MULTI-TASK SEQUENCE TO SEQUENCE LEARNING

Minh-Thang Luong[*], **Quoc V. Le, Ilya Sutskever, Oriol Vinyals, Lukasz Kaiser**
Google Brain
lmthang@stanford.edu, {qvl,ilyasu,vinyals,lukaszkaiser}@google.com

ABSTRACT

Sequence to sequence learning has recently emerged as a new paradigm in supervised learning. To date, most of its applications focused on only one task and not much work explored this framework for multiple tasks. This paper examines 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. Furthermore, we have established a new *state-of-the-art* result in constituent parsing with 93.0 F_1. 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.

1 INTRODUCTION

Multi-task learning (MTL) is an important machine learning paradigm that aims at improving the generalization performance of a task using other related tasks. Such framework has been widely studied by Thrun (1996); Caruana (1997); Evgeniou & Pontil (2004); Ando & Zhang (2005); Argyriou et al. (2007); Kumar & III (2012), among many others. In the context of deep neural networks, MTL has been applied successfully to various problems ranging from language (Liu et al., 2015), to vision (Donahue et al., 2014), and speech (Heigold et al., 2013; Huang et al., 2013).

Recently, sequence to sequence (*seq2seq*) learning, proposed by Kalchbrenner & Blunsom (2013), Sutskever et al. (2014), and Cho et al. (2014), emerges as an effective paradigm for dealing with variable-length inputs and outputs. *seq2seq* learning, at its core, uses recurrent neural networks to map variable-length input sequences to variable-length output sequences. While relatively new, the *seq2seq* approach has achieved state-of-the-art results in not only its original application – machine translation – (Luong et al., 2015b; Jean et al., 2015a; Luong et al., 2015a; Jean et al., 2015b; Luong & Manning, 2015), but also image caption generation (Vinyals et al., 2015b), and constituency parsing (Vinyals et al., 2015a).

Despite the popularity of multi-task learning and sequence to sequence learning, there has been little work in combining MTL with *seq2seq* learning. To the best of our knowledge, there is only one recent publication by Dong et al. (2015) which applies a *seq2seq* models for machine translation, where the goal is to translate from one language to multiple languages. In this work, we propose three MTL approaches that complement one another: (a) the *one-to-many* approach – for tasks that can have an encoder in common, such as translation and parsing; this applies to the multi-target translation setting in (Dong et al., 2015) as well, (b) the *many-to-one* approach – useful for multi-source translation or tasks in which only the decoder can be easily shared, such as translation and image captioning, and lastly, (c) the *many-to-many* approach – which share multiple encoders and decoders through which we study the effect of unsupervised learning in translation. We show that syntactic parsing and image caption generation improves the translation quality between English

[*]Minh-Thang Luong is also a student at Stanford University.

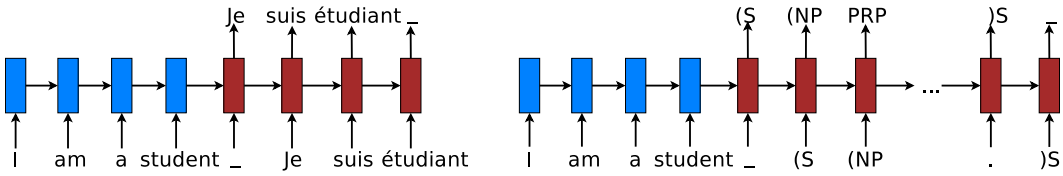

Figure 1: **Sequence to sequence learning examples** – (left) machine translation (Sutskever et al., 2014) and (*right*) constituent parsing (Vinyals et al., 2015a).

and German by up to +1.5 BLEU points over strong single-task baselines on the WMT benchmarks. Furthermore, we have established a new *state-of-the-art* result in constituent parsing with 93.0 F_1. We also explore two unsupervised learning objectives, sequence autoencoders (Dai & Le, 2015) and skip-thought vectors (Kiros et al., 2015), and reveal their interesting properties in the MTL setting: autoencoder helps less in terms of perplexities but more on BLEU scores compared to skip-thought.

