YuQ: A Chinese-Uyghur Medical-domain Neural Machine Translation Dataset Towards Knowledge-driven

Qing Yu Xinjiang University yuqing0131@126.com
Zhe Li Xinjiang University lizhe@stu.xju.edu.cn
Jiabao Sheng Xinjiang University jiabao@stu.xju.edu.cn
Jing Sun Xinjiang University sunjing@xju.edu.cn
Wushour Slamu Xinjiang University wushour@xju.edu.cn

Abstract

Recent advances of deep learning have been successful in delivering state-of-the-art performance in medical analysis, However, deep neural networks (DNNs) require a large amount of training data with a high-quality annotation which is not available or expensive in the field of the medical domain. The research of medical domain neural machine translation(NMT) is largely limited due to the lack of parallel sentences that consist of medical domain background knowledge annotations. To this end, we propose a Chinese-Uyghur NMT knowledge-driven dataset, **YuQ**, which refers to a ground medical domain knowledge graphs. Our corpus contains 65K parallel sentences from the medical domain and 130K utterances. By introduce medical domain glossary knowledge to the training model, we can win the challenge of low translation accuracy in Chinese-Uyghur machine translation professional terms. We provide several benchmark models. Ablation study results show that the models can be enhanced by introducing domain knowledge.

1 Introduction

Knowledge can improve the translation quality in NMT models where background knowledge plays a vital role in the success of text generation(Shang et al., 2015; Li et al., 2016; Shao et al., 2016). In neural machine translation systems, background knowledge is defined as slot-value pairs, which provide key information for proper noun translation, and has been well defined and thoroughly studied in conversational systems (Wen et al., 2015; Zhou et al., 2016). However, in neural machine translation of terminology, it is important but challenging to leverage background knowledge, which is represented as either knowledge graphs (Zhu et al., 2017; Zhou et al., 2018a) or unstructured texts (Ghazvininejad et al., 2018), for making improve the accuracy of proper noun translation especially medical domain.

Freshly, a variety of knowledge-based text generation corpora have been proposed (Zhou et al., 2018b; Dinan et al., 2018; Moghe et al., 2018) to fill the gap where previous datasets do not provide knowledge grounding of the text generation (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017). However, these datasets are not suitable for the medical domain or knowledge planning through neural machine translation based on knowledge. OpenDialKG (Moon et al., 2019) and DuConv (Wu et al., 2019) use knowledge graphs as knowledge resources. However, for knowledge-grounded NMT datasets still have the gap.

In this paper, As given in Figure-1, we propose YuQ, a Chinese-Uyghur neural machine translation dataset towards the medical domain, which is suitable for modeling knowledge interactions in machine translation in the medical domain, including knowledge planning, knowledge grounding, knowledge adaptations, etc. YuQ contains 65K utterances and 130K parallel corpus in the medical domain. Each sentence is annotated with related knowledge entities in the knowledge graph, Its effect is as supervision for knowledge interaction modeling. Furthermore, YuQ contains medical topics, which manually annotated accurately with higher quality than other datasets. The relations of entity are explicitly defined in the knowledge graph. We provide a benchmark to evaluate both generation- and retrieval-based neural machine translation

