Evaluation of Dictionary Creating Methods for Under-Resourced Languages

Eszter Simon and Iván Mittelholcz

Research Institute for Linguistics, Hungarian Academy of Sciences,
H-1068 Budapest, Benczúr u. 33.
{simon.eszter,mittelholcz.ivan}@nytud.mta.hu

Abstract. In this paper, we present several bilingual dictionary building methods applied for Northern Saami–{English, Finnish, Hungarian, Russian} language pairs. Since Northern Saami is an under-resourced language and standard dictionary building methods require a large amount of pre-processed data, we had to find alternative methods. In a thorough evaluation, we compared the results for each method, which proved our expectations that the precision of standard lexicon building methods is quite low. The most precise method is utilizing Wikipedia title pairs extracted via inter-language links, but Wiktionary-based methods also provided useful result.

Keywords: bilingual dictionaries, evaluation, under-resourced languages, dictionary building methods

1 Introduction

Bilingual dictionaries play a critical role not only in machine translation [5] and cross-language information retrieval [8], but also in other NLP applications such as computational semantics and several tasks requiring reliable lexical semantic information [16]. Since manual dictionary building is time-consuming and takes a significant amount of skilled work, it is not affordable in the case of lesser used languages. However, completely automatic generation of clean bilingual resources is not possible according to the state of the art, but it is possible to create certain lexical resources, termed proto-dictionaries, that can support lexicographic and NLP work. Proto-dictionaries contain candidate translation pairs produced by bilingual dictionary building methods. Depending on the method used, they either comprise more incorrect translation candidates and provide greater coverage, or provide precise word pairs at the expense of some decrease in recall; their right size depends on the specific needs.

The standard dictionary building methods are based on parallel corpora. However, such corpora are still available only for the best-resourced language pairs – this is the reason of the increased interest in compiling comparable corpora. The standard approach of bilingual lexicon extraction from comparable corpora is based on context similarity methods (e.g. [7,11]). Recently, source and target vectors are learned as word embeddings in neural networks based on gigaword corpora (e.g. [15]). These methods need a large amount of (pre-processed) data and a seed lexicon which is then used to acquire additional translations of the context words. One of the shortcomings of this approach is that it is sensitive to the choice of parameters such as the size of the context, the size of the corpus, the size of the seed lexicon, and the choice of the association and similarity measures.

The research demonstrated in this paper is part of a project whose general objective is to provide linguistically based support for several small Finno-Ugric (FU) digital communities in generating online content and help revitalize the digital functions of some endangered FU languages. The practical objective of the project is to create bilingual dictionaries for six small FU languages (Udmurt, Komi-Permyak, Komi-Zyrian, Hill Mari, Meadow Mari and Northern Sami) paired with four major languages which are important for these small communities (English, Finnish, Hungarian, Russian).

The status of each language of the world is usually described using the Expanded Graded Intergenerational Disruption Scale (EGIDS) [9], which gives an estimate of the overall development versus endangerment of the language. In this scale – quite counterintuitively – the highest level is 0, where languages are world-wide used *koiné*s, while languages on level 10 are already extinct. Northern Saami is on the highest level among the aforementioned FU languages: its level is 2 (provincial), thus it is used in education, work, mass media, and government within some officially bilingual region of Norway, Sweden and Finland. In the case of the Meadow Mari language, the EGIDS level is 4 (educational), which means that it is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education. The EGIDS level of the other FU languages (Komi-Zyrian, Komi-Permyak, Hill Mari, Udmurt) is 5, i.e. they are developing, which means that there is literature which is available in a standardized form, though it is not yet widespread or sustainable.

Consequently, all these languages are under-resourced, therefore we could not collect enough data for building parallel and comparable corpora. Even if we found some text material in these languages, we could not automatically pre-process them, since – with only rare exceptions – standard text processing tools for these languages are lacking. For these reasons, the aforementioned standard dictionary building methods cannot be used for these languages. Therefore, conducting experiments with alternative methods was needed. We made experiments with several lexicon building methods utilizing crowd-sourced language resources, such as Wikipedia and Wiktionary [3,12].

