COMMERCE SCHOLARS PROGRAM
Final Research Paper

Heuristic Pivot Languages in Machine Translation

Stoyan Hristov[1], **Woonghee Tim Huh**[2]

[1,2] Sauder School of Business, University of British Columbia

April 2023

Abstract

Machine translation is the automatic translation of text from one language to another by representing words as vectors, called embeddings. While translation can be accomplished by directly mapping a source language's embedding space into the target's, it is possible to transit through a pivot language. Translating through a pivot language can increase computational efficiency by relaxing the requirement to learn mappings between every language pair, avoiding combinatorial explosion. However, translating through a poorly chosen pivot language can severely degrade translation quality. This study aims to find the best pivot language to translate through for any given pair of source and target languages by conducting an experiment on six European languages using the unsupervised multilingual alignment of word embeddings and the Bellman-Ford algorithm. Results indicate that translation from Portuguese to any of the other languages will see 58-75% improvement when English is used as a pivot. Translation from English to Portuguese benefits only from the direct translation. Translation between any of the other languages sees a 1-7% improvement when using a pivot language. The results of this study challenge the transitivity and symmetry of machine translation by showing that translation between almost every language pair will benefit from a pivot language.

Keywords: *Machine translation; pivot languages; unsupervised multilingual word embeddings; multilingual unsupervised and supervised embeddings; Bellman-Ford algorithm*

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

I would also like to acknowledge the following people:

Prof. Tim Huh, for his mentorship, support, and guidance in helping me choose an interesting yet feasible research question, as well as for poking holes in my ideas along the way.

Prof. Marc-David Seidel, for facilitating my first exposure to academic research and for cultivating a spirit of kindness and curiosity within me.

Prof. Young-Heon Kim, for getting me started on the research process by recommending a paper that explores machine translation using Optimal Transport Theory, a field of mathematics that he pioneered.

Colton Osterlund, for the countless hours we spent on-call debugging code and getting dependencies to line up. Without his support in the data generation process, this research would not have been possible.

The **Commerce Scholars Program Steering Committee**, for creating such an intellectually stimulating community and pushing all of us to become more capable researchers.

Stefanus Soegiarto, for sneaking me into a Commerce Scholars Program seminar before I was admitted into the program and for his enthusiasm throughout the following three years.

Jessie Lam, for welcoming me into that aforementioned seminar and for her empathy throughout the research process.

My parents, **Hristo Hristov** and **Tanya Hristova**, for taking a leap of faith by coming to Canada in 2006. For ingraining a culture of discipline and ambition in me. For pushing me to work hard in high school and forcing me to take breaks in university. And finally, for asking how my research was going every weekend despite having no clue what I was talking about. I would not be here without your support — thank you, I love you.

Contents

1. Introduction

Machine translation (MT) is the task of automatically converting text from one language to text in another language (Brownnlee, 2019).

In the 1970s, Rule-Based Machine Translation (RBMT) was the primary focus of research. This type of MT involves rules at the lexical, syntactic, and semantic levels that dictate the conversion of text from the source language to the target language (Brownnlee, 2019). However, translation requires a detailed understanding of the source text, which linguistic rules are inadequate to capture. As such, RBMT often needed to be augmented with detailed domain knowledge, increasing the cost of translation (Garg & Agarwal, 2018). Finally, RBMT proved difficult for languages with vastly different rules, such as English to Japanese (Garg & Agarwal, 2018).

To fill this gap, Statistical Machine Translation (SMT) was invented in the 1980s. SMT involves the use of statistical models that learn to translate text from a source language to a target language given a large corpus of examples (Brownnlee, 2019). Formally, every sentence S in a source language has a possible translation T in the target language. Then, each pair (S, T) is assigned a probability, $P(T|S)$, which is the probability that sentence T is the translated equivalent of sentence S (Garg & Agarwal, 2018). The problem of SMT was defined as:

$$T = \arg\max_T P(T|S) = \arg\max_T P(T)P(S|T) \tag{1}$$

SMT strongly depends on the distributional hypothesis, which states that words occurring in similar contexts have similar meanings (Harris, 1954). Given that this type of translation is data-driven, linguists were no longer required to specify the rules of translation. SMT was further innovated in the early 2000s by Neural Machine Translation (NMT). NMT uses neural networks to learn all the parameters in statistical models. NMT leverages a single system that can be trained directly on source and target text, meaning that the pipeline of specialized systems used in SMT is no longer required.

