NUIG at TIAD 2021: Cross-lingual word embeddings for translation inference

NUIG at TIAD 2021: Cross-lingual Word Embeddings for Translation Inference

Sina Ahmadi[0000−0001−7904−6551], Atul Kr. Ojha[0000−0002−9800−9833], Shubhanker Banerjee[0000−0002−3969−5183], and John P. McCrae[0000−0002−7227−1331]

Data Science Institute, National University of Ireland Galway
firstname.lastname@insight-centre.org

Abstract. Inducing new translation pairs across dictionaries is an important task that facilitates processing and maintaining lexicographical data. This paper describes our submissions to the Translation Inference Across Dictionaries (TIAD) shared task of 2021. Our systems mainly rely on the MUSE and VecMap cross-lingual word embedding mapping to create new translation pairs between English, French and Portuguese data. We also create two regression models based on the graph analysis features. Our systems perform above the baseline systems.

Keywords: translation inference · bilingual lexicon induction · distributional semantics

1 Introduction

New words are coined regularly in our languages, many change meaning or usage depending on various social effects and some other words are considered obsolete over time. Documenting words and their senses by compiling dictionaries and collecting lexicographical data is a costly and time-consuming task which is carried out by lexicographers who are passionate about words in a language. One particular task that can facilitate the usage of dictionaries across languages and further add to the content value is bilingual lexicon induction, also known as translation inference.

Given a set of words in two different dictionaries, the translation inference task aims at aligning dictionary headwords that refer to the same concepts in an unsupervised manner. This task is deemed nontrivial and challenging due to the inconstant level of polysemy of words in a language and different coverage of senses in various resources [17]. In this context, the Translation Inference Across Dictionaries (TIAD) shared task aims to incentivize researchers to propose new techniques and approaches to translation inference in an unsupervised way.

This task is beneficial not only to align existing lexicographical data but also to create new dictionaries for less-resourced and under-represented languages which lack such resources. In addition, inducing new translation pairs enable lexicographers to document words more efficiently and therefore, facilitate the dictionary compilation process.

Our submissions to this year's shared task include the following systems:

1. `ONETA` based on orthonormal explicit topic analysis
2. `graph-SVR` and `ONETA-SVR` which are two support vector regression models, respectively based on the translation graph and the previous ONETA system
3. `MUSE` based on the MUSE unsupervised cross-lingual embedding mappings [12,5]
4. `VecMap` based on the VecMap unsupervised cross-lingual word embedding mappings [3]

We further describe our systems in the following sections.

2 Related Work

Year	Target dictionaries	Paper	Approach	External resources
TIAD 2017	German-Portuguese Danish-Spanish Dutch-French	[1]	graph analysis	-
		[16]	graph analysis and collocation-based models	Europarl corpus
		[7]	Support Vector Machine using features based on the translation graph and string similarity	-
TIAD 2019	English French Portuguese	[2]	multi-way neural machine translation	corpora of languages from the same family and Wiktionary
		[21]	graph analysis and neural machine translation	Directorate General for Translation corpus [18]
		[9]	pivot-based and cross-lingual word embeddings	monolingual corpora
		[6]	multi-lingual word embedding	pretrained embedding model
		[14]	unsupervised document embedding using Orthonormal Explicit Topic Analysis	Wikipedia corpora
TIAD 2020	English French Portuguese	[15]	unsupervised multi-way neural machine translation and unsupervised document embedding	Directorate General for Translation corpus [18]
		[4]	propagation of concepts over a graph of interconnected dictionaries using WordNet synsets and lexical entries as concepts	WordNet
		[13]	graph analysis and cross-lingual word embeddings	monolingual corpora of Common Crawl and Wikipedia
		[8]	graph analysis relying on paths, synonyms, similarities and cardinality in the translation graph	-

Table 1. An overview of the approaches proposed in the previous TIAD shared tasks

Table 1 summarizes the previously proposed techniques in the TIAD shared tasks. Graph analysis techniques rely on the analysis of translation graphs to determine a possible connection between two words. Among the famous techniques, pivot-based [21], cycle-based [7] and One Time Inverse Consultation approaches [13] are applied. On the other hand, external resources are used in an unsupervised way to train multi-way machine translation models and cross-lingual word embedding mappings. Although these techniques align translations without being trained on parallel corpora, they face challenges in retrieving part-of-speech tags and lemmatizing various word forms [2].

The datasets provided this year contain 44 languages and 53 language pairs, with a total number of 1,540,996 translations between 1,750,917 lexical entries. In should be noted that the datasets have been changed over years making the comparison of these techniques difficult.

