The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Machine-Translated Knowledge Transfer for Commonsense Causal Reasoning

Jinyoung Yeo,[†] Geungyu Wang,[‡] Hyunsouk Cho,[†] Seungtaek Choi,[‡] Seung-won Hwang[‡]
[†]Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea
[‡]Yonsei Univeristy, Seoul, Republic of Korea
{jinyeo,prory}@postech.edu {posuer,hist0613,seungwonh}@yonsei.ac.kr

Abstract

This paper studies the problem of multilingual causal reasoning in resource-poor languages. Existing approaches, translating into the most probable resource-rich language such as English, suffer in the presence of translation and language gaps between different cultural area, which leads to the loss of causality. To overcome these challenges, our goal is thus to identify key techniques to construct a new causality network of cause-effect terms, targeted for the machine-translated English, but without any language-specific knowledge of resource-poor languages. In our evaluations with three languages, Korean, Chinese, and French, our proposed method consistently outperforms all baselines, achieving up-to 69.0% reasoning accuracy, which is close to the state-of-the-art accuracy 70.2% achieved on English.

Introduction

There is growing interest in the commonsense causal reasoning, to explain past observations or predict future events by understanding their general causal dependency. Recent efforts for such understanding are focused on measuring the plausibility of one event statistically leading to another, and are competing on Choice of Plausible Alternatives (COPA) evaluation (Roemmele, Bejan, and Gordon 2011), which is to select the more plausible alternative as a cause (or effect) of the premise as:

Example 1 *Premise:* The girl met her favorite actor. *What happened as an effect?*

Alternative 1: She went to see his new film.

Alternative 2: She asked him for his autograph.

For the purpose of this reasoning, the state-of-the-art, called CausalNet (Luo et al. 2016), harvests causality scores of cause-effect term pairs, *e.g.*, ('actor','autograph'), by mining their linguistic causal patterns, *e.g.*, "If...actor ..., then... autograph...", from an extremely large corpus. As a result, it achieved a remarkable accuracy (70.2%) from COPA. However, although this approach, building on a large corpus, shows higher reasoning accuracy as the corpus size grows, it is applicable only to resource-rich language (*e.g.*, English) and not to resource-poor language.

To overcome this challenge, we propose to train multilingual causal reasoning, *e.g.*, COPA task translated into non-English questions and answers, from English resources. We show that, even under extreme scenarios with no language-specific corpus, we can achieve comparable COPA accuracy using only English resources.

So far, most successful approaches on existing multilingual work such as question answering (Ture and Boschee 2016), sentiment classification (Zhou, Wan, and Xiao 2016), relation extraction (Faruqui and Kumar 2015) adopt *one-best MT* (1MT) which translates all contents into its most probable English via Machine Translation (MT) systems. Then, a certain multilingual task is performed on the translated English in the same manner with its monolingual task on English. Intuitively, this baseline may be strong for our work with the huge success of both English causal reasoning and Neural Machine Translation (NMT) (Johnson et al. 2016) adopted in off-the-shelf translators (*e.g.*, Google Translate[1], Microsoft Translator[2], and Naver Papago[3]).

However, 1MT does not fully capture causality in a target language, *e.g.*, its causal reasoning accuracy on the machine-translated English (ME) is up-to 64.2% below 70.2% on native English (NE). Our goal is thus to enable 1MT to achieve comparable reasoning performance. For that, we observe the following limitations in Example 1:

- **[L1] Translation gap:** Korean sentence [*Alternative 1*][4] is incorrectly translated into "She went to see a new release". Due to this error, its plausibility is weakened as the ME causality ('actor','release') has a lower causality score than the NE causality ('actor','film').
- **[L2] Language gap:** Korean sentence [*Alternative 2*] is correctly translated into "She asked him a signature" because 'autograph' and 'signature' map to the same word in Korean expressing both. As a result, the critical causality ('actor','autograph') is lost by 'signature' not frequently used around 'actor' in NE.

To bridge these gaps, we propose a framework, PSG, to (1) understand ME with the above gaps, and model such difference in both (2) word alignment and (3) score propagation to transfer causality knowledge from NE to ME. These

[1]https://translate.google.com
[2]https://translator.microsoft.com
[3]https://papago.naver.com
[4][*EN*] represents a foreign language sentence meaning "EN".

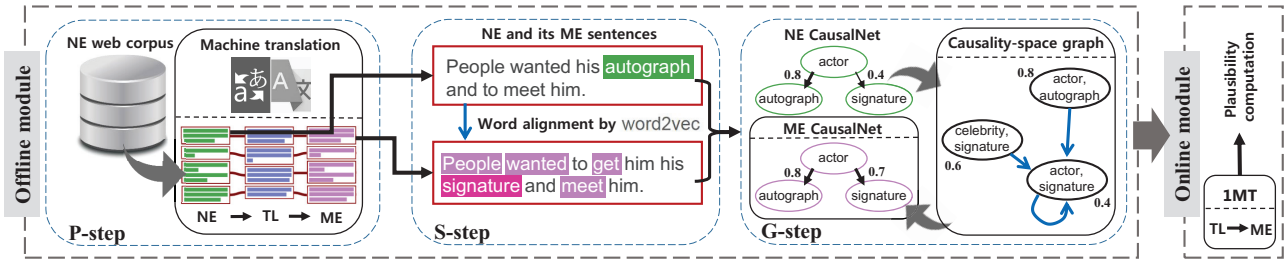

Figure 1: The sketch of PSG for multilingual commonsense causal reasoning

three requirements motivate three steps P-S-G respectively, as Figure 1 shows the overview of our reasoning framework. More specifically:

