PivotAlign: Leveraging High-Precision Word Alignments for Bilingual Dictionary Inference

Steinþór Steingrímsson[1], Hrafn Loftsson[1], and Andy Way[2]

[1] Department of Computer Science, Reykjavik University, Iceland
steinthor18@ru.is, hrafn@ru.is
[2] ADAPT Centre, School of Computing, Dublin City University, Ireland
andy.way@adaptcentre.ie

Abstract. This paper describes our contribution to the TIAD 2021 shared task for Translation Inference Across Dictionaries. Our system, PivotAlign, approaches the problem from two directions. First, we collect translation candidates by pivoting through intermediary dictionaries, made available by the task organizers. Second, we decide which candidates to keep by applying scores to the candidate list, obtained by running an ensemble of word alignment tools on parallel corpora and comparing frequency of alignments to frequency of word co-occurrence in the parallel texts. Our approach outperforms all other participating systems with respect to F1 measure and recall, as well as having a very competitive precision score, showing the usefulness of a scoring mechanism based on highly accurate word alignments for this kind of task.

Keywords: Translation inference · Word alignment · Dictionary building.

1 Introduction

The growing availability of open, high-quality lexical resources and semantic data, monolingual as well as multilingual, opens a wide range of possibilities for new methods and approaches in building resources for end users or to improve machine learning systems. Dictionary compilation is traditionally labour intensive and expensive, and research into automatic methods to aid that process can thus be of great practical value. The 4^{th} Translation Inference Across Dictionaries (TIAD) shared task aims at generating new translations automatically among three languages, English, French and Portuguese, based on known translations contained in Apertium dictionaries. Participants were allowed to use Apertium data and other freely available sources of background knowledge to

improve performance, as long as no direct translation was applied. After applying their methods, participants were to submit dictionaries containing pairs of source and target languages words, along with a score indicating the estimated probability of the pair being equivalent.

Our contribution is based on two approaches applied in parallel and combined in the final step to induce the bilingual dictionary aimed for. On the one hand, we pivot through intermediary dictionaries without any attempt to discern a possible translation candidate from other candidates. On the other hand, we run an ensemble of word alignment tools on a parallel corpus to create a probability score for all aligned words.

Word alignments have previously been used for automatically inducing bilingual dictionaries, see i.e. [11], [1], [15]. This is expected as it is easy to regard the outputs of word alignment models as hypotheses for translation equivalence. However, the problem with word alignments has been that these hypotheses are not necessarily very accurate, both due to the limitations of the aligners themselves and to the limitations of the data being aligned. We try to circumvent these limitations by using CombAlign [16], a tool that combines the output of an ensemble of word aligners and returns high-precision or high-recall alignments, according to the needs of the user and the task at hand. Furthermore, the CombAlign outputs are used to produce a confidence score for each pair, which can be applied to filter and remove the most improbable pairs. This setup results in a very competitive system, with better recall and F_1 scores than other participating systems.

2 System Description

We start by collecting as many lexical translations as possible. We use a subset of Apertium RDF v2 [6] (see Figure 1). Our main approach (described in Section 2.2) is pivoting through either one or two intermediary languages for each language pair. In order to score the candidate lexical translations, we extract sentence pairs from a parallel corpus, align them on word level and calculate a word alignment score (described in Section 2.3) for each aligned pair of words.

2.1 Datasets

We use the TSV versions of the Apertium dictionaries provided by the task organizers. The dictionaries we use are represented by edges in Figure 1.

Pivoting through the Apertium graph results in a high number of translation candidates. In order to filter that output and estimate which of the candidates have the highest probability of being correct translations, we use parallel corpora and word alignments to create a probability score. We acquire parallel corpora from OPUS [18] and, for each language pair, create sets that consist of 1 million sentence pairs. The sentence pairs are sampled from the Europarl, OpenSubtitles [9], Tatoeba and TildeMODEL [14] corpora. This procedure is described in Section 2.3.

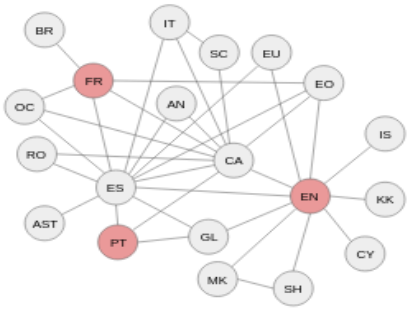

Fig. 1. The subset of the Apertium dictionaries used in the work described herein. Each bilingual dictionary is represented by an edge between vertices in the graph.