3 Systems

3.1 Graph-based regression models

Our graph-based methods are based on the analysis that was performed previously in McCrae and Arcan [15], where the algorithm for extracting the connections between two nodes was applied as previously. We further extended this algorithm to extract the following measures from the graph:

- $d_{\min}(n, m)$: The minimum distance in the graph between the two nodes.
- $N_*(n, m)$: The number of paths between the nodes of any length.
- $N_2(n, m)$: The number of paths between the nodes of length 2.
- $a_*(n)$: The number of nodes reachable from node n.
- $a_1(n)$: The number of nodes directly connected to node n.

We used $d_{\min}$, N_* and N_2 directly as features in our system and we added to methods based on the One-Time Inverse Consultation [20, OTIC] as follows:

$$\frac{N_*(m,n)}{a_*(n)a_*(m)} \qquad \frac{N_2(m,n)}{a_1(n)a_1(m)}$$

This leads to five features in total which could be combined as a linear model.

Given that no training data is provided in the task, we apply our graph-based approach on the English-Spanish translation pairs to extract features for training. This data set is then used train two Support Vector Regression[1] models with a linear kernel, namely `ULD_graphSVR` and `ULD_OnetaSVR`. Given a new data instance based on our target languages, the regression models predict a score corresponding to the confidence score of the shared task. It is worth noting that all features and normalized and scaled properly.

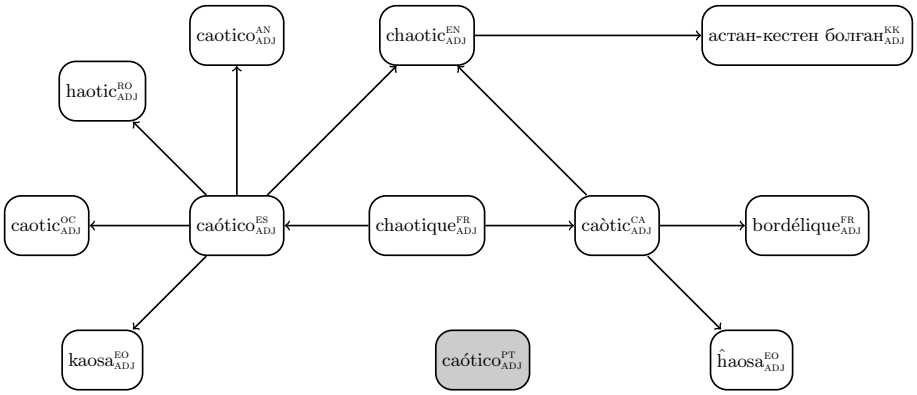

Fig. 1. Paths starting from '*chaotique*' (adjective in French) in the Apertium translation data. Language codes and part-of-speech tags are respectively provided in subscript and superscript.

3.2 Cross-lingual embedding mappings

One major limitation of graph-based methods is due to the limited coverage of connectivity between certain translations, i.e. nodes. Figure 1 illustrates some of the translations that can be retrieved for the word 'chaotic' (adjective) in the Apertium translation graph [10] where the Portuguese translation '*caótico*' ('chaotic') is not retrievable by traversing intermediate nodes.

In order to tackle this limitation, we use two unsupervised cross-lingual word embedding mapping techniques, namely VecMap [3] and MUSE [12]. These techniques find a mapping between the monolingual word embedding spaces of the source and target languages. Figure 2 shows a visualization of '*chaotique*' ('chaotic') in French and its closest words in both the French and Portuguese vector spaces.

VecMap based cross-lingual embedding was built on the unsupervised method using pre-trained French and English fastText monolingual embedding models[2]. After building the cross-lingual embedding and achieving confidence scores, we used monolingual pre-trained UDPipe 2.5 models [19][3] to generate the part-of-speech features only of the target(French) language. Furthermore, the generated parts-of-speech tags were mapped with parts-of-speech tags of the shared task.

In the same vein, a mapping is learned using the MUSE unsupervised method and fastText monolingual embeddings of French, English and Portuguese which takes use of adversarial learning followed by iterative Procrustes refinement (de-

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html`

[2] `https://fasttext.cc/docs/en/crawl-vectors.html`

[3] UDPipe 2.5 models were trained on 94 treebanks of 61 languages of Universal Dependencies (UD) framework (`https://universaldependencies.org/`) framework. This framework contains lemma, parts-of-speech, morphological and syntactic dependencies features.

fault configuration of $n\_refinements = 5$)[4]. Ultimately, these mappings are to create new translation pairs between the 10 most nearest translations in the target language using cosine similarity. The cosine similarity score is then considered as the confidence score in the final submission and the part-of-speech of the source word is used for the target predictions as well.

