Pivot-based multilingual dictionary building using Wiktionary

Judit Ács
Research Institute for Linguistics
Hungarian Academy of Sciences
acs.judit@nytud.mta.hu

Abstract

We describe a method for expanding existing dictionaries in several languages by discovering previously non-existent links between translations. We call this method *triangulation* and we present and compare several variations of it. We assess precision manually, and recall by comparing the extracted dictionaries with independently obtained basic vocabulary sets. We featurize the translation candidates and train a maximum entropy classifier to identify correct translations in the noisy data.

Keywords: triangulation, Wiktionary, dictionary building

1. Introduction

Bilingual dictionaries are required for a variety of tasks, yet they are very hard to find, aside from a few major languages. Fully machine readable dictionaries, three star or better in the 'five star data' scheme of the W3C (Berners-Lee, 2009) are particularly rare. One of the most common ways to deal with this problem is to find a common language that has dictionaries with both languages and use it as a *pivot* language.

Constructing bilingual dictionaries by a pivot (usually English) has been tried only for a small number of scattered languages pairs – the first systematic attempt to extend the method to all pairs in a larger set is Soderland et al. (2009), discussed below. The main problem is noise due to polysemy. This was first addressed by Tanaka and Umemura (1994), who introduced a method called *Inverse Consultation* (IC) and applied it on Japanese–English–French. Here we are extending IC, which originally relied on a single pivot language, to using up to 53 pivots, exploiting the fact that pairs found via several pivot languages are more precise than those found via one.

Kaji et al. (2008) introduced *distributional similarity* (DS) as a measure for pruning noisy translations found via triangulating. Distributional similarity acquires context information about words, and compares the context vectors to compute a similarity measure. Saralegi et al. (2001) compared IC and DS and found out that DS yields good precision with considerably higher recall. In this paper we measure recall on basic vocabulary. Unfortunately, DS requires comparable corpora in all languages, which is very hard to attain for such a large number of languages.

Soderland et al. (2009) applied triangulation on a large number of languages and created PanDictionary. Unfortunately, PanDictionary has not been released to the research community. While our methods are inferior in data size, the dictionaries are available on our website.[1]

2. Wiktionary

Wiktionary is a crowdsourced dictionary aiming at eventually defining 'all words'. Similarly to Wikipedia, Wiktionary has different language editions which differ in size and detail as well. Wiktionary was created and populated by human editors (with bots introduced only recently) making machine parsing difficult. It comes in different language editions following the pattern of Wikipedia (*en.wiktionary.org, hu.wiktionary.org*). The editors are expected to follow a set of standards characterizing a Wiktionary edition. These standards may vary greatly among different editions, often making a parser for one edition unsuitable for others. A notable attempt to build a machine-readable ontology of Wiktionary is DBPedia Wiktionary, now fully supporting four language editions and two more in testing (Lehmann et al. (2013)). JWKTL (Zesch et al., 2008) is a Java-based API for accessing Wikipedia and Wiktionary, but it only supports three Wiktionary editions.

Since our method requires only parsing the translation sections in every article and ignores the rest, and we want to parse more (at least 40) editions to this level, we developed a tool for extracting translations from the so-called translation tables. The tool, *wikt2dict* currently supports 43 Wiktionary editions and is available on GitHub.[2] Wiktionary is a rapidly growing data source, therefore harvesting it again and again can yield significantly better results. For example, the Limburgish Wiktionary grew more than a 100% in less than 9 months. In this paper we present the results harvested from Wiktionary dumps made in February 2014.

We chose 53 languages to work with: Arabic, Azerbaijani, Basque, Bulgarian, Catalan, Chinese (Mandarin), Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Finnish, French, Galician, Georgian, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Ido, Indonesian, Italian, Japanese, Kazakh, Korean, Kurdish, Latin, Limburgish, Lithuanian, Macedonian, Malagasy, Malay, Norwegian, Occitan, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swahili, Swedish, Thai, Turkish, Ukrainian and Vietnamese. The extracted dictionaries are available on our website.

3. Triangulation

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. The idea is presented in

[1] http://www.nytud.hu/depts/mathling

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

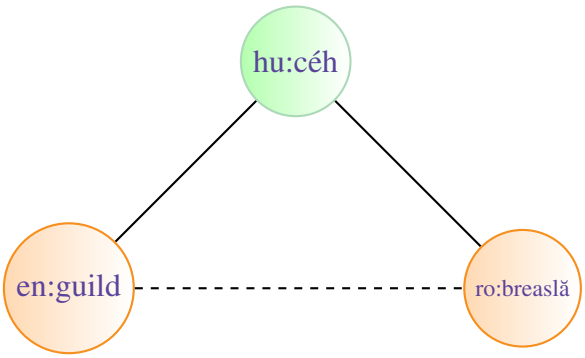

Figure 1: Straight edges represent translation pairs extracted directly from the Wiktionaries. The pair *guild–breaslă* was found via triangulating.

