applied sciences

MDPI

Article

Constructing Uyghur Commonsense Knowledge Base by Knowledge Projection

Azmat Anwar [1,2,3], Xiao Li [1,2,3,*], Yating Yang [1,2,3] and Yajuan Wang [1,2,3,4]

1 Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China
2 Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China
3 University of Chinese Academy of Sciences, Beijing 100049, China
4 Department of Information Security Engineering, Xinjiang Police College, Urumqi 830011, China
* Correspondence: xiaoli@ms.xjb.ac.cn; Tel.: +86-136-0993-8871

Received: 24 July 2019; Accepted: 9 August 2019; Published: 13 August 2019

Abstract: Although considerable effort has been devoted to building commonsense knowledge bases (CKB), it is still not available for many low-resource languages such as Uyghur because of expensive construction cost. Focusing on this issue, we proposed a cross-lingual knowledge-projection method to construct an Uyghur CKB by projecting ConceptNet's Chinese facts into Uyghur. We used a Chinese–Uyghur bilingual dictionary to get high-quality entity translation in facts and employed a back-translation method to eliminate the entity-translation ambiguity. Moreover, to tackle the inner relation ambiguity in translated facts, we made a hand-crafted rule to convert the structured facts into natural-language phrases and built the Chinese–Uyghur lingual phrases based on the similarity of phrases that corresponded to the bilingual semantic similarity scoring model. Experimental results show that the accuracy of our semantic similarity scoring model reached 94.75% for our task, and they successfully project 55,872 Chinese facts into Uyghur as well as obtain 67,375 Uyghur facts within a very short period.

Keywords: commonsense knowledge bases; Uyghur; knowledge projection; low-resource languages

1. Introduction

Knowledge Bases (KBs) play an important role in many natural-language processing (NLP) tasks such as question answering, web searching and dialog tasks [1,2]. KBs describe the knowledge about entities, relations, and their attributes. In a KB, each fact is a triple of the form (*h*, *r*, *t*) that indicates the head entity *h* and tail entity *t* are connected with a relationship named *r*, e.g., (*Bat*, *CapableOf*, *Fly*). As a part of KB, commonsense knowledge (CSK) is mainly referred to as background knowledge and used in natural-language processing tasks that require reasoning based on implicit knowledge [3–6]. However, many languages, especially low-resource languages including Uyghur, have no existing commonsense knowledge base (CKB) to use [7–9].

Constructing CKBs from scratch is very time-consuming and labor-intensive. Instead, there are many available CKB resources in other rich-resource languages such as English and Chinese. A straightforward way to construct an Uyghur CKB is to directly translate Chinese KB to Uyghur based on the surface texts of a fact with the existing machine translation (MT) system or bilingual dictionary. However, we find that this method suffers from the problem of ambiguity. For example, consider translating Chinese fact (主机 <host computer>, *CapableOf*, 发热 <heat>) shown in Figure 1. The head entity "主机" (host computer) has six Uyghur translation candidates, including ئاساسى ماتۇر(main engine) and مەركىزىي كومپيۇتېر(central computer). The tail entity "发热" (heat) also has six translation candidates including قىزىش(heat) and قىزىتما(have a fever). Thus, (主机, *CapableOf*, 发热) will generate $6 \times 6 = 36$ Uyghur translation candidates in total.

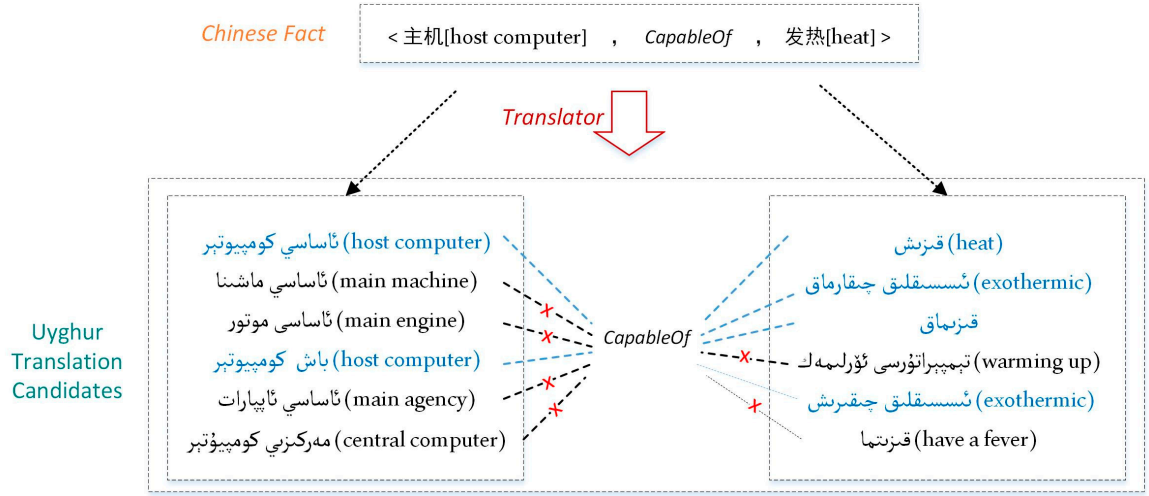

Figure 1. An example of translation ambiguity in Chinese–Uyghur commonsense knowledge base (CKB) translation. Black words represent the semantically unrelated translation candidates while blue words represent the retained candidates after back-translation.

There are two main challenges to effectively disambiguate these translation triples. The first challenge is how to remove the semantic unrelated translation candidates for a single entity. The second challenge is how to effectively model the semantics of the Chinese fact and Uyghur fact in common semantic space.

