Extracting Location and Creator-related Information from Wikipedia-based Information-rich Taxonomy for ConceptNet Expansion

Marek Krawczyk, Rafal Rzepka, Kenji Araki

Hokkaido University, Graduate School of Information Science and Technology
Kita-ku, Kita 14, Nishi 9, Sapporo, Japan
{marek,rzepka,araki}@ist.hokudai.ac.jp

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, 6,217 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.

Keywords: common sense knowledge, knowledge extraction, ConceptNet

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

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

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

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 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 (Figure 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.

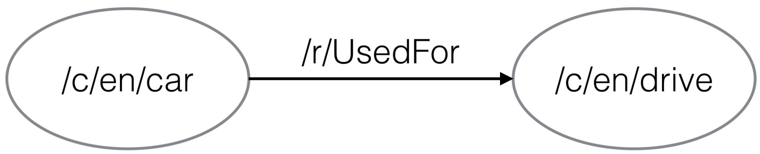

Figure 1: Example of a single edge connecting two nodes. Symbols between slashes indicate the role and language of the respective items - 'c' stands for concept and 'r' for relation.

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

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

[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.

The procedure of acquiring additional information is presented in Figure 2 and exemplified in Figure 3. First (Step 1), we scan the G-INTER using our handcrafted primary rule base in search of tags referring to locations or creators, for example [city], [district], [cartoonist], [writer] and so on. In the case of acquiring LocatedNear pairs, we confirm that the basic hypernym contains a marker indicating physical proximity (such as the Chinese character meaning 'neighboring'). Next (Step 2), we filter the basic hypernym through a secondary rule base to exclude items that would introduce noise. For example, we can extract information about the birthplaces of famous people; however this does not mean that we can build an AtLocation kind of relationship between the person and his or her birthplace. If so, hypernyms indicating people are excluded from the analysis of location. When analysing LocatedNear pairs we filter out ambiguous items. If the basic hypernym is positively assessed by the secondary rule base, then (Step 3) we assume that the phrase acquired by deleting the basic hypernym from the G-INTER is a valid location or creator tag. Using the first example from Figure 2, we check that 'county in England' is a valid tag to describe a location. In the next stage (Step 4) we compare the validated location or creator tag with the content of the T-INTER. This way, using the previous example, we can extract the knowledge that the county we refer to is East Sussex. Finally (Step 5), we join the newly acquired information to the base hyponym with a proper relationship tag to extract a new relation, for example Uckfield-AtLocation-East Sussex.

The effectiveness of the method mainly depends on the number and nature of introduced rules to both the primary and secondary rule bases. Our method is still work in progress, and at this stage we used 55 primary rules and 14 secondary rules, which allowed us to extract assertions concerning location, neighboring locations and creators. The manually crafted rules have been

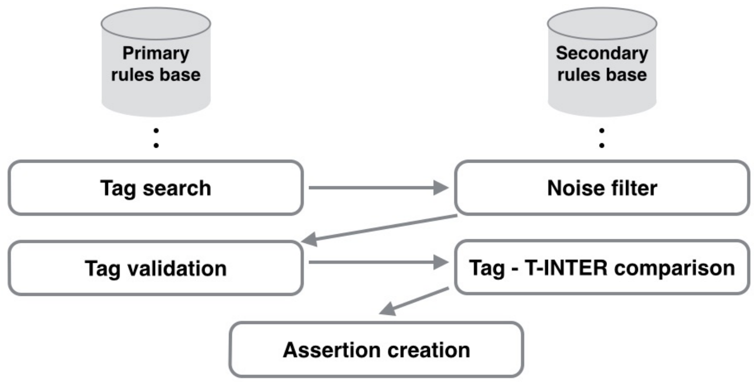

Figure 2: Flowchart of our proposed method.

created using heuristics after analysis of the input data. The reason why we chose this kind of approach is because the information units contain Chinese characters indicating a type of location, a city, province, school or a creator. We use the rules to detect these characters, and this way we are able to obtain the named entities referring to locations and creators. Due to the qualities of the Japanese language's writing system these rules are often very simple, containing a single character, but are still effective for detecting the language units we want to extract. For example, the secondary rules used for detecting people include the suffix '∼sha', which describes different professions. For English such a shortcut would be harder to apply, and therefore person detection would require a much larger rule base covering a long list of names of professions and appropriate suffixes (like '∼er', '∼or' or '∼ist').

