Cross-lingual Knowledge Projection Using Machine Translation and Target-side Knowledge Base Completion

Naoki Otani*
Carnegie Mellon University
notani@cs.cmu.edu

Hirokazu Kiyomaru
Kyoto University
kiyomaru@nlp.ist.i.kyoto-u.ac.jp

Daisuke Kawahara
Kyoto University
dk@i.kyoto-u.ac.jp

Sadao Kurohashi
Kyoto University
kuro@i.kyoto-u.ac.jp

Abstract

Considerable effort has been devoted to building commonsense knowledge bases. However, they are not available in many languages because the construction of KBs is expensive. To bridge the gap between languages, this paper addresses the problem of projecting the knowledge in English, a resource-rich language, into other languages, where the main challenge lies in projection ambiguity. This ambiguity is partially solved by machine translation and target-side knowledge base completion, but neither of them is adequately reliable by itself. We show their combination can project English commonsense knowledge into Japanese and Chinese with high precision. Our method also achieves a top-10 accuracy of 90% on the crowdsourced English–Japanese benchmark. Furthermore, we use our method to obtain 18,747 facts of accurate Japanese commonsense within a very short period.

1 Introduction

Commonsense has been considered to play a vital role in language understanding (LoBue and Yates, 2011), and considerable effort has been devoted to building knowledge bases (KBs) that organize commonsense (Zang et al., 2013). The largest multi-lingual commonsense KB is ConceptNet (Liu and Singh, 2004b). ConceptNet maintains knowledge as a triple of two concepts and relationship between them, which we call *fact*. The characteristic of ConceptNet is that concepts are represented in undisambiguated forms of words or phrases, which facilitates commonsense acquisition and inference in practice (Liu and Singh, 2004a) and recently has proven useful for building word representations (Speer et al., 2017; Camacho-Collados et al., 2017). We target ConceptNet in this paper.

A major problem lies in a large gap of quantity and quality between languages. The latest release (v5.5.0) of ConceptNet has 2,828,394 unique English facts[1], but the number of Japanese facts in ConceptNet is only 69,902 ($\approx 2.5\%$) even though Japanese knowledge takes up the eighth-largest portion of the database. This problem is not specific to ConceptNet. English KBs are typically larger and of higher quality than other languages. Although an adequate amount of knowledge of named entities is often available in many languages thanks to semi-structured text on the web such as Wikipedia infobox (Lehmann et al., 2014), commonsense is hard to obtain due to the lack of tractable and objective information (Gordon and Van Durme, 2013).

It is not realistic to develop large knowledge resources in every language from scratch because of cost constraints. Instead, this paper focuses on cross-lingual knowledge projection. We translate English commonsense facts into a target language, aiming to gain large commonsense resources in the target language efficiently.

The main challenge is projection ambiguity. For example, consider translating (bat, *CapableOf*, fly) shown in Figure 1. Bat has four Japanese translations such as koumori [bat (animal)] and batto [bat (stick)]. Fly has 27 translations such as tobu [fly (verb)] and hae [fly (insect)]. Thus, (bat, *CapableOf*, fly) results in $4 \times 27 = 108$ Japanese translation candidates in total, and even after filtering them

*This work was conducted while the first author was at Kyoto University.

[1] An English fact is a fact with two English concepts.

Proceedings of the 27th International Conference on Computational Linguistics, pages 1508–1520
Santa Fe, New Mexico, USA, August 20-26, 2018.

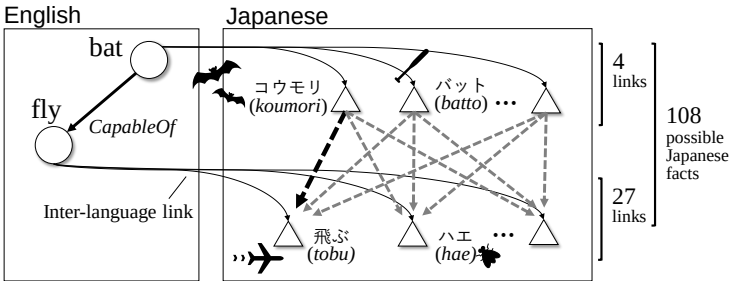

Figure 1: **Ambiguity of knowledge projection.** (bat, *CapableOf*, fly) in English and 108 possible translations in Japanese. The task is to identify the correct link between Japanese words (dash line) that corresponds to the English link (solid line.)

by considering part-of-speech constraints we still have 64 candidates. This problem happens very frequently because 42% of English concepts appearing in inter-language links have more than one Japanese translation.

This is in contrast to previous studies explored cross-lingual knowledge projection focused on knowledge of named entities (Feng et al., 2016; Klein et al., 2017; Chen et al., 2017). Their methods assume one-to-one mapping of concepts across languages. This assumption is reasonable if concepts are named entities because the majority of named entities has one or a few translations, *e.g.*, France (English) and Francia (Spanish).