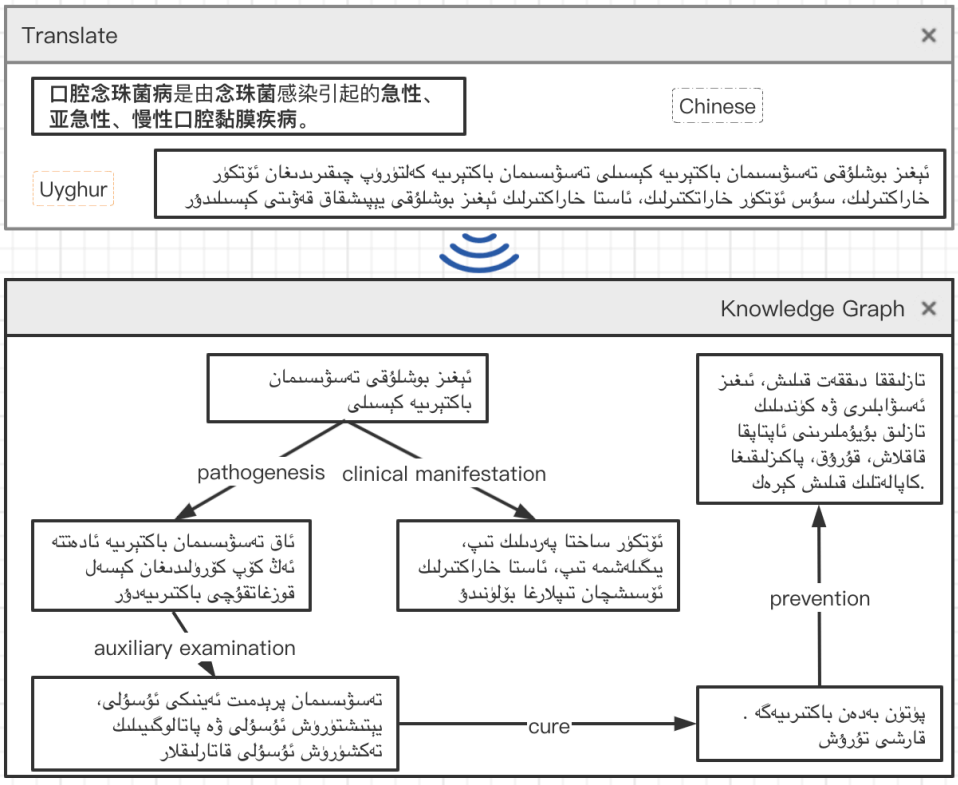

Figure 1: An example in YuQ from the medical domain. The **bold** text is the related knowledge that is utilized in NMT.

models on the YuQ dataset with/without access to the medical knowledge. Results show that knowledge-based contributes to the advancement of these models while existing models are still not strong enough to deliver knowledge-coherent NMT, indicating a large space for future work.

In summary, this paper makes the following contributions:

- We construct a new dataset, YuQ, for knowledge-driven neural machine translation in Chinese-Uyghur. YuQ contains 130K utterances in medical domains.
- YuQ provides a benchmark to evaluate the ability of neural machine translation with access to the corresponding knowledge in medical domains. The corpus can empower the research of not only knowledge-grounded machine translation text generation but also domain adaptation or transfer learning between similar domain or dissimilar domains.
- We provide benchmark models on this corpus to facilitate further research and conduct extensive experiments. Results show that the models can be enhanced by introducing background knowledge, but there is still much room for further research.

2 Related work

Recently, neural machine translation has been largely advanced due to the increase of publicly available machine translation data (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017). However, the lack of annotation of background information or related knowledge results in a significant bottleneck in medical term translation, where the translation accuracy of medical terms needs to improve. These models produce a translation that is substantially different from those humans translate, which largely rely on background knowledge.

To facilitate the development of NMT models that mimic human translate, there have been several knowledge-grounded corpora proposed. (Duan et al., 2020) proposes a new NMT method that is based on no parallel sentences but can refer to a ground-truth bilingual dictionary. This

new task can effectively improve the accuracy of the translation of specialized words in the medical domain. However, the Perplexity of translated sentences is not as well as Seq2Seq architecture. (Chen et al., 2020) considers the importance of the word in the sentence meaning and design a content word-aware NMT to improve translation performance. However, the accuracy of generated machine translation for medical terminology is often not controllable, resulting in some mistakes in the generated results. (Hao et al., 2019) presents multi-granularity self-attention (MG-SA): a neural network that combines multi-head self-attention and phrase modeling and can capture useful medical-domain phrase information at various levels of granularities. (Sokolov and Filimonov, 2020) presents an automatic natural language generation system, capable of generating both human-like interactions and annotations by the means of paraphrasing to solve manual annotations are expensive and time-consuming.

To obtain the high-quality knowledge-grounded datasets, some studies construct from scratch with human annotators, based on the unstructured text or structured knowledge graphs. For instance, several datasets (Zhou et al., 2020; Zhou et al., 2018b; Gopalakrishnan et al., 2019) have human conversations where participants have access to the unstructured text of related background knowledge. while OpenDialKG (Moon et al., 2019) and DuConv (Wu et al., 2019) build up their corpora based on structured knowledge graphs. (Young et al., 2018) proposes to explicitly augment input text with concepts expanded via 1-hop relations where KG triples are represented in the sentence embeddings space. (He et al., 2017) propose a system which iteratively updates KG embeddings and attends over connected entities for response generation. However, several challenges remain to scale the simulated knowledge graph, for knowledge augmented text generation, (Parthasarathi and Pineau, 2018; Ghazvininejad et al., 2018; Long et al., 2017) uses embedding vectors obtained from external knowledge sources, Wikipedia, free-form text, etc. as an auxiliary input to the model in dialog generation. Knowledge graphs can provide rich structured knowledge facts for better language understanding, (Zhang et al., 2019) utilize both large-scale textual corpora and KGs to train an enhanced language representation model (ERNIE), which can take full advantage of lexical, syntactic, and knowledge information simultaneously.

3 Datasets

The general method of constructing a parallel corpus is to collect, sort, mark, preserve and utilize professional corpus software for parallel processing and retrieval of the bilingual corpus. This paper is slightly different. In the processing of Chinese corpus, automatic line partitioning is carried out first, and the text is translated manually according to the line partition, which avoids the line labeling and alignment processing of the corpus. In the later retrieval, the method of combining professional corpus software and self-built retrieval system is adopted.

