Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 32–41
Marseille, 20 June 2022

Compiling a Highly Accurate Bilingual Lexicon by Combining Different Approaches

Steinþór Steingrímsson[1,2], Luke O'Brien[1], Finnur Ingimundarson[2], Hrafn Loftsson[1], and Andy Way[3]
[1]Department of Computer Science, Reykjavik University, Iceland
[2]The Árni Magnússon Institute for Icelandic Studies, Iceland
[3]ADAPT Centre, School of Computing, Dublin City University, Ireland
steinthor18@ru.is, luke20@ru.is, fai@hi.is
hrafn@ru.is, andy.way@adaptcentre.ie

Abstract

Bilingual lexicons can be generated automatically using a wide variety of approaches. We perform a rigorous manual evaluation of four different methods: word alignments on different types of bilingual data, pivoting, machine translation and cross-lingual word embeddings. We investigate how the different setups perform using publicly available data for the English-Icelandic language pair, doing separate evaluations for each method, dataset and confidence class where it can be calculated. The results are validated by human experts, working with a random sample from all our experiments. By combining the most promising approaches and data sets, using confidence scores calculated from the data and the results of manually evaluating samples from our manual evaluation as indicators, we are able to induce lists of translations with a very high acceptance rate. We show how multiple different combinations generate lists with well over 90% acceptance rate, substantially exceeding the results for each individual approach, while still generating reasonably large candidate lists. All manually evaluated equivalence pairs are published in a new lexicon of over 232,000 pairs under an open license.

Keywords: Bilingual Lexicon Induction, Dictionary, Bilingual Corpora, Pivoting, Machine Translation

1. Introduction

Bilingual lexicons are useful for an array of different tasks. First, they can be used for harvesting bitexts from multilingual websites or corpora. For example, *Bicleaner* (Ramírez-Sánchez et al., 2020), a popular tool used for that task, requires a probabilistic lexicon for training. Second, they can be used for cross-language information retrieval (see e.g. Bonab et al. (2020), Steingrímsson et al. (2021b)). Third, they can be exploited in machine translation (MT), e.g. as an additional scoring component (Arthur et al., 2016), for initializing unsupervised MT (Artetxe et al., 2018b; Lample et al., 2018b; Duan et al., 2020), for substituting words in source sentences in pre-training (Lin et al., 2020), for annotating source sentences with possible translations from lexicons (Dinu et al., 2019; Niehues, 2021), or for inputting prior knowledge into the self-attention module of the encoder (Chen et al., 2021).

Among the different approaches to the bilingual lexicon induction (BLI) task are extracting bilingual lexicons from parallel corpora using word alignments (Mihalcea and Pedersen, 2003; Och and Ney, 2003), mining comparable corpora, commonly using cross-lingual word embeddings (Rapp et al., 2020), and pivoting through intermediary languages in available dictionaries (Gracia et al., 2019). The different approaches have contrasting limitations. Pivoting is limited by the availability of dictionaries that connect the source and target languages, and while bitext mining can produce very many candidates it is prone to giving noisy results, both when using word embeddings and candidate pair extraction using word alignments.

We present a methodology to build a moderately large lexicon for the English-Icelandic language pair, a language pair that has basic resources available allowing us to approach the problem from different angles. Previously, only the *Wiktionary*[1] and *Apertium* (Forcada et al., 2011) dictionaries were publicly available for this language pair, containing approximately 18,000 and 23,000 word pairs, respectively. While a wide variety of approaches to automatic bilingual lexicon induction

[1]https://www.wiktionary.org/

Translation Pair		Probabilities	
Icelandic	**English**	**is→en**	**en→is**
ananas	pineapple	1.0	0.82
ananasjurt	pineapple	1.0	0.15
granaldin	pineapple	1.0	0.03
regnhlíf	umbrella	0.70	0.73
regnhlíf	brolly	0.30	1.0
hlífð	umbrella	0.02	0.01
sólhlíf	umbrella	0.31	0.26
sólhlíf	parasol	0.48	1.0
sólhlíf	sunshade	0.21	0.46

Table 1: Example of translation pairs with probability scores from the lexicon resulting from the project. If there is only one translation for a word, the probability is 1.0, if there are many translations the probabilities sum to 1.0, as for the English word *pineapple* or the Icelandic word *regnhlíf*.

(BLI) have been shown to be effective, we experiment extensively with four different methods and perform rigorous manual evaluation with human experts validating a random sample of candidate pair lists from all our experiments. As our goal is to find a quick and efficient way to compile a glossary, we also assess the effectiveness of combining the most promising strategies in order to compile a manually approved lexicon as fast as possible.

