Crowdsourcing Localization of Ontology and Geographical Names

Amarsanaa Ganbold, Altangerel Chagnaa
Department of Information and Computer Science, School of Engineering and Applied Sciences, National University of Mongolia,
{amarsanaag, altangerel}@num.edu.mn

Abstract

Building and enriching knowledge base techniques via crowdsourcing have been broadly investigated. However, crowdsourcing localization of ontology of diversity-aware knowledge base and of geospatial entities are studied less. In this paper, we show preliminary experiments on ontology localization task and transliteration of geographical names through crowd force. In the result of asking bilingual web users human intelligence task to translate ontology terms in English into Mongolian, we obtained 77% accuracy of localized ontology. Also 60% of location names have correctly localized by one crowdsourcing activity.

Keywords: crowdsourcing, ontology localization, transliteration of geographical names

1. Introduction

Building Knowledge Base (KB) used in semantic web requires multilingual knowledge and local cultures which includes diversity [1]. However, there are common concepts across languages, specific concepts can only be used in a particular language that is called diversity. KBs capturing this broad knowledge are able to process over all the knowledge of world. Some multilingual knowledge bases such as YAGO [2], DBpedia [3] have built by extracting of semi-structured data, they are not precisely capturing diversity in terms of single concept or notion.

In order to enrich the UKC [4, 5] knowledge base, ontology localization capturing diversity [6] done with manual approach have investigated. But this method is very time consuming and expensive because it always requires linguistic and domain experts in case of building very large knowledge base. In contrast, automatic extraction is unfeasible for low resourced languages such as Mongolian. Thus, this paper aims to examine possibility of ontology localization and geographical names transliteration via crowdsourcing. Crowdsourcing is a process of obtaining human intellectual work by collecting a number of contributions from each individual of crowd. In this work, we created human intelligence tasks (HIT) for translating space ontology terms of the UKC from English to Mongolian and asked web users as well as transliteration of geospatial names from GeoNames [15]. In the result, we obtained 78 of 99 synsets and 4233 of 7002 Mongolian geographical names.

The rest of the paper is organized as follows. Section 2 introduces several studies related the crowdsourcing translation process and its evaluation techniques. Section 3 introduces the HIT task design and evaluation methodology for ontology localization. The transliteration task is described in Section 4. Section 5 shows experimental result of the HIT tasks. Finally, we conclude the paper. This paper is extended version of “An experiment of ontology localization via crowdsourcing” paper which published in the proceedings of MMT-2015 domestic conference, in Mongolian language.

2. Related work

A workflow design of collaborative translation [7], translation tasks in library cataloguing [8] and a quality control model for non-professional translation [9], [10] referred to crowdsourcing translation challenges and evaluation of translation output.

The collaborative workflow is to translate words in a sentence one by one at first, and in the second phase complete the sentence by a bilingual contributor in an assistive way. Finally, a monolingual collaborator ranks the best suitable translation of the sentence [7]. The main idea of the effort was to provide translations through a precise collaboration on every constituent of a sentence and it is not independent translations by an individual contributor. Moreover, the method has a bit higher quality than traditional translation according to evaluation method BLEU (Bilingual Evaluation Understudy) which is one of the most popular metrics estimating the quality of output of machine translation.

The evaluation control of crowdsourcing translation can be carried out through reading-comprehension and it is a more effective approach could converge our needs mentioned in the previous section. Because a translation of concept is more context-sensitive and should be a natural interpretation for what comprehension human being see. The interesting method evaluates how understandable the output translation and it is better than other evaluation methods based relative ranking [9].

Furthermore, manual evaluation of translation has guaranteed that is feasible to obtain high-quality translations as an expert did, and is a low-cost output via crowdsourcing platform [10]. These manual evaluation studies were conducted on Amazon's Mechanical Turk (sometimes called MTurk) which is the most popular online marketplace for human labor.

The literature agrees that iterative tasks in crowdsourcing translation including manual evaluations are inexpensive and achieve robust results with the language translation task.

