h_t^l = o_t^l \circ \tanh(c_t^l) \quad (2.26)$$

where sigm and tanh are applied element-wise, \circ denotes element-wise multiplication, and $T_{4n \times 2n}$ is a $4n \times 2n$ matrix of weights that depends on l but not t .⁵ If $l = 1$ then h_t^{l-1} is the input vector x_t . If $t = 1$ then h_{t-1}^l and c_{t-1}^l are taken to be zero.

An LSTM block at layer $l \in \{1, \dots, L\}$ and time $t \in \{1, \dots, T\}$ consists of:

- The hidden state $h_t^l \in \mathbb{R}^n$
- The memory cell $c_t^l \in \mathbb{R}^n$
- The input gate $i_t^l \in [0, 1]^n$
- The forget gate $f_t^l \in [0, 1]^n$
- The output gate $o_t^l \in [0, 1]^n$
- The input modulation gate $\hat{h}_t^l \in [0, 1]^n$

We call n the LSTM block size.

LSTM Backpropagation $\delta_{c^{(2)}}$

$$\delta_{h^{(2)}}$$

$$\delta_{c^{(1)}}$$

$$\delta_{h^{(1)}}$$

$$\delta_c += \delta_h o_t \tanh'(c_t)$$

⁵Note: Sometimes these equations are written omitting the superscript l and writing h_t^{l-1} as x_t , but for the purposes of deriving the back-propagation equations, we need to refer to the layer l explicitly.

$$\delta_c = \delta_c \circ f_t$$

$$\delta_h += \text{upper grad}$$

2.2 Neural Machine Translation

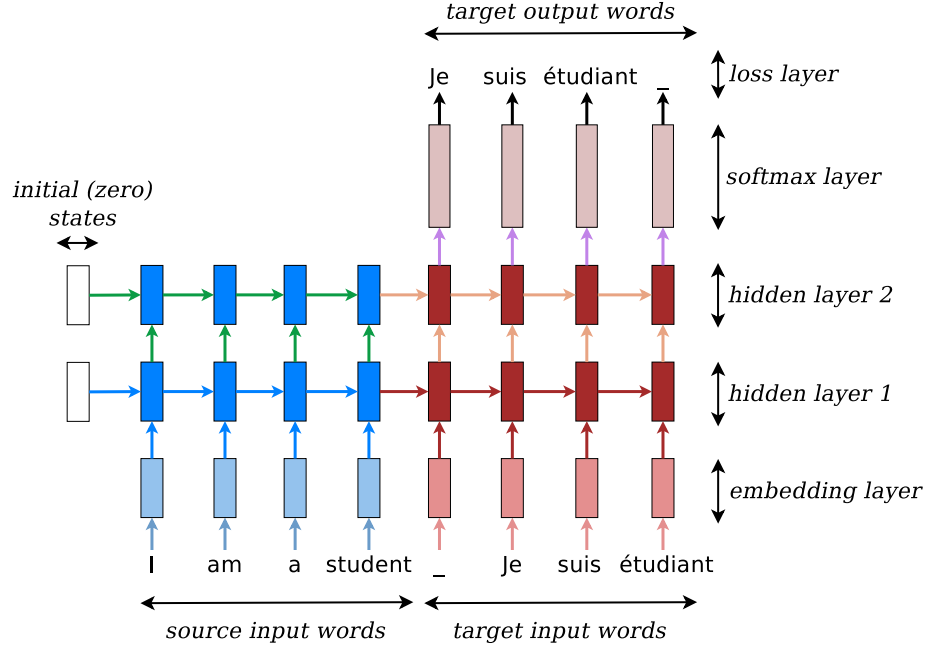


Figure 2.3: **Neural machine translation** – example of a deep recurrent architecture proposed by Sutskever et al. [54] for translating a source sentence “I am a student” into a target sentence “Je suis étudiant”. Here, “_” marks the end of a sentence.

Neural machine translation aims to directly model the conditional probability $p(y|x)$ of translating a source sentence, x_1, \dots, x_n , to a target sentence, y_1, \dots, y_m . It accomplishes such goal through the *encoder-decoder* framework [10, 54]. The *encoder* computes a representation s for each source sentence. Based on that source representation, the *decoder* generates a translation, one target word at a time, and hence, decomposes the conditional probability as:

$$\log p(y|x) = \sum_{j=1}^m \log p(y_j | y_{<j}, x, s) \quad (2.27)$$

A natural choice to model such a decomposition in the decoder is to use a recurrent

neural network (RNN) architecture, which most of the recent NMT work have in common. They, however, differ in terms of the RNN architectures used and how the encoder computes the source representation \mathbf{s} .

