Knowledge-Based Systems 108 (2016) 125–131

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Extracting location and creator-related information from Wikipedia-based information-rich taxonomy for ConceptNet expansion

Marek Krawczyk*, Rafal Rzepka, Kenji Araki

Graduate School of Information Science and Technology Kita-ku, Hokkaido University, Kita 14, Nishi 9, Sapporo, Japan

ARTICLE INFO

Article history:
Received 14 November 2015
Revised 23 April 2016
Accepted 6 May 2016
Available online 9 May 2016

Keywords:
Common sense knowledge
Knowledge extraction
Conceptet

ABSTRACT

Our research goal is to generate new assertions suitable for introduction to the Japanese part of the ConceptNet common sense knowledge ontology. In this paper we present a method for extracting IsA assertions (hyponymy relations), AtLocation assertions (informing of the location of an object or place), LocatedNear assertions (informing of neighboring locations) and CreatedBy assertions (informing of the creator of an object) automatically from Japanese Wikipedia XML dump files. We use the Hyponymy extraction tool v1.0, which analyzes definition, category and hierarchy structures of Wikipedia articles to extract IsA assertions and produce an information-rich taxonomy. From this taxonomy we extract additional information, in this case AtLocation, LocatedNear and CreatedBy types of assertions, using our original method. The presented experiments prove that we achieved our research goal on a large scale: both methods produce satisfactory results, and we were able to acquire 5,866,680 IsA assertions with 96.0% reliability, 131,760 AtLocation assertion pairs with 93.5% reliability, 6217 LocatedNear assertion pairs with 98.5% reliability and 270,230 CreatedBy assertion pairs with 78.5% reliability. Our method surpassed the baseline system in terms of both precision and the number of acquired assertions.

1. Introduction

The effectiveness of systems dealing with textual-reasoning tasks depends on the scope of the large-scale general knowledge bases they utilize. A few examples of such bases include Cyc [1], YAGO [2] and ConceptNet [3]. In this paper we will focus on the last of these three - ConceptNet, a knowledge representation project that provides a large semantic graph describing general human knowledge. We have chosen ConceptNet for its superiority in key aspects: it captures a wide range of common sense concepts and relations, and its simple semantic network structure makes it easy to use and manipulate [4]. ConceptNet was designed to contain knowledge collected by the Open Mind Common Sense project's website [5]. Later versions incorporated knowledge from similar websites and online word games which automatically collect general knowledge in several languages. The current goal of ConceptNet is to expand the knowledge base with data mined from Wiktionary[1] [6] and Wikipedia[2] [7]. This open-source knowledge base is used for many applications such as topic-gisting [8], affect-sensing [9], dialog systems [10], daily activities recognition [11], social media analysis [12] and handwriting recognition [13]. ConceptNet is also applied to open-domain sentiment analysis as an integral element of a common and common sense knowledge core, which is then transformed into more compact multidimensional vector space [14]. Manual expansion of the knowledge base would be a long and labor-intensive process, as seen in nadya.jp [15], an online project that aims to gather knowledge by using a game with a purpose [16]. Since its launch in 2010, nadya.jp has been able to introduce a little over 43,500 entries to ConceptNet. It is therefore evident that we need to employ automatic methods to gather new data.

Projects such as NELL [17] or KNEXT [18] aim to extract semantic assertions from unstructured text data found on the Internet. Alternatively, we could transfer information from the existing semi-structured sources into a knowledge base. As a considerable amount of human validation has already been involved in the process of creating such sources, the reliability of information gathered in this way would be considerably higher. Wikipedia is probably the best example of an open-source, large-scale information pool. Apart from the previously-mentioned YAGO, DBpedia project also aims to transfer knowledge gathered in Wikipedia into a more formalized, digitally processable form [19]. English part of DBpedia has already been merged to ConceptNet, however the Japanese part has not been transferred yet, leaving this part of the

* Corresponding author.
E-mail addresses: marek@ist.hokudai.ac.jp (M. Krawczyk), rzepka@ist.hokudai.ac.jp (R. Rzepka), araki@ist.hokudai.ac.jp (K. Araki).

[1] A multilingual, web-based free content dictionary.

[2] A free-access, free content Internet encyclopedia.

http://dx.doi.org/10.1016/j.knosys.2016.05.004
0950-7051/© 2016 Elsevier B.V. All rights reserved.

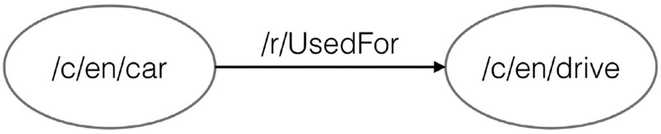

Fig. 1. Example of a single edge connecting two nodes. Symbols between slashes indicate the role and language of the respective items - 'c' stands for concept and 'r' for relation.

knowledge base at the size of roughly 1/10th of the English language domain. The problem with using the DBpedia repository is that the information gathering algorithms used to prepare the knowledge base were designed for multilingual input processing and therefore introduce a considerable amount of noise. As the knowledge gathered in ConceptNet is in large part language-specific, it is vital to widen the scope of the Japanese part independently.