In contrast, commonsense concepts are represented by common nouns, verbs, and phrases, and those words/phrases have many translations by nature as shown in Figure 1. Such translation ambiguity, that is, the knowledge projection ambiguity can be partially solved by machine translation (MT) and knowledge base completion (KBC) techniques. Cross-lingual knowledge projection can be seen as a structured version of an MT task. KBC models complete missing relations between concepts based on existing relations, which are also closely related to our task. Neither of them, however, can disambiguate knowledge projection with adequate precision. We do not have sufficient training data for building a translation model of facts because MT systems are generally developed not for structured knowledge but for unstructured text. KBC models need to be trained on a sufficiently large KB in a target language.

To alleviate these problems, we combine MT and target-side KBC. The MT and KBC models are trained on separate datasets, and our model weights the estimates from the two models to generate final results. To compute translation probabilities of facts with MT, we propose to convert a fact into plain text with hand-crafted templates.

Our contributions are three-fold.

1. We propose a cross-lingual projection method for undisambiguated forms of commonsense. Our method combines MT and target-side KBC to disambiguate knowledge projection across languages. To utilize an MT model trained on unstructured text, we develop rule-based conversion of structured knowledge.
2. We demonstrate that our method outperforms a projection method that assumes one-to-one mapping of concepts, and single KBC and MT models. Furthermore, an experiment on a crowdsourced dataset shows our method can find correct translations with a top-10 accuracy of 90%.
3. We obtained 18,747 accurate facts of Japanese commonsense using our method and crowdsourcing, which are an equivalent or larger amount of the existing facts in ConceptNet for 12 relation types. We release the resulting datasets as well as code to reproduce our experiments to the research community.[2]

2 Related Work

Developing human language technologies for low-resource languages has been an important challenge for years, and several studies attempted to bridge the resource gap across languages by cross-lingual

[2] https://github.com/notani/CLKP-MTKBC

Relation	e_1	e_2	English	Japanese	Chinese
AtLocation	NP	NP	You are likely to find e_1 in e_2.	e_2 de e_1 wo miru koto ga aru.	Ni keyi zai e_2 zhaodao e_1.
CapableOf	NP	VP	e_1 can e_2	e_1 wa e_2 koto ga dekiru .	e_1 hui e_1.
MadeOf	NP	NP	e_1 is made of e_2.	e_1 wa e_2 kara tsukurareru.	e_1 keyi yong e_2 zhi cheng.
UsedFor	NP	VP	You can use e_1 to e_2.	e_1 wa e_2 tame ni tsukawareru.	e_2 de shihou keneng hui yong dao e_1.

Table 1: **Examples of templates for converting facts into sentences.** Constraints of part-of-speech on e_1 and e_2 (Speer and Havasi, 2012) are also presented. Some templates were developed by the ConceptNet organizers. The rest of the templates can be found in the released code.

knowledge projection.

Klein et al. (2017) and Chen et al. (2017) represented concepts in multiple languages in a unified vector space, and built knowledge base completion models based on vector representations. Their methods ensure a concept in the source language has a similar vector representation to its target-side counterpart, assuming each concept in the source language corresponds to exactly one concept in the target language.

There is a rich body of work on sense embedding, which allows one surface form of a word to have sense-specific vectors (Neelakantan et al., 2014; Iacobacci et al., 2015). However, to the best of our knowledge, previous studies in this field do not target sense vectors of concepts for cross-lingual knowledge projection.

Several studies proposed methods for one-to-one projection of facts (Kuo and Hsu, 2010; Feng et al., 2016). The work by Feng et al. (2016) is the most related to our study. Their model learns mappings between English and Chinese facts by manually annotated alignments. Their experimental result showed the model successfully resolved the projection ambiguity. Their experiment was, however, limited to a narrow domain due to the cost of manual annotations, indicating the difficulty of obtaining sufficient resources for learning a model.

Various types of commonsense is vital to understanding languages in a wide range of tasks such as recognizing textual entailment (LoBue and Yates, 2011). Researchers have compiled resources to maintain such knowledge. Cyc (Lenat, 1995) is a seminal big project that aims to organize commonsense in logical forms. Logical forms are suitable for disambiguating the meaning of language, but we need high expertise to acquire or utilize them. In contrast, ConceptNet (Liu and Singh, 2004b; Speer et al., 2017) adopted natural language expressions such as words and phrases that may have ambiguities to represent knowledge, which made it possible to collect millions of commonsense facts in multiple languages via crowdsourcing.

3 Problem Setting

Suppose we project a fact f^s in a source language into a target language. We obtain n candidate translations by following inter-language links. In ConceptNet, the links are built from data sources such as Wiktionary and WordNet. We denote these candidates as $f_1^t, \cdots, f_n^t$.

Our goal is to estimate a projection score $h(f_i^t|f^s)$, and find the most appropriate target-side fact that maximizes the score.

$$\hat{f}^t = \underset{f_i^t}{\operatorname{argmax}}\, h(f_i^t|f^s) \tag{1}$$

4 Method

We propose two methods to combine MT and target-side KBC models for estimating projection scores.

