A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation

Raúl Vázquez
University of Helsinki
Department of Digital Humanities
raul.vazquez@helsinki.fi

Alessandro Raganato
University of Helsinki
Department of Digital Humanities
alessandro.raganato@helsinki.fi

Mathias Creutz
University of Helsinki
Department of Digital Humanities
mathias.creutz@helsinki.fi

Jörg Tiedemann
University of Helsinki
Department of Digital Humanities
jorg.tiedemann@helsinki.fi

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Neural machine translation has considerably improved the quality of automatic translations by learning good representations of input sentences. In this article, we explore a multilingual translation model capable of producing fixed-size sentence representations by incorporating an intermediate crosslingual shared layer, which we refer to as attention bridge. This layer exploits the semantics from each language and develops into a language-agnostic meaning representation that can be efficiently used for transfer learning. We systematically study the impact of the size of the attention bridge and the effect of including additional languages in the model. In contrast to related previous work, we demonstrate that there is no conflict between translation performance and the use of sentence representations in downstream tasks. In particular, we show that larger intermediate layers not only improve translation quality, especially for long sentences, but also push the accuracy of trainable classification tasks. Nevertheless, shorter representations lead to increased compression that is beneficial in non-trainable similarity tasks. Similarly, we show that trainable downstream tasks benefit from multilingual models, whereas additional language signals do not improve performance in non-trainable benchmarks. This is an important insight

Submission received: 21 March 2019; revised version received: 12 November 2019; accepted for publication: 29 January 2020.

https://doi.org/10.1162/COLI_a_00377

© 2020 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

that helps to properly design models for specific applications. Finally, we also include an in-depth analysis of the proposed attention bridge and its ability to encode linguistic properties. We carefully analyze the information that is captured by individual attention heads and identify interesting patterns that explain the performance of specific settings in linguistic probing tasks.

1. Introduction

Neural machine translation (NMT) has rapidly become the new machine translation (MT) standard, significantly improving over the traditional statistical machine translation model (Bojar et al. 2018). In only about four years, several architectures and approaches have been proposed, with increasing research efforts toward multilingual machine translation (Firat et al. 2016; Lakew, Cettolo, and Federico 2018; Wang et al. 2018). Inasmuch as MT is described as the task of translating a sentence from one language to another, at the recent conferences on MT (WMT18 and WMT19)[1] much interest was put on multilingualism, where a sub-track on multilingual systems was introduced with the aim of exploiting a third language to improve a bilingual model.

Multilingual neural machine translation comes in many flavors with different architectures and ways of sharing parameters (Luong et al. 2016; Zoph and Knight 2016; Lee, Cho, and Hofmann 2017; Dong et al. 2015; Firat, Cho, and Bengio 2016; Lu et al. 2018; Blackwood, Ballesteros, and Ward 2018). The main motivation of multilingual models is the effect of transfer learning that enables machine translation systems to benefit from relationships between languages and training signals that come from different data sets. Common techniques explore multisource encoders, multitarget decoders, or combinations of both. Multilingual models can push the translation performance of low-resource language pairs but also enable the translation between unseen language pairs, so-called zero-shot translation (Ha, Niehues, and Waibel 2016; Johnson et al. 2017; Gu et al. 2018a).

The effective computation of sentence representations using the translation task as an auxiliary semantic signal has also drawn interest to MT models (Hill, Cho, and Korhonen 2016; McCann et al. 2017; Schwenk and Douze 2017; Subramanian et al. 2018). Indeed, recent work makes use of machine translation models to capture syntactic and semantic properties of the input sentences, later to be used for learning general-purpose sentence representations (Shi, Padhi, and Knight 2016; Belinkov et al. 2017; Dalvi et al. 2017; Poliak et al. 2018; Bau et al. 2019). An important feature that enables an immediate use of the MT-based representations in other downstream tasks is the effective reduction to a fixed-sized vector; it enables functionality, at the expense of hampering the performance in the MT task (Britz, Guan, and Luong 2017; Cífka and Bojar 2018). However, it is not fully clear how the properties of the fixed-sized vector influence the tradeoff between the performance of the model in MT and the information it encodes as a meaning representation vector. Recent studies either focus on the usage of such MT-based vector representations in other tasks (Schwenk 2018), on translation quality (Lu et al. 2018), on speed comparison (Britz, Guan, and Luong 2017), or only explore a bilingual scenario (Cífka and Bojar 2018).

For this study, we focus on exploring a crosslingual intermediate shared layer in an MT model. We apply an architecture based on shared inner-attention with

1 http://www.statmt.org/wmt18/translation-task.html.
http://www.statmt.org/wmt19/translation-task.html.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

language-specific encoders and decoders that can easily scale to a large number of languages (more details about the architecture in Section 2). Simultaneously, it addresses the task of obtaining language-agnostic sentence embeddings (Lin et al. 2017; Cífka and Bojar 2018; Lu et al. 2018) that can be straightforwardly applied to downstream tasks. In Sections 4 and 5, we examine this model with a systematic evaluation on different sizes of the shared layer and extensive experiments to study the abstractions it learns from multiple translation tasks.

In contrast to previous work (Cífka and Bojar 2018), we demonstrate that there is a direct relation between the translation performance and the scores attained on trainable downstream tasks when adjusting the size of the intermediate layer. The trend is different for non-trainable tasks that benefit from the increased compression that denser representations achieve, which typically hurts the translation performance because of the decreased capacity of the model. We also show that multilingual models improve trainable downstream tasks, even further demonstrating the additional abstraction that is pushed into the representations through additional translation tasks involved in training. This even holds in low-resource scenarios as we show empirically in Section 4.4. Moreover, we find that multilingual training leads to a better encoding of linguistic properties of the sentence, and that a larger size of the shared inner-attention layer leads to a better syntactic understanding of the sentence rather than semantic (see Section 5). Furthermore, we include an in-depth analysis of the attention bridge on the ability of encoding linguistic properties, investigating systematically each component of the shared inner-attention layer.

In the following, we will first introduce the architecture that we apply in our experiments. Thereafter, we will discuss translation quality before diving into the detailed analyses of sentence representations and their applications, which will be the main focus of this article.

2. Model Architecture

The model we use follows the standard set-up of an encoder–decoder model of machine translation with a traditional attention mechanism (Bahdanau, Cho, and Bengio 2015; Luong et al. 2016). However, to enable multilingual training we augment the network with language-specific encoders and decoders trainable with a language-rotating scheduler (Dong et al. 2015; Schwenk and Douze 2017). We also incorporate an intermediate inner-attention layer, which summarizes the encoder information in a fixed-size vector representation, to serve as a language-agnostic layer (Cífka and Bojar 2018; Lu et al. 2018). Because of the attentive connection between encoders and decoders we call this layer **attention bridge**, and its architecture is a multilingual adaptation from the model proposed by Cífka and Bojar (2018). The overall architecture is illustrated in Figure 1.

2.1 Background: Attention Mechanism

Given an input $X = (x_1, \ldots, x_n)$, a sequence of embedded tokens into the vector space $\mathbb{R}^{d_x}$, our goal is to generate a translation $Y = (y_1, \ldots, y_m)$. For the sake of clarity, we assume a recurrent **encoder** in the following even though the mechanism is not restricted to this particular type of encoder. A recurrent neural network (RNN)-based encoder reads each element in X to generate a context vector c. Generally, for each token the

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

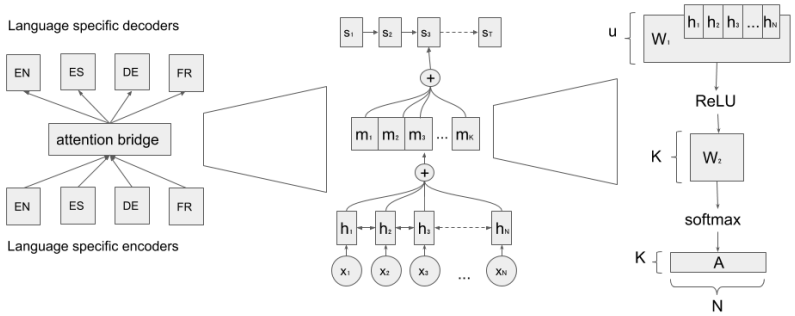

Figure 1
Architecture of the proposed multilingual NMT system. *Left:* The attention bridge connects the language-specific encoders and decoders. *Center:* Input $x_1 \dots x_n$ is translated into the decoder states $s_1 \dots s_t$ via the encoder states $h_1 \dots h_n$ and the attention bridge $M = m_1 \dots m_k$. *Right:* Computation of the fixed-size attentive matrix A.

RNN generates a hidden state $h_t \in \mathbb{R}^{d_h}$ where the last hidden state of the RNN often defines c:

$$h_t = f(x_t, h_{t-1}) \tag{1}$$

$$c = h_n \tag{2}$$

and $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_h} \longrightarrow \mathbb{R}^{d_h}$ is a non-linear activation function. We use bidirectional long short-term memory (LSTM) units (Graves and Schmidhuber 2005) as f in this article.

Then, the **decoder** network sequentially computes $(y_1, \dots, y_m)$ by optimizing

$$p(Y|X) = \prod_{t=1}^{m} p(y_t|c, Y_{t-1}) \tag{3}$$

where $Y_{t-1} = (y_1, \dots, y_{t-1})$. Each distribution $p_t = p(y_t|c, Y_{t-1}) \in \mathbb{R}^{d_v}$ is usually computed with a softmax function over all the words in the vocabulary, taking into account the current hidden state of the decoder s_t:

$$p_t = softmax(y_{t-1}, s_t) \tag{4}$$

$$s_t = \varphi(c, y_{t-1}, s_{t-1}) \tag{5}$$

where φ is another non-linear activation function and d_v is the size of the vocabulary.

Including an **attention mechanism** in the decoder implies that a different context vector c_t will be computed at each step t, instead of fixing c as in Equation (2) for generating all output words. This alignment method allows the decoder to assign different weights to each part of the input at every decoding step by defining c_t as the weighted

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

sum of hidden states of the encoder $c_t = \sum_{i=1}^{n} \alpha_{t,i} h_t$, where $\alpha_{t,i}$ indicates how much the i-th input word contributes to generating the t-th output word, and is usually defined as

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{k=1}^{n} exp(e_{t,k})} \tag{6}$$

$$e_{t,i} = g(s_t, h_i) \tag{7}$$

where g is a feedforward neural network.

2.2 Inner-Attention as Semantic Bridge

To enable multilingual training and the possibility to obtain a fixed-size sentence representation from the model, we propose to extend the attention-based network (Section 2.1) with the following modifications:

1. the incorporation of the *attention bridge*: an inner-attention layer shared among all language pairs, that serves as a neural "interlingua";
2. the use of language-specific encoders and decoders for each language pair, trainable with a language-rotating scheduler; and
3. the introduction of a penalty term in the loss function to avoid redundancy in the shared inner-attention.

(1) Attention bridge: Each encoder takes as input a sequence of tokens $(x_1, \ldots, x_n)$ and produces n hidden states $H = (h_1, \ldots, h_n)$ with $h_i \in \mathbb{R}^{d_h}$, in our case, using a bidirectional LSTM (Graves and Schmidhuber 2005).[2] Next, we encode this variable length sentence-embedding matrix H into a fixed size $M \in \mathbb{R}^{d_h \times k}$ capable of focusing on k different components of the sentence (Lin et al. 2017; Chen, Ling, and Zhu 2018; Cífka and Bojar 2018), using self-attention as follows:

$$A = softmax\left(W_2 \mathrm{ReLU}(W_1 H^T)\right) \tag{8}$$

$$M = AH \tag{9}$$

where $W_1 \in \mathbb{R}^{d_w \times d_h}$ and $W_2 \in \mathbb{R}^{k \times d_w}$ are weight matrices, with d_w a hyperparameter set arbitrarily, and k the number of *attention heads* in the attention bridge. Note that each column of M, m_i, is a component focusing on a portion of the sentence, so all of them together should reflect the overall semantics of the sentence.

Each decoder follows a common attention mechanism in NMT (Luong, Pham, and Manning 2015), with an initial state computed by mean pooling over M, and using M instead of the hidden states of the encoder for computing the context vector. Formally, we only need to compute Equations (6) and (7) using the columns of M instead of the encoder states h_i.

