Concept Equalization to Guide Correct Training of Neural Machine Translation

Kangil Kim
Konkuk University
Republic of Korea

Jong-Hun Shin
ETRI
Republic of Korea

Seung-Hoon Na*
Chonbuk National University
Republic of Korea

SangKeun Jung
SK Telecom
Republic of Korea

Abstract

Neural machine translation decoders are usually conditional language models to sequentially generate words for target sentences. This approach is limited to find the best word composition and requires help of explicit methods as beam search. To help learning correct compositional mechanisms in NMTs, we propose concept equalization using direct mapping distributed representations of source and target sentences. In a translation experiment from English to French, the concept equalization significantly improved translation quality by 3.00 BLEU points compared to a state-of-the-art NMT model.

1 Introduction with Related Works

After the possibility of learning end-to-end translation model (Sutskever et al., 2014) was reported, there has been a surge of research to apply recurrent neural networks (RNN) with long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) to machine translation. After intensive developments, this approach becomes the state-of-the-art of machine translation called as neural machine translation (NMT). Remarkable approaches are bidirectional LSTM using both forward and backward sequences (Sutskever et al., 2014), attention model to learn explicit alignment models (Bahdanau et al., 2014; Luong et al., 2015a), rare word modeling to estimate unknown word through alignment information (Luong et al., 2014). Many other detailed following techniques improved the performance such as batch normalization (Ioffe and Szegedy, 2015), ensembles, beam search, input feature specialization, and input feeding, which are all aggregated into Google's NMT report (Wu et al., 2016).

Most decoders of NMTs are conditional language models, which sequentially generate target words and its proceeding correct target words in the condition of a given source sentence. This approach is a greedy algorithm, so dependency of selected words to subsequent words may restrict selecting the best target word composition. Beam search is a promising method to approximate the correct compositions. However, inversely, promising results imply that NMTs are still weak to learn the dependency between words in a target sentence. This limitation in training process creates fundamental barrier of representing correct translation process in a neural network. In (Ranzato et al., 2015), this issue was discussed as a problem of maximum likelihood estimation ignoring the dependency between selected target words and an approach penalizing the likelihood with sentence-level distance has been proposed (Shen et al., 2015).

Beyond correct target word generation, translation may be regarded as a subproblem to find mapping of sentence-level semantics between two languages. Accuracy of this mapping is often limited because of ambiguity caused by many-to-many mapping relations. It is difficult to find exact mapping with simple and direct mapping models as the reported difficult in mapping simpler word-level semantics (AP et al., 2014; Luong et al., 2015b; Upadhyay et al., 2016).The framework of NMTs is the most successful model to find an exact mapping of sentence-level semantics so far. However, its huge expression power leads to inefficiency in training which requires addition connections to transfer related information such as attention or input feeding models (Bahdanau et al., 2014; Wu et al., 2016). This inefficiency may be removed by using simple di-

*corresponding author: nash@jbnu.ac.kr

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 302–307,
Taipei, Taiwan, November 27 – December 1, 2017

rect models to restrict unnecessary area to search in training step.

In this paper, we propose *concept equalization* method to apply the direct semantic-mapping model to existing NMT frameworks for guiding training of model parameters. Distinguished contributions of this method are to introduce 1) an effective penalty function and 2) a plug-in framework to transfer the constraint information of the penalty to LSTM stacks. In practical translation tasks from English to French, this method improves translation quality and convergence speed to local optima. Extra benefit is easy adaptation to any type of NMT frameworks.

Paper structure is as follows. Section 2 explains motivations and details of concept equalization. Section 3 shows experimental configurations on data, model, and runs and Section 4 interprets the results. Section 5 is conclusion and future work.

2 Concept Equalization

2.1 Limit of Learning Target Composition

Beam search is a promising method for NMTs by overcoming the problem of greedy search in sequential target word generation. On the other hand, the impact of beam search inversely implies that the sequential decisions by the model are likely to be incorrect to select the best sentences in many cases. There are many possible causes for the inaccurate prediction of composition of target words such as inaccurate model representation, complex parameter landscapes, and noise data.