4.1 Machine Translation (MT)

MT models consider contexts to find bilingual mapping of sentences. Given "I saw a bat in the zoo." as a source sentence, the model will assign a higher translation probability to "doubutsuen de koumori wo

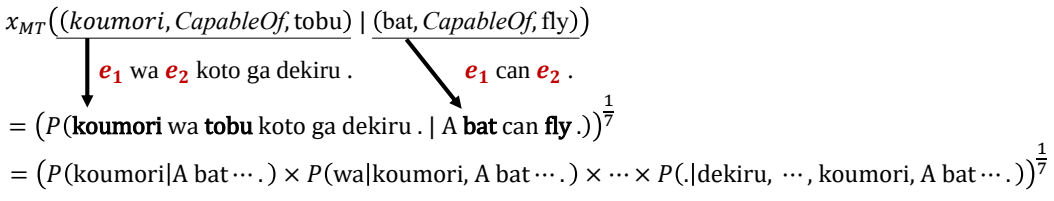

Figure 2: **Calculating a translation probability of a fact using an MT model.** For simplicity, this example does not use subword units. In addition, we omit special symbols that represent the beginning and end of a sentence in this figure.

mita." [I saw a bat (animal) in the zoo.] than to "doubutsuen de batto wo mita." [I saw a bat (stick) in the zoo.]

In contrast to our problem, typical MT focuses on plain texts, and only unstructured parallel texts are normally available for training MT models. Thus, we convert facts into natural language expressions beforehand. We use a rule-based approach to generate an expression for each fact, for example, "e_1 can e_2" corresponding to (e_1, *CapableOf*, e_2). Fortunately, some facts in ConceptNet already have such language expressions (Speer and Havasi, 2012). For the rest of the facts, we develop simple templates based on the existing expressions. Table 1 shows examples of templates in English, Japanese and Chinese.[3] We refer to part-of-speech tags of concepts to generate natural-sounding sentences.

Using sentences of facts, we define a score from the MT model as a translation probability of the language expression normalized by the target-side length m.

$$x_{\mathrm{MT}}(f^t|f^s) = \left(\mathrm{P}(W^t|W^s)\right)^{1/m}, \tag{2}$$

where W^t and W^s are sentences of f^t and f^s, respectively. To define $\mathrm{P}(W^t|W^s)$, we employ an off-the-shelf sequence-to-sequence model with an attention mechanism (Bahdanau et al., 2014), which is one of the recent successful MT models.

Figure 2 illustrates the translation probability of a Japanese fact (koumori, *CapableOf*, tobu) given an English fact (bat, *CapableOf*, fly). We first obtain language expressions of the facts using hand-crafted templates and compute a translation probability with the translation model.

4.2 Knowledge Base Completion (KBC)

KBC models evaluate the plausibility of a given fact based on existing information on the KB. For example, if we already know many animals with wings can fly and bats have wings, we can imagine that bats also can fly. We train a KBC model on the target-side KB.

We use a bilinear model used in several previous studies (*e.g.*, (Li et al., 2016)) as a component of our model, where concepts and relations are represented as vectors and matrices, respectively. This component can also be replaced with other KBC models.[4]

Given a fact $f^t = (e_1, r, e_2)$, the bilinear model outputs the value of plausibility as follows.

$$x_{\mathrm{KBC}}(f^t) = \sigma(\boldsymbol{u}_1^T \boldsymbol{M}_r \boldsymbol{u}_2), \tag{3}$$

where σ is a sigmoid function, $\boldsymbol{u}_i \in \mathbb{R}^d$ $(i = 1, 2)$ corresponds to vectors of concepts e_1 and e_2, $\boldsymbol{M}_r \in \mathbb{R}^{d \times d}$ corresponds to a matrix of relation r, and d is a hyper parameter. We construct concept vectors by averaging pre-trained d'-dimensional word embeddings as several previous studies did to boost the predictive performance (Socher et al., 2013; Li et al., 2016). The following nonlinear transformation reduces the dimensionality for computational efficiency.

$$\boldsymbol{u}_i = \tanh(\boldsymbol{W}\boldsymbol{v}_i + \boldsymbol{b}), \tag{4}$$

[3] We provide the templates in the released code.

[4] Potential alternatives can be found in the survey paper by Wang et al. (2017).

where $\boldsymbol{v}_i \in \mathbb{R}^{d'} (i = 1, 2)$ is a pre-trained concept vector, and $\boldsymbol{W} \in \mathbb{R}^{d \times d'}$ is a weight, $\boldsymbol{b} \in \mathbb{R}^d (d < d')$ is a bias term. The model parameters are learned to minimize a cross-entropy function on training facts.

4.3 Combination of Scores

We combine the two scores explained above to generate a score for each pair of f_i^t and f^s. Our model h takes $\boldsymbol{x}(f_i^t|f^s) = \left(x_{\text{KBC}}(f_i^t), x_{\text{MT}}(f_i^t|f^s)\right)$ as an input (for simplicity, we omit f_i^t and f^s, hereafter), and calculates a projection score, where x_{KBC} and x_{MT} are normalized before calculation.

We first describe two options for $h(\boldsymbol{x})$, (1) a linear transformation and (2) a multi-layer perceptron, and next explain the inference procedure of parameters below.