(2) Language-specific encoders and decoders: To deal with additional language pairs, we incorporate an encoder for each input language and an attentive decoder for each

2 Note that the attention bridge is independent of the underlying encoder and decoder (Lu et al. 2018). Although we use a BiLSTM, it could be replaced with a GRU (Cho et al. 2014), a transformer type network (Vaswani et al. 2017), or with a CNN (Gehring et al. 2017)

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

output language to be connected via the attention bridge. This adjusts the parameters of the bridge layer with multilingual information.

Figure 1 shows a basic diagram on the left-hand side to illustrate the use of several encoders and decoders that are plugged in and out at every change of batch. To avoid over-fitting the attention bridge layer toward one specific language-pair, we cycle through the available target and source languages at each batch uniformly as in Lu et al. (2018).

(3) Penalty term: The attention bridge matrix M from Equation (9) could potentially suffer from redundancy problems by learning repetitive information for different attention heads. To address this issue, we add a penalty term to the loss function, proven effective in related work (Lin et al. 2017; Chen, Ling, and Zhu 2018; Tao et al. 2018):

$$\mathcal{L} = -\log\left(p\left(Y|X\right)\right) + \left\|AA^T - I\right\|_F^2 \tag{10}$$

where A is as in Equation (8) and I is the identity matrix. Note that this term forces each vector to focus on different aspects of the sentence by making the columns of A to be approximately orthogonal in the Frobenius norm.

The advantage of the fixed-size representation is the straightforward application in downstream tasks. However, selecting a reasonable size of the attention bridge in terms of attention heads is crucial for the performance both in a bilingual and multilingual scenario as we will see in our experiments in Sections 3.2 and 4.

3. Translation Quality

Before applying and analyzing sentence representations that can be learned with the proposed architecture from the previous section, we ought to verify that the model is indeed capable of learning multilingual translation—the original training objective. For this, we apply the model in two scenarios: a low-resource scenario with a multilingual image caption translation task (Elliott et al. 2016) and the application to considerably larger data sets based on experiments with Europarl (Koehn 2005) and news translation tasks (Callison-Burch et al. 2007). In the following we will first discuss multilingual transfer learning in the low-resource scenario before we analyze the effect of the attention bridge size on translation quality in the large-data setting.

3.1 Multilingual Translation of Image Captions

Multi30K (Elliott et al. 2016) is a parallel data set containing 29k image captions for training and 1k sentences for validation in four European languages; Czech (cs), German (de), French (fr), and English (en). We test the trained model with the flickr 2016 test data of the same data set and obtain BLEU scores using the sacreBLEU script[3] (Post 2018). The preprocessing pipeline consists of lowercasing, normalizing, and tokenizing using the scripts provided in the Moses decoder (Koehn et al. 2007), together with learning and applying a 10k operations byte-pair-encoding (BPE) model per language (Sennrich, Haddow, and Birch 2016). Each encoder consists of two stacked BiLSTMs of size $d_h = 512$ (i.e., the hidden states per direction are of size 256). Each decoder is composed of two stacked unidirectional LSTMs with hidden states of size 512. For the

3 With signature BLEU+case.lc+numrefs.1+smooth.exp+tok.13a+version.1.2.11.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 1
BLEU scores obtained in experiments on the Multi30k data set. *Left:* Bilingual models, our baselines. *Center:* Models trained on {De,Fr,Cs}↔En, with zero-shot translations in italics. *Right:* Many-to-many model. Both zero-shot and M ↔ M translations improve significantly when including monolingual data. (Best results shown in bold font).

	BILINGUAL				{DE,FR,CS} ↔ EN				M ↔ M			
src/tgt	EN	DE	CS	FR	EN	DE	CS	FR	EN	DE	CS	FR
EN	–	36.78	28.00	55.96	–	37.85	29.51	57.87	–	37.70	29.67	55.78
DE	39.00	–	23.44	38.22	39.39	–	*0.35*	*0.83*	40.68	–	26.78	41.07
CS	35.89	28.98	–	36.44	37.20	*0.65*	–	*1.02*	38.42	31.07	–	40.27
FR	49.54	32.92	25.98	–	48.49	*0.60*	*0.30*	–	49.92	34.63	26.92	–
	BILINGUAL + ATT BRIDGE				{DE,FR,CS} ↔ EN + MONOLING				M ↔ M + MONOLINGUAL			
	EN	DE	CS	FR	EN	DE	CS	FR	EN	DE	CS	FR
EN	–	35.85	27.10	53.03	–	**38.92**	30.27	**57.87**	–	38.48	**30.47**	57.35
DE	38.19	–	23.97	37.40	40.17	–	*19.50*	*26.46*	**41.82**	–	**26.90**	**41.49**
CS	36.41	27.28	–	36.41	37.30	*22.13*	–	*22.80*	**39.58**	**31.51**	–	**40.87**
FR	48.93	31.70	25.96	–	50.41	25.96	20.09	–	**50.94**	**35.25**	**28.80**	–

model input and output, the word embeddings have dimension $d_x = d_y = 512$. We use an attention bridge layer with $k = 10$ attention heads with $d_w = 1,024$, dimensions of W_1 and W_2 from Equation (8).

We use a stochastic gradient descent optimizer with a learning rate of 1.0 and batch size 64, and for each experiment, we select the best model on the development set. We implement our model on top of an OpenNMT-py (Klein et al. 2017) fork, which we make available for reproducibility purposes.[4]

3.1.1 Baselines. The first experiment we conduct is to corroborate that the proposed architecture works correctly, and we assess performance in a bilingual setting. We expect that the models slightly drop in performance when the fixed-size attention bridge is introduced, because there are no direct crosslingual attention links between the source and target languages. However, we want to see whether the architecture is robust enough to carry over the essential information needed for translation with the inclusion of the additional intermediate abstraction layer.

In Table 1 we present a comparison of our architecture in contrast with a strong bilingual baseline consisting of an architecture with the same specifications, without the components of our model. The table presents the scores obtained for each of the 12 bilingual models trained on each language pair. In this case, we note that the basic bilingual models without any attention bridge have a slightly better performance in most cases. The most significant drop occurs when translating English to French, with a difference of over 2 BLEU points, but this case is exceptional. Typically the BLEU score decreases by less than 1 point.

This behavior is expected because the information from the encoder has to be summarized in the 10 heads of the inner-attention layer without (multilingual) information from other encoders to boost the states of this bridge. Nevertheless, these tests justify the validity of the architecture; namely, that the attention bridge does not cause

4 https://github.com/Helsinki-NLP/OpenNMT-py/tree/neural-interlingua.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

a significant problem for the translation model in the bilingual case. We will use the results of bilingual models both with and without attention bridge as our baselines for the comparison to the multilingual models that we describe subsequently.

3.1.2 Many-To-One and One-To-Many Models. The expected power of the attention bridge comes from its ability to share information across various language pairs. We now look at the effect of including additional languages during training on the translation performance of individual language pairs. We start by training models that include many-to-one and one-to-many settings with English as target and source, respectively. This set-up makes it possible to study the ability of zero-shot translation, that is, the translation between languages that have not been seen together in the training data. By performing zero-shot translation, we can test the abstraction potential of the attention bridge and its effectiveness in encoding multilingual information.

For the first experiment, we use the many-to-one and one-to-many strategy to train a {De,Fr,Cs}↔En model. As depicted in Table 1, this attempt already results in substantial improvements for the language pairs seen during training.

The model exceeds both bilingual baselines from the previous section. However, this model is entirely incapable of performing zero-shot translations. We believe that this inability of the model to generalize to unseen language pairs arises from the fact that every non-English encoder (or decoder) only learns to process information that is to be decoded into English (or encoded from English input). This finding is consistent with Lu et al. (2018); so, to address this problem, we incorporate monolingual data during training, that is, for each available language A, we train $A \rightarrow A$ with identical copies of the input sentence as the target. Hence, we do not include any additional data, but we reincorporate examples from the same parallel training corpus used in all other experiments. As a consequence, we see a remarkable increase in the BLEU scores, including a substantial boost for the language pairs not seen during training. In short, the monolingual data informs the model that other languages can be produced besides English, and that English is not the unique source language.

Additionally, there is a positive effect on the seen language pairs, the cause of which is not immediately evident. One possibility may be that the shared layer acquires additional information that can be included in the abstraction process yet not available to the other models.

3.1.3 Many-to-Many Models. To further examine the capabilities of the proposed architecture we conduct two experiments under a many-to-many scenario.

First, we test the architecture in a many-to-many setting with all language pairs included. Table 1 summarizes the results of our experiments. As in the previous case, we compare settings that include monolingual data with their counterparts that do not include it.

On a first note, the inclusion of language pairs results in an improved performance when compared to the bilingual baselines, as well as the many-to-one and one-to-many cases. The only exception is the En→Fr task. Moreover, the addition of monolingual data during training leads to even higher scores, producing the overall best model. The improvements in BLEU range from 1.40 to 4.43 compared to the standard bilingual model.

Next, we perform a systematic evaluation on zero-shot translation. For this, we train six different models where we include all but one of the available language pairs (e.g., En↔De). Then, we test our models while also performing bidirectional zero-shot translations for the unseen language pairs. Figure 2 summarizes the results.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

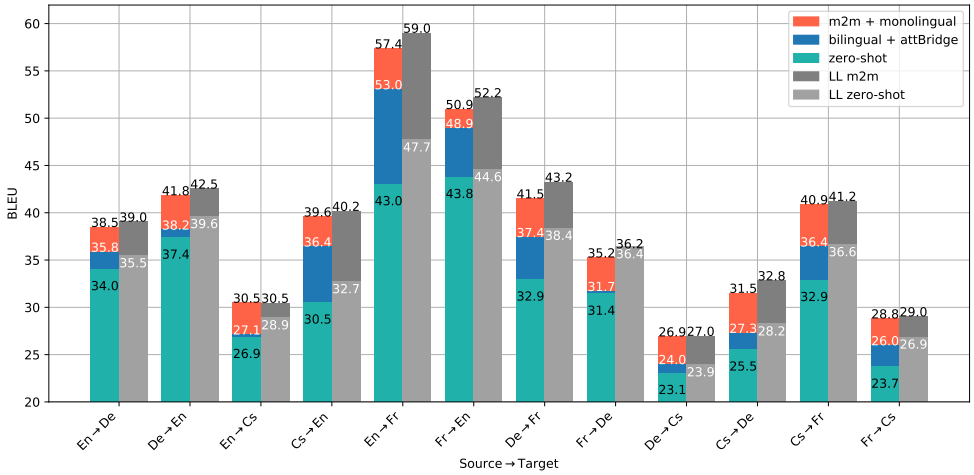

Figure 2
For every language pair, we compare the BLEU scores between models trained and tested on the Multi30k data set: (1) our best model (M ↔ M plus monolingual data), (2) the bilingual model of that language pair, (3) the zero-shot translation of the many-to-many model trained without that specific language pair, (4) the Johnson et al.(2017) model using language labels (LL) trained in a many-to-many scenario, and (5) the zero-shot of of LL without that specific language pair.

We observe that these zero-shot translation scores are generally better than the ones from the previous {De,Fr,Cs}↔En model with monolingual data (Table 1). We also note that the zero-shot models perform relatively well in comparison with the MANY-TO-MANY model. Furthermore, these zero-shot models almost reach the scores of the bilingual models trained only on the zero-shot language pairs.

As a point of comparison, we also implemented the approach of Johnson et al. (2017), using a language label at the beginning of the input sentence to specify the required target language and a single shared model with joint vocabulary. We will refer to this model as the *LL* approach herein. We used a combined 40k BPE operations model trained on the combined corpora and the same architecture specifications from Section 3.1, without the components of the attention bridge model. The results are shown in Figure 2 in the gray bars next to our attention bridge scores. We can see that the many-to-many LL models perform slightly better than our attention bridge model. This is not very surprising as they are based on a model architecture that also performs better in the bilingual case as we have seen in the comparison between bilingual models with and without attention bridge in Table 1. Section 3.2.2 will also show that this is basically caused by long sentences that are not as well covered by the attention-bridge model.