3.1 Data Collection

By searching a huge number of literatures and investigating in the hospital, a Chinese corpus from the general practitioner diagnosis and treatment system is finally determined. The data collected covered seven clinical disciplines: internal medicine, surgery, pediatrics, obstetrics and Gynecology, infectious diseases, dermatology, and Venereology, and five sense organs science. Each diagnosis and treatment article was retrieved by using the word crawl tool text, and a storage directory is established according to the department name and disease type. The disease name of a single diagnosis and treatment article was stored as a TXT file name, and the storage format was UTF-8 A total of 593 articles, 7 department catalogs, 65 disease catalogs, and 593 disease diagnosis and treatment corpora have been built. The corpus data is from clinical diagnosis in the hospital, and the content is authentic and representative. The balance of the corpus is fully considered in the collection. The proportion of data collected by each department is respectively Results: internal medicine 26.78%, surgery 15.17%, pediatrics 13.59%, obstetrics and Gynecology 10.92%, infectious diseases 12.09%, dermatology and Venereology 10.13%, facial

science 11.30%, basically meet the actual needs of patients, and reflect the medical language style and characteristics.

3.2 Corpus Preprocessing

Chinese medical and health data were collected manually, totaling 45,216 sentences. The data cover 12 major clinical disciplines: infectious diseases, dermatology, and venereology, facial features, epidemiology, internal medicine, surgery, pediatrics, obstetrics and gynecology, neuropathy, psychiatry, ophthalmology, and stomatology, totaling 739 diseases. The collection contents for each disease include etiology and pathology; Diagnosis and differential diagnosis; Clinical manifestations; Inspection, auxiliary inspection, and laboratory inspection; Therapy and physical therapy; Prevention, etc. The acquisition of medical texts is a relatively difficult task, and its text preprocessing is also quite difficult. General data preprocessing methods are applied to medical texts, but the effect is not significant, and medical words are often scattered. For example, the word "da chang gan jun" is divided into two words "da chang" and "gan jun" in the preprocessing process. The obtained processing results cannot be directly used for translation model training. The input data set suitable for model training needs to be obtained through text garbled code filtering, length ratio filtering, text word segmentation, and other steps in advance. After denoising, the corpus is divided into three levels according to the UTF-8 format: root directory, Department directory and disease category directory.

3.3 Annotation

The actual work of translation processing is after the corpus is automatically entered into the database. At this time, the work of line segmentation and text entry into the database has been completed. Translators translate according to the prescribed format, avoiding the problem of alignment.

3.4 Knowledge Graph Construction

The sparsity and the huge scale of the knowledge are difficult to handle, the annotated medical corpus is expensive, and the knowledge of these medical entities contains both structured knowledge triples and unstructured knowledge texts, which make the task more general but challenging. After filtering the start entities which have few knowledge triples, the medical domain contains 215 start entities, respectively. After filtering the start entities, we built the knowledge graph. Given the start entities as seed, we build their neighbor entities within three hops. We merged the start entities and these build entities (nodes in the graph) and relations (edges in the graph) into a domain-specific knowledge graph for medical domains. The statistics of the knowledge graphs used in constructing YuQ are provided in Table-1 and Table-2.

Entity Type	Explain	Number	Example
Test	Diagnostic Inspection Items	76	blood sugar, urinary ketone body
Disease	Disease	23	diabetic cardiomyopathy
Drug	Drug	73	glibenclamide, repaglinide
Food	Food	19	protein, fat
Symptom	Symptoms of disease	24	Drink more, eat more, urinate more
Total	Total	215	

Table 1: Statistics of the knowledge graph entity types of YuQ

4 Corpus Analysis

Chinese-Uyghur medical parallel corpus is a special corpus. By building a thesaurus, analyzing the frequency of words, we can make an objective analysis of the lexical features, determine the position and nature of different words in the lexical list in the medical corpus, and reveal the

Relationship Type	Explain	Number	Example
Belongs_to	Belong to	2	<type 1 diabetes, belonging_to, diabetes>
Acompany_with	complicating disease	18	< diabetic cardiomyopathy, Acompany_with, diabetic microangiopathy >
Cure	Therapeutic	101	< metformin,cure,glyburide >
No_eat	Avoid food for diseases	10	<disease,No_eat,wine>
prevention	therapy method	9	<disease,prevention,sea fish>
Symptom	Disease symptoms	28	<disease,Symptom,urine>
auxiliary_examination	Check the diagnosis	287	<disease,auxiliary_examination,insulin>
Total	Total	465	

Table 2: Statistics of the knowledge graph relationship types of YuQ

distribution law of lexical frequency phenomenon. At the same time, we compare the self-built corpus with other large-scale general corpora to further statistically analyze the importance of different words in the special corpus.