Figure 1, using the Hungarian word *céh* as a pivot for joining its English and Romanian translations, thus creating the previously non-existent translation pair, *guild – breaslă*.
As pointed out by Saralegi et al. (2012), the initial results obtained via triangulation are quite noisy. We distinguish four classes of translation pair candidates:

1. Correct candidates
2. Wrong candidates due to polysemy
3. Wrong candidates due to errors in the original dictionary
4. Wrong candidates due to parsing errors in the extracted dictionary

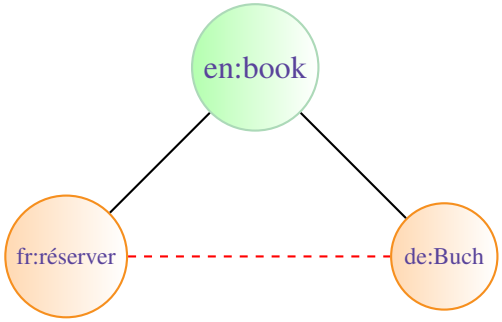

Figure 2: Error due to polysemy

The main source of errors is the polysemous nature of words. An example of this would be to join the German word *Buch* with the French word *réserver* through the polysemous English word *book* (see Figure 2).
The simplest filtering method, IC, amounts to accepting only pairs found via at least two pivots (see Figure 3). Unfortunately this aggressive filtering greatly reduces the number of triangulated pairs. It also does not solve the issue of *parallel noise* in the original data. Let's assume that we extract the English-Greek pair *dog–XXX*, where *XXX* is used as a placeholder for future translations (this is actually used in the Greek Wiktionary). If the placeholder is widely used, it is possible that we have an entirely different pair with the same Greek side, such as the German-Greek pair

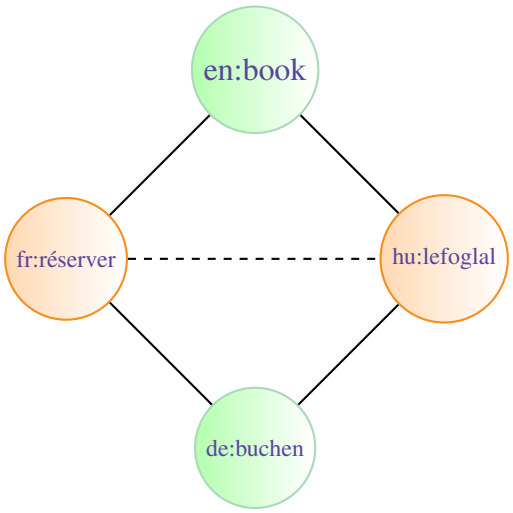

Figure 3: Translation graph with two pivots

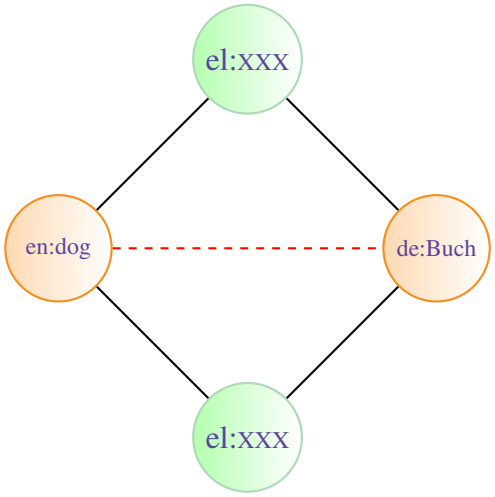

Figure 4: Error due to parallel noise

Buch–XXX. It is easy to imagine the same case for many words, which results in erroneous translation pairs found via several *XXX* pivots (see Figure 4). Although we tried to filter these placeholders, there is a high chance that some of them were overlooked by us in the 43 Wiktionary editions. To solve this issue, we examine the source Wiktionary edition of the pairs (i.e. the Wiktionary they were extracted from). All pairs are considered symmetrical but we order them alphabetically by the Wiktionary codes, thus creating a *left* and a *right* side of a triangle. In Figure 1 the pair *céh–guild* is the left pair and the pair *céh–breaslă* the right pair. We consider a candidate pair to be more reliable based on the following:

1. its left and right side were extracted from different Wiktionaries,
2. either side was found in more than one Wiktionary,
3. the pair was found via more than one pivot.