2 Sequence to Sequence Learning

Sequence to sequence learning (*seq2seq*) aims to directly model the conditional probability $p(y|x)$ of mapping an input sequence, $x_1, \ldots, x_n$, into an output sequence, $y_1, \ldots, y_m$. It accomplishes such goal through the *encoder-decoder* framework proposed by Sutskever et al. (2014) and Cho et al. (2014). As illustrated in Figure 1, the *encoder* computes a representation $\boldsymbol{s}$ for each input sequence. Based on that input representation, the *decoder* generates an output sequence, one unit at a time, and hence, decomposes the conditional probability as:

$$\log p(y|x) = \sum_{j=1}^{m} \log p\left(y_j|y_{<j}, x, \boldsymbol{s}\right) \tag{1}$$

A natural model for sequential data is the recurrent neural network (RNN), which is used by most of the recent *seq2seq* work. These work, however, differ in terms of: (a) *architecture* – from unidirectional, to bidirectional, and deep multi-layer RNNs; and (b) *RNN type* – which are long-short term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and the gated recurrent unit (Cho et al., 2014).

Another important difference between *seq2seq* work lies in what constitutes the input representation $\boldsymbol{s}$. The early *seq2seq* work (Sutskever et al., 2014; Cho et al., 2014; Luong et al., 2015b; Vinyals et al., 2015b) uses only the last encoder state to initialize the decoder and sets $\boldsymbol{s} = [\,]$ in Eq. (1). Recently, Bahdanau et al. (2015) proposes an *attention mechanism*, a way to provide *seq2seq* models with a random access memory, to handle long input sequences. This is accomplished by setting $\boldsymbol{s}$ in Eq. (1) to be the set of encoder hidden states already computed. On the decoder side, at each time step, the attention mechanism will decide how much information to retrieve from that memory by learning where to focus, i.e., computing the alignment weights for all input positions. Recent work such as (Xu et al., 2015; Jean et al., 2015a; Luong et al., 2015a; Vinyals et al., 2015a) has found that it is crucial to empower *seq2seq* models with the attention mechanism.

3 Multi-task Sequence-to-Sequence Learning

We generalize the work of Dong et al. (2015) to the multi-task sequence-to-sequence learning setting that includes the tasks of machine translation (MT), constituency parsing, and image caption generation. Depending which tasks involved, we propose to categorize multi-task *seq2seq* learning into three general settings. In addition, we will discuss the unsupervised learning tasks considered as well as the learning process.

3.1 One-to-Many Setting

This scheme involves *one encoder* and *multiple decoders* for tasks in which the encoder can be shared, as illustrated in Figure 2. The input to each task is a sequence of English words. A separate decoder is used to generate each sequence of output units which can be either (a) a sequence of tags

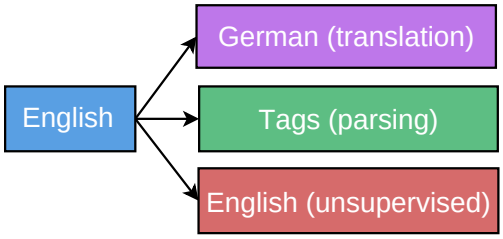

Figure 2: **One-to-many Setting** – one encoder, multiple decoders. This scheme is useful for either multi-target translation as in Dong et al. (2015) or between different tasks. Here, English and German imply sequences of words in the respective languages. The α values give the proportions of parameter updates that are allocated for the different tasks.

for constituency parsing as used in (Vinyals et al., 2015a), (b) a sequence of German words for machine translation (Luong et al., 2015a), and (c) the same sequence of English words for autoencoders or a related sequence of English words for the skip-thought objective (Kiros et al., 2015).

3.2 Many-to-One Setting

This scheme is the opposite of the *one-to-many* setting. As illustrated in Figure 3, it consists of *multiple encoders* and *one decoder*. This is useful for tasks in which only the decoder can be shared, for example, when our tasks include machine translation and image caption generation (Vinyals et al., 2015b). In addition, from a machine translation perspective, this setting can benefit from a large amount of monolingual data on the target side, which is a standard practice in machine translation system and has also been explored for neural MT by Gulcehre et al. (2015).