However, our experiments revealed that extracting creator information is more complex and creates some challenges. While extracting location-related information, the introduced rules may be simple and straightforward. In the case of creators, the rules not only have to cover the qualities of the writing system, but also take into consideration the importance of particular roles while creating a given piece of work. For example our annotators indicated that a number of professionals taking part in the creation of films may not be con-

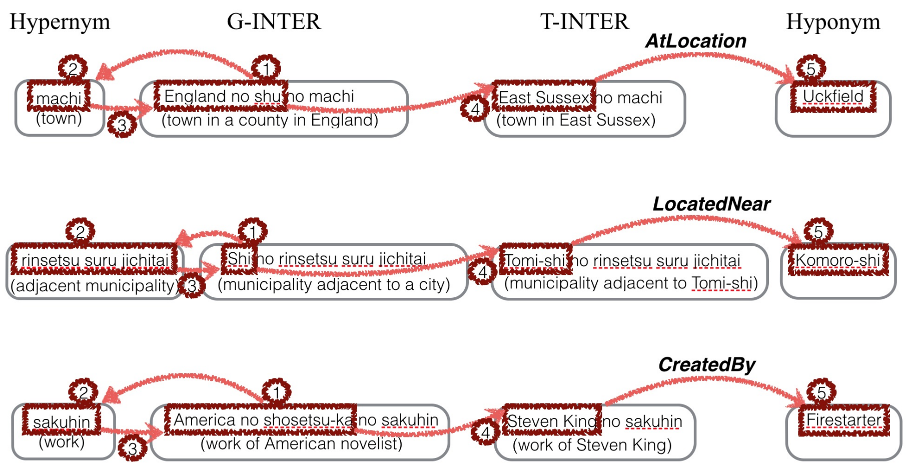

Figure 3: Procedure of our proposed method exemplified on the extracted relations.

sidered as the creators of these films. Actors, actresses and voice actors, even if they make a great contribution to the work, should not be labeled as its creators. Further experiments showed that similarly animators, animation directors, sound directors, and storyboard creators do not qualify to be included in the common sense CreatedBy assertions.

In future we would like to investigate the possibility of combining heuristics with automated rule discovery methods in order to achieve higher precision and recall. The number and reliability level of the data acquired with our method is presented in the Evaluation section.

5. Evaluation

We used the 2014-11-04 version of the Japanese Wikipedia dump data to verify the reliability level declared by Sumida [25] and evaluate our proposed method for obtaining additional relations. We ran the definition, category and hierarchy modules of the Hyponymy extraction tool v1.0 at 93% precision rate using the biggest available training set, and obtained 6,014,194 hypernym-hyponym pairs. The number of unique hyponymy pairs was 5,866,680, which indicates that 147,514 pairs have been extracted by more than one module. The 93% reliability level declared by the authors of the method has been verified by three human annotators, whose task was to evaluate a sample of the data and decide whether the extracted pairs a) represent a correct hyponymy relation, b) represent related concepts, but not in a hyponymy relation, or c) represent unrelated concepts. The annotators assigned 1, 0.5 and 0 points respectively to 300 randomly selected assertions. We decided to assign 0.5 points to related concepts as they may be used to create correct assertions (see Future Work section). If two or more annotators assessed an item as belonging to one category, their decision was regarded as the evaluation output. In cases where their decisions varied (which happened 10 times), the first author decided the score. The procedure follows a modified Sumida *et al.* [25] evaluation method.