Our work results in a manually verified lexicon of over 232,000 pairs, with a probability score attached to each pair for both translation directions. The probability scores are an attempt to order the translations for a given source word from most common to least common. The probability is calculated by tallying the number of times the pair was suggested by our methods and comparing that to how often other translations for the same word were suggested. An example of the lexicon format is shown in Table 1.

Our main contributions are:

- doing rigorous manually verified experiments on four different BLI approaches: 1) using cross-lingual word embeddings trained on comparable corpora, 2) pivoting through available dictionaries, 3) mining bitexts using word alignments, and 4) translating using available MT systems.
- showing that combining outputs of diverse approaches can greatly improve the rate of acceptable candidate pairs, while still retaining a large portion of the acceptable candidate pairs, if the combined approaches are carefully selected.

Furthermore, we publish a new, manually verified English–Icelandic lexicon (Steingrímsson et al., 2021), substantially larger than what was previously available, with probability scores for each translation pair. The lexicon and its availability is described in Section 5.

2. Related Work

A variety of approaches to automatically compile bilingual lexicons have been shown to be successful. Bilingual lexicons have been mined from parallel corpora using word alignments (Mihalcea and Pedersen, 2003; Vulić and Moens, 2012), and from comparable corpora with a variety of approaches, most commonly by learning cross-lingual word embeddings (Lample et al., 2018a; Rapp et al., 2020). Artetxe et al. (2019) use an unsupervised MT system to create a synthetic corpus which they extract the lexicon from.

Comparable corpora can also be exploited by identifying word pairs in the corpus using word alignments. For this purpose, sentence pairs first have to be extracted from the comparable corpora. This has been carried out using various approaches, e.g. using bilingual word embeddings to help calculate a BLEU score (Papineni et al., 2002) to estimate semantic similarity (Bouamor and Sajjad, 2018), using a BERT model (Devlin et al., 2019) to generate a similarity score based on contextualized sentence embeddings (Feng et al., 2020), or using cross-language information retrieval to limit the search space and a classifier, based on a word alignment score and a contextualized embedding score, to select the sentence pairs (Steingrímsson et al., 2021b).

Shi et al. (2021) show that lexicon induction performance correlates with bitext quality, although they are still able to induce a reasonably good bilingual lexicon from their lowest quality bitexts. They also observe that a better word aligner usually leads to a better induced lexicon.

Pivoting through existing dictionaries to infer translations between two languages using an intermediary language, e.g. using L1→L2 and L2→L3 dictionaries to infer translations between L1→L3, can produce a useful lexicon if measures are taken to filter the output of such an approach, as often a monosemous lexical item in one language can be polysemous in its corresponding translation into another language (Ordan et al., 2017). Tanaka and Umemura (1994) consult an inverse dictionary after pivoting and select equivalences based on common elements when source and target language words are translated into the intermediary language.

Mausam et al. (2009) tackle the problem by using multiple Wiktionary dictionaries to build graphs, identify sense cliques and try to identify ambiguity sets to be able to disambiguate between senses. The problem has also been approached by using MT systems to translate the words between languages (Arcan et al., 2019). The highest scoring system in the *2021 shared task for Translation Inference Across Dictionaries* (TIAD 2021) used a combination of pivoting and bitext extraction (Steingrímsson et al., 2021c).

3. Experimental Settings

We designed a number of experiments to explore three research questions:

1. How accurately can we produce equivalence pairs using four different methods: using cross-lingual word embeddings trained on comparable corpora, pivoting through available dictionaries, mining bitexts using word alignments, and translating using available MT systems?
2. To what extent does the frequency of words affect the results in corpus-based approaches?
3. How can we best combine the different approaches to increase accuracy while not reducing the size of the resulting lexicon too much?

Each experiment resulted in a list of translation candidates from which we extracted a random sample for evaluation. The evaluation was carried out by first comparing the list against the following manually curated Icelandic-English/English-Icelandic dictionaries and word lists: English-Icelandic Wiktionary and

Apertium dictionaries, titles of common pages in the Icelandic and English Wikipedia, the Icelandic Term Bank[2], and the Terminology Database of the Ministry of Foreign Affairs[3].

If the candidate pairs were found in these data sets they were accepted, otherwise a human annotator manually evaluated them and categorized into the following categories: *acceptable*, *unacceptable*, *rectifiable/partial*. Four annotators worked on the project, all Icelandic native speakers, educated in linguistics and with excellent knowledge of English. The criteria given to the annotators was that if the word in either language could be translated to the other word, in any environment the annotators could think of, the pair should be categorized as *acceptable*. The *rectifiable/partial* category was used when there was a minor error in one of the words, e.g. a spelling error, lemmatization error or a typo, or when a word in one language had to be translated into a multiword unit, and the translation given only has a part of that unit. Words that fell into neither of these categories were categorized as *unacceptable*.

