Understanding Translationese in Multi-view Embedding Spaces

Koel Dutta Chowdhury
Saarland University
koeldc@lst.uni-saarland.de

Cristina España-Bonet
DFKI GmbH, Saarbrücken
cristinae@dfki.de

Josef van Genabith
Saarland University
DFKI GmbH, Saarbrücken
Josef.Van_Genabith@dfki.de

Abstract

Recent studies use a combination of lexical and syntactic features to show that footprints of the source language remain visible in translations, to the extent that it is possible to predict the original source language from the translation. In this paper, we focus on embedding-based semantic spaces, exploiting departures from isomorphism between spaces built from original target language and translations into this target language to predict relations between languages in an unsupervised way. We use different views of the data — words, parts of speech, semantic tags and synsets — to track translationese. Our analysis shows that (i) semantic distances between original target language and translations into this target language can be detected using the notion of isomorphism, (ii) language family ties with characteristics similar to linguistically motivated phylogenetic trees can be inferred from the distances and (iii) with delexicalised embeddings exhibiting source-language interference most significantly, other levels of abstraction display the same tendency, indicating the lexicalised results to be not "just" due to possible topic differences between original and translated texts. To the best of our knowledge, this is the first time departures from isomorphism between embedding spaces are used to track translationese.

1 Introduction

The term "translationese" refers to systematic differences between translations and text originally authored in the target language of the translation, in the same genre and style (Gellerstam, 1986; Baker and others, 1993; Baroni and Bernardini, 2005; Volansky et al., 2015). Characteristics such as simplification, over-adherence to conventions of the target language, and explicitation can occur as a communicative process itself. This is contrasted with "interference" or "shining-through" (Teich, 2003), described as "phenomena pertaining to the make-up of the source text tend to be transferred to the target text" (Toury, 2012). Prominent evidence for shining-through as a translationese effect is found in the work of Rabinovich et al. (2017), who show that footprints of the source language remain visible in translations, to the extent that it is possible to predict the original source language from the translation. In the similar vein, a significant amount of work has gone into training classifiers to distinguish between translations and originally authored text and then investigating the contributions of individual features to the result of the classification (Baroni and Bernardini, 2005; Koppel and Ordan, 2011; Volansky et al., 2015; Avner et al., 2016). Features that contribute strongly to classification are interpreted as indicating important dimensions of translationese.

In contrast, in this work, we leverage departures from isomorphism between embedding-based semantic spaces to detect translationese. We construct embedding spaces from original English (O) data and translations into English (T) from comparable data in a number of languages. We hypothesize that the closer the source language is to English, the more isomorphic the embedding spaces are. In other words, departure from isomorphism is an indicator of language distance. We use eigenvector similarity (Søgaard et al., 2018) to quantify departure from isomorphism. If our hypothesis is correct, we should be able to reconstruct phylogentic trees from measures of departure from isomorphism. We show that this is indeed

Proceedings of the 28th International Conference on Computational Linguistics, pages 6056–6062
Barcelona, Spain (Online), December 8-13, 2020

the case and compare our embedding-based results with previous results of reconstructing phylogenetic trees (Serva and Petroni, 2008). In order to show that our outcomes are not the result of topic divergences between O and T data, we explore delexicalised views[1] of our data, using embeddings based on parts of speech (PoS), semantic tags (ST), and synsets (SS), rather than word tokens (Raw). We show that our results are robust under delexicalisation.

Our paper is structured as follows: Section 2 discusses related work and inference of phylogenetic trees. Section 3 introduces our experimental setup. The distance measure is described in Section 4. We present our results and analysis in Section 5, followed by conclusions in Section 6.

2 Phylogenetics and Shining-through

Historical comparative linguistics determines genetic relationships between languages using concept lists of words that share a common origin, similar meaning and pronunciation across multiple languages (Swadesh, 1952; Dyen et al., 1992). By contrast, computational analysis methods aim to reconstruct language phylogeny based on measurable linguistic patterns (Rabinovich et al., 2017; Bjerva et al., 2019).

