Evaluation of Dictionary Creating Methods for Under-Resourced Languages

Eszter Simon & Iván Mittelholcz
Research Institute for Linguistics, Hungarian Academy of Sciences
{simon.eszter,mittelholcz.ivan}@nytud.mta.hu

Abstract

Here 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. Here we compare the results for each method, which proves 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 provide useful result.

Introduction

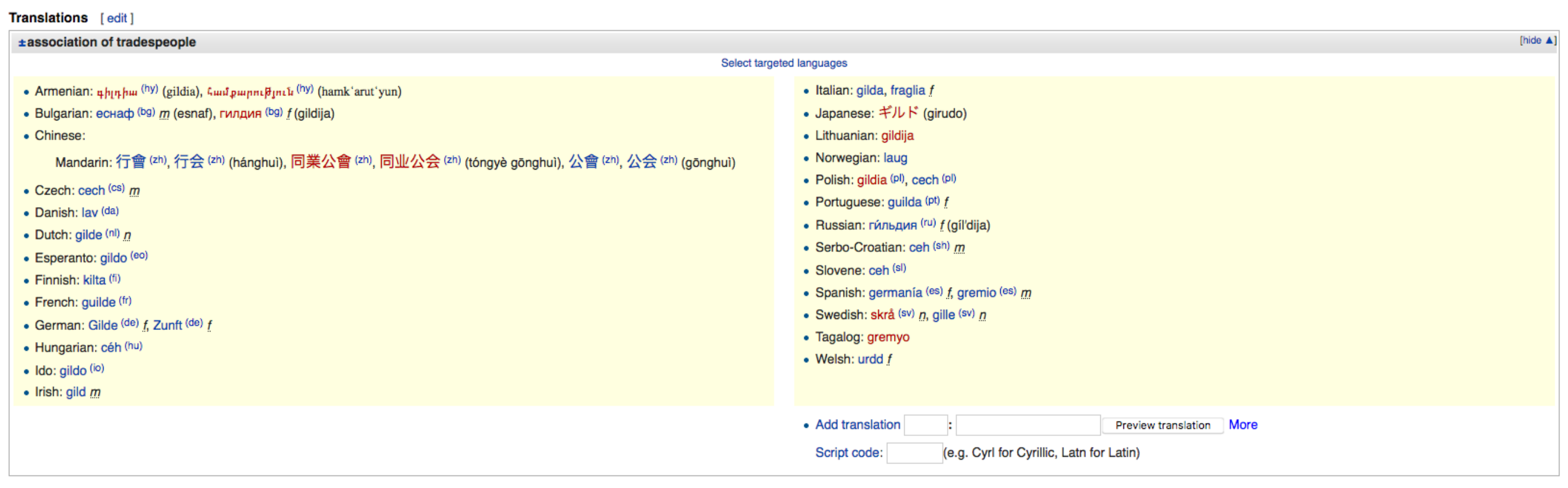
Since manual dictionary building is time-consuming and takes a significant amount of skilled work, it is not affordable in the case of lesser used languages. However, completely automatic generation of clean bilingual resources is not possible according to the state of the art, but it is possible to create so-called **proto-dictionaries** which contain candidate translation pairs produced by dictionary building methods.

The **standard dictionary building methods** are based on parallel or comparable corpora, thus they 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.

Northern Saami is under-resourced, and the standard text processing tools are lacking (with the exception of Giellatekno). Therefore, the standard dictionary building methods cannot be used for this language, thus 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.

Creating the proto-dictionaries

1. **Wikipedia title pairs:** we created bilingual dictionaries from Wikipedia title pairs using the interwiki links
2. **Wiktionary-based methods** using the Wikt2dict tool:
   `https://github.com/juditacs/wikt2dict`
   (a) **Wikt2dict extraction:** we parsed the English, Finnish, Russian and Hungarian editions of Wiktionary and extracted translations from the translation tables



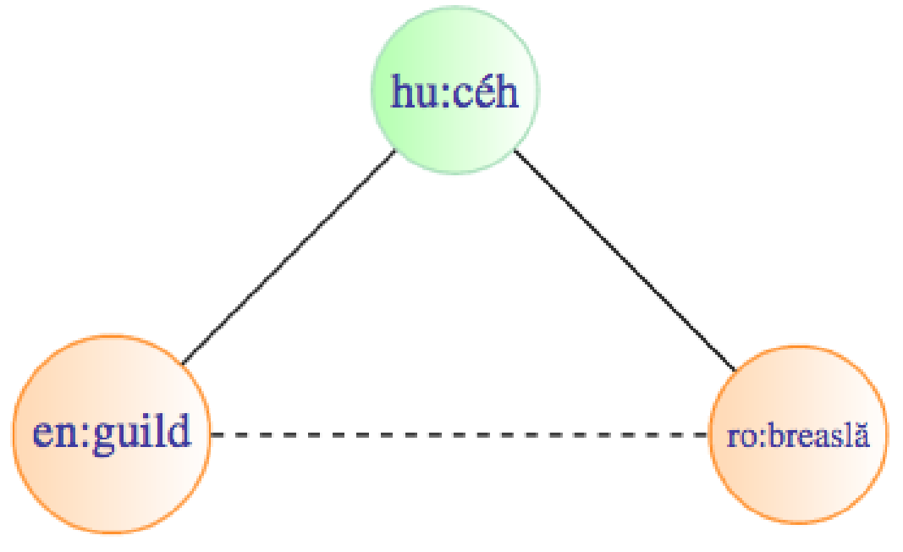
Translations [edit]

