Mining Verb-Oriented Commonsense Knowledge

Jingping Liu[†‡], Yuanfu Zhou[†], Dan Wu[†], Chao Wang[†], Haiyun Jiang[†], Sheng Zhang[¶], Bo Xu[§], Yanghua Xiao[†*]
[†]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, China
[‡]CETC Big Data Research Institute Co.,Ltd., Guiyang, China
[§]School of Computer Science and Technology, Donghua University, Shanghai, China
[¶]Ping An Health Technology, Shanghai, China
{jpliu17, yfzhou17, danwu15, cwang17, 16110240020, shawyh}@fudan.edu.cn
xubo@dhu.edu.cn, zhangsheng108@pingan.com.cn

***Abstract*—Commonsense knowledge acquisition is one of the fundamental issues in the implementation of human-level AI. However, commonsense is difficult to obtain, because it is a human consensus and rarely explicitly appears in texts or other data. In this paper, we focus on the automatic acquisition of a typical kind of implicit verb-oriented commonsense knowledge (e.g., "*person eats food*"), which is the concept level knowledge of verb phrases. For this purpose, we propose a knowledge-driven approach to mine verb-oriented commonsense knowledge from verb phrases with the help of taxonomy. First, we design an entropy-based filter to cope with noisy input verb phrases. Then, we propose a joint model based on minimum description length and a neural language model to generate verb-oriented commonsense knowledge. We conduct extensive experiments to show that our solution is more effective to mine verb-oriented commonsense knowledge than competitors, and finally, we harvest 18K verb-oriented commonsense knowledge.**

***Index Terms*—commonsense knowledge, verb phrases, probabilistic taxonomy**

I. Introduction

Commonsense knowledge (CK) plays an important role in artificial intelligence and linguistics, such as "*bird eats insects*" and "*poet writes poems*". However, most commonsense knowledge (CK) (e.g., "*person eats food*") is difficult to be acquired since it is rarely explicitly mentioned in spoken or written language [1].

Existing work on harvesting CK has been conducted in two mainstream directions. First, pursuing high-quality of CK usually requires massive manual collection. Typical commonsense knowledge bases along this direction include WordNet [2], Cyc [3] and ConceptNet [4]. However, the hand-crafted methods are not scalable due to the time-consuming and costly process. Second, harvesting CK from large corpora with data-driven approaches seeks to overcome the limitation in the recall of the first direction. Typical efforts along the second direction include the discovery of inference rules from the text [5], extraction of class attributes from query logs [6], etc. However, since most of these methods are limited to acquire knowledge directly from corpora, they can only extract explicit CK, without the ability to harvest implicit knowledge (e.g., "*person eats food*") that rarely appears in corpora. In particular, the number of relations in the existing commonsense knowledge bases is often limited since most of them are pre-defined by hand. For example, there are only about 6 relations in WordNet and 36 relations in ConceptNet 5.5 [7].

In this paper, we focus on the automatic acquisition of a typical kind of implicit CK that is verb-oriented (e.g., "*person eats food*). The most fundamental CK about a specific verb (e.g., `eat`) is what kind of subjects (e.g., `person`) will act on what kind of objects (e.g., `food`). One distinguishing characteristic of verb-oriented CK is that it is only meaningful when expressed at the concept level (we denote this knowledge, like "*person eats food*", as a kind of implicit CK that rarely appears directly in texts). It is due to that verb-oriented knowledge at instance level is so specific that it can only be *factual knowledge*. For example, "*John eats bread*" is just a trivial fact and cannot be considered as commonsense knowledge.

We propose a knowledge-driven approach to mine verb-oriented CK from verb phrases (VPs), which consist of subject, verb and object. The core idea of our approach is to conceptualize subject and object separately with concepts in a large-scale probabilistic taxonomy, known as Probase [8]. For example, "*person eats food*" can be acquired by conceptualizing the VP "*John eats bread*" with *isA*(`John, person`) and *isA*(`bread, food`). However, there are still several challenges for the acquisition of verb-oriented CK. First, *How to determine which VPs can be generalized to the verb-oriented CK.* In fact, not all input VPs need to be conceptualized as CK since some VPs are already abstract enough. Second, *How to select the appropriate concepts for an instance.* In general, an instance always has thousands of concepts, some concepts are overly specific and others are overly abstract. Third, *How to measure the plausibility of the generated verb-oriented CK in terms of semantics after conceptualizing subjects and objects, respectively.* For example, although `name` can be viewed as an appropriate concept for `John`, "*name eats food*" is implausible.

