Creating seed lexicons for under-resourced languages

Ivett Benyeda, Péter Koczka, Tamás Váradi
Research Institute for Linguistics of the Hungarian Academy of Sciences
H-1068 Benczúr utca 33., Budapest
{benyeda.ivett, koczka.peter, varadi.tamas} @nytud.mta.hu

Abstract

In this paper we present methods of creating seed dictionaries for an under-resourced language, Udmurt, paired with four thriving languages. As reliable machine readable dictionaries do not exist in desired quantities this step is crucial to enable further NLP tasks, as seed dictionaries can be considered the first connecting element between two sets of texts. For the language pairs discussed in this paper, detailed description will be given of various methods of translation pair extraction, namely Wik2Dict, triangulation, Wikipedia article title pair extraction and handling the problematic aspects, such as multiword expressions (MWUs) among others. After merging the created dictionaries we were able to create seed dictionaries for all language pairs with approximately a thousand entries, which will be used for sentence alignment in future steps and thus will aid the extraction of larger dictionaries.

Keywords: under-resourced languages, dictionary extraction, seed dictionaries, comparable corpora

1. Introduction

In this paper we will present a method of creating seed dictionaries for four language pairs: Udmurt–Russian, Udmurt–Finnish, Udmurt–English and Udmurt–Hungarian. The research demonstrated in this paper is part of a project whose aim is to support small Finno-Ugric languages in generating on-line content. The goal of this project is to create bilingual dictionaries and parallel corpora for six small Finno-Ugric (Udmurt, Komi-Permyak, Komi-Zyrian, Hill Mari, Meadow Mari and Northern Sámi) languages paired with four thriving ones which are important for these small communities. For creating these sources a seed dictionary is essential in the process. In this paper we are focusing on the Udmurt language and demonstrate the process of creating seed dictionaries for language pairs where Lang1 is Udmurt, of which a detailed introduction is given in section 2., and Lang2 is {English, Finnish, Hungarian, Russian}.

As reliable machine readable dictionaries are not available for Udmurt in sufficient size, we had to create these lexicons ourselves. The lack of parallel corpora for these language pairs makes the process challenging. We created comparable corpora for the above language pairs ranging from 96 133 tokens (Udmurt–Hungarian) to 225 914 tokens (Udmurt–English) in size.

So-called seed dictionaries play a significant role in extracting bilingual information from parallel and especially from comparable corpora. Seed dictionaries can be considered the first connecting elements between two sets of texts, allowing the extraction of parallel sentences from comparable corpora among others. Context similarity methods, the standard approach to bilingual lexicon extraction from comparable corpora (e.g. (Fung and Yee, 1998)), crucially rely on seed lexicons so the quality of these dictionaries is critical even if they are created automatically without supervision. Although bilingual dictionaries are easily accessible for widely-spoken languages (which can be used easily as a seed lexicon), it is still a challenge even to get a small set of bilingual dictionaries for endangered languages as these are rarely available in digital format and their quality is often questionable.

Fortunately we could download dictionaries for two of the language pairs. The first step was processing these sources. Using these lexicons we could create dictionaries for Udmurt–Russian and Udmurt–Finnish language pairs. This method is discussed in section 3.

An additional source was the Wiktionary dictionary. The Wik2dict tool made us able to extract translation pairs for language pairs which are in our interest.

For creating more translation pairs from Wiktionary we used the triangulation method (Ács, 2014). This technique uses a pivot translation to get additional word pairs. “Triangulation is based on the assumptions that two expressions are likely to be translations if they are translations of the same word in a third language” – (Ács, 2014). As these word pairs were created automatically we consider the output less reliable and these pairs will be processed later with Wikipedia title dictionary.

As the first and second lexicons were made manually we considered the entries from them reliable and these were not validated by experts.

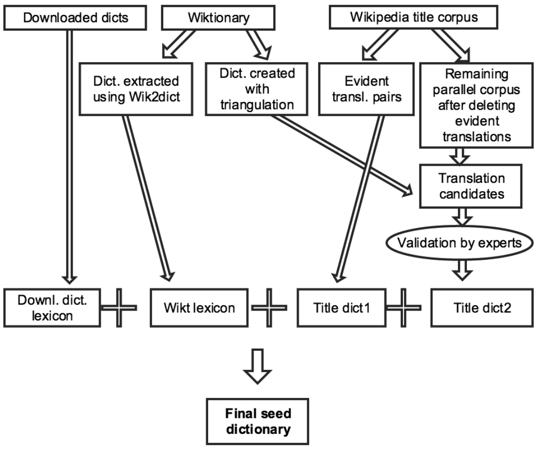

Figure 1: The process of creating seed dictionaries.