3.1. Extracting Word Pairs from Bilingual Corpora

We extracted word alignments as accurately as possible using the CombAlign tool (Steingrímsson et al., 2021a), which uses a voting system employing multiple different word aligners, Giza++ (Och and Ney, 2003), fast_align (Dyer et al., 2013), eflomal (Östling and Tiedemann, 2016), two SimAlign (Masoud et al., 2020) models and AWESoME (Dou and Neubig, 2021). If four models agreed on an alignment, it was accepted. In order to increase alignment accuracy and to reduce noise, we lemmatized all the data and collected lemma pairs from the lemmatized sentence pairs. We used SpaCy[4] for lemmatizing English, and after PoS-tagging the Icelandic texts using ABLTagger (Steingrímsson et al., 2019), we lemmatized them using Nefnir (Ingólfsdóttir et al., 2019), which is trained on the Database of Icelandic Morphology (DIM) (Bjarnadóttir et al., 2019). We then calculated a confidence score for each aligned word pair $\langle s,t \rangle$ using Equation (1), as employed by Steingrímsson et al. (2021c):

$$\rho\,(s,t) = \frac{match\,(s,t)}{coc\,(s,t) + \lambda} \tag{1}$$

In Equation (1), $match\,(s,t)$ is the one-to-one matching count, i.e. how often the words are aligned in the corpus, and $coc\,(s,t)$ is the number of one-to-one co-occurrences, i.e. count of $\langle s,t \rangle$ appearing in a sentence pair in the corpus. λ is a non-negative smoothing term. The equation was proposed by Shi et al. (2021). While they set the smoothing variable λ to 20, here it is set to $\log_2 s$ where s is the number of sentence pairs in the corpus under consideration. This way the score is more comparable between corpora of different sizes.

The score is used as a filtering mechanism, by finding cutoff thresholds for six different bilingual corpora of three types: a parallel corpus, comparable corpora, and synthetic corpora. We describe the corpora in the following subsections.

3.1.1. Parallel Corpus

We used the English-Icelandic ParIce corpus (Barkarson and Steingrímsson, 2019), containing 3.6 million sentence pairs, 80% of which are sourced from official EEA documents or movie subtitles.

3.1.2. Comparable Corpora

ParaCrawl (Bañón et al., 2020) is a large project to create parallel corpora by crawling the web. They publish document pairs and sentence pairs extracted from the documents, using various tools in their pipeline, including Bitextor[5] for document alignment, hunalign (Varga et al., 2005), Vecalign (Thompson and Koehn, 2019) and Bleualign (Sennrich and Volk, 2011) for sentence alignment and Bicleaner (Ramírez-Sánchez et al., 2020) for filtering. ParaCrawl has published data for more than 40 languages, low resource and high resource, most of which are paired with English. WikiMatrix (Schwenk et al., 2021) is another publicly available set of sentence pairs, mined from Wikipedia using an approach based on massively multilingual sentence embeddings (Artetxe and Schwenk, 2019b) and a margin criterion (Artetxe and Schwenk, 2019a). WikiMatrix was published for 85 different languages and 1620 language pairs.

The methods applied in these two projects could be applied to most languages that have available monolingual data, comparable to data in another language, although the size of the available monolingual data limits the size of the resulting datasets. As these two publicly available datasets, WikiMatrix and ParaCrawl, have English–Icelandic sentence pairs collected from comparable corpora, we opt to use them instead of creating our own. WikiMatrix has 86K sentence pairs, but ParaCrawl is considerably larger and has 2.4M sentence pairs for version 7.1 and 5.7M sentence pairs for version 8, the two versions we experiment with.

3.1.3. Synthetic Corpora

For synthetic corpora, we used the same methodology as before, i.e. extract word pairs from aligned sentence pairs using word alignment tools. Our synthetic corpora are two back-translated corpora consisting of source sentences and back-translations generated using a transformer network (Símonarson et al., 2020). 44.7M English source sentences were retrieved from Wikipedia, Newscrawl and Europarl, while the 31.3M Icelandic sentences were sourced from the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018).