Table 2 presents the evaluation results. 283 pairs were assessed as representing a correct hyponymy relation, 10 pairs as related concepts, but not in a hyponymy relation and 7 as unrelated concepts.This results in 96.0% precision value of the tested sample, which surpasses the 93% declared by Sumida *et al.* The level of overall agreement between annotators was 86.9%, and the Kappa value[4] was 0.80, which indicates that the annotation judgement was in substantial agreement [28].

Table 2: Evaluation results for IsA relations.

Correct hyponymy	Related concepts	Unrelated concepts	Precision	Total number of pairs
0.943 (283/300)	0.033 (10/300)	0.023 (7/300)	0.960	5,866,680

[4]To measure the agreement level between judges, we used Randolph's free marginal multirater kappa instead of Fleiss' fixed-marginal multirater kappa, due to high agreement low kappa paradox.

Running the fourth 'extended' module of the Hyponymy extraction tool v1.0 on the same Wikipedia dump data resulted in obtaining 2,738,211 basic hypernym–G-INTER–T-INTER–basic hyponym sets. By applying our method for obtaining additional information, we were able to produce 131,760 pairs representing AtLocation relation, 6,217 pairs representing LocatedNear relation and 270,230 pairs representing CreatedBy relation. For comparison, nadya.jp, the baseline system, has provided only 8,706 AtLocation relations and no LocatedNear or CreatedBy relations in four years of its operation. In the case of AtLocation pairs, we evaluated 100 pairs[5] randomly selected from our method's output and 100 pairs randomly selected from nadya.jp's AtLocation assertions [16]. While evaluating LocatedNear and CreatedBy relations, a comparison with the baseline was not possible, as ConceptNet 5.3 does not yet contain any LocatedNear or CreatedBy pairs in its Japanese language section. These assertions were therefore evaluated independently. The evaluation procedure follows the previously applied one: 1 point being applied to correct AtLocation, LocatedNear or CreatedBy assertions, 0.5 point to related concepts, but not in the evaluated relation, and 0 points to unrelated concepts. In 13 cases the annotators' evaluation was inconsistent, and therefore the first author decided the score.

Table 3 shows the evaluation results of our AtLocation pairs generation method in comparison with the baseline system. 88 pairs generated by our method were evaluated as representing a correct AtLocation relation, 11 pairs as related concepts, but not in an AtLocation relation, and 1 as unrelated concepts. This results in a 93.5% precision value. In the case of the baseline system, 64 pairs were evaluated as correct AtLocation assertions, 20 as related concepts, but not in an AtLocation relation, and 16 as unrelated concepts. The precision value for the baseline system is 74.0%. The level of overall agreement between annotators was 73.6% and the Kappa value was 0.60, which indicates that the

[5] We adjusted the number of evaluated pairs to balance the proportion between the total number of pairs and the test sample.

annotation judgment was in moderate agreement. Examples of the extracted AtLocation assertions are presented in Table 4.

Table 3: Evaluation results for AtLocation relations in comparison with the nadya.jp baseline.

	Correct AtLocation	Related concepts	Unrelated concepts	Precision	Total number of pairs
Proposed	0.880 (88/100)	0.110 (11/100)	0.010 (1/100)	0.935	131,760
Baseline	0.640 (64/100)	0.200 (20/100)	0.160 (16/100)	0.740	8,706

$p < 0.001$, t-score = 4.6291

Table 5 contains the evaluation result of the generated LocatedNear relations. 97 pairs were evaluated as correct LocatedNear pairs, 3 as related concepts and none as unrelated concepts, which results in 98.5% precision. The level of overall agreement between annotators was 86.6% and the Kappa value was 0.80, which indicates that the annotation judgment was in substantial agreement. Examples of the extracted LocatedNear assertions are presented in Table 6.

Table 7 contains the evaluation result of the generated CreatedBy relations. 60 pairs were evaluated as correct CreatedBy pairs, 37 as related concepts and 3 as unrelated concepts, which results in 78.5% precision. The level of overall agreement between annotators was 71.6% and the Kappa value was 0.57, which indicates that the annotation judgment was in moderate agreement. Examples of the extracted LocatedNear assertions are presented in Table 8.

