Using Crowd Agreement for Wordnet Localization

Amarsanaa Ganbold*, Altangerel Chagnaa*, Gábor Bella†
*Department of Information and Computer Science, SEAS, National University of Mongolia
Ikh surguuliin gudamj – 3, Sukhbaatar district, 14201 Ulaanbaatar, Mongolia
†Department of Information and Computer Science, University of Trento
via Sommarive, 5, 38123 Trento, Italy
{amarsanaag, altangerel}@num.edu.mn, gabor.bella@unitn.it

Abstract

Building a wordnet from scratch is a huge task, especially for languages less equipped with pre-existing lexical resources such as thesauri or bilingual dictionaries. We address the issue of costliness of human supervision through crowdsourcing that offers a good trade-off between quality of output and speed of progress. In this paper, we demonstrate a two-phase crowdsourcing workflow that consists of a synset localization step followed by a validation step. Validation is performed using the inter-rater agreement metrics *Fleiss' kappa* and *Krippendorf's alpha*, which allow us to estimate the precision of the result, as well as to set a balance between precision and recall. In our experiment, 947 synsets were localized from English to Mongolian and evaluated through crowdsourcing with the precision of 0.74.

Keywords: crowdsourcing evaluation, inter-rater agreement, synset localization, wordnet

1. Introduction

Lexical-semantic resources like WordNet are a fundamental resource for many NLP and semantic applications. Building such resources traditionally involves the collaboration of a large number of professionals, such as psycholinguists and lexicographers, in order to obtain a high-quality end result. To mitigate the cost of WordNet construction, automatic and semi-automatic approaches have been successfully used, based on existing bilingual resources such as dictionaries and thesauri. Such methods, however, are hardly applicable to less-resourced languages that do not already have rich thesauri and dictionaries.

In this paper we introduce an alternative approach of localizing wordnet synsets through crowdsourcing. The process consists of two phases: in phase one, workers build synsets by contributing synonymous words while in phase two they validate the correctness of words provided earlier. Validation results are combined and assessed using two alternative statistical metrics of *inter-rater agreement*, the idea being that higher levels of agreement corresponds to higher precision. We evaluate this hypothesis with the help of a gold standard corpus. The evaluation provides us with an insight on the efficiency of crowdsourcing and allows us to optimize the coefficient parameters of these metrics for increased precision.

2. Related Work

Ontology localization, as described by Espinoza et al. (2009) and Cimiano et al. (2010), presents an alternative approach to addressing the cost of building lexical-semantic resources. An approach based on ontology label translation (Arcan and Buitelaar, 2013) was developed to provide a knowledge-based extension to a statistical machine translation system. However, automatically translated multilingual terms often suffer from quality issues. A preliminary crowdsourcing model for less-resourced languages (Benjamin and Radetzky, 2014) allows the building of lexicons with the help of Internet users. They concluded that extensive manipulation and review by language experts was necessary in order to obtain high-quality linguistic data and to capture a large diversity of knowledge. Crowdsourcing techniques were used to build wordnets from scratch through word sense acquisition (Biemann and Nygaard, 2010), to bootstrap a wordnet through translation between two languages (Wijesiri et al., 2014), and to annotate the correctness of synset words of an automatically developed wordnet (Fišer et al., 2014). In (Lanser et al., 2016), a two-stage workflow for translation-based crowdsourcing of ontology lexicons was designed and evaluated, with positive results. The difference of our approach lies in the fact that we evaluate synset words using inter-rater agreement.

3. Synset Localization

A wordnet synset is a set of synonymous words in a natural language that represents a lexical concept. A lot of such concepts can be mapped to equivalent or very similar concepts across languages, a principle without which bilingual dictionaries would not be possible. For instance, the English synset {`hovel, hut, hutch, shank, shanty`} is equivalent to the Mongolian synset {овоохой, оромж} and both synsets represents the concept `small crude shelter used as a dwelling`. This equivalence allows us to reuse the semantic structure of an existing wordnet (in this case, English) and to produce a partial wordnet (in this case, Mongolian) by following some basic principles of *ontology localization* (Ganbold et al., 2014b; Espinoza et al., 2009). While localization also involves well-known problems related to language diversity, such as conceptualizational mismatches or lexical gaps (Giunchiglia et al., 2017; Bentivogli and Pianta, 2000), we consider these problems as a separate research issue and prefer to focus in this paper on the more straightforward problem of synset localization. Synset localization provides a set of words in a target language that is equivalent to a given synset in a source language. Rather than a literal translation of words between languages, the localization of synsets involves finding the most appropriate words in the target language to express the

same meaning. Providing the words of a synset is a non-trivial task that involves the understanding of precise semantic distinctions with respect to hypernym and hyponym synsets, often based on psycholinguistic and lexicographic expertise. In this paper we examine to what extent crowdsourcing may offer a viable alternative to expert involvement for the specific task of synset localization.