1. **Pseudo-parallel corpus generation:** First, we can simulate **L1** and **L2** on so-called *pseudo-parallel corpus*, which consists of English and its corresponding machine translation (to a target language, *e.g.*, Korean, then to English again, *e.g.*, Korean ME). One issue is whether the ME involves the standard and frequent expressions in a given target language. Our hypothesis is that NMT partly achieves perfect liberal translation such that the translated Korean includes native expressions in Korean, and is thus effective for looking up the difference of term occurrence and causality patterns between NE and ME.
2. **Selective word alignment:** From the NE-ME corpus, we can align their similar word pairs, taking account of **L1** and **L2**. For example, a NE term 'meet' can be trivially aligned as it appears in the exact same form in the ME sentence in Figure 1. Meanwhile, 'autograph' can be aligned to 'signature' but with lower *alignment confidence*, which indicates how reliably the NE term is aligned to a ME term to identify **L2**. That is, based on this confidence metric, we selectively extract high confidence of alignments, *e.g.*, 'autograph' to 'signature', while filtering incorrect alignments from **L1** such as 'film' to 'release'.
3. **Graph transformation:** Finally, we can transform the original NE CausalNet into a ME-friendly form, which has causality scores reducing the loss of causality from **L1** and **L2** between NE and ME. Specifically, we first transform CausalNet into an intermediate graph structure, called *causality-space graph*, of aggregating selective alignment results to propagate causality scores, *i.e.*, ('actor','autograph') to ('actor','signature'). Then, after such propagation on the structure, we finally transform the causality-space graph into ME CausalNet, with updated causality scores for ME, *e.g.*, 0.4 to 0.7.

Our framework PSG combining the above steps achieves the state-of-the-art accuracy of 69.0%, 66.2%, and 66.2%, while the baseline 1MT achieves 64.2%, 63.4%, and 60.4% accuracy, on the COPA evaluation for three multiple languages, Korean, Chinese, and French, respectively. Especially, the accuracy on Korean is close to 70.2% accuracy on NE using the original CausalNet, which can be roughly considered as the upper-bound performance of our problem.

Our main contributions are summarized as follows.

- To the best of our knowledge, this is the first work that investigates *multilingual commonsense causal reasoning* problem on any resource-poor language.
- We formalize machine-translated English with its intrinsic characteristics for which the use of 1MT can be improved for our research problem.
- We propose a novel framework called PSG that generates ME-friendly causality network of cause-effect terms with no extra knowledge on resource-poor language.
- We validate that our proposed method PSG outperforms all state-of-the-art baselines, consistently achieving the highest accuracy on multiple languages.

Preliminaries

English Causality Network: CausalNet

CausalNet is a weighted and directed graph $G(\mathbb{N}, \mathcal{C}_{\mathbb{N}}, W_{\mathbb{N}})$ with nodes (lemmatized English terms) $\mathbb{N} = \{n_1, n_2, ...\}$ and edges (causal relations) $\mathcal{C}_{\mathbb{N}}$. The edge weights are captured by the function $W_{\mathbb{N}} : \mathcal{C}_{\mathbb{N}} \rightarrow [0, 1]$. The weight $w_{i,j}$ associated with an edge (n_i, n_j) represents the causality score, denoted as $CS(n_i, n_j)$, of a cause n_i and an effect n_j. Causality scores depend on the number of occurrences that two terms n_i and n_j are in linguistic patterns known as causal cues (Chang and Choi 2004) identifying precise cause/effect roles, *e.g.*, "If...n_i ..., then... n_j..." and "... n_j..., because...n_i ...". That is, as more occurrences of (n_i, n_j) in causal cues, its causality score is higher as:

$$W_{\mathbb{N}} : w(n_i, n_j) = CS(n_i, n_j) \propto freq(n_i, n_j) \quad (1)$$

where $freq(n_i, n_j)$ is the frequency of observing the causal pair (n_i, n_j) from an English corpus. We omit the details of the list of causal cues and Eq. 1 in (Luo et al. 2016).

Despite building on a rather simple and shallow text analysis, by leveraging the scale and richness from a extremely large (10TB) text corpus, CausalNet achieves the state-of-the-arts accuracy on COPA tasks. The corpus contains 1.6B web pages, which result in 64,436 nodes in CausalNet.

Multilingual Commonsense Causal Reasoning

Commonsense causal reasoning problem is defined based on COPA, which consists of one thousand multiple-choice questions requiring causal reasoning to answer correctly. In

Example 1, given an English sentence pair of premise $\mathcal{N}_1$ and alternative $\mathcal{N}_2$, existing English causal reasoning work computes the plausibility $P(\mathcal{N}_1, \mathcal{N}_2)$ aggregating causality scores of all possible causal pairs between $\mathcal{N}_1$ and $\mathcal{N}_2$:

$$P(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{|\mathcal{N}_1| + |\mathcal{N}_2|} \sum_{n_i \in \mathcal{N}_1} \sum_{n_j \in \mathcal{N}_2} CS(n_i, n_j) \quad (2)$$

where $\mathcal{N}_k$ is segmented into lemmatized word terms, and the causality score $CS(n_i, n_j)$ is extracted from CausalNet.

In our work, this causal reasoning problem is extended into a multilingual setting, *e.g.*, COPA in another target language TL, for which the corpus availability is far more limited than in English. Naive solutions include (a) translating CausalNet for TL and (b) translating COPA tasks in TL into English and use English CausalNet. For (a), although there are many graph translation techniques (Feng et al. 2016; Sun, Hu, and Li 2017; Chen et al. 2017), we still need to match terms in the sentence with the translated graph, which requires language-specific tools, such as lemmatizer. We empirically validate this baseline fails to gain high reasoning accuracy due to this problem. In contrast, (b), known as 1MT, does not require such tools and thus has been more widely adopted in related problem settings (Ture and Boschee 2016). Specifically, given a non-English question, we convert the multilingual problem into a monolingual task by translating the question into ME by NMT[5], to use CausalNet. 1MT achieves higher accuracy in our evaluation.