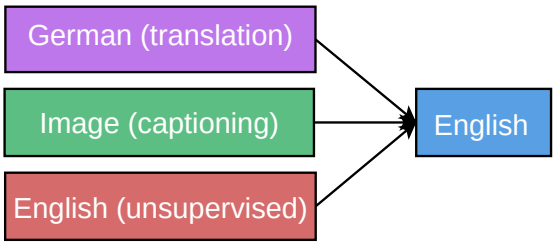

Figure 3: **Many-to-one setting** – multiple encoders, one decoder. This scheme is handy for tasks in which only the decoders can be shared.

3.3 Many-to-Many Setting

Lastly, as the name describes, this category is the most general one, consisting of multiple encoders and multiple decoders. We will explore this scheme in a translation setting that involves sharing multiple encoders and multiple decoders. In addition to the machine translation task, we will include two unsupervised objectives over the source and target languages as illustrated in Figure 4.

3.4 Unsupervised Learning Tasks

Our very first unsupervised learning task involves learning *autoencoders* from monolingual corpora, which has recently been applied to sequence to sequence learning (Dai & Le, 2015). However, in Dai & Le (2015)'s work, the authors only experiment with pretraining and then finetuning, but not joint training which can be viewed as a form of multi-task learning (MTL). As such, we are very interested in knowing whether the same trend extends to our MTL settings.

Additionally, we investigate the use of the *skip-thought* vectors (Kiros et al., 2015) in the context of our MTL framework. Skip-thought vectors are trained by training sequence to sequence models on pairs of consecutive sentences, which makes the skip-thought objective a natural *seq2seq* learning candidate. A minor technical difficulty with skip-thought objective is that the training data must

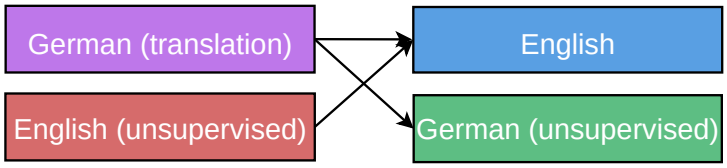

Figure 4: **Many-to-many setting** – multiple encoders, multiple decoders. We consider this scheme in a limited context of machine translation to utilize the large monolingual corpora in both the source and the target languages. Here, we consider a single translation task and two unsupervised autoencoder tasks.

consist of ordered sentences, e.g., paragraphs. Unfortunately, in many applications that include machine translation, we only have sentence-level data where the sentences are unordered. To address that, we split each sentence into two halves; we then use one half to predict the other half.

3.5 Learning

Dong et al. (2015) adopted an *alternating* training approach, where they optimize each task for a fixed number of parameter updates (or mini-batches) before switching to the next task (which is a different language pair). In our setting, our tasks are more diverse and contain different amounts of training data. As a result, we allocate different numbers of parameter updates for each task, which are expressed with the *mixing* ratio values α_i (for each task i). Each parameter update consists of training data from one task only. When switching between tasks, we select randomly a new task i with probability $\frac{\alpha_i}{\sum_j \alpha_j}$.

Our convention is that the first task is the *reference* task with $\alpha_1 = 1.0$ and the number of training parameter updates for that task is prespecified to be N. A typical task i will then be trained for $\frac{\alpha_i}{\alpha_1} \cdot N$ parameter updates. Such convention makes it easier for us to fairly compare the same reference task in a single-task setting which has also been trained for exactly N parameter updates.

When sharing an encoder or a decoder, we share both the recurrent connections and the corresponding embeddings.

4 Experiments

We evaluate the multi-task learning setup on a wide variety of sequence-to-sequence tasks: constituency parsing, image caption generation, machine translation, and a number of unsupervised learning as summarized in Table 1.

4.1 Data

Our experiments are centered around the *translation* task, where we aim to determine whether other tasks can improve translation and vice versa. We use the WMT'15 data (Bojar et al., 2015) for the English⇆German translation problem. Following Luong et al. (2015a), we use the 50K most frequent words for each language from the training corpus.[1] These vocabularies are then shared with other tasks, except for parsing in which the target "language" has a vocabulary of 104 tags. We use newstest2013 (3000 sentences) as a validation set to select our hyperparameters, e.g., mixing coefficients. For testing, to be comparable with existing results in (Luong et al., 2015a), we use the filtered newstest2014 (2737 sentences)[2] for the English→German translation task and newstest2015 (2169 sentences)[3] for the German→English task. See the summary in Table 1.

