CLICKER: ATTENTION-BASED CROSS-LINGUAL COMMONSENSE KNOWLEDGE TRANSFER

Ruolin Su[1*], *Zhongkai Sun*[2], *Sixing Lu*[2], *Chengyuan Ma*[2], *Chenlei Guo*[2]

[1]Georgia Institute of Technology, [2]Amazon Alexa AI
ruolinsu@gatech.edu, {zhongkas, cynthilu, mchengyu, guochenl}@amazon.com

ABSTRACT

Recent advances in cross-lingual commonsense reasoning (CSR) are facilitated by the development of multilingual pre-trained models (mPTMs). While mPTMs show the potential to encode commonsense knowledge for different languages, transferring commonsense knowledge learned in large-scale English corpus to other languages is challenging. To address this problem, we propose an attention-based Cross-LIngual Commonsense Knowledge transfER (CLICKER) framework for minimizing the performance gaps between English and non-English languages on commonsense question-answering tasks. CLICKER can effectively improve commonsense reasoning for non-English languages by differentiating language-specific knowledge from commonsense knowledge. Experimental results on public benchmarks demonstrate that CLICKER is effective in decoupling commonsense knowledge from non-commonsense knowledge and achieving remarkable improvements in the cross-lingual CSR task of languages other than English.

***Index Terms*—** Commonsense reasoning, multilingual, pre-trained language model, knowledge extraction, self-attention

1. INTRODUCTION

Commonsense reasoning (CSR) relies on shared and unchanged knowledge across different languages and cultures to help computers understand and interact with humans naturally [1]. CSR is a crucial problem in natural language processing that has proved important for artificial intelligence systems [2, 3].

Cross-lingual CSR aims to reason commonsense across languages, which is the key to bridging the language barrier in natural language understanding and generalizing CSR to a broader scope [4]. Recently, several cross-lingual datasets are proposed amidst the surging interests in cross-lingual CSR, *e.g.* XCOPA [5], X-CSQA [6], and X-CODAH [6]. Multilingual pre-trained language models (mPTMs) based on Transformer [7], such as mBERT [8], XLM [9], XLM-R [10] and InfoXLM [11], have also been demonstrated to have the potentials of conducting CSR in multiple languages [12, 4, 5, 6]. The performance of mPTMs for non-English CSR, however, is typically worse than that for English CSR due to data scarcity issues [13, 10, 5]. Furthermore, mPTMs continue to raise concerns regarding their ability to transfer commonsense knowledge given that they do not 1) explicitly extract commonsense knowledge and 2) improve the CSR for any specific language in multilingual scenarios [10].

To address the above issues, we propose a Cross-LIngual Commonsense Knowledge transfER (**CLICKER**) to bridge the performance gap of mPTMs for CSR between the source (**English**) and the target (**non-English**) language by eliciting commonsense knowledge explicitly via cross-lingual task-adaptive pre-training [14]. Specifically, CLICKER is a three-step framework based on XLM-R [10]. First, the task-adaptive pre-training is conducted on the multilingual commonsense corpora to help XLM-R adapt to the CSR task. A self-attention [7] mechanism is applied to learn the knowledge related to commonsense. Second, commonsense and non-commonsense knowledge are differentiated by jointly optimizing their similarity measurements. Third, the extracted commonsense knowledge representation is then fine-tuned on the downstream CSR tasks.

Experiments on public benchmark demonstrate our approach closes the discrepancies between English and German on CSR and outperform the XLM-R baseline on X-CSQA and X-CODAH benchmarks [6]. Further analysis indicates that the CLICKER can extract cross-lingual commonsense representations more effectively, and with better interpretability.

2. METHOD

This section introduces the CLICKER model based on XLM-R for cross-lingual CSR. As illustrated in Figure 1, CLICKER extracts commonsense knowledge of English to benefit non-English languages[1] in three steps: 1) task-adaptive pre-training, 2) commonsense differentiation, and 3) knowledge-transfer fine-tuning.

---

* Work done during internship at Amazon Alexa AI.

[1]In this paper, we take German as an example of a foreign language that is not up to par with English for CSR.