Characteristics of Machine-Translated English

Regarding 1MT, its success depends on the assumption that machine translation preserves causality. Our research question is thus, is this assumption valid? In Figure 2, to analyze this, we translate 1,500 Korean and Chinese sentences of the same meaning into English, respectively, by NMT, and label all translation results into three categories, wrong, direct, and liberal translation.

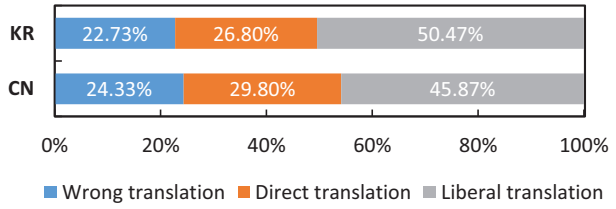

Figure 2: Distribution of machine translation to English

As a result, we find that common characteristics of MEs from any foreign language are mainly two types: The first one is the translation gap (**L1**) resulted from average 23.53% of wrong translation. Despite the recent improvement of NMT (Johnson et al. 2016) on resource-poor languages, it is reported that machine cannot win human yet in this dimension, which suggests the inherent occurrences of machine-translation error. The second one is the language gap (**L2**) observed when translating between two languages of different culture (*e.g.*, Asia and Europe). Especially, this gap is multiplied with average 28.30% of direct translation with infrequent expressions, *e.g.*, Korean term [*Morning call*] is translated into 'Morning call', while being reduced with its liberal translation 'Wake-up call'.

These gaps suggest using CausalNet trained from NE may harm causality inference on ME due to different term occurrence patterns, as we later confirm empirically as well. As a possible solution, such differences of term occurrences may be alleviated by annotating synonym relation (Qu, Ren, and Han 2017) between NE and ME words. However, we find that not all synonyms are interchangeable to represent the same causality context, for example 'film' and 'picture', and they can often multiply translation gaps in case of incorrect translation, for example, 'release' and 'publication'. We empirically validate considering these synonyms for reasoning is not effective later. In contrast, we propose a more systematic way to annotate the bias introduced from **L1** and **L2**, so that we can factor in such relations in causality computation.

Our Approach

This section develops our systematic framework PSG designed to generate a ME-friendly CausalNet suitable for ME causal reasoning. For that, PSG aims to transfer causality knowledge from NE to ME, following the three main steps, P-S-G, in Algorithm 1. We consider only English text corpus and CausalNet, which are publicly available, as input.

Algorithm 1 Offline module in PSG

Require: corpus $\mathcal{D}_{NE}$: NE text corpus
Require: graph G: NE CausalNet
Ensure: graph G': ME CausalNet
1: $\mathcal{D}_{ME} \leftarrow$ P-Step($\mathcal{D}_{NE}$)
2: $\mathbb{A} \leftarrow$ S-Step($\mathcal{D}_{NE}, \mathcal{D}_{ME}$)
3: $G' \leftarrow$ G-Step($G, \mathbb{A}$)
4: **return** G'

P-Step: Pseudo-Parallel Corpus Generation

We first propose *Pseudo-parallel corpus* (PP-corpus), which simulates and looks up **L1** and **L2** between NE and ME. Formally, let $\mathcal{D}_{NE}$ be a NE corpus, each of which is a NE sentence $\mathcal{N}_i$, *i.e.*, $\mathcal{D}_{NE} = \{\mathcal{N}_i | 1 \leq i \leq |\mathcal{D}_{NE}|\}$. Then, we generate a corpus $\mathcal{D}_{TL} = \{\mathcal{T}_i | \mathcal{N}_i \Rightarrow_t \mathcal{T}_i; 1 \leq i\}$ as PP-corpus with $\mathcal{D}_{NE}$, where $\mathcal{N}_i \Rightarrow_t \mathcal{T}_i$ represents the machine translation from a NE sentence $\mathcal{N}_i$ to a TL sentence $\mathcal{T}_i$.

This step would be redundant if machine translation to TL does not convey the sense of the original English sentences with natural expressions of TL, *i.e.*, direct translation. However, as analyzed in Figure 2, NMT is capable of achieving about 50% of perfect liberal translation, from which we can quantify **L1** and **L2** after the liberal translation to TL is translated back to ME. For example, by liberal translation to TL, "We had lunch" $\Rightarrow_t$ [*We ate lunch*] in Korean, in which [*eat*] is far more natural than [*have*] with [*lunch*], and "I asked for a wake-up call" $\Rightarrow_t$ [*I asked for a morning call*] in Korean, in which Korean people do not use a term

[5]We employ Naver Papago for all translations in this paper.

[*wake-up call*]. Then, by direct translation to ME, "We had lunch" $\Rightarrow_t$ [*We ate lunch*] $\Rightarrow_t$ "We ate lunch". In this sense, we generate a ME corpus $\mathcal{D}_{ME}$ to match with $\mathcal{D}_{NE}$ as:

$$\mathcal{D}_{ME} = \{\mathcal{M}_i | \mathcal{N}_i \Rightarrow_t \mathcal{T}_i; \mathcal{T}_i \Rightarrow_t \mathcal{M}_i; 1 \leq i\} \quad (3)$$

The subsequent S-step covers each corresponding sentence pair between $\mathcal{D}_{NE}$ and $\mathcal{D}_{ME}$.

S-Step: Selective Word Alignment

As for the second step, we propose *selective word alignment* to identify different term occurrence patterns between NE and ME. The advantage of using ME corpus is that we can leverage state-of-the-art tools for English regardless of a given target language, while its disadvantage is that we need to filter out causality errors due to **L1** and **L2**, specifically by "alignment" reflecting translation confidence.

