WMT 2016 Multimodal Translation System Description based on Bidirectional Recurrent Neural Networks with Double-Embeddings

Sergio Rodríguez Guasch and Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
sergio.rodriguez.guasch@est.fib.upc.edu,marta.ruiz@upc.edu

Abstract

Bidirectional Recurrent Neural Networks (BiRNNs) have shown outstanding results on sequence-to-sequence learning tasks. This architecture becomes specially interesting for multimodal machine translation task, since BiRNNs can deal with images and text. On most translation systems the same word embedding is fed to both BiRNN units. In this paper, we present several experiments to enhance a baseline sequence-to-sequence system (Elliott et al., 2015), for example, by using double embeddings. These embeddings are trained on the forward and backward direction of the input sequence. Our system is trained, validated and tested on the Multi30K dataset (Elliott et al., 2016) in the context of the WMT 2016 Multimodal Translation Task. The obtained results show that the double-embedding approach performs significantly better than the traditional single-embedding one.

1 Introduction

Sequence-to-sequence learning is a new common approach to translation problems (Sutskever et al., 2014). The basic idea consists in mapping the input sentence into a vector of fixed dimensionality with a Recurrent Neural Network (RNN) and, then, do the reverse step to map the vector to the target sequence. From this new perspective, multimodal translation (Elliott et al., 2015) has become a feasible task. In particular, we are referring to the WMT 2016 multimodal task that consists in translating English sentences into German, given the English sentence itself and the image that it describes. This paper describes our participation in this task using a translation scheme based on Bidirectional RNNs (BiRNNs) which allows to combine both information from image and text.

In this paper, we take as baseline system the one from (Elliott et al., 2015) and focus on experimenting with the word embedding system and encoding techniques.

The rest of the paper is organised as follows. Section 2 briefly describes related work on image captioning and machine translation. Section 3 gives details about the architecture of the multimodal translation system. Section 4 reports details on the experimental framework including the parameters of our model and the results obtained. Finally, Section 5 concludes and comments on further work.

2 Related work

Image captioning has gained interest in the community and deep learning has been applied in this area. The two most common caption-related problems are caption generation (Vinyals et al., 2014) and caption translation (Elliott et al., 2015).

Similarly, machine translation approaches based on neural networks (Sutskever et al., 2014; Cho et al., 2014) are competing with standard phrase-based systems (Koehn et al., 2003). Neural machine translation uses an encoder-decoder structure (Cho et al., 2014). The implementation of an attention-based mechanism (Bahdanau et al., 2015) has allowed to achieve state-of-the-art results. The community is actively investigating in this approach and there have been enhancements related to addressing unknown words (Luong et al., 2015), integrating language modeling (Gülçehre et al., 2015), using character information in addition to words (Costa-jussà and Fonollosa, 2016) or even combining different languages (Firat et al., 2016), among others.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 655–659,
Berlin, Germany, August 11-12, 2016.