In this paper, we address these two challenges by presenting a cross-lingual knowledge-projection method to translate the Chinese CKB into Uyghur. Given a Chinese fact, first, the method uses a Chinese–Uyghur bilingual dictionary to get entity translations in fact and use back-translation to remove the semantically unrelated translation candidates. Then, the method converts each Chinese and Uyghur fact to a parallel sentence using a hand-crafted rule template and achieves their sentence representation by a recursive autoencoder. Finally, the method encodes the source and target sentence in the same semantic space using the bidimensional attention network, and calculates the distance between them to get the semantic similarity score.

Being the largest multilingual CKB, ConceptNet [10] connects words and phrases of natural language with labeled edges and maintains knowledge as a triple of two concepts and relations between them. The relations come from a fixed set. The latest release (v5.6.0) of ConceptNet has 369,687 unique Chinese facts (both head and tail nodes are Chinese) while the number of Uyghur facts is only 3872 (≈1.05%). We focus on ConceptNet in this paper. We project ConceptNet's Chinese facts to Uyghur by using a Chinese–Uyghur bilingual dictionary and a bilingual semantic similarity scoring model, automatically building an Uyghur CKB from existing Chinese CKBs, taking the advantages of the cross-lingual knowledge projection.

Suppose we project a Chinese fact f^s into a target-side Uyghur fact, and obtain n candidate translations by using dictionary translation. We denote these candidates as $f_1^t, f_2^t, \ldots, f_n^t$. Our goal is to estimate a projection score $h(f_i^t|f^s)$ and find the most appropriate Uyghur fact that maximizes the score, which can be formulated in Equation (1).

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

The structure of this paper is organized as follows: in Section 2 we discuss the related works; in Section 3 we introduce our cross-lingual knowledge-projection method; in Section 4 we present the experiments and analysis of the results. Conclusions will be given in the last section.

2. Related Works

Low-resource languages often suffer from a lack of annotated corpora to estimate high-performing neural network models for many NLP tasks. Cross-lingual knowledge projection is an efficient way to bridge the gap across languages.

Named-entity recognition (NER) for low-resource languages has received great benefit from cross-lingual language projection. Bharadwaj et al. [11] built a transfer model using phonetic features instead of lexical features. These features are not strictly language-independent but work well when languages share vocabulary but have spelling variations, as in the case of Turkish, Uzbek, and Uyghur. Mayhew et al. [12] used lexicon to translate the available annotated data in one or several high-resource language(s), and learned a standard monolingual NER model. They evaluated their model on 7 diverse languages and improved the state of the art (SOTA) by an average of 5.5% F1 points. To improve the mapping of lexical items across languages, Xie et al. [13] proposed a method that finds translations based on bilingual word embeddings and uses self-attention to improve the robustness for word-order differences. Their method achieved a SOTA NER performance on commonly tested languages.

Chen et al. [14], Wang et al. [15], and Klein et al. [16] represented concepts in multiple languages in a common vector space and ensured a concept in source language has a similar vector representation to its target-side counterpart. Xu et al. [17] treated the cross-lingual knowledge projection as a graph-matching problem and proposed a graph-attention-based solution, which matches all the entities in two topic entity graphs and jointly models the local matching information to derive a graph-level matching vector.

Manaal et al. [18] presents a system that performs relation extraction (RE) on a sentence in the source language by translating the sentence into English then performing RE in English and projecting the relation phrase back to the source language sentence. Their method only needs a MT system from the source language to English without any other analysis tools for the source language and can extract relationships for any source languages.

Due to the lack of training data for sentiment analysis, Jeremy et al. [19] introduced a bilingual sentiment embedding model for cross-lingual sentiment classification. Their model only requires a small bilingual lexicon, a source-language corpus annotated for sentiment, and monolingual word embeddings for each language. Experiments on three language combinations for sentence-level cross-lingual sentiment outperforms the SOTA methods.

Several studies proposed methods for the one-to-one projection of facts. To expand Chinese KB by leveraging English KB resources, Feng et al. [20] presented a gated neural network approach to map the source triples and target triples in the same semantic vector space. Their experimental result showed the model can successfully alleviate the projection ambiguity. The work by Naoki et al. [21] is the closest related to our study. They treated cross-lingual knowledge projection as a structured version of the MT task and generated a training corpus from ConceptNet using hand-crafted rules for every type of relationship. By combining MT and a target-side knowledge-base completion model, they projected the English CSK into Japanese and Chinese with high precision.

3. Method

3.1. Data Preprocessing

Being a multilingual commonsense KB, the ConceptNet contains facts from hundreds of languages. Each fact consists of five parts: The URI of the whole edge, the relationship expressed by the edge, the node at the head of the edge, the node at the tail of the edge and JSON-structured additional information. To project the Chinese facts into Uyghur, first, we need to filter out the facts in which both the head and tail node entities are Chinese, and convert them into a format of $f^s = (e_1^s, r, e_2^s)$, where $e_1^s \in E_1$, $r \in R$, $e_2^s \in E_2$. Symbol E_1 and E_2 represent the set of the head and the tail nodes while R represents the set of the relationships between nodes.

3.2. Dictionary-Based Entity Translation

To project a given Chinese fact $f^s = (e_1^s, r, e_2^s)$ into Uyghur, we need to translate the head and tail entity into Uyghur separately. To get a high-quality entity translation, we use a Chinese–Uyghur bilingual dictionary to get the target fact $f^t = (e_1^t, r, e_2^t)$, where $e_1^t \in \left\{e_{11}^t, e_{12}^t, \ldots, e_{1n}^t\right\}$ and $e_2^t \in \left\{e_{21}^t, e_{22}^t, \ldots, e_{2m}^t\right\}$. From the example shown in Figure 1, it can be seen that for most cases we can get more than one candidate Uyghur entity for each Chinese entity after translation, but some of them are semantically unrelated to the original Chinese entity.

