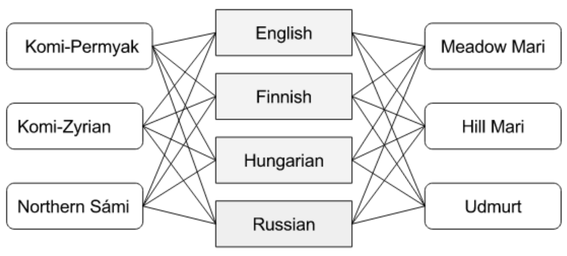
Evaluation of Dictionary Creating Methods for Finno-Ugric Minority Languages

Zsanett Ferenczi, Iván Mittelholcz, Eszter Simon, Tamás Váradi
Research Institute for Linguistics, Hungarian Academy of Sciences
{FAMILY_NAME.FIRST_NAME}@nytud.mta.hu

Introduction

General objective: to provide linguistically based support for several small FU digital communities to generate online content and help to revitalize the digital functions of some FU minority languages.
Practical objective: to create bilingual dictionaries for six FU minority languages paired with four major languages which are important for these communities and to deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary.



Language	ISO	EGIDS	Population	Location	Writing
Saami, North	sme	2	26,000	Norway, Sweden, Finland	Latin
Mari, Meadow	mhr	4	470,000	Russia	Cyrillic
Mari, Hill	mrj	5	30,000	Russia	Cyrillic
Komi-Zyrian	kpv	5	156,000	Russia	Cyrillic
Komi-Permyak	koi	5	63,000	Russia	Cyrillic
Udmurt	udm	5	340,000	Russia	Cyrillic

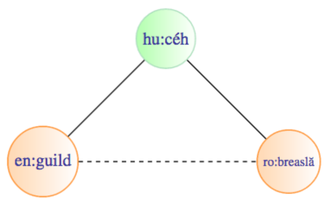
The **standard dictionary building methods** are based on parallel or comparable corpora, thus they need a large amount of (pre-processed) data and a seed lexicon which is then used to acquire additional translations of the context words. The aforementioned FU languages, however, are under-resourced, and the standard text processing tools are lacking (with the exception of Giellatekno). Therefore, the standard dictionary building methods cannot be used for these languages, thus conducting experiments with **alternative methods** was needed.

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.

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

   (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

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 word pairs extracted from the automatic word alignment **created with GIZA++ and the Moses toolkit**.

Manual Validation

The proto-dictionaries for each language pair were merged, and repeated lines were filtered out. These merged files were then manually validated by native speakers and linguist experts of the FU languages. The following categories come from the validation:

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

Precision

Results for each method were summed up across all language pairs (Table 1). The total number of dictionary entries of proto-dictionaries is presented in the first column. The ratio of the number of the ok-ok word pairs to the total number of word pairs can be treated as the **precision** of a method.

method	all (#)	ok-ok (%)	ok-nd (%)	nd-ok (%)	nd-nd (%)	ok-wr (%)	nd-wr (%)	wr-xx (%)
W2D ext	1,965	71.76	1.22	5.75	15.17	4.63	0.36	0.76
W2D tri	23,066	56.61	1.79	2.98	3.06	30.38	1.1	3.82
WikiTit	16,854	54.11	2.97	5.57	32.5	2.92	0.49	0.75
Opus	8,401	27.57	3.99	10.4	18.64	13.99	14.57	10.69

Table 1: Results for the methods.

The large merged dictionary of each language pair was evaluated for each category of the manual validation (Table 2). The first column shows the total number of word pairs gathered with all methods for the language pair. Since the validated dictionaries are the input of generating new Wiktionary entries, we need to extract all useful word pairs from the merged dictionary for each language pair. The second column of the table contains the ratio of the useful word pairs comprising all word pairs minus the wr-xx category.