To solve the above challenges, we design two modules including entropy-based triple filter and verb-oriented CK generator. In the first module, we introduce an entropy-based metric to solve the first challenge by measuring the abstract degree of the subject and object for each VP. In the second module, we carefully design a joint optimization model with

*Yanghua Xiao is the corresponding author. This paper was supported by Shanghai Science and Technology Innovation Action Plan (No. 19511120400), National Natural Science Foundation of China (No.61732004, No.61906035) and Shanghai Sailing Program (No.19YF1402300).

2375-026X/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDE48307.2020.00181

objectives developed upon minimum description length (MDL) and neural language model (NLM) to solve the second and third challenges, where MDL is used to measure the quality of concepts for an instance, and NLM is used to evaluate the plausibility of the generated verb-oriented CK.

Contributions. The contributions of this paper are summarized as follows:

- We mined a typical kind of verb-oriented CK at concept level. Most of our target CK cannot be acquired directly from corpora.
- We proposed a joint model based on minimum description length and neural language model to generate verb-oriented CK. We also proposed an entropy-based metric to filter out noisy input verb phrases.
- We conducted extensive experiments on real-world datasets to justify the effectiveness of our approach. Finally, we harvested 18K verb-oriented CK.

II. Overview

In this section, we first formalize the problem and then outline the solution for generating verb-oriented CK. The CK generation of different verbs is independent of each other. Hence, we analyze each verb and its phrases separately. In the following paragraph, we discuss our solution for a given verb.

A. Problem Definition.

Given a verb v, we obtain a set Q that contains triples[1] (s, v, o), where $s \in S$ (a set of subjects) and $o \in O$ (a set of objects) are the subject and the object of v from corpora. With *isA* relations in Probase, we have two sets C_S and C_O that are the concepts of the subjects S and objects O, respectively. For each concept $c_s \in C_S$ (or $c_o \in C_O$), there exists a $s \in S$ ($o \in O$) satisfying *isA*(s, c_s) (or *isA*(o, c_o)). Our goal is to obtain a set of verb-oriented commonsense triples $\{(c_s, v, c_o)\}$, where $c_s \in C_S$ and $c_o \in C_O$.

B. Framework.

Our commonsense acquisition solution consists of two main modules. The first is an *entropy-based triple filter* used to get rid of noisy verb triples that are not appropriate for conceptualization. The second is a *verb-oriented CK generator* used to conceptualize filtered verb triples. The procedure is illustrated in Fig. 1.

Module 1: Entropy-based triple filter. If the subjects or objects of an input verb triple is overly abstract, we consider the triple to be a noisy input since it is already abstract. We filter a noise by measuring the abstract degree of its subject and object. In general, terms at lower-levels tend to be more abstract (the root is at the 1st level). Furthermore, *the level distribution of subjects' (objects') hyponyms at lower-levels tends to be uniform, while at higher-levels is usually concentrated.* Thus, an entropy-based metric is designed to capture this regularity.

[1]We canonicalize verb phrases to verb triples in the form of (subject, verb, object), such as (*John*, *eat*, *bread*).

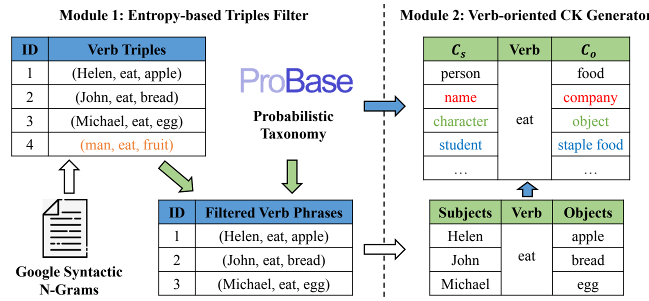

Fig. 1: Solution framework of verb-oriented CK generation. C_s is the concept set of subjects, and C_o is the concept set of objects. The words with brown, blue, green and red color represent the noisy verb triples, overly specific concepts, overly abstract concepts and implausible concepts of candidate commonsense triples, respectively. The arrows with green and blue color represent module 1 and module 2, respectively.