[1]The corpus has already been tokenized using the default tokenizer from Moses. Words not in these vocabularies are represented by the token <unk>.

[2]http://statmt.org/wmt14/test-filtered.tgz

[3]http://statmt.org/wmt15/test.tgz

Task	Train Size	Valid Size	Test Size	Vocab Size		Train Epoch	Finetune	
				Source	Target		Start	Cycle
English→German Translation	4.5M	3000	3003	50K	50K	12	8	1
German→English Translation	4.5M	3000	2169	50K	50K	12	8	1
English unsupervised	12.1M	Details in text		50K	50K	6	4	0.5
German unsupervised	13.8M			50K	50K	6	4	0.5
Penn Tree Bank Parsing	40K	1700	2416	50K	104	40	20	4
High-Confidence Corpus Parsing	11.0M	1700	2416	50K	104	6	4	0.5
Image Captioning	596K	4115	-	-	50K	10	5	1

Table 1: **Data & Training Details** – Information about the different datasets used in this work. For each task, we display the following statistics: (a) the number of training examples, (b) the sizes of the vocabulary, (c) the number of training epochs, and (d) details on when and how frequent we halve the learning rates (*finetuning*).

For the *unsupervised* tasks, we use the English and German monolingual corpora from WMT'15.[4] Since in our experiments, unsupervised tasks are always coupled with translation tasks, we use the same validation and test sets as the accompanied translation tasks.

For *constituency parsing*, we experiment with two types of corpora:

1. a small corpus – the widely used Penn Tree Bank (PTB) dataset (Marcus et al., 1993) and,
2. a large corpus – the high-confidence (HC) parse trees provided by Vinyals et al. (2015a).

The two parsing tasks, however, are evaluated on the same validation (section 22) and test (section 23) sets from the PTB data. Note also that the parse trees have been linearized following Vinyals et al. (2015a). Lastly, for *image caption generation*, we use a dataset of image and caption pairs provided by Vinyals et al. (2015b).

4.2 Training Details

In all experiments, following Sutskever et al. (2014) and Luong et al. (2015b), we train deep LSTM models as follows: (a) we use 4 LSTM layers each of which has 1000-dimensional cells and embeddings,[5] (b) parameters are uniformly initialized in [-0.06, 0.06], (c) we use a mini-batch size of 128, (d) dropout is applied with probability of 0.2 over vertical connections (Pham et al., 2014), (e) we use SGD with a fixed learning rate of 0.7, (f) input sequences are reversed, and lastly, (g) we use a simple finetuning schedule – after x epochs, we halve the learning rate every y epochs. The values x and y are referred as *finetune start* and *finetune cycle* in Table 1 together with the number of training epochs per task.

As described in Section 3, for each multi-task experiment, we need to choose one task to be the *reference task* (which corresponds to $\alpha_1 = 1$). The choice of the reference task helps specify the number of training epochs and the finetune start/cycle values which we also when training that reference task alone for fair comparison. To make sure our findings are reliable, we run each experimental configuration twice and report the average performance in the format *mean (stddev)*.

4.3 Results

We explore several multi-task learning scenarios by combining a *large* task (machine translation) with: (a) a *small* task – Penn Tree Bank (PTB) parsing, (b) a *medium-sized* task – image caption generation, (c) another *large* task – parsing on the high-confidence (HC) corpus, and (d) lastly, *unsupervised tasks*, such as autoencoders and skip-thought vectors. In terms of evaluation metrics, we report both validation and test perplexities for all tasks. Additionally, we also compute test BLEU scores (Papineni et al., 2002) for the translation task.

[4]The training sizes reported for the unsupervised tasks are only 10% of the original WMT'15 monolingual corpora which we randomly sample from. Such reduced sizes are for faster training time and already about three times larger than that of the parallel data. We consider using all the monolingual data in future work.

[5]For image caption generation, we use 1024 dimensions, which is also the size of the image embeddings.