Kalchbrenner and Blunsom [23] used an RNN with the vanilla RNN unit for the decoder and a convolutional neural network for encoding the source. On the other hand, Sutskever et al. [54] and Luong et al. [31, 32] built deep RNNs with the Long Short-Term Memory (LSTM) unit [20] for both the encoder and the decoder. Cho et al., [10], Bahdanau et al., [2], and Jean et al. [22] all adopted an LSTM-inspired hidden unit, the gated recurrent unit (GRU), and used bidirectional RNNs for the encoder.

In more details, considering the top recurrent layer in a deep RNN architecture, one can compute the probability of decoding each target word y_j as:

$$p(y_j | y_{<j}, x, \mathbf{s}) = \text{softmax}(\mathbf{h}_j) \quad (2.28)$$

with \mathbf{h}_j being the current target hidden state computed as:

$$\mathbf{h}_j = f(\mathbf{h}_{j-1}, y_{j-1}, \mathbf{s}) \quad (2.29)$$

Here, f derives the current state given the previous state \mathbf{h}_{j-1} , the current input (often the previous word y_{t-1}), and optionally, the source representation \mathbf{s} . f can be a vanilla RNN unit, a GRU, or an LSTM. The early NMT approach [10, 23, 32, 54] uses the last source hidden state $\mathbf{s} = \bar{\mathbf{h}}_n$ once to initialize the decoder hidden state and sets $\mathbf{s} = [\]$ in Eq. (2.29).

2.2.1 Training

The training objective is formulated as follows:

$$J = \sum_{(x,y) \in \mathbb{D}} -\log p(y|x) \quad (2.30)$$

with \mathbb{D} being our parallel training corpus.

mention bucketing and batching

2.2.2 Testing

mention beam-search, ensemble

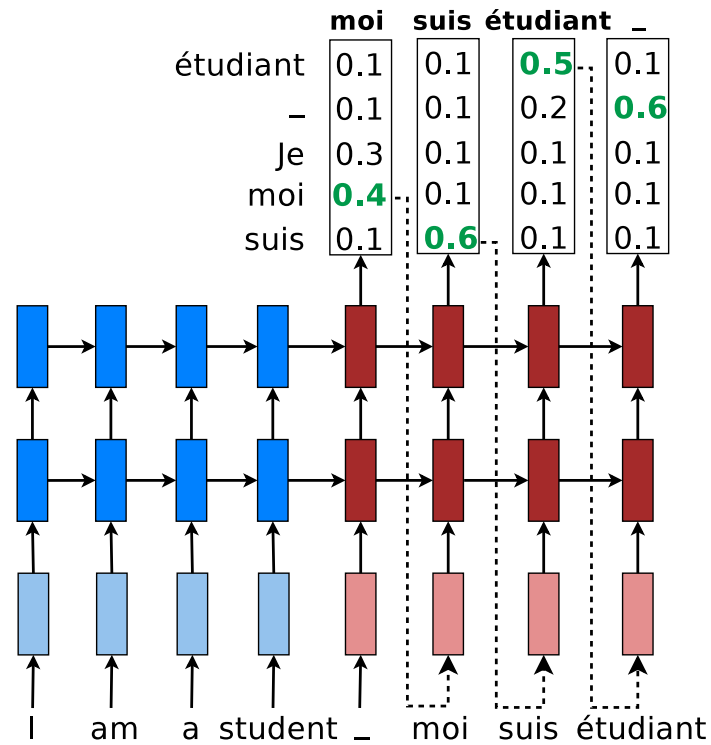


Figure 2.4: **Neural machine translation** – example of a deep recurrent architecture proposed by Sutskever et al. [54] for translating a source sentence “I am a student” into a target sentence “Je suis étudiant”. Here, “_” marks the end of a sentence.