A similar effect is visible in the zero-shot results that we obtain in the same way as with our attention bridge model (i.e., leaving one language pair out of the training data). The differences to our model are sometimes larger than in the supervised set-up. This can be explained by the positive effect of sharing all encoder and decoder parameters in the case of related languages. Having a small data set to start with, the additional data from the other language pairs seems to be very beneficial and in some cases the zero-shot performance comes very close to the supervised model with all data included. In future work, we would like to investigate the effect on more distant languages and increasing numbers of languages involved in our comparison.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Also note that the language labeling technique does not produce crosslingual sentence representations, the main advantage of our approach, which we will test in multilingual downstream tasks (see Section 4.2). The language label makes the encoder effectively depending on the target language, which makes it difficult to apply the representations produced by that system to unrelated downstream tasks. These drawbacks and the fact that we produce competitive results with our architecture while producing directly applicable crosslingual sentence representations motivate the use of our architecture in multilingual set-ups. Furthermore, we can also show that the drop in performance mainly comes from long sentences that are not covered as well as shorter ones. More details on this effect can be found in Section 3.2.2.

3.1.4 Effect of the Penalty Term. In order to study the effect of the penalty term, we have trained additional bilingual and multilingual models, where the penalty term Equation (10) was excluded from the loss function. We re-ran all the 36 tests in the lower row of sub-tables in Table 1. We then compared the BLEU scores between corresponding set-ups where the penalty term was present and absent. We discovered that in 21 out of the 36 tests (58 %) the presence of the penalty term was beneficial. On average, the penalty term improves the BLEU scores by 0.11 points, across all tested types of models and language pairs.

As discussed in Lin et al. (2017), the quantitative effect of the penalty term might not be significant for some tasks, yet still it maintains the positive effect of encouraging the attentive matrix to be focused on different aspects of the sentence rather than picking up redundant information. Indeed, as we will see in Section 5.3, adding the penalty term effectively helps the model to spread the attention of the individual attention heads once the sentence is covered with token-specific attention. This leads us to keeping it in the remaining experiments.

3.2 The Effect of the Attention Bridge Size on MT Quality

The study on the Multi30K data set demonstrates the general ability of the attention bridge model to learn multilingual translation models capable of sharing knowledge between the various language pairs also enabling zero-shot translation similar to other multilingual NMT architectures. In the following, we investigate the impact of the size of the attention bridge on translation performance. For this study, we choose a data set of a realistic size and a more challenging benchmark with a larger vocabulary and a greater variety of sentence lengths as one of the most crucial properties influencing the quality of machine translation. In particular, we apply the Europarl Corpus v7 (Koehn 2005) with a selection of four languages and news test sets from the the ACL-WMT07 shared task (Callison-Burch et al. 2007), using dev2006 as validation data and devtest2006 plus test2006 as blind test data, ending up with 2K and 4K sentences, respectively. We focus on six language pair directions: English–French (EN–FR), English–German (EN–DE), and English–Spanish (EN–ES), with training data of approximately 2M sentences each.[5] The data are pre-processed following the standard MT pipeline, including tokenization and truecasing. Sentences are then encoded using BPE (Sennrich, Haddow, and Birch 2016), with 32,000 merge operations for each language. BLEU scores are computed case-insensitively using SACREBLEU as before.

5 As before, we trained all models including monolingual data, and because of the small size of the Czech Europarl data, we include Spanish instead.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 2
BLEU scores for bilingual Europarl models.

		BILINGUAL MODELS				
		baseline	k=1	k=10	k=25	k=50
EN	DE	22.72	15.04	20.25	21.26	21.87
	ES	30.28	22.8	27.3	28.52	29.15
	FR	25.88	18.97	23.49	24.42	25.07
DE	EN	24.28	17.22	22.53	23.18	23.59
ES		28.16	19.33	25.2	26.49	28.16
FR		25.39	17.46	22.1	22.4	24.22

We will first look at the impact of attention bridge size on bilingual and multilingual models before we discuss the impact of sentence length on our model. In general, we expect that the positive effect of transfer learning in translation will fade out as the bilingual baseline models become stronger and outperform the attention bridge model with their additional bottleneck of a fixed size intermediate representation. This will mainly affect long sentences that are not properly summarized in the shared layer, causing a less effective access to encoder information through the crosslingual attention (more detailed analyses are presented in Section 3.2.2).

3.2.1 The Impact of Attention Bridge Size on Bilingual and Multilingual Models. For the following experiments, we apply the same architecture and hyperparameters as in Section 3.1. Regarding the attention bridge, we experiment with four different numbers of attention heads: $k = 1, 10, 25, 50$. In the training we use the Adam optimizer (Kingma and Ba 2015) with a learning rate of 0.0004 and batch size 256, for at most 100,000 steps per language pair. We select the best model according to the BLEU score on the validation set. For multilingual systems, we select only one model with the best overall BLEU score across the validation set of all the language pairs involved.

We adopt different training strategies: First we train bilingual models for the language pairs of interest; then we train a {DE,ES,FR}↔EN model using the many-to-one and one-to-many strategy; lastly we train a many-to-many model involving all translation directions between the three languages (i.e., we also include DE–ES, DE–FR, and ES–FR).[6]

Table 2 shows the BLEU scores of our models with varying k, the number of the attention heads in the attention bridge, compared with a baseline, a traditional encoder-decoder model with attention mechanism (Luong, Pham, and Manning 2015). Among the attention bridge models, we can see that the performance consistently increases when k grows. The model with 50 heads achieves the best results among our models. It obtains scores that range in the same ballpark as the baseline, only in a few cases there is a degradation of around 1 BLEU point. Furthermore, the performance of this model compared with the one with one attention head is substantial: more than 6 BLEU points on average, corroborating previous findings (Britz, Guan, and Luong 2017; Cífka and Bojar 2018). Intuitively, a large number of attention heads manage to encode richer

6 Data coming from the same Europarl source: `http://opus.nlpl.eu/Europarl.php`.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 3
BLEU scores for multilingual Europarl models with various sizes k of the attention bridge. For comparison, the table includes results of a multi-way multilingual NMT model (Firat, Cho, and Bengio 2016) and a completely shared architecture with language labels: LL (Johnson et al. 2017) .

		M ↔ EN						M ↔ M		
		k=1	k=10	k=25	k=50	Firat	LL	k=1	k=50	LL
EN	DE	14.66	19.87	20.61	20.83	18.49	21.63	14.89	20.47	21.7
	ES	21.82	27.55	28.41	28.13	27.73	29.48	21.4	27.6	29.53
	FR	17.8	23.35	24.36	23.79	23.22	25.56	17.62	24.15	25.51
DE		16.97	21.39	23.42	24	24.8	25.96	17.38	24.4	25.84
ES	EN	18.38	25.39	27.01	27.12	25.7	28.41	19.43	26.98	28.67
FR		17.52	21.93	24.4	23.9	24.52	26.93	17.47	24.47	25.47

information about the source sentence improving the performance of the model for MT. Those results verify that BLEU and meaning representations do not have to be in opposition, as suggested by Cífka and Bojar (2018).

For the multilingual settings, we train a {DE,ES,FR}↔EN model using the many-to-one and one-to-many strategy, and a many-to-many model as discussed in Sections 3.1.2 and 3.1.3. Table 3 shows the comparison between the multilingual models. In general, we observe the same trend as in the bilingual evaluation concerning the size of the attention bridge. Namely, more attention heads lead to a higher BLEU score. Notably, we do not see any increase in translation quality from the {DE,ES,FR}↔EN model to the many-to-many model; the BLEU scores for all six translation directions are statistically equivalent. Besides, when we compare the bilingual and multilingual models for a given k, we do not note any apparent degradation or improvement regarding the BLEU score when incorporating multilingual data into the models.

For comparison, we again add results from the language-labeling approach by Johnson et al. (2017) and also from another popular approach that has been proposed by Firat, Cho, and Bengio (2016). The latter refers to a multi-way multilingual NMT system with a shared crosslingual attention mechanism, a model that is quite similar in spirit with our approach but without a fixed-size shared layer between encoder and decoder that bridges the crosslingual attention.

The multiway architecture produces lower scores for most language pairs. Note that we only show results for the {DE,ES,FR}↔EN set-up as the used implementation no longer holds for current standards,[7] training is very slow and would be prohibitively expensive in the many-to-many set-up. We expect that the trend will be the same and the scores are below our proposed architecture. The language-label approach by Johnson et al. (2017), on the other hand, is again very effective and produces the overall best results. Sharing all parameters is also beneficial in the Europarl experiments similar to what we have seen in the Multi30K results. Again, we have to note that the bilingual baseline will be higher and that we also focus on related languages again that benefit from a strong overlap in linguistic properties. However, once again, we can see that our model produces competitive results with the additional benefit of producing crosslingual fixed-size representations that are directly applicable in downstream tasks including crosslingual ones.

7 https://github.com/nyu-dl/dl4mt-multi.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

3.2.2 Length Analysis. In the previous section, we could see that there is a strong correlation between the size of the attention bridge and the quality of the translations produced. We could also see that the attention bridge model is capable of translating with a similar performance even though it creates an additional bottleneck of fixed-size representations. Nevertheless, the performance drops slightly and, in this section, we would like to investigate the reasons for that drop by looking at the effect on different subsets of the test data.

One of the main motivations for having more attention heads lies in the better support of longer sentences. To study the effect, following previous work (Bahdanau, Cho, and Bengio 2015; Tu et al. 2017; Dou et al. 2018), we group sentences of similar length and compute the BLEU score for each group. As we can see from Figure 3, a larger number of attention heads has, indeed, a positive impact when translating longer sentences. Long sentences do require a bigger attention bridge, and it affects both bilingual and multilingual models. Interestingly enough, on sentences with up to 45 words, there is no real gap between the results of the baseline model and our bridge models with a high number of attention heads. It looks like the performance drop of the attention bridge models is entirely due to sentences longer than 45 words. The same is true in comparison to the language-label approach. This also suggests that the increased performance of that model is due to the better coverage of long sentences.

Furthermore, we notice that multilingual models with 50 attention heads lose more on long sentences than bilingual ones. We hypothesize that this might be due to the increasing syntactic divergences between the languages that have to be encoded. The shared inner-attention layer needs to learn to focus on different parts of a sentence depending on the language it reads and, with increasing lengths of a sentence, this ability becomes harder and more difficult to pick up from the data alone.

3.3 Discussion

Our results demonstrate that the attention bridge model proposed in this paper implements an effective approach to multilingual machine translation. The shared layer successfully bridges language-dependent encoder and decoder networks enabling efficient transfer learning and improved sentence representation learning. Using the multi30k benchmark, the results of the multilingual models consistently outperform a strong

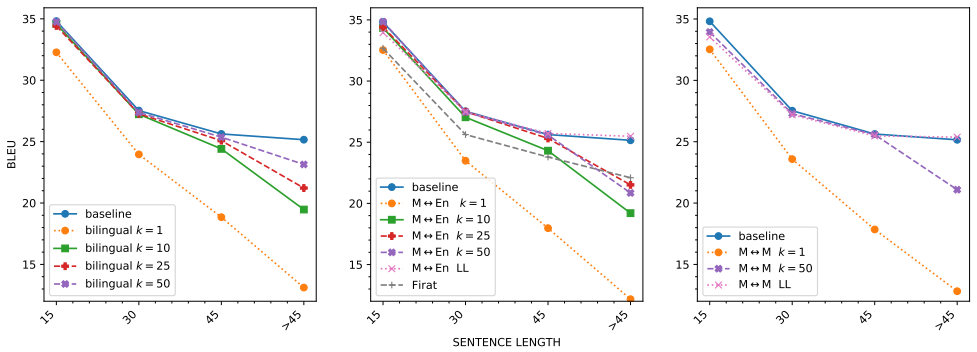

Figure 3
The BLEU scores obtained by the models with respect to different sentence length. The left figure shows the bilingual models, the middle one the many-to-one models, and the figure to the right illustrates the many-to-many models.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

bilingual model. This advantage, however, fades out with larger data sets. This is expected because of the limits of the fixed-size representations that bridge the gap between the various languages. But our analysis shows that this is mainly due to the problem with long sentences, an issue that needs to be addressed in future work. Our analysis also reveals that the size of the attention bridge plays a crucial role on translation quality and we will further discuss this below in the application of the sentence representations to unrelated downstream tasks. This brings us to the main point of this article, namely, the discussion of the quality of representations that can be learned from translations using the proposed multilingual architecture.