The third source of seed dictionary building was the Wikipedia article title pairs. This parallel corpus was processed in two steps. In the first step the evident translation pairs were extracted. This method resulted in another lexicon. The details of this process can be found in section 4.2. As these pairs were made by Wikipedia users we also considered the output reliable (these contains title pairs where both title are one word long or a one word long title is paired with a multi-word expression (MWE)).

After these translations were deleted from the parallel corpora the remaining pairs were processed and additional translations were extracted. These translation candidates were merged with another lexicon which was extracted using the triangulation method and these translation pairs were processed together. The result of this step was another lexicon which was validated by experts.

At the end of the dictionary building all of the created small lexicons were concatenated and this resulted one dictionary with approximately 1000 entries for each language pairs. After filtering out duplicates these dictionaries could be used as seed dictionaries.

2. The sociolinguistic situation of Udmurt

The language in centre of this paper is Udmurt, among the so-called thriving languages (Russian, English, Finnish and Hungarian) which are also mentioned. While the thriving languages are well known, Udmurt might need some introduction. Even if Udmurt is considered as the most visible and one of the bigger of the Finno-Ugric languages of the Russian Federation (Pischlöger, 2014), it is, unfortunately, still classified by the UNESCO as definitely endangered (UNESCO Atlas 2014)[1]. The sociolinguistic situation of the language is clearly supporting this classification. According to the 1989 Russian Census, 747.000 people declared themselves to be of Udmurt origin and of these people circa 70% (520.000 people) speak Udmurt as their mother tongue (Winkler 2001). The 2002 Census showed a significant drop in both the number of speakers and people who identified themselves to be of Udmurt origin, 637.000 people with around 73% claimed to be able to speak the language (464.000 people) (from Perepis 2002)[2]. The most recent Russian Census shows even more alarming numbers, only 59%, 324.000 people of the self-identified Udmurts (550.000) could speak the language to a certain degree, but not exclusively fluently. Younger people, especially in urban areas, are prone to Russification, the generation that has the most access to new technology. Udmurts living in scattered settlements usually form a majority in said communities and thus preserved their language very well, but given the location and infrastructural features of these villages, along with the demographic composition of the community (younger people tend to give up village life and move to urban areas where Russian is the language of everyday life) the speakers there are unlikely to have a significant web presence.

For Udmurt, there are prescriptive rules and a standardized orthography (Winkler, 2001), which makes it possible to publish Udmurt language materials, including mass media (TV and radio broadcast, books, newspapers, etc.) and most importantly, from a language revitalization point of view, Web 2.0 and especially the Social Network Sites (Pischlöger, 2014) can increase the visibility of the language and provide material for research. While Social Network Sites have a more relaxed atmosphere and hardly any sign of linguistic purism, Wikipedia articles, Wikipedia being another exceptional example of a community driven Web 2.0 project, are expected to be well written, following the orthographic rules of the language.

Udmurt, being a Finno-Ugric language, is heavily agglutinating. This means that morphological analysers have to deal with rather complex constructions and while there is a well performing tool available for years[3], unfortunately it is not open source. There are initiatives to create a HFST-based analyser for Udmurt, among many other Uralic languages, at Giellatekno[4] in Tromsø, but the development of such tools is very laborious.

3. Extracting word pairs from existing lexicons

As it was mentioned in the previous paragraphs Udmurt is a severely under-resourced language. Considering this, it is no surprise that we only have Wikipedia texts as comparable corpora for the mentioned language pairs. Unfortunately we have not found any translation texts (which would be suitable for parallel corpora) in electronic form. For processing the texts of Wikipedia article pairs it is necessary to have a reliable and relatively large seed dictionary. Although there are some existing e-dictionaries, these are quite small and we decided to expand them. We also used Wiktionary entries to have more translation pairs which also resulted in a few additional dictionary entries. As a first step, we extracted translation pairs from downloaded lexicons which were in different formats. Using these resources we created additional lexicons with a few hundred entries.