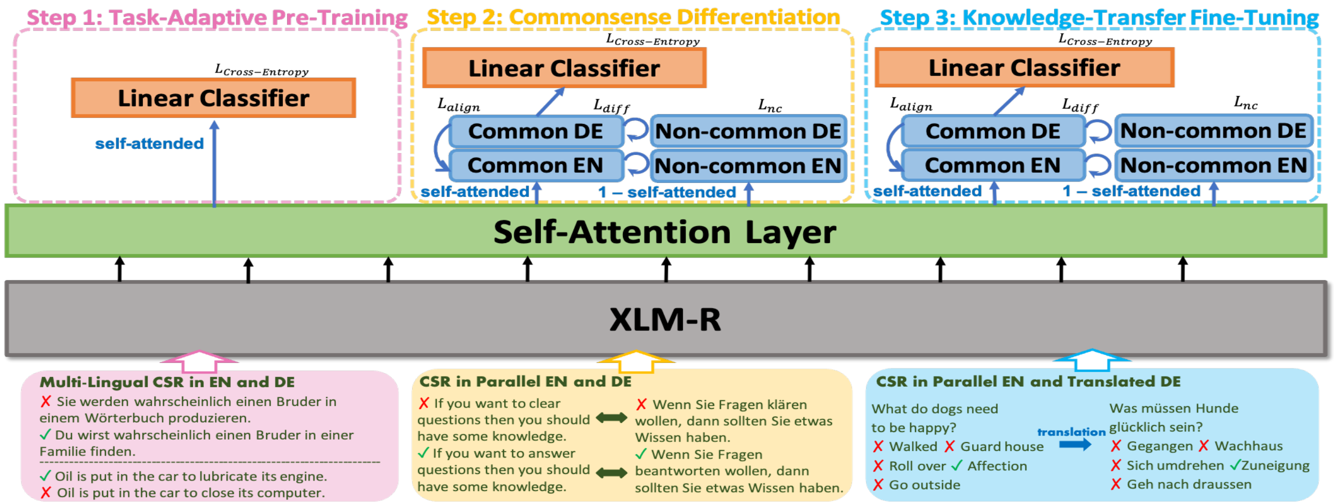

Fig. 1. The overview of CLICKER framework, which is trained in three steps with joint objectives for cross-lingual commonsense reasoning. Parameters in XLM-R and self-attentions layers are shared in all steps. Each check or cross represents whether the following choice is commonsense reasonable or not, respectively.

2.1. Problem Definition

The CSR task aims to select one from multiple choices that are most reasonable in commonsense given the previous statement or question. For example, the plausible choice of answer to *"What is a great place to lay in the sun?"* is *"beach"* rather than choices like *"in the basement"* or *"solar system"*. Denoting a set of choices for CSR input as $\mathbf{S}^{(j)}$'s, $j \in [1,\ldots,|C|]$, where the number of choices for each input as $|C|$, the goal is to predict the common sense choice:

$$\tilde{y} = \text{argmax}_j P(y = j|\{\mathbf{S}^{(1)},\ldots,\mathbf{S}^{(|C|)}\}) \tag{1}$$

2.2. Step One: Task-Adaptive Pre-Training

Task-adaptive pre-training uses self-attention to enable our model to learn preliminary commonsense knowledge of English and German, respectively, from the multilingual corpora. Specifically, the input is a tokenized utterance, *i.e.* $\mathbf{S}_i = \{[\text{CLS}], q_i^{(1)}, \ldots, q_i^{(k)}, [\text{SEP}]\}$, where $i \in [1,\ldots,N]$, N is the size of dataset, and k is the length of sequence. A self-attention layer is built on top of the Transformer to obtain attention toward commonsense knowledge from the Transformer's pooled output states. The self-attended outputs are optimized by the Cross-Entropy loss through a multiple-choice classifier to select the commonsense-reasonable choices. Our model is trained on multilingual CSR datasets containing examples of both English (**EN**) and German (**DE**).