A parallel corpus is a piece of text that is identical in the source language and the target language. NMT that uses supervised learning requires some parallel corpus to build the translation. Currently, an exploding area of research is NMT using unsupervised learning. Through unsupervised learning, the requirement of a large parallel corpus to train NMT systems is relaxed, enabling translation to occur with less data. This is especially important considering the number

of languages in the world and the limited amount of data available for them (Garg & Agarwal, 2018).

2. Literature Review and Research Question

2.1. Literature Review

In the early 2000s, the increasing application of computational techniques to linguistics created controversy in the field. In caricature, computational linguists believed that throwing raw text into statistical black boxes could dispense of linguists altogether, whereas linguists believed that these statistical black boxes lacked the substance required to advance our understanding. However, the ultimate goal of linguistics is to understand language in a broad sense, and the mathematical techniques that computational linguists have begun using are yielding strong progress on previously intransient problems (Abney, 2002).

Mikolov et al. (2013) recognized that words could be translated by learning language structures based on large monolingual data and mapping between languages using a small bilingual dataset. By transforming words from a language into vectors (called distributed representations or embeddings) and projecting the vector space into two dimensions, they recognized that the vector spaces of languages share similar shapes. Then, they used these distributed representations in combination with a small starting dictionary to learn a linear mapping between the vector spaces of the source and target languages. Old SMT methods relied on dictionaries and phrase tables, requiring much effort to generate and resulting in performance far behind the performance of expert human translators. In contrast, this method made few assumptions about the languages, meaning that it could be used to extend and refine translation for any language pair, especially for languages that are substantially different, such as English to Japanese (Mikolov, 2013). This research was a breakthrough in the field of NMT and paved the way for numerous studies refining the performance of NMT methods.

Senrich & Haddow (2016) argued that, although NMT achieves impressive results with little use of external linguistic information, the strong learning capability of NMT models does not make linguistic features redundant. In fact, linguistic features can easily be incorporated to provide further improvements in performance. Later, it was proven that the linear transformation between two spaces should be orthogonal and that the transformation can be obtained from the singular value decomposition (SVD) on a dictionary of translation pairs (Smith et al., 2017). In an English-

Italian translation experiment, Smith et al. replaced the typical 5000-word dictionary with a pseudo dictionary acquired from identical characters that appear in both languages. While Mikolov's method only reached 1% precision, Smith's method achieved 40% precision. Smith et al. argued that the more robust orthogonal transformations enable us to learn translations from a pseudo-dictionary without the need of expert bilingual signal. Further, Artetxe et al. (2017) exploited the structural similarity of embedding spaces to create a translation using only a 25-word bilingual dictionary or even an automatically generated list of numerals. Their bootstrapping method worked well with low-dimensional pre-trained embeddings, which are currently widely used, and obtained results comparable to systems that used significantly richer resources. In sum, since Mikolov's study, lots of MT research revolved around achieving similarly-high quality translations with less data.

This trend continued until Conneau et al. (2018) developed a method called Multilingual Unsupervised Embeddings (MUSE) for translating words without any parallel data. They recognized that state-of-the-art methods at the time relied on bilingual dictionaries or parallel corpora to leverage the distributional hypothesis in translation. Further, although these requirements could be alleviated with character level information, the results were not on par with the supervised counterparts and the models were limited to pairs of languages sharing a common alphabet. Visualized in Figure 1, Conneau et al.'s method leverages unsupervised learning to align monolingual word embedding spaces and build a translation between two languages without using parallel corpora.