Back-translation [22] method is first used in MT to enrich the training corpus. In this paper, we use it to eliminate the entity-translation ambiguity. Firstly, we use an Uyghur–Chinese bilingual dictionary to translate the translated candidate entities back to Chinese, then compare that with the original one, and keep the candidate entities if they are equal.

After back-translation, we get translated head and tail entity $\acute{e}_1^t$, $\acute{e}_2^t$, which $\acute{e}_1^t \in \{\acute{e}_{11}^t, \acute{e}_{12}^t, \ldots, \acute{e}_{1\acute{n}}^t\}$, $\acute{e}_2^t \in \{\acute{e}_{21}^t, \acute{e}_{22}^t, \ldots, \acute{e}_{2\acute{m}}^t\}$, for a Chinese fact f^s. Combining with relationship r, the translated fact and the count of translated fact can be expressed as $\acute{f}^t = (\acute{e}_1^t, r, \acute{e}_2^t)$ and $L(\acute{f}^t) = L(\acute{e}_1^t) \times L(\acute{e}_2^t) = \acute{n} \times m \cdot$, where $\acute{n}$ and $\acute{m}$ is the count of the translated candidate head and tail entity, respectively.

3.3. Rule-Based Conversion of Structured Knowledge

Although we get the correct translation of the head and the tail entity by the dictionary-based translation separately, when combining with relationship r, the generated Uyghur fact also displays semantic ambiguity between the head and tail entity with the relationship. For the given example above, the generated Uyghur fact (كومپيوتېر باش <host computer>, *CapableOf*, قىزىتما <have a fever>) does not semantically have a *CapableOf* relationship. It is challenging to solve this inner relation ambiguity in a single projected fact.

We suppose that the original Chinese fact does not have any ambiguity, and by calculating the semantic similarity of the original and the projected fact, we can tackle this inner relationship ambiguity in translated fact. However, it is difficult to calculate the semantic similarity of two facts while all facts in ConceptNet are in triple structure. Therefore, we make hand-crafted templates for every relationship in Chinese and Uyghur separately to convert the structured facts into phrases. Thus, we can get the similarity of the facts by calculating the semantic similarity of the parallel phrases generated by templates. Hand-crafted rule templates for Chinese and Uyghur are shown in Table 1. *e*1 and *e*2 in templates will be programmatically replaced by head and tail entities in fact.

Table 1. Templates for converting Chinese and Uyghur facts into phrases. The English templates were developed by ConceptNet organizers (https://github.com/commonsense/conceptnet5/wiki). The content in parentheses after the entity in the Uyghur template represents the variations of affix in the Uyghur template.

Relationship	English	Chinese	Uyghur
IsA	e1 is part of e2	e1是e2的一种	بىر تۈرى (نىڭ) **e2**بولسا **e1**
Causes	The effect of e1 is e2	e1会e2	بولىدۇ **e2**بولسا **e1**
Desires	e2 wants to e1	e1需要e2	مۇھتاج (غا، قا، گە، كە)**e2 e1**
CapableOf	e1 can e2	e1会e2	قىلالايدۇ (نى) **e2 e1**
SymbolOf	e1 represents e2	e1代表e2	ئىپادىلەيدۇ (نى) **e2 e1**
HasProperty	e2 is e1	e1是e2的	**e1 e2**
RelatedTo	e1 is related to e2	e1跟e2有关	مۇناسىۋەتلىك **e2**بىلەن **e1**
UsedFor	You can use e1 to e2	e2的时候可能会用到e1	ئىشلىتىشى مومكىن (نى) **e2 e1**
CausesDesire	e1 makes you want to e2	e1让你想要e2	**e2**سىزنى **e1**
MadeOf	e1 is made of e2	e1可以用e2制成	ئارقىلىق ياسىلىدۇ **e1 e2**
NotDesires	e1 not desires e2	e1不想e2	خالىمايدۇ (نى) **e1 e2**
AtLocation	You are likely to find e1 in e2	你可以在e2找到e1	تاپالايسىز (نى)**e1** (دىن، تىن)**e2**سىز

Table 1. *Cont.*

Relationship	English	Chinese	Uyghur
DerivedFrom	e2 is derived from e1	e2源自e1	**e1 e2** (دىن ، تىن) كەلگەن
partOf	e2 is part of e1	e2是e1的一部分	**e1 e2** (نىڭ) بىر قىسمى
HasSubevent	One of the things you do when you e1 is e2	当e1时,可能会e2	**e1** بولسا **e2**مومكىن
Synonym	11 and e2 are synonymous	e1和e2是同义词	**e1** مەنداش سۆز **e2**بىلەن
HasA	e2 has e1	e2有e1	**e1** بار (دا ، تا ، دە ، تە)**e2**

Being an agglutinative language, Uyghur has many affixes which play an important role in syntax information. In Uyghur, there are multivariant affixes with different variants of one affix added to harmonize the phonetic characteristics of the particular stem. For example, the plural affix has two variants "لار/لەر" and they must be chosen based on the phonetic harmony rule between stem and variants [23]. Aizimaiti et al. [24] proposed a rule-based variant-selection algorithm for Uyghur affixes based on Uyghur phonetic harmony. We use their method while replacing entities in a template to select a correct affix variation to combine with the entity for each Uyghur entity.