The reason for the relatively low precision score of the assessed CreatedBy assertions is as follows: in 24 cases it was the annotators' opinion that actors, voice actors, animators, storyboard creators or sound directors cannot be considered as creators of works they contribute to. Although it would be valid to include such persons in the RelatedTo kind of relationship with the work they

Table 4: Examples of generated AtLocation assertions.

Tomato Ginkou (Tomato Bank)	AtLocation	*Okayama-shi* (Okayama city)
Mariina Oudouri (Marina Boulevard)	AtLocation	A Coruna
Warren Shinrin-kyoku Kukou (Warren USFS Airport)	AtLocation	*Aidaho-gun* (Idaho County)
Hoshinomiya Jinja (Hoshinomiya Temple)	AtLocation	*Minami-mura* (Minami village)
Otao hoikuen (Outao nursery)	AtLocation	*Sakai-shi* (Sakai city)
Shindzutsumi Shizen Kouen (Shinzutsumi nature park)	AtLocation	*Kurihara-shi* (Kurihara city)
Sandifukku (Sandy Hook)	AtLocation	*Eriotto-gun* (Elliott County)
Hoteru Kadoya (Kadoya Hotel)	AtLocation	*Tochigi-shi* Tochigi city)

Table 5: Evaluation results for LocatedNear relations

Correct Located-Near	Related concepts	Unrelated concepts	Precision	Total number of pairs
0.970 (97/100)	0.030 (3/100)	0.000 (0/100)	0.985	6,217

helped to create, defining them as creators would go against common sense. This is a valid observation and it will be taken into consideration when re-designing

Table 6: Examples of generated LocatedNear assertions.

Ougoe-machi (Ogoe city)	LocatedNear	*Ono-machi* (Ono city)
Iseri-gawa (Iseri river)	LocatedNear	*Konoha-gawa* Konoha river
Shin Edo-gawa Kouen (New Edo River Park)	LocatedNear	*Koudansha Noma Kinenkan* (Kodansha Noma Memorial Museum)
Daiting	LocatedNear	Monheim
Sahoro Yuusu Hosteru (Sahoro Youth Hostel)	LocatedNear	*Obihiro Yachiyo Yuusu Hosteru* (Obihiro Yachiyo Youth Hostel)
Kumotori-yama (Mount Kumotori)	LocatedNear	*Karamatsuo-yama* (Mount Karamatsuo)
Goshogawara-shi (Goshogawara city)	LocatedNear	*Sotogahama-machi* (Sotogahama town)
Gujou Keisatsujo (Gujou Police Station)	LocatedNear	*Ouno Keisatsusho* (Ohno Police Station)

and expanding the rule base for the next version of the algorithm. There were also cases of assertions assessed as invalid due to errors passed from the output of the Hyponymy extraction tool to the proposed method. Table 9 contains examples of assertions that were assessed as erroneous by the annotators.

The results show that IsA relation pairs generated by the definition, cate-

Table 7: Evaluation results for CreatedBy relations

Correct CreatedBy	Related concepts	Unrelated concepts	Precision	Total number of pairs
0.600 (60/100)	0.370 (37/100)	0.030 (3/100)	0.785	270,230

Table 8: Examples of generated CreatedBy assertions.

Dark Horse	CreatedBy	George Harrison
Kaze (Wind)	CreatedBy	*Kubota Koutarou*
Manuke-na Oukami (Sheep Wrecked)	CreatedBy	Michael Lah
The Point of View	CreatedBy	Alan Crosland
Boom, Boom, Boom, Boom!!	CreatedBy	Vengaboys
Genki-na Buroukun Haato (Healthy Broken Heart)	CreatedBy	*Matsumoto Takashi*
Haru-no Hi (Spring Day)	CreatedBy	*Watanabe Takuya*
When the Birds Fly South	CreatedBy	Stanton A. Coblentz

gory and hierarchy of the Hyponymy extraction tool v1.0, as well as AtLocation and LocatedNear relation pairs extracted by our proposed method may be incorporated into ConceptNet. Considering the number of the newly acquired assertions as well as reliability of the data in comparison with the resources already present in the knowledge base, such operation would be beneficial for ConceptNet. CreatedBy relation pairs could also be added after the revision of introduced rules and a substantial increase of the precision rate.