Having the proto-dictionaries, they were merged for each language pair, and repeated lines were filtered out. These files were then the object of manual validation by native speakers and experts of the languages. In the last phase of the project, we will deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary.

The rest of the article is as follows. In Section 2, the methods used for creating the proto-dictionaries are presented. We conducted thorough evaluation of the resulted dictionaries for language pairs where the source language is Northern Saami and the target language is English, Finnish, Hungarian or Russian. In Section 3, the results of the evaluation is presented: Section 3.1 contains the description of the process of the manual validation of the merged proto-dictionaries, while in Section 3.2, we detail the performance of each dictionary creating method applied here. The article ends with some conclusions and future directions in Section 4.

2 Creating the Proto-dictionaries

2.1 Wikipedia Title Pairs

Wikipedia is not only the largest publicly available database of comparable documents, but it also can be used for bilingual lexicon extraction in several ways. For example, Erdmann et al. [6] used pairs of article titles for creating bilingual dictionaries, which were later expanded with translation pairs extracted from the article texts. Mohammadi and Ghasem-Aghaee [10] extracted parallel sentences from the English and Persian Wikipedia using a bilingual dictionary generated from Wikipedia titles as a seed lexicon. We followed the approach which is common in both articles, thus we created bilingual dictionaries from Wikipedia title pairs using the interwiki links, which resulted in a few hundred candidates for each language pair.

Opinions differ in the literature on how the set of the resulting title pairs is viewed. Some (such as [6]) consider it as a dictionary on its own with a significant amount of multi-word expressions, while others (such as [4]) regard as a parallel corpus and proceed with further steps to extract word translations using methods based on word co-occurrences. In our work, entries where both the source and target language words are one-word units are considered as entries of a bilingual dictionary. The remaining pairs were handled as a parallel corpus, and additional word translations were extracted from it using a procedure based on word co-occurrences (for details, see [3]), but the proto-dictionaries coming from this method are not part of the evaluation presented in this paper.

2.2 Wiktionary-Based Methods

Besides Wikipedia, Wiktionary is also considered as a crowd-sourced language resource which can serve as a source of bilingual dictionary extraction. Although Wiktionary is primarily for human audience, the extraction of underlying data can be automated to a certain degree. Ács et al. [2] extracted translations from the so-called translation tables. Since their tool Wikt2dict is freely available[1], we could apply it for our language pairs. We parsed the English, Finnish, Russian and Hungarian editions of Wiktionary looking for translations in the small FU languages we deal with. With this method, we gathered several translation candidates for almost all language pairs.

Ács [1] expanded the collection of translation pairs, discovering previously non-existent links between translations with a triangulation method. It is based on the assumption that two expressions are likely to be translations, if they are translations of the same word in a third language. With the triangulation mode of Wikt2dict, we could create proto-dictionaries with a few hundred candidates for each language pair.

3 Evaluation

The proto-dictionaries for each language pair were merged, and repeated lines were filtered out. These merged files were then manually validated by a linguist expert of

[1] https://github.com/juditacs/wikt2dict

Northern Saami. The instructions for the validator were as follows. The source and the target word must be a valid word in the language concerned, they must be dictionary forms, and they must be translations of each other. If the source word is not a valid Northern Saami word, the word pair is treated as wrong. If the source word is a valid word but not a dictionary form, the correct dictionary form should be manually added. If the target word is a good translation of the source word but is not a dictionary form, similarly to the former case, the correct dictionary form should be added. If the target word is not a good translation, a new translation should be given.

The following categories come from these instructions:

- ok-ok: The source and the target word are valid words, they are dictionary forms, and they are translations of each other.
- ok-nd: The source and the target word are valid words, they are translations of each other, but the target word is not a dictionary form.
- nd-ok: The source and the target word are valid words, they are translations of each other, but the source word is not a dictionary form.
- nd-nd: The source and the target word are valid words, they are translations of each other, but none of them are dictionary forms.
- ok-wr: The source word is a valid word, it is a dictionary form, but the target word is not a valid word or it is not a correct translation of the source word.
- nd-wr: The source word is a valid word but not a dictionary form, and the target word is not a valid word or it is not a correct translation of the source word.
- wr-xx: The source word is not a valid word.