3.4. Bilingual Semantic Similarity Scoring Model

Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings (BattRAE) were first proposed by Zhang et al. to evaluate the semantic similarity between a source phrase and a target phrase in an MT task [25,26]. We introduced the BattRAE model to score the semantic similarity of parallel phrases generated from the original and projected facts using the hand-crafted template. This model learns bilingual phrase embeddings according to the strengths of interactions between the linguistic items at different levels of granularity on the source side and the target side. Figure 2 shows the overall architecture of the BattRAE model.

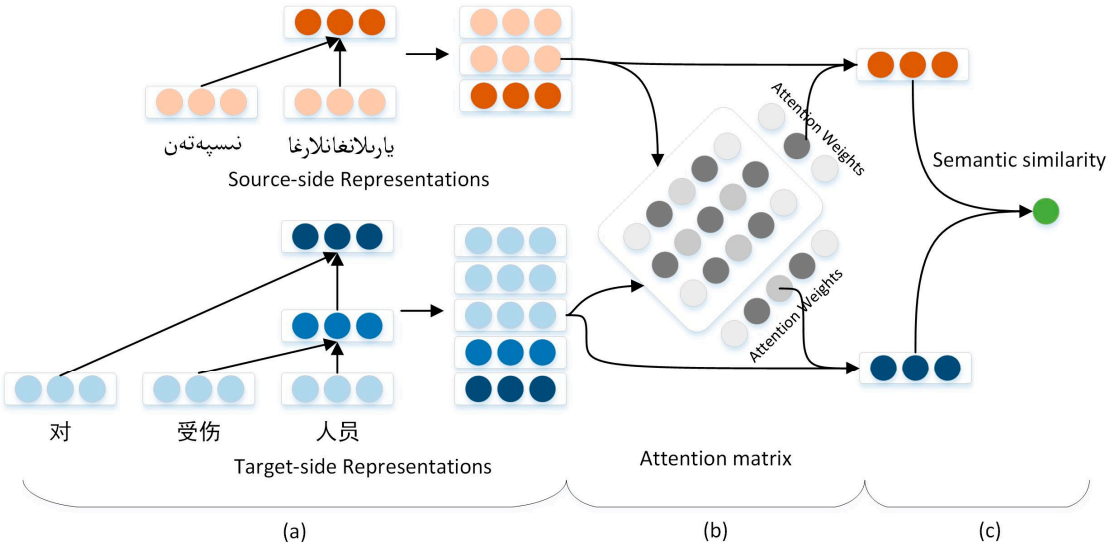

Figure 2. Overall architecture for the model: (**a**) Recursive autoencoders (**b**) bidimensional attention network; and (**c**) semantic similarity layer.

3.4.1. Learning Multilevel Phrase Embeddings

We use recursive autoencoders (RAE, Figure 2a) to learn initial embeddings at different levels of phrases. By combining two children vectors from the bottom up recursively, RAE can generate low-dimensional vector representations for variable-sized sequences. The recursion procedure usually consists of two main steps: composition and reconstruction.

Composition: Generally, for a list of words in a phrase (x_1, x_2, x_3), each of them will be embedded into a d-dimensional continuous vector, RAE selects two neighboring children (e.g., $c_1 = x_1$ and $c_2 = x_2$) via some selection criterion, and then composes them into a parent embedding y_1, which can be computed by Equation (2).

$$y_1 = f(W^{(1)}[c_1 : c_2] + b^{(1)}) \tag{2}$$

where $[c_1 : c_2] \in \mathbb{R}^{2d}$ is the concatenation of c_1 and c_2, $W^{(1)} \in \mathbb{R}^{d \times 2d}$ is a parameter matrix, $b^{(1)} \in \mathbb{R}^d$ is a bias term, f is element-wise activation function such as $\tan\mathrm{h}(\cdot)$, which is used in our experiments.

Reconstruction: After getting the d-dimensional representation for parent y_1 in the composition step, to measure how well the parent y_1 represents its children, we reconstruct the original child nodes via a reconstruction layer formulated in Equation (3).

$$\left[c_1' : c_2'\right] = f(W^{(2)} y_1 + b^{(2)}) \tag{3}$$

where c_1' and c_2' are the reconstructed children, $W^{(2)} \in \mathbb{R}^{2d \times d}$ and $b^{(2)} \in \mathbb{R}^{2d}$, The minimum Euclidean distance between $[c_1 : c_2]$, and $\left[c_1' : c_2'\right]$ is usually used as the selection criterion during composition.

These two steps repeat until the embedding of the entire phrase is generated. While embedding, RAE also constructs a binary tree. The structure of the tree is determined by the used selection criterion in composition. We use a greedy algorithm [27] based on the following reconstruction error, which can be seen as Equation (4).

$$E_{rec}(\chi) = \sum_{y \in T(x)} \frac{1}{2} \left\| [c_1 : c_2]_y - \left[c_1' : c_2'\right]_y \right\|^2 \tag{4}$$

where y is an intermediate node of the binary tree $T(x)$, and parameters $W^{(1)}$ and $W^{(2)}$ are learned to minimize the sum of reconstruction errors.

Given a binary tree learned by RAE, the leaf, internal nodes, and root of the tree which represents the representations of words, sub-phrases, and phrases separately, we can use RAE to produce the embeddings of phrases at different levels. As shown in Figure 1, RAE learns representations of the source and target phrases in different semantic spaces, marked as d_s and d_t, respectively.

