

NMT Enhancement based on Knowledge Graph Mining with Pre-trained Language Model

Hao Yang
Translate Service Center
Huawei Co. Ltd.
Beijing, China
yanghao30@huawei.com

Yao Deng
Translate Service Center
Huawei Co. Ltd.
Beijing, China
dengyao@huawei.com

Ying Qin
Translate Service Center
Huawei Co. Ltd.
Dongguan, China
qinying@huawei.com

Minghan Wang
Translate Service Center
Huawei Co. Ltd.
Beijing, China
wangminghan@huawei.com

Abstract —Pre-trained language models like Bert, RoBERTa, GPT, etc. have achieved SOTA effects on multiple NLP tasks (e.g. sentiment classification, information extraction, event extraction, etc.). We propose a simple method based on knowledge graph to improve the quality of machine translation. First, we propose a multi-task learning model that learns subjects, objects, and predicates at the same time. Second, we treat different predicates as different fields, and improve the recognition ability of NMT models in different fields through classification labels. Finally, beam search combined with L2R, R2L rearranges results through entities. Based on the CWMT2018 experimental data, using the predicate's domain classification identifier, the BLEU score increased from 33.58% to 37.63%, and through L2R, R2L rearrangement, the BLEU score increased to 39.25%, overall improvement is more than 5 percentage

Keywords —NMT, Pre-trained Language Model, Knowledge Graph

I. INTRODUCTION

Neural Machine Translation (NMT for short) is a neural network models to learn machine translation. The most famous models of NMT are Encoder-Decoder model [1] and Transformer [2]. In the Encoder-Decoder model, the encoder neural network reads the source sentence and encodes it into a fixed-length vector. The decoder predicts output with the fixed-length vector. The entire encoder-decoder system (including encoders and decoders for language pairs) is jointly trained to maximize the likelihood of correct translation given a source sentence. Although effective, the encoder-decoder architecture suffers from long translation sequences, and the performance of Transformer is much better. The problem stems from the fixed-length internal representation that must be used to decode every word in the output sequence. The solution is to use an attention mechanism that allows the model to learn how to focus on the input sequence when decoding each word of the output sequence, the name "Attention is what you need" or "Transformer".

Knowledge Mining or Knowledge Graph Mining[3] is very useful to promote NMT performance with two aspects. The

correct identification of subject and object can effectively reduce the error of NER. The effective identification of the subject-object relationship can avoid domain or common sense errors and significantly improve the effect of machine translation.

NLP pre-training model, such as Bert[4], RoBERTa[5], GPT, etc., has been shown to be effective for improving many natural language processing tasks. BERT, which stands for Bidirectional Encoder Representations from Transformers, is designed to pre-train deep bidirectional representations from self-supervised text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained model create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications., including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% point absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement)[6][7].

Considering Pretrained Language Model(PLM)[8][9] shows good performance in text classification and sequence annotation, we propose a simple way to improve the quality of machine translation using PLM. Firstly, we can learn a multi-task learning

SPO triple (Subject-Predicate-Object) model[10][11], which distills subject/object entities, and predicate classification; Secondly, Predicates are used as a domain tag, directly fine-tunes model to fit special domain. Finally, L2R and R2L reranking is provided based on by subject/object list with pairwise Learning To Rank (LTR). Experiments show Bert based knowledge distilling model can improve NMT performance effectively.

The structure of this paper is as follows. The second part introduces the knowledge extraction model based on PLM. The third part introduces Transformer based on Predicate Classification and domain fine-tuning, and bi-direction reranking methods based on Subject/Object in the fourth part. The fifth part analyzes the experiments domain classification and LTR Re-ranking performance. Final part is summarization and future research.

II. PLM-BASED KNOWLEDGE EXTRACTION

The knowledge extraction of aligned corpus should generally include SPO (Subject-Predicate-Object) list extraction in two languages.

2.1 PLM-based classification and sequence Annotation model

By following a PLM model with a Dense layer (complete connection layer) and specifying the parameters of the Dense layer as the number of classification categories, the PLM classification model can be easily obtained. The PLM annotation model can be a PLM connected to the LSTM and CRF layers, and can be annotated and trained based on serialized BIO annotations.