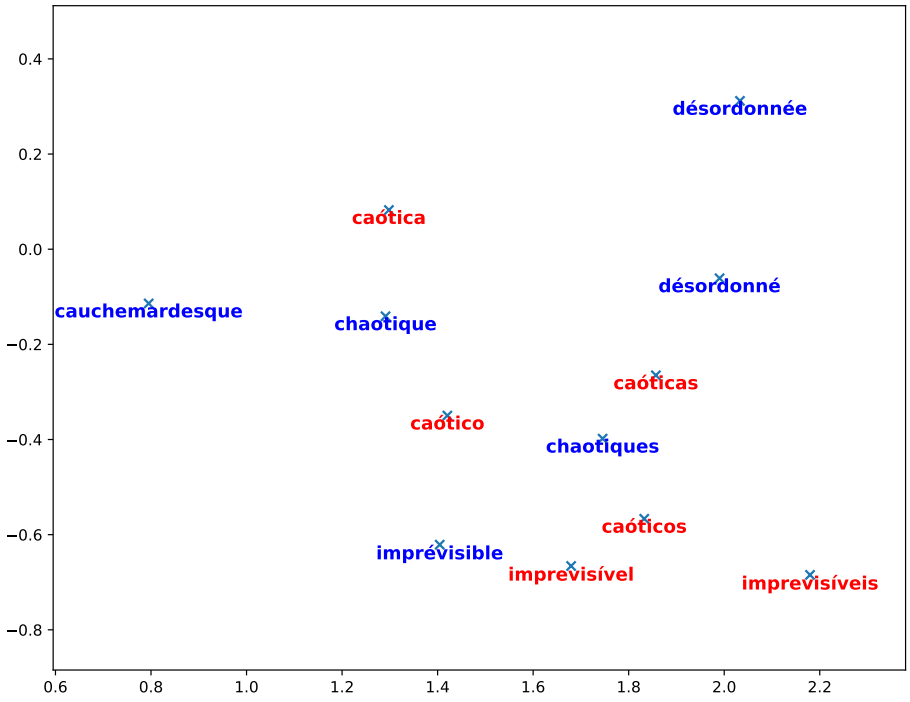

Fig. 2. t-SNE visualization of *chaotique* (adjective in French) in the MUSE multilingual word embeddings of French and Portuguese

4 Evaluation

Table 2 provides the evaluation results of our systems in comparison to the baseline systems. Results are averaged for every system and correspond to an arbitrary 0.5 threshold. In addition to the ULD_GraphSVR system, the ULD_MUSE system covers over half of the dictionary entries. Furthermore, Figure 3 (to the right) provides a comparison of the F1-measure of our systems and the baseline ones.

[4] We use the official codes of MUSE for training models: `https://github.com/facebookresearch/MUSE`

System	Precision	Recall	F1-measure	Coverage
ULD_GraphSVR	**0.7**	**0.49**	**0.57**	**0.69**
baseline-Word2Vec	0.69	0.23	0.33	0.4
ULD_MUSE	0.29	0.41	0.33	0.65
baseline-OTIC	0.78	0.18	0.29	0.28
ULD_OnetaSVR	0.76	0.1	0.17	0.14
ULD_Oneta	0.64	0.07	0.13	0.11
ULD_VecMap	0.36	0.01	0.01	0.02

Table 2. The performance of our systems in comparison to the baselines

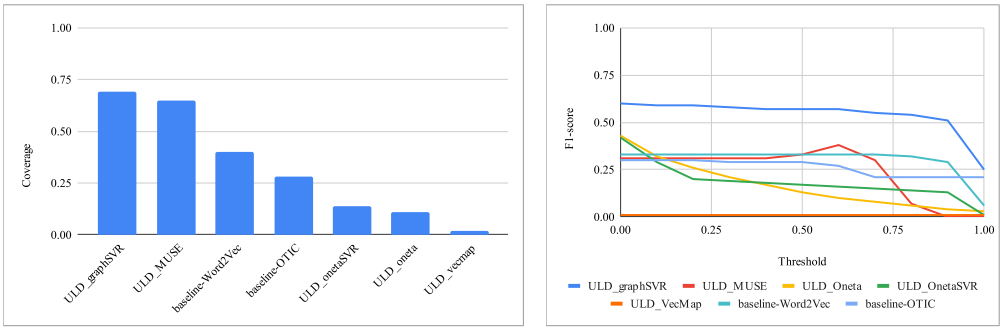

Fig. 3. A comparison of F1-score of our systems and the baselines. Coverage and performance of the English-French system based on the thresholds are respectively provided in the left and right plots.