Rabinovich et al. (2017) showed that source language interference is visible in translation. Specifically, they leverage interference (PoS trigrams and function words) and translation universal features (cohesive markers) to construct phylogenetic trees. Agglomerative clustering with variance minimisation (Ward Jr, 1963) is used as linkage procedure to cluster the data. The result is compared to the pruned gold tree of Serva and Petroni (2008) (henceforth referred to as SP08) used as the linguistic phylogenetic gold standard tree. Their comparison metric, which is based on the L2 norm, is basically the sum of squared deviations between each pair's gold-tree distance g and computed distance P:

$$Dist(P, g) = \sum_{i,j} \left(D_P(l_i, l_j) - D_g(l_i, l_j)\right)^2 \quad (1)$$

SP08 was constructed by computing the Levenshtein (edit) distance between words from an open cross-lingual list (Dyen et al., 1992) to compare linguistic divergence through time and thus partially encodes lexical similarity of languages (Oncevay et al., 2020). Rabinovich et al. (2017) also acknowledges that SP08 has been disputed and researchers have not yet agreed on a commonly accepted tree of the Indo-European languages (Ringe et al., 2002). More recently, Bjerva et al. (2019) built on this work and compared different languages based on distance metrics computed on phrase structure trees and dependency relations. They claimed that such language representations correlate better with structural family distances between languages than genetic similarities. These examples show that phylogenetic reconstruction approaches and in particular, the evaluation of generated trees remains a highly debated topic in the history of linguistics and is beyond the scope of this study.

3 Experimental Settings

Data. We use the comparable portion of Europarl (Koehn, 2005) with translations from 21 European Union languages into English to minimise the impact of domain difference. The amount of tokens per language varies, ranging from 67 k tokens for Maltese to 7.2 M for German. We refer to the multiple translations into English as L_j's, where $j = 1, 2, ..., n$; and to originally written text in English as L_e. We select the subset of translations from 16 languages covering three language families: *Romance* (French, Italian, Spanish, Romanian, Portuguese), *Germanic* (Dutch, German, Swedish, Danish) and *Balto-Slavic* (Latvian, Lithuanian, Czech, Slovak, Slovenian, Polish and Bulgarian) into English and English original text.

Abstractions. In addition to using raw word tokens, we create multiple views of the data at the morphological (PoS), lexical semantic (ST) and conceptual-semantic (SS) levels. We use the spaCy tagger (Honnibal and Johnson, 2015) with the OntoNotes 5 version (Weischedel et al., 2013) of the Penn Treebank PoS tag set. For ST (Bjerva et al., 2016; Abzianidze et al., 2017), we use the best model of Brants

[1]Since these views represent diversified and complementary information of the same data, we refer to them as *multi-view* representations.

Feature	Annotated Output	Vocabulary
PoS	DT NN VBD DT NN IN NN	37
ST	DEF CON EPS DIS CON REL CON	57
Synset	ministry.n.04 send.v.02 answer.n.04 inquiry.n.0.1	1667
Raw	the ministry sent an answer to inquiry	6739

Table 1: Examples of the level of abstraction with the vocabulary size.

(2000) which works directly on the words as input, and determines formal lexical semantics. Their implementation achieves around 95% accuracy, when evaluated on short Parallel Meaning Bank sentences. For SS, we follow España-Bonet and van Genabith (2018) to retrieve synsets according to the PoS of a token using the knowledge base of WordNet (Miller, 1998). We only select a subset of PoS tags, namely *NN, ADV, ADJ* and *VB* and consider the first synset for each word/tag combination. Table 1 presents an overview and examples of each type of annotation used in this study.

Vector Spaces. For each view, we induce a separate monolingual word embedding space (both L_e and L_j's) by treating each token or tag as a word using fastText (Bojanowski et al., 2017). Embeddings have 300 dimensions and only words with more than 5 occurrences are retained for training. We use skip-gram with negative sampling (Mikolov et al., 2013) and standard hyper-parameters.

4 Measuring Isomorphism

An empirical measure of semantic proximity between two languages is often computed using the degree of isomorphism, that is, how similar the structures of two languages are in topological space (Søgaard et al., 2018). Research in cross-lingual transfer tasks shows that linguistic differences across languages often make spaces depart from isomorphism (Nakashole and Flauger, 2018; Søgaard et al., 2018; Patra et al., 2019; Vulić et al., 2020). While this degrades the quality of bilingual embeddings, it is a desired characteristic in our case: since our task involves processing of (multi-view) representations of *monolingual* text, *departures from isomorphism* indicate diversity in the source that generates them.

To quantify isomorphism, we compute embeddings on a corpus in language L. Embeddings reflect distributional properties in the data: words in similar contexts have similar meanings (Harris, 1954) and should be close in embedding space. We then view the points representing words or tags in the resulting hyperdimensional embedding space as a graph and compare different spaces (e.g. for data from different languages, or originals and translations into the language of the originals) in terms of how similar or dissimilar the corresponding graphs are. This is measured in terms of a well established metric, the eigenvector similarity.