The current paper elaborates on the efforts of [20]. We extended the scope of acquired assertions and explored the possibilities of deriving common sense knowledge from instance-related information triplets.

2. Graph structure of ConceptNet

In order to discuss the proposed method for expanding ConceptNet, it is necessary to introduce some basic information about the ontology's structure. ConceptNet is a network of nodes and the edges that connect them [21]. Each node is a concept described by a singe word, a word sense or a short phrase written in a natural language. Edges, as mentioned before, are the connections established between the nodes (Fig. 1 shows an example edge). The fundamental element of an edge is a relation: a codified description of a relationship between the two connected nodes. A few main examples of relations present in ConceptNet include a general RelatedTo relation, hierarchical IsA relation, PartOf, UsedFor, AtLocation, LocatedNear, HasProperty, CreatedBy, TranslationOf, etc. In total there are 52 kinds of relations. Each edge also contains information about sources of the underlying relation, surface text describing this relation and other additional features. One or more edges create an assertion - the proposition expressed by a relation between two concepts. Our goal is to find data to create new edges for the graph, which would lead to the establishment of new, meaningful assertions about the surrounding reality.

3. Hyponymy relation as IsA relation

In our approach we use the Hyponymy extraction tool v1.0 [22], an open-source program for extracting hyponymy relation pairs from Wikipedia's XML dump files. The tool has been developed specifically to process Japanese language entries. It consists of four modules, three of which deal with extraction of hyponymy pairs from different parts of Wikipedia content: definition, category and hierarchy structures [23]. The program utilizes the Pecco library [24] (SVM-like machine learning tool) to assess the plausibility level of the extracted hyponymy relation pairs and boost the precision and recall of the system [25]. The extracted hyponymy pairs may be transferred to ConceptNet as two concepts related to each other by IsA relationship (Table 1 lists examples of the extracted pairs). According to [26] these pairs are not informative enough to be useful for NLP tasks such as Question Answering; however they do fall into the scope of ConceptNet, a domain representing common sense and general knowledge. They are simple enough not to interfere with the ConceptNet's usage flexibility, yet informative enough to introduce new and valuable input to the knowledge base.

Table 1
Examples of extracted 'IsA' relationship pairs.

Hypernym	Hyponym
kouen[3] (park)	*Motomiya-kouen* (Motomiya Park)
koukyou-shisetsu (public institution)	*roujin-fukushi-sentaa* (welfare center for the elderly)
kougu (tool)	*baisu* (vice)
saiji (festival)	*unagi-matsuri* (eel festival)
Werudaa Bureemen-no senshu (Werder Bremen player)	Klaus Allofs
Nihon-no futsuu kitte (Japanese definitive stamp)	*dai-ni-ji Shouwa kitte* (second Showa stamp)
Nihon-no SF shousetsu (Japanese SF novel)	*Maikai Suikoden* (Hell's Water Margin)
josei (female)	*Sakurai Ikuko*

4. Extracting other relations

The fourth module of the Hyponymy extraction tool v1.0 generates intermediate concepts of hyponymy relations using the output of the first three modules [26]. The tool executes the following procedure: first it acquires basic hyponymy relations from Wikipedia using the method proposed by [25]. Next, it augments each acquired hypernym with the title of the Wikipedia article from which the basic hyponymy relation was extracted and consolidates the basic hypernym with the newly generated augmented hypernym (so-called 'T-INTER'). Finally, it generates an additional intermediate concept ('G-INTER') by generalizing the enriched hypernym. As a result, it acquires four-level, information-rich hyponymy relations. We can envisage the procedure producing even more additional intermediate concepts by generalizing G-INTER, and further generalizing over acquired concepts. However, it would be difficult to decide the depth to which these generalizations should continue, and therefore the choice to make one generalization seems reasonable from the point of view of output data size. In cases where such further generalizations are required, they could be achieved by traversing the graph structure of ConceptNet.

Examples of augmented hyponymy relations include: *tojo-jinbutsu* (character) – *SF eiga no tojo-jinbutsu* (character of SF movie) – *WALL-E no tojo-jinbutsu* (character of WALL-E) – M.O; *seihin* (product) – *kigyo no seihin* (product of a company) – *Silicon Graphics no seihin* (product of Silicon Graphics, Inc.) – IRIS Crimson; *sakuhin* (work) – *America no shosestu-ka no sakuhin* (work of American novelist) – *J.D. Salinger no sakuhin* (work of J.D. Salinger) – A boy in France; *machi* (town) – *England no shu no machi* (town in a county in England) – *East Sussex no machi* (town in East Sussex) – Uckfield. As we can see from the examples, the generated augmented hypernyms are too specific to be incorporated into ConceptNet directly. However some additional information about their corresponding hyponyms may be extracted from them, such as information concerning location, neighboring locations, creator and so on. Knowledge about location and creator may be directly transferred into ConceptNet through already built-in AtLocation, LocatedNear and CreatedBy relations. It should be noted that according to the ConceptNet documentation [27] the CreatedBy relation relates to processes, however inspection of the existing CreatedBy assertions show that they include creations and their authors as well. The remaining part of the acquired information related to the hyponyms may be represented by a more general RelatedTo relation.

[3] All Japanese language phrases are transliterated and written in italics.