4.1 Lexical feature analysis

4.1.1 Construct Vocabulary

Using the EmEditor tool to replace all part-of-speech tags in the segmented corpus with spaces, A corpus separated by spaces is formed. According to the decreasing order of the occurrence frequency of each word, i.e. High-frequency words are ranked first and low-frequency words are ranked second, and the words are numbered with natural numbers. The highest occurrence frequency is level 1, followed by level 2. Rank is used to represent the word-level sequence and freq is used to represent the occurrence frequency of words in the corpus, thus constructing the vocabulary shown in Table-3:

Rank	Word	Freq	P	Ln(r)	Ln(f)
1	Treatment	2840	0.014196167	0	7.9515595
2	Onset	1488	0.007437992	0.6931472	7.305188
3	Symptoms	1428	0.0071380725	1.098612 3	7.26403
4	Occurrence	1162	0.005808431 7	1.386294 4	7.057898
5	Cause	1095	0.005473522	1.609 438	6.9985094
6	Patient	1033	0.005163606	1.7917595	6.9402223
7	General	957	0.0047837086	1.9459101	6.8638034
8	Serious	916	0.0045787636	2.0794415	6.8200164
9	Operation	866	0.004328831	2.1972246	6.763885
10	mg	832	0.004158877 3	2.3025851	6.7238326

Table 3: Word frequency statistics. Word is a segmented word in the corpus. P is the probability that words appear in the corpus; Ln(r) and Ln(f) are used to calculate the logarithm of Rank and freq respectively.

4.1.2 Statistical analysis of word frequency

Make statistics on the vocabulary, A total of 14,470 different words were acquired, Of these, 5,703 words appear only once, 39.41% of the total. Different from the general corpus, Most of the words with frequency 1 in this corpus are professional words in the medical field. Meaning. 121 words appear twice, 14.66 percent of that total. The word appearing more than three-time, 45.93% of the total. After analyzing the results of word frequency, That is, the 5% word appears only once, 20% word appears twice. But there is a slight gap, The main reason is that there are many professional terms in the medical corpus, Word segmentation algorithm needs to be further improved, In addition, the corpus segmented by the current word segmentation algorithm, It also contains a large number of English strings, Chinese-English and English-Chinese mixed

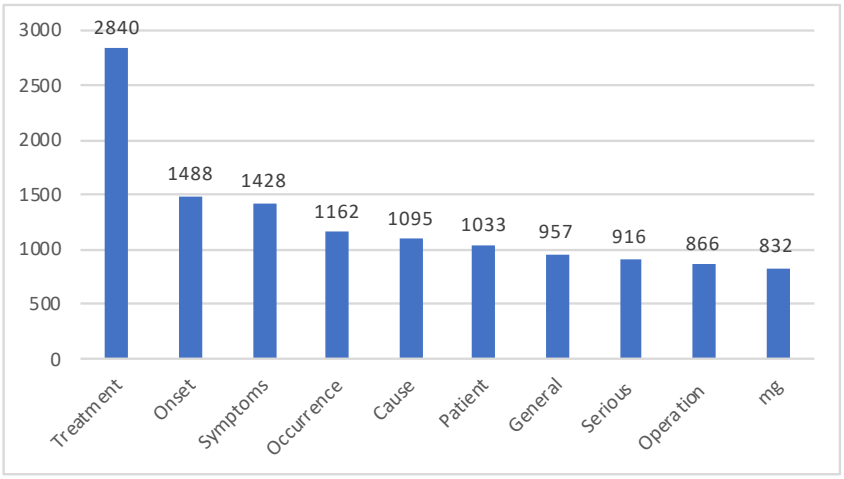

Figure 2: The top ten high-frequency words and frequency of Chinese Medical Corpus.

words, 323 with the frequency of 1, 104 with the frequency of 2 and 167 with other frequencies, which also conforms to the characteristics of medical corpus and includes a large number of transliterated and abbreviated foreign words, such as tc, r globulin, etc. The 10 words with the highest frequency in the vocabulary list and their frequencies are shown in Table-2, which fully reflect the medical characteristics of corpus data.

4.2 Contrastive Analysis of Lexical Features

The People's Daily has a corpus of nearly 2 million words in January, with a wide range of contents and a huge amount of data. Taking this as a reference corpus, the corpus retrieval software AntConc is used to compare and analyze the frequency of each word and word in the Chinese medical corpus word and word frequency list with the frequency of the word and word in the reference corpus to reflect the importance and particularity of the word and word to the medical corpus.

The list of word and word frequency includes information: Rank (word/word level), freq (frequency), word (word/word) and keyness (significant difference in frequency, the frequency difference between the same word and word in the two corpora, the greater the difference, the greater the keyness value). Some comparative statistical results are shown in Table-4.

Rank	Freq	Keyness	Word
1	6 396	12213.213	Can
2	5 462	9 915.505	Or
3	2 840	6 381.669	Treatment
4	2 260	5 228.570	those
5	1 846	4 577.089	sex
6	2 661	3 756.796	and
7	1 428	3 519.548	Symptoms
8	1 493	2 949.540	Heart
9	1 707	2 469.261	as
10	1 144	2 386.947	Disease

Table 4: Statistical Analysis for Contrast of Word Frequency Characteristics between our corpus and People's Daily corpora.