At word level, a straightforward method for alignment is using word similarity metrics based on neural embedding model such as Word2Vec (Mikolov et al. 2013). Formally, given a NE sentence $\mathcal{N} = \{n_1, n_2, ..., n_{|\mathcal{N}|}\}$ and its ME sentence $\mathcal{M} = \{m_1, m_2, ..., m_{|\mathcal{M}|}\}$ on PP-corpus, the alignment probability of a pair (n_i, m_j) can be computed as:

$$Pr(m_j|n_i) = \frac{sim(n_i, m_j)}{\sum_{m_k \in \mathcal{M}} sim(n_i, m_k)} \quad (4)$$

where $sim(n_i, m_j)$ is word similarity by Word2Vec model. Based on $Pr(m_j|n_i)$, we can intuitively align word pairs.

$$[n_i \Rightarrow_a m_j^*] = \arg\max_{m_j \in \mathcal{M}} Pr(m_j|n_i) \quad (5)$$

where $[n_i \Rightarrow_a m_j^*]$ represents term n_i is aligned to term m_j^*. However, not always the maximum alignment probabilities guarantee correct alignments. We thus measure the alignment confidence of a given n_i that represents how reliably n_i can be aligned with $\mathcal{M}$. Let $\Theta(n_i, \mathcal{M})$ be the distribution of $Pr(m_j|n_i)$ varying ME word $m_j \in \mathcal{M}$, *i.e.*, $\Theta(n_i, \mathcal{M}) = \{Pr(m_j|n_i) \mid j = 1, .., |\mathcal{M}|\}$. As $\Theta(n_i, \mathcal{M})$ is more different from an uniform distribution, n_i is likely to be accurately replaced by its matched ME word having the maximum alignment probability. Therefore, we define the alignment confidence (denoted as AC) by using Shannon Wavelet Entropy (Rosso et al. 2001) (denoted as $H(\cdot)$), which measures the entropy of $\Theta(n_i, \mathcal{M})$.

$$\begin{aligned} AC(n_i) &= 1 - H(\Theta(n_i, \mathcal{M})) \\ &= 1 + \sum_{\forall j} Pr(m_j|n_i) log Pr(m_j|n_i) \end{aligned} \quad (6)$$

As $\Theta(n_i, \mathcal{M})$ is closer to uniform distribution making its alignment decision difficult, $AC(n_i)$ decreases with increasing $H(\Theta(n_i, \mathcal{M}))$. For example, in Figure 3, 'detention' ($AC = 0.28$) is likely to be more clearly aligned with its strong confidence than 'received' ($AC = 0.21$) and 'his' ($AC = 0.18$). Using this confidence measure, given a NE sentence $\mathcal{N}$ and its ME sentence $\mathcal{M}$, we accurately extract word-level alignments $(\mathcal{N} \times \mathcal{M})$ as follows:

$$(\mathcal{N} \times \mathcal{M}) = \{a_{ij} | 1 \leq i \leq |\mathcal{N}|; 1 \leq j \leq |\mathcal{M}|; AC(n_i) > \theta\} \quad (7)$$

$\mathcal{N}$: "The student received detention by his teacher"
$\mathcal{M}$: "The student was punished by the teacher"

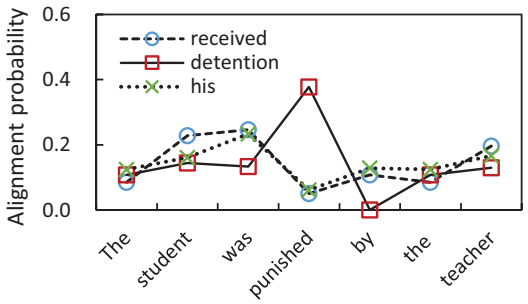

Figure 3: Distribution example of alignment probabilities

where an alignment a_{ij} represents that n_i is aligned to m_j, and θ is a pre-defined threshold. We empirically analyze the effect of varying this threshold later.

However, the word-level alignment is insufficient, for example, ['detention'$\Rightarrow_a$'punish'] has different causal meaning between around 'teacher' or around 'police'. That is, the alignments between single words may be too ambiguous to represent accurate causality context. Therefore, beyond word-level, we more selectively align words by disambiguating their causality context. Let $\mathcal{P}$ be a set of pairs of NE sentences $\mathcal{N}_1$ ($\Rightarrow_t \mathcal{M}_1$) and $\mathcal{N}_2$ ($\Rightarrow_t \mathcal{M}_2$) on PP-corpus with cause and effect roles, respectively, in causal relation, *e.g.*, "If $\mathcal{N}_1$, then $\mathcal{N}_2$" (Luo et al. 2016). Then, we define the word-pair level alignments $\mathbb{A}$ using $\mathcal{P}$ as:

$$\begin{aligned} \mathbb{A} = \{ &[(n_i, n_j) \Rightarrow_a (m_k, m_q)] \\ &| a_{ik} \in (\mathcal{N}_1 \times \mathcal{M}_1); a_{jq} \in (\mathcal{N}_2 \times \mathcal{M}_2); (\mathcal{N}_1, \mathcal{N}_2) \in \mathcal{P}\} \end{aligned} \quad (8)$$

where $\mathbb{A}$ permits duplicate alignments for the next G-step.

G-Step: Graph Transformation

As for the third step, to reduce the gap of different causality context between NE and ME, we propose *graph transformation*, which converts NE CausalNet into ME CausalNet. For that, this step propagates causality scores from NE to ME by aggregating the alignment results in a graphical way.

The first stage of graph transformation is that we construct the graph structure that takes advantages for the causality propagation. We refer to this as *causality-space graph*. Formally, given a NE CausalNet $G(\mathbb{N}, \mathcal{C}_{\mathbb{N}}, W_{\mathbb{N}})$ and a set $\mathbb{A}$ of causality alignments, inspired by graph space transformation (Evans and Lambiotte 2009), we can construct its causality-space graph $\mathcal{G}(\mathbb{C}, \mathcal{E}, \mathcal{W})$ such that:

- A node $c_{n_i,n_j} \in \mathbb{C}$ of $\mathcal{G}$ represents the edge $c_{(n_i,n_j)}$ between the causal pair (n_i, n_j) of CausalNet G;
- A node c_i is adjacent to a node c_j (including c_i) in $\mathcal{G}$ if and only if c_i is aligned to c_j at least once in $\mathbb{A}$; and
- The weight $\mathcal{W} : \bar{\omega}(c_i, c_j)$ on the edge$(c_i, c_j) \in \mathcal{E}$ is assigned by the number of alignments $[c_i \Rightarrow_a c_j]$ in $\mathbb{A}$.