Table 9: Examples of erroneous CreatedBy assertions.

Shishi-no ketsumyaku (Lion bloodline)	CreatedBy	*Ozawa Hitoshi* (actor)
Road 88	CreatedBy	*Tomita Yasuko* (actress)
Tsurupika Hagemaru (Little Baldy Hagemaru)	CreatedBy	*Zen Souichirou* (storyboard creator)
Kaiketsu Zorori (Incredible Zorori)	CreatedBy	*Yamada Etsuji* (sound director)
Kishin Douji Zenki (Zenki)	CreatedBy	*Hayashi Akemi* (animator)
Human (incomplete name error)	CreatedBy	Nicholson Baker

6. Discussion: case study of potential application

In order to verify the potential applicability of the acquired data to a working system, we considered a book recommendation system scenario. The reason for choosing such an approach is that recommendation systems are usually knowledge-based and, especially at the beginning of the operation, suffer from an insufficient amount of available data vectors [29]. We considered a Japanese book recommendation system currently being created at Hokkaido University. The system is being designed to consist of five modules, each performing book recommendation based on a different set of data: attributes (title, author, publisher, sales date, genre, price), content description, users' reviews, Amazon sales-based suggestions, and attributes plus reviews. A preliminary survey performed among the system's test users revealed that the attribute-based module represents the lowest reliability: the test users' opinions suggested that recommendations made on the basis of the authors' name and title similarity were very often misleading. However, to improve the effectiveness of attribute-based recommendation, the system could be provided with more input for building additional vectors. Therefore we decided to verify whether the data extracted by our method could potentially be applied for this purpose. We took the system's working data, consisting of 106,415 book titles accompanied with authors' names. The data was gathered from the Amazon Japan website [30]. In order to test our data against books that are popular in Japanese society, we have filtered out texts which had less than 30 reviews at a Japanese book review sharing site, Dokusho Meter [31]. By doing so we received a list of 14,055 book titles accompanied by their 18,988 authors' names. We created and ran a script to search the title and author data using the IsA and CreatedBy relation pairs. As a result we were able to find additional information about the author or authors of 13,007 books (92.5% of the studied sample), to be more precise, concerning 15,685 authors' names (82.6%). The additional information includes other works created by the authors, the authors' place of birth, occupations and other characteristics included in the IsA and CreatedBy relation bases. These clues may be used to create more detailed profile of each author, which could be utilized when comparing them with other authors to make book recommendations. We also extracted further information concerning the title of 538 volumes (3.8%). In total we were able to provide the system with additional, useful information concerning 13,038 positions, which is 92.7% of the analyzed sample. Each book found in our data received an average of 28 additional information vectors. On the basis of these findings, we could propose a hypothesis that the data acquired by our method have a strong potential for applcation to a practical use. As the approach of the creators of the discussed book recommendation system is to move away from conventional collaborative filtering to more complex and innovative semantic feature analysis-based recommendation, the data produced by our method would provide the fundamental element necessary for realizing that approach. Proving the aforementioned hypothesis, however, would have to be the object of a separate, extensive study performed upon the completion of the current system.