4.2.1 Comparative Result Analysis

Corpus data using existing segmentation tools, based on Chinese medicine, shard 16637 words, which "Word" word frequency is greater than the reference corpus, 14050, 2587 less than the reference corpus, respectively constructed two-word frequency table.

Observe word frequency table of 14050 words, found keyness values by the maximum first became smaller, close to zero, until it is equal to zero, according to the keyness values change, the analysis of word frequency table is as follows:

In the first part, the value of keyness is very large at the beginning. Keyness >5 is taken as the boundary, and there are 7840 words, which are most commonly used in medical treatment, such as treatment and patient.

The second part, in order to 0 or less keyness 5 or less as the boundary, a total of 6210 words, at this time is divided into two cases: (1) to 0 or less keyness 5 or less and freq = 1, a total of 5089 words, observed that these words not only keyness value is small, gradually tends to zero, word frequency and minimum, these words are not commonly used for the two corpora, several medical field has the characteristics of medical is not commonly used words, such as early focal infarction disease, diffuse peritoneal infection, etc. With 321 as the letter combinations, such as athabasca, arvd. Athabasca, Arvd is acute obstructive suppurative cholangitis, respectively, the abbreviation of right ventricular cardiomyopathy arrhythmia caused by sex. (2) 0 or less keyness 5 or less and freq > 1, a total of 1121 words, this part of the vocabulary, freq value is very high, when keyness approach to find these words, such as, in the early morning, belong to the more commonly used words, basic is commonly used in the medical corpus, also commonly used in People's Daily corpus, frequency is similar in the two corpora.

Then Observe word frequency table of 2587 words:

(1) Keyness value is the largest at first and then decreases from the maximum. Contrary to the first part, when keyness value is large, it is all the data with high word frequency in the corpus of People's Daily, such as China, problems, development, etc.

(2) When keyness value is small and FREq =1, the specificity of words cannot be seen, which is related to the fact that People's Daily is a general corpus. (3) when the keyness value smaller and larger freq, a total of 618 words, belong to two corpora are more frequently used vocabulary, but inadequate medical characteristics, such as a hospital bed, etc. After statistical analysis, combined with artificial proofreading, easily from 7840 words and 5089 words, sort out the medical special corpus theme vocabulary. Through the analysis of the above characteristics, not only reflects the corpus itself vocabulary characteristics, common vocabulary, vocabulary, etc. That validates whether corpus construction and late for further study of natural language processing technology to lay the foundation of medicine.

5 Experiments

5.1 Models

As provided baseline models for knowledge-driven NMT modeling, we evaluate such models on our corpus generation-based and retrieval-based models. To investigate the knowledge information annotation results, we evaluate the models with/without introducing to the knowledge graph of our dataset.

5.1.1 Generation-based Models

Language Model(LM) (Bengio et al., 2003): We train a language model that maximizes the log likelihood: $logP(x) = \sum_t logP(x_t|x < t)$, where x denotes a long sentence that sequentially concatenates all the utterances of a machine translation.

Seq2Seq (Sutskever et al., 2014): An encoder-decoder model. The input of the encoder is the concatenation of the past $k-1$ utterances, while the target output of the decoder was the $k-th$ utterance. If there are fewer than $k-1$ sentences in the NMT history, all the past utterances would be used as input.

RNNSearch (Bahdanau et al., 2014) RNNSearch is to improve the performance of Seq2Seq by the attention mechanism, where each word in Y is conditioned on different context vector c, with the observation that each word in Y may relate to different parts in x. In particular, y_i corresponds to a context vector c_i, and c_i is a weighted average of the encoder hidden states $h_1, ..., h_T$:

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{1}$$

where $a_{i,j}$ is computed by:

$$\alpha = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{2}$$

$$e_{ij} = g(s_{t-1}, h_j) \tag{3}$$

where g is a multilayer perceptron.

Transformer (Vaswani et al., 2017): Transformer abandons the recurrent network structure of RNN and models a piece of text entirely based on attention mechanisms. The most important module of the coding unit is the Self-Attention module, which can be described as:

$$Attention(Q, K, V) = Softmax\left(\frac{Q\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{4}$$

To extend the ability of the model to focus on different locations and to increase the representation learning capacity of subspaces for attention units, Transformer adopts the "multi-head" mode that can be expressed as:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{5}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^K) \tag{6}$$

THUMT (Zhang et al., 2017): THUMT is an open-source toolkit for neural machine translation developed by the Natural Language Processing Group at Tsinghua University and a new implementation developed with TensorFlow.

5.1.2 Retrieval-based Model

BERT (Devlin et al., 2019): We adapt this deep bidirectional transformers (Vaswani et al., 2017) as a retrieval-based model. For each utterance, we extract medical keywords and retrieve 10 translation candidates. The training task is to predict whether a candidate target utterance is the fitting source utterance given the source utterance where a sigmoid function is used to output the conditional probability $p(x_t|x_{0:t-1})$ can be modeled by a probability distribution over the vocabulary given linguistic context $x_{0:t-1}$. The context $x_{0:t-1}$ is modeled by neural encoder $f_{enc}(\cdot)$, and the conditional probability:

$$p(x_t|x_{0:t-1}) = g_{LM}\Big(f_{enc}(x_{0:t-1})\Big) \tag{7}$$

where gLM($\cdot$) is the prediction layer. We select the candidate sentence with the largest probability.