However, for the CRF layer, the loss calculation is a bitwise multi-class calculation. In fact, consider using two pointers to point to the beginning of the entity and the end of the entity, respectively. The bitwise multi-classification problem becomes



Pic 1: PLM based Information Extraction Model

two-pointer based on two Classification issues, efficiency and accuracy improved.

2.2 PLM MULTITASK LEARNING FOR KNOWLEDGE MINING

Similar to the probability map idea of the NMT classic seq2seq[12], the actual modeling function of the decoder that generates the sequence is as follows:

$$P(y_1, y_2, \dots, y_n) = P(y_2|y_1)P(y_3|y_1, y_2) \dots P(y_n|y_1, y_2, \dots, y_{n-1}) \quad (1)$$

The SPO triplet can be identified by first identifying the Subject based on Sentence, then identifying the Object based on the Subject, and finally classifying the Predicate by the Subject/Object pair (Formula 2).

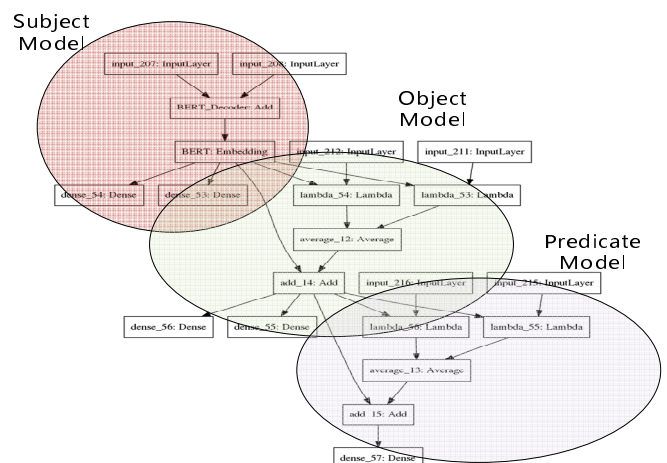
$$P(\text{Subject}, \text{Predicate}, \text{Object} | \text{Sentence}) \\ = P(\text{Subject} | \text{Sentence}) P(\text{Object} | \text{Subject}, \text{Sentence}) \\ P(\text{Predicate} | S, O, \text{Sentence}) \quad (2)$$

Therefore, the mining of triples can be made into a multi-task training problem, including two sequence annotation problems (Subject annotation, Object annotation), and a text classification problem (Predicate classification), as shown in Figure 1.

2.3.1 SPO models Design

Traditional sequence labeling tasks, such as NER identification, are based on BIO entity annotation. CRF or BILSTM+CRF mode is used to classify SOFTMAX word by word. Each TOKEN is determined to be entity start (B_TAG), entity middle (I_TAG) and none Entity Object.[13]

The NER model of verbatim SOFTMAX, considering two constraints, (1) the sum of BIO labeling probabilities of a word is 1, (2) the front of `I_TAG` can only be `B_TAG`, not `O_TAG`[14], ie The first word is `I_TAG`, then the label of the first -1 word must be `B_TAG` or `I_TAG`, see Equation 3 and Equation 4.



Pic 2: PLM Based Information Extraction Model

$$\forall i: \sum_{y \in Y} Z_i^y = 1 \quad (3)$$

$$\forall i, \forall X: Z_{i-1}^{Bx} + Z_{i-1}^{Lx} - Z_i^{Lx} \geq 0 \quad (4)$$

Based on the above two constraints, we can change the label of the verbatim CRF to two two-category labels with only 0,1. The first label, the start bit of the entity is identified as 1, and the other identifiers are 0. For the label, the next bit of the entity end bit is 1, and the others are all 0, that is, the entity and object labels in Figure 2. The advantage of the two 01 annotations is that subsequent models only need PLM + 2 DENSE models to achieve physical or object annotation[15][16].

We build a PLM SPO extraction knowledge extraction model, as shown in Figure 2, including three sub-models:

(1) The upper left corner is the main model. Based on the two inputs of PLM encoding, two DENSE classifiers are combined to predict the start and end positions of the object.

(2) In the middle is the object model, which predicts the start and end positions of the object based on the original sentence encoding and the start and end positions of the topic.