Chapter 3

Copy Mechanisms

Chapter 4

Attention Mechanisms

Chapter 5

Hybrid Models

Chapter 6

NMT Future

Chapter 7

Conclusion

Appendix A

Miscellaneous

Lemma 1. *Let \mathbf{u}, \mathbf{v} be any vectors and \odot be element-wise vector multiplication, we have:*

$$\text{diag}(\mathbf{u}) \cdot \mathbf{v} = \mathbf{u} \odot \mathbf{v} \quad (\text{A.1})$$

Lemma 2. *Let l be a loss value that in which we already knew how to compute its gradient $d\mathbf{v}$ with respect to a vector \mathbf{v} . Given that $\mathbf{v} = f(\mathbf{W}\mathbf{h})$, the gradients $d\mathbf{h}, d\mathbf{W}$ of the loss l with respect to the vector \mathbf{h} and the matrix \mathbf{W} can be derived as follows:*

$$d\mathbf{h} = \mathbf{W}^\top \cdot (f'(\mathbf{W}\mathbf{h}) \odot d\mathbf{v}) \quad (\text{A.2})$$

$$d\mathbf{W} = (f'(\mathbf{W}\mathbf{h}) \odot d\mathbf{v}) \cdot \mathbf{h}^\top \quad (\text{A.3})$$

Proof. Let $\mathbf{z} = \mathbf{W}\mathbf{h}$, we have the following derivations:

$$\begin{aligned} d\mathbf{h} &= \frac{\partial \mathbf{z}}{\partial \mathbf{h}} \cdot \frac{\partial \mathbf{v}}{\partial \mathbf{z}} \cdot d\mathbf{v} && [\text{Vector calculus chain rules}] \\ &= \frac{\partial \mathbf{W}\mathbf{h}}{\partial \mathbf{h}} \cdot \frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \cdot d\mathbf{v} \\ &= \mathbf{W}^\top \cdot \text{diag}(f'(\mathbf{z})) \cdot d\mathbf{v} \\ &= \mathbf{W}^\top \cdot (f'(\mathbf{W}\mathbf{h}) \odot d\mathbf{v}) && [\text{Lemma 1}] \end{aligned}$$

Let \mathbf{w}_i^\top be the i^{th} row vector of matrix \mathbf{W} and v_i, z_i be the i^{th} elements of vectors \mathbf{v}, \mathbf{z} .

Also denoting $d\mathbf{w}_i, dv_i$ to be the gradients of l with respect to \mathbf{w}_i, v_i , we have:

$$\begin{aligned}
 d\mathbf{w}_i &= \frac{\partial z_i}{\partial \mathbf{w}_i} \cdot \frac{\partial v_i}{\partial z_i} \cdot dv_i && \text{[Vector calculus chain rules]} \\
 &= \frac{\partial \mathbf{w}_i^\top \mathbf{h}}{\partial \mathbf{w}_i} \cdot f'(z_i) \cdot dv_i \\
 &= \mathbf{h} \cdot f'(z_i) \cdot dv_i \\
 d\mathbf{w}_i^\top &= (f'(z_i) \cdot dv_i) \cdot \mathbf{h}^\top && \text{[Tranposing]} \\
 d\mathbf{W} &= (f'(\mathbf{W}\mathbf{h}) \odot d\mathbf{v}) \cdot \mathbf{h}^\top && \text{[Concatenating row derivatives]}
 \end{aligned}$$

□

Corollary 1. *As a special case of Lemma 2, when f is an identity function, i.e., $\mathbf{v} = \mathbf{W}\mathbf{h}$, we have:*

$$d\mathbf{h} = \mathbf{W}^\top \cdot d\mathbf{v} \tag{A.4}$$

$$d\mathbf{W} = d\mathbf{v} \cdot \mathbf{h}^\top \tag{A.5}$$

Bibliography

- [1] Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. Joint language and translation modeling with recurrent neural networks. In *ACL*, 2013.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.
- [3] Yoshua Bengio and Jean-Sébastien Senécal. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Trans. Neural Networks*, 19(4):713–722, 2008.
- [4] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2): 157–166, 1994.
- [5] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *JMLR*, 3:1137–1155, 2003.
- [6] Christopher M. Bishop. Mixture density networks. Technical report, Aston University, 1994.
- [7] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 06 1993.
- [8] Daniel Cer, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. Phrasal: A statistical machine translation toolkit for exploring new model features. In *ACL, Demonstration Session*, 2010.