3. Ontology localization

Ontology localization is the adaptation of a piece of knowledge to a particular language and culture [11]. Although high quality localization can be done with collaborations of psycholinguists and domain experts, we study crowdsourcing capability of localizing ontology by non-experts in inexpensive and fast way. The UKC that we use consists of independent language vocabularies represented as hierarchical synset bases such as the WordNet [16] and the Italian part of MultiWordNet [4] [17] where one vocabulary represents the whole knowledge base. Each concept in the KB is linked to a synset which is a set of synonymous words in a language. For instance, the concept city can be represented as the synset of *city, metropolis, urban center* in English, *città, metropoli* in Italian. The concept represented by the synsets has same meaning in the languages which means that is linked to several synsets in different languages. A synset has a descriptions which clearly distinguishes it from other similar ones. This requires the most appropriate translations of synset and its description. Because a word may have multiple meanings and human can disambiguate the polysemous words much better than automatic translations. Thus, human intelligence is a potential for this task.

3.1. Synset translation

In order to define a synset in target language in this case Mongolian that matches to another synset in source language in this case English, its words can be directly translated into the target language. But, some words of the synset might be a lack of translation [6]. A number of web users doing HIT task for the translation can produce appropriate synsets. Synset words can be words or idioms or collocations. For example, the following synset denotes the concept *dwelling* shown in Figure 1.

dwelling, home, domicile, abode, dwelling house, habitation (housing that someone is living in)

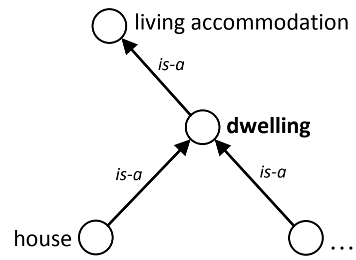

Figure 1. A small subset of the space ontology

This synset can be translated into Mongolian as follows:

гэр орон, орон гэр, гэр, орон байр, оршин суугаа газар (хүн амьдарч байгаа байр)

(in English), ger oron, oron ger, ger, oron bair, orshin suugaa gazar (a residence where a human being lives)

During the localization activity there might be untranslatable synsets which is called lexical gaps [12]. A lexical gap means that the concept is not used or not lexicalized in the target language and culture. For example, *submarine furrow (a closed, linear, narrow, shallow depression)* might be a lexical gap and there is no expression in Mongolian language.

We ask normal web users to define words of synset in the target language which is equivalent to a given synset in the source language or to mark the given synset as a lexical gap. The words can be vary in both synsets that means there is an added word in the target synset without direct translation of the source. This task will be asked by multiple users and they can produce many words for the target synset. Of course, user can skip a task which might be difficult to translate or unknown for him or her. Afterwards, we again ask users to validate all distinct words in the synset that is a separate HIT task.

3.2. Translation evaluation

There are two types of rating in crowdsourcing: quantitative and qualitative. Target synset words can be classified into three categories: *Correct*, *Wrong* and *Unknown* and the qualitative method can be used for this translation evaluation. When a number of users evaluate the words, we can use the statistical metric Fleiss' kappa [13] in order to measure inter-rater agreement on the validations. Thereby, we are able to know how synset translation is correctly done by the correlation coefficient (1) of each target synset.

$$k = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \qquad (1)$$

Where $\overline{P} - \overline{P}_e$ is the degree of achievable agreement, and, $1 - \overline{P}_e$ is the degree of agreement actually achieved. $\kappa=1$ means the raters have complete agreement of Correct or Wrong or Unknown, $\kappa \leq 0$ is that there is no agreement among raters. If the coefficient value is between 0.01 and 1.0, there is somewhat agreements among raters. But the optimal value of coefficient for moderate agreement can be depend on factors such as words, categories and raters in our case.

For each synset with any agreement, we extract words have majority votes of Correct that can be the most suitable for the target synset. In this way, we combine the users' contributions for translation task.

4. Transliteration of geographical names

In this task, we design a crowdsourcing task to localize geographical names from one language to another one. Geographical names have always various transliterations in different languages which makes incorrect forms of the names. It also happens even in a language. For instance, the Mongolian word *сум* (*sum*) has many transliteration forms in English such as *somon*, *suma* and *sumu*. This causes from phonetic ability of certain language speakers. GeoNames includes many alternative names for each name and alternatives are from various languages (e.g., Chinese, Russian). Actually people write names in free forms when they transliterate into a script. Therefore this task can be carried out via crowdsourcing with human level accuracy.