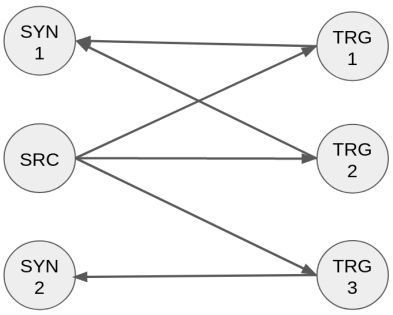

Fig. 2. Synonyms are derived by finding all target language equivalents for a source word and translating them back to the source language.

To validate our experiment, we acquired validation data from the task organizers. The validation sets contain 5% of randomly selected translations from the evaluation data source, for each language pair. According to the task organizers, the validation data is not used for the final evaluation.

2.2 Pivoting

We start by defining a list of all language pairs connected in the graph between the source language and the target language of the dictionary to be inferred. We read all the dictionary data to memory, creating two dictionaries for each language pair: SRC→TRG and TRG→SRC.

Optionally, the system can create "synonyms" for each word in the dictionaries. They are induced by using a set of all dictionaries containing the language in question. Within each dictionary, we look for all translations of a given word and then find back-translations into the source language again, thus finding words that may be related to the original word. The induction process is illustrated in Figure 2. The synonyms can then be used to create new translation candidates, by copying entries with the source word and replacing it with a synonym.

When using the Apertium dictionaries to infer translation candidates, we start by defining the source and target language for the dictionary to be inferred, and decide how many intermediary dictionaries are allowed. Our default is two intermediary languages. In the case of EN→PT this means 10 different paths from the source language to the target language, as illustrated in Figure 3, using edges of different colors for each path.

It has been demonstrated that by using a method called One Time Inverse Consultation (OTIC) it is possible to get a list of candidates with a good likelihood of the candidates being relevant [17]. OTIC induces a candidate list through

a pivot language, but sets restrictions that result in pruning of unlikely candidates. OTIC is used in one of the baseline systems for this shared task. Another method is for the algorithm to be absolutely naive and accept all words inferred through an intermediary dictionary. This means that for each source language word, we look up the intermediary words, and then look up the resulting target words from the intermediary words. This is illustrated in Figure 4 for one intermediary language. In the case of two intermediary languages another layer is added.

As our method relies on a scoring mechanism external to the Apertium dictionaries, our goal in this module is to extract as many potential candidates as possible. We thus opt for the naive approach as that gives us a larger candidate list than the OTIC method. After pivoting, we have large unfiltered dictionaries for each language pair we are working with. We add an extra pivoting step and repeat the process using our new induced dictionaries, enlarging them and adding even more translation candidates.

This results in candidate lists of 50-100K candidates, depending on the language pair. Our final step is to filter that list, but first we have to build the word alignment filter.

2.3 Filtering with Word Alignments

For filtering the inferred translation candidates produced by the pivoting process, we create a list of translation pairs with scores based on the likelihood of the words being aligned by a high-precision word alignment process. While we would probably get more accurate scores if we used all the parallel corpora we could

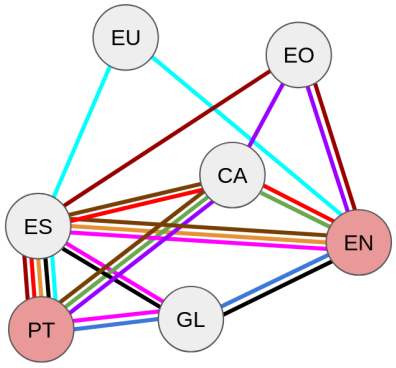

Fig. 3. An example of pivoting paths in the Apertium model between English and Portuguese. Paths using one or two intermediary languages are shown, each path in a different color.