- [9] David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33 (2):201–228, 2007.
- [10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 2014.
- [11] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *ACL*, 2014.
- [12] 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.
- [13] Jeffrey L. Elman. Finding structure in time. In *Cognitive Science*, 1990.
- [14] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech*, 2008.
- [15] Mikel L. Forcada and Ramón Neco. Recursive hetero-associative memories for translation. *Biological and Artificial Computation: From Neuroscience to Technology*, pages 453–462, 1997.
- [16] C. Goller and A. Kehler. Learning task-dependent distributed representations by back-propagation through structure. *IEEE Transactions on Neural Networks*, 1:347–352, 1996.
- [17] A. Graves. Generating sequences with recurrent neural networks. In *Arxiv preprint arXiv:1308.0850*, 2013.
- [18] Kenneth Heafield. KenLM: faster and smaller language model queries. In *WMT*, 2011.

- [19] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *ACL*, 2013.
- [20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [21] W. John Hutchins. *Machine translation: A concise history*, 2007.
- [22] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *ACL*, 2015.
- [23] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *EMNLP*, 2013.
- [24] Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks. <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>, 2015. Accessed: 2016-07-05.
- [25] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL*, 2003.
- [26] 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.
- [27] Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. An end-to-end discriminative approach to machine translation. In *ACL*, 2006.
- [28] Minh-Thang Luong and Christopher D. Manning. Stanford neural machine translation systems for spoken language domain. In *IWSLT*, 2015.
- [29] Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *ACL*, 2016.
- [30] Minh-Thang Luong, Michael Kayser, and Christopher D. Manning. Deep neural language models for machine translation. In *CoNLL*, 2015.

- [31] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.
- [32] 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.
- [33] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *ICLR*, 2016.
- [34] James Martens and Ilya Sutskever. Learning recurrent neural networks with Hessian-free optimization. In *ICML*, 2011.
- [35] Tomáš Mikolov. *Statistical Language Models Based on Neural Networks*. PhD thesis, Brno University of Technology, 2012.
- [36] Tomáš Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. In *SLT*, 2012.
- [37] Tomáš Mikolov, Martin Karafit, Lukas Burget, Jan Cernock, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, 2010.
- [38] Tomáš Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernock, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *ICASSP*, 2011.
- [39] Andriy Mnih and Geoffrey Hinton. A scalable hierarchical distributed language model. In *NIPS*, 2009.
- [40] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *ICML*, 2012.
- [41] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *AISTATS*, 2005.
- [42] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

- [43] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *ICML*, 2013.
- [44] Adam Pauls and Dan Klein. Faster and smaller n-gram language models. In *ACL*, 2011.
- [45] Ronald Rosenfeld. Two decades of statistical language modeling: Where do we go from here? In *IEEE*, volume 88, pages 1270–1278, 2000.
- [46] David E. Rumelhart and James L. McClelland. On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, and PDP Research Group, editors, *Parallel Distributed Processing. Volume 2: Psychological and Biological Models*, pages 216–271. MIT Press, 1986.
- [47] Holger Schwenk. Continuous space language models. *Computer Speech and Languages*, 21(3):492–518, 2007.
- [48] Holger Schwenk. Continuous space translation models for phrase-based statistical machine translation. In *COLING*, 2012.
- [49] Abigail See, Minh-Thang Luong, and Christopher D. Manning. Compression of neural machine translation models via pruning. In *CoNLL*, 2016.
- [50] Le Hai Son, Alexandre Allauzen, and Francois Yvon. Continuous space translation models with neural networks. In *NAACL-HLT*, 2012.
- [51] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *ICSLP*, 2002.
- [52] Ilya Sutskever. *Training Recurrent Neural Networks*. PhD thesis, University of Toronto, 2012.
- [53] Ilya Sutskever, James Martens, and Geoffrey Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.
- [54] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

- [55] Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *ACL*, 2006.
- [56] Ashish Vaswani, Yingdong Zhao, Victoria Fossom, and David Chiang. Decoding with large-scale neural language models improves translation. In *EMNLP*, 2013.
- [57] 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.
- [58] Paul J. Werbos. Back propagation through time: What it does and how to do it. In *Proceedings of the IEEE*, volume 78, pages 1550–1560, 1990.
- [59] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.