We call this group of measures *edge diversity*.
The performance of our parser and the precision and quality of a given Wiktionary edition can greatly influence the pre-

cision of the candidates based on that Wiktionary, hence the third and fourth category of erroneous candidates. Assigning a quality score manually to all 43 Wiktionaries would be next to impossible in the absence of speakers. Instead, we store the number of left edges found in each Wiktionary for each language separately, yielding 53 parameters. Although we chose 53 languages to work with, we only parse 43 corresponding Wiktionary editions and extracted pairs where both sides' languages were in the 53.
It is important to note that we did not perform any stemming or normalization on the extracted words. For now we disregard POS differences in translation candidates.

4. Applying classification on the noisy data

We trained a maximum entropy classifier to identify correct translations among the candidates. In the absence of a gold standard acquiring high quality training data is a hard problem.

4.1. Training data

We consider most Wiktionaries to be high quality, around 90% according to manual evaluation. Since the triangulation usually yields the original pairs, especially the common words, we can choose a fraction of the results that are over 90% correct. We used these pairs as positive training data, excluding the ones classified as negative training data. Out of 32.5M triangles, 1.77M was labeled as positive training sample.
As for acquiring negative training data, we collected anomalies appearing in the triangulation output.

Punctuation filtering pairs containing more than two punctuation symbols are usually due to parsing errors. The punctuation filter included all punctuation marks except: hyphen, question mark (there were idioms in the data), dot, comma, apostrophe and quote mark. Any pair that had more than one other punctuation mark was considered incorrect.

Unigram filtering we computed character unigram frequencies from the Wiktionary results and then searched for anomalies in the triangulation output. This filtering mostly yields pairs where one side is in a different language (script) than it is supposed to be.

The punctuation filter labeled 320k triangles as negative sample. Examples include:

English:	some	–	Serbian:	[[koji]]
English:	Allah#Allah	–	Spanish:	Alá

The unigram filter labeled 70k triangles as negative sample. Example:

English	almost	–	Russian:	presque

4.2. Features

We use a group of features to measure a triangles edge diversity. Let us assume that the pair *en:dog – de:Hund* has the translation graph presented in Figure 5. The features of this triangle would be the following.

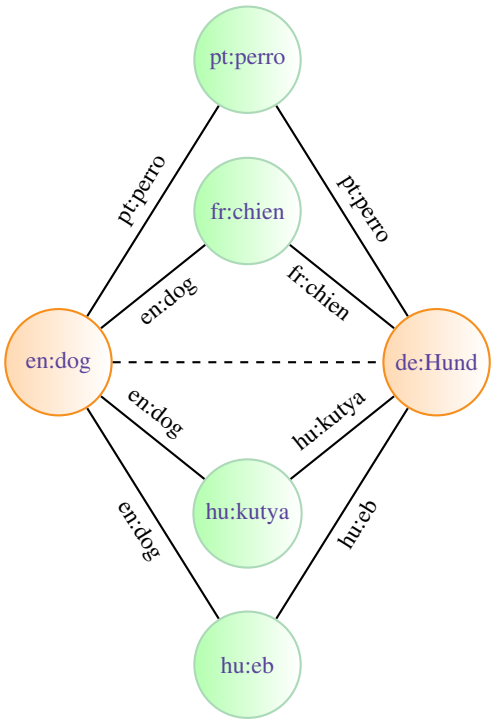

Figure 5: Translation graph with many pivots. The edge labels denote the source Wiktionary and article of the translation pair.

Pivot languages How many different pivot languages it has. In this example, this number is 3 (French, Portuguese, Hungarian).

Number of pivots Number of pivot words: 4.

Left/right languages Number of languages appearing on the left/right side: 2 left (English, Portuguese), 3 right (French, Hungarian, Portuguese).

Left/right edges Number of left/right edges: 4 left, 4 right.

Left/right disjunct edge languages The number of languages that appear among left/right edges but do not appear among right/left edges: 1 left (English), 2 right (French and Hungarian). Portuguese appears on both sides.

All features were used per-language as well, such as how many English left edges does a candidate have. We acquired more than 2000 features in this way clearly among them many are irrelevant. By discarding the features that had zero or very low weights in the maximum entropy model, we reduced this set to 200 features.