Linear Transformation: First, a *linear transformation* (LIN) model combines x_{KBC} and x_{MT} linearly. The model has a different weight vector and a bias term for each relation because the accuracy of KBC and MT varies for different relation types.

$$h(\boldsymbol{x}) = \boldsymbol{w}_r^T \boldsymbol{x} + b_r \quad \boldsymbol{w}_r \in \mathbb{R}^2, b_r \in \mathbb{R}. \tag{5}$$

Multi-layer Perceptron: LIN is a very simple model and may cause underfitting. Thus, we introduce a *multi-layer perceptron* (MLP) model with one hidden layer to increase the model capacity. The model has an input layer, one hidden layer, and an output layer. MLP calculates $h(\boldsymbol{x})$ by the equation below.

$$h(\boldsymbol{x}) = {\boldsymbol{w}_r^{(2)}}^T \boldsymbol{z}(\boldsymbol{x}) + b_r^{(2)} \tag{6}$$

$$\boldsymbol{z}(\boldsymbol{x}) = \tanh\left({\boldsymbol{W}^{(1)}}^T \boldsymbol{x} + \boldsymbol{b}^{(1)}\right) \tag{7}$$

$$\boldsymbol{W}^{(1)} \in \mathbb{R}^{2 \times c}, \boldsymbol{b}^{(1)} \in \mathbb{R}^c, \boldsymbol{w}_r^{(2)} \in \mathbb{R}^c, b_r^{(2)} \in \mathbb{R}.$$

Note that the common weight matrix and bias are used across relations in Equation (7) in order to capture the global intermediate representations of projections.

Inference: Given training instances of fact-to-fact projection, we estimate the model parameters that compute high scores to correct translations and low scores to incorrect translations. The training data consists of a correct translation set T_+ and an incorrect translation set T_-. $T_-(\boldsymbol{x}_+)$ denotes a set of incorrect translations that have the same English fact as $\boldsymbol{x}_+ \in T_+$.

For each $x_+ \in T_+$, we define the following margin-based loss function.

$$\text{loss}(\boldsymbol{x}_+) = \max\left(0, 1 + h(\boldsymbol{x}_+) - h(\boldsymbol{x}_-)\right), \tag{8}$$

where x_- is randomly extracted from $T_-(\boldsymbol{x}_+)$. We sum this loss function over T_+, and obtain the model parameters Θ by minimizing the summed loss function.

$$\hat{\Theta} = \underset{\Theta}{\text{argmin}} \sum_{\boldsymbol{x}_+ \in T_+} \text{loss}(\boldsymbol{x}_+) \tag{9}$$

5 Experiments

We conducted two experiments to compare our method with baseline methods.

5.1 Data

We use automatically and manually constructed datasets for evaluating knowledge projection methods. Throughout the experiments, facts are obtained from ConceptNet version 5.5.0.[5]

5.1.1 Automatically Built Datasets

The first experiment used semi-automatically built datasets of English–Japanese and English-Chinese projection. We call these AUTO datasets. The English–Japanese AUTO dataset was constructed in three steps:

[5] https://github.com/commonsense/conceptnet5/wiki/Downloads

Language	Fact (unique) f^t	Fact (unique) f^s	Translation T_+	Translation T_-
En→Ja	22,508	2,767,450	33,724	5,612,716
En→Zh	2,294	11,648	2,861	12,348

(a) AUTO

Fact (unique) f^t	Fact (unique) f^s	Translation T_+	Translation T_-
13194	200	832	12,365

(b) MANUAL (En → Ja)

Table 2: **Statistics of dataset.**

1. We translated each English fact f^s into Japanese facts $f^t_1, f^t_2, \cdots$ with inter-language links in ConceptNet. Those that violated part of speech constraints (Speer and Havasi, 2012) were discarded.
2. If a translated fact f^t_i already existed in Japanese ConceptNet, we considered the pair of the English and Japanese facts to be a positive projection, *i.e.*, $(f^s, f^t_i) \in T_+$. Otherwise, we include the pair in a set of negative projection T_-.
3. The previous step resulted in millions of obvious negative projection, which is often directed to rare Japanese words. To reduce such projection, we counted co-occurrences of all pairs of concepts in 200 million Japanese web sentences and discarded Japanese facts whose concepts do not occur together.

We applied the same procedure to Chinese facts, where we used the Chinese Gigaword Fifth Edition[6] in step 3. The size of the AUTO dataset is reported in Table 2(a). The English–Japanese dataset is larger than the English–Chinese dataset because the number of English–Japanese inter-language links in ConceptNet is four times larger than English–Chinese links. We can gain data by harvesting links from lexical resources such as dictionaries and multi-lingual WordNet, which is left as future work.

We conducted five-fold cross validation by splitting the datasets into training (60%), validation (20%) and test (20%) sets.

5.1.2 Manually Built Dataset

The AUTO datasets are large but may not be accurate enough to test methods because the target KBs were small by nature, and many true Japanese facts were not identified as correct projection in step 2. Thus, we next built an accurate but small testing dataset annotated by humans. We call this dataset MANUAL. Due to the cost constraint, this dataset was only built for English–Japanese projection.

We used crowdsourcing to annotate the data. Human workers were gathered in a Japanese crowdsourcing platform Yahoo! Crowdsourcing[7].