2.3. Step Two: Commonsense Differentiation

In this step, the representation of the commonsense knowledge shared across EN and DE is differentiated from the non-commonsense representation leveraging paralleled EN and DE datasets. The inputs are similar to those in Sec 2.2 but examples with the same semantics in different languages are mapped together. We note here that the parallel datasets are not necessarily restricted to CSR datasets, but can be generalized to any bi-lingual datasets for mapping semantics of English and the non-English language, e.g. bi-lingual dictionaries or textbooks. The output states of the Transformer are pooled and weighted by the attention matrix followed by a linear projection, being extracted as **commonsense** embeddings X_i and **non-commonsense** embeddings $\tilde{X}_i$, respectively.

$$X_i = FFN(\text{Attention}(\mathbf{O}_i)) \tag{2}$$

$$\tilde{X}_i = FFN(1 - \text{Attention}(\mathbf{O}_i)) \tag{3}$$

where $\mathbf{O}_i$ is output hidden states from the last layer of the Transformer for the *i-th* input, and FFN represents a *Feed-Forward* layer. For brevity, we omit index i in the following equations.

We use X^{EN} and X^{DE} to denote commonsense embeddings of English and German inputs, respectively. Similarly, $\tilde{X}^{EN}$ and $\tilde{X}^{DE}$ represents non-commonsense embeddings. Knowledge mapping is made by measuring the similarities between commonsense embeddings and non-commonsense embeddings. Specifically, we maximize the cosine similarity between English and German embeddings that share the same and valid commonsense knowledge, *i.e.* $X^{EN_{j^*}}$ and $X^{DE_{j^*}}$, as in Eq. (4). And we minimize the cosine similarity between $X^{EN_{j^*}}$ and X^{EN_j}, as in Eq. (5). j^* is the index of the choice that is reasonable in commonsense, $j \in [1,\ldots,|C|]$ and $j \neq j^*$. Such that similar commonsense knowledge in both languages is projected into the same position in the semantic space.

$$\mathcal{L}_{align} = 1 - \cos(X^{EN_{j^*}}, X^{DE_{j^*}}) \tag{4}$$

$$\begin{aligned}\mathcal{L}_{diff} = \sum_{j=1, j\neq j^*}^{|C|} & (\max(0, \cos(X^{EN_{j^*}}, X^{EN_j})) \\ & + \max(0, \cos(X^{DE_{j^*}}, X^{DE_j})))\end{aligned} \tag{5}$$

On the other hand, the non-commonsense embeddings represent knowledge unrelated to cross-lingual commonsense. Assuming the correct choice and other incorrect choices of the input share similar non-commonsense knowledge, we maximize the intra-language cosine similarity of non-commonsense embeddings. Moreover, the correct choice of different languages should share the same non-commonsense knowledge so that we maximize inter-language cosine similarity jointly, as defined in Eq. (6).

$$\begin{aligned}\mathcal{L}_{nc} = & \sum_{j=1, j\neq j^*}^{|C|} (1 - \cos(\tilde{X}_i^{EN_{j^*}}, \tilde{X}_i^{EN_j})) \\ & + \sum_{j=1, j\neq j^*}^{|C|} (1 - \cos(\tilde{X}_i^{DE_{j^*}}, \tilde{X}_i^{DE_j})) \\ & + 1 - \cos(\tilde{X}_i^{EN_{j^*}}, \tilde{X}_i^{DE_{j^*}})\end{aligned} \tag{6}$$

All the losses above and the Cross-Entropy loss are optimized as the joint training objective of the cross-lingual CSR. We use output commonsense embeddings $X^{DE_{j^*}}$ and X^{DE_j} to calculate the Cross-Entropy loss.

2.4. Step Three: Knowledge-Transfer Fine-Tuning

Finally, our model is fine-tuned with the training objectives similar to Sec 2.3 for evaluating CSR on the multiple-choice question-answering (QA) and the clause-selection tasks, leveraging the parallel CSR datasets of English (**EN**) and German translated from English (**EN_DE**) as inputs. Different from previous steps, each input of XLM-R is the concatenation of question and each choice of answer which are then split into tokens with additional special ones, *i.e.* $\mathbf{S}_i = \{\text{[CLS]}, q_i^{(1)}, \ldots, q_i^{(m)}, \text{[SEP]}, \text{[CLS\_Q]}, a_i^{(1)}, \ldots, a_i^{(n)}, \text{[SEP]}\}$, where [CLS_Q] is the beginning special token of the answer spans, q_i and a_i are tokens of the question and answer, and m, n are numbers of question and answer tokens, respectively.