3.1 Evaluation of the Merged Dictionaries

We made experiments with several lexicon building methods, as detailed above. Applying each method resulted in bilingual resources containing translation candidates for all language pairs. These dictionary files will then be used as the starting point to create the final dictionaries.

Besides the aforementioned proto-dictionaries, the large merged file also contains a proto-dictionary which was not created by us but was downloaded from the Opus corpus [13]. For the Northern Saami–{English, Finnish, Hungarian} language pairs, there are available dictionaries which are lists of “reliable” alphabetic token links extracted from the automatic word alignment created with GIZA++ and the Moses toolkit. First, word pairs where the source and target words were character-level equivalents of each other were removed, since they are probably incorrect word pairs and remaining parts after (or in the lack of) boilerplate removal. The remaining part of the dictionary was also merged into the large dictionary, serving as an interesting example of applying standard lexicon extraction tools for an under-resourced language. The text material from which the Opus proto-dictionaries come is a parallel corpus of KDE4 localization files, where the Northern Saami–English parallel data contain 0.9M tokens, the Northern Saami–Finnish data contain 0.6M tokens, and the Northern Saami–Hungarian data contain 0.8M tokens. At the time of creating the proto-dictionaries, there was no available dic file for Northern Saami–Russian in the Opus corpus.

The large merged dictionaries were evaluated for each category described above; the results can be seen in Table 1. The first impression is that the ok-ok category is much better for sme–rus[2] than for the other language pairs, whose reason is that the sme–rus merged dictionary does not contain translation candidates from the automatically generated Opus dictionary (KDE4). As expected, the standard dictionary creation methods based on parallel texts do not have good performance for under-resourced languages, as pointed out in Section 3.2. This is also proved by the fact that the total number of wrong word pairs (ok-wr + nd-wr + wr-xx) is more than 10% lower for sme–rus than for the other language pairs. Similarly, the total number of word pairs from whose words at least one is not a dictionary form (ok-nd + nd-ok + nd-nd) is also significantly lower in the case of sme–rus. It may be because the KDE4 dictionaries were generated from running text containing suffixed word forms as well, while Wikipedia titles and Wiktionary entries usually are lemmas.

As mentioned in Section 1, the manually validated word pairs will be used as the source material of newly created Wiktionary entries, which contain several obligatory elements. These elements containing morphological, etymological and lexico-semantic information will be generated as automatically as possible. For instance, in the case of the Northern Saami–English language pair, the title of the entry will be the Northern Saami word, while its English definition will be its English translation equivalent.

For this purpose, we need to extract all useful word pairs from the merged dictionary for each language pair. Table 1 contains the number of all word pairs for each language pair and the ratio of the number of useful word pairs and the number of all word pairs. In this case, useful word pairs comprise all word pairs minus the wr-xx category, since correct dictionary forms and translation equivalents were manually added by the human validator. Repeated lines were filtered out; so that the number of lines in the remaining part is the number of useful word pairs.

Table 1. Results for the merged dictionaries

lang pair	**all**	**useful**	**ok-ok**	**ok-nd**	**nd-ok**	**nd-nd**	**ok-wr**	**nd-wr**	**wr-xx**
	(#)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
sme–eng	6,042	92.29	53.26	0.43	9.17	4.10	20.94	4.39	7.71
sme–fin	7,100	91.44	42.28	3.59	6.17	12.48	19.31	7.59	8.56
sme–hun	4,969	90.72	49.57	1.99	6.72	6.36	16.28	9.80	9.28
sme–rus	4,373	95.95	71.74	0.57	3.27	0.14	19.48	0.75	4.05

3.2 Evaluation of the Methods

Category tags given to word pairs in the merged dictionaries were projected onto the corresponding word pairs in the proto-dictionaries. Results for each method were then

[2] We use ISO 639-3 language codes in the article: sme: Northern Saami, eng: English, fin: Finnish, hun: Hungarian, rus: Russian.

summed up across all language pairs, as can be seen in Table 2. Abbreviations of the name of the methods are as follows: WikiTitle: Wikipedia title pairs, W2D ext: Wikt2dict extraction mode, W2D tri: Wikt2dict triangulation mode, KDE4: dic files generated from KDE4 parallel files. Besides category tags, the total number of dictionary entries of proto-dictionaries is presented in the first column.

