Exploiting Wikipedia-based Information-rich Taxonomy for Extracting Location and Creator Related Information for ConceptNet Expansion

Marek Krawczyk, Rafal Rzepka, Kenji Araki

Hokkaido University
Kita-ku, Kita 14, Nishi 9
Sapporo, Japan
{marek, rzepka, araki}@ist.hokudai.ac.jp

Abstract

In this paper we present a method of extracting IsA assertions (hyponymy relations), AtLocation assertions (informing of location of objects or places), LocatedNear assertions (informing of neighboring locations) and CreatedBy assertions (informing of creator of objects) automatically from Japanese Wikipedia XML dump files. We use the Hyponymy extraction tool v1.0, which analyses definition, category and hierarchy structures of Wikipedia articles to extract IsA assertions and produce information-rich taxonomy. From this taxonomy we extract additional information, in this case AtLocation, LocatedNear and CreatedBy type of assertions, using our original method. Presented experiments prove that both methods produce satisfactory results: we were able to acquire 5,866,680 IsA assertions with 99% reliability, 131,760 AtLocation assertion pairs with 93% reliability, 6,217 LocatedNear assertion pairs with 99% reliability and 270,230 CreatedBy assertion pairs with 80% 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 large-scale general knowledge bases they utilize. Just to enumerate few examples of such bases we could mention Cyc (Lenat, 1995), YAGO (Suchanek et al., 2007) and ConceptNet (Liu and Singh, 2004). In this paper we will focus on the last of the three - ConceptNet, a knowledge representation project that provides a large semantic graph describing general human knowledge. ConceptNet was designed to contain knowledge collected by Open Mind Common Sense project's website (Singh et al., 2002). Further releases incorporated knowledge from similar websites and online word games which automatically collect general knowledge in several languages. Current goal of ConceptNet is to expand the knowledge base with data mined from Wiktionary[1], a multilingual, web-based free content dictionary, and Wikipedia[2], a free-access, free content Internet encyclopedia. This open-source knowledge base is used for many applications such as topic-gisting (Speer et al., 2010), affect-sensing (Cambria et al., 2010), dialog systems (Korner and Brumm, 2009) and so on. Manual expansion of the knowledge base would be a long and labor-intensive process, as seen in nadya.jp[3], an online project aiming at gathering knowledge by using a game with a purpose (Nakahara and Yamada, 2011). Since its launch in 2010 it was able to introduce little over 43,500 entries to the ConceptNet. It is therefore evident that we need to employ automatic methods to gather new data.

Projects such as NELL (Carlson et al., 2010) or KNEXT (Schubert, 2002) aim at extracting 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 this way would be considerably higher. Wikipedia is probably the best example of open-source, large-scale information pools. Apart from previously mentioned YAGO, DBpedia project also aims at transferring knowledge gathered in Wikipedia into more formalized, digitally processable form (Mendes et al., 2011). 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 1/10 of the English language domain. The problem with using 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 considerable proportion language-specific, it is vital to widen the scope of Japanese part independently.

The current paper elaborates on efforts of (Krawczyk et al., 2015). We extended the scope of acquired assertions as well as explored possibilities of deriving commonsense knowledge from instance related information triplets.

2. Hyponymy relation as IsA relation

In our approach we use the Hyponymy extraction tool v1.0[4], 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 (Sumida and Torisawa, 2008). The program utilizes Pecco library[5] (SVM-like machine learning tool) to

[1] http://www.wiktionary.org/
[2] http://www.wikipedia.org/
[3] http://nadya.jp/
[4] http://alaginrc.nict.go.jp/hyponymy/
[5] http://www.tkl.iis.u-tokyo.ac.jp/∼ynaga/pecco/

assess the plausibility level of the extracted hyponymy relation pairs' and boost the precision and recall of the system (Sumida et al., 2008). The extracted hyponymy pairs may be transferred to ConceptNet as two concepts related to each other by IsA relationship. According to (Yamada et al., 2010) 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 commonsense 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.

3. 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 (Yamada et al., 2010). The tool executes the following procedure: first it acquires basic hyponymy relations from Wikipedia using the method proposed by (Sumida et al., 2008). 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 additional intermediate concept ('G-INTER') by generalizing the enriched hypernym. As a result, it acquires four-level, information-rich hyponymy relations.

Examples of augmented hyponymy relations include: *tojo-jinbutsu*[6] (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, 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[7] 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 on Figure 1 and exemplified on Figure 2. First (Step 1), we scan the G-INTER using our handcrafted primary rules base in search of tags referring to locations or creators, for example [city], [district], [cartoonist], [writer] and so on. In case of acquiring LocatedNear pairs we confirm that the basic hypernym contains a marker indicating physical proximity (such as Chinese character meaning 'neighboring'). Next (Step 2), we filter the basic hypernym through a secondary rules base to exclude items that would introduce noise. For example we can extract information about the birthplace 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 rules 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 example from Figure 2, we check that 'county in England' is a valid tag to describe a location. In 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.