4. MT-Based Representations in Downstream Tasks

The main motivation for our study is to investigate the sentence representations that the MT model picks up during training. Therefore, the most important part is the assessment of these representations in unrelated downstream tasks and the analyses of the internal structure (which we will discuss in Section 5). In the following, we will first briefly introduce the tasks we consider before applying our models to each of them. Our MT models are trained on the Europarl data. However, in Section 4.4 we also include a study on downstream tasks with representations learned from limited resources, using the Multi30K data set, to further demonstrate that useful representations can be picked up even from tiny data sets. This is in contrast to related work where huge amounts of training data are typically applied to obtain reasonable performance.

Our assumption is that multilinguality contributes to a higher level of semantic abstraction that can be learned from the translation objective. To test this claim, we apply standard benchmarks collected in the SentEval toolkit (Conneau and Kiela 2018), the XNLI evaluation corpus (Conneau et al. 2018c), as well as the Yelp challenge data set.[8]

The SentEval toolkit contains three benchmark types: classification, similarity, and linguistic probing tasks. In the classification tasks, a classifier is trained on top of a sentence embedding involving various data sets: CR—product reviews (Hu and Liu 2004), MR—movie reviews (Pang and Lee 2005), MPQA—opinion polarity (Wiebe, Wilson, and Cardie 2005), SUBJ—subjectivity/objectivity status (Pang and Lee 2004), SST—binary and fine-grained sentiment analysis (Socher et al. 2013), TREC—question-type classification (Voorhees and Tice 2000), MRPC—paraphrase detection (Dolan, Quirk, and Brockett 2004), and SICK and SNLI—textual entailment and natural language inference (Marelli et al. 2014; Bowman et al. 2015).

In contrast to the classification tasks mentioned above, the similarity tasks do not involve any training and, instead, correlate the cosine distance between two sentence representations with a human labeled score using Pearson and Spearman coefficients. The data sets come from the SemEval Semantic Textual Similarity (STS) task series, from 2012 to 2016 (Agirre et al. 2012, 2013, 2014, 2015, 2016). The only exceptions are the SICK and STSB data set (Marelli et al. 2014; Cer et al. 2017), where training data are provided.

In addition, the SentEval toolkit contains probing tasks to study how linguistic features are encoded within a fixed-size vector (Conneau et al. 2018a).

All SentEval tasks are designed for English only. Therefore, we find it valuable to evaluate our sentence representations on multilingual classification tasks as well. For this purpose we make use of the XNLI evaluation corpus (Conneau et al. 2018c)

8 http://www.yelp.com/dataset.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

for language transfer and crosslingual sentence classification, as well as a multilingual subset of the Yelp challenge data set.

We run the evaluation following the recommended default settings, that is, training a logistic regression classifier for the classification tasks, with the Adam optimizer (batch size: 64, epoch size: 4). For the probing tasks we use a multilayer perceptron classifier with sigmoid nonlinearity, 200 hidden units, and 0.1 dropout rate. In order to obtain a sentence vector out of multiple attention heads we apply mean pooling over M, as in Lu et al. (2018).

We present our experiments and their results in the following order: First we present the classification tasks, the SentEval classification tasks on English (Section 4.1) as well as multilingual classification based on XNLI and Yelp reviews (Section 4.2). Next, we turn to the similarity tasks in SentEval (Section 4.3). In all these set-ups we use models trained on Europarl data. Afterwards, in Section 4.4 we turn to a low-resource scenario and study SentEval classification and similarity on the Multi30k data set. The SentEval probing tasks are studied in depth as part of the analysis in Section 5.

4.1 SentEval Classification Tasks

Figure 4 shows the average performance of our models on the various classification downstream tasks. The most frequent baseline achieves an average score of 48.19, which all our models beat by a wide margin. We can see that the multilingual models work best with the many-to-many model, clearly outperforming the rest on average. The figure also illustrates the impact of increasing the number of attention heads. Let us have a closer look at individual classification tasks to get a more detailed picture of the performance in the various settings.

Tables 4 and 5 show the performance of our models on the different downstream tasks. We report the accuracy on each individual test set, including the following comparison scores: a baseline of the most frequent class; a bag-of-vectors baseline obtained by averaging GloVe word embeddings (Pennington, Socher, and Manning 2014); an average of word embeddings as well as the CLS fixed-size sentence vector representation obtained from the large-scale pretrained language model BERT (Devlin et al. 2019; Reimers and Gurevych 2019); a state of the art general-purpose model that exploits large-scale multitask learning on different tasks including machine translation (Subramanian et al. 2018); and the performance from other MT systems by Hill, Cho, and Korhonen (2016) and Conneau et al. (2018a).[9]

The experiments reveal two important findings:

1. In contrast with the results from Cífka and Bojar (2018), our scores demonstrate that an increasing number of attention heads is beneficial for classification-based downstream tasks. All models perform best with more than one attention head and the general trend is that the accuracies improve with larger representations. The previous claim was that there is the opposite effect and that lower numbers of attention heads lead to

9 We only report the best result across the various NMT systems presented by Hill, Cho, and Korhonen (2016) and Conneau et al. (2018a).

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

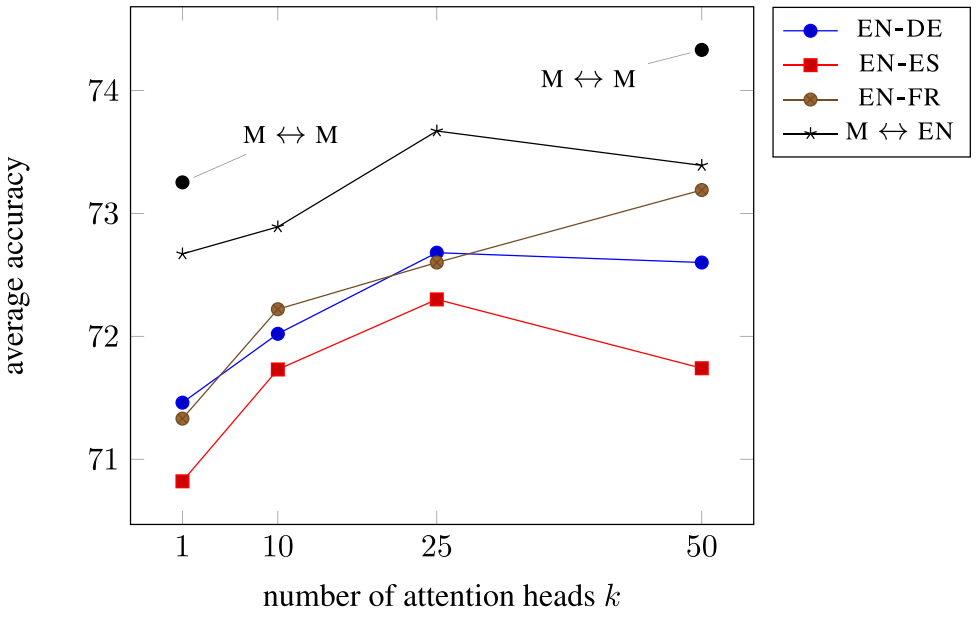

Figure 4
Average scores over the 10 SentEval classification tasks. Results shown for the different trained models.

higher performances in downstream tasks, but we do not see that effect in our set-up, at least not in the classification tasks.

2. The second outcome is the positive effect of multilingual training. We can see that multilingual training objectives are generally helpful for the trainable downstream tasks.

Previous work has focused more on the evaluation of translation alone in the multilingual set-up (Dong et al. 2015) and with our results we can now demonstrate that multilinguality indeed boosts the abstraction power of a fixed-size sentence vector that can be trained with the machine translation objective. Particularly interesting is the fact that the many-to-many model performs best on average even though it does not add any further training examples for English (compared to the other multilingual models), which is the target language of the downstream tasks. This suggests that the model is able to improve generalizations even from other language pairs (DE–ES, FR–ES, FR–DE) that are not directly involved in training the representations of English sentences.

For completeness we also include a comparison with other approaches, even though the comparison is only partly fair, for several reasons (different underlying architecture, different set of hyperparameters, different training data and preprocessing techniques).

First of all, as a sanity check, we observe that our best model reaches far better results than the majority-class baseline. Next, we can see that the results of our best model are better than the best systems by Hill, Cho, and Korhonen (2016) and Cífka and Bojar (2018).

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 4
Accuracies of different models in eight different classification tasks. The average accuracy in the right-most column illustrates the overall trend that a higher number of attention heads and multilingual models are beneficial. Results marked with † taken from Cífka and Bojar (2018); with ‡ from Conneau et al. (2018a); with ⊎ from Reimers and Gurevych (2019).

		CR	MR	MPQA	SUBJ	SSTB	SSTF	TREC	MRPC	AVG
EN-DE	k=1	75.47	68.10	87.49	85.25	71.77	37.15	77.60	70.84	71.71
EN-DE	k=10	74.33	69.29	87.66	85.89	75.12	38.37	73.60	71.83	72.01
EN-DE	k=25	73.93	69.38	87.86	86.13	72.98	38.19	81.40	72.46	**72.79**
EN-DE	k=50	74.41	68.42	87.63	87.42	73.26	38.28	79.60	72.35	72.67
EN-ES	k=1	74.41	66.67	86.95	84.75	70.90	35.93	78.40	70.67	71.08
EN-ES	k=10	72.21	68.72	87.93	86.38	72.05	37.33	77.20	71.83	71.70
EN-ES	k=25	73.72	67.96	87.75	85.79	73.59	36.83	80.00	72.87	**72.31**
EN-ES	k=50	74.06	67.38	87.80	85.84	72.16	36.65	81.20	67.94	71.63
EN-FR	k=1	75.68	68.77	87.27	85.18	71.72	36.97	77.40	70.38	71.67
EN-FR	k=10	74.67	68.89	87.72	86.62	73.59	39.77	75.60	71.65	72.31
EN-FR	k=25	74.41	67.98	87.67	86.33	74.19	38.64	80.00	71.13	72.54
EN-FR	k=50	74.86	69.25	88.26	87.02	75.29	38.06	82.06	72.52	**73.42**
M ↔ EN	k=1	75.28	69.58	88.15	86.98	74.46	38.96	79.60	70.20	72.90
M ↔ EN	k=10	74.07	70.66	88.42	87.63	75.84	38.55	75.80	71.48	72.81
M ↔ EN	k=25	75.36	69.43	88.21	87.33	75.67	39.19	81.80	72.93	**73.74**
M ↔ EN	k=50	75.28	69.87	88.26	87.71	75.12	39.64	80.00	70.14	73.25
M ↔ M	k=1	75.92	71.23	88.07	87.64	75.84	39.73	78.8	73.28	73.81
M ↔ M	k=50	74.72	70.47	88.39	87.98	77.16	40.14	83.00	72.58	**74.31**
Most frequent baseline†		63.80	50.00	68.80	50.00	49.90	23.10	18.80	66.50	48.86
Hill, Cho, and Korhonen (2016) en→fr†		70.10	64.70	81.50	84.90	–	–	82.80	69.10	–
en→cs (2018) †		76.00	68.20	84.90	86.90	72.00	35.70	89.00	70.70	72.92
GloVe-BOW†		78.20	77.00	87.90	91.10	81.00	44.40	82.00	72.30	76.74
Conneau et al. (2018a) en→fi‡		81.10	77.00	90.00	91.50	80.30	43.40	87.20	75.00	–
Subramanian et al. (2018)‡		88.60	82.40	90.70	93.80	85.10	–	94.00	78.30	–
Avg. BERT embeddings⊎		86.25	78.66	88.66	94.37	84.40	–	92.80	69.45	–
BERT CLS-vector⊎		84.85	78.68	88.23	94.21	84.13	–	91.40	71.13	–

Hill, Cho, and Korhonen (2016) train a standard RNN encoder–decoder based system (Cho et al. 2014) on all available English–French data from the 2015 Workshop on Statistical Machine Translation (WMT'15).[10] Similarly to our system, their training set incorporates 2 million English–French sentence pairs from the Europarl corpus. They use additional English–French data, whereas we train on additional English–German and English–Spanish data. We outperform their system in every single classification task (Table 4) when we use multilingual data. Even if we limit ourselves to English–French data, as they did, we outperform them in all tasks but TREC. This suggests that our model is superior in both its way to exploit multilingual data and in its architecture.

