Extracting ConceptNet Knowledge Triplets from Japanese Wikipedia

Marek Krawczyk
Hokkaido University
Kita-ku, Kita 14, Nishi 9
Sapporo, Japan
marek@ist.hokudai.ac.jp

Rafal Rzepka
Hokkaido University
Kita-ku, Kita 14, Nishi 9
Sapporo, Japan
rzepka@ist.hokudai.ac.jp

Kenji Araki
Hokkaido University
Kita-ku, Kita 14, Nishi 9
Sapporo, Japan
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.

1 Introduction

Access to large-scale general knowledge bases, such as the ConceptNet, is an important factor in developing effective programs performing textual-reasoning tasks (Speer and Havasi 2012). A growing number of projects utilizing this open-source knowledge base represent applications such as topic-gisting, affect-sensing, analogy-making, text summarization 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[1], an online project aiming at gathering knowledge by means of a game with a purpose (Nakahara and Yamada 2011), 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. Our method allows automatic, large-scale extraction of new ConceptNet triplets from a dynamically expanding source – Japanese Wikipedia.

2 Hyponymy relation as 'IsA' relation

In our approach we use the Hyponymy extraction tool v1.0[2], 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 fourth module generates intermediate concepts of hyponymy relations using the output of the first three modules (Yamada et al. 2010). The program utilizes Pecco library[3] (SVM-like machine learning tool) to estimate the extracted hyponymy relation pairs' plausibility level and boost the precision and recall of the system (Sumida et al. 2008).

The hyponymy pairs extracted using the definition, category and hierarchy modules may be transferred to ConceptNet as two concepts related to

[1] http://nadya.jp/
[2] http://alaginrc.nict.go.jp/hyponymy/
[3] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/pecco/

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

Table 1. Examples of extracted 'IsA' relationship pairs.

Hypernym	Hyponym
kouen[4] (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)

each other by 'IsA' relationship (Table 1 lists examples of the extracted pairs). Yamada et al. (2010) 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.

3 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 has been described in detail by Yamada et al. (2010).

As we can see from the examples in Table 2, the generated augmented hypernyms are too specific to be incorporated into ConceptNet without compromising the knowledge base's use versatility. 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 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 hy-

Table 2. Examples of augmented hyponymy relations generated by Yamada et al. (2010) method.

Original Hypernym	G-INTER	T-INTER	Hyponym
tojo-jinbutsu	*SF eiga no tojo-jinbutsu*	*WALL-E no tojo-jinbutsu*	
(character)	(character of SF movie)	(character of WALL-E)	M.O
seihin	*kigyo no seihin*	*Silicon Graphics no seihin*	
(product)	(product of a company)	(product of Silicon Graphics, Inc.)	IRIS Crimson
sakuhin	*America no shosestu-ka no sakuhin*	*J.D. Salinger no sakuhin*	
(work)	(work of American novelist)	(work of J.D. Salinger)	A boy in France
machi	*England no shu no machi*	*East Sussex no machi*	
(town)	(town in a county in England)	(town in East Sussex)	Uckfield
kantoku	*musical eiga no kantoku*	*Ame ni Utaeba no kantoku*	
(director)	(director of a musical)	(director of Singin' in the Rain)	Stanley Donen
ibento	*Hoso-kyoku no ibento*	*Fuji Television no ibento*	*Odaiba dotto komu*
(enent)	(event of a broadcasting station)	(event of Fuji Television Co., Ltd.)	(Odaiba dot com)

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

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

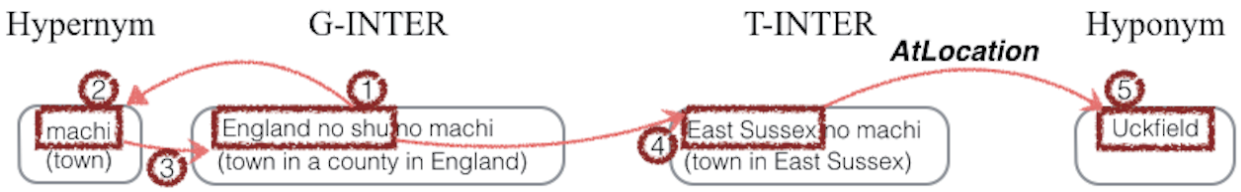

Figure 1. Procedure of our proposed method.

pernym contains a marker indicating physical proximity (such as character '隣', 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 number and reliability level of the acquired data is presented in the evaluation section.

4 Evaluation

In order to verify the reliability level declared by Sumida et al. (2008) 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. (2008) evaluation method.

Table 3 shows the evaluation results. 97 pairs were 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.

Table 3. 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 (Nakahara and Yamada 2010). 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 Locat-

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

edNear 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 4 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%.

Table 4. Evaluation results for 'AtLocation' relations.

	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

Table 5 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.

Table 5. 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.

5 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.

References

Nakahara, K. and Yamada, S. (2011). "Development and Evaluation of a Web−Based Game for Common−Sense Knowledge Acquisition in Japan." *Unisys Technology Review* no.107, pp. 295-305.

Speer, R. and Havasi, C. (2012). "Representing general relational knowledge in ConceptNet 5." In *Proceedings of LREC Conference*, pp. 3679-3686.

Sumida, A. and Torisawa, K. (2008). "Hacking Wikipedia for hyponymy relation acquisition." In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 883-888.

Sumida, A., Yoshinaga, N., and Torisawa, K. (2008). "Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia." In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 2462-2469.

Yamada, I., Hashimoto, C., Oh, J. H., Torisawa, K., Kuroda, K., De Saeger, S., Tsuchida, M., and Kazama, J. I. (2010). "Generating Information-Rich Taxonomy from Wikipedia." In *Proceedings of the 4th International Universal Communication Symposium*, pp. 96-103.

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.