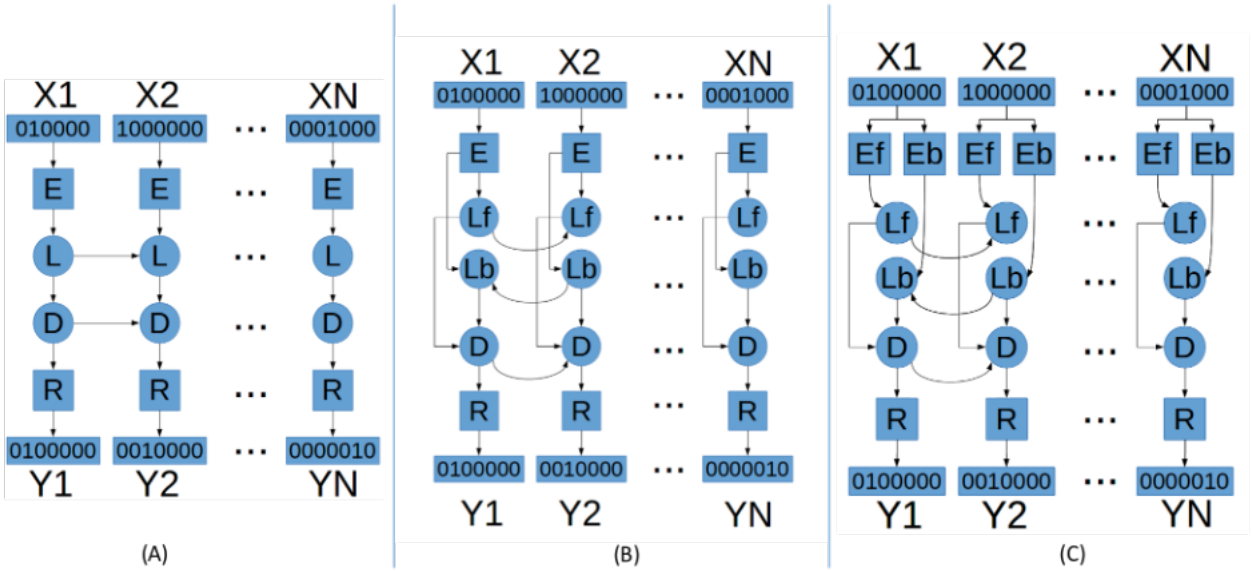

Figure 1: NMT architectures: (A) using unidirectional RNNs, (B) using BiRNNs, (C) adding double embedding.

3 System description

This section describes the main architectures that have been tested to build the final system.

3.1 Baseline approach

The baseline system is a RNN model over word sequences (Elliott et al., 2015), which can use visual and linguistic modalities. The core model is a RNN over word sequences, trained to predict the next word in the sequence, given the sequence so far. The input sequence is codified in *1-of-K* vector, which is embedded into a high-dimensional vector. Then, a unidirectional RNN is used. Finally, in the output layer, the softmax function is used to predict the next word. This model is extended to a multimodal language model, where sequence generation in addition to be conditioned on the previously seen words, are conditioned on image features. The translation model simply adds features from the source language model, following work from (Sutskever et al., 2014; Cho et al., 2014) and calling the source language model the *encoder* and the target language model the *decoder*.

3.2 Sequence-to-sequence approach and enhancements

Inspired by the architecture presented in (Sutskever et al., 2014), we train a system based on the many-to-many encoder-decoder architecture. It accepts a sequence $x_1, .., x_N$ as input and returns a sequence $y_1, .., y_N$, where N is the maximum sequence length allowed.

The architectures that we have tested start in a unidirectional encoder-decoder, then we use a bidirectional encoder-decoder, a bidirectional encoder-decoder with double embeddings, and a final architecture that accepts a combination of input text and image. See Figure 1 (A), (B) and (C) and Figure 3.2 (D) for an schematic representation of these architectures.

Architecture (A) The model receives as input the codifications *1-of-K* of the source sequence $x_1...x_n$, then the word embedding is computed, obtaining a new representation $E(x_1)...E(x_n)$. This new sequence is processed by a RNN L, obtaining the vectors $L_1...L_n$. These vectors are processed by another RNN D, obtaining the sequence $D_1...D_n$, which is processed by a conventional neural network obtaining the target vectors which are normalised using $softmax$.

Architecture (B) The main difference is that we are using BiRNNs, processing the input sentence forward and backward. The BiRNN is implemented with LSTMs (Long Short Term Memories) for better long-term dependencies handling (Hochreiter and Schmidhuber, 1997; Chung et al., 2014). The BiRNN are represented by unit L, but in this case, one in each direction, generating two vectors Lf_i and Lb_i, corresponding to each input x_i.

Architecture (C) In addition to using BiRNNs, each input codification is processed by two different feed-forward neural networks E_f and E_b, generating two vectors $E_f(x_1)...E_f(x_n)$ and $E_b(x_1)...E_b(x_n)$ of size H, where H is a constant. At each timestep the pair of vectors are fed to the BiRNN Lf and Lb.

