2015 IEEE International Conference on Systems, Man, and Cybernetics

Populating ConceptNet knowledge base with Information Acquired from Japanese Wikipedia

Marek Krawczyk Rafal Rzepka Kenji Araki
Hokkaido University
Kita-ku, Kita 14, Nishi 9
Sapporo, Japan
marek@ist.hokudai.ac.jp rzepka@ist.hokudai.ac.jp araki@ist.hokudai.ac.jp

***Abstract*—This paper presents a method of acquiring IsA assertions (hyponymy relations), AtLocation assertions (informing of location of objects) and LocatedNear assertions (informing of neighboring locations) automatically from Japanese Wikipedia XML dump files. To extract IsA assertions, we use the Hyponymy extraction tool v1.0, which analyses definition, category and hierarchy structures of Wikipedia articles. The tool also produces information-rich taxonomy from which, using our original method, we can extract additional information, in this case AtLocation and LocatedNear type of assertions. Experiments showed that both methods produce positive results: we were able to acquire 5,866,680 IsA assertions with 99.0% reliability, 131,760 AtLocation assertion pairs with 93.0% reliability and 6,217 LocatedNear assertion pairs with 99.0% reliability. Our method exceeded the baseline system considering both precision and the number of acquired assertions.**

***Keywords*-Knowledge Acquisition; Pattern Recognition; Information Retrieval;**

I. Introduction

Access to large-scale general knowledge bases, is an important factor in developing effective programs performing textual-reasoning tasks. One of the examples of such bases is ConceptNet, a knowledge representation project that provides a large semantic graph describing general human knowledge [1]. ConceptNet was initiated to represent knowledge collected by Open Mind Common Sense project [2], which utilizes an interactive website to collect new knowledge. Further releases incorporated knowledge from similar websites as well as online word games that automatically collect general knowledge in, besides English, Japanese, Chinese, Portuguese and Dutch. 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, which are both hosted by non-profit Wikimedia Foundation. A growing number of projects utilizing this open-source knowledge base represent applications such as topic-gisting [3], affect-sensing [4], dialog systems [5], daily activities recognition [6] and so on. However the effectiveness of such programs depends on the size of the knowledge base: the more extensive it is, the higher recall. Populating knowledge bases manually would be a long and labor-intensive process. For example, nadya.jp[3], an online project aiming at gathering knowledge by means of a game with a purpose [7], since its launch in 2010 was able to introduce little over 43,500 entries to the ConceptNet. It is therefore clear that there exists a strong need to develop methods of introducing new pieces of information automatically.

Projects such as NELL [8] or KNEXT [9] focus on extracting meaningful assertions from unstructured text data found on the Internet. Alternative to that approach would be to transfer information from the existing semi-structured sources into a knowledge base. The biggest advantage of semi-structured sources analysis is that a considerable amount of human validation has already been involved in the creation of such sources, which transfers to higher reliability of included information. Wikipedia is probably one of the first sources that come to mind while thinking about large-scale information pools. There are working projects, such as DBpedia, focusing on transferring knowledge gathered in Wikipedia into more formalized, digitally processable form [10]. English part of DBpedia has already been introduced to ConceptNet, greatly expanding the number concepts described in that language. The Japanese part however has not been transferred yet, leaving this part of ConceptNet 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 great proportion, language and culture-specific, it is vital to widen the scope of Japanese part independently.

[1]http://www.wiktionary.org/
[2]http://www.wikipedia.org/
[3]http://nadya.jp/

978-1-4799-8697-2/15 $31.00 © 2015 IEEE
DOI 10.1109/SMC.2015.519

II. 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 [11]. The fourth module generates intermediate concepts of hyponymy relations using the output of the first three modules [12]. The program utilizes Pecco library[5] (SVM-like machine learning tool) to estimate the extracted hyponymy relation pairs' plausibility level and boost the precision and recall of the system [13]. The hyponymy pairs extracted using the definition, category and hierarchy modules may be transferred to ConceptNet as two concepts related to each other by 'IsA' relationship (Table I lists examples of the extracted pairs). Yamada *et al.* [12] argues, that these pairs are not informative enough to be useful for such NLP tasks as Question Answering, however they do fall into the scope of ConceptNet, a domain representing commonsense and general knowledge. They are simple and general enough not to interfere with the ConceptNet's usage flexibility, yet informative enough to introduce new and valuable knowledge to the knowledge base.