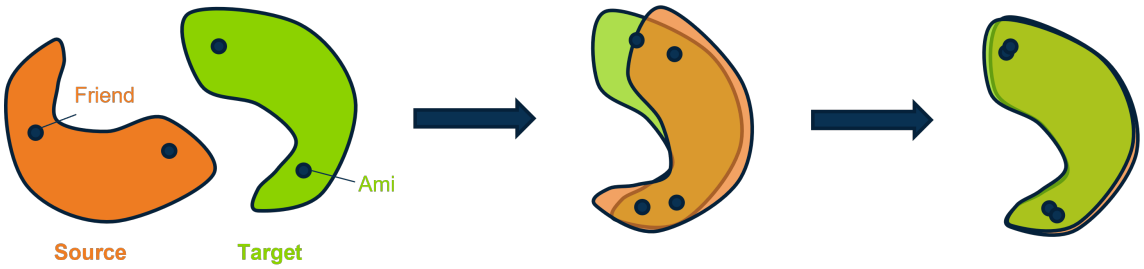


Figure 1. Visualization of the multilingual unsupervised and supervised embeddings algorithm, adapted from Conneau et al. (2018). The first step shows the embedding spaces of the source and target languages. The second step involves learning the multilingual alignment of the source embedding space into the target embedding space. The third step involves refinement of the alignment, then learning an optimal mapping between the two spaces.

In addition to previous approaches lagging in performance and requiring parallel data, evaluation of the algorithms suffered from the hubness problem. This is due to the asymmetric properties of nearest neighbour algorithms in higher dimensions. Namely, if y is a K-nearest-neighbour of x, it is not true in general that x is a K-nearest neighbour of y. This leads to a phenomenon such that some vectors are nearest neighbours to many points, whereas others are not nearest neighbours

to any points. Conneau et al. introduced an evaluation metric that mitigates this problem called cross-domain similarity local scaling (CSLS). In sum, Conneau et al.'s model achieves state-of-the-art performance machine translation between language pairs without requiring any parallel data and is extendable to low resource language pairs.

MUSE inspired computational linguists to continue challenging the distributional hypothesis. Grave et al. (2018) recognized that previous approaches to train distributed representations leveraged the distributional hypothesis by using the context around a word to train its vector. They then used a continuous bag of words (CBOW) model to represent words as bags of character n-grams and learn vector representations of words by only using web data from Wikipedia and the Common Crawl Project. These resources provide high quality data that is comparable across languages, making them a popular choice for training natural language processing models. The CBOW model proved fruitful and enabled them to train high quality 300-dimensional word vectors for 157 languages. These vectors accurately capture the syntactic and semantic properties of the words. Although these vectors are currently state-of-the-art, they have some limitations. Most prominently, the quality of the word vectors depends strongly on the availability of data for a given language. Low-resource languages such as Hindi lack web data to build vectors from, so their representations lag in performance.

Unsupervised multilingual word embeddings (UMWEs) encompass methods to represent words from multiple languages in a single distributional vector space without any cross-lingual supervision, which is a significant advantage over traditional supervised approaches (Chen & Cardie, 2018). Chen & Cardie (2018) recognize that MUSE relies on a number of independently trained unsupervised bilingual word embeddings, which failed to leverage the interdependencies that exist among languages. They then proposed a fully unsupervised framework for learning UMWEs that directly exploits the relations between all language pairs by aligning the monolingual word embeddings of all source languages to the space of the target language. This results in a shared multilingual embedding space in which they can learn the mappings between all pairs of languages. This method, visualized in Figure 2, substantially outperformed previous approaches and even beat supervised approaches trained with cross-lingual resources. However, there is a drawback to this method when using a single fixed target language with no interaction between the sources languages. For example, French and Italian are very similar. Converting them both to a less similar language such as English could degrade the translation quality (Chen & Cardie, 2018).

The approach of aligning many languages to the space of one in an unsupervised and simul-

Figure 2. Visualization of unsupervised multilingual word embeddings alignment algorithm, adapted from Chen & Cardie (2018). In contrast to Figure 1, this involves mapping a set of languages, in this case the source and the target, into the space of the pivot language. Then, a mapping can be learned between every pair of languages. Through this model, translating from the source to the target involves the longer translation path of source to pivot to target.

taneous way gained popularity. In 2020, Lian et al. recognized that neural language modeling has shown that word embeddings can capture both semantic and syntactic information. They further recognized that learning a small dictionary and leveraging statistical similarities between two languages had become a dated approach; NMT was now the state-of-the-art technique for machine translation. However, they argued that aligning many languages to the space of one common pivot language was a naïve approach because a poorly chosen shared language could degrade the translation quality. As such, they proposed using the Wasserstein barycenter (WB) of each language as the pivot language. The WB is an artificial language that minimizes the Wasserstein distance between embedding spaces, essentially minimizing the transportation cost, resulting in a strong translation. This method improved the accuracy of pairwise translations compared to other methods by optimizing a meaningful objective function independent from any bilingual data.