Algorithm 2 Transformation to causality-space graph

Require: graph $G(\mathbb{N}, \mathcal{C}_{\mathbb{N}}, W_{\mathbb{N}})$: NE CausalNet
Require: set $\mathbb{A}$: Word-pair alignments
Ensure: graph $\mathcal{G}(\mathbb{C}, \mathcal{E}, \mathcal{W})$: Causality-space graph
1: /* create the nodes of $\mathcal{G}$ */
2: **for** each $(n_i, n_j) \in \mathcal{C}_{\mathbb{N}}$ **do**
3: $\quad \mathbb{C} \leftarrow \mathbb{C} \cup c_{(n_i,n_j)}$
4: **end for**
5: /* create the edges of $\mathcal{G}$ */
6: **for** each $c_i \in \mathbb{C}$ **do**
7: $\quad$ **for** each $c_j \in \mathbb{C}$ **do**
8: $\quad\quad$ **if** $[c_i \Rightarrow_a c_j] \in \mathbb{A}$ **then**
9: $\quad\quad\quad \mathcal{E} \leftarrow \mathcal{E} \cup (c_i, c_j)$
10: $\quad\quad$ **end if**
11: $\quad$ **end for**
12: **end for**
13: /* assign a weight to each edge of $\mathcal{G}$ */
14: **for** each $(c_i, c_j) \in \mathcal{E}$ **do**
15: $\quad \mathcal{W} : \bar{\omega}(c_i, c_j) \leftarrow \frac{\#[c_i \Rightarrow_a c_j]}{\sum_{c_k \in N(c_j)} \#[c_k \Rightarrow_a c_j]}$ // Eq. 9
16: **end for**
17: **return** $\mathcal{G}(\mathbb{C}, \mathcal{E}, \mathcal{W})$

Algorithm 3 Transformation to ME CausalNet

Require: graph $G(\mathbb{N}, \mathcal{C}_{\mathbb{N}}, W_{\mathbb{N}})$: NE CausalNet
Require: graph $\mathcal{G}(\mathbb{C}, \mathcal{E}, \mathcal{W})$: Causality-space graph
Ensure: graph $G'(\mathbb{M}, \mathcal{C}_{\mathbb{M}}, W_{\mathbb{M}})$: ME CausalNet
1: /* create the nodes and edges of G' */
2: initialize $G'(\mathbb{M} \leftarrow \mathbb{N}, \mathcal{C}_{\mathbb{M}} \leftarrow \mathcal{C}_{\mathbb{N}}, W_{\mathbb{M}} \leftarrow W_{\mathbb{N}})$
3: **for** each $[c_i \Rightarrow_a c_j] \in \mathbb{A}$ **do**
4: $\quad (m_k, m_q) \leftarrow c_j$
5: $\quad \mathbb{M} \leftarrow \mathbb{M} \cup \{m_k, m_q\}$
6: $\quad \mathcal{C}_{\mathbb{M}} \leftarrow \mathcal{C}_{\mathbb{M}} \cup (m_k, m_q)$
7: **end for**
8: /* propagate causality scores through $\mathcal{G}$ */
9: initialize $\forall c_i \in \mathbb{C} : CS^{(0)}(c_i) \leftarrow W_{\mathbb{N}} : w(c_i)$
10: **while** $W_{\mathbb{M}}^{(z+1)} \neq W_{\mathbb{M}}^{(z)}$ **do**
11: $\quad$ **for** each $c_i \in \mathbb{C}$ **do**
12: $\quad\quad CS^{(z+1)}(c_i) \leftarrow \sum_{c_j \in N(c_i)} \bar{\omega}(c_j, c_i) \cdot CS^{(z)}(c_j)$
13: $\quad$ **end for**
14: **end while**
15: /* assign a weight to each edge of G' */
16: **for** each $c_i \in \mathbb{C}$ **do**
17: $\quad W_{\mathbb{M}} : w(c_i) \leftarrow CS^{(*)}(c_i)$ // Eq. 11 and 12
18: **end for**
19: **return** $G'(\mathbb{M}, \mathcal{C}_{\mathbb{M}}, W_{\mathbb{M}})$

For example, as shown in Figure 1, the edge ('actor', 'autograph') in NE CausalNet corresponds to the node ('actor', 'autograph') in causality-space graph. The weight of the edge between $c_{(n_i,n_j)}$ and $c_{(n_k,n_q)}$ is calculated by statistics of alignments $\mathbb{A}$ as:

$$\bar{\omega}(c_{(n_i,n_j)}, c_{(n_k,n_q)}) = \frac{\#[(n_i, n_j) \Rightarrow_a (n_k, n_q)]}{\sum_{c_{(n'_i,n'_j)} \in \mathbb{C}} \#[(n'_i, n'_j) \Rightarrow_a (n_k, n_q)]} \quad (9)$$

where $\#[c_i \Rightarrow_a c_j]$ is the number of alignments $[c_i \Rightarrow_a c_j]$ in $\mathbb{A}$. For efficiency, $\mathbb{C}$ can be replaced by $N(c_{(n_k,n_q)})$ representing the neighbors adjacent to $c_{(n_k,n_q)}$. In Eq. 9, we also take account of the self alignments that have $n_i = n_k$ and $n_j = n_q$, *i.e.*, NE and ME have the same causality pattern. Algorithm 2 describes the details of constructing the causality-space graph $\mathcal{G}$, but we omit detailed explanations.