Table 2. Results for the methods

method	**all**	**ok-ok**	**ok-nd**	**nd-ok**	**nd-nd**	**ok-wr**	**nd-wr**	**wr-xx**
	(#)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
WikiTitle	2,989	94.58	0.33	1.20	0.70	1.97	0.33	0.67
W2D ext	921	91.75	0.00	3.69	0.00	3.04	0.33	1.09
W2D tri	11,714	60.94	0.79	4.23	0.20	26.26	1.05	6.49
KDE4	8,401	29.23	3.61	11.25	16.83	13.81	14.13	10.97

Methods are presented in a descending order based on their performance in the ok-ok category. This score is the *precision* of a method, i.e. the ratio of the number of the correct word pairs and the total number of word pairs. Depending on the research purpose, word pairs containing non-dictionary forms can also be treated as correct translations, thus precision metrics may vary among approaches. Here we use it in a strict sense, thus a word pair is correct iff it is in the ok-ok category.

Some precision-like metrics are generally used for the evaluation of automatically generated bilingual dictionaries. For example, [14] use Precision@1 score, which is the percentage of words where the first word from the list of translations is the correct one, and mean reciprocal rank (MRR), where for a source word w, $rank_w$ denotes the rank of its correct translation within the retrieved list of potential translations. All these metrics are based on the assumption that the method used produces a list of translation candidates along with some confidence or probability measures. Even though it is not the case in our work, we can treat figures in the ok-ok column in Table 2 as Precision@1 scores calculated for a one-unit list of translation candidates.

Not surprisingly, using Wikipedia title pairs as a dictionary is proved to be the most precise method. This resource has very valuable translation texts since these translations were manually made by Wikipedia editors. The second most precise method is using Wikt2dict in extraction mode thus extracting translation equivalents from Wiktionary translation tables. Similarly to that of in the case of Wikipedia, word pairs coming from this method are quite reliable, since Wiktionary entries are manually created. The third method is using Wikt2dict in triangulation method, but there is a 30% decrease in the performance of this method compared to that of the first two ones. As this method does not directly use manually created links, its output may contain incorrect translations. The ok-wr figure for this method is the highest, mainly due to polysemy. The worst result was produced by the method used in the Opus corpus, which is a standard dictionary building method based on parallel text material, using standard alignment and word pair extraction tools developed for well-resourced languages.

Figures of the last method are more flat, i.e. word pairs more uniformly spread among the categories compared to the other methods. It may have several reasons. First, the KDE4 dictionaries were generated from running text containing inflected and derived word forms and lemmas as well. Therefore, the number of non-dictionary forms and wrong translations is higher. (Inflected word forms were treated as valid words in non-dictionary form, while derived forms were categorized as wrong by the validator.) Second, the tools used within the Opus corpus project are not really feasible for under-resourced languages therefore produced more non-dictionary forms and wrong word pairs.

If the number of created dictionary entries can be treated as a kind of *coverage*, it can be said that the Wikt2dict triangulation method has the best coverage, since it produced the largest number of translation candidates. As expected, the method with the worst precision has a quite good coverage. Reversing this logic, the method with the best precision should have the worst coverage, but this is not the case. That is a sign of that evaluating the coverage of a dictionary is greatly challenging. We could gather much more word pairs from Wikipedia titles than from Wiktionary translation tables, which is likely due to the fact that Wikipedia contains more articles compared to the number of translations in Wiktionary's translation tables. Moreover, the number of articles and entries highly depends on the activity of editors knowing the Northern Saami language and willing to create new articles and entries. Coverage of a dictionary can also be measured by comparing the number of its entries to that of another – ideally hand-crafted – dictionary, such as in [4]. For this purpose, we plan to use Wiktionary, which is not an expert-built lexicon but manually edited by thousands of contributors.