[2] https://idordabanki.arnastofnun.is/
[3] https://hugtakasafn.utn.stjr.is/
[4] https://spacy.io
[5] https://github.com/bitextor/bitextor

	Sample from 10,000 most frequent				Sample from 100,000 most frequent			
Corpus	Accept	Unacc.	Partial	Accuracy	Accept	Unacc.	Partial	Accuracy
ParIce	202	**170**	128	0.40	178	**214**	108	0.36
Paracrawl 7.1	**279**	190	31	**0.56**	**212**	228	60	**0.42**
Paracrawl 8	143	339	18	0.29	134	334	32	0.27
WikiMatrix	232	220	48	0.46				
Synthetic is-en	205	258	37	0.41	167	225	108	0.33
Synthetic en-is	272	195	33	0.54	202	227	71	0.40

Table 2: Accuracy of candidate pairs sampled from two different frequency classes in six bilingual corpora. 500 pairs were randomly selected from each frequency class. The table gives numbers for equivalents (accepted), non-equivalents (unaccepted) and partial equivalents in the manually evaluated data. Accuracy is the acceptance ratio, i.e. the number of accepted pairs divided by the total number of pairs.

Synthetic corpora like these can be created for any language pair if an MT model is available, or even by building and using an unsupervised MT model, see e.g. Artetxe et al. (2019).

3.2. Pivoting

We used dictionaries with Icelandic as a source language and pivoted through an intermediate language into English. For collecting translations from Icelandic into intermediary languages we used the ISLEX (Úlfarsdóttir, 2014) and LEXIA dictionaries (Icelandic-Danish / Swedish / Norwegian / Finnish / French) and dict.cc[6] for Icelandic-German. For collecting translations from the intermediary languages into English we used Apertium (Forcada et al., 2011) (Finnish / French / Norwegian / Swedish-English) and dict.cc (German/Finnish/Norwegian/Swedish/French/English). For each Icelandic source word, we collected all possible translations in the intermediary languages and, for each of the intermediary translations, we collected all English translations.

3.3. Machine Translation

Our most simple approach was translating words into English using four available MT models: Google Translate[7], Microsoft Translator[8], OPUS-MT (Tiedemann and Thottingal, 2020) and M2M100 M2M (Fan et al., 2020). First, we translated the Icelandic source words of the ISLEX/LEXIA dictionaries into English, thereby creating a candidate list. Second, we also translated into English the target language equivalents in these dictionaries, Danish, Swedish, Norwegian, Finnish and French, and then paired the source Icelandic word to the translation of the target words.

While this method is simple and accessible for many languages, using existing commercial MT services can make it difficult to replicate the results of the experiments. As one of our goals is to compile a lexicon as fast as possible we decided to use these services anyway, to see if they could be useful for this purpose.

3.4. Cross-lingual Word Embeddings

Icelandic news texts collected from the IGC and English news texts collected from Newscrawl[9] were used to train two word2vec models (Mikolov et al., 2013), one for English and the other for Icelandic. VecMap (Artetxe et al., 2018a) was then used to build cross-lingual word embeddings by mapping the models to a common vector space.

Three candidate lists were generated. One is based on the most frequent English and Icelandic words in their respective corpus, with the nearest neighbour (NN) to each word in terms of cosine distance. The other two lists contain, on the one hand, words selected based on the lowest cosine distance to a word in the other language and, on the other hand, based on the highest Cross-domain Similarity Local Scaling (CSLS) method, which alleviates the problem of hubs of incorrect translations polluting the vector space (Dinu and Baroni, 2015).

This unsupervised approach is available for all languages if monolingual corpora are available.

4. Evaluation

We performed a thorough evaluation of the different methods, comparing the word pairs against available manually compiled datasets and by performing a manual evaluation as described in Section 3.

For the corpus-based approaches we created classes that could be expected to correlate with the likelihood of the candidate pairs being equivalents. The classes were either based on frequency or similarity as estimated by cross-lingual word embedding models. We tested each of these classes manually. Candidates generated by pivoting and MT were evaluated on a random sample of 500 pairs from each method and class of data evaluated.

[6] https://www.dict.cc/

[7] https://translate.google.com/, accessed in May 2021

[8] https://translator.microsoft.com/, accessed in May 2021

[9] https://data.statmt.org/news-crawl/en/

4.1. Bilingual Corpora

We extracted word pairs from six different bilingual corpora, as shown in Table 2, only considering pairs that appear more than five times in each corpus. We created two frequency classes, i.e. for the 10,000 and 100,000 most frequent words in the corpora, respectively. Frequency was calculated as an average of the total count of the Icelandic words in the Icelandic part of the corpus and the English words in the English part. We randomly sampled 500 pairs from both frequency classes in each corpus. For WikiMatrix we did not take a sample from the 100,000 most frequent, as the corpus was too small for us to collect that many samples.