Table I
Examples of extracted 'IsA' relationship pairs.

Hypernym	Hyponym
kouen [6] (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)

III. Extracting other relations

As mentioned before, the fourth, so called 'extended' module of the Hyponymy extraction tool v1.0 enriches the taxonomy acquired by the previous modules. The procedure is as follows: first it acquires hyponymy relations, further referred to as basic hyponymy relations, from Wikipedia with the method proposed by Sumida *et al.* [13]. Next, it augments each acquired hypernym with the title of the Wikipedia article holding the basic hyponymy relation and consolidates the basic hypernym with the newly generated augmented hypernym (so called 'T-INTER'), creating a 'T- hyponymy relation'. In the final step, it generates additional intermediate concept ('G-INTER') by generalizing the enriched hypernym. As a result, it acquires four-level, information-rich hyponymy relations ('G-hyponymy relations'). Figure 1 shows the process, depicted by the example used by the authors [12]. We could imagine 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 on how deep these generalizations should continue, and therefore a choice to make one generalization seems reasonable from the point of view of output data size. If such further generalizations would be required, they could be achieved by traversing the graph structure of ConceptNet.

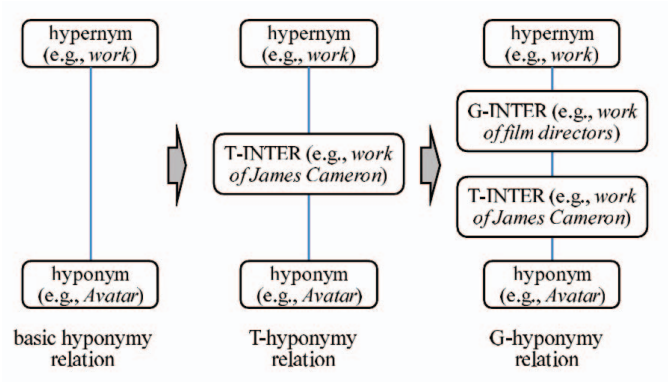

Figure 1. Procedure of Yamada *et al.* [12] method.

As we can see from the examples in Table II, the generated augmented hypernyms are too specific to be incorporated into ConceptNet without compromising the knowledge base's versatility of use . However they may be used for acquiring additional information about their corresponding hyponyms, such as location, neighboring locations, creator and so on. Information about location and creator may be directly transferred into ConceptNet through already built-in 'AtLocation', 'LocatedNear' and 'CreatedBy' relations. The rest of the acquired information related to the hyponyms may be represented by a more general 'RelatedTo' relation.

The procedure of acquiring additional information is shown 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. Next (Step 2), we filter the basic hypernym through a secondary rules base to exclude items that would introduce noise to the output. For example we can acquire information about the birthplace of famous people, however this does not mean that we can build an 'AtLocation' kind of relationship between the name of the person and his or her birthplace. Therefore hypernyms indicating people are excluded from the analysis of location. If the basic hypernym is positively assessed by the secondary

[4]http://alaginrc.nict.go.jp/hyponymy/
[5]http://www.tkl.iis.u-tokyo.ac.jp/∼ynaga/pecco/
[6]All Japanese language phrases are transliterated and written in italics.

Table II
EXAMPLES OF AUGMENTED HYPONYMY RELATIONS GENERATED BY YAMADA *et al.* [12] METHOD.