4 Conclusions and Future Work

We presented several bilingual dictionary building methods applied for the Northern Saami–{English, Finnish, Hungarian, Russian} language pairs. Since Northern Saami is an under-resourced language and standard dictionary building methods require a large amount of pre-processed data, we had to find alternative methods. In a thorough evaluation, we compared the results for each method, which proved our expectations that the precision of standard lexicon building methods is quite low. The most precise method is using Wikipedia title pairs extracted via inter-language links, but Wiktionary-based methods also provided useful result.

Wiktionary is not only used for extracting data from it, but we want to give our results back to the community, thus translation pairs enriched with obligatory pieces of linguistic information will be uploaded as new entries into Wiktionary. Before uploading new entries, it must be checked whether an entry with the same word already exists in Wiktionary. From this, the number of brand new entries created by us can be easily counted, along with a kind of coverage, if we compare the number of the word pairs in the merged dictionaries to the number of the Northern Saami words in the version of Wiktionary in the language concerned. This, however, remains for future work.

Acknowledgements. The research reported in the paper was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #107885.

References

1. Ács, J.: Pivot-based multilingual dictionary building using Wiktionary. In: 9th Language Resources and Evaluation Conference. ELRA, Reykjavik (2014)
2. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: 6th Workshop on Building and Using Comparable Corpora. pp. 52–58. ACL, Sofia (2013)
3. Benyeda, I., Koczka, P., Váradi, T.: Creating seed lexicons for under-resourced languages. In: GLOBALEX 2016 workshop. ELRA, Portorož (2016)
4. Bharadwaj, G.R., Tandon, N., Varma, V.: An iterative approach to extract dictionaries from Wikipedia for under-resourced languages. In: 8th International Conference on Natural Language Processing. Macmillan Publishers, India (2010)
5. Brown, R.D.: Automated dictionary extraction for "knowledge-free" example-based translation. In: 7th International Conference on Theoretical and Methodological Issues in Machine Translation. pp. 111–118 (1997)
6. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: An Approach for Extracting Bilingual Terminology from Wikipedia. ACM Transactions on Multimedia Computing, Communications, and Applications 5(4), 1–17 (2009)
7. Fung, P., Yee, L.Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In: 17th International Conference on Computational Linguistics. pp. 414–420. ACL, Stroudsburg (1998)
8. Grefenstette, G.: The Problem of Cross-Language Information Retrieval. In: Grefenstette, G. (ed.) Cross-Language Information Retrieval, pp. 1–9. Kluwer Academic Publishers, Boston (1998)
9. Lewis, M.P., Simons, G.F.: Assessing endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique 55(2), 103–120 (2010)
10. Mohammadi, M., Ghasem-Aghaee, N.: Building Bilingual Parallel Corpora Based on Wikipedia. In: 2nd International Conference on Computer Engineering and Applications. pp. 264–268 (2010)
11. Rapp, R.: Identifying word translations in non-parallel texts. In: 33rd Annual Meeting of the Association for Computational Linguistics. pp. 320–322. ACL, Stroudsburg (1995)
12. Simon, E., Benyeda, I., Koczka, P., Ludányi, Zs.: Automatic creation of bilingual dictionaries for Finno-Ugric languages. In: 1st International Workshop on Computational Linguistics for Uralic Languages. Tromsø (2015)
13. Tiedemann, J.: News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In: Nicolov, N., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007, pp. 237–248. John Benjamins, Borovets (2009)
14. Vulić, I., De Smet, W., Moens, M.F.: Identifying word translations from comparable corpora using latent topic models. In: 49th Annual Meeting of the Association for Computational Linguistics. pp. 479–484. ACL, Stroudsburg (2011)
15. Vulić, I., Moens, M.F.: Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In: 53rd Annual Meeting of the Association for Computational Linguistics. pp. 719–725. ACL, Stroudsburg (2015)
16. Zesch, T., Müller, C., Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: 6th Language Resources and Evaluation Conference. ELRA, Marrakech (2008)