Table 2 shows that the highest accuracy was achieved on the ParaCrawl 7.1 corpus. While it could have been expected to attain the highest scores from ParIce, the parallel corpus, due to it being compiled from known parallel documents, we can see that it has a very high percentage of pairs categorized as partially correct. This may indicate that the texts in ParIce have a higher ratio of multiword units and that if we would extract not only single words from the bilingual corpora, the accuracy might change for this corpus. There is a noticable difference between ParaCrawl 7.1 and 8. As version 8 is more than twice the size of version 7.1, this may indicate that the additional sentence pairs are of lower quality, although this would have to be investigated further.

We used the confidence score (see Equation 1), calculated for each of the word pair candidates, to create ten confidence bands, with the lowest having a score of less than 0.1 and the highest with a score higher than 0.9. We evaluated 250 pairs in each band for each of the corpora. Figure 1 shows that the confidence scores do not represent the same level of accuracy for all corpora. While more than half of the pairs with a confidence score higher than 0.4 were accepted for ParIce, WikiMatrix and ParaCrawl 8, the confidence score for

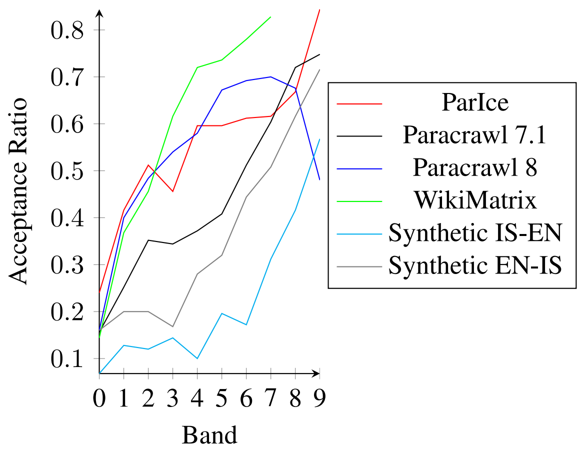

Figure 1: Bilingual corpora. Manually evaluated acceptability of candidate pairs at different bands of confidence, as automatically assessed by our confidence score.

	Apertium		dict.cc	
	acc. ratio	no. pairs	acc. ratio	no. pairs
se	0.64	34,915	0.76	26,622
fi	0.43	214,659	0.75	19,304
no	0.53	15,261	0.74	31,213
fr	0.63	20,865	0.64	39,590
de			0.54	137,970

Table 3: Pivoting. Acceptance ratio and number of pairs yielded by each pivoting path from Icelandic to English connected by an intermediary language in ISLEX and the Apertium and dict.cc dictionaries.

	acc. ratio	no. pairs
se (A+D)	0.85	10,805
fi (A+D)	0.89	12,969
fr (A+D)	0.83	11,012
fi (A) + de (D)	0.91	17,681
fi (A) + se (D)	0.93	13,962
fi (A) + no (D)	0.93	14,750
fi (A) + fr (D)	0.94	13,743

Table 4: Pivoting combinations. Acceptance ratio and number of candidate pairs yielded with different combinations of two pivoting paths. A=Apertium, D=dict.cc.

the synthetic corpora had to be at least 0.7 in order to obtain the same results.

4.2. Pivoting

We compiled candidate lists for each of the intermediary languages, using both Apertium and dict.cc for obtaining English translations from the intermediary language words. The dictionaries vary in size and that is reflected in the candidate lists. For each list, 500 randomly selected candidate pairs were evaluated and the acceptance ratio calculated. Results are shown in Table 3. The smaller lists tend to have higher acceptance ratios. This may be because the smaller lists more often only have the most common translation for any given word, and when multiple senses are given for a word, some of these are likely to have different translations in a third language (see e.g. Tanaka and Umemura (1994)).

As seen in Table 3, up to 76% of the translations are acceptable, depending on the language and dictionary used. In order to increase the accuracy even further, we can require the pairs to be suggested by two or more pivoting paths. We combined two pivoting approaches by selecting an intersection of the result of each. This substantially raised the accuracy, especially when two different language pairs and dictionaries are combined. Table 4 shows the accuracy and number of candidate pairs for all combinations that yield more than 10,000 pairs.