(3) The lower right is the predicate model. The predicate classification is based on the original sentence encoding, the start and end positions of the topic, and the start and end positions of the object.

The PLM-based <Subject, Predicate, Object> model is built as table 1:

	Model	Input	Output
Subject	PLM-Annotation with two Pointers	Sentence	Subject Pointers
Object	PLM-Annotation with two Pointers	(Sentence, Subject Pointers)	Object Pointers
Predicate	PLM-CLS with	(Sentence, Subject Pointers, Object Pointers)	Predicate Classifications

Table 1: PLM-based SPO Distilling model

2.3.2 LOSS FUNCTION DESIGN

We define the loss function as the sum of three losses, including the subject loss function, the object loss function, and the predicate loss function. Topic loss is two binary cross entropy (BCE), while object loss is two binary cross entropy based on topic prediction, and predicate LOSS is cross entropy (CE) multi-task training based on topic and object prediction. For knowledge extraction of the LOSS function, see Equation 5.

$$\begin{aligned} \text{Loss} &= \text{Loss}_{\text{subject}} + \text{Loss}_{\text{object}} + \text{Loss}_{\text{predicate}} \\ &= \arg\max(\text{BCE}(\text{sub}, \text{sub}_p) \\ &\quad + \text{BCE}(\text{obj}, \text{obj}_p) | \text{sub}_p) \\ &\quad + \text{CE}(\text{predict}, \text{predict}_p) | (\text{sub}_p, \text{obj}_p) \end{aligned} \quad (5)$$

A piece of text may include multiple subjects. One subject may correspond to multiple objects, and a pair of subject objects may

correspond to multiple categories. Therefore, the starting position sequence of the subject may have multiple 1 identifiers, based on the first one, continuing to After that, find the first one corresponding to the end position sequence, which is the first entity. In the same way, multiple objects can be found. Because the object model itself is based on the conditional probability of the subject model, the PLM knowledge extraction model is an architecture that can support multiple subjects and correspondingly find multiple objects and multiple relationships. The overall architecture diagram is shown in Figure 4.

III. NMT OPTIMIZATION BASED ON KNOWLEDGE TRIPLES

Empirical NMT optimization includes: (1) model feature optimization (2) domain classification optimization (3) ranking optimization. According to our previous analysis, the two-way encoding mechanism of PLM is not suitable for NMT, so parameter-based optimization is not feasible. We mainly consider domain-based optimization and bleu-based paired reordering.

First, NMT optimization with predicate classification is provided. Introducing predicate classification into NMT is actually very simple, just take the predicate from the source sentence and add the predicate in front of the target sentence during training. When making predictions, just delete the first word of the new target sentence.

Second, NMT reordering is based on topics / objects. Learn ranking methods, including point-by-point method, pairwise method, and list method.

Pointwise is a method for single document, extracting features, performing single document weight regression analysis, obtaining absolute correlation of single document, and sorting documents based on absolute relative score. Typical algorithms include Pranking, OAP-BPM, etc. The Pairwise method is based on the relevance of the document, and the functionality of the document is reduced before and after. If the former is higher than the latter, it is marked as 1. If the latter is higher than the former, it is marked as 0, and then a second classification training is performed. Therefore, a weight of the feature is obtained. Typical algorithms include SVMRank. Listwise is to ensure the entire set of ordered training by fitting the entire set of feature weights. Typical algorithms include Lambda MART. Because paired models are simple, easy to understand, and used in the industry.

Because PLM has strong text correlation analysis ability, building a PLM-based Pairwise LTR only needs to connect Subject/Object with Sentence1 and Sentence2. If the former BLEU is higher than the latter, it is marked as 1; otherwise, it is marked

as 0. LTR The sorting problem becomes the PLM's 2 classification problem.

IV. EXPERIMENT ANALYSIS

We conducted experiments based on CWMT18 data, including:

- (1) GPU configuration: Titan XP 12G GPU * 4
- (2) System environment: Python3.6, CUDA-10.0
- (3) Deep learning environment: Tensorflow-gpu 1.11.0, Keras 2.2.4, keras_bert 0.66.0, Pytorch 1.0.0
- (4) NMT Tool: Fairseq Transformer Big,

Because CWMT18's ZH-EN predication is News Corpus, and CCKS2019's Information Extraction is also News domain, we choose ZH-EN task as our experiment. Total data volume: (1) CWMT corpus, remove non-Chinese and English, remove duplicate and alignment corpus length is too far apart (Bilingual word length is greater than 10, and the word length ratio is greater than 3), cleaning about 6.5 million. (2) UN corpus, cleaning with CWMT, about 20 million.