We aim to obtain correct geographical names in a script (e.g., Mongolian Cyrillic as target) from another one (e.g., Latin script as source) by crowdsourcing transliteration. This approach helps to build gazetteer data in low resource languages. We use GeoNames and each name has coordinates, classes, relations and metadata. We show name location and its alternative names on a map to the users by providing some metadata such as its administrative division and its class (e.g., populated place, lake, and mountain). Then we ask users to provide only one correct name in the target script. If a number of users provide same transliterations on a geographical name, there might be perfect agreement among the users and the result will most likely be correct. Users also can skip a task when he or she is unfamiliar with the given name. If the task is skipped by most of the users who run on the task, it can be difficult or wrong names. In addition, we exclude users who cheated the crowdsourcing activity or is not able to do this task in terms of his or her background knowledge about locations.

Users also may have fair agreement on certain names. For example, if users provide many different transliterations on a name, the users cannot reach any agreement. These names are main challenges for this kind of crowdsourcing task. That is not solved by one activity but may requires another task.

5. Experimental result

The one of crowdsourcing challenge is how to recruit users [14]. In this work, 44 English-Mongolian bilingual web users localize 99 English synsets into Mongolian language. We ask 10 users for each synset and 990 synset contributions have collected in total. There were 639 (364 distinct) words, 20 lexical gaps, and 411 skip answers from the users.

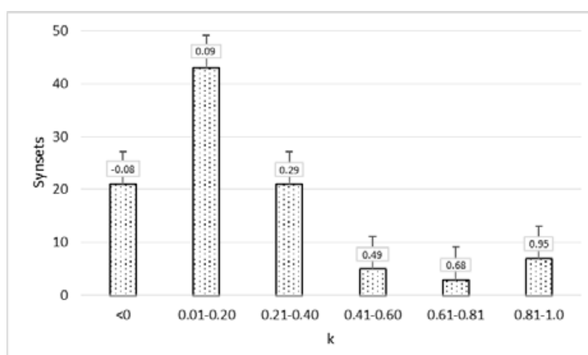

Figure 2. Agreement on localized synsets

Figure 2 shows a number of synsets for each kappa coefficient interpretation range. Synsets with some agreement are 80%. Synsets with the minimum agreement are 43 and their average kappa value is 0.09.

While words of synset increase, the kappa decreases that shown in Figure 3.

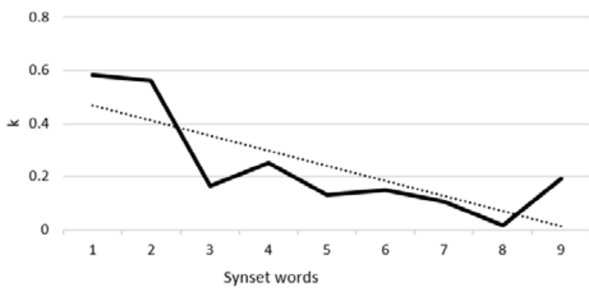

Figure 3. Kappa averages by synset words

After selecting k > 0 synsets we have 381 words. 145 of them have majority vote of Correct annotation. These words have evaluated by an expert who have good command of English language and familiar with common knowledge and in the result we obtain 112 words as true positives.

During this experiment, users produce error-prone translations such as free combination of text instead of lexical unit (word), incorrect spelling and so on. And, this shows that we need to design more suitable HIT task and its constraints on the user interface. In addition, there is no lexical gap found after crowdsourcing evaluation. This means that the web users cannot define or distinguish lexical gaps and more expert users are needed in this case.

The second crowdsourcing experiment have conducted on transliteration of location names. In this two-week process, 7002 task were presented in CrowdCrafting [18] platform and 171 volunteers have contributed.

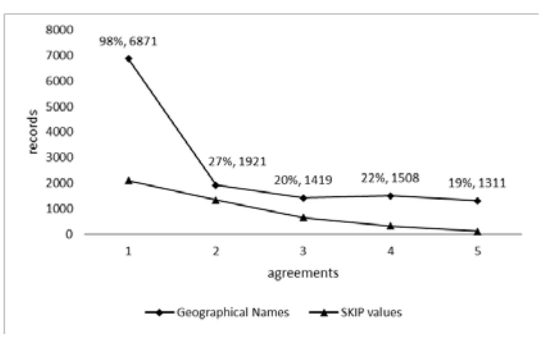

Figure 4. Frequency of agreement

We ask 5 users for each name then we obtained 2818 names with perfect agreement which means all the 5 or 4 users provide same answers. While these names were 100% accuracy, names with 3-user (triple) agreement were provided 91%. Thus, we successfully obtained 4233 geographical names in the one-phase crowdsourcing. Figure 4, shows frequency of agreement along Skip answers conversely with agreement numbers.