4.3.1 LARGE TASKS WITH SMALL TASKS

In this setting, we want to understand if a small task such as *PTB parsing* can help improve the performance of a large task such as translation. Since the parsing task maps from a sequence of English words to a sequence of parsing tags (Vinyals et al., 2015a), only the encoder can be shared with an English→German translation task. As a result, this is a *one-to-many* MTL scenario (§3.1).

To our surprise, the results in Table 2 suggest that by adding a very small number of parsing mini-batches (with mixing ratio 0.01, i.e., one parsing mini-batch per 100 translation mini-batches), we can improve the translation quality substantially. More concretely, our best multi-task model yields a gain of +1.5 BLEU points over the single-task baseline. It is worth pointing out that as shown in Table 2, our single-task baseline is very strong, even better than the equivalent non-attention model reported in (Luong et al., 2015a). Larger mixing coefficients, however, overfit the small PTB corpus; hence, achieve smaller gains in translation quality.

For parsing, as Vinyals et al. (2015a) have shown that attention is crucial to achieve good parsing performance when training on the small PTB corpus, we do not set a high bar for our attention-free systems in this setup (better performances are reported in Section 4.3.3). Nevertheless, the parsing results in Table 2 indicate that MTL is also beneficial for parsing, yielding an improvement of up to $+8.9$ F_1 points over the baseline.[6] It would be interesting to study how MTL can be useful with the presence of the *attention* mechanism, which we leave for future work.

Task	**Translation**			**Parsing**
	Valid ppl	Test ppl	Test BLEU	Test F_1
(Luong et al., 2015a)	-	8.1	14.0	-
Our single-task systems				
Translation	8.8 (0.3)	8.3 (0.2)	14.3 (0.3)	-
PTB Parsing	-	-	-	43.3 (1.7)
Our multi-task systems				
Translation + PTB Parsing (1x)	8.5 (0.0)	8.2 (0.0)	14.7 (0.1)	54.5 (0.4)
Translation + PTB Parsing (0.1x)	8.3 (0.1)	7.9 (0.0)	15.1 (0.0)	**55.2 (0.0)**
Translation + PTB Parsing (0.01x)	**8.2** (0.2)	**7.7** (0.2)	**15.8** (0.4)	39.8 (2.7)

Table 2: **English→German WMT'14 translation & Penn Tree Bank parsing results** – shown are perplexities (ppl), BLEU scores, and parsing F_1 for various systems. For muli-task models, *reference* tasks are in italic with the mixing ratio in parentheses. Our results are averaged over two runs in the format *mean (stddev)*. Best results are highlighted in boldface.

4.3.2 LARGE TASKS WITH MEDIUM TASKS

We investigate whether the same pattern carries over to a medium task such as *image caption generation*. Since the image caption generation task maps images to a sequence of English words (Vinyals et al., 2015b; Xu et al., 2015), only the decoder can be shared with a German→English translation task. Hence, this setting falls under the *many-to-one* MTL setting (§3.2).

The results in Table 3 show the same trend we observed before, that is, by training on another task for a very small fraction of time, the model improves its performance on its main task. Specifically, with 5 parameter updates for image caption generation per 100 updates for translation (so the mixing ratio of 0.05), we obtain a gain of $+0.7$ BLEU scores over a strong single-task baseline. Our baseline is almost a BLEU point better than the equivalent non-attention model reported in Luong et al. (2015a).

4.3.3 LARGE TASKS WITH LARGE TASKS

Our first set of experiments is almost the same as the one-to-many setting in Section 4.3.1 which combines *translation*, as the reference task, with parsing. The only difference is in terms of parsing

[6] While perplexities correlate well with BLEU scores as shown in (Luong et al., 2015b), we observe empirically in Section 4.3.3 that parsing perplexities are only reliable if it is less than 1.3. Hence, we omit parsing perplexities in Table 2 for clarity. The parsing test perplexities (averaged over two runs) for the last four rows in Table 2 are 1.95, 3.05, 2.14, and 1.66. Valid perplexities are similar.