1. We extracted the 200 most confident English facts based on the scores in ConceptNet. We only used English concepts with fewer than 20 inter-language links.[8] In addition, we removed facts containing dirty words.[9]
2. We projected the English facts into Japanese with inter-language links as we did in step 1 of the AUTO datasets.
3. Crowd workers annotated the Japanese facts with five-level labels: (1) "false, or does not make sense", (2) "true only in a few contexts", (3) "true in several contexts", (4) "true in many contexts", and (5) "true". Each Japanese fact was judged by five workers.
4. We aggregated the collected judgments by taking median.

The size of the resulting dataset is reported in Table 2(b). An English fact had 66 translations on average. We used this dataset only for evaluation. To conduct this evaluation, we trained our models on the whole AUTO datasets.

[6]The LDC catalog number is LDC2011T13.

[7]`http://crowdsourcing.yahoo.co.jp`

[8]We found a few concepts had extremely many links (*e.g.*, `/c/en/person`), and most of the links are inappropriate.

[9]We use a dirty word list on google_twunter_lol (`https://gist.github.com/jamiew/1112488`).

	MRR	Acc@1	Acc@10
PPMI	.183	.081	.406
MT	.306	.190	.546
KBC	.352	.232	.610
MTransE	.185	.097	.372
LIN (EQ)	.363†	.233	.626†
LIN	.363†	.230	.633†
MLP	**.370**†	**.234**	**.644**†

(a) En→Ja

	MRR	Acc@1	Acc@10
PPMI	.667	.466	.980
MT	.742	.585	.982
KBC	.673	.480	.975
MTransE	.762	.626	.976
LIN (EQ)	.768	.624	.986
LIN	.766	.620	.985
MLP	**.772**	**.629**	**.988**†

(b) En→Zh

Table 3: **Results on the AUTO dataset.** † denotes LIN (EQ), LIN and MLP outperformed all the baselines significantly (paired t-test with $\alpha = 0.05$.)

5.2 Baselines and Proposed Methods

We compare the performance of our proposed methods with the following baselines.

- **PPMI:** Positive pointwise mutual information of two concepts consisting of a target-side fact f^t. We count the co-occurrence of the concepts in the 200 million web sentences for Japanese, and in the Chinese Gigaword Fifth Edition for Chinese concepts.
- **MT:** The neural MT model with an attention mechanism, which computes x_{MT} in the proposed methods. We used an implementation by Neubig (2015) and train a model on 3.25M (en-ja) and 2.97M (en-zh) sentence pairs from dictionaries and newswire corpora. BPE (Sennrich et al., 2016) was used to reduce the vocabulary size.
- **KBC:** The target-side bilinear KBC model which was used as the component to produce x_{KBC}. The Japanese and Chinese models were trained on 59,274 and 318,361 facts, respectively.
- **MTransE:** The multi-lingual translation-based KBC model (Chen et al., 2017) which learns TransE (Bordes et al., 2013) and concept-to-concept alignment jointly. Chen et al. (2017) proposed five different alignment models and reported the fourth variant performed best in their experiments. Thus we use the variant in our experiments.

The proposed methods **LIN** and **MLP** use estimates of MT and KBC models above. We also include **LIN (EQ)**, a variant of LIN, which equally combines scores from MT and KBC after normalizing each score to a $[0, 1]$-range. We use the Adam optimizer (Kingma and Ba, 2014) for training models. We provide implementation details and hyperparameter settings in the appendix.

5.3 Results

We report mean reciprocal rank (MRR), top-1 and -10 accuracy (Acc@1 and Acc@10) on the test set.[10] To calculate these metrics on the MANUAL dataset with five-way labels, we binarized the labels by considering (5) true label as positive and the others as negative because we aim to find the most appropriate projection for each English fact. We removed English facts which only had positive/negative Japanese translations when calculating MRR and accuracy. Besides, we also report nDCG (normalized discounted cumulated gain) using the five-way labels.

5.3.1 AUTO

Table 3 shows MRR, Acc@10 and Acc@1 on the test datasets for English–Japanese and English–Chinese projection. LIN (EQ), LIN and MLP outperformed MT and KBC in most cases, indicating combining them helped find correct translations. LIN (EQ) performed on par with LIN even though LIN (EQ) does not learn combination weights from the training data. This result could be attributed to the limited capacity of a linear model and motivates us to use MLP.

MTransE achieved high precision on the English–Chinese dataset, but failed to yield correct predictions on the English–Japanese dataset. The essential difference between the two language pairs is in

[10] Some previous studies reported meanrank, but MRR is more robust to outliers than meanrank.

En→Ja	MRR	Acc@1	Acc@10	nDCG@10
PPMI	.450	.282	.859	.632
MT	.544	.380	.866	.689
KBC	.448	.303	.775	.610
MTransE	.260	.148	.521	.473
LIN (EQ)	.600†	**.472**†	.894	.711†
LIN	.602†	**.472**†	.894	.711†
MLP	**.606**†	**.472**†	.901	**.714**†
#facts$_{en}$		142		200

Table 4: **Result on the MANUAL dataset.** † denotes LIN (EQ), LIN and MLP outperformed all the baselines significantly (paired t-test with $\alpha = 0.05$.)