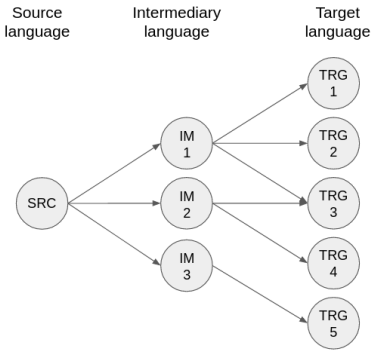

Fig. 4. Pivoting through intermediaries to collect translation candidates. The figure shows how five target language words are found for one source language word.

get, we also wanted to use the same process for all the language pairs we are working with in order to make the results more comparable between the inferred dictionaries. We thus decided to have all parallel corpora the same size, 1 million sentence pairs. The corpora were obtained from OPUS (see Section 2.1) and the sentence pairs selected using a greedy algorithm that only accepts a sentence if it contains a word from a word of lists in the Apertium source or target language dictionary. In order to get a decent coverage, after a word has been found in 10 sentences it is removed from the list. When we have collected the sentence pairs, the sentences are lemmatized using spaCy [7], and word alignments found for each sentence using CombAlign [16]. CombAlign combines multiple alignment tools in order to obtain maximum recall or maximum precision, according to the needs of the user. In our settings, we use six word alignment models using five different word alignment systems.[3] The system has different settings for reaching high recall or high precision. The settings we use are the following:

- Maximum Recall: Obtained by creating a union of all alignment hypotheses from all six models.
- Maximum Precision: Obtained by a simple majority vote. Word alignments suggested by four or more models are accepted.

Our scoring formula uses the word alignment information, combined with a count of word co-occurrences in the sentence pairs. Our confidence score is calculated for each word pair $\langle s, t \rangle$ using Equation (1):

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

where $mat(s,t)$ 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 [15], but we use it with a slight variation. While [15] set the smoothing variable λ to 20, we set it to $\log_2 s$ where s is the number of sentence pairs in the corpus under consideration. The scores should be in the range [0.0 ... 1.0]. When Equation (1) returns a number higher than 1, the score is set to 1.

2.4 PivotAlign

We call our combined system PivotAlign.[4] It executes the pivoting process and the scoring mechanism. It then combines their output by applying the word alignment scores to each inferred translation candidate in the unfiltered dictionary and removes candidates that are below a certain threshold. The combined system is illustrated in Figure 5.

[3] SimAlign [10] (two models: one based on mBERT [3] the other on XLM-R [2]), Giza++ [12], fast_align [5], eflomal [13], AWESoME [4]

[4] Available at `https://github.com/steinst/PivotAlign`

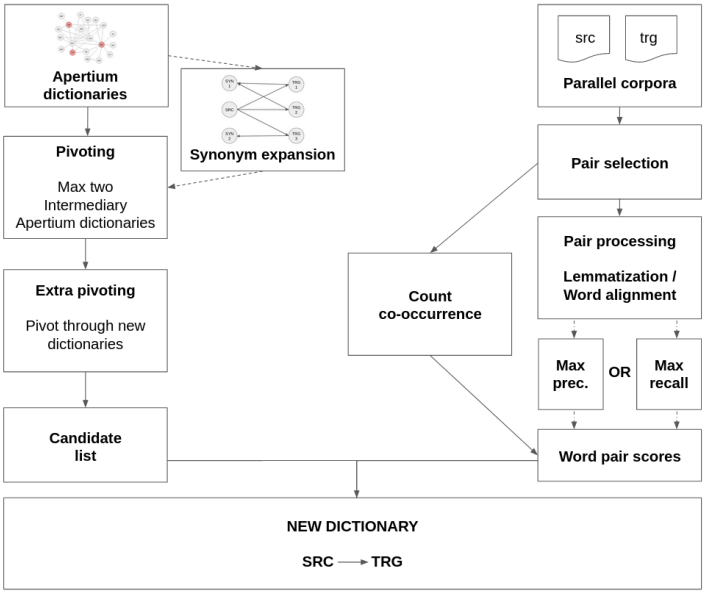

Fig. 5. PivotAlign

2.5 Different PivotAlign Variants

We submitted three variants of PivotAlign to the TIAD 2021 shared task. One system variant aiming for high precision, another aiming for high recall and the third aiming for a high F_1-score. We try to achieve these aims through settings in the system: whether we apply the induced synonyms, how many dictionaries we use as intermediaries when pivoting, whether we set CombAlign for high recall or high precision, and how we set our thresholds. We set two types of threshold, the alignment score, and the alignment score combined with an absolute alignment count in the parallel data. That combination is used to try to raise the recall in cases where alignment is common, even though the two words co-occurred very often in the data without the alignment being suggested. Our hypothesis is that this helps with very common words. In order to decide settings and thresholds, we used the verification data sets and selected the ones reaching the highest precision, recall or F_1, depending on the aim for that system variant.