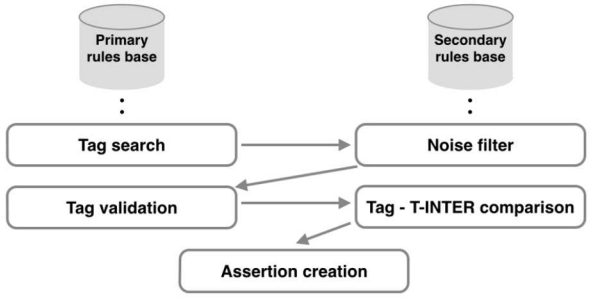

Figure 1: Flowchart of our proposed method.

The effectiveness of the method mainly depends on the number and nature of introduced rules to both primary and secondary rules base. 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 created using heuristics after the 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 get the named entities referring to locations and creators. Because of the qualities of Japanese language writing system these rules are often very simple, containing a single character, but still effective for detecting language units we want to extract. For example secondary rules used for detecting people include suffix '∼sha', which describes different professions. For English such shortcut would be harder to apply, and therefore person detection

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

[7] https://github.com/commonsense/conceptnet5/wiki/Relations

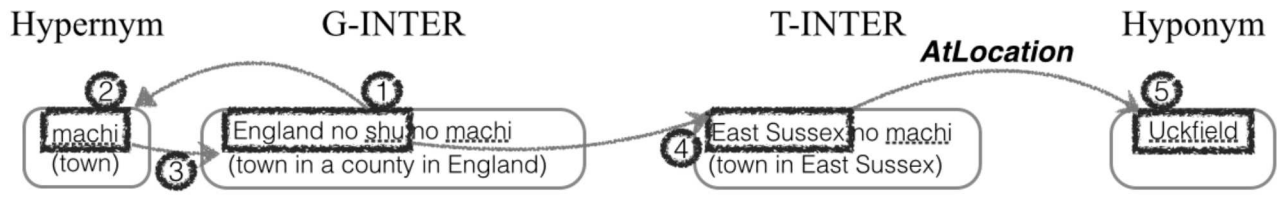

Figure 2: Procedure of our proposed method.

would require a much larger rules base covering a long list of names of professions and appropriate suffixes (like '∼er', '∼or' or '∼ist').

Our experiments revealed however 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 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 creation of films may not be considered as creators. Actors, actresses and voice actors, even if they make great contribution to the work, should not be labeled as creators. Further experiments showed that similarly animators, animation directors, sound directors, and people making storyboards do not qualify to be included in the CreatedBy assertions.

In future we would like to investigate the possibility of combining heuristics with automated rules 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.

4. Evaluation

We used the 2014-11-04 version of the Japanese Wikipedia dump data to verify the reliability level declared by Sumida (Sumida et al., 2008) and evaluate our proposed method of obtaining additional relations. We run the definition, category and hierarchy modules of the Hyponymy extraction tool v1.0 at 93% precision rate and 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 200 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 case when their decisions varied (two cases), one of the authors decided the score. The procedure follows a modified Sumida *et al.* (Sumida et al., 2008) evaluation method. Table 1 presents the evaluation results. 97 pairs were assessed as representing correct hyponymy relation, 2 pairs as related concepts, but not in a hyponymy relation and 1 as unrelated concepts.This results in 99% precision value of the tested sample, which surpasses 93% declared by Sumida *et al.* The level of overall agreement between annotators was 91% and the Kappa value[8] was 0.86, which indicates that the annotation judgement was in almost perfect agreement (Randolph, 2005).

Table 1: Evaluation results for IsA relations.

Correct hyponymy	Related concepts	Unrelated concepts	Precision	Total number of pairs
0.985 (197/200)	0.010 (2/200)	0.005 (1/200)	0.990	5,866,680

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 of 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, provided only 8,706 AtLocation relations and no LocatedNear nor CreatedBy relations in four years of its operation. In case of AtLocation pairs, we evaluated 50 pairs[9]randomly selected from our method's output and 50 pairs randomly selected from nadya.jp's AtLocation assertions (Nakahara and Yamada, 2011). While evaluating LocatedNear and CreatedBy relations, a comparison with baseline was not possible, as the current version of ConceptNet does not contain any LocatedNear nor CreatedBy pairs in its Japanese language section yet. 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 by the evaluated relation, and 0

[8]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.

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

points to unrelated concepts. In seven cases the annotators' evaluation was inconsistent, and therefore one of the authors decided the score. Table 2 shows the evaluation results of our AtLocation pairs generation method in comparison with the baseline system. 43 pairs generated by our method were evaluated as representing correct AtLocation relation, 7 pairs as related concepts, but not in an AtLocation relation. None of the pairs were assessed as unrelated concepts. This results in 93% precision value. In case of the baseline system, 32 pairs were evaluated as correct AtLocation assertions, 12 as related concepts, but not in an AtLocation relation, and 6 as unrelated concepts. The precision value for the baseline system is 76%. The level of overall agreement between annotators was 71.7% and the Kappa value was 0.57, which indicates that the annotation judgement was in moderate agreement.