Module 2: Verb-oriented CK generator. The generator is realized as an optimization model with an objective function developed from MDL and NLM. MDL in this paper is used to select the appropriate concepts that are neither *specific* nor *abstract* for a given set of subjects or objects. MDL in information theory is applied to the compression of data transmission. Concept selection can be considered as a process of data compression. That is, we find the best set of concepts (models) to provide a compact description of subjects or objects (data). Thus, with MDL, we need to minimize the overall code length for encoding the bag-of-subjects (objects) with the help of the concepts. The reason we employ MDL is as follows. First, unlike traditional approaches (e.g., topic model), the number of concepts does not need to be determined in advance. Second, we can control the abstract degree of concepts by introducing an extra parameter.

NLM in this paper is used to measure the plausibility of generated triples. *A conceptualized triple is reasonable only if it can be stated as a simple but plausible natural language sentence.*

III. Entropy-based Triples Filter

In this section, we detail the entropy-based triple filter with the help of a probabilistic taxonomy. In practice, we filter the noisy verb triples by measuring the abstract degree of its subject and object. The basic idea is to use an *absolute* level to measure the abstract degree of a term. In general, the lower the level, the more abstract the terms. However, there are lots of isolated sub-taxonomies in probabilistic taxonomy, which makes it unreasonable to directly compare the absolute levels of two terms in different sub-taxonomies. Thus, we need a more flexible method to measure the abstract degree of a term. In probabilistic taxonomy, there is a phenomenon where the level distribution of hyponyms of a term at lower-levels tends to be uniform, while at higher-levels is usually concentrated.

Formally, we employ an entropy-based metric f to formalize this regularity, which is defined as follows:

$$f(t) = \begin{cases} 0, & \text{if } t \in E \\ -\sum_{i=1}^{l} p_i(t) \cdot \log p_i(t), & \text{otherwise} \end{cases}, \quad (1)$$

where E is a set of the leaf nodes, l is the maximum level in probability taxonomy, and $p_i(t)$ is the ratio of the number of t's hyponyms at the i-th level to the total number of t's hyponyms. In general, the entropy of abstract terms tends to be larger.

IV. Verb-Oriented CK Generator

In this section, we elaborate on our generator for acquiring verb-oriented CK. First, we employ MDL to select appropriate concepts for a bag-of-subjects (objects). We then use NLM to measure the plausibility of candidate triples. Finally, we present a joint model based on MDL and NLM to generate verb-oriented CK for a given set of filtered verb triples.

A. MDL-based Conceptualization

We formalize the conceptualization of a bag-of-subjects (objects) based on minimum description length [9]. MDL is widely used to find the best model for a given set of data that leads to the best compression of the data. Our task can be also considered as a process of data compression. That is, we need to select the best set of concepts (models) to provide a compact description of terms (data).

Given a set $T = \{t_1, ..., t_n\}$ of subjects (objects), our goal is to seek for a much small set $C = \{c_1, ..., c_m\}$ of concepts. Let $L(C, T)$ be the description length to encode subjects (objects) T with the concepts C. We find the set of concepts through minimizing the code length, i.e.,

$$\arg\min_{C} \; L(C, T). \quad (2)$$

To solve this function, we use a two-stage encoding schema. The key idea is to first encode the concepts C (the encoding length is defined as $L(C)$) and then encode the subjects (objects) T with the help of C (the encoding length is defined as $L(T|C)$). Thus, we have

$$L(C, T) = L(C) + L(T|C) = \sum_{c_i \in C} L(c_i) + \sum_{t_i \in T} L(t_i|C). \quad (3)$$

Based on the principle of MDL, we encode concepts with

$$L(c) = -\log \Pr(c) \quad (4)$$

bits, where $\Pr(c) = \frac{\sum_{t_i} w(t_i, c)}{\sum_{(t_j, c_k)} w(t_j, c_k)}$ is the prior probability of concept c and $w(t, c)$ is the weight of *isA*(t, c) in Probase. To encode subjects (objects), we need to know not only the code length of the subject (object) t for a given concept c, but also *the code length of the index*, which enables us to know which c is used to encode t. Since we have a total of $|C|$ concepts, each index is encoded with $\log |C|$ bits. Thus, we have

$$L(t|C) = \log |C| + \min_{c \in C}\{-\log \Pr(t|c)\}, \quad (5)$$

where $\Pr(t|c) = \frac{w(t, c)}{\sum_{t_i} w(t_i, c)}$ is the conditional probability of t for a given c, and we call it as the typicality score of t with respect to c.