The second stage of graph transformation consists of propagating causality scores through $\mathcal{G}$ and reconstructing a ME-friendly CausalNet $G'(\mathbb{M}, \mathcal{C}_{\mathbb{M}}, W_{\mathbb{M}})$. Algorithm 3 describes its detailed procedure. First, ME CausalNet G' is initialized by G to share the same node and edge sets and their weight function (line 2). Then, we update the nodes $\mathbb{M}$ and edges $\mathcal{C}_{\mathbb{M}}$ with additional terms in ME (line 3-7). For score propagation, we adopt an iterative score propagation method (Qin et al. 2005) for simplicity (line 9-14):

$$CS^{(z+1)}(c_i) = \sum_{c_j \in N(c_i)} \bar{\omega}(c_j, c_i) \cdot CS^{(z)}(c_j) \quad (10)$$

where $\sum_{c_j \in N(c_i)} \bar{\omega}(c_j, c_i) = 1$ in every z-th iteration. This procedure is repeated until the scores are converged or predefined maximum iterations (line 10).

For implementation, we use the much smaller text corpus $\mathcal{D}_{NE}$ for alignment, compared to the 10TB text corpus for NE CausalNet construction. Considering this case, we adopt a simple smoothing method of linear interpolation between the original score $CS^{(0)}$ and the converged score $CS^{(*)}$:

$$CS^{(*)}(c_i) \approx \begin{cases} (1-\lambda) CS^{(0)}(c_i) + \lambda\, CS^{(*)}(c_i) & \text{if } N(c_i) \neq \{c_i\} \\ CS^{(0)}(c_i) & \text{otherwise (no alignment)} \end{cases} \quad (11)$$

where $\lambda \in [0, 1]$ is used to adjust to tradeoff between $CS^{(0)}$ and $CS^{(*)}$. We empirically analyze the effect of varying this systematic parameter later. If the new causality scores are determined, the weight $w(m_i, m_j)$ associated with an edge (m_i, m_j) is assigned with the causality score $CS^{(*)}(m_i, m_j)$ of ME (line 16-18):

$$W_{\mathbb{M}} : w(m_i, m_j) = CS^{(*)}(m_i, m_j) \quad (12)$$

Once G-step constructs ME CausalNet G' for a target foreign language on the offline module, as shown in Figure 1, we can perform the reasoning task (*i.e.*, plausibility computation) on the online module by cooperating with 1MT.

Experimental Evaluation

Experiment Setup

Datasets To validate the effectiveness and robustness of our proposed method, we select three target languages, Korean, Chinese, and French, to cover diverse cultural and linguistic characteristics. We manually translate COPA dataset, *i.e.*, 1,000 commonsense causal reasoning questions, into each language, dividing into development and test question set of 500 each.

As additional development datasets for PSG, we leverage CausalNet[6] and about 1M English web pages for word alignment. One implementation issue is that the reasoning coverage of PSG depends on the alignment datasize. To resolve this scale constraint, we first identify ME terms translated from our COPA data by NMT, then intensively search for the articles that include NE terms translated into the ME terms of interest.

[6]NE CausalNet: https://cs-zyluo.github.io/CausalNet

Table 1: Accuracy and its benchmarking ratio (%)

Method	Korean	Chinese	French
TransCP	.568 (80.9%)	.572 (81.5%)	.560 (79.8%)
TransCN	.574 (81.8%)	.580 (82.6%)	.568 (80.9%)
TransQA	.642 (91.5%)	.634 (90.3%)	.604 (86.0%)
TransQA+	.546 (77.9%)	.570 (81.2%)	.580 (82.6%)
PSG	**.690** (98.3%)	**.662** (94.3%)	**.662** (94.3%)

Evaluation measure As a main evaluation metric, we use the accuracy adopted in much reasoning work as:

$$\text{Accuracy} = \frac{\#\text{correctly answered questions}}{\#\text{answered questions}} \quad (13)$$

Considering our work depends on the achievement on English, we also design another evaluation measure, *Benchmarking ratio*, to quantify relative improvements of multilingual reasoning, which we define as:

$$\text{Benchmarking ratio} = \frac{\text{Accuracy}}{\text{Accuracy}_{EN}} \quad (14)$$

where Accuracy_{EN} is the English reasoning accuracy.

Baselines We compare PSG with the following four baseline methods, but using no NLP tool on any target language:

- **TransCP (Luo et al. 2016):** This method leverages PP-corpus to reconstruct CausalNet for ME in the same manner of NE CausalNet, then performs the causal reasoning on the machine-translated questions by 1MT.
- **TransCN (Chen et al. 2017):** This method translates NE CausalNet from English to a target foreign language, projecting the causality scores by identifying corresponding causality pairs with graph alignment techniques.
- **TransQA (Ture and Boschee 2016):** This method is a standard 1MT approach to perform the causal reasoning on the machine-translated questions, then leverage the original NE CausalNet to extract causality scores.
- **TransQA+ (Qu, Ren, and Han 2017):** This method extends TransQA with a synonym extractor to update the plausibility by replacing a score of each causal pair with the average score of its top-3 synonymous causal pairs.

Evaluation Results and Discussion

We investigate the empirical findings for the following research questions:
RQ1: Does our framework outperform the baselines?
RQ2: How does our framework capture ME causality?
RQ3: What causes the error? How to recover them?

Baseline comparison (RQ1) Table 1 shows that our proposed method PSG consistently outperforms all the above baselines in the three languages. Especially, in Korean, the benchmarking ratio is 98.3%, which means that the reasoning accuracy of foreign language is almost close to that of English. Regarding these improvements, we describe our strength over each of the baselines.