Sources used for creating bilingual seed lexicons:

- Small downloaded dictionaries from the web
  We could download 90 translation word pairs for Udmurt–Finnish from Goldendict[5] and 1466 pairs for Udmurt–Russian. Another 136 translations could be downloaded from Apertium[6,7], another relevant site.

- Word pair extraction from Wiktionary
  Using the Wikt2dict tool we extracted translation pairs for three of the language pairs.

- Extracting additional word pairs from Wiktionary using the triangulation method
  The Triangulating method is also based on Wiktionary,

[1] http://www.unesco.org/culture/languages-atlas/

[2] http://www.perepis2002.ru

[3] http://www.morphologic.hu/urali/

[4] http://giellatekno.uit.no

[5] http://yoshkarola.bezformata.ru/listnews/slovari-dlya-goldendict/

[6] https://svn.code.sf.net/p/apertium/svn/nursery/apertium-udm-rus/apertium-udm-rus.udm-rus.dix

[7] https://svn.code.sf.net/p/apertium/svn/incubator/apertium-fin-udm/apertium-fin-udm.fin-udm.dix

but it deals with extracting more translations using so-called pivot elements. Using this we were able to extract an another set of translations.

Language pair (L1-L2)	E-dict	Wikt2dict	Wikt triang.
Udmurt-English	-	102	1202
Udmurt-Finnish	90	-	1213
Udmurt-Russian	1602	276	811
Udmurt-Hungarian	-	11	723

Table 1: The number of translation word pairs in the extracted lexicons

As Wikipedia texts are highly varied in their topics, utilizing a similarly comprehensive seed dictionary is essential. As it can be seen in the table above, we were able to download existing lexicons for Udmurt-Finnish and Udmurt-Russian, and extract a number of translation pairs from Wiktionary for Udmurt-English, Udmurt-Russian and Udmurt-Hungarian using the Wikt2dict tool. The other approach based on Wiktionary is the so-called triangulating method. Using this we could extract approximately a thousand of word translations.

4. Using Wikipedia title corpus to extract translation word pairs

While there are no extensive parallel corpora for language pairs formed with Udmurt, we can still find a minuscule parallel subset of the Wikipedia articles, their titles. Wikipedia title translation pairs can be easily extracted using the so-called interwiki links, or otherwise called Wikipedia interlanguage links (ILL). This resource has very valuable translation texts since these translations are manually made by Wikipedia contributors (Hara et al., 2008). Unfortunately processing them is not as obvious as it seems at first sight. While it is quite often the case that both of the titles are one word long, sometimes one of the languages appear to have a multiword expression. When titles in both languages are single words, they can be directly treated as a bilingual dictionary entry. In some cases this could be true for a number of title-based translation pairs even where we find multiword expressions, phrases or sentence fragments, for which reason we can consider a subset of the title pairs a comparable corpora.

4.1. Preprocessing title pairs

4.1.1. Language Identification

Although these title pairs are made manually by Wikipedia users or editors, allowing them to be considered a reliable and valuable source, there are some pairs which are of hardly any use when it comes to bilingual dictionary building. This is the case, for example, if the text is not in the expected language as it often happens with articles about plants and animals where one can find the scientific, latin name instead of the generally used term in a given language. Since it is quite frequent that the pair of the Udmurt title in the other language (in our case these are the English, Finnish, Hungarian and Russian titles) is the Latin name, we decided to filter out these using a language identification tool. As expected, these language identification tools are well performing if the input is longer, but title texts have a tendency to be rather short, which causes this identification and filtering process to become more difficult and less reliable without the use of any precautionary measure. This means that if we used LANGID[8] in a way when everything was filtered out from the corpora which were not written in the given language (according to langid) not taking into account the possibility of falsely identified texts, a remarkable number of good translation candidates were left out. Because of this reason we decided to filter out candidates where texts in L2 were written in Latin language. This technique allows the more careful, more precise filtering of titles that are of no interest.

4.1.2. Filtering Out Stopwords

For L2 titles we used stopword lists to make the output better. This was done using the stopword lists of PYTHON's NLTK[9] module. For Udmurt, we had to avoid using any stopword lists. Using the highest frequency words from an Udmurt Wikipedia based frequency list, the resulting output had an easily noticeable drop in quality as the list used was noisy, contaminated with strings that cannot be considered stopwords.