Eigenvector Similiarity (EV). Søgaard et al. (2018) proposed this spectral metric based on Laplacian eigenvalues (Shigehalli and Shettar, 2011) to estimate the extent to which nearest neighbor graphs are isomorphic. We use the same idea to model differences between two spaces: original $\mathcal{X}$ and translationese $\mathcal{Y}$ for the single target language translations from different source languages. First, we compute the nearest neighbour graphs G_i of the two embedding spaces for the *most frequent* overlapping vocabulary.[2] We then compute the eigenvector similarity of the Laplacians of the nearest neighbor graphs, $\mathcal{L}(\mathcal{X})$ and $\mathcal{L}(\mathcal{Y})$ in original and translationese respectively. The degree of graph similarity is given by the distance among the eigenvectors λ of the Laplacian of G:

$$\Delta = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2 \tag{2}$$

Following Søgaard et al. (2018), we find the smallest k_1 in Equation 2 such that the sum of its k_1 largest eigenvalues $\sum_{i=1}^{k_1} \lambda_{1i}$ is at least 90% of the sum of all its eigenvalues. Analogously, we find k_2

[2]All tokens were converted to lowercase, and stop words and non-alphanumeric characters were filtered out.

Families	PoS	ST	SS	Raw
Germanic	1.00	0.95	0.70	0.56
Romance	2.86	1.27	0.75	0.59
Balto-Slavic	4.15	3.81	3.98	3.62

Table 2: Sum of normalised EV distance scores between original English and English target translations from each family (lower means closer to original) on the different views.

.

	PoS	ST	SS	Raw
τ	0.55	0.44	0.39	0.37

Table 3: Correlations between similarities (SPO8 and EV) on the different views of linguistic representations (the higher the better).

and set $k = min(k_1, k_2)$. The graph similarity metric returns a value in the half-open interval $[0, \infty)$, where values (Δ) closer to zero indicate more isomorphic embedding spaces.

To control the impact of data size for different L_j's, we choose the size of common overlapping vocabulary list corresponding to the range of the most resource-poor language in each view (see last column of Table 1) and run EV on this size for the rest of L_{j-1}.

5 Results and Analysis

To analyse the computed distances between original English and English target translations on different levels of linguistic analysis, in a first step, we calculate the sum of the normalised EV scores per language family (i.e., Germanic, Romance, Balto-Slavic) shown in Table 2. Translations from Germanic languages are the closest ones to original English (itself a Germanic language) regardless of the level of linguistic representation, followed by translations from Romance and finally from Balto-Slavic source languages. This shows that language distance in vector space is higher for etymologically distant language pairs in translation providing evidence that languages with similar topological semantic structure exhibit less interference. The fact that deviation from isomorphism between multi-view semantic spaces of translation into English and original English changes with respect to source language shows that source language interference is a strong characteristic of translated texts, adding new semantic space based support for the findings in Rabinovich et al. (2017).

Footprints of the source language into the translation product allow us to construct phylogenetic trees. Figure 1 shows the result of clustering the distance scores using agglomerative clustering with variance minimisation (Ward Jr, 1963) for four views considered in this study. Consistent results across all the trees demonstrate strong presence of the shining-through. We identify some of the well known language–language relationships in all four trees, such as for example, English is grouped with Germanic and Romance languages while Balto-Slavic languages are always put together. Some interesting divergences, with respect to *Balkan Sprachbund (BS)* are visible as well. The geographical location of Romanian opens it to cross-pollination with the other languages of the BS area and the figures provide some evidence for that.

Table 3 shows the Kendall τ correlation coefficients between how close languages are in SP08 and in our difference-from-isomorphism based reconstructions. Our results show that both lexicalised and delexicalised structures correlate reasonably well ($\tau \in [0.37, 0.55]$) with the linguistically motivated phylogenetic tree, indicating that the lexicalised results are not "just" due to possible differences in topic between original and translated texts. In fact, the delexicalised representations (PoS and ST) which are less influenced by topic and have fewer *types* show more functional similarities with SP08 than the fine-grained semantic representations.