4.1 INFORMATION EXTRACTION SCHEMA

The CCKS 2019 Knowledge Extraction Schema includes 50 articles, of which 49 are independent Predicate, such as "Capital" (首都), "Registered Capital" (注册资本), "Director" (导演) and so on. Independent Subject 16 such as "Attraction" (景点), "Country", (国家) "Enterprise" (企业), etc. Independent Object 16 includes two major types, such as "Website" (网站), "Publisher" (出版社), "Location" (地址), etc., and also includes attribute values such as "Number", "Text", "Date".

Schema examples are as follows

{"object_type": "国家", "predicate": "国籍", "subject_type": "人物": 某某人物的国籍是...

{"object_type": "Country", "predicate": "Nationality", "subject_type": "Person": Some person's country is ...

{"object_type": "Number", "predicate": "海拔", "subject_type": "地点": 某某地点的海拔是...

{"object_type": "Number", "predicate": "Elevation", "subject_type": "Location": The elevation of a certain place is...

4.2 INFORMATION EXTRACTION SCORE

The accuracy and training duration of the SPO triple are listed in Table 2. The accuracy of the PLM-based MTL can reach 82%, and the accuracy of the predicate classification is higher than 90%, but the training duration is much higher than that of the embedded Embedding and Word2Vec Embedding, increased from 12 seconds to 140 seconds, as in Table 3.

MTL	F1	Predicate F1	Entity F1
bare embedding + CNN + CRF	0.5275	0.6569	0.6805
Word2vec + CNN + CRF	0.5042	0.6686	0.7341
BERT	0.8212	0.9043	0.7633

Table 2: Information Extraction F1 Score

MTL	Training time
bare embedding	12s
Word2vec	15s
BERT	140s

Table 3: Training Time for each Epoch

Example

SPO results are like below:

(1) Sentence ZH: “此后不久，戈登·布朗便受到了女王陛下的单独接见，被正式确认为英国新一任首相。”

(2) Sentence EN: “Shortly thereafter, Gordon Brown was summoned for a private audience with the Queen to be confirmed as Britain's new prime minister.”

(3) SPO ZH: {"object_type": "国家", "object": "英国", "predicate": "首相", "subject_type": "人物", "subject": "戈登·布朗"}

(4) SPO EN: {"object_type": "Country", "object": "Britain", "predicate": "PrimeMinister", "subject_type": "Person", "subject": "Gordon Brown"}

4.3 NMT TRAINING WITH PREDICATE CLASSIFICATION TAGS

NMT training with classification labels consists of six forms, namely:

- (1) Unlabeled, standard Transformer model
- (2) Unlabeled, Transformer Big model
- (3) UN/CWMT tag
- (4) Predicate tag
- (5) UN/CWMT + Predicate label, 2 tags

After 10 epoches and 30 epoches of training, the 6 types of BLEU values are listed in table 4, UN/CWMT + Predicate model gets SOTA results.

domain	BLEU (10 epochs)	BLEU(30 epochs)	
Transformer	29.8	30.4	-3.18
Transformer-big(Baseline)	33.13	33.58	-
UN/CWMT	33.85	36.27	+2.69
Predicate	33.5	35.85	+2.37
UN/CWMT + Predicate	34.1	37.63	+4.05
UN/CWMT + Pre1 + pre2	33.83	36.22	+2.64

Table 4: Multi-tags for NMT Promotion

4.5 SUBJECT/OBJECT BASED LTR TRAINING

In the LTR experiment, considering the PLM performance reasons, and combining the previous Reranking SOTA, we only select the top 10 translation candidate sets for reranking. The three strategies are used to do the experiment, including:

- (1) Standard BeamSearch, as the base line, (2) SVM Rank, because the target language is English, separated by spaces when doing Rerank, a weight of all the vocabulary actually obtained
- (3) Subject/Object + PLM Rerank strategy, through PLM's own participle, the final result is 21K vocabulary weights.