2.2. Gaps in the Literature

Let a "source language" (src) be the language we translate from, a "target language" (tgt) be the language we translate to, and a "pivot language" (pvt) be a language used as an intermediary for translation between the source and the target. Using a pivot language can reduce computational time required for training by relaxing the requirement to learn translations between all language pairs. However, retranslation can introduce possible mistakes and ambiguities. It is well known that pivot languages can improve translation by leveraging the interdependencies between languages (Chen & Cardie, 2018); however, translation through a poorly chosen pivot language can degrade the translation quality (Lian et al., 2020).

There has been research on artificial pivot languages improving machine translation; Lian et al

(2020) successfully leveraged the Wasserstein barycenter to improve machine translation through minimizing the transportation cost between languages. However, the Wasserstein barycenter of languages is not a natural language and, therefore, has limited implications beyond computational efficiency.

Although Chen & Cardie (2018) developed UMWE to leverage pivot languages, to the best of our knowledge, there is no literature aimed at discovering which is the best natural pivot language to choose given a source-target pair.

2.3. Research Question

This paper aims to answer the following question:

For a given source language and target language, which pivot language (if any) yields a better translation of words than the direct approach?

Given existing literature, we expect pivot languages to yield improved translations due to linguistic and algorithmic features. First, linguistically, using a pivot language would leverage the interdependencies between the source, target, and pivot languages (Chen & Cardie, 2018). Further, translation through a poorly chosen pivot language can degrade translation quality. Second, algorithmically, it has been shown that creating a pivot language to minimize transport cost between a source and target language will improve translation quality (Lian et al., 2020). We adopt the view that this Wasserstein barycenter is the optimal pivot language, similar to an exact solution. It follows that, if a natural pivot language were selected, it would act as a heuristic optimum. As such, we expect that, in general, there exists some strategically chosen natural pivot language that can improve the translation between a source and a target language. This would require the language to be linguistically "similar" or "compatible" with the source, target, and/or Wasserstein barycenter so as to not degrade the translation quality.

One way to interpret this question is to adapt the following colloquial paraphrase: “to translate between two languages, should I translate between them directly, or is there some third language that I can use as a bridge?”

3. Methodology

Given the lack of existing research on which pivot language to use given a fixed source and target, this paper adopts an inductive approach with the goal to infer theoretical concepts and patterns from the observed data.

3.1. Data

Although alignment algorithms can be extended to low-resource languages, the word representations of low-resource languages are of lower quality than those of high-resource languages. As such, we conduct experiments on six high-resource European languages: English (en), French (fr), Italian (it), Spanish (es), German (de), and Portuguese (pt).

First, we download high-quality vector representations of each language from FastText. This is an open-source resource developed by Grave et al. (2018) that allows users working on standard hardware to learn text representations. Word vectors are typically between 100- and 1000-dimensional. However, since low-dimensional pre-trained embeddings are widely used in NMT, we opt to use 300-dimensional word vectors for each language. All embeddings we used were downloaded in January 2023.

Second, we implement Chen & Cardie's (2018) UMWE model to align the vector spaces of source languages to one target language. This model is publicly available through GitHub. We choose to use unsupervised NMT because it is the most recent technique, has achieved higher performance than its predecessors, is widely used in industry, and requires the least data and linguistic proficiency to use. In sum, it is the lowest-cost and highest-performance method to answer our question. Further, the recent explosions in NMT literature increase the usefulness and future applicability of our study. We choose to use Chen & Cardie's UMWE model specifically because, to the best of our knowledge, it is the only NMT model that is compatible with the use of pivot languages. Further, it has topped state-of-the-art performance, resulting in high-quality translations.

To clarify the nomenclature, what is referred to as a "target language" in Chen & Cardie's model is referred to as a "pivot language" in this paper. For example, say we run UMWE with English and French as source languages and Portuguese as a target language. This would align English and French to the embedding space of Portuguese. Then, UMWE would learn a mapping between all language pairs. In this case, the en-fr and fr-en mappings would use Portuguese as a

pivot language. Further, the en-pt and fr-pt mappings would be equivalent to directly translating English and French, respectively, to Portuguese. Similarly, the pt-en and pt-fr mappings would be equivalent to directly translating Portuguese to English and French, respectively.