Architecture (D) Finally, the last architecture proposes to introduce an image. See Figure 3.2. This is the main advantage of using a machine translation system based on neural networks: we can use multimodal inputs. In this case, image and text. The model in this case has two inputs: the input text sequence $x_1...x_n$ and the image vector, which is the result of intermediate layers of a pre-trained convolutional neural network (Simonyan and Zisserman, 2014).

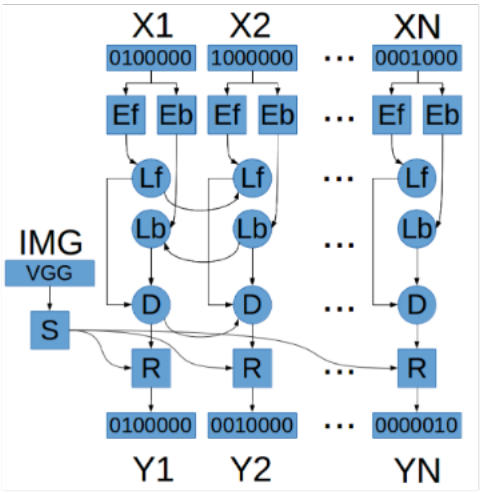

Figure 2: Diagram of NMT architecture (D) using image and text.

4 Experiments and results

4.1 Data

The system is developed, trained and tested with the Multi30K dataset provided by the WMT organization. On our experiments, all characters are converted to lower case. The chosen vocabulary consists on all the training source words and all the training target words that appear more than once. This choice is made to minimise the number of unknown tokens at the source sentences and to avoid an excessive model size and training time.

4.2 Model training

Each source sentence is encoded onto a $N \times V$ matrix M, where each row represents a *1-of-K* encoding of a word over a source vocabulary with V words. An unknown word is replaced by an special <U> token and a <E> token is appended at the end of the sequence. If the sequence length (including <E>) is less than N the remaining rows will be zeros. If the sequence is too long, then it is truncated in order to suit the input size restrictions. During the training phase, target sentences also have a <B> token before the first word. For a given example, the generated prediction is considered to be all the words generated between the <B> and <E> tokens. Unknown tokens are replaced by the second highest probability word.

Parameter	Description	Value
N	Maximum sequence length	45
V	Source vocabulary words	10364
T	Target vocabulary words	8012
H	Embedding size	512
DROP	Dropout rate	0.25
L2	L2 regularizer	10^{-8}

Table 1: Model parameters value

Dropout rate of 0.25 is applied to all non-recurrent units and a L2 regularization is applied to all weights and units.

Training is performed on batches of size 10000 and on mini-batches of size 128. The target metric is the categorical cross entropy and the used optimiser is Adam (Kingma and Ba, 2014). Results are validated at each epoch on the dataset validation split using the BLEU metric (Papineni et al., 2002), along with model perplexity.

BLEU scores during validation are also used as an early stop criteria in case the maximum score so-far is not surpassed on the following 10 epochs. In order to evaluate our system performance obtained results are compared against a single-embedding system trained under the same conditions and parameters. Their BLEU score monitorization can be observed in Figure 3 and the chosen parameter set is summarised in Table 1.

4.3 Results

Table 2 shows the BLEU and METEOR (Lavie and Denkowski, 2009) results for the main architectures described in section 3 for the official test set of the WMT 2016 Multimodal Translation

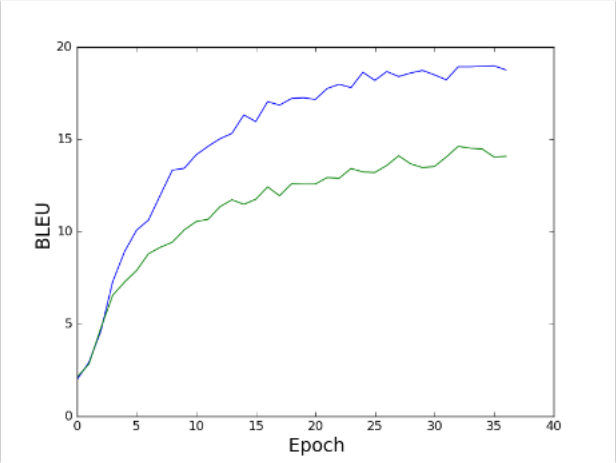

Figure 3: Evolution of BLEU scores (y-axis) on the validation split for the double-embedding system (top blue line) and the single-embedding one (bottom green line).