10 www.statmt.org/wmt15/translation-task.html.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 5
Results of the two natural language inference (NLI) tasks in SentEval. SICKE = SICK entailment set. Results marked with † taken from Cífka and Bojar (2018); with ‡ from Conneau et al. (2018a).

		SNLI	SICKE
EN-DE	k=1	63.86	77.09
EN-DE	k=10	**65.30**	78.77
EN-DE	k=25	65.13	79.34
EN-DE	k=50	**65.30**	**79.36**
EN-ES	k=1	62.79	76.76
EN-ES	k=10	**66.02**	77.65
EN-ES	k=25	65.20	**79.30**
EN-ES	k=50	65.49	78.83
EN-FR	k=1	63.71	76.19
EN-FR	k=10	65.64	78.08
EN-FR	k=25	**65.68**	**79.97**
EN-FR	k=50	65.47	79.14

		SNLI	SICKE
M ↔ EN	k=1	65.56	77.96
M ↔ EN	k=10	67.01	79.48
M ↔ EN	k=25	66.94	79.85
M ↔ EN	k=50	**67.38**	**80.54**
M ↔ M	k=1	66.92	77.82
M ↔ M	k=50	**67.73**	**81.12**
Most frequent baseline†		34.30	56.70
GloVe-BOW†		66.00	78.20
en→cs (2018)†		69.30	80.80
en→fi (2018a)‡		-	81.70
Subramanian et al. (2018)‡		-	87.40

Hill, Cho, and Korhonen (2016) use the last state of the encoder as their sentence representation, whereas we use the attention bridge layer.

The model by Cífka and Bojar (2018) is based on a very similar architecture as ours, but they train on bilingual data, 57 million English–Czech sentence pairs. We train on a considerably smaller, but multilingual, data set (3 times 2 million sentence pairs of EN–FR, EN–DE, and EN-ES). Yet our system outperforms theirs in six out of nine tasks listed in Tables 4 and 5. This again demonstrates the power of multilingual models.

In futher comparisons, we can see that our model outperforms the competitive baseline of GloVe-BOW (Kruszewski et al. 2015; Arora, Liang, and Ma 2017; Adi et al. 2017) in five tasks out of ten. However, Conneau et al. (2018a) and Subramanian et al. (2018) perform better than us in all the classification and NLI tasks. We believe that the strong performance of the latter models is explained by orders of magnitudes of more training data. GloVe-BOW and the Conneau et al. (2018a) model are based on word embeddings, which have been pretrained on several billions of words of text. The large vocabularies of the pretrained embeddings provide better representations for low-frequency as well as out-of-vocabulary words. Subramanian et al. (2018) use 124 million sentence pairs for training, which is 20 times more than we have. The BERT models, trained on 3.3 billion words, do not quite reach the level of Subramanian et al. (2018).

Although our aim is not to beat the state of the art, but rather to understand the impact of various sizes of attention heads in a bilingual and multilingual scenario, we argue that a larger attention bridge and multilinguality constitute a preferable starting point to learn more meaningful sentence representations. With this, we can contrast and extend previous findings, leading the way to further extensions of the MT-based framework for crosslingual representation learning.

4.2 Multilingual Classification Tasks

In the previous section, we focused on downstream tasks that consider English only. The main point was to show that reasonable representations can be learned from the

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

translation objective and that multilingual data help to improve the abstractions that can be derived. Even more intriguing is the fact that our shared representation combines language-specific encoders with language agnostic representations. This makes it possible to directly test crosslingual downstream tasks, which we will focus on in this section.

The interest in crosslingual NLP leads to a number of benchmarks and downstream applications and here we will consider the framework of crosslingual NLI as defined by the XNLI challenge (Conneau et al. 2018c) and crosslingual review classification as proposed by Lu et al. (2018). We start with the XNLI results and then turn to the multilingual classifier based on Yelp reviews.[11]

4.2.1 XNLI. The idea of the XNLI challenge is that the provided corpus enables us to test natural language inference across different languages. The test pairs are all translated into 14 languages, which makes it possible to obtain comparable results across various language pairs. Hence, a classifier can be trained on one language and be tested on another one. In order to make this work, one essentially needs to produce crosslingual sentence representations that are useful for the task in all test languages. Table 6 summarizes the results obtained for different settings. We rely on our multilingual attention bridge model trained in a many-to-many fashion.

For comparison, we include representations obtained from large pretrained word embeddings. Note that those embeddings are trained on vastly more data than our model, which is trained on the parallel Europarl corpus. In particular, we use the multilingual word embeddings from the fastText (Grave et al. 2018) and the MUSE (Conneau et al. 2018b) libraries. The fastText algorithm is based on word2vec and produces word embeddings compounded from character *n*-grams (Bojanowski et al. 2017), which is to be preferred for morphologically richer languages in a multilingual setting. The fastText word vectors are pretrained in CommonCrawl and Wikipedia using CBOW with position weights, whereas MUSE word embeddings are Wikipedia fastText vectors from 30 languages aligned in a supervised way into a single vector space. Because fastText vectors are not aligned into the same space we only present the accuracies on the relevant languages for each case. To obtain sentence representations, we compute the average of the individual word vectors.

We use the XNLI corpus to train multilingual classifiers that are then to be tested for zero-shot classification. Logistic regression classifiers are trained on top of the sentence embeddings produced with the English, German, French, and Spanish training data, or a combination of these, and then tested in all four languages. We observe that our model is clearly better than the fastText and MUSE benchmarks. Besides, it reaches results equal to the XCBOW model presented in Conneau et al. (2018c)—a model that incorporates an additional multilingual loss to enhance the performance in the task, and is based on a feed-forward neural network classifier instead of simple logistic regression. Note that the XCBOW results are taken directly from the original paper and only available for a classifier trained on English. Even though the comparison with XCBOW scores is not completely fair because the data sets we applied in training are much smaller and narrow in domain (Europarl only), we reach similar performance. It is certainly re-assuring that our model is capable of creating language-agnostic representations that are properly aligned with language-specific encoders. Furthermore, our model outperforms the crosslingually aligned MUSE word embeddings by a significant margin, which demonstrates the importance of a proper sentence encoder also in the

11 http://www.yelp.com/dataset.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 6
Accuracy obtained in the XNLI task comparing our multilingual model with pretrained vector embedings. Results marked with † taken from Conneau et al. (2018c) and use a feed-forward neural network as a classifier as opposed to all others that use a logistic regression.

	Classifier trained on	**EN**	**DE**	**FR**	**ES**
M ↔ M k=50	EN	65.0	59.0	55.9	58.0
MUSE		55.9	47.8	42.5	42.8
fastText		53.1	–	–	–
X-CBOW†		64.5	61.0	60.3	60.7
M ↔ M k=50	DE	58.9	62.6	54.8	57.9
MUSE		50.5	53.1	39.2	42.5
fastText		–	52.9	–	–
M ↔ M k=50	FR	62.6	60.9	63.5	61.9
MUSE		45.9	43.3	53.0	41.8
fastText		–	–	48.6	–
M ↔ M k=50	ES	61.2	59.0	56.3	63.8
MUSE		49.2	44.5	43.1	52.5
fastText		–	–	–	50.6
M ↔ M k=50	EN+DE	65.3	62.8	58.6	60.8
MUSE		55.4	53.4	40.1	43.1
M ↔ M k=50	EN+FR	65.2	61.6	63.8	61.7
MUSE		55.0	46.5	52.8	44.5
M ↔ M k=50	EN+ES	64.9	60.9	58.5	64.6
MUSE		54.5	46.9	43.4	52.0
M ↔ M k=50	EN+DE+FR+ES	64.8	62.8	63.0	63.6
MUSE		54.8	51.7	51.6	51.2

crosslingual setting. The same effect can be seen in comparison to the non-aligned language-specific fastText word embeddings that have been trained on huge amounts of training data.

Looking at the different combinations of training data and the difference between supervised and zero-shot classification, we can see that our model is quite robust across the different settings. The drop in performance when moving to zero-shot classification is rather modest in most cases and in some of them we achieve scores that are close to the fully supervised mode (see, for example, the results for Spanish with the classifier trained on French). The results also show that all languages seem to be equally covered and the supervised scores end up to be around 63–65% accuracy with zero-shot scores ranging from roughly 56–62%. Interesting to note is that a combination of training languages can lead to improvements of zero-shot classification. For example, results for Spanish and French are improved when combining English and German in the training data. The same happens for German and Spanish when combining English and French, and so forth.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

4.2.2 Yelp. Another crosslingual task is the multilingual review classification task proposed by Lu et al. (2018). The idea is to train a classifier to label online reviews based on their ratings to decide whether reviews in another language are received in a positive or in a negative way. Table 7 shows the scores achieved by an English-reviews classifier when tested on French, German, and Spanish sentences. To make the results as comparable as possible to Lu et al. (2018) we use the same settings as they did, with the addition of including Spanish in our experiments. Namely, we took a subset of 5,000 reviews in English from the Yelp review data set (Round 13) to train a simple logistic regression classifier, as well as test sets of French, English, German, and Spanish reviews, of 1,000 sentences each. As in their approach, we extracted the non-English reviews by applying a language detection tool (Joulin et al. 2017). We use binary review scores, where 4- and 5-star reviews are labeled as positive, we do not use 3 stars, and 1- and 2-star reviews are labeled as negative. We treat each review as a sentence and use the shared intermediate representations produced by our multilingual systems as input to the classifier. As before, we compare our results to MUSE and fastText baselines, obtaining the sentence representations by averaging the word embeddings of a full review.

In Table 7 we present models with different sizes of the attention bridge for the many-to-English set-up and, finally, also the results for representations coming from a many-to-many model. We only include classifiers trained on English as there are not enough data for the other languages to train a reliable classifier.

The results show that our representations perform well across all languages with scores mostly around 83–87% accuracy. The most interesting thing is that there is basically no deterioration when moving to zero-shot classification for languages other than English. Note that the class distribution is quite skewed here and the majority class baseline is quite high, especially for German and French. This makes it rather difficult to interpret the results at least for those languages with a heavy over-representation of positive reviews. Spanish might be the most reliable zero-shot language in our test set with a more balanced distribution and, here, we can see a clear improvement over the majority-class baseline when applying the representations coming from our multilingual translation model. We also clearly outperform MUSE in that case demonstrating the ability of our sentence encoders in creating reliable crosslingual representations. Again, MUSE and fastText are not directly comparable as they have been trained on much larger and more diverse data sets. Nevertheless, we also outperform fastText-based representations in the supervised case in all settings except one. The picture about

Table 7
Accuracy for crosslingual Yelp binary review classification.

		EN	DE	FR	ES
% Positive		75.3	86.6	83.0	69.7
MUSE		81.5	86.7	83.1	76.0
fastText		82.5	–	–	–
M ↔ EN	k=1	**84.9**	84.6	83.5	79.6
M ↔ EN	k=10	83.7	84.4	83.6	**83.6**
M ↔ EN	k=25	81.2	85.7	76.1	83.8
M ↔ EN	k=50	83.3	79.6	83.8	83.4
M ↔ M	k=50	82.6	**87.1**	**84.1**	81.8

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

the perfect size for the attention bridge is not very clear. In some cases, a small size is preferable whereas in others a larger size is beneficial. Also, the effect of additional language pairs is not entirely clear but on average the many-to-many set-up produces better scores compared with the similar set-up with many-to-English models. Some additional studies might reveal further insights, which we will leave for future work.

4.3 SentEval Similarity Tasks

The next evaluation refers to the similarity tasks of SentEval—that is, English data only. Table 8 summarizes the results using Pearson's correlation coefficient as well as the average on all tasks. As comparison we include the bag-of-vectors baseline (GloVe-BOW) as in the earlier SentEval classification tasks, the best model from Cífka and Bojar (2018), and the InferSent model (Conneau et al. 2017) as a state-of-the-art model that is pretrained on a natural language inference (NLI) task. As discussed earlier, note that the SICK and STSB benchmarks provide training data where a classifier learns to predict the probability distribution of the relatedness scores (Tai, Socher, and Manning 2015). Two different trends become visible:

i) On the unsupervised textual similarity tasks, having fewer attention heads is beneficial. Contrary to the results in the classification tasks, the best overall model is

Table 8
Results from seven similarity tasks, measured using Pearson's correlation coefficients; the Spearman's coefficients exhibit the same trend. The average values are displayed in the right-most column. Results marked with † taken from Cífka and Bojar (2018).