PivotAlign-P Our system variant aiming for high precision obtained an average precision of 0.75 on the verification sets. The settings used were:

- Pivot: Max two intermediary languages.
- Alignment filtering: Maximum precision (majority vote)
- Score: $\rho(s,t) > 0.9$; $\rho(s,t) > 0.6$ and count > 200; $\rho(s,t) > 0.15$ and count > 300.

PivotAlign-R Our system variant aiming for high recall obtained an average recall of 0.61 on the verification sets. The settings used were:

- Pivot: Max two intermediary languages; inferred synonyms added.
- Alignment filtering: Maximum recall (union of all)
- Score: $\rho(s,t) > 0.15$.

PivotAlign-F Our system variant aiming for high F_1 obtained an average F-score of 0.49 on the verification sets. The settings used were:

- Pivot: Max two intermediary languages.
- Alignment filtering: Maximum recall (union of all)
- Score: $\rho(s,t) > 0.28$; $\rho(s,t) > 0.15$ and count > 200.

3 Shared Task Evaluation

Using the three different settings of PivotAlign described in Section 2.5, we generated translations in both directions for each language pair EN→PT, PT→EN, PT→FR, FR→PT, EN→FR and FR→EN. Evaluation of the results was carried out by the organisers against a gold standard with translations extracted from manually compiled pairs from an outside source, K Dictionaries (KD). To allow for comparison, only a subset of KD that is covered by Apertium was used to build the gold standard. Fourteen systems were submitted to the shared task and one of our submissions, PivotAlign-R scored highest both in term of recall and F_1-measure. That system also had the highest coverage of all the systems submitted.

Using our preferred threshold, PivotAlign-P, aiming for high precision, achieved a precision of 0.85, which was the third highest. While two systems, TUANMUSEes and TUANMUSEca, had slightly higher precision, 0.86 and 0.87, they had much lower recall, 0.10 and 0.08 respectively, compared to 0.24 for PivotAlign-P. When the threshold for PivotAlign-P was raised, precision went up to 0.88, while recall went down to 0.15. By changing the thresholds for TUANMUSEes and TUANMUSEca these systems also reached a maximum presicion of 0.88,

Top 5 Systems				
System	Precision	Recall	F_1-score	Coverage
PivotAlign-R	0.71	0.58	0.64	0.77
PivotAlign-F	0.81	0.51	0.62	0.68
ACDcat	0.75	0.53	0.61	0.75
TUANWEsg	0.81	0.47	0.59	0.76
TUANWEcb	0.81	0.47	0.59	0.76

Table 1. The five highest ranking systems with regards to F_1 score.

while their recall went even further down. Thus, our system also had the potentially best precision score, outperforming other systems reaching the same precision score in terms of recall.

As previously stated, PivotAlign-R reached the highest recall of all systems in the shared task, 0.58, considerably higher than the next system ACDcat with 0.53. Surprisingly, PivotAlign-R also had the highest F_1-measure, 0.64, higher than the 0.62 that PivotAlign-F reached, but PivotAlign-F was aiming for more balance between P and R and through that the highest F_1-score. While this goal was achieved against the validation data sets, the composition of the test sets seems to be slightly different and when the system was run on the test sets the reduction in recall outweighed the increase in precision, thus having lower F_1-measure than PivotAlign-R.

Finally, PivotAlign-R also had the highest coverage of all the participating systems, coverage being a measure of how many entries in the source language were translated with respect to the gold standard. The scores for the five highest ranking systems, with regards to F_1 score, are shown in Table 1.

The candidate translations accepted by PivotAlign-R contained scores from 0.15 and up to 1.00. These scores measured against the evaluation sets show how the precision score rises linearly as the threshold rises, while recall goes down, see Figure 6. This indicates a good correlation between our alignment score and translation inference, showing that a scoring mechanism based on accurate alignments from an ensemble of word alignment tools, can be highly valuable for tasks such as this one.