Table 2: Evaluation results for AtLocation relations in comparison with nadya.jp baseline.

	Correct AtLocation	Related concepts	Unrelated concepts	Precision	Total number of pairs
Proposed	0.860 (43/50)	0.140 (7/50)	0.000 (0/50)	0.930	131,760
Baseline	0.640 (32/50)	0.240 (12/50)	0.120 (6/50)	0.760	8,706

$p = 0.003$, t-score $= 3.2097$

Table 3 contains the evaluation result of the generated LocatedNear relations. 49 pairs were evaluated as correct LocatedNear pairs, 1 as related concepts and none as unrelated concepts, which results in 99% precision. The level of overall agreement between annotators was 84% and the Kappa value was 0.76, which indicates that the annotation judgement was in substantial agreement.

Table 3: Evaluation results for LocatedNear relations

Correct LocatedNear	Related concepts	Unrelated concepts	Precision	Total number of pairs
0.980 (49/50)	0.020 (1/50)	0.000 (0/50)	0.990	6,217

Table 4 contains the evaluation result of the generated CreatedBy relations. 30 pairs were evaluated as correct CreatedBy pairs, 20 as related concepts and none as unrelated concepts, which results in 80% precision. The level of overall agreement between annotators was 82.6% and the Kappa value was 0.74, which indicates that the annotation judgement was in substantial agreement.

The reason for the relatively low precision score of the assessed CreatedBy assertions is as follows: in 15 cases it was the annotators' opinion that voice actors or animators can not be considered as creators of works they take part in. This is a valid observation and it will be taken into consideration while re-designing and expanding the rules base for the next version of the algorithm.

Table 4: Evaluation results for CreatedBy relations

Correct CreatedBy	Related concepts	Unrelated concepts	Precision	Total number of pairs
0.600 (30/50)	0.400 (20/50)	0.000 (0/50)	0.800	270,230

The results show that IsA relation pairs generated by the definition, category 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.

5. Generalizing over assertions

Wikipedia contains a lot of information about instances of certain concepts, such as Salvador Dali as an instance of a painter. Filling up ConceptNet with instances is a valid task, as it is very hard to establish the boundaries of commonsense knowledge - facts obvious for one group of people in large proportion overlap with 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's A hypernyms and concept's B hypernyms. We have 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, commonsense 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 5 for the examples of generated general assertions. The procedure requires further development in terms of the method of frequency calculations and automatic filtering of non-general assertions.

6. Conclusion

In this paper we presented a method for automatic acquisition of commonsense knowledge triplets from

Table 5: 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)

Japanese Wikipedia. It allowed us to mine IsA, AtLocation, LocatedNear and CreatedBy assertions with precision at the level of 99%, 93%, 99% and 80% respectively. We also demonstrated a possibility of formulating commonsense assertions on the basis of generated instances data. As the Japanese part of 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.

7. 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 machine learning algorithm for automatic rule generation combined with already present heuristics. Such 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 evaluation of the method's output by Japanese native speakers, which would allow us to utilize the pairs representing related concepts.

8. References

Cambria, Erik, Amir Hussain, Catherine Havasi, and Chris Eckl, 2010. Senticspace: visualizing opinions and sentiments in a multi-dimensional vector space. In *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, pages 385–393.

Carlson, Andrew, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell, 2010. Toward an architecture for never-ending language learning. In *AAAI*, volume 5.

Korner, Sven J and Torben Brumm, 2009. Resi-a natural language specification improver. In *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*. IEEE.

Krawczyk, Marek, Rafal Rzepka, and Kenji Araki, 2015. Extracting conceptnet knowledge triplets from japanese wikipedia. In *Proceedings of the 21st Annual Meeting of The Association for Natural Language Processing*.

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

Liu, Hugo and Push Singh, 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Mendes, Pablo N, Max Jakob, Andrés García-Silva, and Christian Bizer, 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*. ACM.

Nakahara, Kazuhiro and Shigeo Yamada, 2011. Development and evaluation of a web-based game for commonsense knowledge acquisition in japan. In *Unisys Technology Review no. 107*. pages 295–305.

Randolph, Justus J, 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online Submission*.

Schubert, Lenhart, 2002. Can we derive general world knowledge from texts? In *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc.

Singh, Push, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu, 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*. Springer, pages 1223–1237.

Speer, Robert H, Catherine Havasi, K Nichole Treadway, and Henry Lieberman, 2010. Finding your way in a multi-dimensional semantic space with luminoso. In *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM.

Suchanek, Fabian M, Gjergji Kasneci, and Gerhard Weikum, 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM.

Sumida, Asuka and Kentaro Torisawa, 2008. Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, volume 8. Citeseer.

Sumida, Asuka, Naoki Yoshinaga, and Kentaro Torisawa, 2008. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in wikipedia. In *LREC*.

Yamada, Ichiro, Chikara Hashimoto, Jong-Hoon Oh, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaaki Tsuchida, and Junichi Kazama, 2010. Generating information-rich taxonomy from wikipedia. In *Universal Communication Symposium (IUCS), 2010 4th International*. IEEE.