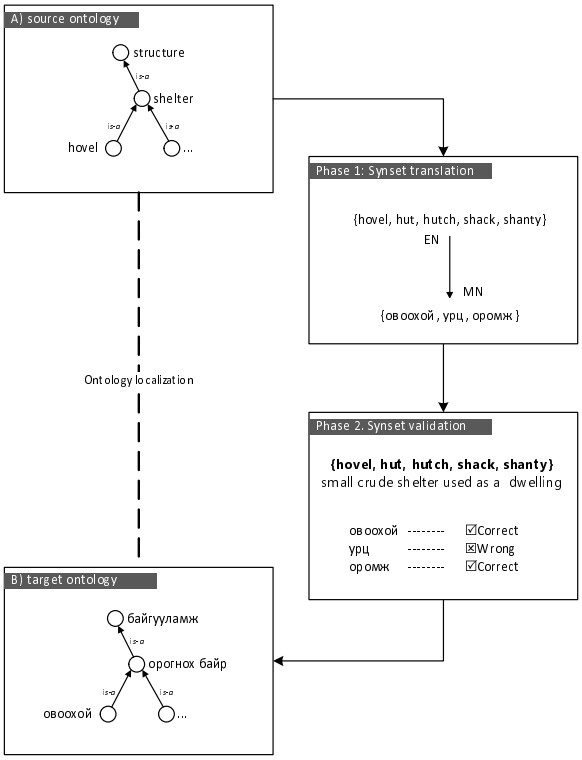

Figure 1: Crowdsourcing workflow of the synset localization. The concepts represented in the *ontology* boxes are labeled by a single word of their representative synsets.

We specify a two-phase crowdsourcing workflow (Figure 1) which consists of a translation and a validation phase. In the translation phase we ask several (in our case, five) workers to provide words for a synset in the target language (in our case, Mongolian, MN) that is equivalent to a given synset in the source language (English, EN). The number of words can be different in the two synsets. Thus, the same synset will be assigned to several workers who may end up suggesting many candidate words for the target synset.
The translation task allocated to workers provides them with synset words, the gloss of the synset, example sentences, the part of speech, as well as images from ImageNet[1] which might help workers in understanding the underlying meaning. If the concept cannot be localized because of a nonexistent lexicalization in the target language (e.g., the concept of *sea port* is not lexicalized in Mongolian) , a worker may mark it as a lexical gap. Workers can also skip tasks that they find difficult to translate or that contain words that are unknown to them. The HIT (Human Intelligence Task) had the following instructions:
Question: *Provide the most appropriate Mongolian word(s) for the following concept represented in English. Please follow the instructions below:*

- *A word must be a lexical unit which is one of a headword in dictionary, a phrase or a restricted collocation.*
- *A word must be written in Mongolian Cyrillic.*
- *A word can be case insensitive.*
- *Words should be separated by semicolon (;).*
- *If you think the concept cannot be localized, please mark it as a lexical GAP.*

A word is not acceptable if includes the following:

- *a Latin letter or words, or digits*
- *a special character (dot, colon etc.) except hyphens (-) which is acceptable in some Mongolian words*

Warning:

- *Image(s) may not exactly represent the concept in some cases.*
- *Please skip the task if you are not familiar with the concept and try to provide words for the next concept.*

4. Synset Validation

In the second phase, workers are asked to validate all distinct candidate words in the synset by annotating the words with three categories: *Correct*, *Wrong*, and *Unknown*.
Our validation task design had the following instructions:
Question: *Validate the Mongolian word(s) which can represent following concept. Please follow the instructions:*

- *assign a category of Correct, Wrong and Unknown for each word;*
- *assign the Wrong category if the word has spelling errors;*
- *assign the Wrong or the Unknown category for all words, if you think the concept is GAP and, assign the Correct category to the GAP entry.*

This evaluation is not acceptable if you do following:

- *randomly evaluate the words;*
- *assign a single category to all words and repeated several times.*