In practice, the number of concepts affects its degree of abstraction. It is due to that with fewer concepts, these concepts are more abstract to better cover the meaning of all subjects (objects). We thus add an extra parameter to balance the importance of concepts and subjects (objects). We reformulate $L(C, T)$ defined in Eq. (5) as

$$L(C, T) = \alpha \sum_{c_i \in C} L(c_i) + (1 - \alpha) \sum_{t_i \in T} L(t_i|C), \quad (6)$$

where α is an adjustable parameter that is used to balance abstraction and specificity of concepts.

B. NLM-based Plausibility Measurement

After conceptualizing subjects and objects separately, we use NLM to measure the plausibility of the candidate triples, which are generated by combining the concepts of subjects and objects with the verb using indexes. Because a triple can be viewed as a simple sentence containing subject-verb-object. In recent years, Long Short-Term Memory (LSTM) has achieved state-of-the-art performance on NLM, we thus employ LSTM to build a language model.

Assume that the candidate triple (c_s, v, c_o) contains a sequence $w_{1:q} = [w_1, ..., w_q]$ of words, we use $\Pr(w_{1:q}) = \prod_{i=1}^{q} \Pr(w_i|w_1, ..., w_{i-1})$ to determine how likely the sentence formed by the candidate triple (c_s, v, c_o) appears in natural language. In practice, *perplexity* (PPL) is often used to evaluate the performance of NLM, which is defined as the geometric average of $1/\Pr(w_{1:q})$. That is,

$$\text{PPL}(c_s, v, c_o) = \Pr(w_{1:q})^{-\frac{1}{q}}. \quad (7)$$

In general, a lower PPL indicates that NLM has a better performance.

C. Joint Conceptualization

The naive method of generating verb-oriented CK is to combine MDL and NLM using a pipeline approach. Such a pipeline model suffers a major problem. That is, the errors generated by MDL may propagate to NLM. For example, when conceptualizing subjects {`Helen`, `John`, `Michael`}, only the concept `name` is selected using MDL. But for NLM, (`name`, `eat`, `food`) is an unplausible triple, resulting in a *NULL* result. It is due to that other concepts (e.g., `person`) are dropped with MDL. Thus, we need to design a joint optimization model with objectives reflecting both MDL and NLM, which is defined as follows:

$$\arg\min_{C_s, C_o} L(C_s, S) + L(C_o, O) + \lambda \frac{1}{z} \sum_{i=1}^{z} \text{PPL}_i(c_s, v, c_o), \quad (8)$$

where S and O are the set of subjects and objects of input verb triples with the verb v, C_s and C_o are two sets that are the concepts of subjects and objects, z is the number of generated candidate triples, $c_s \in C_s$ and $c_o \in C_o$ are the concept of subject $s \in S$ and object $o \in O$, λ is a parameter used to adjust the balance between MDL and NLM.

V. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the effectiveness of our approach.

A. Experiment Setup

Dataset. We use three real-world datasets to realize our solution: 1) Probabilistic taxonomy. We choose Probase to provide concepts for the subjects and objects of input verb triples. 2) Verb phrase data. We use the "English All" part in Google Syntactic N-Grams[2] to extract VPs, and 7476 verbs and 277,556 verb phrases are extracted. 3) Training corpus. We use Wikipedia [3] to pre-train NLM.

Baselines. To evaluate the effectiveness of our approach, we compare it with three baselines: 1) MDL. This is the solution by optimizing Eq. (8) without the third part, i.e., PPL($\cdot$). 2) NLM. For each VP, we first select the Top-10 typical concepts of the subject and object with the typicality score. Then, we choose the top-1 CK ranked by NLM. 3) CMNLM. We combine the conceptualization mechanism proposed in [10] and our NLM as a new method called CMNLM.