Third, although UMWE can learn mappings in a fully unsupervised way without parallel corpora, some dictionary between the language pairs is needed to evaluate the performance of the final translation. As such, we use the dictionaries developed by Glavas et al. (2019). These dictionaries were created using bilingual lexicon induction (BLI): the task of inducing word translations from monolingual corpora in two languages (Irvine & Callison-Burch, 2019). Although the evaluation of NMT models is still in the early stages of research, BLI is currently the most comprehensive and accurate method for evaluating unsupervised multilingual word embeddings.

Finally, the evaluation of these alignments outputs a normalized CSLS score between 0 and 1, which can be interpreted as the percentage accuracy of translation.

3.2. Training Algorithm

Use of UMWE proved complex. First, we had to download all proper dependencies to run the algorithm and ensure that it leveraged our GPU rather than CPU to save time. This required installing a mirror that had the exact version of each library that UMWE uses. Then, we were able to run UMWE by coding a PowerShell script through a Linux emulator, which then generated text logs of the UMWE results.

A crucial aspect of selecting high quality machine learning models is their validation. Typically, model validation is achieved by splitting a dataset into training and testing data — the model is trained on the training set and is then evaluated on the testing set, which it has never seen before. However, since UMWE uses unsupervised learning, there does not exist a training set or a validation set. As such, Chen & Cardie (2018) used a surrogate validation criterion that does not depend on bilingual data. This surrogate showed promising results when used by Lample et al. (2018) to evaluate the performance of their unsupervised machine translation model. Then, Chen & Cardie (2018) adapted the criterion to fit in the multilingual setting and hence be compatible with pivot languages. This results in using the most accurately trained model in our alignment and mapping of embedding spaces, yielding high internal and external validity.

3.3. Experiment 1

We ran UMWE for each triplet of source, pivot, and target languages as described above. Then, we evaluated the translation and obtained a CSLS score for each triplet. Finally, we analyzed translation performance on four fronts using Python. First, we evaluated how each language performed as a pivot by taking its mean CSLS when used as a pivot language. Second, we evaluated the percentage improvement of each language when used as a pivot by comparing the pivot translation to the direct translation. For example, if src-pvt-tgt yields a CSLS of 0.9 whereas src-tgt yields a CSLS of 0.8, the percentage improvement of using a pivot language in this specific case would be $\frac{0.9-0.8}{0.8} = 12.5\%$. Our last two metrics, consistent with current literature on NMT, evaluate how each language performs as a source and target language by analyzing each language's mean CSLS score when used as a source and target, respectively.

3.4. Experiment 2

We modeled language translation as a complete bidirectional graph. Each node represents a language and each edge represents translation from the source node to the target node. Further, the weight on each edge represents the CSLS for direct translation from the source language to the target language. We reformulated the research question as "what is the shortest path from the source node to the target node".

Arbitrage is defined as the purchase and sale of foreign exchange in different markets in order to profit from price discrepancies (Mahajan et al., 2020). For example, if one has CAD and can achieve a greater amount of USD by first converting CAD to JPY and then to USD rather than directly converting CAD to USD, then there is an arbitrage opportunity. The parallels between foreign exchange arbitrage and machine translation are evident, however this topic has only briefly been mentioned as a possibility in any of the literature we have come across. Mahajan et al. (2020) leveraged the Bellman-Ford algorithm to detect arbitrage opportunities. The Bellman-Ford algorithm calculates the shortest path in a bottom-to-top manner. First, it calculates the shortest path with the least number of edges involved. Then, it calculates the shortest path with two edges, and iterates the process up to $|V| - 1$ edges, where $|V|$ is the cardinality of the set of vertices.

To calculate the shortest path, the Bellman-Ford algorithm is minimizing a sum of the edges. However, the goal of foreign exchange arbitrage is to maximize a product of the edges. To see the extent to which this applies in NMT, we adopt a similar approach. For the Bellman-Ford

algorithm to detect a path that maximizes the product of the edge weights, we leverage Mahajan et al.'s (2020) approach and update the edge weights with $ln(\frac{1}{CSLS})$.