		SICK	**STSB**	**STS12**	**STS13**	**STS14**	**STS15**	**STS16**	AVG
EN-DE	k=1	0.74	**0.69**	0.57	0.46	0.58	0.63	0.62	**0.61**
EN-DE	k=10	0.76	**0.69**	0.52	0.41	0.54	0.58	0.56	0.58
EN-DE	k=25	**0.78**	0.67	0.50	0.39	0.50	0.57	0.50	0.56
EN-DE	k=50	0.78	0.65	0.47	0.36	0.46	0.54	0.46	0.53
EN-ES	k=1	0.73	**0.68**	0.54	0.42	0.53	0.60	0.60	**0.58**
EN-ES	k=10	0.77	0.68	0.53	0.37	0.52	0.58	0.55	0.57
EN-ES	k=25	0.77	0.64	0.50	0.35	0.47	0.56	0.48	0.54
EN-ES	k=50	**0.78**	0.63	0.48	0.31	0.40	0.49	0.44	0.51
EN-FR	k=1	0.74	0.66	0.56	0.44	0.57	0.62	0.62	**0.59**
EN-FR	k=10	0.75	**0.67**	0.53	0.36	0.51	0.57	0.58	0.57
EN-FR	k=25	0.77	0.65	0.50	0.36	0.48	0.55	0.48	0.54
EN-FR	k=50	**0.77**	0.63	0.46	0.38	0.45	0.53	0.45	0.52
M ↔ EN	k=1	0.76	0.69	0.53	0.38	0.52	0.56	0.57	**0.57**
M ↔ EN	k=10	0.78	**0.69**	0.51	0.34	0.47	0.56	0.55	0.56
M ↔ EN	k=25	0.78	0.68	0.46	0.32	0.42	0.51	0.43	0.51
M ↔ EN	k=50	**0.79**	0.66	0.45	0.30	0.36	0.47	0.43	0.50
M ↔ M	k=1	0.77	**0.71**	0.53	0.39	0.52	0.58	0.59	**0.59**
M ↔ M	k=50	**0.79**	0.69	0.47	0.30	0.35	0.45	0.41	0.50
GloVe-BOW†		0.80	0.64	0.52	0.50	0.55	0.56	0.51	0.59
en→cs (2018)†		0.81	0.73	0.46	0.32	0.45	0.53	0.47	0.54
InferSent†		0.88	0.76	0.59	0.59	0.70	0.71	0.71	0.70

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

provided by a bilingual setting with only one attention head. This is in line with the findings of Cífka and Bojar (2018) and could also be expected as the model is more strongly pushed into a dense semantic abstraction that is beneficial for measuring similarities without further training. More surprising is the negative effect of the multilingual models. We believe that the multilingual information encoded jointly in the attention bridge hampers the results for the monolingual semantic similarity measured with the cosine distance, while it becomes easier in a bilingual scenario where the vector encodes only one source language data, English in this case.

ii) On the supervised textual similarity tasks, we find a similar trend as in the image-caption models for SICK: Both a higher number of attention heads and multilinguality contribute to better scores. For STSB, we notice a different pattern. Namely, including multilingual data in the models helps the performance in this task. The many-to-many models score better than the best bilingual models. Moreover, when increasing k, the multilingual models scores are not as badly hampered as the bilingual ones.

Comparing against two baseline systems, GloVe-BOW and the best model by Cífka and Bojar (2018), our best model on average achieves better results showing the potentials of our approach. The strength of the InferSent model is probably explained by the pre-trained word embeddings extracted from billions of words as well as their training data, which was taken from the NLI domain rather than parallel corpora of translated sentences.

4.4 Downstream Tasks in a Low-Resource Setting

Another important test is whether our model is capable of learning reasonable representations from limited resources. In order to study this, we applied the Multi30k models from Section 3.1 to the SentEval tasks in the same way as the Europarl models discussed above. Note that these scores are not directly comparable to other models trained on large-scale data sets because the Multi30K models were trained on very limited data sets; they contain 30k sentences on the specific domain of image captioning. Tables 9 and 10 summarize the scores on downstream tasks we obtain for bilingual and for multilingual models. We ran each experiment with five different seeds, and we present the average of these scores.

We notice that for the classification and NLI tasks of the SentEval collection, the sentence embeddings produced by the multilingual models show consistent improvements, with only two exceptions. Moreover, we observe that our many-to-many model obtains better results in the SICK Relatedness (SICKR) and STS-Benchmark (STS-B), that

Table 9
Accuracies of models trained with limited data in eight different classification tasks. Including more languages during training boosts the performance.

	CR	MR	MPQA	SUBJ	SSTB	SSTF	TREC	MRPC
EN-DE	68.37	59.76	73.19	75.26	61.87	31.15	67.44	69.13
EN-CS	67.79	59.71	73.16	75.32	61.92	30.43	67.75	68.23
EN-FR	68.52	60.08	**73.51**	77.25	61.91	30.55	61.04	**70.96**
M ↔ EN	68.32	60.4	72.98	78.64	62.02	**32.10**	**69.84**	68.83
M ↔ M	**69.01**	**61.80**	73.28	**80.88**	**62.24**	31.83	66.4	70.43

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 10
SentEval NLI and semantic similarity tasks results for the models trained with limited data obtained. We present the Pearson's correlation coefficient for the similarity tasks. Models trained with more languages get better scores in trainable tasks, whereas the non-trainable tasks show a different behavior.

	NLI Tasks		Semantic Similarity Tasks						
			Trainable		non-Trainable				
	SNLI	**SICKE**	**SICKR**	**STS-B**	**STS12**	**STS13**	**STS14**	**STS15**	**STS16**
EN-DE	61.45	72.82	0.618	0.564	**0.393**	**0.265**	**0.426**	0.489	**0.430**
EN-CS	61.75	73.89	0.652	0.616	0.385	0.234	0.380	**0.505**	0.422
EN-FR	60.95	74.85	0.646	0.574	0.359	0.220	0.428	0.476	0.430
M ↔ EN	64.52	75.46	0.659	0.618	0.323	0.190	0.353	0.418	0.375
M ↔ M	**65.12**	**76.92**	**0.677**	**0.630**	0.327	0.256	0.415	0.460	0.400

is, the trainable semantic similarity tasks. However, the results of the non-trainable (semantic similarity) tasks exhibit a different behavior (see the rightmost part of Table 10), which can be explained by the fact that the additional information encoded in multilingually trained embeddings cannot be effectively separated from the information that is necessary for monolingual similarity measures. In other words, the attention bridge layer of the multilingual models outputs vectors that contain information shared across languages, rendering them incomparable with the cosine similarity.

5. Linguistic Analyses of Inner-Attention-Based Sentence Representations

In this section, we take a closer look at the representations and what they actually encode. For that, we use the probing tasks that are available in the SentEval toolkit, which make it possible to study specific linguistic features of a given representation (see Sections 5.1 and 5.2). Furthermore, we add a careful analysis of the individual attention heads to further study the information that is encoded by the internal shared representation layer (see Section 5.3).

5.1 Probing Tasks in SentEval

The SentEval probing tasks inspect three different linguistic categories: surface information on the length and word content of the sentence (Adi et al. 2017); several probing tasks regarding syntactic properties, such as word ordering, top constituent task (Shi, Padhi, and Knight 2016); and semantic properties, such as subject/object number, odd man out, and coordination inversion, to name a few. To assess how our representations correlate with linguistic properties, we carry out an evaluation on the various categories of these tasks. Table 11 shows the accuracies on the three linguistic classes.

The results show a clear correlation between the number of attention heads and the syntactic power of the representations: An increased number of attention heads corresponds to a higher syntactic score (up to 5 points). Intuitively, more attention heads can learn long-range dependencies better, which are essential for a better syntactic understanding of the sentence. For the semantic probing test, the best results are provided by the models with one attention head, in line with the findings of

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 11
Average accuracy of the syntactic and semantic probing tasks (Europarl models). The accuracy on the surface information tasks is shown on the rightmost columns.

		Average syntactic	Average semantic	Surface info	
				Length	WC
EN-DE	k=1	55.87	**72.52**	82.20	**63.80**
EN-DE	k=10	57.97	71.04	85.40	61.40
EN-DE	k=25	59.10	70.58	**86.90**	56.70
EN-DE	k=50	**59.97**	71.84	86.00	49.60
EN-ES	k=1	54.90	**72.48**	82.30	56.50
EN-ES	k=10	58.37	71.36	87.50	**60.80**
EN-ES	k=25	59.27	71.44	**88.00**	52.80
EN-ES	k=50	**59.70**	72.26	87.30	46.50
EN-FR	k=1	55.37	72.04	82.30	60.60
EN-FR	k=10	58.13	71.16	**86.70**	**61.50**
EN-FR	k=25	58.63	71.30	86.40	53.80
EN-FR	k=50	**59.40**	**72.42**	86.30	45.70
M ↔ EN	k=1	57.83	**73.86**	89.10	**58.70**
M ↔ EN	k=10	61.33	72.60	90.30	54.20
M ↔ EN	k=25	60.87	72.78	**91.50**	41.20
M ↔ EN	k=50	**61.70**	73.60	92.10	34.50
M ↔ M	k=1	59.43	**74.20**	**91.30**	**58.90**
M ↔ M	k=50	**60.97**	73.40	90.70	35.60

Section 4.3, except that we do not observe a high degradation with more attention heads. Regarding the surface information, it is clear that more attention heads contribute to understanding the length of the sentence better, while forgetting about word content.

5.2 Probing Tasks with Limited Resources

Similar to Section 4.4, we also apply the same probing tasks to the Multi30K models to demonstrate the ability of the models to learn from limited resources.

Table 12 compares scores between bilingual and multilingual models. Again, we observe improvements in the majority of cases when adding multiple languages to the training procedure. Remarkably, we observe a significant increment on the accuracy for the specific tasks of Length (superficial property), Top Constituents (syntactic property), and Object Number (semantic information) when training the encoders with multilingual data. Some of the tests result in better scores with the many-to-English setting compared to the many-to-many set-up, which is slightly surprising. Nevertheless, multilingual models outperform the bilingual models in all but one test.

5.3 Analysis of Individual Attention Heads

So far we have evaluated sentence representations that are based on attention bridges of different sizes, ranging from 1 to 50 attention heads. The attention bridge matrix as

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 12
Scores for the SentEval probing tasks (Multi30k models).

	Length	WC	Depth	TopConst	BShift	Tense	SubjNum	ObjNum	SOMO	CoordInv
EN-DE	75.69	**10.02**	31.46	39.74	57.41	65.69	66.18	67.94	49.70	61.05
EN-CS	80.76	9.55	30.21	39.38	56.76	66.07	64.21	67.22	49.11	60.74
EN-FR	80.00	9.66	32.14	40.12	57.22	67.61	68.55	70.01	49.90	61.38
M↔EN	84.76	9.56	**33.05**	**44.04**	58.35	**69.36**	69.67	72.19	49.46	60.57
M ↔ M	**85.41**	9.13	31.6	39.76	**59.76**	68.27	**69.89**	**73.29**	**50.12**	**62.21**

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

a whole has been used in our experiments. An interesting further step is to investigate whether specialized roles are assigned to the different attention heads (columns in the matrix) and whether they effectively learn to focus on different parts of the sentence.

In order to analyze this, we conduct additional studies on the Europarl models, and we further assess the performance of individual heads on the SentEval linguistic probing tasks. Our aim is to identify whether some attention heads are particularly important in some probing task, and whether we see differences in how the roles of the heads are distributed depending on the size of the attention bridge.

We compare attention bridges of sizes 10, 25, and 50, and, to study the effect of the individual attention heads, we detach one attention head at a time to be used as the representation of the sentence and apply them in the various probing tasks. Figure 5 shows two scenarios, attention bridges of sizes 10 and 50 taken from the many-to-many Europarl model. In both scenarios, the accuracy of each probing task is shown for each attention head, separately. To test stability, we trained the probing tasks with five different seeds and present the average accuracy score in the figure. The variance is very small (of order $\leq$1e-1) in all cases and, hence, the trends shown in the illustration are reliable.