Metrics. Since no ground truth is available, we ask three volunteers to manually rank the quality for the generated verb-oriented CK. For each verb, we report the mean average quality (MAQ) for a set of verb-oriented CK, which is the mean of the average scores of volunteers for each verb. The higher the score, the more plausible verb-oriented CK is.

B. Statistics of Verb-oriented CK.

Now, we report an overview of our mined verb-oriented CK. We filter the low-frequency verbs to ensure the accuracy of the inputs. In practice, we filter the noisy verbs with $freq(v) \leq 3$, where $freq(v)$ is the frequency of verb v in all verb phrases. Finally, we extract 2423 verbs and 126,367 verb phrases. With our solution, *we harvest 259 verbs and 18406 verb-oriented CK.* In Table I, we list several typical verbs and their CK.

TABLE I: Some extracted commonsense triples.

Verbs	**(subject, object)**
eat	(person, food), (bird, insect), (rabbit, grass) ...
write	(poet, poem), (writer, novel), (composer, song) ...
read	(child, story), (poet, epic), (person, document) ...
wear	(woman, garment), (emperor, ornament) ...
cause	(bacterium, infection), (antibiotic, side effect) ...

C. Model Comparison

We select the top 100 verbs of the highest frequency to compare our model with the baselines including MDL, NLM and CMNLM. The result is shown in Table II. We can see that our model has achieved a better performance than competitors. The reason for the poor performance of CMNLM is that the input of this model is a single triple, while that of our model is multiple triples. In general, multiple triples can be generalized to CK better.

[2]http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html

[3]https://dumps.wikimedia.org/enwiki/

TABLE II: The results of compared models. Our model outperforms competitors.

Models	NLM	MDL	CMNLM	Our Model
MAQ	1.18	1.73	1.63	2.10

TABLE III: The comparison of several typical commonsense knowledge bases.

Database	**# Relations**	**Database**	**# Relations**
OMCSNet [11]	20	WordNet [2]	6
WebChild [12]	19	MindNet [13]	24
ConceptNet5.5 [7]	36	Ours	259

Furthermore, we compare our mined verb-oriented CK with other typical commonsense knowledge bases in terms of relations. The results are shown in Table III. Compared to other commonsense knowledge bases, our mined CK has two characteristics. First, the number of our relations is up to 259, far exceeding that of other knowledge bases. Second, most of the relations found by our solution are not in the competitors. That is, most verb-oriented CK is new knowledge never found.

VI. CONCLUSION

In this paper, we focus on a typical kind of implicit verb-oriented CK, which rarely appears in corpora. To this end, we propose a knowledge-driven approach to mine CK automatically from verb phrases with the help of the probability taxonomy. In the experiment, we implement our solution and harvest 18K verb-oriented CK.

REFERENCES

[1] H. G. Widdowson, *Discourse analysis*. Oxford University Press Oxford, 2007, vol. 133.
[2] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[3] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.
[4] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.
[5] D. Lin and P. Pantel, "Dirt @sbt@discovery of inference rules from text," in *KDD*, 2001.
[6] M. Pasca and B. Van Durme, "What you seek is what you get: Extraction of class attributes from query logs." in *IJCAI*, vol. 7, 2007, pp. 2832–2837.
[7] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
[8] W. Wu, H. Li, H. Wang, and K. Zhu, "Towards a probabilistic taxonomy of many concepts," *Microsoft Res. Redmond, WA, USA, Tech. Rep. MSR-TR-2011-25*, 2011.
[9] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.
[10] Z. Wang, H. Wang, J.-R. Wen, and Y. Xiao, "An inference approach to basic level of categorization," in *CIKM*, 2015.
[11] H. Liu and P. Singh, "Omcsnet: A commonsense inference toolkit," 2003.
[12] N. Tandon, G. de Melo, F. M. Suchanek, and G. Weikum, "Webchild: harvesting and organizing commonsense knowledge from the web," in *WSDM*, 2014.
[13] S. D. Richardson, W. B. Dolan, and L. Vanderwende, "Mindnet: Acquiring and structuring semantic information from text," in *COLING-ACL*, 1998.