After modeling language translation as a graph and finding the shortest path between nodes, we analyze the paths that the Bellman-Ford algorithm detects and note their improvements over the direct translations. All implementation and analysis is conducted with Python. Finally, we compare and contrast the results of this experiment with the paths identified in our first experiment to see the extent to which language translation follows a similar pattern as foreign exchange arbitrage.

4. Results

4.1. Experiment 1

Our first experiment involved aligning languages to the embeddings of different pivot languages, repeating this procedure for all triplets of source, pivot, and target languages, and then analyzing the results.

First, we examined the mean CSLS of each language when used as a pivot. This metric, shown in Figure 3, indicates the average performance of each language as a pivot. We observe that German has the highest mean CSLS score, whereas English has the lowest CSLS score. French, Spanish, and Italian all have similar CSLS scores, whereas Portuguese is lagging.

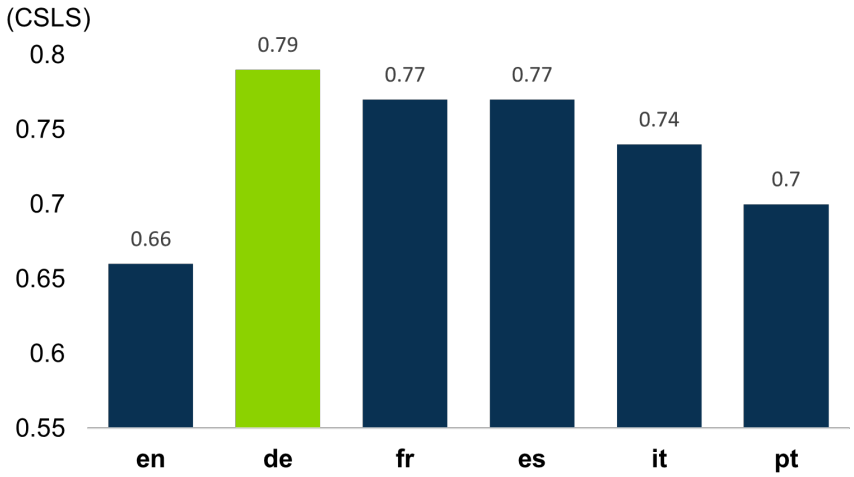


Figure 3. Mean CSLS between all translation pairs when each language is introduced as a pivot in translation.

These results are counterintuitive. Given the abundance of English data, many computational

linguists use English as a pivot language to reduce computing time (Lian et al., 2020). However, these results indicate that, as the worst pivot language of the group, using English as a pivot language does, in general, degrade translation quality when compared to other possible pivot languages.

Second, we examined the percentage improvement of CSLS when a pivot language was introduced relative to the direct translation. As seen in Figure 4, we note that German, French, Spanish, Italian, and Portuguese all result in low single-digit percentage improvements. However, using English as a pivot resulted in on average 67.01% improvement in CSLS.

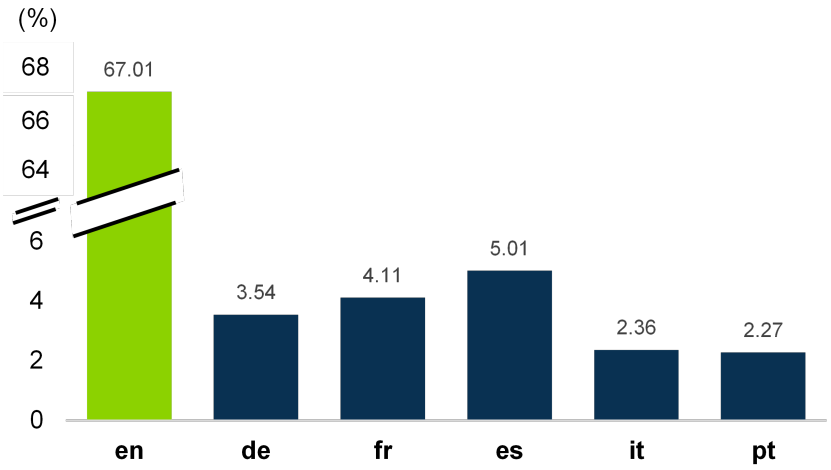


Figure 4. Mean percentage improvement in CSLS between all translation pairs when each language is introduced as a pivot in translation.