3.4.2. Bidimensional Attention Network

We propose the bidimensional attention network (Figure 2b) to incorporate a multilevel representation of embeddings from RAE into phrase embeddings and further into the semantic similarity of bilingual phrases. We can put vectors from all nodes of a tree into the columns of a matrix of size $(2n-1) \times d$ (n_s, d_s for source and n_t, d_t for target), where d is the dimension of embeddings and n is the length of phrase (there are $n-1$ steps in RAE to construction and therefore are $2n-1$ nodes in total). Let us denote these matrices by M_s and M_t for the source and target tree, respectively. Then we can project all the embeddings into a common attention space by using by a non-linear projection function $f(Wx + b)$. In this attention space, all the embeddings from the source tree can "interact" with all the embeddings from the target tree. We will measure the interaction strength between the i-th projected source embedding and the j-th projected target embedding by Equation (7).

$$A_s = f(W^{(3)} M_s + b^A) \tag{5}$$

$$A_t = f(W^{(4)} M_t + b^A) \tag{6}$$

$$B(i,j) = g(A_{s,i}^T A_{t,i}) \tag{7}$$

where $g(\cdot)$ and $f(\cdot)$ are non-linear activation functions, e.g., the $\mathrm{sigmoid}(\cdot)$ and the $\tan\mathrm{h}(\cdot)$ functions are used in this paper, A_s (Equation (5)) and A_t (Equation (6)) are projections of M_s and M_t to the attention space, $W^{(3)} \in \mathbb{R}^{d_a \times d_s}$ and $W^{(4)} \in \mathbb{R}^{d_a \times d_t}$ are transformation matrix, $b^A \in \mathbb{R}^{d_a}$ is the bias

term. We will use the same bias-term force model to learn to encode the attention semantics into transformation matrices, rather than the bias term.

It can be seen that we define a $(2n_s - 1) \times (2n_t - 1)$ matrix B, which is called the bidimensional attention matrix represented by Equation (7). Intuitively, this matrix is a result of handshakes between source and target phrases at a multilevel representation. We can interpret the sum of the i-th row as the total strength that the i-th source node has on the semantic similarity between the two considered phrases.

$$\widetilde{a}_{s,i} = \sum_j B_{i,j}\ ,\ \ \widetilde{a}_{t,j} = \sum_i B_{i,j} \tag{8}$$

where $\widetilde{a}_s \in \mathbb{R}^{n_s}$ and $\widetilde{a}_t \in \mathbb{R}^{n_t}$ are the semantic matching score vectors.

Because of phrase length uncertainty, we can normalize all these strengths using a $softmax$ function: $a_s = Softmax(\widetilde{a}_s)$, $a_t = Softmax(\widetilde{a}_t)$. This forces a_s and a_t to become real-valued distributions in the attention space, known as attention weights. Then, we use them to obtain the final phrase representations by the following Equation (9).

$$p_s = \sum_i a_{s,i} M_{s,i},\ \ p_t = \sum_j a_{t,j} M_{t,j} \tag{9}$$

where $p_s \in \mathbb{R}^{d_s}$, and $p_t \in \mathbb{R}^{d_t}$, notice that they still are located in their language-specific vector space.

3.4.3. Semantic Similarity

To measure the semantic similarity of the bilingual phrase, first we transform the learned phrases representations p_s and p_t into common d_{sim}-dimensional semantic space by a non-linear projection formulated in Equations (10) and (11).

$$s_s = f(W^{(5)} p_s + b^s\ ) \tag{10}$$

$$s_t = f(W^{(6)} p_t + b^s\ ) \tag{11}$$

where $W^{(5)} \in \mathbb{R}^{d_{sim} \times d_s}$, $W^{(6)} \in \mathbb{R}^{d_{sim} \times d_t}$ and $b^s \in \mathbb{R}^{d_{sim}}$ are the parameters. We will also use the same bias term as shown in Equation (6).

Then, to get the final semantic similarity of bilingual phrases, we calculate the cosine similarity of p_s and p_t by Equation (12) (Figure 2c).

$$s(f, e) = \frac{s^{Ts}}{\|s_s\|\ \|s_t\|} \tag{12}$$

where f and e indicate the source and target phrase, and $\|\cdot\|$ denotes the L2-norm of a vector.

According to the definition of semantic similarity, the semantic error will be introduced to measure the semantic equivalence of source and target phrase. Given a positive bilingual phrase pair (f, e) with its negative samples (f^-, e) and (f, e^-), we use the following error-based max-margin function, which is formulated in Equation (13).

$$E_{sim}(f, e) = \max(0,\ 1 + s(f, e^-) - s(f, e)\ ) + \max(0,\ 1 + s(f^-, e) - s(f, e)\ ) \tag{13}$$

Intuitively, minimizing this error will maximize the similarity of the positive instance and minimize the similarity of the negative pairs. For each training instance (f, e), the joint objective of BattRAE is defined by Equation (14):

$$J(\theta) = \alpha E_{rec}(f, e) + \ \beta E_{sim}(f, e) + R(\theta) \tag{14}$$

where $E_{rec}(f, e) = \ E_{rec}(f) + \ E_{rec}(e)$, $\alpha + \ \beta = 1$, and $R(\theta)$ is regularization term.