5 Conclusion

In this paper, we describe our submissions to the Translation Inference Across Dictionaries (TIAD) shared task of 2021. We use the VecMap and MUSE unsupervised methods to create mappings between the monolingual embeddings of the target languages of the shared task, i.e. English, French and Portuguese. In addition, we use regression models to generate new translation pairs using graph analysis features.

As future work, we believe that word and contextual embeddings such as BERT should be studied for this task [11,22]. In addition, lemmatization and part-of-speech tagging should also be taken into account when using word and contextual embeddings which lack such information.

Acknowledgments

This work has received funding from the EU's Horizon 2020 Research and Innovation programme through the ELEXIS project under grant agreement No. 731015.

References

1. Alper, M.: Auto-generating Bilingual Dictionaries: Results of the TIAD-2017 Shared Task Baseline Algorithm. In: LDK Workshops. pp. 85–93 (2017)
2. Arcan, M., Torregrosa, D., Ahmadi, S., McCrae, J.P.: Inferring translation candidates for multilingual dictionary generation with multi-way neural machine translation. In: Proceedings of the Translation Inference Across Dictionaries Workshop (TIAD 2019) (2019)
3. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 451–462 (2017)
4. Chiarcos, C., Schenk, N., Fäth, C.: Translation inference by concept propagation. In: Proceedings of the 2020 Globalex Workshop on Linked Lexicography. pp. 98–105. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.globalex-1.16
5. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
6. Donandt, K., Chiarcos, C.: Translation inference through multi-lingual word embedding similarity. In: TIAD@ LDK. pp. 42–53 (2019)
7. Donandt, K., Chiarcos, C., Ionov, M.: Using Machine Learning for Translation Inference Across Dictionaries. In: LDK Workshops. pp. 103–112 (2017)
8. Dranca, L.: Multi-strategy system for translation inference across dictionaries. In: Proceedings of the 2020 Globalex Workshop on Linked Lexicography. pp. 111–115. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.globalex-1.18
9. Garcia, M., García-Salido, M., Alonso, M.A.: Exploring cross-lingual word embeddings for the inference of bilingual dictionaries. In: TIAD@ LDK. pp. 32–41 (2019)
10. Goel, S., Gracia, J., Forcada, M.L.: Bilingual dictionary generation and enrichment via graph exploration. the Semantic Web Journal (2021), preprint under review available at http://www.semantic-web-journal.net/content/bilingual-dictionary-generation-and-enrichment-graph-exploration
11. Gonen, H., Ravfogel, S., Elazar, Y., Goldberg, Y.: It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT. arXiv preprint arXiv:2010.08275 (2020)
12. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043 (2017)
13. 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), https://www.aclweb.org/anthology/2020.globalex-1.17
14. McCrae, J.P.: TIAD Shared Task 2019: orthonormal explicit topic analysis for translation inference across dictionaries. In: TIAD@ LDK. pp. 54–60 (2019)
15. McCrae, J.P., Arcan, M.: NUIG at TIAD: Combining unsupervised NLP and graph metrics for translation inference. In: Proceedings of the 2020 Globalex Workshop on Linked Lexicography. pp. 92–97. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.globalex-1.15

16. Proisl, T., Heinrich, P., Evert, S., Kabashi, B.: Translation Inference across Dictionaries via a Combination of Graph-based Methods and Co-occurrence Statistics. In: LDK Workshops. pp. 94–82 (2017)
17. Søgaard, A., Ruder, S., Vulić, I.: On the limitations of unsupervised bilingual dictionary induction. arXiv preprint arXiv:1805.03620 (2018)
18. Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P.: DGT-TM: A freely available translation memory in 22 languages. arXiv preprint arXiv:1309.5226 (2013)
19. Straka, M., Straková, J.: Universal dependencies 2.5 models for UDPipe (2019-12-06) (2019), http://hdl.handle.net/11234/1-3131, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
20. 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), https://www.aclweb.org/anthology/C94-1048
21. Torregrosa, D., Arcan, M., Ahmadi, S., McCrae, J.P.: TIAD 2019 shared task: Leveraging knowledge graphs with neural machine translation for automatic multilingual dictionary generation. Translation Inference Across Dictionaries (2019)
22. Zhang, J., Ji, B., Xiao, N., Duan, X., Zhang, M., Shi, Y., Luo, W.: Combining Static Word Embeddings and Contextual Representations for Bilingual Lexicon Induction. arXiv preprint arXiv:2106.03084 (2021)