	Opus	M2M	Google	MS	no. pairs
is			0.59	0.60	53,151
da	0.52		0.59	0.63	80,074
sv	0.56	0.32	0.65	0.65	69,884
fi	0.53	0.27	0.66	0.62	62,876
no			0.59	0.61	66,129
fr	0.56	0.35	0.67	0.71	48,533

Table 5: Machine translation. Acceptance ratio in 500 randomly selected candidate pairs for each language and system. For all languages except Icelandic, we pivoted through intermediary languages using dictionaries and translated the intermediary languages to English using MT.

	acc. rate (%)	no. pairs
se+fr	97.7	11,274
se+fi	97.1	14,931
se+de+no	95.8	13,151
fr+fi	97.7	9,914

Table 6: Machine translation combinations. Acceptance rate and number of pairs yielded by an intersection of MT outputs. All combinations listed are an intersection of both Google Translate and Microsoft Translate for each of the languages listed.

4.3. Machine Translation

As described in Section 3.3, we employed MT using two approaches. The more straightforward one was to translate the Icelandic source words from the ISLEX dictionary into English using two different MT engines. The other one was translating the target language words in the ISLEX dictionary into English using up to four different MT engines, and then replacing the ISLEX target word with the Icelandic source word to create an Icelandic–English candidate list. All the systems except M2M resulted in over 50% acceptable translations for all languages. The pivoting process yielded a different number of words to translate, depending on the dictionary, ranging from 48,000-80,000 words. For most languages, Microsoft Translator gave the best results, as shown in Table 5. By combining results from multiple systems and using multiple intermediary languages, accuracy can be raised substantially. We tried taking an intersection of candidate pairs produced for all six languages using both Microsoft Translator and Google Translate. When all these twelve outputs were in agreement, the human annotators agreed with the outputs 99.6% of the time, but the number of candidate pairs yielded went down to only only 2,358. By combining fewer outputs, a higher number of candidates is produced while the acceptance rate is still very high. For the experiments yielding such high accuracy we raised the number of pairs to evaluate to 2,000 for each combination. Table 5 shows the highest resulting

Lang. Direction	Retrieval method	Classification High	Medium	Low
en-is	NN	0.39	0.20	0.03
	CSLS	0.59	0.38	0.14
	freq.	0.71	0.50	0.14
is-en	NN	0.48	0.26	0.15
	CSLS	0.63	0.40	0.19
	freq.	0.67	0.44	0.22

Table 7: Cross-lingual word embeddings. Acceptance ratio for candidate lists in different similarity or frequency classes, for each of the methods employed.

combinations of 2-3 languages yielding close to 10,000 candidate pairs or more.

While Table 6 shows that combining the results of different MT systems can yield a highly acceptable list of candidate pairs, a downside to the MT approach is that each system only outputs one equivalence suggestion for each source word, which when correct is usually a very common translation. Accordingly, this does not seem to be an effective way to obtain translations for low-frequency senses or rare words.

4.4. Cross-Lingual Word Embeddings

Three approaches are used to extract word pairs from our cross-lingual word embeddings, as described in Section 3.4. For each of these approaches we divide the results into three classes: *High*, for the top 2,000 pairs, *Medium*, for the next 8,000 pairs, and *Low* for the next 90,000 pairs. The pairs are ordered by similarity in terms of NN or CSLS, or by frequency in the corpora used to train the embedding models. Table 7 shows that while we obtain decent scores for the most frequent words in the corpora and most similar word pairs according to the model, the scores fall sharply as word frequency and similarity decrease.

4.5. Combining different approaches

Based on the results presented above, we created two lists. One contains all candidate pairs obtained through pivoting or MT, being in classes where acceptance rate of candidate pairs is over 50%. The other list was created from all six bilingual corpora, but only from confidence bands with over 50% acceptance rate (see Figure 1). Taking an intersection of these resulted in a list of 29,609 candidates, of which 93.2% were accepted after manual evaluation. Detailed results are shown in Table 8.

Furthermore, if the confidence bands are ignored and the second list has all pairs from the six bilingual corpora, the intersection of the two lists results in a list of 57,818 candidates, of which 84.1% were accepted.

5. Availability

We publish all word pairs accepted in the evaluation process. The final dataset, resulting from evaluation

		Confidence Scores with over 50% Acceptability			Also in Pivoting/MT Candidate Lists		
Corpus	Total Pairs	Acceptance Ratio (%)	Number of Pairs	Estimated Correct	Acceptance Ratio (%)	Number of Pairs	Estimated Correct
ParIce	346,723	51.6	45,646	23,553	90.4	3,713	3,356
Paracrawl 7.1	107,989	59.6	70,281	41,887	95.8	18,836	18,045
Paracrawl 8	342,444	62.6	93,850	58,750	96.2	16,522	15,894
WikiMatrix	15,781	77.2	6,944	5,360	97.4	3,343	3,256
Synthetic is–en	191,934	67.2	13,215	8,880	97.3	4,986	4,851
Synthetic en–is	229,661	60.2	132,381	79,693	94.4	19,423	18,335
Total	938.354	46.6	249,872	116,440	93.2	29,609	27,595