Task	Translation			Captioning
	Valid ppl	Test ppl	Test BLEU	Valid ppl
(Luong et al., 2015a)	-	14.3	16.9	-
Our single-task systems				
Translation	11.0 (0.0)	12.5 (0.2)	17.8 (0.1)	-
Captioning	-	-	-	30.8 (1.3)
Our multi-task systems				
Translation + Captioning (1x)	11.9	14.0	16.7	43.3
Translation + Captioning (0.1x)	10.5 (0.4)	12.1 (0.4)	18.0 (0.6)	**28.4** (0.3)
Translation + Captioning (0.05x)	**10.3** (0.1)	**11.8** (0.0)	**18.5** (0.0)	30.1 (0.3)
Translation + Captioning (0.01x)	10.6 (0.0)	12.3 (0.1)	18.1 (0.4)	35.2 (1.4)

Table 3: **German→English WMT'15 translation & captioning results** – shown are perplexities (ppl) and BLEU scores for various tasks with similar format as in Table 2. *Reference* tasks are in italic with mixing ratios in parentheses. The average results of 2 runs are in *mean (stddev)* format.

data. Instead of using the small Penn Tree Bank corpus, we consider a large parsing resource, the high-confidence (HC) corpus, which is provided by Vinyals et al. (2015a). As highlighted in Table 4, the trend is consistent; MTL helps boost translation quality by up to +0.9 BLEU points.

Task	Translation		
	Valid ppl	Test ppl	Test BLEU
(Luong et al., 2015a)	-	8.1	14.0
Our systems			
Translation	8.8 (0.3)	8.3 (0.2)	14.3 (0.3)
Translation + HC Parsing (1x)	8.5 (0.0)	8.1 (0.1)	15.0 (0.6)
Translation + HC Parsing (0.1x)	**8.2** (0.3)	**7.7** (0.2)	**15.2** (0.6)
Translation + HC Parsing (0.05x)	8.4 (0.0)	8.0 (0.1)	14.8 (0.2)

Table 4: **English→German WMT'14 translation** – shown are perplexities (ppl) and BLEU scores of various translation models. Our multi-task systems combine translation and parsing on the high-confidence corpus together. Mixing ratios are in parentheses and the average results over 2 runs are in *mean (stddev)* format. Best results are bolded.

The second set of experiments shifts the attention to *parsing* by having it as the reference task. We show in Table 5 results that combine parsing with either (a) the English autoencoder task or (b) the English→German translation task. Our models are compared against the best attention-based systems in (Vinyals et al., 2015a), including the state-of-the-art result of 92.8 F_1.

Before discussing the multi-task results, we note a few interesting observations. First, very small parsing perplexities, close to 1.1, can be achieved with large training data.[7] Second, our baseline system can obtain a very competitive F_1 score of 92.2, rivaling Vinyals et al. (2015a)'s systems. This is rather surprising since our models do not use any attention mechanism. A closer look into these models reveal that there seems to be an architectural difference: Vinyals et al. (2015a) use 3-layer LSTM with 256 cells and 512-dimensional embeddings; whereas our models use 4-layer LSTM with 1000 cells and 1000-dimensional embeddings. This further supports findings in (Jozefowicz et al., 2016) that larger networks matter for sequence models.

For the multi-task results, while autoencoder does not seem to help parsing, translation does. At the mixing ratio of 0.05, we obtain a non-negligible boost of 0.2 F_1 over the baseline and with 92.4 F_1, our multi-task system is on par with the best single system reported in (Vinyals et al., 2015a). Furthermore, by ensembling 6 different multi-task models (trained with the translation task at mixing ratios of 0.1, 0.05, and 0.01), we are able to establish a new *state-of-the-art* result in English constituent parsing with **93.0** F_1 score.

[7] Training solely on the small Penn Tree Bank corpus can only reduce the perplexity to at most 1.6, as evidenced by poor parsing results in Table 2. At the same time, these parsing perplexities are much smaller than what can be achieved by a translation task. This is because parsing only has 104 tags in the target vocabulary compared to 50K words in the translation case. Note that 1.0 is the theoretical lower bound.