Validation results were collected on a word-by-word basis. We did not consider word ranks (the order of relevance of words) for this paper.
On the synset level, validation results were aggregated to compute the *inter-rater agreement* using the statistical metrics *Fleiss' kappa* (eq. 1) and *Krippendorff's alpha* (eq. 2). These metrics can measure agreement between a fixed number of workers who provided words. The higher the values of these metrics (close to 1), the more certainty we have that the workers have correctly classified the synset

[1] http://image-net.org/

words into correct and incorrect ones, and ultimately the more confident we are of the precision of the result.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (1)$$

where $\bar{P} - \bar{P}_e$ is the degree of achievable agreement and $1 - \bar{P}_e$ is the degree of agreement actually achieved. $\bar{P}$ is the mean of all proportions, where a proportion is the agreement for a given candidate word. $\bar{P}_e$ is the sum of squares of weighted votes of each category. $\kappa = 1$ means the raters have total agreement over *Correct* or *Wrong* or *Unknown*, while $\kappa < 0$ means that there is no agreement among raters. If the value is $0 < \kappa \leq 1.0$, there is some level of agreement among raters. In our case, the optimal value for moderate agreement can depend on factors such as words, categories, and raters.

$$\alpha = 1 - \frac{D_o}{1 - D_e} \quad (2)$$

While κ calculates the number of categories of each word in a synset, α computes the agreement from the data that includes category labels per annotator of the words in the synset. The range of α values is same as κ . In eq. 2, D_o is the disagreement observed and D_e is the disagreement expected by chance.

Alpha allows missing data, i.e., no categorization for some words. The *Unknown* category can be considered as missing data as this option is selected by annotators who do not know what category to assign to the word.

For each synset, when a certain agreement is reached, we extract words that have a majority of *Correct* votes, considered as the most suitable for the target synset. This way we combine the workers' contributions for the translation task. To the best of our knowledge, there are no experiments on the use of Fleiss' kappa and Krippendorff's alpha which can be used to combine workers' contributions. These metrics meet our requirements of validation of synset localization.

5. Experiments and Results

We used *CrowdCrafting*[2] for recruiting workers because of a limited presence of Mongolian speakers on platforms such as *Amazon Mechanical Turk* and *CrowdFlower*. CrowdCrafting is free for scientific projects with volunteer contributors. In phase 1, the total of 77 web users were asked to translate 947 manually built synsets from the *space domain*, that is, the subtree under the high-level synsets of *space* in (Ganbold et al., 2014b; Giunchiglia et al., 2009). In phase 2, 75 web users were asked to validate the results of phase 1. In total, contributors have completed 9,490 tasks and have introduced 6,442 words[3].

In order to evaluate contributions from the crowd, we compiled a gold standard from the space domain in Mongolian, covering all synsets that were included in the crowdsourcing experiment. The gold standard corpus was created by five language experts using an *expert sourcing* methodology described in (Ganbold et al., 2014b; Ganbold et al., 2014a). The methodology involved translation by bilingual translators and a two-level validation process where as many iterations were performed on each translation as necessary for reaching a high-quality result through inter-expert agreement.

Source	# of synsets	# of words
Space domain	943	1,436
Expert validation	889	1,813
Total	**943**	**2,627**

Table 1: Size of the gold standard *space domain* corpus and of the crowd-translated corpus post-validated by experts.

While expert sourcing provides a high-quality reference corpus, it does not guarantee exhaustiveness: crowd workers may come up with alternative yet acceptable lexicalizations. For this reason we completed the evaluation based on our *a priori* gold standard by an *a posteriori* expert validation step on the output of workers, carried out by three language experts. Table 1 shows the size of the gold standard corpus and of the crowd translations for which expert validation was provided. The higher number of words in the expert-validated corpus indicates that the crowds managed to provide correct new lexicalizations that were not present in the gold standard, i.e., that had not been though of by experts.

Evaluation of the entire crowdsourced corpus yielded an overall precision of 0.74. Our main research goal, however, was the evaluation of inter-rater agreement as a tool for controlling the precision of the output. For this purpose, we first computed inter-rater agreements on each synset separately. Then we created increasingly smaller subsets of synsets corresponding to a certain minimum level of inter-rater agreement. For example, to understand the effect of imposing an agreement threshold of 0.5, we ran an evaluation only on synsets with an agreement of 0.5 or higher. The result of these evaluations is shown in Figure 2 where the *x* axis indicates agreement thresholds and the *y* axis indicates precision, recall, and F-measure.