Table 8: Combining different methods. Evaluation of the combination of different approaches, using bitexts on the one hand and pivoting/MT on the other.

of all the experiments carried out during this research, contains 232,950 pairs, with 105,442 different Icelandic lexical items, of which 84,812 are single words and 20,630 multiword units, and 116,744 different English items, of which 45,147 are unique English words and 71,597 multiword units. The published dataset includes the probability scores described in Section 1 and word class information, in cases where that could be retrieved automatically from Wiktionary or the DIM (Bjarnadóttir et al., 2019). The published dataset also contains information on which methods produced the pairs included in the dataset and how often. The data is available for download at a CLARIN repository[10].

6. Conclusion and Future Work

We have compared four different approaches to automatically compile an English-Icelandic bilingual lexicon. We have shown that by using a combination of bilingual corpora, pivoting and MT approaches, we can build a highly accurate candidate list for lexicon translations between languages. Our combined approach yields a candidate list of almost 30,000 pairs of which 93.2% are acceptable translations. Using individual approaches yields more data, but with less accuracy. Very high accuracy can be achieved using individual approaches by combining the resulting candidate pairs from different data sets, while still yielding a decently sized candidate lists, as shown in Table 4 for pivoting combinations and Table 6 for MT combinations. While using an unsupervised approach such as cross-lingual word embeddings did not result in many useful candidate pairs, extracting candidate pairs from back-translated data using word alignments gives promising results for our language pair.

The results indicate that there are multiple feasible ways to extend the lexicon. Adding more dictionaries for pivoting and by pivoting through more than one intermediary language would produce more candidates. To limit the noise as much as possible we could use a variant of inverse consultation (Tanaka and Umemura, 1994).

While pivoting and MT can yield multiword units, our methods for extracting from bilingual corpora only identifies single word units. The high number of partial equivalents in our parallel corpus is an indication that there is still room for improvement in extracting equivalence pairs from bitexts with the help of word alignments if we have a mechanism for retrieving not only single words but multiword units. We want to explore that further using a similar hybrid approach as Semmar (2018). We are also interested in extracting candidate pairs from other bilingual corpora, e.g. version 9 of ParaCrawl, and creating additional synthetic corpora.

Furthermore, the new compiled lexicon can be a valuable asset to better align and filter parallel corpora or for better extracting parallel sentences from comparable corpora. It could be worthwhile to use the dataset created in this project to explore an iterative approach, where the new English-Icelandic lexicon is used to refine the parallel and comparable corpora used, and then to repeat this experiment and investigate if it then yields more candidates or more accurate candidate lists.

7. Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019–2023. The programme, which is managed and coordinated by Almannarómur[11], is funded by the Icelandic Ministry of Education, Science and Culture and by the ADAPT Centre for Digital Content Technology which is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. We would also like to thank all the annotators that helped evaluate the data: Árni Davíð Magnússon, Þórdís Dröfn Andrésdóttir and Inga Guðrún Eiríksdóttir.

[10] `https://repository.clarin.is/repository/xmlui/handle/20.500.12537/144`

[11] `https://almannaromur.is/`

8. Bibliographical References

Arcan, M., Torregrosa, D., Ahmadi, S., and McCrae, J. P. (2019). Inferring Translation Candidates for Multilingual Dictionary Generation with Multi-Way Neural Machine Translation. In *Proc. of TIAD-2019 Shared Task - Translation Inference Across Dictionaries co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019)*, pages 13–23, Leipzig, Germany.

Artetxe, M. and Schwenk, H. (2019a). Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy.

Artetxe, M. and Schwenk, H. (2019b). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, March.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018b). Unsupervised Neural Machine Translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Artetxe, M., Labaka, G., and Agirre, E. (2019). Bilingual Lexicon Induction through Unsupervised Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy.

Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating Discrete Translation Lexicons into Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online.

Barkarson, S. and Steingrímsson, S. (2019). Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland.

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. (2019). DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland.

Bonab, H., Sarwar, S. M., and Allan, J. (2020). Training Effective Neural CLIR by Bridging the Translation Gap. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 9–18.

Bouamor, H. and Sajjad, H. (2018). H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, pages 43–47, Miyazaki, Japan.

Chen, K., Wang, R., Utiyama, M., and Sumita, E. (2021). Integrating Prior Translation Knowledge into Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Dinu, G. and Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training Neural Machine Translation to Apply Terminology Constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy.