Task	Parsing	
	Valid ppl	Test F_1
LSTM+A (Vinyals et al., 2015a)	-	92.5
LSTM+A+E (Vinyals et al., 2015a)	-	**92.8**
Our systems		
HC Parsing	1.12/1.12	92.2 (0.1)
HC Parsing + Autoencoder (1x)	1.12/1.12	92.1 (0.1)
HC Parsing + Autoencoder (0.1x)	1.12/1.12	92.1 (0.1)
HC Parsing + Autoencoder (0.01x)	1.12/1.13	92.0 (0.1)
HC Parsing + Translation (1x)	1.12/1.13	91.5 (0.2)
HC Parsing + Translation (0.1x)	1.13/1.13	92.0 (0.2)
HC Parsing + Translation (0.05x)	**1.11/1.12**	**92.4 (0.1)**
HC Parsing + Translation (0.01x)	1.12/1.12	92.2 (0.0)
Ensemble of 6 multi-task systems	-	**93.0**

Table 5: **Large-Corpus parsing results** – shown are perplexities (ppl) and F_1 scores for various parsing models. Mixing ratios are in parentheses and the average results over 2 runs are in *mean (stddev)* format. We show the individual perplexities for all runs due to small differences among them. For Vinyals et al. (2015a)'s parsing results, LSTM+A represents a single LSTM with attention, whereas LSTM+A+E indicates an ensemble of 5 systems. Important results are bolded.

4.3.4 MULTI-TASKS AND UNSUPERVISED LEARNING

Our main focus in this section is to determine whether unsupervised learning can help improve translation. Specifically, we follow the *many-to-many* approach described in Section 3.3 to couple the German→English translation task with two unsupervised learning tasks on monolingual corpora, one per language. The results in Tables 6 show a similar trend as before, a small amount of other tasks, in this case the *autoencoder* objective with mixing coefficient 0.05, improves the translation quality by $+0.5$ BLEU scores. However, as we train more on the autoencoder task, i.e. with larger mixing ratios, the translation performance gets worse.

Task	Translation			German	English
	Valid ppl	Test ppl	Test BLEU	Test ppl	Test ppl
(Luong et al., 2015a)	-	14.3	16.9	-	-
Our single-task systems					
Translation	11.0 (0.0)	12.5 (0.2)	17.8 (0.1)	-	-
Our multi-task systems with Autoencoders					
Translation + autoencoders (1.0x)	12.3	13.9	16.0	**1.01**	2.10
Translation + autoencoders (0.1x)	11.4	12.7	17.7	1.13	**1.44**
Translation + autoencoders (0.05x)	**10.9** (0.1)	**12.0** (0.0)	**18.3** (0.4)	1.40 (0.01)	2.38 (0.39)
Our multi-task systems with Skip-thought Vectors					
Translation + skip-thought (1x)	**10.4** (0.1)	**10.8** (0.1)	17.3 (0.2)	**36.9** (0.1)	**31.5** (0.4)
Translation + skip-thought (0.1x)	10.7 (0.0)	11.4 (0.2)	17.8 (0.4)	52.8 (0.3)	53.7 (0.4)
Translation + skip-thought (0.01x)	11.0 (0.1)	12.2 (0.0)	**17.8** (0.3)	76.3 (0.8)	142.4 (2.7)

Table 6: **German→English WMT'15 translation & unsupervised learning results** – shown are perplexities for translation and unsupervised learning tasks. We experiment with both *autoencoders* and *skip-thought vectors* for the unsupervised objectives. Numbers in *mean (stddev)* format are the average results of 2 runs; others are for 1 run only.

Skip-thought objectives, on the other hand, behave differently. If we merely look at the perplexity metric, the results are very encouraging: with more skip-thought data, we perform better consistently across both the translation and the unsupervised tasks. However, when computing the BLEU scores, the translation quality degrades as we increase the mixing coefficients. We anticipate that this is due to the fact that the skip-thought objective changes the nature of the translation task when using one half of a sentence to predict the other half. It is not a problem for the autoencoder objectives, however, since one can think of autoencoding a sentence as translating into the same language.

We believe these findings pose interesting challenges in the quest towards better unsupervised objectives, which should satisfy the following criteria: (a) a desirable objective should be compatible with the supervised task in focus, e.g., autoencoders can be viewed as a special case of translation, and (b) with more unsupervised data, both intrinsic and extrinsic metrics should be improved; skip-thought objectives satisfy this criterion in terms of the intrinsic metric but not the extrinsic one.