These values were computed separately for the three different agreement metrics, namely, *Fleiss' kappa*, *Krippendorff's alpha*, and *Krippendorff's alpha/NA*. *Alpha/NA* is an alternative form of Krippendorf's alpha where the *Unknown* category is considered as missing data and is not included in calculations.

The first observation we can make from the results is that precision monotonically increases with inter-rater agreement for all three metrics. From this we conclude that inter-rater agreement generally is a reliable tool to control the precision of the outcome, at least for synset localization tasks. We did not obtain a significant difference related to the three different metrics employed (alpha tends to report somewhat more conservative agreement results than the two other metrics for the same precision). At the same time, we observe that for higher agreement levels (above 0) the rapidly increasing precision is accompanied by an even more rapidly falling recall. This indicates to us that, at

[2]https://crowdcrafting.org

[3]Data collected during the two phases are available at https://crowdcrafting.org/project/mongolian-lkc and at https://crowdcrafting.org/project/mongolian-lkc-evaluation under CC-BY-SA license.

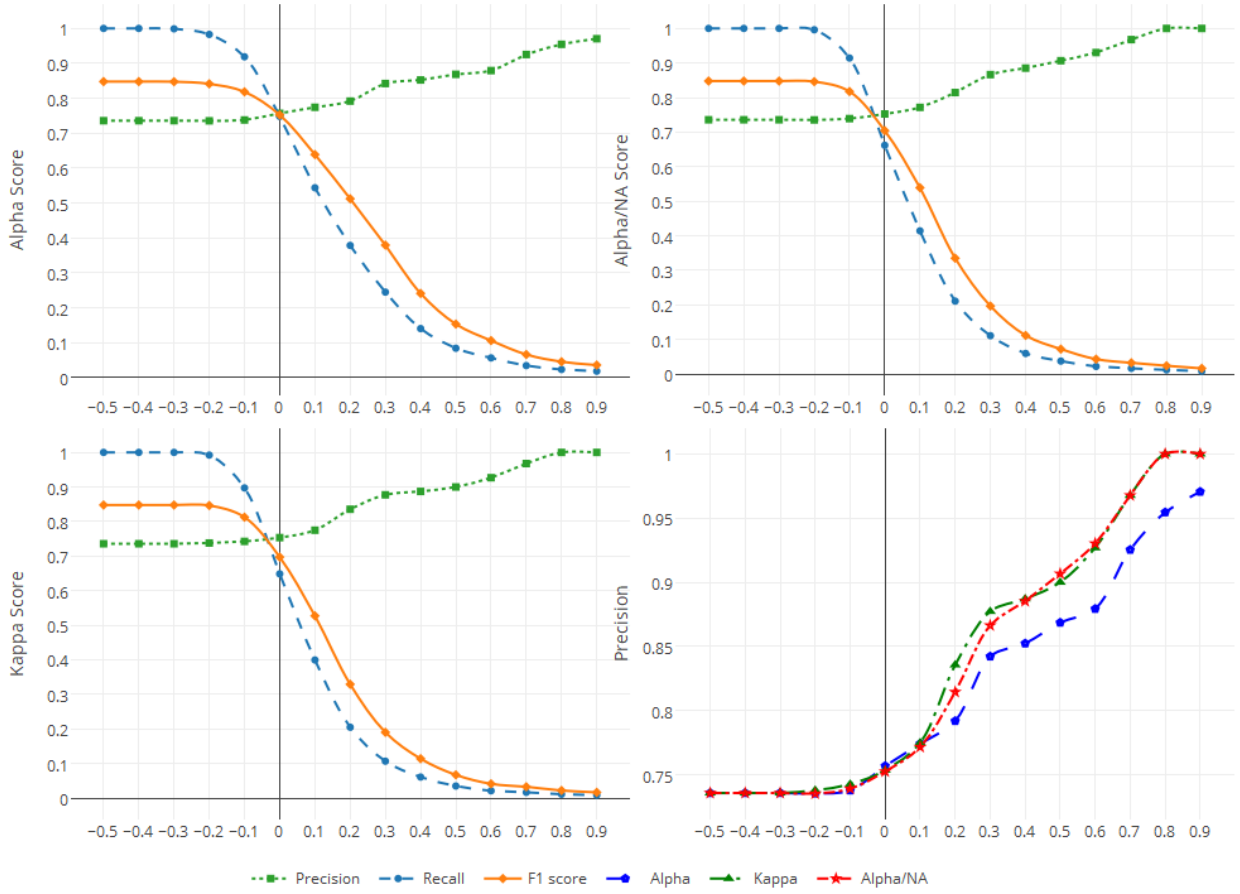

Figure 2: Experiment scores versus agreement values (x-axis)

least in our experiment, inter-annotator agreement was not frequent. Our hypothesis is that both the level of agreement and the overall precision could be improved in at least two distinct ways: by a more careful selection of workers based on competence (which may result in a less heterogeneous output) and by a more sophisticated allocation of tasks to workers. For example, by allocating all hyponyms of a synset to the same worker the *differentia* among them would become more evident and would possibly result in better-quality localizations. We leave these improvements for future work.