If the attention heads obtain different, specialized roles, we expect to see a higher accuracy on one probing task for a specific attention head and a higher accuracy on some other task for another one. In the case of 10 attention heads ($k = 10$), it is hard to see any clear pattern. In contrast, with 50 heads ($k = 50$) there are some tasks in which the performance of the heads with a higher index (> 40) is clearly better than the performance of the ones with low indexes (< 15). The probing tasks, in which this phenomenon is particularly pronounced, are *word content*, *top constituents*, *number of the subject* (subjnumber), *number of the object* (objnumber), and possibly *verb tense* (tense). However, rather than differentiation there is strong relation between these tasks, such that an attention head that performs well on one task also performs well on the other tasks.

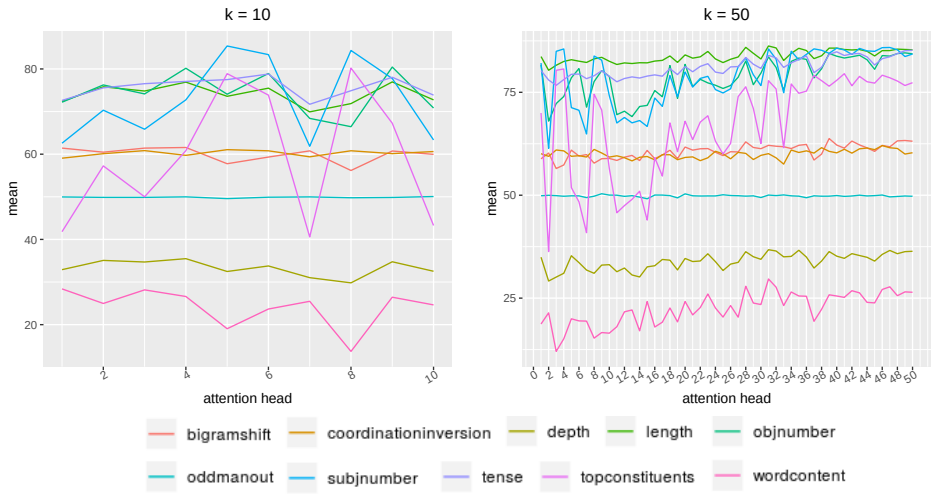

Figure 5
Performance of each attention head on the probing tasks. Mean accuracy along different runs of SentEval probing tasks using the trained {DE,FR,CS} $\leftrightarrow$ EN models with $k = 10$ and 50.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

From the illustration, we can see that there is an interesting pattern with increasing performance on attention heads with higher index. This can be seen especially on tasks like *wordcontent*, *topconstituents*, *subjnumber*, and *objnumber*. The differences are small but this trend is still clearly visible and rather unexpected. In order to check whether this observation is purely coincidental, we computed the average accuracy scores over all Europarl models in our experiments averaging the numbers from the ones with the same size of the attention bridge (in number of heads). This includes bilingual, many-to-one, and many-to-many models. Figure 6 shows the average accuracy scores and their variance on four tasks: *wordcontent*, *topconstituents*, *subjnumber*, and *objnumber*. In this case, we also show figures from the $k = 25$ settings. The most striking outcome of this study is that the same positive trend can be observed for all independently trained models with larger attention bridges without aligning the inner-attention heads or enforcing this behavior, instead of consistently having a large range and a stationary mean accuracy, as one would expect from combining without any alignment the results of the inner-attention layers of models trained independently. In the smallest attention bridge ($k = 10$) it is difficult to see whether any head performs better than any other. In the medium size attention bridge ($k = 25$), there is already a hint at higher performance on the tasks *topconstituents* and *subjnumber* for the attention heads with the highest

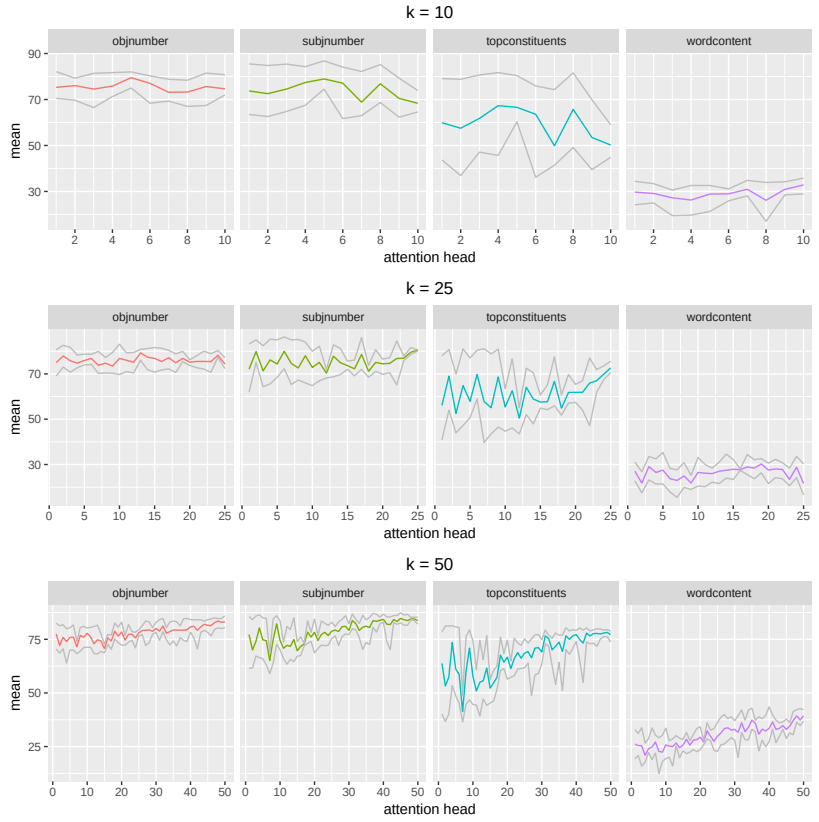

Figure 6
Mean accuracy across the different models with $k = 10, 25, 50$ on four probing tasks. Gray lines surrounding the mean values are the min and max values obtained for each head across models.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

indexes. In the largest attention bridge ($k = 50$), on all four tasks, accuracies clearly increase with high indexes.

We can also see that the variance is generally higher in the case of a small attention bridge, while the scores become more stable with larger attention bridges and to some extent higher attention head indexes. This seems to suggest that there are indeed certain attention heads that specialize on specific tasks. The system does not have any explicit mechanism to ensure that the attention heads will be aligned. This means that it is possible that exactly the same information encoded by head i in one training would be encoded by head j during another training. We observe that the first attention heads indeed follow this behavior, whereas the later ones start behaving in a similar way. Why part of this specialization is happening in the same order is less clear and this is probably due to some internals of the underlying implementation of the training algorithms.

To have a better understanding on why the last attention heads seem to be the best ones in some of the probing tasks, we plotted heat maps of which areas of a sentence each of them focuses on. An example sentence (*"We cannot afford to lose more of the momentum that existed in the beginning of the Nineties."*) is shown in Figure 7. For each scenario ($k = 1, 10, 25, 50$) the patterns for the attention heads are shown, such that every line illustrates the main focus areas of one of them. We obtain similar pictures for

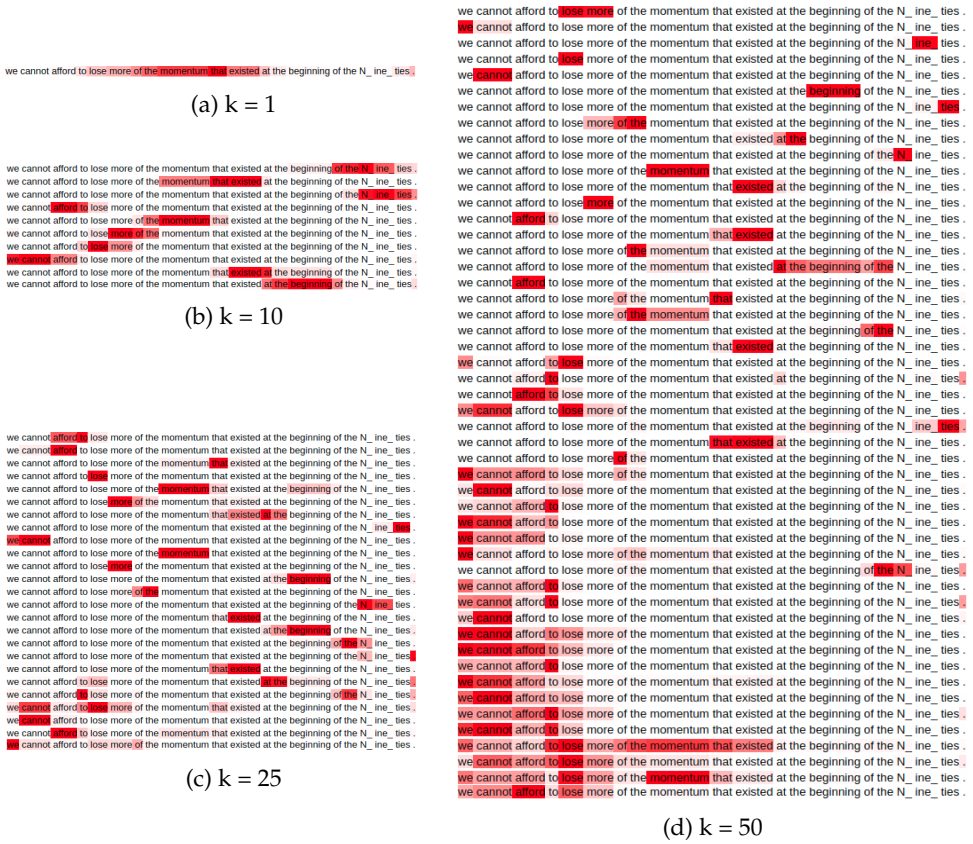

Figure 7
Heat map showing where each attention head (rows) focuses, using an example sentence. Four different attention bridges from the {DE,FR,CS} ↔ EN models are used in the visualization, where k is the number of heads in the bridge.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

other sentences and different models (bilingual, many-to-one, or many-to-many) and the figure is just a good example of the general picture we get.

The illustration suggests that the model learns to distribute attention to individual words or bigrams among the attention heads until the sentence is sufficiently covered. The order of positions seems rather random and not consistent as the variance of probing task results suggest. Once there are additional attention heads available, the remaining heads start picking up longer word sequences, predominantly from the beginning of the sentence. This effect seems to explain the increasing performance on the probing tasks that we have observed above.

The main clause is often in the beginning of a sentence, which explains why we see high accuracies on the tasks *top constituents*, *number of subject*, and *number of object*. The goal in the *top constituents* task is to determine the high level syntactic structure of the sentence in terms of the top constituents immediately below the sentence (S) node in a phrase structure grammar; an example of such a structure is: ADVP NP VP, as in *"Then it happened."* In the *number of subject* task, we need to determine whether the subject of the main clause of the sentence is in singular or plural number. The number of the subject is marked on the subject itself (plural -s), but also in determiners (a, an) and sometimes in the predicate verb (third person singular -s). Similarly, in the task *number of object*, we need to decide whether the object of the main clause is in singular or plural. In order to excel in these tasks, it is crucial to be able to identify the area of the sentence where the main clause is located, which apparently the last attention heads are often capable of doing.

Furthermore, especially in the tasks *top constituents* and *word content* it is necessary to analyze multiple words. The aim of the *word content* task is to identify which words in a set of 1,000 preselected mid-frequency words are present in the sentence. An attention head covering a larger area is naturally more likely to recognize a larger number of words and, hence, performs better on that task. This also explains why the *word content* task works well with an attention bridge consisting of only one attention head ($k = 1$). If there is only one head available, the attention needs to be spread over a broader range of words, which is beneficial for some tasks and detrimental for other tasks (see Section 5.1).

Our hypothesis of differentiated roles of the separate attention heads has been partly confirmed. Different attention heads do focus on different areas of the sentence. However, they do not seem to specialize on information that is beneficial for a specific downstream probing task. Rather, we see significant correlation, such that some attention heads are particularly strong at multiple probing tasks concurrently.