Task 1. Baseline results are kindly provided by the organisers, referred in the evaluation official results as *1_GroundedTranslation_C*.

We see that using BiRNNs improve vs RNNs, and double-embeddings improves over single-embeddings. Finally, adding the image information does not improve results. Therefore, the best architecture (C) is the one that participated in WMT 2016 Multimodal Translation Task. Official results ranked our system in the 14th position out of 16. We priorised participating with a pure multimodal extensible architecture. However, we know it would have improved our ranking just performing a simple technique as rescoring our system with a standard Moses (Koehn et al., 2007).

System	BLEU	METEOR
Baseline	9.41	24.71
Architecture (A)	19.16	34.23
Architecture (B)	20.89	35.97
Architecture (C)	**22.74**	**37.68**
Architecture (D)	17.74	32.39

Table 2: BLEU and METEOR Results. Official baseline *1_GroundedTranslation_C* kindly provided by the organisers.

The best architecture (C) (compared to using one embedding) is capable of solving problems like unknown words or chosing the appropriate word. Table 3 shows an example that shows the word fixation problem.

However, our generated translations have often many repeated words or end prematurely, mainly due to the differences in lengths and alignments between source and target sentences and the lack of feedback from previous timesteps. In any case, our system is still capable to generate readable translations and to replace unknown words with similar ones.

Source	a man sleeping in a green room on a couch
Generated	ein mann schlaft in einem grünen grünen auf einem sofa
Reference	ein mann schlaft in einem grünen raum auf einem sofa

Table 3: An example that shows the word fixation problem

Also, our system performance drastically decreases on long sentences, or on sentences where the length of the source and target sentences differ too much.

5 Conclusions

Our system is not competitive compared to standard phrase-based system (Koehn et al., 2003) or the auto-encoder neural machine translation system (Bahdanau et al., 2015) as shown by our ranking in the official evaluation (14 position out of 16). However, the architecture of our system makes it feasible to introduce image information. Maybe in a larger corpus we would get competitive results.

All software is freely available in github[1].

The main contribution of this paper is that we show that double embeddings (trained on forward and backward input sequence) provides a significant improvement over single embeddings.

As further work, we are considering experimenting towards replacing the word based encoder for a character-based embedding (Costa-jussà and Fonollosa, 2016), or to introduce attention-based decoders (Bahdanau et al., 2014). Due to the system's modularity, it is also possible to reuse intermediate outputs to train additional models. For example, it is possible to extract the BiRNN intermediate outputs and fed them to another decoder model, thus reducing training time.

[1] https://github.com/srgrr/Neural-Translation

Acknowledgements

This work is supported by the 7th Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951) and also by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, contract TEC2015-69266-P (MINECO/FEDER, UE).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

Dimitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR, abs/1406.1078.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555.

Marta R. Costa-jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. In Proc. of the ACL.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. CoRR, abs/1510.04709.

Desmond Elliott, Stella Frank, Khalid Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. CoRR, abs/1605.00459.

Orhan Firat, KyungHyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. CoRR, abs/1601.01073.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. CoRR, abs/1503.03535.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. CoRR, abs/1412.6980.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In Proceedings of the ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicolas Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proc. of the 45th Annual Meeting of the Association for Computational Linguistics, pages 177–180.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. Machine Translation, 23(2-3):105–115, September.

Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pages 11–19.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. CoRR, abs/1409.3215.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. CoRR, abs/1411.4555.