First, while TransCP suffers from not much large corpus generating a low-coverage ME CausalNet, PSG generates high-quality ME CausalNet transformed from the original CausalNet. Also, unlike PSG, TransCN leverages CausalNet of a target language, namely TL CausalNet, which has difficulty in matching corresponding terms between questions and the TL CausalNet due to absence of lemmatization. Using these low-quality CausalNets of TransCP and TransCN does not lead to high reasoning accuracy.

Compared to TransCP and TransCN, TransQA achieves higher accuracy by using NE CausalNet, which suggests that adopting 1MT with NE CausalNet is a better starting point as stated in our problem definition. However, because TransQA does not consider the challenges **L1** and **L2**, there is a significant performance gap between TransQA and PSG, which indicates the impact of reducing the penalty from **L1** and **L2**. To overcome the challenges, TransQA+ additionally considers synonyms of ME words. Nevertheless, it fails to improve TransQA because causality scores of synonyms often disqualify the plausibility for accurate reasoning with term ambiguity. For example, given a ME alternative "She jumped off the plane", 'plane' is extended with not only a synonym 'aircraft' of the same causality context but also another synonym 'flat' of different noise context.

Component study (RQ2) To see how PSG achieves such improvements, this section investigates the effectiveness of components, P-S-G, overcoming the limitations on ME. Due to lack of space, we cover only Korean and Chinese.

For P-step, we analyze the effect of corpus size on word alignment in Figure 4, where we randomly sample a part of our web corpus and thus perform PSG per different size of corpus. One can observe a positive correlation between the size of the data and the ability to reason about commonsense causality: R-squared is 0.91 and 0.92 for Korean and Chinese, respectively. This correlation indicates our PP-corpus is effective to simulate and look up **L1** and **L2** to improve the reasoning performance. Also, the monotonic increase of accuracy suggests that more alignment data may contribute to increase the accuracy even higher.

For SG-steps, Figure 5 shows the average reasoning accuracy of Korean and Chinese when varying threshold θ and weight parameter λ. Pearson correlation coefficient (PCC) between Korean and Chinese results varying the parameters is 0.702. In this heatmap, we can make the following three

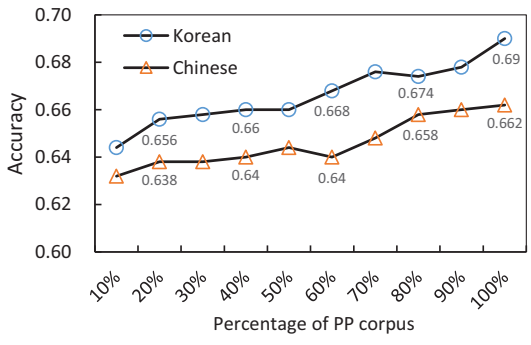

Figure 4: The effect of varying pseudo-parallel corpus size

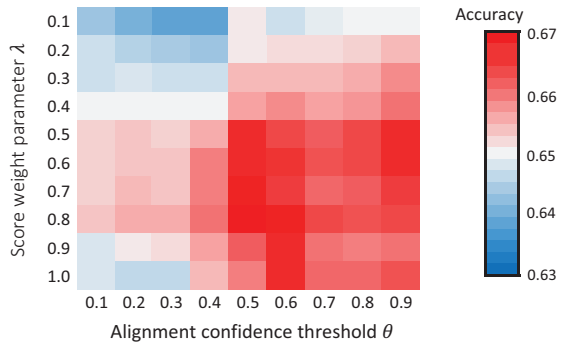

Figure 5: Influence of threshold θ and weight parameter λ

observations: (i) Regarding θ, the narrow range around 0.5 to 0.6 is optimal. However, the accuracy of PSG, by effectively selecting word alignments, remains rather insensitive to this value. (ii) Regarding λ, its optimal values are focused on the range around 0.5 to 0.8, which indicates that the updated causality scores by propagation are more effective for ME than the original scores. (iii) Supervised learning of these parameters is reliable as such trends are commonly found in two languages with high positive PCC.

Lastly, for G-step, Figure 6 shows that the accuracy of PSG is improved by performing iterations of causality propagation in Eq 11. This means that the causality-space graph is suitable to propagate causality knowledge. As a result, PSG is converged to the accuracy of Table 1 within the small number of iterations, *i.e.*, around 9-th iteration. The worse improvement of Chinese through iterations is resulted from less accurate translation to ME (indicated by accuracy of TransQA in Table 1), which complicates the next step of word alignment. We omit the evaluation of runtime, as G-step is performed offline.

Error analysis (RQ3) We investigate the cases where the inference was inaccurate in PSG, and find that negative expressions, *i.e.*, 'not A', are more generated by MT, which causes the loss of causality as CausalNet does not consider the sentence context in Eq 2.

Example 2 *Premise:* The table wobbled. *What is its cause?*
Alternative 1: The floor was uneven.
Alternative 2: The floor was slippery.

In Example 2, Korean and Chinese sentences, [*The floor was uneven*], are translated into a ME sentence "The floor was not flat" with lower causality, *e.g.*, $\frac{CS(\text{'uneven'},\text{'wobble'})}{CS(\text{'flat'},\text{'wobble'})} = 3533$. However, this problem can be resolved by replacing 'not A' by 'B' that is the antonym of 'A', by existing antonym detector (Ono, Miwa, and Sasaki 2015). This solution improves the reasoning accuracy by 0.010, 0.014, and 0.004 in Korean, Chinese, and French, respectively.