5.1.3 Knowledge-aware Models

A key-value memory module (Miller et al., 2016) is introduced to the aforementioned models to utilize the knowledge information. We treat all knowledge triples mentioned in an NMT as the knowledge information in the memory module. For a triple that is indexed by i, we represent the key memory and the value memory respectively as a key vector k_i and a value

vector v_i, where k_i is the average word embeddings of the head entity and the relation, and v_i is those of the tail entity. We use a query vector q to attend to the key vectors $k_i(i = 1, 2, ...)$: $a_i = softmaxx_i(q^T k_i)$, then the weight sum of the value vectors $vi(i = 1, 2, ...), v = \sum_i a_i v_i$, is incorporated into the decoding process (for the generation-based models, concatenat with the initial state of the decoder) or the classification (for the retrieval-based model, concatenat with the <CLS> vector). For Seq2Seq, q is the final hidden state of the encoder. For RNNSearch and Transfomer, we treat the context vector as the query, while for BERT, the output vector of <CLS> is used.

Our dataset has a sentence-level annotation of the triples of knowledge used by each utterance. In order to compel the knowledge-aware models to attend to the KG triples, we applied an extra loss of focus.

$$L_att = -\frac{1}{|truth|} \sum_{i \in truth} \log a_i \tag{8}$$

where truth is the set of indexes of triples that are used in the true response. The total loss are the weighted sum of $L^{(l)}$ and L_att:

$$L_{tot}^{(l)} = L_0^{(l)} + \lambda L_{att}, l \in g, r. \tag{9}$$

The knowledge-enhanced BERT is initialized from the fine-tuned BERT, and the transformer parameters are frozen during training the knowledge related modules. The purpose is to exclude the impact of the deep transformers but only examine the potential effects introduced by the background knowledge.

5.2 Setup

We implement the above models with Pytorch while THUMT implement by tensorflow. The type of RNN network units is all GRU and the number of hidden units of GRU cells is all set to 200. ADAM as used to optimize all the models with the initial learning rate of 1×10^{-5} for BERT and 1×10^{-3} for others. The mini-batch sizes are set to 2 sentences for LM and 32 pairs of source- and target-sentence for Seq2Seq.

5.3 Automatic Evaluation

5.3.1 Metrics

We adopt BLEU, Rouge, and Perplexity as the evaluation metrics to measure the quality of the generated response. For BLEU, we employ the values of BLEU 1-4 and show the value of Rouge-1/2/L. Intuitively, the higher BLEU score and Rouge score mean more n-gram overlaps between the generated responses, and thereby indicate the better performance. Nevertheless, Perplexity is a well-established performance metric for generative text generation models. On the other hand, Perplexity explicitly measures the ability of the model to account for the syntactic structure of the dialogue, and the syntactic structure of each utterance and lower perplexity is indicative of a better model.

5.3.2 Results

The results are shown in Table-5. We analyze the results from the following viewpoints:

The influence of knowledge: In the medical domains, the knowledge-aware BERT model achieves the best performance in all of the metrics, as it retrieves the golden-truth response at a fairly high rate. The transformer-based models perform best in BLEU-k among all the generation-based baselines without considering the knowledge. Knowledge-aware Transformer has better results of BLEU-k and better results of PPL, while the knowledge-enhanced Transformer–based models achieve the best metrics scores among all the generation-based models.

Comparison between models: In the medical domains, the knowledge-aware BERT model achieves the best performance in all of the metrics, as it retrieves the golden-truth response at a fairly high rate. The transformer-based models perform best in BLEU-k among all the generation-based baselines without considering the knowledge. Knowledge-aware Transformer has better results of BLEU-k and better results of PPL, while the knowledge-enhanced Transformer–based models achieve the best metrics scores among all the generation-based models.

Model	**PPL**	BLEU-1/2/3/4				Rouge-1/2/L		
LM	45.44	10.27	2.31	0.34	0.09	0.271	0.162	0.259
Seq2Seq	41.13	17.19	6.67	1.06	0.16	0.368	0.167	0.273
RNNSearch	40.45	20.97	8.40	1.71	1.27	0.387	0.124	0.248
Transformer	39.28	25.08	10.37	2.43	2.75	0.394	0.158	0.279
THUMT	21.91	24.22	12.40	2.71	2.27	0.384	0.207	0.313
BERT	37.32	27.63	14.32	3.35	3.13	0.427	0.216	0.314
Transformer+know	37.24	30.29	15.79	3.15	3.02	0.453	0.205	0.317
THUMT+know	37.91	30.41	18.43	3.72	3.01	0.498	0.237	0.349
BERT+know	33.11	33.14	20.54	4.93	3.91	0.592	0.481	0.591

Table 5: Automatic evaluation. The best results of generative models and retrieval models are in bold and underlined respectively. "+ know" means the models enhanced by the knowledge base.