The rerank scores of the three types of strategies are shown in Table 5. The PLM Rerank based on Subject/Object achieved the best results, with a relative increase of 1.62%.

Rerank	BLEU	
Beam Search (Base line)	37.63	
Rank SVM	38.07	+0.44
Bert Rerank	39.25	+1.62

Table 5: Bert Rerank for NMT Promotion

V. CONCLUSIONS AND FUTURE WORK

Knowledge embedding and pre-training are typical ways of improving natural language processing and machine translation. How to smoothen and improve machine translation effects end-to-end is our next research plan.

Reference

- [1]I. V. Serban et al., 'A hierarchical latent variable encoder-decoder model for generating dialogues', in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
- [2]A. Vaswani et al., 'Attention is all you need', in Advances in neural information processing systems, 2017, pp. 5998–6008.
- [1]W. L. Taylor, "'Cloze procedure": A new tool for measuring readability', Journalism Bulletin, vol. 30, no. 4, pp. 415–433, 1953.
- [3]H. Yang, G. Xie, Y. Qin, and S. Peng, 'Domain Specific NMT based on Knowledge Graph Embedding and Attention', presented at the 2019 21st International Conference on Advanced Communication Technology (ICACTION), 2019, pp. 516–521.
- [2]C. Alberti, K. Lee, and M. Collins, 'A

- BERT Baseline for the Natural Questions', ArXiv, vol. abs/1901.08634, 2019.
- [4]J. Yang et al., 'Towards Making the Most of BERT in Neural Machine Translation', arXiv preprint arXiv:1908.05672, 2019.
- [5]S. R. Bowman et al., 'Looking for ELMo's friends: Sentence-Level Pretraining Beyond Language Modeling', arXiv preprint arXiv:1812.10860, 2018.
- [6]D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, 'Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment', arXiv preprint arXiv:1907.11932, 2019.
- [7]N. Nangia and S. R. Bowman, 'Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark', arXiv preprint arXiv:1905.10425, 2019.
- [8]W. Huang, X. Cheng, T. Wang, and W. Chu, 'BERT-Based Multi-Head Selection for Joint Entity-Relation Extraction', arXiv preprint arXiv:1908.05908, 2019.
- [9]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint arXiv:1810.04805, 2018.
- [10]Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, 'Understanding the Behaviors of BERT in Ranking', arXiv preprint arXiv:1904.07531, 2019.
- [7]Z. Huang, W. Xu, and K. Yu, 'Bidirectional LSTM-CRF models for sequence tagging', arXiv preprint arXiv:1508.01991, 2015.
- [11]D. Zhang, H. Xu, Z. Su, and Y. Xu, 'Chinese comments sentiment classification based on word2vec and SVMperf', Expert Systems with Applications, vol. 42, no. 4, pp. 1857–1863, 2015.
- [12]S. Gupta and E. Hulburd, 'Exploring Neural Net Augmentation to BERT for Question Answering on SQUAD 2.0', arXiv preprint arXiv:1908.01767, 2019.
- [13]K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, and P. He, 'Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text', arXiv preprint arXiv:1908.07721, 2019.
- [14]W. Peng, J. Liu, L. Li, and Q. Liu, 'Huawei's NMT Systems for the WMT 2019 Biomedical Translation Task', in Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), 2019, pp. 164–168.
- [15]Z. Liu, Z.-Y. Niu, H. Wu, and H. Wang, 'Knowledge Aware Conversation Generation with Reasoning on Augmented Graph', arXiv preprint arXiv:1903.10245, 2019.
- [16]I. Sutskever, O. Vinyals, and Q. V. Le, 'Sequence to sequence learning with neural networks', in Advances in neural information processing systems, 2014, pp. 3104–3112.



Hao Yang, Research Scientist of 1202 Lab, Huawei Company limited. This author became a Member (M) of IEEE in 2005, a Senior Member (SM) in 2009. Dr Yang achieved Phd from Beijing University of Posts and Telecommunications in 2009, and majoring research fields includes Natural Language Processing, Neural Network Translation and Deep Learning in Text areas.