We can also observe that the attention of individual heads is in general very focused with a low kurtosis in its distribution until the sentence is covered. Once this point is reached, the penalty term forces the inner-attention to explore collocational relationships typically from the beginning of the sentence incrementally extending its span. Interestingly, this behavior is accentuated by the penalty term in the loss function that we introduced in Section 2.2. To illustrate this, we computed the entropy over the attention heads of a model trained with and without penalty term using the following procedure:

We randomly sampled 10K sentences of length up to 40 tokens from the English monolingual NewsCrawl corpus 2018,[12] and calculated the entropy values over the weights A from Equation (6) using the EN-DE models with 50 attention heads. In

12 http://www.statmt.org/wmt19/translation-task.html.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Table 13
Comparison of entropy values for models trained with and without penalty term.

	With Penalty Term	Without Penalty Term
$k \in [1, 50]$	1.20	0.63
$k \in [1, n_i]$	0.49	0.43
$k \in [n_i, 50]$	1.73	0.82

Table 13 we present three different entropy values using *(i)* all the attention heads, $k \in [1, 50]$, *(ii)* the initial n_i attention heads until reaching the length of the sentence, $k \in [1, n_i]$, and *(iii)* the final $50 - n_i$ attention heads, $k \in [n_i, 50]$; where n_i is the length of sentence i.

A higher entropy value indicates a broader spread of the attention, which leads us to the conclusion that the penalty term is indeed forcing to scatter the attention once the initial (position-specific) attention heads have covered the whole sentence. Up to the length of the sentence ($k \in [1, n_i]$) its presence does not make a significant difference (entropy values 0.49 vs. 0.43). However, for the attention heads beyond the sentence length ($k \in [n_i, 50]$), we see that the model without it ends up focusing on a much smaller portion of the sentence (entropy 0.82) compared with the model incorporating the penalty term (entropy 1.73). Manual inspection of selected heat maps verifies this result with redundant attention mostly fixed on the first token of the sentence in models where the penalty term is left out.

In future work, we would like to investigate the behavior of attention in our model with a larger diversity of languages and tasks. An interesting question is whether we can see the same trend with less related languages included in our data and whether we can force specialization using certain constraints or augmented loss functions during training. We also want to explore the effect of adding other tasks that can be modeled with the same architecture including speech recognition, sequence labeling, or parsing.

6. Related Work

Before concluding the paper, we briefly summarize related work.

Multilingual NMT has been widely studied and developed in different pathways during recent years (Dong et al. 2015; Luong et al. 2016; Chen et al. 2017; Johnson et al. 2017). Work has been done with networks that use language specific encoders and decoders, such as Dong et al. (2015), who used a separate attention mechanism for each decoder on one-to-many translation. Zoph and Knight (2016) exploited a multi-way parallel corpus in a many-to-one multilingual scenario, while Firat, Cho, and Bengio (2016) used language-specific encoders and decoders that share a traditional attention mechanism in a many-to-many scheme. Another approach is the use of universal encoder–decoder networks that share embedding spaces to improve the performance of the model, like the one proposed by Gu et al. (2018b) for improving translation on low-resourced languages and the one from Johnson et al. (2017), where the term *zero-shot translation* was coined. Even though the latter model is proven to be effective and appealing thanks to its simplicity, it does not scale well to the use of many languages. Sharing all parameters of the model leads to a strongly degrading performance when covering many diverse languages and new languages cannot easily be added on neither

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

encoder nor decoder side. Furthermore, it does not provide a straightforward meaning representation that can be used for downstream tasks.

Sentence meaning representation has also been vastly studied under NMT settings. When introducing the encoder–decoder architectures for MT, Sutskever, Vinyals, and Le (2014) showed that seq2seq models are better at encoding the meaning of sentences into vector spaces than the bag-of-words model. Recent work includes that of Schwenk and Douze (2017), who used multiple encoders and decoders that are connected through a shared layer, albeit with a different purpose than performing translations; Platanios et al. (2018) showed an intermediate representation that can be decoded to any target language while describing a parameter generation method for universal NMT; Britz, Guan, and Luong (2017) made a computational efficiency analysis for MT using a fixed-size attention layer; Artetxe and Schwenk (2019) used a shared LSTM with max pooling to learn sentence embeddings on 93 translation directions; Cífka and Bojar (2018) introduced an architecture with a self-attentive layer to extract sentence meaning representations of fixed size. Here we use a similar architecture but in a multilingual setting.

Our work on multilingual MT and sentence representations is closely related to the study from Lu et al. (2018). There, the authors attempt to build a neural interlingua by using language-independent encoders and decoders which share an attentive LSTM layer. Our approach differs on the choice of the crosslingual shared layer; we use a shared inner-attention mechanism in contrast to having a feedforward layer on top of a shared LSTM. Additionally, we also experiment in a multilingual many-to-many setting, instead of only exploring the one-to-many or many-to-one scenarios.

7. Conclusions

We have shown that fixed-size sentence representations can effectively be learned with multilingual machine translation using an inner-attention layer and scheduled training with multiple translation tasks. The performance of the model heavily depends on the size of the intermediate representation layer and we can show that a higher number of attention heads leads to improved translation and stronger representations in supervised downstream tasks, a result that contradicts earlier findings.

Multilinguality also helps boost the performance in the aforesaid downstream tasks although it does not necessarily contribute to improve translation performance when trained with large data sets. However, multilinguality does substantially benefit translation performance in low-resource scenarios. The multilingual training objectives enable effective transfer learning that leads to an improvement of up to 4.43 absolute BLEU points in an established image caption translation task. We can observe that even languages that are different from both, the source and the target language of the test case, help the overall model to achieve a better translation quality. This indirectly hints at the additional semantic abstraction that the model can pick up from the multilingual signal.

Our further analysis reveals that the attention bridge model mainly struggles with long sentences. In fact, the decrease in performance that we observe on the Europarl domain and the large-data scenario is entirely due to the drop in performance on sentences above 45 tokens. Here we can see the limits of the fixed-size attention bridge and the results are not particularly surprising. A future research direction is to study ways to prevent this behavior and to investigate the impact of increasing the linguistic diversity in translation performance.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

The main purpose of the model is to act as an efficient way of learning language-agnostic representations using translation as the auxiliary training objective. We verify on the larger data set that the system is able to produce sentence embeddings that encode essential syntactic and semantic information in a sentence. The results on downstream tasks in the large data set-up follow the same trend as we have shown in the low-resource scenario. Punctually, multilinguality helps in supervised tasks. Generalizations can be improved even from language pairs not involving the language we test on (English).

We make a careful study on the impact of the size of the attention bridge and the role of individual attention heads in encoding linguistic information. In short, increasing the size of the bridge does not only lead to better translation performance but also to improved results in supervised downstream tasks. To further study the impact of individual attention heads on sentence encoding, we detach each of them in turn and test their ability to perform a number of linguistic probing tasks. This experiment reveals a surprising trend in which the attention heads with a high index in the attention bridge perform better on various tests. Looking at the regions of sentences where these higher-index heads focus on shows that they spread their attention more broadly than their lower-index counterparts across the encoded sentence, which is beneficial for those probing tasks. A similar explanation can be found for attention bridges of limited size that perform well on the same probing tasks, in which an attention head spreads the focus over larger regions of the sentence. This analysis provides important insights about the process of encoding and its applicability in other tasks.

Our findings open many directions for future research. First of all, we would like to study the impact of increasingly diverse training material on sentence representations that can be learned in the translation set-up. This includes the use of additional languages from different language families and the integration of multiple domains and textual genres, as well as the effect of using resource-rich language pairs on the performance of low-resource languages. We would like to further investigate the linguistic properties that a system picks up and, for example, study the implicit relational structure (syntactic and semantic) that is captured by inner-attention in a multilingual encoder. For this, we will also increase the complexity of the shared architecture with additional layers and connections. We will systematically look at different training schedules that vary the amounts of shared parameters depending on linguistic and typological properties that we want to investigate.

Acknowledgments

This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 771113). The authors gratefully acknowledge the support of the Academy of Finland through project 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence and projects 270354 and 273457. Finally, we would also like to acknowledge CSC – IT Center for Science, Finland, for computational resources, as well as NVIDIA and their GPU grant.

References

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *5th International Conference on Learning Representations, ICLR 2017*, Conference Track (Poster), Toulon.

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea et al. 2015. SemEval-2015 Task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop*

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

on Semantic Evaluation (SemEval 2015), pages 252–263, Denver, CO.

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin.

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, CA.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 32–43, Atlanta, GA.

Agirre, Eneko, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393, Montreal.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *5th International Conference on Learning Representations, ICLR 2017*, Conference Track (Poster), Toulon.

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*, Conference Track, San Diego, Ca.

Bau, Anthony, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. *7th International Conference on Learning Representations, ICLR 2019*, Conference Track (Poster), New Orleans, LA.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 861–872, Vancouver.

Blackwood, Graeme, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122, Santa Fe, NM.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor. 2018. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon.

Britz, Denny, Melody Guan, and Minh-Thang Luong. 2017. Efficient attention using a fixed-size memory representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 392–400, Copenhagen.

Callison-Burch, Chris, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors. 2007. *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague.

Cer, Daniel, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the*

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

11th International Workshop on Semantic Evaluation (SemEval-2017), pages 1–14, Vancouver.

Chen, Qian, Zhen-Hua Ling, and Xiaodan Zhu. 2018. Enhancing sentence embedding with generalized pooling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1815–1826, Santa Fe, NM.

Chen, Yun, Yang Liu, Yong Cheng, and Victor O. K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha.

Cífka, Ondřej and Ondřej Bojar. 2018. Are BLEU and meaning representation in opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1362–1371.

Conneau, Alexis and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1699–1704.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen.

Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. *6th International Conference on Learning Representations, ICLR 2018*, Conference Track (Poster), Vancouver.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018c. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels.

Dalvi, Fahim, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 142–151.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN.

Dolan, Bill, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Dou, Zi Yi, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang. 2018. Exploiting deep representations for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4253–4262.

Elliott, D., S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin.

Firat, Orhan, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

attention mechanism. In *Proceedings of NAACL-HLT*, pages 866–875, San Diego, CA.

Firat, Orhan, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, TX.

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1243–1252, Sydney.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487, Miyazaki.

Graves, Alex and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6):602–610.

Gu, Jiatao, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018a. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 344–354, New Orleans, LA.

Gu, Jiatao, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018b. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, LA.

Ha, Thanh Le, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *Proceedings of the International Workshop on Spoken Language Translation, IWSLT 2016*.

Hill, Felix, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, CA.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5(1):339–351.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia.

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015*, Conference Track (Poster), San Diego, CA.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, Vancouver.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, and Alexandra Constantinand Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*, Demo and Poster Sessions.

Kruszewski, Germán, Angeliki Lazaridou, Marco Baroni, et al. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 971–981.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Lakew, Surafel Melaku, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, NM.

Lee, Jason, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Lin, Zhouhan, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017*, Conference Track (Poster).

Lu, Yichao, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels.

Luong, Minh Thang, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *4th International Conference on Learning Representations, ICLR 2016*, Conference Track (Poster), San Juan.

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223, Reykjavik.

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, Long Beach, CA.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Barcelona.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, MI.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVE: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha.

Platanios, Emmanouil Antonios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Doha.

Poliak, Adam, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 513–523, New Orleans, LA.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983, Hong Kong.

Schwenk, Holger. 2018. Filtering and mining parallel data in a joint multilingual space. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234.

Schwenk, Holger and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *ACL Workshop on Representation Learning for NLP*, pages 157–167, Vancouver.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024

Linguistics (Volume 1: Long Papers), volume 1, pages 1715–1725, Berlin.

Shi, Xing, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, TX.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, WA.

Subramanian, Sandeep, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *6th International Conference on Learning Representations, ICLR 2018*, Conference Track (Poster), Vancouver.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Ghahramani, Z., M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 3104–3112.

Tai, Kai Sheng, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1556–1566, Beijing.

Tao, Chongyang, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4418–4424, Stockholm.

Tu, Zhaopeng, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Voorhees, Ellen M. and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, Athens.

Wang, Yining, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960, Brussels.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.

Zoph, Barret and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of NAACL-HLT*, pages 30–34, San Diego, CA.

Downloaded from http://direct.mit.edu/coli/article-pdf/46/2/387/1847640/coli_a_00377.pdf by guest on 10 August 2024