Rank			Concepts		
MLP	MT	KBC	e_1	e_2	Label
1	6	8	ao [blue]	shikisai [color]	**true**
2	9	4	heki [blue]	shikisai [color]	**true**
3	2	35	brû [blue]	shikisai [color]	**true**

(a) (blue, RelatedTo, color)

MLP	MT	KBC	e_1	e_2	Label
1	5	1	rokku [rock (music)]	myûjikku [music]	**true**
2	9	2	rokku [rock (music)]	ongaku [music]	**true**
3	25	5	rokku [rock (music)]	fumen [music score]	true in several contexts
			...	...	
15	1	49	iwa [rock (stone)]	myûjikku [music]	false

(b) (rock, IsA, music)

Table 5: **Improved examples.** Top ranked projections by MLP are reported with labels and ranks given by MT and KBC.

the degree of the ambiguity, that is, the English–Chinese projection is not as ambiguous as the English–Japanese projection on our dataset because of the lack of English–Chinese links in ConceptNet v5.5.0. This characteristic boosted the performance of MTransE, which assumes one-to-one projection of concepts.

We observed the performance of the baselines varied across relations. MT was inaccurate at lexical relations such as *Antonym* and *Synonym* but outperformed KBC on *HasFirstSubevent*, *HasLastSubevent*, and *UsedFor*. MLP outpeformed single MT and KBC for the most of the relations by combining them.

5.3.2 MANUAL

Table 4 shows the result on the MANUAL dataset. The differences between the proposed methods and the baselines are statistically significant except for Acc@10 (paired t-test with $\alpha = 0.05$.) All the methods resulted in better scores on this dataset than on the AUTO dataset. Although PPMI appears to be accurate, in fact it failed to provide valid scores to 8,894 out of 13,197 examples as the co-occurrences of their concepts were not observed in the corpus. Likewise, the MANUAL dataset had many concepts that were not in the training data for MTransE, which seriously degraded its performance.

The examples in Table 5 show that MLP well combined the strength of MT and KBC models as we hypothesized. In Table 5(a), MLP put a weight on MT since it learned on the training set that MT tends to be more reliable at *RelatedTo* relation than KBC. The ratio *RelatedTo* facts was higher on the MANUAL dataset than the AUTO dataset, and we think this was the reason why MT outperformed KBC. Table 5(b) is an example in which KBC mitigated erroneous predictions by MT. MT preferred iwa [stone] to rokku [rock music], whereas KBC provided the high score to the latter. We found a pre-trained word vector of rokku was similar to those of music genres such as hevimetaru [heavy metal] and jazu [jazz], which could help KBC to identify the word sense of rock occurring with music.

Rank			Concepts		
MLP	KBC	MT	e_1	e_2	Label
1	11	5	*iso [rocky coast]	*sunago [sand/grit]	false
2	1	61	*iso [rocky coast]	suna [sand/grit]	true only in a few contexts
3	4	43	*umibe [seashore]	sand [sand/grit]	true in several contexts

(beach, *RelatedTo*, sand)

Table 6: **Failed example.** Top ranked projections by MLP are reported with labels and ranks given by MT and KBC. * denotes a rare Japanese word.

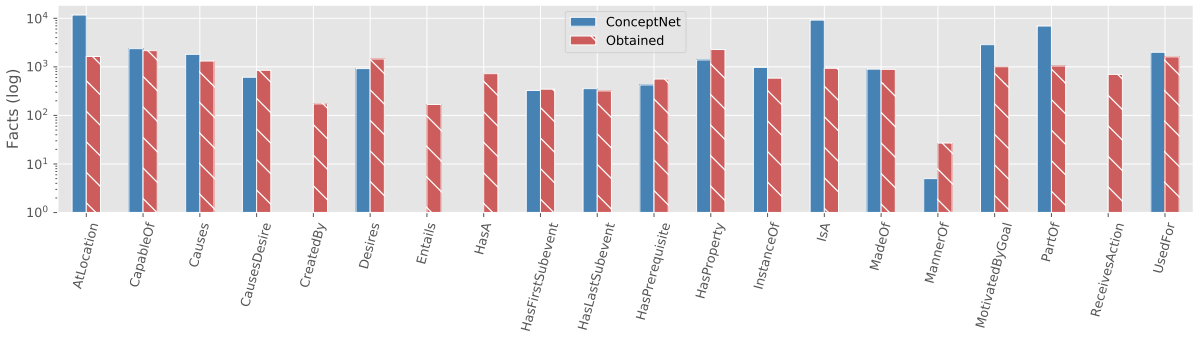

Figure 3: **Number of obtained and existing Japanese facts.**

Table 6 shows failed examples. KBC, MT, and the proposed methods produced low scores for the correct facts in these examples. This was because some negative examples contained rare words, and both MT and KBC gave them high scores. We found KBC was particularly inaccurate for facts containing OOV words. In practice, discarding rare words would be a reasonable choice in order to achieve high precision.

6 Japanese Commonsense Knowledge Construction

The previous experiments have shown our method can score projection candidates with high precision. We now use our method to collect Japanese commonsense resources of high quality.