5 CONCLUSION

In this paper, we showed that multi-task learning (MTL) can improve the performance of the attention-free sequence to sequence model of (Sutskever et al., 2014). We found it surprising that training on syntactic parsing and image caption data improved our translation performance, given that these datasets are orders of magnitude smaller than typical translation datasets. Furthermore, we have established a new *state-of-the-art* result in constituent parsing with an ensemble of multi-task models. We also show that the two unsupervised learning objectives, autoencoder and skip-thought, behave differently in the MTL context involving translation. We hope that these interesting findings will motivate future work in utilizing unsupervised data for sequence to sequence learning. A criticism of our work is that our sequence to sequence models do not employ the attention mechanism (Bahdanau et al., 2015). We leave the exploration of MTL with attention for future work.

ACKNOWLEDGMENTS

We thank Chris Manning for helpful feedback on the paper and members of the Google Brain team for thoughtful discussions and insights.

REFERENCES

Ando, Rie Kubota and Zhang, Tong. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 6:1817–1853, 2005.

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Multi-task feature learning. In *NIPS*, 2007.

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

Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Haddow, Barry, Huck, Matthias, Hokamp, Chris, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Post, Matt, Scarton, Carolina, Specia, Lucia, and Turchi, Marco. Findings of the 2015 workshop on statistical machine translation. In *WMT*, 2015.

Caruana, Rich. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

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

Dai, Andrew M. and Le, Quoc V. Semi-supervised sequence learning. In *NIPS*, 2015.

Donahue, Jeff, Jia, Yangqing, Vinyals, Oriol, Hoffman, Judy, Zhang, Ning, Tzeng, Eric, and Darrell, Trevor. DeCAF: A deep convolutional activation feature for generic visual recognition, 2014.

Dong, Daxiang, Wu, Hua, He, Wei, Yu, Dianhai, and Wang, Haifeng. Multi-task learning for multiple language translation. In *ACL*, 2015.

Evgeniou, Theodoros and Pontil, Massimiliano. Regularized multi–task learning. In *SIGKDD*, 2004.

Gulcehre, Caglar, Firat, Orhan, Xu, Kelvin, Cho, Kyunghyun, Barrault, Loic, Lin, Huei-Chi, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*, 2015.

Heigold, Georg, Vanhoucke, Vincent, Senior, Alan, Nguyen, Patrick, Ranzato, Marc'Aurelio, Devin, Matthieu, and Dean, Jeffrey. Multilingual acoustic models using distributed deep neural networks. In *ICASSP*, 2013.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Huang, Jui-Ting, Li, Jinyu, Yu, Dong, Deng, Li, and Gong, Yifan. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *ICASSP*, 2013.

Jean, Sébastien, Cho, Kyunghyun, Memisevic, Roland, and Bengio, Yoshua. On using very large target vocabulary for neural machine translation. In *ACL*, 2015a.

Jean, Sébastien, Firat, Orhan, Cho, Kyunghyun, Memisevic, Roland, and Bengio, Yoshua. Montreal neural machine translation systems for WMT'15. In *WMT*, 2015b.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

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

Kiros, Ryan, Zhu, Yukun, Salakhutdinov, Ruslan, Zemel, Richard S., Torralba, Antonio, Urtasun, Raquel, and Fidler, Sanja. Skip-thought vectors. In *NIPS*, 2015.

Kumar, Abhishek and III, Hal Daumé. Learning task grouping and overlap in multi-task learning. In *ICML*, 2012.

Liu, Xiaodong, Gao, Jianfeng, He, Xiaodong, Deng, Li, Duh, Kevin, and Wang, Ye-Yi. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *NAACL*, 2015.

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

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

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

Marcus, Mitchell P., Marcinkiewicz, Mary Ann, and Santorini, Beatrice. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

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

Pham, Vu, Bluche, Théodore, Kermorvant, Christopher, and Louradour, Jérôme. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 285–290. IEEE, 2014.

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

Thrun, Sebastian. Is learning the n-th thing any easier than learning the first? In *NIPS*, 1996.

Vinyals, Oriol, Kaiser, Lukasz, Koo, Terry, Petrov, Slav, Sutskever, Ilya, and Hinton, Geoffrey. Grammar as a foreign language. In *NIPS*, 2015a.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. In *CVPR*, 2015b.

Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron C., Salakhutdinov, Ruslan, Zemel, Richard S., and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.