4.3. Results

We than used the model to classify the rest of the new triangles. The trained maxent model classified 59.6% as correct translation candidates.

5. Measuring relevance

The size of a dictionary does not solely depend on the number of pairs found, especially if a large ratio of the words are

Table 1: Maxent classification results with full feature set and with the reduced feature set.

Feature set	Prec	Recall	F1
Full	0.9229	0.9463	0.9345
200 features	0.9237	0.946	0.9347

rare words, therefore it is important to measure how much of the most relevant translations are extracted. We define recall as the ratio of a basic vocabulary covered by the multilingual dictionary. We have a collection of 3,500 common words forming a concept lexicon, that we used to measure recall. For the lexicographic principles used to build this lexicon see Ács, 2013.

Table 2: Recall of dictionaries for all 53 languages, its variance, most covered 40 and 10 languages

Dataset	All langs	Var	Top 40	Top 10
Wiktionary	67.4%	0.21	76.4%	93.6%
Triangles	71.7%	0.23	83.7%	93.3%
Wikt + Triangles	82.4%	0.13	88%	95.6%
Maxent correct	80.2%	0.14	86.6%	95.4%

The lexicon, *4lang* currently has bindings in 4 languages (English, Hungarian, Polish and Latin) 91% complete. We counted how many of these words are translated from either language to a given language, obtaining a ratio for each language and their average is listed in Table 2. Recall varies greatly among languages (third column).

5.1. Evaluation

Table 3: Manual evaluation results. Languages: Chinese(zh), Dutch(nl), English(en), French(fr), German(de), Hungarian(hu), Japanese(ja), Korean(ko), Portuguese(pt), Russian(ru), Slovak(sk), Slovenian(sl)

Langs	Wiktionary			Triangles		
	Ok	Small	Bad	Ok	Small	Bad
de-hu	95	3	2	50	17	33
en-hu	92	5	3	43	14	41
en-pt	77	14	9	48	12	36
fr-hu	89	5	6	38	18	39
hu-ja	91	6	3	54	9	25
hu-ko	81	15	4	47	18	24
hu-sk	89	6	5	52	1	32
hu-sl	92	3	5	52	5	43
hu-zh	86	5	8	52	6	31
nl-ru	92	0	8	43	13	43
Avg.	**88.4**	6.2	5.3	**47.9**	11.3	34.7%

We used manual spot-checking for a few language pairs. For each language pair, the annotators received 100 translation candidates parsed from Wiktionary and 100 translation candidates obtained via triangulating. The latter was sampled from the triangles not appearing in the original Wiktionary data (e.g. added translations). The annotators were asked to assign the pairs into three categories:

1. Correct translations
2. Small difference
3. Incorrect translations

Table 3 presents the results of the evaluation.

6. Conclusions

While Wiktionary is an invaluable resource with an active and growing community, its size and coverage substantially drops after the first dozen editions. We proposed a method called *triangulation* to automatically expand translations to new, often underresourced languages. Triangulation uses one or more pivot languages to find translations. As pointed out previously, triangulation yields noisy results mainly due to polysemy. Filtering results that were found via less than two pivots would reduce the number of translation candidates to less than its quarter. According to manual evaluation, almost half of the new candidates are correct.
We built an undirected graph of the translations and assigned features to the translation candidates. We trained a maximum entropy classifier, which currently yields 0.9347 F-score.

Table 4: Summary of dictionaries built

Data set	size
Wiktionary	4,092,995
Triangles	32,551,335
Triangles excl. Wiktionary	29,643,801
Triangles 2+ pivots	7,629,713
Classified as correct	19,386,537

The dictionaries built are summarized in Table 4 and are available for download.

7. Acknowledgments

I would like to thank my advisor, András Kornai for his insightful advice on theory and his constant help. I also thank my human annotators.

8. References

Ács, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August. Association for Computational Linguistics.

Berners-Lee, T. (2009). Linked data. *W3C design document http://www.w3.org/DesignIssues/LinkedData.html*.

Kaji, H., Tamamura, S., and Erdenebat, D. (2008). Automatic construction of a Japanese-Chinese dictionary via English. In *LREC*, volume 2008, pages 699–706.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al. (2013). DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal (under review, 2013)*.

Saralegi, X., Manterola, I., and Vicente, I. S. (2011). Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics.

Saralegi, X., Manterola, I., and Vicente, I. S. (2012). Building a Basque-Chinese dictionary by using English as pivot. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., Bilmes, J., et al. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 262–270. Association for Computational Linguistics.

Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646–1652.