7. Generalizing over assertions

Wikipedia contains a lot of information about instances of certain concepts, such as Salvador Dali as an instance of an artist. Filling up ConceptNet with instances is a valid task, as it is very hard to establish the boundaries of common sense knowledge – facts that are obvious to one group of people overlap to a large proportion with the knowledge of another group, but there is always a discrepancy. This issue raises a question: would it be possible to come to more general conclusions on the basis of the numerous instances? In order to solve this problem we created and performed an initial test of the following method: we took each of the additional information lists (representing LocatedAt, LocatedNear and CreatedBy relations) and analyzed each assertion one by one. For both concepts in the assertion we found their hypernyms in the generated IsA relations list. Next, we generated assertions representing all possible combinations between concept A's hypernyms and concept B's hypernyms. We repeated the process for all assertions in the additional information list and calculated the generated hypernym assertions' occurrence frequency. As predicted, the assertions with the highest occurrence frequency represent general, common sense observations. This is true for AtLocation and CreatedBy lists, but it is not the case when processing the LocatedNear list, because of the relatively low number of LocatedNear assertions. It became apparent that the higher number of initial assertions increases the probability of generating meaningful general assertions. See Table 10 for the examples of generated general assertions. The procedure requires further development in terms of the method for frequency calculations and automatic filtering of non-general assertions.

8. Conclusion

In this paper we presented a method for automatic acquisition of common sense knowledge triplets from the Japanese Wikipedia. It allowed us to mine IsA, AtLocation, LocatedNear and CreatedBy assertions with precision estimated at the levels of 96.0%, 93.5%, 98.5% and 78.5% respectively. We also demonstrated

Table 10: Examples of generated general assertions.

toshi oyobi machi (city and town)	AtLocation	*gun* (province)
shougakkou (elementary school)	AtLocation	*machi* (city)
douro (road)	AtLocation	*machi* (city)
sakuhin (work)	CreatedBy	*zonmei jinbutsu* (living person)
anime sakuhin (anime)	CreatedBy	*anime kankeisha* (people involved in making anime)
shutsuen sakuhin (performance art)	CreatedBy	*bunkajin* (cultural figure)

a case study of a practical use of the acquired data, as well as the possibility of formulating common sense assertions on the basis of generated instances data. As the Japanese part of the current ConceptNet 5.3 consists of 1,071,046 assertions, a contribution of 6,274,887 new assertions would be significant. It would mean an almost sixfold increase and could potentially make ConceptNet applicable to many Japanese language analysis problems. Moreover, as Wikipedia is a constantly expanding source, we could acquire more assertions simply by applying our method to the updated Wikipedia XML dump files.

The applicability of ConceptNet is not limited to any particular branch of data analysis. Therefore we could speculate that the results of our method may not only augment the effectiveness and scope of already created tools, but also may contribute to the development of new directions and approaches, as depicted by the presented book recommendation system example.

9. Future work

In order to extend the functionality of our proposed method, we intend to update the primary and secondary rules, which would allow the system to increase its precision and the scope of extracted information. We would also like to explore the possibility of using a machine learning algorithm for automatic rule generation combined with the already present heuristics. Such a combination could potentially be more effective in increasing precision and recall, as well as finding new rules to extract even more relations.

We also plan to create an interface for the evaluation of the method's output by Japanese native speakers, which would allow us to utilize the pairs representing related concepts.

10. Acknowledgements

The authors would like to thank Shouhei Kitahara from Hokkaido University for providing the dataset used in the presented potential application case study.

References

[1] D. B. Lenat, Cyc: A large-scale investment in knowledge infrastructure, Communications of the ACM 38 (11) (1995) 33–38.

[2] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: Proceedings of the 16th international conference on World Wide Web, ACM, 2007, pp. 697–706.

[3] H. Liu, P. Singh, Conceptneta practical commonsense reasoning tool-kit, BT technology journal 22 (4) (2004) 211–226.

[4] H. Liu, P. Singh, Commonsense reasoning in and over natural language, in: Knowledge-based intelligent information and engineering systems, Springer, 2004, pp. 293–306.

[5] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, W. L. Zhu, Open mind common sense: Knowledge acquisition from the general public, in: On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE, Springer, 2002, pp. 1223–1237.

[6] Wiktionary, Website, last checked: 13.11.2015.
URL `https://www.wiktionary.org/`

[7] Wikipedia, Website, last checked: 13.11.2015.
URL `https://www.wikipedia.org/`

[8] R. H. Speer, C. Havasi, K. N. Treadway, H. Lieberman, Finding your way in a multi-dimensional semantic space with luminoso, in: Proceedings of the 15th international conference on Intelligent user interfaces, ACM, 2010, pp. 385–388.