4. Experiment

4.1. Setup

- Facts dataset: Through the experiments, we will use the facts obtained from ConceptNet version 5.6.0 (https://github.com/commonsense/conceptnet5/wiki/Downloads).
- Dictionary: We use the Chinese–Uyghur bilingual dictionary, which contains 328,000 unique Chinese terms and 531,000 unique Uyghur terms, to translate entity.
- Word embeddings dataset: We use the toolkit Word2Vec (https://github.com/tmikolov/word2vec) to pretrain the word embeddings, which contains 11,500,000 Chinese sentences provided by Sogou (http://www.sogou.com/labs/resource/list_yuliao.php) to train the Chinese word embeddings and 1,500,000 Uyghur sentences crawled from the Tianshan website (http://uy.ts.cn/) to train the Uyghur word embeddings.
- Semantic similarity model training dataset: To obtain high-quality bilingual phrases to train the semantic scoring model, we use the Moses decoder (http://www.statmt.org/moses/) to force decoding [28] on CWMT2013 Chinese–Uyghur parallel corpus (https://www.cis.um.edu.mo/cwmt2014/en/cfp.html) which contains 109,000 parallel sentences, and the extra collected 1,380,000 bilingual phrases. To generate negative samples for each training phrase, we used the following two strategies introduced by Ondrej [29]: (1) taking a completely different phrase; and (2) choosing a random word from the phrase and replacing it with its farthest word by calculating the cosine distance all over the vocabulary.
- Semantic similarity model hyperparameters: we set $d_s = d_s = d_a = d_{sim} = 50$, $\alpha = 0.125$ (so that $\beta = 0.875$), use L-BFGS algorithm (libLBFGS (http://www.chokkan.org/software/liblbfgs/)) to optimize the objective function.

4.2. Experiment

4.2.1. Entity Filtering and Translation Performance

The filtering and dictionary-based entity-translation results are shown in Table 2. We obtain 369,687 Chinese facts after filtering, which contains 67,400 unique start entities and 85,800 unique end entities with 24 relations. Through dictionary-based entity translation, we translate the 99,600 Chinese facts to Uyghur and get 2,900,000 translated Uyghur facts with 24,800 head and 27,900 tail entities with 23 relations. It can be seen that the translation generated from many incorrect facts leads to entity-translation ambiguity. By using the dictionary-based entity back-translation, we can filter out the incorrect or semantically unrelated entities on the source and translated side. The back-translation result shows that we effectively remove 95% of the incorrectly translated facts while losing only 36.4% of facts on the source side.

Table 2. Experimental results of dictionary-based entity translation.

Method	Source (Chinese)			Translated (Uyghur)			Relation Count
	Head	Tail	Fact	Head	Tail	Fact	
Original	67,464	85,832	369,687		0		24
Translation	12,708	15,373	99,649	24,863	27,921	2,944,802	23
Back-translation	7336	8089	63,406	8835	9873	143,926	17

4.2.2. Bilingual Semantic Similarity Scoring Model Analysis

1. Semantic Accuracy

We obtain 143,900 translated Uyghur facts for 63,400 Chinese facts in the dictionary-based translation step. By using the hand-crafted rule template, we can generate the 143,900 Uyghur–Chinese

parallel sentences and score the semantic similarity of each sentence by trained bilingual semantic similarity scoring model.

There is no publicly available test set for Uyghur–Chinese bilingual phrase similarity measurement, so we randomly select 2000 scored bilingual sentences, using a combination of automatic test and manual check, to get the accuracy of the semantic scoring model.

Automatic Test:

We also focus on whether our model can recognize correct bilingual phrases; in other words, we assign higher semantic similarity scores for the correct bilingual phrases. We use semantic accuracy metrics for this evaluation, which is mentioned in [25]. Formally, given a pair of correct bilingual phrases (f, e) and its incorrect counterpart (f, e^-) (replaced with a non-translation target phrase) or (f^-, e) (replaced with a non-translation source phrase), the semantic accuracy (*SAcc*) of the bilingual phrase is defined as follows: (we take (f, e^-) for example)

$$SAcc = \begin{cases} True & if\ s(f,e) > s(f,e^-) \\ False & otherwise \end{cases} \tag{15}$$

Manual Check:

We use crowdsourcing to check semantic scores, detailed as follows:

- As we use cosine similarity as a scoring metric, whose values are distributed from −1 to 1, we set zero as the semantic similarity threshold.
- Workers check the Uyghur facts with labels: (1) "True, makes sense in every context", (2): "False, does not make sense, or does not make sense in some contexts".
- Each Uyghur fact is judged by three workers.
- We aggregate the collected judgments by taking the median.

Finally, we kept the Uyghur facts that have been through the automatic and manual test verification. Performance of the scoring model is shown in Figure 3. It can be observed that the model can achieve 94.75% accuracy for our task, which works well for many relationship templates, except for *SymbolOf* and *Synonym* relationship.