±association of tradespeople [hide ▲]

Select targeted languages

- Armenian: գիլդիա (hy) (gildia), համքարություն (hy) (hamkʻarutʻyun)
- Bulgarian: еснаф (bg) *m* (esnaf), гилдия (bg) *f* (gildija)
- Chinese:
  Mandarin: 行會 (zh), 行会 (zh) (hánghuì), 同業公會 (zh), 同业公会 (zh) (tóngyè gōnghuì), 公會 (zh), 公会 (zh) (gōnghuì)
- Czech: cech (cs) *m*
- Danish: lav (da)
- Dutch: gilde (nl) *n*
- Esperanto: gildo (eo)
- Finnish: kilta (fi)
- French: guilde (fr)
- German: Gilde (de) *f*, Zunft (de) *f*
- Hungarian: céh (hu)
- Ido: gildo (io)
- Irish: gild *m*
- Italian: gilda, fraglia *f*
- Japanese: ギルド (girudo)
- Lithuanian: gildija
- Norwegian: laug
- Polish: gildia (pl), cech (pl)
- Portuguese: guilda (pt) *f*
- Russian: ги́льдия (ru) *f* (gílʹdija)
- Serbo-Croatian: ceh (sh) *m*
- Slovene: ceh (sl)
- Spanish: germanía (es) *f*, gremio (es) *m*
- Swedish: skrå (sv) *n*, gille (sv) *n*
- Tagalog: gremyo
- Welsh: urdd *f*
- Add translation [ ] : [ ] Preview translation More
  Script code: [ ] (e.g. Cyrl for Cyrillic, Latn for Latin)

   (b) **Wikt2dict triangulation:** discovering previously non-existent links between translations with a triangulation method, which 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



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 Northern Saami. The following categories come from the validation:

- **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

Evaluation of the merged dictionaries

Serving as an interesting example of applying standard lexicon extraction tools for an under-resourced language, proto-dictionaries which were not created by us but were downloaded from the **Opus corpus** were added to the merged dictionaries. 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**. The text material from which the Opus proto-dictionaries come is a parallel corpus of **KDE4** localization files.

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

Table 1: Results for the merged dictionaries

The standard dictionary creation methods have a **lower performance** on less-resourced language pairs, which is proved by the following facts:

- the ok-ok category is much better for sme–rus than for the other language pairs, because the sme–rus merged dictionary does not contain translation candidates from the automatically generated Opus dictionary
- 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

We also extracted 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.

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 summed up across all language pairs, as can be seen in Table 2. The total number of dictionary entries of proto-dictionaries is also presented in the first column.

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

Table 2: Results for the methods

The ratio of the number of the correct (ok-ok) word pairs and the total number of word pairs can be treated as the **precision** of a method.

I. **Wikipedia title pairs:** valuable, since translations were manually made by Wikipedia editors
II. **Wikt2dict extraction:** reliable, since Wiktionary entries are manually created
III. **Wikt2dict triangulation:** 30% decrease, since it does not directly use manually created links
IV. **KDE4:** the lowest result, since it is a standard method developed for well-resourced languages

The number of the created dictionary entries can be treated as a kind of **coverage**.

I. **Wikt2dict triangulation:** methods with low precision scores have a quite good coverage
II. **KDE4:** methods with low precision scores have a quite good coverage
III. **Wikipedia title pairs:** Wikipedia contains more articles
IV. **Wikt2dict extraction:** Wiktionary's translation tables contain less translations

Discussion & Conclusion

- Northern Saami is an under-resourced language & standard dictionary building methods require a large amount of pre-processed data → we had to find alternative methods
- the results proved our expectations: the precision of the standard lexicon building methods is quite low
- the most precise method is using Wikipedia title pairs, but Wiktionary-based methods also provided useful result
- the KDE4 dictionaries were generated from running text → the number of non-dictionary forms and wrong translations is higher
- the ok-wr figure for the Wikt2dict triangulation method is the highest → polysemy
- 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

Future work

- precision and coverage scores for the other under-resourced languages we deal with
- another kind of coverage: comparing 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
- translation pairs enriched with obligatory pieces of linguistic information will be uploaded as new entries into Wiktionary

Acknowledgements

The research reported here was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #107885. The project's 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).