lang pair	all (#)	useful (%)	ok-ok (%)	ok-nd (%)	nd-ok (%)	nd-nd (%)	ok-wr (%)	nd-wr (%)	wr-xx (%)
koi–eng	1,251	96.64	74.82	0.16	7.83	0.00	13.67	0.16	3.36
koi–fin	592	98.82	65.20	3.04	9.97	0.84	19.59	0.17	1.18
koi–hun	540	93.15	70.19	3.33	4.63	1.30	13.52	0.19	6.85
koi–rus	611	98.85	65.47	2.95	16.69	1.47	11.62	0.65	1.15
kpv–eng	902	100.00	66.30	0.22	0.55	30.16	2.55	0.22	0.00
kpv–fin	577	100.00	57.89	3.29	0.69	37.09	0.87	0.17	0.00
kpv–hun	523	99.81	49.71	1.34	0.96	43.98	3.82	0.00	0.19
kpv–rus	544	100.00	63.60	8.64	9.93	14.52	3.31	0.00	0.00
mhr–eng	2,549	100.00	44.41	2.55	4.04	22.40	26.09	0.51	0.00
mhr–fin	2,565	100.00	50.80	1.05	3.31	20.74	23.63	0.47	0.00
mhr–hun	1,647	100.00	52.64	0.97	5.89	25.20	14.15	1.15	0.00
mhr–rus	1,707	100.00	40.01	2.11	4.28	17.28	35.56	0.76	0.00
mrj–eng	2,334	100.00	44.09	0.17	9.04	43.10	3.08	0.51	0.00
mrj–fin	1,013	100.00	20.24	7.70	9.77	52.32	8.59	1.38	0.00
mrj–hun	942	100.00	34.18	4.99	12.95	41.08	5.20	1.59	0.00
mrj–rus	835	100.00	27.07	11.26	9.58	31.38	16.89	3.83	0.00
sme–eng	6,041	91.97	47.57	3.77	7.33	6.56	21.65	5.08	8.03
sme–fin	7,100	91.03	42.03	3.42	5.42	12.56	19.92	7.66	8.97
sme–hun	4,969	90.78	48.48	1.67	6.72	6.62	17.05	10.24	9.22
sme–rus	4,373	95.40	71.35	0.50	2.56	0.18	20.05	0.75	4.60
udm–eng	2,087	99.14	77.19	3.07	0.91	0.29	17.59	0.10	0.86
udm–fin	1,700	99.65	49.12	2.06	1.06	18.82	28.06	0.53	0.35
udm–hun	1,204	99.50	57.14	1.74	1.50	23.17	15.45	0.50	0.50
udm–rus	1,226	98.78	8.56	2.04	0.98	65.25	20.64	1.31	1.22

Table 2: Results for the merged dictionaries.

Coverage

The number of the created dictionary entries can be treated as a kind of **coverage** (Table 1). Coverage of a dictionary can also be measured by comparing the number of its entries to that of a hand-crafted dictionary. Since our newly created word pairs are to be transformed into Wiktionary articles, for this purpose, here we used Wiktionary (Table 3).

lang pair	all (#)	useful (%)	useful (#)	remaining (#)	Wiktionary (#)	common (#)	new (#)	coverage (%)	improvement (%)
kom–eng	2,153	95.26	2,051	655	54	25	630	46.30	1166.67
kom–fin	1,169	95.54	1,117	687	42	27	660	64.29	1571.43
kom–hun	1,063	95.29	1,013	699	152	35	664	23.03	436.84
kom–rus	1,155	92.54	1,069	672	465	223	449	47.96	96.56
chm–eng	4,883	98.83	4,826	1,670	347	53	1,617	15.27	465.99
chm–fin	3,578	98.57	3,527	1,903	443	213	1,690	48.08	381.49
chm–hun	2,589	98.29	2,545	1,633	34	12	1,621	35.29	4767.65
chm–rus	2,542	98.11	2,494	1,493	848	201	1,292	23.70	152.36
sme–eng	6,041	91.97	5,556	2,531	4,072	882	1,649	21.66	40.50
sme–fin	7,100	91.03	6,463	2,862	817	422	2,440	51.65	298.65
sme–hun	4,969	90.78	4,510	2,392	206	146	2,246	70.87	1090.29
sme–rus	4,373	95.40	4,172	2,034	306	237	1,797	77.45	587.25
udm–eng	2,087	99.14	2,069	751	32	15	736	46.88	2300.00
udm–fin	1,700	99.65	1,694	828	55	45	783	81.82	1423.64
udm–hun	1,204	99.50	1,198	729	128	69	660	53.91	515.62
udm–rus	1,226	98.78	1,211	578	643	247	331	38.41	51.48

Table 3: Results for the language pairs.

Conclusions and Future Work

- FU minority languages are under-resourced languages, and standard dictionary building methods require a large amount of pre-processed data, thus we had to find alternative methods.
- The results proved our expectations: the precision of the standard lexicon building methods is quite low, while Wiktionary-based methods are proved to be the most precise methods.
- We provide freely available professional online multilingual lexical data for digital communities of some FU minority languages via Wiktionary and via our website (http://finnotka.nytud.hu)

Acknowledgements

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