A possible cause is the simple representation of the correctness of target words. In current NMTs, cross-entropy is the most popular cost function composed of probabilities of selecting each correct word of a target sequence. Therefore, only one variable is responsible for representing whether the selected target word is correct. Using only one variable may be risky because the second probable word and its highly probable following sequences may give higher cross-entropy than any sequences derived from the correct word selection. This case is a deceptive example of restricting accurate word composition in decoders.

Another cause is slow parameter update in NMT structures. In LSTMs, the gradient vanishing (Bengio et al., 1994) over time steps is resolved by using memory cells and over vertical structures by addition input feeding or multidimensional memories (Wu et al., 2016; Kalchbrenner et al., 2015). They are applied to the encoder and decoder, but the interface part is often a feedforward layer suffering from the gradient vanishing. This vanishing limits the achievable translation quality in general and may restrict learning the correct composition.

2.2 Motivation: Concept Equalization

To resolve the limited target word composition, two approaches are proposed: 1) direct linking the interface vector to the cost function and 2) increasing the dimension for representing correctness of target words. The first approach defines a cost function to directly use the transferred interface vector, so that the depth from the cost the interface vector decreases and reduce the effect of gradient vanishing. In the second approach, if we train NMT to select a correct word only if its all variables are the most probable, then the deception of selecting the second probable word in one perspective is easily excluded. This second approach is somewhat new in NMT literature, because the dimension of probability vectors is already so high that increasing the dimension becomes serious burden.

To use the two approaches without problems, we introduce a *concept equalization* for training NMT where the concept indicates the semantics of source and target sentences and represented as two vectors. A expected role of this approach is to guide NMTs not to train obviously wrong sentences in explicit direct mapping models. The method is illustrated in Fig. 1 and described in the following sections.

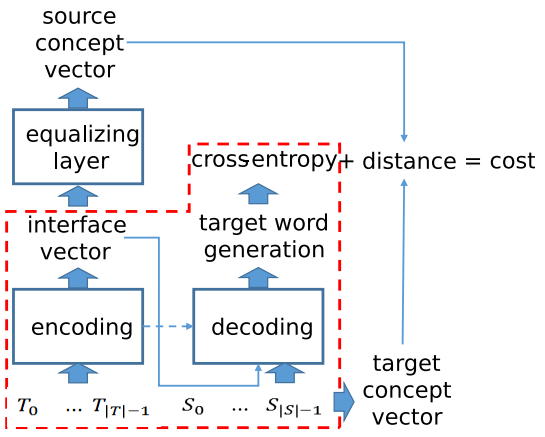

Figure 1: Concept Equalization Model Plugged In to Typical Neural Machine Translation (red and dashed line: typical model)

2.3 Concept Equalization Model

We newly propose the following three parts compared to typical NMT frameworks.

Raw Concept Vector Representation Representing a target word as a vector, concept equalization assigns many variables to indicate the correctness of target word without increasing the dimension of existing probability vectors for calculating cross-entropy. The concept vector $\mathbf{c}_S$ of source sentence S and $\mathbf{c}_T$ of target sentence T are defined as

$$\mathbf{c}_S = \sum_{t=1}^{|S|} \mathbf{h}_t \quad , \quad \mathbf{c}_T = \sum_{t=1}^{|T|} \mathbf{w}_t$$

where $\mathbf{h}_t$ is the hidden vector generated from the top LSTM stacks at time step t in a NMT encoder and $\mathbf{w}_t$ is the word vector at time t in a decoder. This raw vector generation process can be easily extended to existing NMTs. For example, in bidirectional models, $\mathbf{h}_t$ is replaced by $\mathbf{h}_t^f \| \mathbf{h}_t^b$. In bidirectional attention models, interface vectors are transformed vectors of $\mathbf{h}_t$ with alignment model and target word, but we can still use $\mathbf{h}_t^f \| \mathbf{h}_t^b$.