3. EXPERIMENTS AND ANALYSES

We use English and German subsets of Mickey Corpus [6] for Step 1 to warm up the multilingual language model for cross-lingual CSR tasks. Then we take advantage of parallel corpora of English and German in the Mickey Corpus again for Step 2 to obtain their semantic mappings and differentiate commonsense and non-commonsense embeddings. For Step 3, the CLICKER model is fine-tuned on the English and machine-translated German training set of X-CSQA and X-CODAH [6], which are in the style of multiple-choice QA and selection of appropriate clauses, respectively.

We compare our model with the multilingual contrastive pre-training (MCP) [6] model based on XLM-R$_B$ [10]. MCP model is trained on permuted Mickey Corpus for multilingual contrastive training and fine-tunes on cross-lingual CSR training set in English only. Instead, we re-implement it to train on the combination of English and German Mickey Corpus. Then we fine-tune it on both English and machine-translated German CSR training sets and evaluate it on the test set in German to make a fair comparison with our model.

The following subsections describe the experimental results and analyze CLICKER models on the cross-lingual CSR benchmarks X-CSQA and X-CODAH in German. Note that our experiments are conducted for commonsense knowledge transfer from English to German, but the approach can be extended to other languages.

3.1. Experimental Results

Table 1 shows the test accuracy of baselines and CLICKER models for CSR in German. Different combinations of losses are applied in experiments for optimizing commonsense differentiation. We observe consistent improvements with our three-step framework by extracting commonsense knowledge with self-attentions (*i.e.* CLICKER - *base*) on both datasets compared to baselines.

Results show that the *align* loss further improves the base CLICK model on X-CSQA. And the *non-commonsense (nc)* loss is proved effective on both datasets. The best performance on X-CSQA is achieved when using the *align* loss with or without the *diff* loss, which illustrates lining up embeddings of the same commonsense between English and German dominates the performance of CSR. Besides, the model with *align* and *nc* loss is slightly inferior to the model with *nc* loss only on X-CSQA. On X-CODAH, our CLICK models perform the best with the *nc* loss which maximizes the cosine similarity of non-commonsense embeddings, improving 1.6% on accuracy.

3.2. Discussion

Our models address the alignment of extracted embeddings with various combinations of objectives. The fact that *align+nc* loss is not as good as *nc* loss alone suggests a conflict between aligning the commonsense embeddings and aligning the non-commonsense embeddings. This can be explained as both objectives aiming to maximize the cosine similarity of embeddings, making it harder for the model to discern different commonsense knowledge in them. From

Models	Acc
X-CSQA	
MCP(XLM-R$_B$)* [6]	48.8
CLICKER - *base*	49.6 (+0.8)
CLICKER - *align*	**50.6 (+1.8)**
CLICKER - *align+diff*	**50.6 (+1.8)**
CLICKER - *nc*	49.8 (+1.0)
CLICKER - *align+nc*	49.6 (+0.8)
X-CODAH	
MCP(XLM-R$_B$)* [6]	49.2
CLICKER - *base*	50.2 (+1.0)
CLICKER - *align*	49.6 (+0.4)
CLICKER - *align+diff*	50.3 (+1.1)
CLICKER - *nc*	**50.8 (+1.6)**
CLICKER - *align+nc*	49.6 (+0.4)

Table 1. Accuracy on the test set of X-CSQA and X-CODAH in German. MCP(XLM-R$_B$)* model is trained in English and machine-translated German. The $align, diff, nc$ refer to the equation (4), (5), and (6), respectively.

the best accuracy achieved on two datasets, we conjecture the quality of commonsense embeddings (optimized by *align* and *diff* losses) dominates CSR on X-CSQA, while non-commonsense embeddings (optimized by *nc* loss) dominates that on X-CODAH. The reason for this may be that extracting commonsense knowledge for clause selection in X-CODAH is somewhat more challenging than multiple-choice QA in X-CSQA, whereas separating the non-commonsense embeddings help the multiple-choice classifier predict the remaining part with less noise. We also observe that using *align* and *nc* losses together is not the best practice according to our experiments. We interpret that the major knowledge in question answering is commonsense, in that the *align* loss dominates in the question-answering task of X-CSQA. However, separating the non-commonsense embeddings with *nc* loss dominates the sentence-selection task of X-CODAH.