Original Hypernym	G-INTER	T-INTER	Hyponym
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
kantoku (director)	*musical eiga no kantoku* (director of a musical)	*Ame ni Utaeba no kantoku* (director of Singin? in the Rain)	Stanley Donen
ibento (enent)	*Hoso-kyoku no ibento* (event of a broadcasting station)	*Fuji Television no ibento* (event of Fuji Television Co., Ltd.)	*Odaiba dotto komu* (Odaiba dot com)

rules base, then (Step 3) we assume that the phrase generated by deleting the basic hypernym from the G-INTER is a valid location or creator tag. Using the example from Figure 2, we validate that 'county in England' is a proper tag to describe a location. In next stage (Step 4) we compare the validated location or creator tag with the information included in the T-INTER. This way, using the previous example, we can extract the knowledge that in this case the county we refer to is East Sussex. Finally (Step 5), we connect the newly acquired information to the base hyponym with an appropriate relationship tag to extract a new relation, for example Uckfield-AtLocation-East Sussex. In case of acquiring 'LocatedNear' pairs we confirm that the basic hypernym contains a marker indicating physical proximity (such as Chinese character meaning 'neighboring'). In (Step 2) we filter out items that introduce noise due to ambiguity and then perform (Step 3)–(Step 5) as described above.

The effectiveness of the method greatly depends on the number of introduced rules to both primary and secondary rules base. As our method is still work in progress, this time we used 21 primary rules and 14 secondary rules, which allowed us to extract assertions concerning location and neighboring locations. The rules have been created manually using heuristics after the analysis of the input data. The reason why we chose this kind of approach is because Wikipedia entries analyzed by the system contain information about named entities referring to locations written in a formal format. It means that the information units contain Chinese characters indicating a type of location, a city, province, school etc. We use manually crafted rules to detect these characters, which in turn make us able to get the named entities referring to locations. 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 are interested in. For example secondary rules used for detecting people include suffix '∼sha', which describes different professions. For languages such as English such shortcut would be harder to apply, and therefore person detection would require a much larger rules base covering an extensive list of names of professions and appropriate suffixes (like '∼er', '∼or' or '∼ist'). In future we would like to explore 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.

IV. EVALUATION

In order to verify the reliability level declared by Sumida *et al.* [13] and evaluate our proposed method of obtaining additional relations, we used the 2014-11-04 version of the Japanese Wikipedia dump data. We obtained 6,014,194 hypernym-hyponym pairs by running the definition, category and hierarchy modules of the Hyponymy extraction tool v1.0 at 93.0% precision rate and using the biggest available training set. The number of unique hyponymy pairs was 5,866,680, which means that 147,514 pairs have been extracted by more than one module. This may be treated as an additional reliability level of those pairs. The 93.0% reliability level declared by the authors of the method has been verified by three human annotators, whose task was to evaluate 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 pairs. We assigned 0.5 points to related concepts as they may be used to create correct assertions (see Future Work section). The ratings provided by more than one annotator were regarded as the evaluation output. In case of every annotator giving a different score to a particular pair (two cases), one of the authors decided the score. The procedure follows a modified Sumida *et al.* [13] evaluation method. Table III shows the evaluation results. 97 pairs were

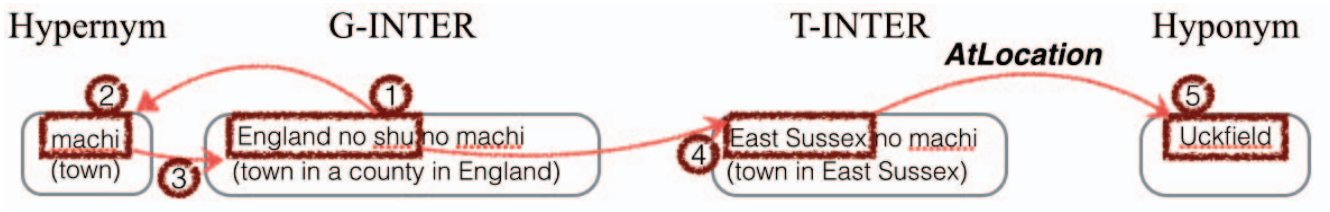

Figure 2. Procedure of our proposed method.

evaluated as representing correct hyponymy relation, 2 pairs as related concepts, but not in a hyponymy relation and 1 as unrelated concepts.This results in 99.0% precision value, which surpasses 93.0% declared by Sumida *et al.* The level of overall agreement between annotators was 91.0% and the Kappa value[7] was 0.86, which indicates that the annotation judgement was in almost perfect agreement [14].