Equalizing Layer In equalization, the biggest risk is the conflict of vector distribution of $\mathbf{h}_t$ desirable for minimizing cross-entropy and for concept distance because of quality decrease of local optima. From the definition, many sentences can be mapped to a concept vector in both sides, which increases ambiguity and their average distances as well. If the average is relatively larger than cross-entropy, distance-cost dominates the parameter updates. If this phenomenon is maintained near optimal, the converged model will be far from the true optimal. Pros of this mapping is restricting generation of undesirable sentences using many variables and potential cons is the accuracy decrease by the conflict. To reduce the negative, we use one more linear combination layer for more flexible mapping, which will reduce average distance between the concepts.

$$\mathbf{v}_S = \mathbf{W}_e \mathbf{c}_S + \mathbf{b}_e \quad , \quad \mathbf{v}_T = \mathbf{c}_T \qquad (1)$$

The equalizing layer is composed of parameters $\mathbf{W}_e$ and $\mathbf{b}_e$ and the concept vector $\mathbf{v}_S$ is the output vector of the layer from the given raw concept vector $\mathbf{r}_S$.

Cost Function In sentence-level translation, underlying assumption is that the semantics of matching sentences are so equal that any distributed representation of two sentences in a vector space should be equal. In the assumption, reducing the distance of the concepts is a perfect goal for maintaining the same true optimal of NMT and therefore we can directly use it as a cost function. To use it, we set a cost function as following equation.

$$\text{new cost}(\theta, D) = \text{cost}(\theta, D) + ||\mathbf{v}_S - \mathbf{v}_T||_2 \quad (2)$$

which uses Euclidean distance of the concept vectors as a penalty. θ is a parameter set to represent a model and D is a given training data set. This method adds cost and distance without any scaling factors because the equalizing layer implicitly adapts its scale in updates. In early stages of the updates, the layer gives large distance for all vectors by random initialization but the large distance dominates updates and makes NMTs rapidly converge to a model to generate small distance over all vectors. Then, the impact of cross-entropy increases and the model moves to the true optimal determined by the entropy. Therefore, if the optimal distance is sufficiently small, then this method will guide the training in early updates and preserve the true optimal with respect to cross entropy with restriction of generating negative sentences.

3 Experiment Setting

We performed translation experiments from English to French to evaluate the impact of concept equalization in a state-of-the-art NMT.

3.1 Data Preparation

We used WMT14 Europarl parallel corpus for training [1] and applied tokenizing, lowercasing, and limiting token numbers by 40 in a sentence through using scripts provided by a machine translation package, MOSES (Koehn et al., 2007) [2]. Starting and ending symbol are attached to each source sentence. Test set is the first 10,000 sentences of the news-commentary set released with the Europarl corpus. Data statistics are shown in Table 1. We extracted word vectors from the training set using a neural network language model implemented in word2vec [3] for English and French. Extracted word vector dictionaries are imported in training phases.

[1] http://www.statmt.org/wmt14/
[2] http://www.statmt.org/moses/
[3] https://code.google.com/archive/p/word2vec/

Table 1: Data Statistics

	training		test	
	En	Fr	En	Fr
tokens	30.7M	34.2M	0.2M	0.2M
sentences	1.5M		10,000	

3.2 Neural Network Structure

Because of the lack of space, we drop full mathematical description of our model. We built bidirection model and passed the **h** and **c** from the forward to backward pass of the encoder. Then the $\mathbf{h}_t$ of forward and $\mathbf{h}_{|S|-t}$ are concatenated to derive $\mathbf{r}_s$. Attention model is equal to (Bahdanau et al., 2014) except additionally passing **c** for initialization of the decoder. The input word vectors are fed on to the second shallowest LSTM stack of the encoder. To boost converging speed, we applied batch normalization through weighted average of original and normalized vectors. The weight is decayed by multiplying 0.8 at each epoch, which becomes almost 0 after 16 epochs. The concept equalization is only applied to the training phase. Sample phase is equal to typical NMTs.