Names with 1- (single) or 2-user (double) agreement shown in Figure 5 are hard to transliterate because users have no moderate agreement on those names.

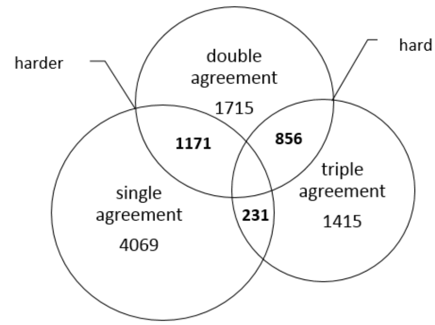

Figure 5. Disagreed names

These names requires another crowdsourcing task such as evaluation of candidate names.

6. Conclusion

In this work, we conducted an experiment on localization of ontology and geographical names. We obtained 78 of 99 English synsets with 77% accuracy in the result of evaluation by English-Mongolian bilingual expert. These localization were carried out by normal web users and if we recruit more potential users in this task we could achieve better result in inexpensive and fast way. For gazetteer data, we showed that the crowdsourcing approach is more feasible and have high quality of transliteration. Furthermore, calibration questions to check user cheating and quality control on runtime would be next issue in our both crowdsourcing activities.

7. References

[1] Ganbold, A., Farazi, F., Reyad, M., Nyamdavaa, O., and Giunchiglia, F., “Managing Language Diversity Across Cultures : the English-Mongolian Case Study,” *Int. J. Adv. Life Sci.*, Vol. 6(3), 2014, pp. 167–176.

[2] Suchanek, F. M., Kasneci G., and Weikum, G., “YAGO: A Large Ontology from Wikipedia and WordNet,” *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 6(3), 2008, pp. 203–217.

[3] Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z., “DBpedia : A Nucleus for a Web of Open Data,” *Nucleus*, Vol. 4825, 2007, pp. 722–735.

[4] Giunchiglia, F., Maltese, V., and Biswanath, D., "Domains and context: first steps towards managing diversity in knowledge," *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 12, 2012, pp. 53–63.

[5] Giunchiglia, F., Dutta, B., Maltese, V., and Farazi, F., "A facet-based methodology for the construction of a large-scale geospatial ontology," *Journal on data semantics*, Vol. 1(1), 2012, pp. 57–73.

[6] Ganbold A., Farazi, F., and F. Giunchiglia, "An Experiment in Managing Language Diversity Across Cultures," in *eKNOW 2014 : The Sixth International Conference on Information, Process, and Knowledge Management*, 2014, pp. 51–57.

[7] Ambati, V., Vogel, S., and Carbonell, J., "Collaborative workflow for crowdsourcing translation," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012, pp. 1191–1194.

[8] Corney, J., Lynn, A., Torres, C., Di Maio, P., Regli, W., Forbes, G., and Tobin, L., "Towards crowdsourcing translation tasks in library cataloguing, a pilot study," in *Digital Ecosystems and Technologies DEST 2010 4th IEEE International Conference on*, 2010, pp. 572–577.

[9] Callison-Burch, C., "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk," *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol. 1, 2009, pp. 286-295.

[10] Zaidan, O. F., and Callison-Burch, C., "Crowdsourcing translation: professional quality from non-professionals," *in HLT'11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol 1, 2011, pp. 1220–1229.

[11] Cimiano, P., Montiel-ponsoda, E., Buitelaar, P., and Espinoza, M., "A Note on Ontology Localization," *Journal of Applied Ontology (JAO),* Vol. 5, 2010, pp. 127–137.

[12] Bentivogli, L. and Pianta, E., "Looking for lexical gaps" *In Proceedings of the Ninth EURALEX International Congress*, 2000, pp. 663–669.

[13] Fleiss, J. L., "Measuring nominal scale agreement among many raters," *Psychological bulletin*, Vol.76(5), 1971, pp. 378–382.

[14] Doan, A., Ramakrishnan, R., and Halevy, A. Y., "Crowdsourcing systems on the World-Wide Web," *Communications of the ACM*, Vol. 54(4), 2011, pp. 86-96.

[15] http://www.geonames.org

[16] http://wordnet.princeton.edu/

[17] http://multiwordnet.fbk.eu/

[18] http://crowdcrafting.org/