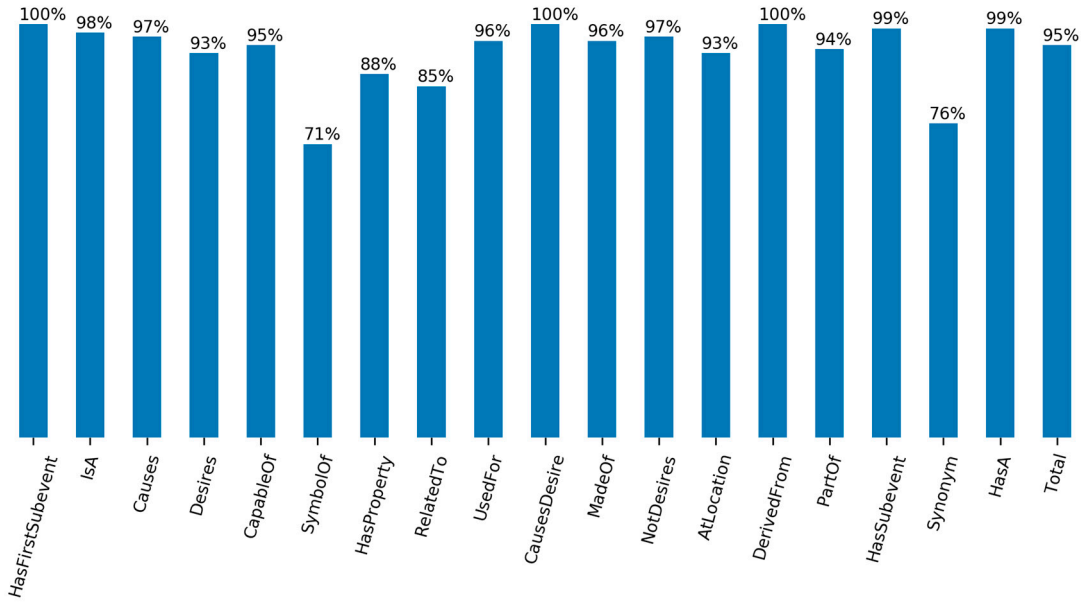

Figure 3. Accuracy of bilingual phrase semantic similarity model for each relation type.

2. Error Analysis

After analyzing the semantic scoring result, we find two types of errors, as follows:

- Unknown Word Error: Although we have pretrained word embeddings on a fairly large corpus, we also find that being an agglutinative language, Uyghur still has some words that could not be included. They affect the accuracy of the model due to random initialization. For example, the words سىياسىئون(politician) and پرېزىدېنت(president) could not get a correct embedding when training and testing, as the sentence which contains this word gets a low score.
- Template Error: We define a single template for each relationship type, which works well for most facts. However, for some verbs, the dictionary-translated entities format does not match with the hand-crafted template, so it generates ungrammatical sentences, especially for Uyghur. For example, Table 3 shows the generated sentences of grammatical errors for the *Causes* relation. Because we get translations of Uyghur verbs with the incorrect tense according to the dictionary, the template generates ungrammatical sentences and gets incorrect score while testing.

Table 3. Failed Uyghur sentences generation example for *Causes* relation.

Method	Chinese	Uyghur (Incorrect)	Uyghur (correct)
Sentences	难过会忧郁	كۆڭلى بۇزۇلماق بولسا قايغۇلۇق بولىدۇ	كۆڭلى بۇزۇلسا قايغۇرىدۇ
	愚弄会翻脸	كولدۇرلاتماق بولسا يۈز ئۆرۈمەك بولىدۇ	كولدۇرلاتسا يۈز ئۆرىدۇ

4.2.3. Construct Uyghur CKB

To get the final projected Uyghur facts, we need to filter the bilingual phrases according to the semantic similarity score. We use two filtering strategies, as follows:

- For Chinese facts which only have a single candidate projected Uyghur fact, we will keep this fact if the semantic score is greater than zero.
- For Chinese facts which have multiple candidate Uyghur facts, we will sort the scores and keep the highest one.

The results of filtered Uyghur facts are shown in Figure 4. After all the above steps, we can filter out the projected 55,872 Chinese facts into Uyghur successfully and get the 67,375 facts.

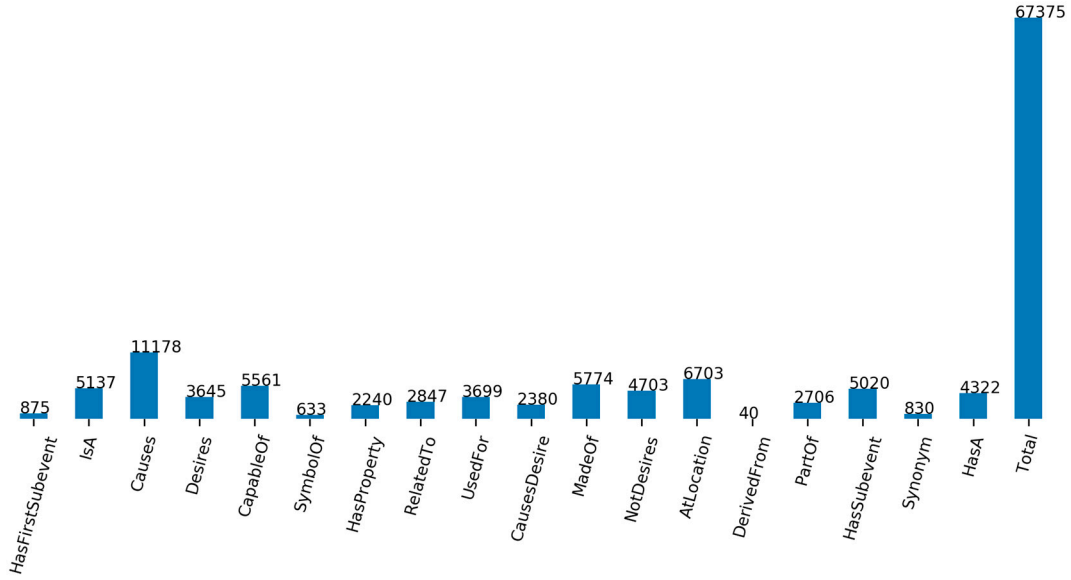

Figure 4. Results of filtered Uyghur facts for each relationship type.

5. Conclusions and Future Work

We propose a method to project knowledge stored in Chinese into the Uyghur language. We focus on CSK that is required to understand human communications. The main challenge of this work is entity ambiguity and inner relationship ambiguity. To get the entity projection, our method uses a Chinese–Uyghur dictionary for the entity translation and employs back-translation for entity ambiguity. To resolve the inner relationship ambiguity, we make relationship templates to convert facts to bilingual phrases and use a semantic similarity scoring model to filter facts. Experiments show that our method works well for our projection task. Finally, we projected 55,872 Chinese facts into Uyghur and got 67,375 Uyghur facts successfully. There are still more than 300,000 Chinese facts that cannot be translated correctly by the dictionary and we are planning to project them into Uyghur by using the existing Chinese–Uyghur MT system combined with the proposed semantic similarity scoring model.