Table III
EVALUATION RESULTS FOR 'ISA' RELATIONS.

Correct hyponymy	Related concepts	Unrelated concepts	Precision	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 for obtaining additional information we were able to generate 131,760 pairs representing AtLocation relation and 6,217 pairs representing LocatedNear relation. For comparison, nadya.jp, the baseline system using online game, provided only 8,706 AtLocation relations and no LocatedNear relations in four years. In case of AtLocation pairs, we evaluated 50 pairs randomly selected from our method's output and 50 pairs randomly selected from nadya.jp's AtLocation assertions [7]. In case of LocatedNear relations, a comparison with baseline was not possible, as the current version of ConceptNet does not contain any LocatedNear pairs in its Japanese language section yet. 50 randomly selected LocatedNear pairs were therefore evaluated independently. The evaluation procedure follows the previously applied one, 1 point being appointed to correct AtLocation or LocatedNear assertions, 0.5 point to related concepts, but not by AtLocation or LocatedNear relation, and 0 points to unrelated concepts. In seven cases the annotators' evaluation was inconsistent, and therefore one of the authors decided the score. Table IV shows the evaluation results of our AtLocation pairs acquisition 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 93.0% 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 76.0%. 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.

[7]We used Randolph's free marginal multirater kappa instead of Fleiss' fixed-marginal multirater kappa due to high agreement low kappa paradox.

Table IV
EVALUATION RESULTS FOR 'ATLOCATION' RELATIONS IN COMPARISON WITH NADYA.JP BASELINE.

	Correct AtLocation	Related concepts	Unrelated concepts	Precision	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 V 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 99.0% precision. The level of overall agreement between annotators was 84.0% and the Kappa value was 0.76, which indicates that the annotation judgement was in substantial agreement.

Table V
EVALUATION RESULTS FOR 'LOCATEDNEAR' RELATIONS

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

The results of our experiments show that both 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. Such operation would be beneficial for the knowledge base, considering the number of the newly introduced assertions as well as reliability of the data in comparison with the resources already present in the knowledge base.

V. Conclusion

This paper presented a method for automatic acquisition of ConceptNet knowledge triplets from Japanese Wikipedia. It allowed us to mine IsA, AtLocation and LocatedNear assertions with precision at the level of 99.0%, 93.0% and 99.0% respectively. Considering the fact that the Japanese part of current ConceptNet 5.3 consists of 1,071,046 assertions, a contribution of 6,004,657 new assertions would be significant. It would mean an increase at the level of 560.6%. As Wikipedia is a constantly expanding source, we would be able to acquire more assertions simply by applying our method to the updated Wikipedia XML dump files.

VI. Future work

In order to extend the functionality of our proposed method we intend to introduce more primary and secondary rules, which would allow the system to increase its precision, as well as the scope of extracted information. The result of expanded module of the Hyponymy extraction tool v1.0 contains, apart from location data, information about tens of thousands of books, films, plays, paintings etc. and their authors. We intend to use our method on this information as well, which would allow generating many reliable CreatedBy type assertions. Additional analysis of the acquired data may reveal further possibilities. As mentioned before, 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 easy evaluation of the method's output by Japanese native speakers. This would allow us to utilize the pairs representing related concepts. Instead of simply filtering them out, we could build new assertion on their basis by modifying the relation type or correcting one of the concepts.

References

[1] R. Speer and C. Havasi, "Representing general relational knowledge in conceptnet 5." in *LREC*, 2012, pp. 3679–3686.

[2] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and 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.

[3] R. H. Speer, C. Havasi, K. N. Treadway, and 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.

[4] E. Cambria, A. Hussain, C. Havasi, and 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.

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

[6] J. Ullberg, S. Coradeschi, and 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.

[7] K. Nakahara and 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.

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

[9] 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.

[10] P. N. Mendes, M. Jakob, A. García-Silva, and 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.

[11] A. Sumida and K. Torisawa, "Hacking wikipedia for hyponymy relation acquisition." in *IJCNLP*, vol. 8. Citeseer, 2008, pp. 883–888.

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

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

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