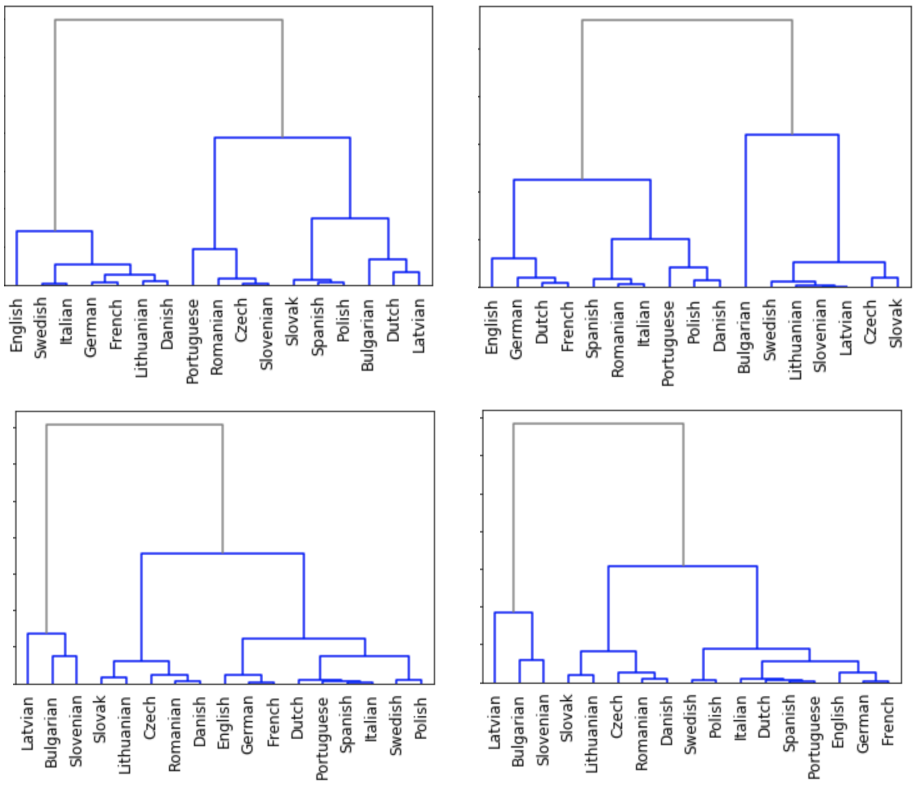

Figure 1: Clustering based on EV distances for the different views: PoS (*top–left*), ST (*top–right*), Synset (*bottom–left*) and Raw (*bottom–right*)

6 Conclusion

We presented an investigation of translationese effects in a single language sourced from multiple translations based on the topology of their multi-view embedding spaces. We explored embedding spaces constructed from word level (Raw), morphological (PoS), lexical semantic (ST) and conceptual-semantic (SS) views of the data. To the best of our knowledge, our study is the first to track translationese using isomorphism in semantic space.

Our translationese-based results can infer phylogenetic language family relations based on divergence from isomorphism between embedding spaces from translations and originally authored text. We find that language distances correlate with the divergence from isomorphism in embedding space. Our analysis demonstrates that while all embedding views exhibit source-language interference, delexicalised embeddings do so most significantly. In turn, this allows us to conclude that the lexicalised results are not just due to possible topic differences between original and translated texts.

Unlike supervised lexicostatistic approaches relying on aligned multilingual cognate lists, our isomorphism analysis is unsupervised and still able to detect important language differences related to linguistic typology. In a sense, and compared to some previous approaches, departure from isomorphism in embedding spaces lets "the data speak more for itself".

As future work, we intend to extend our experiments to capture geometric properties of the embedding features and work on isolated languages. Since we see that spaces with less tags, i.e., a smaller vocabulary, are better predictor of genetic relationships, more thorough robustness tests on the quality of the embeddings that might have an effect in skewing the results will be applied as well.

Acknowledgments

We would like to thank Rik van Noord and Antonio Toral for their feedback and providing us with the semantic tags. This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain, April. Association for Computational Linguistics.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, 31(1):30–54.

Mona Baker et al. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Marco Baroni and Silvia Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic tagging with deep residual networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA, April. Association for Computational Linguistics.

Isidore Dyen, Joseph B Kruskal, and Paul Black. 1992. An indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical society*, 82(5):iii–132.

Cristina España-Bonet and Josef van Genabith. 2018. Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems. In *Proceedings of the LREC 2018, MLP-MomenT Workshop*, pages 8–13, Miyazaki, Japan, May.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Ndapa Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 221–227, Melbourne, Australia, July. Association for Computational Linguistics.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. *arXiv preprint arXiv:2004.14923*.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy, July. Association for Computational Linguistics.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada, July. Association for Computational Linguistics.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-european and computational cladistics. *Transactions of the philological society*, 100(1):59–129.

Maurizio Serva and Filippo Petroni. 2008. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Vijayalaxmi S Shigehalli and Vidya M Shettar. 2011. Spectral techniques using normalized adjacency matrices for graph matching. *International Journal of Computational Science and Mathematics*, 2(4):371–378.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July. Association for Computational Linguistics.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

Elke Teich. 2003. *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Gideon Toury. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070*.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.