5.4 Manual Evaluation

To better understand the quality of the generated responses from the semantic and knowledge perspective, we conducted the manual evaluation for knowledge-aware BERT, knowledge-aware RNNSearch, and Transformer, which have achieved advantageous performance in automatic evaluation.

5.5 Metrics

In terms of the fluency and coherence metrics, human annotators are asked to score a generated response.

Fluency (rating scale 0,1,2) is described as if the answer is normal and fluid:

- Grade 0 (bad): the grammatical mistakes are not articulate and challenging to comprehend.
- Grade 1 (fair): includes but yet clear grammatical errors.
- Grade 2 (good): humans generate it fluently and plausibly.

Coherence (rating scale is 0,1,2) is characterized as whether an answer to the context and knowledge information is valid and coherent:

- Grade 0 (bad): History is meaningless.
- Grade 1 (fair): important to the context, but not consistent with the details on expertise.
- Grade 2 (good): both context-relevant and consistent with background information.

5.6 Annotation Statistics

We randomly sampled about 500 contexts from the test sets and generated sentences by each model. These 1,500 parallel sentences pairs in total and related knowledge graphs are presented to three human annotators.

5.7 Results

The findings are seen in the Table-3. As can be shown, knowledge-aware BERT greatly outperforms other models in all dimensions in the medical realms, which correlates with automated evaluation performance.The Fluency is at 2.00 because all human-written sentences are the collected responses. The fluency scores of both generation-based models are approximately 2.00 suggesting that the translation produced is fluent and grammatical. The BERT and knowledge-aware BERT Coherence scores are higher than 1.00 but still have a big gap of 2.00, meaning that in most instances the translation produced is important to the background but not consistent with knowledge-aware facts. The Coherence score is substantially enhanced after integrating the knowledge information into BERT, as the knowledge information is more reflected in the produced translation.

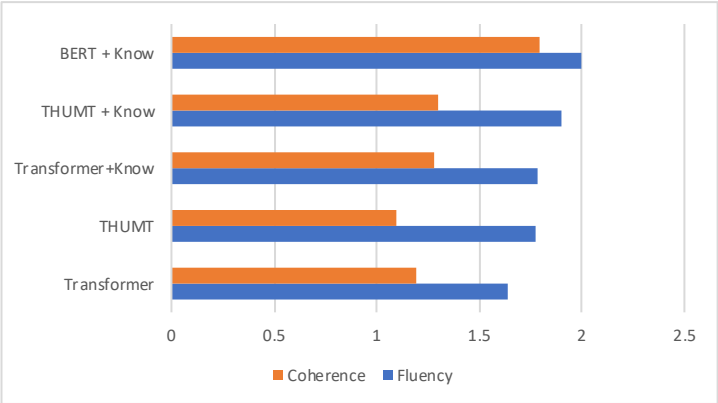

Figure 3: Manual evaluation between three generative models. "+ know" means the models enhanced by knowledge information.

5.8 Case Study

Some sample translations in the medical realms provided by knowledge-aware BERT are seen in Table-6. As we can see, knowledge-aware BERT is able to produce knowledge-based translation after the presentation of knowledge content, such as translation with expertise in the medical domain. However, it is still challenging for information-aware BERT to produce knowledge-coherent responses with respect to unstructured text awareness as modeling knowledge of unstructured texts requires more powerful models.

<table>
<tr><th rowspan="2">Translation</th><th colspan="3">Knowledge Triple</th></tr>
<tr><th>Head Entity</th><th>Relation</th><th>Tail Entity</th></tr>
<tr><td>CN: 眶下间隙蜂窝织炎</td><td rowspan="6">ئېغىز بوشلۇقىنىڭ ئىككى يۈز قىسمى كۆنەكسىمان توقۇلما يالۇغى</td><td rowspan="3">كېلىنكىلىق ئىپادىلىرى
(clinical manifestation)</td><td rowspan="3">يۈقىرىقى لەۋ ئىششىپ چىقىدۇ، بۇرۇن-كالپۇك ئېرىقچىسى يوقاپ كېتىدۇ</td></tr>
<tr><td>Ug: ئېغىز بوشلۇقىنىڭ ئىككى يۈز قىسمى كۆنەكسىمان توقۇلما يالۇغى</td></tr>
<tr><td>CN: 感染发生于眼眶下方，上颌骨前壁与面部表情肌之间。</td></tr>
<tr><td>Ug: يۇقۇملىنىش قاپاقنىڭ ئاستى تەرىپى، ئوستىنكى ئىككىنىڭ ئالدى ۋە يۈز قىسملىرىدا كۆرىنەرلىك بولىدۇ.</td><td rowspan="3">داۋالاش
(cure)</td><td rowspan="3">پۈتۈن بەدەن ئانتىبىئوتىكلىرىنى ماسلاشتۇرۇپ يۇقۇملىنىشقا قارشى تۇرۇش</td></tr>
<tr><td>CN:全身配合抗生素抗感染。</td></tr>
<tr><td>Ug: پۈتۈن بەدەن ئانتىبىئوتىكلىرىنى ماسلاشتۇرۇپ يۇقۇملىنىشقا قارشى تۇرۇش</td></tr>
</table>

Table 6: Cases of the medical domain. Text is the knowledge used by the golden truth or the knowledge correctly utilized by the models.