Commonsense v.s. Non-commonsense. To investigate the effectiveness of our learned commonsense embeddings, we evaluate the accuracy of our CLICKER models on the X-CSQA dev set predicted by commonsense embeddings or non-commonsense embeddings. As seen in Table 2, the performance of commonsense embeddings is significantly better than that of non-commonsense embeddings. It is as expected, as our models are trained with cross-lingual CSR objectives to discern commonsense embeddings and trained with cross-lingual CSR objectives, while maximizing the similarity of non-commonsense embeddings. Non-commonsense embeddings can induce confusion for CSR, such that combining both embeddings performs worse than using commonsense embeddings only.

Does self-attention imply commonsense knowledge? We assume that self-attentions in our models can appropriately attend to tokens that affect the plausibility of commonsense. Figure 2 is the heatmap of the attention head in the self-attention layer evaluated on an example from X-CSQA. It's noteworthy to see that attention weights are given more to commonsense-related tokens, such as "before", "dream" and "sleep" tokens of the example "What do you need to be before you can dream?". A similar phenomenon is observed on X-CODAH as well. These attentions are weighted to create commonsense embeddings and help our models improve accuracy and interpretability on reasoning commonsense knowledge.

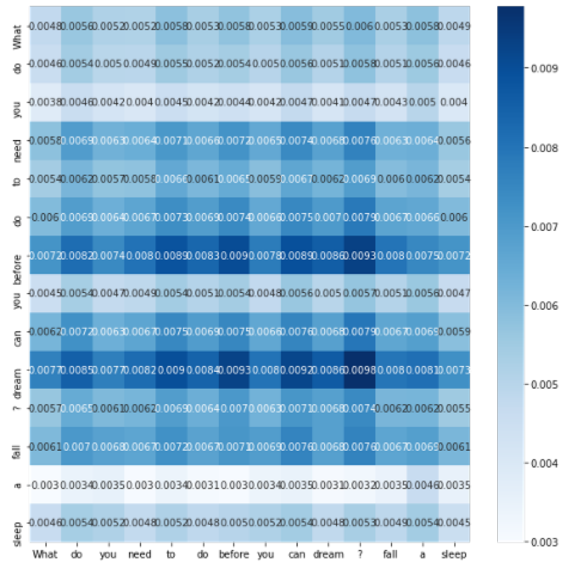

Fig. 2. Attention head of the self-attention layer. The given example is from X-CSQA.

Classifier Input	Dev Acc
Commonsense	**47.8**
Non-Commonsense	11.0
Commonsense + Non-Commonsense	47.6

Table 2. Dev accuracy on X-CSQA taking commonsense or non-commonsense embeddings as inputs for the classifier.

4. CONCLUSION

In this paper, we propose a cross-lingual framework CLICKER for commonsense reasoning. Experiments on X-CSQA and X-CODAH demonstrate the effectiveness of CLICKER in cross-lingual commonsense reasoning as it not only reduces performance discrepancies of commonsense reasoning between English and non-English languages but also improves the interpretability of commonsense knowledge across languages. The potential of our approach to be generalized to other low-resource languages will be beneficial for alleviating data scarcity in cross-lingual commonsense reasoning.

5. REFERENCES

[1] Ernest Davis and Gary Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence," *Communications of the ACM*, vol. 58, no. 9, pp. 92–103, 2015.

[2] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4149–4158, Association for Computational Linguistics.

[3] Shane Storks, Qiaozi Gao, and Joyce Y Chai, "Recent advances in natural language inference: A survey of benchmarks, resources, and approaches," *arXiv preprint arXiv:1904.01172*, 2019.

[4] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization," *CoRR*, vol. abs/2003.11080, 2020.

[5] Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen, "XCOPA: A multilingual dataset for causal commonsense reasoning," *CoRR*, vol. abs/2005.00333, 2020.

[6] Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren, "Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021)*, 2021.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[9] Guillaume Lample and Alexis Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[10] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019.

[11] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou, "InfoXLM: An information-theoretic framework for cross-lingual language model pre-training," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, June 2021, pp. 3576–3588, Association for Computational Linguistics.

[12] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, Association for Computational Linguistics.

[13] Jinyoung Yeo, Geungyu Wang, Hyunsouk Cho, Seungtaek Choi, and Seung-won Hwang, "Machine-translated knowledge transfer for commonsense causal reasoning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[14] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith, "Don't stop pretraining: adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.