Dou, Z.-Y. and Neubig, G. (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online.

Duan, X., Ji, B., Jia, H., Tan, M., Zhang, M., Chen, B., Luo, W., and Zhang, Y. (2020). Bilingual Dictionary Based Neural Machine Translation without Using Parallel Sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, Online.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Çelebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond English-Centric Multilingual Machine Translation. *ArXiv*, abs/2010.11125.

Feng, F., Yang, Y.-F., Cer, D. M., Arivazhagan, N., and Wang, W. (2020). Language-agnostic BERT Sentence Embedding. *ArXiv*, abs/2007.01852.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Rojas, S. O., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Gracia, J., Kabashi, B., Kernerman, I., Lanau-Coronas, M., and Lonke, D. (2019). Results of the Translation Inference Across Dictionaries 2019 Shared Task. In *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*, pages 1–12.

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. (2019). Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 310–315, Turku, Finland.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. (2018a). Word translation without parallel data. In *International Conference on Learning Representations*.

Lample, G., Denoyer, L., and Ranzato, M. (2018b). Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*.

Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., and Li, L. (2020). Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online.

Masoud, J. S., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online.

Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., and Bilmes, J. (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*, pages 262–270, Suntec, Singapore.

Mihalcea, R. and Pedersen, T. (2003). An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Canada.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona.

Niehues, J. (2021). Continuous Learning in Neural Machine Translation using Bilingual Dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Ordan, N., Gracia, J., and Kernerman, I. (2017). Autogenerating Bilingual Dictionaries. In *Proceedings of fifth biennial conference on electronic lexicography (eLex 2017)*, Leiden, Netherlands.

Östling, R. and Tiedemann, J. (2016). Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Ramírez-Sánchez, G., Zaragoza-Bernabeu, J., Bañón, M., and Ortiz-Rojas, S. (2020). Bifixer and Bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal.

Rapp, R., Zweigenbaum, P., and Sharoff, S. (2020). Overview of the Fourth BUCC Shared Task: Bilingual Dictionary Induction from Comparable Corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 6–13, Marseille, France.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online.

Semmar, N. (2018). A Hybrid Approach for Automatic Extraction of Bilingual Multiword Expressions from Parallel Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 311–318, Miyazaki, Japan.

Sennrich, R. and Volk, M. (2011). Iterative, MT-based Sentence Alignment of Parallel Texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*, pages 175–182, Riga, Latvia.

Shi, H., Zettlemoyer, L., and Wang, S. I. (2021). Bilingual Lexicon Induction via Unsupervised Bitext Construction and Word Alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Interna-*

tional Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 813–826, Online.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4361–4366, Miyazaki, Japan.

Steingrímsson, S., Kárason, Ö., and Loftsson, H. (2019). Augmenting a BiLSTM Tagger with a Morphological Lexicon and a Lexical Category Identification Step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1162–1169, Varna, Bulgaria.

Steingrímsson, S., Loftsson, H., and Way, A. (2021a). CombAlign: a Tool for Obtaining High-Quality Word Alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Online.

Steingrímsson, S., Lohar, P., Loftsson, H., and Way, A. (2021b). Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17, Online.

Steingrímsson, S., Loftsson, H., and Way, A. (2021c). Pivotalign: Leveraging High-Precision Word Alignments for Bilingual Dictionary Inference. In *Proc. of LDK 2021 workshops and tutorials [IN PRESS]*. CEUR-WS.

Tanaka, K. and Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, pages 297–303, Kyoto, Japan.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved Sentence Alignment in Linear Time and Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal.

Úlfarsdóttir, Þ. (2014). ISLEX — a Multilingual Web Dictionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2820–2825, Reykjavik, Iceland.

Varga, D., Halaácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596, Borovets, Bulgaria.

Vulić, I. and Moens, M.-F. (2012). Sub-corpora Sampling with an Application to Bilingual Lexicon Extraction. In *Proceedings of COLING 2012*, pages 2721–2738, Mumbai, India.

9. Language Resource References

Símonarson, Haukur Barri and Snæbjarnarson, Vésteinn and Þorsteinsson, Vilhjálmur. (2020). *En-Is Synthetic Parallel Corpus*. CLARIN-IS, http://hdl.handle.net/20.500.12537/70.

Steingrímsson, Steinþór and Obrien, Luke James and Ingimundarson, Finnur Ágúst and Magnússon, Árni Davíð and Andrésdóttir, Þórdís Dröfn and Eiríksdóttir, Inga Guðrún. (2021). *English-Icelandic/Icelandic-English glossary 21.09*. CLARIN-IS, http://hdl.handle.net/20.500.12537/144.