Table 2: Detailed Model and Run Settings

LSTM stacks	4	parameter	
cells per stacks	250	encoder	3.05M
dim. of word	50	decoder	3.10M
dim. of attention	250	output	11M
dim. of equalizer	50	interface	0.19M
batch size	128	epochs	50

4 Results and Discussions

We evaluated BLEU and NIST (Papineni et al., 2002; Doddington, 2002) score as shown in Table 3. In the table, we can confirm that applying

Table 3: Translation Quality of NMT with and without Applying Concept Equalization

	equalized NMT		NMT	
epoch	NIST	BLEU	NIST	BLEU
11	5.7682	22.78	5.3333	20.89
14	6.2776	25.48	5.6247	21.70
17	6.2786	25.31	5.9214	23.98
30	6.5775	26.68	5.9941	23.86
38	6.6466	27.08	6.0282	24.08

concept equalization improves translation quality by 3 BLEU and 0.6186 NIST after 38 epochs compared to a typical state-of-the-art NMT. The baseline BLEU is 24.08 similar to or smaller than the results of (Bahdanau et al., 2014), but the used training sentences are 8 million in the paper while we used 1.5 million, which may be the reason of the baseline gape.

Fig. 2 shows the training accuracy change by epochs. The accuracy is the portion of correctly selected words in the training set through applying argmax to probability vectors generated by the decoder. Applying the concept equalization shows faster convergence before 5 epochs, but converges to lower optimal points after 50 epochs compared to the normal NMT. From this result,

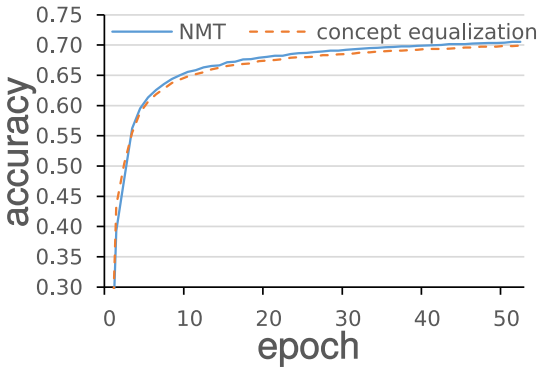

Figure 2: Training Accuracy by Epoch (accuracy: correctly selected token rate in training)

we can confirm the equalization boost the speed of convergence, but the conflict limits the accuracy near optimals. This is consistent behavior to the properties of the method. A notable point is that the translation quality is significantly improved. We guess the cause is the different underlying assumptions in decoding. In the case of training accuracy, the decoder assumes receiving correct input at every time step while it uses previously selected word in actual decoding. Therefore, in actual decoding to generate translation for the test set, a selected wrong word in an intermediate step may cause subsequent errors by strange context generation. Overall, the two results imply that the equalization guides the model to be robust to the unseen errors by restricting generation of strange sequences without any regularization techniques as randomized distortions on some parts of models.

5 Conclusion

In this paper, we raised the issue of limit in learning correct target word composition of current

NMTs. To resolve it, we introduced concept equalization to learn direct mapping of source and target sentences for guiding the NMT training. In the result, translation quality is significantly improved and training speed becomes slightly faster in early epochs. This method is expected to effectively discard wrong target composition from the observations.

6 Future Work

We will generalize this work for various direct mapping models and wider empirical tasks. Impact of concept equalization to cost landscape in parameter optimization will be more rigorously analyzed.

Acknowledgement

This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement) and the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-2016-0-00465) supervised by the IITP(Institute for Information & communications Technology Promotion).

References

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Incorporation.

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

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456.

Nal Kalchbrenner, Ivo Danihelka, and Alex Graves. 2015. Grid long short-term memory. *arXiv preprint arXiv:1507.01526*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2014. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015b. Bilingual word representations with monolingual quality in mind. In *VS@ HLT-NAACL*, pages 151–159.

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*, pages 311–318. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.