[9] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Senticspace: visualizing opinions and sentiments in a multi-dimensional vector space, in: Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2010, pp. 385–393.

[10] S. J. Korner, T. Brumm, Resi-a natural language specification improver, in: Semantic Computing, 2009. ICSC'09. IEEE International Conference on, IEEE, 2009, pp. 1–8.

[11] J. Ullberg, S. Coradeschi, F. Pecora, On-line adl recognition with prior knowledge, in: Proceedings of the 2010 conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers' Symposium, IOS press, 2010, pp. 354–366.

[12] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Sentic computing: exploitation of common sense for the development of emotion-sensitive systems, in: Development of Multimodal Interfaces: Active Listening and Synchrony, Springer, 2010, pp. 148–156.

[13] Q.-F. Wang, E. Cambria, C.-L. Liu, A. Hussain, Common sense knowledge for handwritten chinese text recognition, Cognitive Computation 5 (2) (2013) 234–242.

[14] E. Cambria, Y. Song, H. Wang, N. Howard, Semantic multidimensional scaling for open-domain sentiment analysis, Intelligent Systems, IEEE 29 (2) (2014) 44–51.

[15] Naadya to nazonazo: Conceptnet, Website, last checked: 13.11.2015. URL `http://nadya.jp/`

[16] K. Nakahara, S. Yamada, Development and evaluation of a web-based game for common-sense knowledge acquisition in japan, in: Unisys Technology Review no. 107, 2011, pp. 295–305.

[17] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, T. M. Mitchell, Toward an architecture for never-ending language learning., in: AAAI, Vol. 5, 2010, p. 3.

[18] L. Schubert, Can we derive general world knowledge from texts?, in: Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., 2002, pp. 94–97.

[19] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, Dbpedia spotlight: shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems, ACM, 2011, pp. 1–8.

[20] M. Krawczyk, R. Rzepka, K. Araki, Extracting conceptnet knowledge triplets from japanese wikipedia, in: Proceedings of the 21st Annual Meeting of The Association for Natural Language Processing, 2015, pp. 1052–1055.

[21] R. Speer, C. Havasi, Representing general relational knowledge in conceptnet 5., in: LREC, 2012, pp. 3679–3686.

[22] Version1.0 : Hyponymy extraction tool, Website, last checked: 13.11.2015. URL https://alaginrc.nict.go.jp/hyponymy/

[23] A. Sumida, K. Torisawa, Hacking wikipedia for hyponymy relation acquisition., in: IJCNLP, Vol. 8, Citeseer, 2008, pp. 883–888.

[24] N. Yoshinaga, pecco - c++ library for efficient classification with conjunctive features, Website, last checked: 13.11.2015.
URL http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/pecco/

[25] A. Sumida, N. Yoshinaga, K. Torisawa, Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in wikipedia., in: LREC, 2008.

[26] I. Yamada, C. Hashimoto, J.-H. Oh, K. Torisawa, K. Kuroda, S. De Saeger, M. Tsuchida, J. Kazama, Generating information-rich taxonomy from wikipedia, in: Universal Communication Symposium (IUCS), 2010 4th International, IEEE, 2010, pp. 97–104.

[27] R. Speer, Relations . commonsense/conceptnet5 wiki . github, Website, last checked: 13.11.2015.
URL https://github.com/commonsense/conceptnet5/wiki/Relations

[28] J. J. Randolph, Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa., Online Submission.

[29] A. I. Schein, A. Popescul, L. H. Ungar, D. M. Pennock, Methods and metrics for cold-start recommendations, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2002, pp. 253–260.

[30] Amazon.co.jp, Website, last checked: 13.11.2015.
URL http://www.amazon.co.jp/

[31] Dokusho meter, Website, last checked: 13.11.2015.
URL http://bookmeter.com/