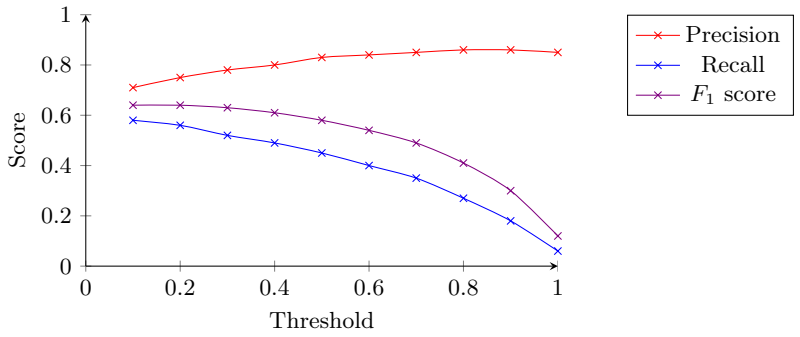

Fig. 6. Precision, recall and F_1 score charted against various threshold settings for PivotAlign-R. The threshold is highly correlated with all three scores.

4 Conclusion and Future Work

We have presented PivotAlign, a system that compiles two types of translation candidate lists and induces a dictionary from them. One translation candidate

list is built from parallel corpora using highly accurate word alignments, with a score indicating the likelihood of a pair being a pair of equivalents. The other candidate list is generated by inferring translations through pivoting via intermediary dictionaries.

In order to be able to have a similar setup for all language pairs, we limited our parallel corpora to one million sentence pairs for each language pair. There are much larger parallel corpora available for these language pairs and our first step in trying to improve our system will be to enlarge the corpora used for word alignment scoring. Adding additional scoring mechanisms and training a binary classifier to select candidate pairs based on multiple different types of scores may also bring improvements. For instance, it may be worthwhile to make use of scores derived from cross-lingual word embeddings, as they have been shown to perform reasonably well for this task [8]. When pivoting through the Apertium graph we pivot through a maximum of two intermediary languages. By allowing more than two intermediary languages our code will run slower but it will likely return more translation candidates. This may raise recall without necessarily lowering precision substantially.

We have shown that using a parallel corpus and high-precision word alignments is a viable mechanism for scoring inferred translation candidates, and that by combining that with a candidate list generated by pivoting, competitive results can be achieved, in our case securing our system the first position among the participating systems in the TIAD 2021 shared task.

Acknowledgements

This work is supported by the Language Technology Programme for Icelandic 2019-2023, funded by the Icelandic government, 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.

References

1. Caseli, H.M., Nunes, M.d.G.V., Forcada, M.L.: Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. Machine Translation **20**, 227–245 (2006)
2. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Online (Jul 2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Minneapolis, Minnesota (Jun 2019)

4. Dou, Z.Y., Neubig, G.: 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. pp. 2112–2128. Online (Apr 2021)
5. Dyer, C., Chahuneau, V., Smith, N.A.: 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. pp. 644–648. Atlanta, Georgia (Jun 2013)
6. Gracia, J., Fäth, C., Hartung, M., Ionov, M., Bosque-Gil, J., Veríssimo, S., Chiarcos, C., Orlikowski, M.: Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain, p. 499–514. Springer (Nov 2020)
7. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020)
8. Lanau-Coronas, M., Gracia, J.: Graph exploration and cross-lingual word embeddings for translation inference across dictionaries. In: Proceedings of the 2020 Globalex Workshop on Linked Lexicography. pp. 106–110. European Language Resources Association, Marseille, France (May 2020)
9. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 923–929. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016)
10. Masoud, J.S., Dufter, P., Yvon, F., Schütze, H.: SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1627–1643. Online (Nov 2020)
11. Melamed, I.D.: Models of translation equivalence among words. Computational Linguistics **26**(2), 221–250 (2000)
12. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics **29**(1), 19–51 (2003)
13. Östling, R., Tiedemann, J.: Efficient word alignment with Markov Chain Monte Carlo. Prague Bulletin of Mathematical Linguistics **106**, 125–146 (October 2016)
14. Rozis, R., Skadiņš, R.: Tilde MODEL - multilingual open data for EU languages. In: Proceedings of the 21st Nordic Conference on Computational Linguistics. pp. 263–265. Association for Computational Linguistics, Gothenburg, Sweden (May 2017)
15. Shi, H., Zettlemoyer, L., Wang, S.I.: Bilingual lexicon induction via unsupervised bitext construction and word alignment. ArXiv **abs/2101.00148** (2021)
16. Steingrímsson, S., Loftsson, H., Way, A.: CombAlign: a Tool for Obtaining High-Quality Word Alignments. In: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa). pp. 64–73. Reykjavik, Iceland (Online) (May 31–2 Jun 2021)
17. Tanaka, K., Umemura, K.: Construction of a bilingual dictionary intermediated by a third language. In: COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics (1994)
18. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2214–2218. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)