4.2. Extracting translation pairs where the correspondence is evident

As it was mentioned above, we considered a part of this resource as a dictionary. Following the pre-processing and modifying the corpora to be case-insensitive, the next processing step was the extraction of word pairs. Extracting the pairs where the title1 and title2 are one word long, we created a dictionary from this title corpora. If only one of the pair is one word long and its translation is longer we consider it also as a dictionary item and the longer translation is handled as an MWU.

Udmurt (L1)	English (L2)
дунай	danube
донецк	donetsk
койык	moose
тӧдьыгыплы	lily of the valley
соборной мечеть	mosque

Table 2: Examples from the lexicon

After the extraction of the dictionaries files were created containing only reliable data. After this process only longer title pairs remained in the corpus.

4.3. Extracting other word pairs from the remaining comparable corpora

4.3.1. The handling of multi-word expressions

This process is quite robust and it is based on word translation co-occurrence. The script for processing these is able to handle multi word units using an n-gram model.

[8] https://github.com/saffsd/langid.py/tree/master/langid

[9] http://www.nltk.org/

L1-L2	Whole title corp.	Extracted dict.
Udm-Eng	2701	1172
Udm-Hun	1428	589
Udm-Rus	2519	265
Udm-Fin	1663	795

Table 3: Dictionary sizes

Our observation is that the longest multi-word expression in these small corpora is three word long. So bigrams and trigrams were used in this process. This multi-word unit extraction method is quite simple. It counts how many times the bi- or trigram occurs in the text and how many times these words are found in other contexts. If it is repeated that these are occurring together, these are handled as multi-word units and marked and concatenated with underscores in the corpora.

4.3.2. Expanding the remaining title corpora with other word pairs and finding translation candidates

As the output candidates of triangulation method are not always reliable, it seems to be a reasonable idea to use these candidates with the remaining title parallel corpora helping to choose the valid translations. To make the next process easier we deleted the words which were already in the extracted corpora. If the L1 word, which is in the existing dictionary, can be found in the longer parallel L1 title text, and it is also the case with the translation word and L2 text, these are deleted. For example, if the extracted dictionary contains the pair *ӟуч* – *language*, and the remaining parallel title corpora contains entries like the pair *ӟуч кыл* – *russian language*, the output of this process will result in the pair *ӟуч* – *russian*.

After this step each L1 word in the actual entry is paired with each L2 word in the same entry. These translation candidates will be scored using a method discussed in the next paragraph.

L1 title	L2 title
калыккуспо телефон код	telephone numbering plan

Table 4: An example entry

L1	L2
калыккуспо	telephone
калыккуспо	numbering
калыккуспо	plan
телефон	telephone
телефон	numbering
телефон	plan
код	telephone
код	numbering
код	plan

Table 5: Candidates created from the previous entry

4.3.3. Calculating scores for translation candidates

Bharadwaj G., Tandon and Varma (Rohit Bharadwaj et al., 2010) used a method for calculating scores which were based on translation co-occurrences. Although the scores in our work are also based on translation co-occurrences among the candidates, there are some plus weights which make the method a bit more complex.

The created candidates are stored in a DICT TYPE in Python (a DICT TYPE is a hash-table). The keys of this dict are the Udmurt (L1) words. Each key have a list value which stores tuples[10] (the tuple contains the L2 translation candidate and its actual score).

KEY: UDMWORD VALUE [(L2TRANSL, SCORE), (L2TRANSL, SCORE), ...]

The first idea is that a candidate is more likely to be reliable if it can be found multiple times in this corpus. Each time when an L1 word is paired in the corpus with an L2 translation it gets plus one score (if this translation pair has not existed it will be created). As in these language pairs it is mostly true that good translations are in the same position L2 candidates which are in the same position as the L1 candidate get another plus 1 score. It is also reasonable that if the title pairs are one word long (because other words were deleted as they existed in our previously created dictionary) it is much more probable that they are good translations. In this case they get another plus score.

4.3.4. Choosing best translations and defining a threshold of candidate scores

As we wanted to have the most reliable translations, the threshold was quite high at the beginning which resulted in a rather small output. The solution to get more good translations was not just lowering the threshold, as it resulted in the reduced quality of results. Because of this reason we decided to run the extraction and scoring method iteratively several times. First time, the threshold is rather high, resulting in only a few translations. Following this we stored these pairs in a list and deleted them from the parallel corpora as described in paragraph 4.3.2. This means that the parallel corpora gets smaller in each iteration and the list of extracted translations grows. The threshold is lowered in each iteration and as the pairs in the parallel corpora are always shorter because of the deletion of good translation word pairs (and as one to one word translations get plus scores) we will get more translations in each iteration. The script ran 9 times and the first threshold was 20 which was decreased by 2 at each iteration. At the end of the process the threshold got as low as 2. This means that if the score of the candidate was above 2 it was moved to the created dictionary.