We first sampled 10,000 English facts that cover 20 relation types.[11] In the same way as step 1 in Section 5.1.1, we obtained Japanese counterparts of them. We then used the MLP model, which achieved the best score on the MANUAL dataset, and computed scores of the projection candidates.

Although top-10 predictions are likely to contain correct projection as shown in Table 4, we further used crowdsourcing to refine the projected knowledge. We showed crowd workers top 10 confident Japanese facts that were generated from the same English fact, and the workers chose all correct Japanese facts if any. Each set of candidates was judged by five workers. Here, we converted facts into natural language by the hand-crafted templates described in Section 4.1 so that workers can easily understand the meaning. All the facts were checked by 838 workers only for 25 hours. Their annotations were aggregated by majority voting.

As a result, we obtained 18,747 facts. Note that one English fact can have multiple Japanese counterparts. Figure 3 shows the distribution of the obtained facts against the existing Japanese facts in ConceptNet. The current Japanese facts concentrated on a few relation types such as *IsA* and *RelatedTo*, and most of the relation types do not have many facts. Indeed, we have already collected an equivalent or larger amount of commonsense knowledge for 12 relation types.[12]

[11]The sampled English facts include *AtLocation*, *CapableOf*, *Causes*, *CausesDesire*, *CreatedBy*, *Desires*, *Entails*, *HasA*, *HasFirstSubevent*, *HasLastSubevent*, *HasPrerequisite*, *HasProperty*, *InstanceOf*, *IsA*, *MadeOf*, *MannerOf*, *MotivatedByGoal*, *PartOf*, *ReceivesAction*, and *UsedFor*.

[12]12 types include *CapableOf*, *CausesDesire*, *CreatedBy*, *Desires*, *Entails*, *HasA*, *HasFirstSubevent*, *HasPrerequisite*, *HasProperty*, *MadeOf*, *MannerOf*, and *ReceivesAction*.

7 Conclusion and Future Work

We proposed a method to project knowledge stored in English into other languages. We focused on commonsense knowledge that is required to understand human communications. The main challenge of cross-lingual knowledge projection is the ambiguity of projection. To resolve this ambiguity, our method combines MT and target-side KBC models. Experiments showed the proposed method outperformed baseline methods by large margins consistently. We projected 10,000 English into Japanese and obtained 18,747 accurate facts using our method and crowdsourcing. There are still more than 450,000 English facts with inter-language links to Japanese, and we are planning to project them into Japanese by our proposed method and crowdsourcing refinement. We will release the resulting resources to research communities in order to facilitate research in many languages.

Acknowledgments

We thank anonymous reviewers for their valuable suggestions. This work was partially supported by Yahoo Japan Corporation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473:1–15.

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2787–2795.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pages 15–26, Vancouver, Canada, August. Association for Computational Linguistics.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multi-lingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, Melbourne, Australia, November.

Xiaocheng Feng, Duyu Tang, Bing Qin, and Ting Liu. 2016. English-chinese knowledge base translation with neural network. In *Proceedings of the 26th International Conference on Computational Linguistics: (COLING)*, pages 2935–2944. The COLING 2016 Organizing Committee, December.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of The 3rd Workshop on Automated Knowledge Base Construction (AKBC)*. ACM Press, October.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 95–105, Beijing, China, July. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6, December.

Patrick Klein, Simone Paolo Ponzetto, and Goran Glavaš. 2017. Improving neural knowledge base completion with cross-lingual projections. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 516–522. Association for Computational Linguistics, April.

Yen-Ling Kuo and Jane Yung-Jen Hsu. 2010. Bridging common sense knowledge bases with analogy by graph similarity. In *Proceedings of the 2nd AAAI Workshop on Collaboratively-Built Knowledge Sources and Artificial Intelligence (WikiAI)*, pages 22–27, Atlanta, Georgia, USA, July. AAAI Press.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. 2014. DBpedia – A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

Douglas B. Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1445–1455, Berlin, Germany, August. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004a. Commonsense Reasoning in and Over Natural Language. In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES)*, pages 293–306, Wellington, New Zealand. Springer, Berlin, Heidelberg.

Hugo Liu and Push Singh. 2004b. ConceptNet A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Peter LoBue and Alexander Yates. 2011. Types of Common-Sense Knowledge Needed for Recognizing Textual Entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 329–334, Portland, Oregon, USA, June. Association for Computational Linguistics.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2292–2297. Association for Computational Linguistics, sep.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar, October. Association for Computational Linguistics.

Graham Neubig. 2015. lamtram: A toolkit for language and translation modeling using neural networks. http://www.github.com/neubig/lamtram.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 926–934. Stateline, Nevada, USA.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 3679–3686, Istanbul, Turkey, May. European Language Resources Association.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 4444–4451, San Francisco, California, USA, February. AAAI Press.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, December.

Liangjun Zang, Cong Cao, Yanan Cao, Yuming Wu, and Cungen Cao. 2013. A Survey of Commonsense Knowledge Acquisition. *Journal of Computer Science and Technology*, 28(4):689–719, July.