6. Conclusion and Future Work

We presented a two-phase crowdsourcing workflow for wordnet localization, consisting of a synset translation phase and of a subsequent validation phase. Two statistical metrics, *Fleiss' kappa* and *Krippendorff's alpha*, were used to compute inter-rater agreement. The overall precision of crowdsourced localizations was 0.74. We found that both agreement metrics can be successfully used to control the precision of the result, the latter monotonically increasing starting from the agreement threshold of about -0.1. However, choosing an agreement threshold significantly higher than this value decreased recall considerably, a fact that hints at a generally low level of worker agreement. In future work we plan to address this point by a more careful selection of workers and of localization tasks. As future work, we will refine our crowdsourcing workflow to improve F1-score as well as our crowd task generation algorithm. Further goal is to apply this crowdsourcing approach to get a usably-sized Mongolian wordnet.

7. Acknowledgements

The research has received funding from the Mongolian Science and Technology Fund under grant agreement SSA_024/2016. The result described in this paper is part of the long-term project *Mongolian Local Knowledge Core*, partially supported by the National University of Mongolia under grant agreement P2017-2383.

8. Bibliographical References

Arcan, M. and Buitelaar, P. (2013). Ontology Label Translation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 40–46. The Association of Computational Linguistics.

Benjamin, M. and Radetzky, P. (2014). Multilingual Lexicography with a Focus on Less-Resourced Languages : Data Mining , Expert Input , Crowdsourcing , and Gamification Acquiring Lexical Data for LRLs. *9th edition of the Language Resources and Evaluation Conference*.

Bentivogli, L. and Pianta, E. (2000). Looking for lexical gaps. In *In Proceedings of the Ninth EURALEX International Congress*, pages 663–669.

Biemann, C. and Nygaard, V. (2010). Crowdsourcing WordNet. In *Proceedings of the 5th Global WordNet conference*, pages 5659–5664, Mumbai, India.

Cimiano, P., Montiel-ponsoda, E., Buitelaar, P., and Espinoza, M. (2010). A Note on Ontology Localization. *Applied Ontology*, 5:127–137.

Espinoza, M., Montiel-Ponsoda, E., and Gómez-Pérez, A. (2009). Ontology localization. In *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*, K-CAP '09, pages 33–40, New York, New York, USA. Universidad Politécnica de Madrid, Spain, ACM Press.

Fišer, D., Tavčar, A., and Erjavec, T. (2014). sloWCrowd: a Crowdsourcing Tool for Lexicographic Tasks. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3471–3475, Reykjavik, Iceland, 5. European Language Resources Association (ELRA).

Ganbold, A., Farazi, F., and Giunchiglia, F. (2014a). An Experiment in Managing Language Diversity Across Cultures. In *eKNOW 2014 : The Sixth International Conference on Information, Process, and Knowledge Management*, number c, pages 51–57, Barcelona, Spain.

Ganbold, A., Farazi, F., Reyad, M., Nyamdavaa, O., and Giunchiglia, F. (2014b). Managing Language Diversity Across Cultures : the English-Mongolian Case Study. *International Journal on Advances in Life Sciences*, 6(3):167–176.

Giunchiglia, F., Dutta, B., and Maltese, V. (2009). Faceted Lightweight Ontologies. In *Conceptual Modeling Foundations and Applications*, volume 5600, pages 36–51.

Giunchiglia, F., Batsuren, K., and Bella, G. (2017). Understanding and exploiting language diversity. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 4009–4017.

Lanser, B., Unger, C., and Cimiano, P. (2016). Crowdsourcing Ontology Lexicons. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3477–3484. European Language Resources Association (ELRA).

Wijesiri, I., Gunathilaka, B., Wimalasuriya, D., Paranavithana, R., Gallage, M., Lakjeewa, M., Dias, G., and de Silva, N. (2014). Building a WordNet for Sinhala. In *GWC 2014: Proceedings of the 7th Global Wordnet Conference*.