4.3.5. Results and evaluation of the method

Using all the processes combined we managed to extract an additional lexicon.

[10] Tuple is a container datatype in PYTHON which is able to store two values.

L1-L2	Number of pairs	Precision
Udm-Eng	68	79,10%
Udm-Fin	45	90,90%
Udm-Hun	40	92,30%
Udm-Rus	59	63,79%

Table 6: Size and quality of the resulting dictionaries

The validation of the lexicons were done manually by experts. Since the word pairs were not lemmatized, the translations were considered good regardless of the suffixes that may have appeared on either word, meaning, for example, if the Udmurt word was in plural, but its translation was in singular, this pair was still considered valid.

5. Final size of the seed dictionaries

As the quality of downloaded dictionaries are good, we consider the outputs of Wik2dict reliable similarly to the first lexicon extracted from the Wikipedia title corpora. The only output that needed to be evaluated was the extracted translation word pairs from the remaining parallel corpora following the first lexicon extraction. After the evaluation we merged the mentioned reliable dictionaries with the evaluated new dictionary. After this step the created big dictionary could contain duplicates for avoiding this we deleted duplicated translations.

L1-L2	D.	W2D	WT1	WT2 good	C.
Udm-Eng	0	102	880	53	1034
Udm-Fin	90	0	496	40	626
Udm-Rus	1602	276	259	37	2172
Udm-Hun	0	11	497	36	543

Table 7: Final dictionaries, where D is the size of downloaded, W2D the Wik2dict, WT1 all Wikipedia titles, WT2 the validated Wikipedia titles and C is the combined, final dictionary

Using the aforementioned methods we could create seed dictionaries for all the language pairs which will allow the extraction of more translations from comparable corpora and additionally aid to parallelize these texts and create parallel corpora for further research.

6. Summary and future plans

The research presented in this paper is part of a bigger project whose aim is to support small Finno-Ugric communities in generating online content. As the role of bilingual dictionaries and parallel corpora is huge in machine translation (Bender et al., 2003), cross-language information retrieval (Grefenstette, 1998) and also language learning (Kilgarriff et al., 2013) creating these sources is a very important step in order to support the digital presence of these small languages. Since we are processing comparable corpora seed dictionaries are essential in our work. In this paper we introduced a method which enabled us to create seed dictionaries for Udmurt–English, Udmurt–Finnish, Udmurt–Hungarian and Udmurt–Russian language pairs. In our work we used open-source software (Wik2dict, Triangulation method) and downloadable sources (Wiktionary, free bilingual dictionaries, Wikipedia) and created seed dictionaries for the mentioned language pairs which will be used for extracting parallel fragments from comparable corpora for creating parallel texts. An additional goal is to extract more translation word pairs from comparable sources in order to create large lexicons which will be uploaded to Wiktionary at the end of the project.

7. Acknowledgements

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

8. References

Ács, J. (2014). Pivot-based multilingual dictionary building using wiktionary. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Bender, O., Och, F. J., and Ney, H. (2003). Maximum entropy models for named entity recognition. In *Conference on Computational Natural Language Learning*, pages 148–152, Edmonton, Canada, May.

Fung, P. and Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 414–420, Stroudsburg, PA, USA. Association for Computational Linguistics.

Grefenstette, G. (1998). *Cross-Language Information Retrieval*. Springer US.

Hara, T., Erdmann, M., Nakayama, K., and Nishio, S. (2008). A bilingual dictionary extracted from the wikipedia link structure. *Database Systems for Advanced Applications*, pages 686–689.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2013). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Pischlöger, C. (2014). Udmurtness in web 2.0: Urban udmurts resisting language shift. In Hasselblatt, C. & Wagner-Nagy, B., editor, *Finnisch-Ugrische Mitteilungen*, volume 38, pages 143–161. Buske.

Rohit Bharadwaj, G., Tandon, N., and Varma, V. (2010). An iterative approach to extract dictionaries from wikipedia for under-resourced languages.

Winkler, E. (2001). *Udmurt*. Lincom Europa, München.