Author Contributions: A.A., Y.W., X.L. and Y.Y. conceived the model, prepared the datasets, and wrote the manuscript. All authors read and approved the final manuscript.

Funding: This work is supported in part by the Xinjiang Uygur Autonomous Region Level talent introduction project (Y839031201), The National Natural Science Foundation of China (U1703133), The Subsidy of the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2017472), the Xinjiang Key Laboratory Fund (Grant No.2018D04018).

Acknowledgments: The authors would like to thank all anonymous reviewers for their constructive advices.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cui, W.; Xiao, Y.; Wang, H.; Song, Y.; Hwang, S.-W.; Wang, W. KBQA: Learning question answering over QA corpora and knowledge bases. *Proc. VLDB Endow.* **2017**, *10*, 565–576. [CrossRef]
2. Xiong, C.; Power, R.; Callan, J. Explicit semantic ranking for academic search via knowledge graph embedding. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 1271–1279.
3. Davis, E.; Marcus, G.J.C.A. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* **2015**, *58*, 92–103. [CrossRef]
4. Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; Huang, M. Augmenting end-to-end dialogue systems with commonsense knowledge. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
5. Ostermann, S.; Roth, M.; Modi, A.; Thater, S.; Pinkal, M. SemEval-2018 Task 11: Machine comprehension using commonsense knowledge. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 747–757.
6. Ma, Y.; Peng, H.; Cambria, E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
7. Qiu, L.; Zhang, H. Review of Development and Construction of Uyghur Knowledge Graph. In Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, China, 21–24 July 2017; pp. 894–897.
8. Abaidulla, Y.; Osman, I.; Tursun, M. Progress on Construction Technology of Uyghur Knowledge Base. In Proceedings of the 2009 International Symposium on Intelligent Ubiquitous Computing & Education, Chengdu, China, 15–16 May 2009.
9. Yilahun, H.; Imam, S.; Hamdulla, A. A survey on uyghur ontology. *Int. J. Database Theory Appl.* **2015**, *8*, 157–168. [CrossRef]
10. Speer, R.; Chin, J.; Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.

11. Bharadwaj, A.; Mortensen, D.; Dyer, C.; Carbonell, J. Phonologically aware neural model for named entity recognition in low resource transfer settings. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1462–1472.
12. Mayhew, S.; Tsai, C.-T.; Roth, D. Cheap translation for cross-lingual named entity recognition. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2536–2545.
13. Xie, J.; Yang, Z.; Neubig, G.; Smith, N.A.; Carbonell, J. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018.
14. Chen, M.; Tian, Y.; Yang, M.; Zaniolo, C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *arXiv* **2016**, arXiv:1611.03954.
15. Wang, Z.; Lv, Q.; Lan, X.; Zhang, Y. Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 349–357.
16. Klein, P.; Ponzetto, S.P.; Glavaš, G. 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, Valencia, Spain, 3–7 April 2017.
17. Xu, K.; Wang, L.; Yu, M.; Feng, Y.; Song, Y.; Wang, Z.; Yu, D. Cross-lingual Knowledge Graph Alignment via Graph Matching Neural Network. *arXiv* **2019**, arXiv:1905.11605.
18. Faruqui, M.; Kumar, S. Multilingual open relation extraction using cross-lingual projection. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics, Denver, CO, USA, 31 May–5 June 2015.
19. Barnes, J.; Klinger, R.; Walde, S.S. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018.
20. Feng, X.; Tang, D.; Qin, B.; Liu, T. English-chinese knowledge base translation with neural network. In Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016), Osaka, Japan, 11–17 December 2016; pp. 2935–2944.
21. Otani, N.; Kiyomaru, H.; Kawahara, D.; Kurohashi, S. Cross-lingual Knowledge Projection Using Machine Translation and Target-side Knowledge Base Completion. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1508–1520.
22. Sennrich, R.; Haddow, B.; Birch, A. Improving neural machine translation models with monolingual data. *arXiv* **2015**, arXiv:1511.06709.
23. Munire, M.; Li, X.; Yang, Y. Construction of the Uyghur Noun Morphological Re-Inflection Model Based on Hybrid Strategy. *Appl. Sci.* **2019**, *9*, 722. [CrossRef]
24. Ainiwaer, A.; Jun, D.; Xiao, L.I. Rules and Algorithms for Uyghur Affix Variant Collocation. *J. Chin. Inf. Process.* **2018**, *32*, 27–33.
25. Zhang, B.; Xiong, D.; Su, J.; Qin, Y. Alignment-Supervised Bidimensional Attention-Based Recursive Autoencoders for Bilingual Phrase Representation. *IEEE Trans. Cybern.* **2018**. [CrossRef] [PubMed]
26. Zhang, B.; Xiong, D.; Su, J. Battrae: Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
27. Socher, R.; Pennington, J.; Huang, E.H.; Ng, A.Y.; Manning, C.D. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Scotland, UK, 27–31 July 2011; pp. 151–161.
28. Wuebker, J.; Mauser, A.; Ney, H. Training phrase translation models with leaving-one-out. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 475–484.
29. Hübsch, O. Core Fidelity of Translation Options in Phrase-Based Machine Translation. Bachelor's Thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, 2017.