5.9 Ablation Study

To evaluate the contributions of key factors in our method, we perform an ablation study.

The influence of BPE on the Morphological segmentation of Uygur language In order to verify the need for morphological segmentation of Uyghur language before using BPE technology, This paper compares the performance of BPE on Uyghur data without morphological segmentation and BPE on data after morphological segmentation under neural machine translation system respectively. According to Table-7, BLEU values of the same data set on the test set are 11.56 and 10.28 respectively. The former is 1.28 higher than the latter, and the improvement is not significant. Therefore, when BPE technology is adopted for the Uyghur language, morphological segmentation can be avoided, and BPE technology can effectively solve the problem of sparse data matrix.

vocabulary	Uyghur Morphological Segmentation	BLEU
12000	+Morphological segmentation	11.56
12000	-Morphological segmentation	10.28

Table 7: The Influence of Uyghur Morphological Segmentation on BLEU Value.

The Effect of Word List Size on Machine Translation Performance The above experimental conclusions show that there is no great influence on whether Uyghur language is morphologically segmented and then BPE technology is used. The performance comparison experiment of neural machine translation methods based on the self-attention mechanism is continued under different vocabulary sizes. Table-8 experimental results show that Uyghur language does not undergo morphological segmentation, and the BLEU value with a vocabulary size of 32000 is 19.89 higher than that of Uyghur language with morphological segmentation and a vocabulary size of 12000. This shows that on the scarce resources and rich forms of Chinese-Uygur data set, Compared with morphological segmentation, The size of the word list can improve the performance of machine translation, The reason may be that morphological segmentation leads to the lack of semantic information at the word level, For enlarging the vocabulary, it can effectively reduce the number of unregistered words and save the effective information at the word level without losing. The neural network based on the self-attention mechanism can better learn the morphological structure and features of words, thus effectively improving the performance of machine translation.

vocabulary	Uyghur Morphological Segmentation	BLEU
32000	-Morphological segmentation	30.17
12000	+Morphological segmentation	11.56

Table 8: The influence of vocabulary size and morphological segmentation on neural machine translation.

6 Conclusion and Future Work

In this paper, we propose a high-quality manually annotated Chinese-Uyghur medical-domain corpus for knowledge-driven neural machine translation, YuQ. It contains 130K utterances and 65K parallel sentences, with an average length of 19.0. Each parallel sentence contains sentence-level annotations that map each utterance with the medical knowledge triples. The dataset provides a benchmark to evaluate the ability to model knowledge-driven translation. We provide generation and retrieval-based benchmark models to facilitate further research. Extensive experiments illustrate that NMT models can be enhanced by introducing knowledge, whereas there is still much room in knowledge-grounded neural machine translation modeling for future work. We hope that this dataset facilitates future research on the medical-domain neural machine translation problem.

7 Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Qing Yu and Zhe Li are contributed equally to this research. This paper support by National Natural Science Foundation of China Research on the Construction of Chinese and Uygur Medical and Health Terms Resource Database Grant Number 61562082, and National Natural Science Foundation of China Research on Key Technologies of Uyghur-Chinese Phonetic Translation System Grant Number U1603262, Xinjiang Uygur Autonomous Region Graduate Research and Innovation Project Grant Number XJ2020G071, Dark Web Intelligence Analysis and User Identification Technology Grant Number 2017YFC0820702-3, and funded by National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

References

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020. Content word aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 358–364.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, pages 4171–4186.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. *arXiv preprint arXiv:2007.02671*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019. Multi-granularity self-attention for neural machine translation. *arXiv preprint arXiv:1909.02222*.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. 2017. A knowledge enhanced generative conversational service agent. In *Proceedings of the 6th Dialog System Technology Challenges (DSTC6) Workshop*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. 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)*, pages 1577–1586.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2016. Generating long and diverse responses with neural conversation models.

Alex Sokolov and Denis Filimonov. 2020. Neural machine translation for paraphrase generation. *arXiv preprint arXiv:2006.14223*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. pages 5998–6008.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. Thumt: An open source toolkit for neural machine translation. *arXiv preprint arXiv:1706.06415*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Hao Zhou, Minlie Huang, and Xiaoyan Zhu. 2016. Context-aware natural language generation for spoken dialogue systems. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2032–2041.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *arXiv preprint arXiv:2004.04100*.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv*, pages arXiv–1709.