A Experimental Setup

In experiments (Section 5), we built the baselines and our combination methods (Section 4) in the following procedures. Facts are from ConceptNet version 5.5.0. Japanese words in concepts are normalized using the morphological analyzer JUMAN++ (Morita et al., 2015). We used the Adam (Kingma and Ba, 2014) for training models. Initial learning rates α are described below.

A.1 MT Model

We built the neural MT model with lamtram (Neubig, 2015). The English–Japanese parallel corpora included 3,253,923 sentence pairs from dictionaries and newswire:

- JEC Basic Sentence Data[13]
- EDICT [14]
- Eijiro[15]
- Tatoeba Project[16]
- Open Multilingual WordNet (Bond and Foster, 2013)
- JENAAD (Utiyama and Isahara, 2003)

We extracted 2,965,845 English–Chinese sentence pairs from LDC[17]. We segmented words by byte-pair encoding (BPE) with vocabulary size 8,003 for each language. This vocabulary includes three special symbols indicating the beginning and the end of the sentence, and the out-of-vocabulary word. We trained the BPE model (Sennrich et al., 2016) on the training set with sentencepiece[18]. Hyper parameters were tuned based on perplexity on randomly selected 1,000 sentence pairs. The best setting was the encoder/decoder of LSTM with 512 hidden nodes, dropout of rate 0.25, and learning rate $\alpha = 0.001$ for the both language pairs.

A.2 KBC Model

A bilinear model (Section 4.2) was trained on 59,274 Japanese and 317,161 Chinese facts in ConceptNet. They are already sorted by a confidence score. We excluded facts in the training and evaluation datasets for knowledge projection (Section 5.1.)

Following Li et al. (2016), we define concept vectors by taking an average of predefined word embeddings. Japanese word embeddings of 256 dimensions were trained on 200 million sentences from the web. Chinese word embeddings of 300 dimensions were obtained from the CoNLL 2016 data.[19] We normalized concept vectors using the mean and variance calculated on the training set.

In hyper parameter tuning, the most confident 600 facts were used for evaluation, the second most confident 600 facts were used for early stopping in training, and the remaining facts were used for training. These facts were treated as positive examples, and we generated negative examples by randomly swapping one component of each fact. Once the best parameters were determined, the most confident 600 facts were used for early stopping, and the other facts were used to train a model.

We removed positive examples with out-of-vocabulary (OOV) words from the datasets. Instead, we added facts with a OOV vector as negative examples to the training data because rare words are often incorrect.

We selected dimensionality $d \in \{100, 150\}$, regularization coefficients for $\boldsymbol{M}_r$ and the other parameters $\lambda_1, \lambda_2 \in \{0.001, 0.0001, 0.00001\}$, learning rate $\alpha \in \{0.1, 0.01\}$, and batch size $\beta \in 200, 400, 800$.

[13] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JEC\%20Basic\%20Sentence\%20Data

[14] http://www.edrdg.org/jmdict/edict.html

[15] ISBN: 978-4757428126

[16] https://tatoeba.org

[17] The catalog numbers are LDC2002L27, LDC2002T01, LDC2003T17, LDC2004T07, LDC2004T08, LDC2005T10, LDC2005T34, LDC2006T04, LDC2007T02, LDC2012T16, LDC2012T20, and LDC2012T24.

[18] https://github.com/google/sentencepiece

[19] CoNLL 2016 Shared Task Multilingual Shallow Discourse Parsing, http://www.cs.brandeis.edu/~clp/conll16st/dataset.html

The resulting parameters for the Japanese facts were $d = 150$, $\lambda_1 = 0.00001$, $\lambda_2 = 0.001$, $\alpha = 0.1$, and $\beta = 200$. Those for the Chinese facts were $d = 100$, $\lambda_1 = 0.001$, $\lambda_2 = 0.0001$, $\alpha = 0.1$, and $\beta = 400$.

A.3 MTransE

MTransE (Chen et al., 2017) was trained with the implementation provided by Chen et al. (2017). We set batch size 100 to speed up training although the original code adopted online learning (*i.e.*, batch size is 1). We empirically found mini batch learning did not impair the performance. We enumerated all combinations reported by Chen et al. (2017) for tuning dimensionality d, weight of the alignment model γ, norm l, learning rate α. The resulting parameters for the English–Japanese dataset are $d = 100$, $\gamma = 2.5$, $l = L2$, and $\alpha = 0.5$. Those for the English–Chinese dataset are $d = 100$, $\gamma = 2.5$, $l = L2$, and $\alpha = 0.1$.

A.4 LIN and MLP

The proposed methods were trained using estimates of KBC and MT described above. The validation set was used for early stopping in training, and selecting the best hyper parameters. We conducted a grid search for learning rate $\alpha \in \{0.5, 0.25, 0.125, 0.0625, 0.03125\}$ and dimensionality $d \in \{8, 16, 32, 64, 128\}$ (only for MLP). The resulting configurations are reported below.

- **LIN:** $\alpha = 0.125$ for the English–Japanese dataset, and $\alpha = 0.5$ for the English–Chinese dataset
- **MLP:** $\alpha = 0.5$ and $d = 16$ for the English–Japanese dataset, and $\alpha = 0.5$ and $d = 32$ for the English–Chinese dataset