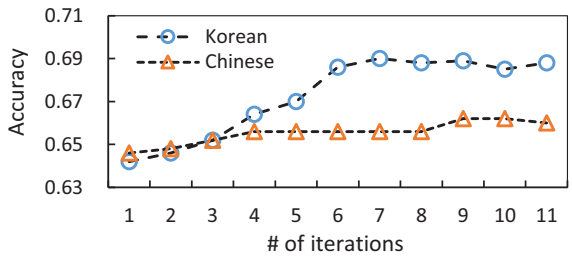

Figure 6: Convergence of PSG

Related Work

Our work can also be viewed as a domain adaptation, *e.g.*, English domain to non-English domain. For example, sentiment classifier trained for movie domain, can adapt to classify for book domain, by identifying common or pivot features of two domains (Andreevskaia and Bergler 2008; Bollegala, Weir, and Carroll 2011; He, Lin, and Alani 2011; Pan and Yang 2010; Li et al. 2013; 2012; Xia and Zong 2011; Yoshida et al. 2011; Wu, Tan, and Cheng 2009; Liu 2012). However, these methods do not accumulate domain-specific knowledge, as we do. Though (Chen, Ma, and Liu 2015) similarly aims for accumulating the difference of two domains, in terms of language models, or distributions, such difference is more complicated in our target problem of adapting between graphs.

When dealing with multilingual collections, most prior approaches (Ture and Boschee 2016; Faruqui and Kumar 2015; Mihalcea, Banea, and Wiebe 2007; Banea et al. 2008) translate all text into English beforehand, then treat the task as monolingual retrieval (previously referred to as 1MT). Such prior work does not focus on the translation component, which is a black box. More closely related work is transferring POS tagging in two graphs (Das and Petrov 2011; Kim and Lee 2012). Commonality is label propagation for such transfer, but we distinguish by combining machine translation and label propagation. Specifically, we generate the projection space of label propagation on "back-translated English" (*i.e.*, ME from PP-corpus), which is adopted in recent paraphrasing work (Mallinson, Sennrich, and Lapata 2017; Wieting, Mallinson, and Gimpel 2017) with the success of NMT. Additionally, we also have another challenge of recomputing the score for the target language (or ME) structure.

Conclusion

We studied the commonsense causal reasoning problem for any resource-poor language. Our proposed approach significantly improves reasoning accuracy in multiple languages, Korean, Chinese, and French, by reducing the causality loss with respect to translation and language gaps. We demonstrate, for the first time, that the cross-lingual knowledge transfer between NE and ME can be an effective technique in growing the causal reasoning ability on resource-poor languages without any target language corpus.

Acknowledgement

This work was supported by Microsoft Research, and Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT)

(No.2017-0-01778,Development of Explainable Human-level Deep Machine Learning Inference Framework). S. Hwang is a corresponding author.

References

Andreevskaia, A., and Bergler, S. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *ACL*.

Banea, C.; Mihalcea, R.; Wiebe, J.; and Hassan, S. 2008. Multilingual subjectivity analysis using machine translation. In *EMNLP*.

Bollegala, D.; Weir, D.; and Carroll, J. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *ACL*.

Chang, D.-S., and Choi, K.-S. 2004. Causal relation extraction using cue phrase and lexical pair probabilities. In *ICON*.

Chen, M.; Tian, Y.; Yang, M.; and Zaniolo, C. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI*.

Chen, Z.; Ma, N.; and Liu, B. 2015. Lifelong learning for sentiment classification. In *ACL*.

Das, D., and Petrov, S. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*.

Evans, T., and Lambiotte, R. 2009. Line graphs, link partitions, and overlapping communities. *Physical Review E* 80(1):016105.

Faruqui, M., and Kumar, S. 2015. Multilingual open relation extraction using cross-lingual projection. In *ACL*.

Feng, X.; Tang, D.; Qin, B.; and Liu, T. 2016. English-chinese knowledge base translation with neural network. In *COLING*.

He, Y.; Lin, C.; and Alani, H. 2011. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *ACL*.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Kim, S., and Lee, G. G. 2012. A graph-based cross-lingual projection approach for weakly supervised relation extraction. In *ACL*.

Li, F.; Pan, S. J.; Jin, O.; Yang, Q.; and Zhu, X. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *ACL*.

Li, S.; Xue, Y.; Wang, Z.; and Zhou, G. 2013. Active learning for cross-domain sentiment classification. In *IJCAI*.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

Luo, Z.; Sha, Y.; Zhu, K. Q.; Hwang, S.-w.; and Wang, Z. 2016. Commonsense causal reasoning between short texts. In *KR*.

Mallinson, J.; Sennrich, R.; and Lapata, M. 2017. Paraphrasing revisited with neural machine translation. In *EACL*.

Mihalcea, R.; Banea, C.; and Wiebe, J. 2007. Learning multilingual subjective language via cross-lingual projections. In *ACL*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Ono, M.; Miwa, M.; and Sasaki, Y. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *HLT-NAACL*.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *TKDE* 22(10):1345–1359.

Qin, T.; Liu, T.-Y.; Zhang, X.-D.; Chen, Z.; and Ma, W.-Y. 2005. A study of relevance propagation for web search. In *SIGIR*.

Qu, M.; Ren, X.; and Han, J. 2017. Automatic synonym discovery with knowledge bases. In *KDD*.

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium*.

Rosso, O. A.; Blanco, S.; Yordanova, J.; Kolev, V.; Figliola, A.; Schürmann, M.; and Başar, E. 2001. Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *Journal of neuroscience methods* 105(1):65–75.

Sun, Z.; Hu, W.; and Li, C. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. *arXiv preprint arXiv:1708.05045*.

Ture, F., and Boschee, E. 2016. Learning to translate for multilingual question answering. *EMNLP*.

Wieting, J.; Mallinson, J.; and Gimpel, K. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *EMNLP*.

Wu, Q.; Tan, S.; and Cheng, X. 2009. Graph ranking for sentiment transfer. In *ACL-IJCNLP*.

Xia, R., and Zong, C. 2011. A pos-based ensemble model for cross-domain sentiment classification. In *IJCNLP*.

Yoshida, Y.; Hirao, T.; Iwata, T.; Nagata, M.; and Matsumoto, Y. 2011. Transfer learning for multiple-domain sentiment analysis-identifying domain dependent/independent word polarity. In *AAAI*.

Zhou, X.; Wan, X.; and Xiao, J. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *ACL*.