These results are paradoxical given commonly held beliefs about machine translation and the above discussion. They indicate that the use of a strategic pivot language does, on average, improve machine translation performance. The performance of English, however, appears to contradict the results discussed above, which claim that English should be the poorest performing pivot language. The explosive improvements when English is introduced as a pivot language arise due to Portuguese, which has severely low CSLS when translating directly between every language except English. In this case, when a pivot language is introduced, English is able to significantly improve CSLS. However, although these improvements are large, English does not help the translation enough to increase the CSLS to the levels of other pivot languages. Further, interestingly, English was never selected as the best pivot language for any non-Portuguese language pairs. This explains the relatively lower mean CSLS score discussed previously.

Third, we examined the mean CSLS of each language when used as a source language, shown in Figure 5. The results show that English, German, French, Spanish, and Italian all performed

similarly. Portuguese, however, scored the lowest by far. Fourth, we examined the mean CSLS of each language when used as a target language, shown in Figure 6. Once again, most languages saw similar performance except for Portuguese, which lagged severely.

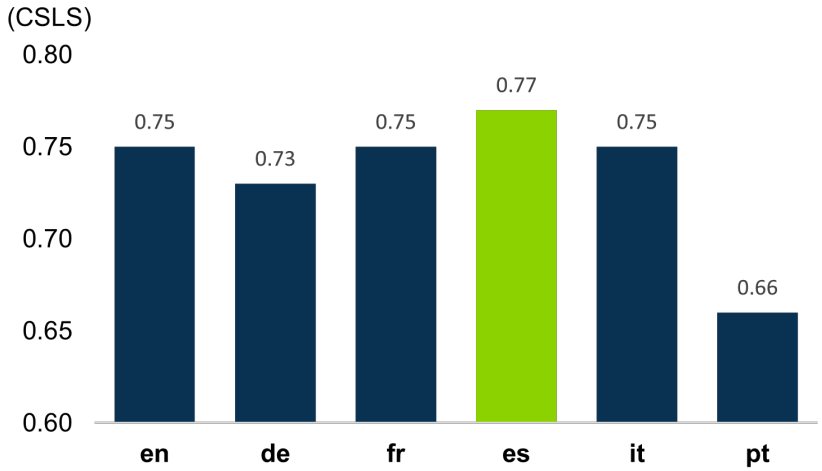


Figure 5. Mean CSLS in translation between the source language and all other languages.

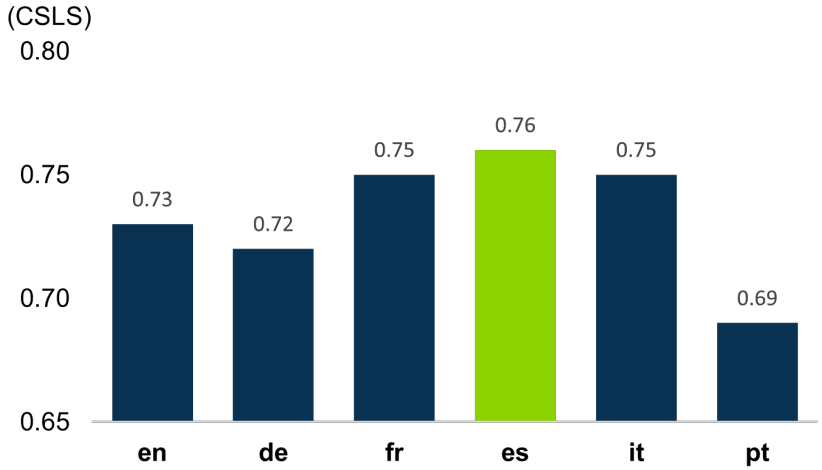


Figure 6. Mean CSLS in translation between all languages and the target language.

These results are interesting; given the previous two discussions, we expected German to have relatively high source and target CSLS scores. This, when combined with the relatively low percentage improvement in CSLS when German is used as a pivot, indicates that German is often chosen for language pairs that already have a decent direct translation. Then, the properties of German's embedding space is able to marginally improve the translation. For example, translation from Spanish to French has a CSLS of 0.76 (the second highest direct translation in the sample); however, when German is introduced as a pivot, the CSLS increases by 4.04%. Although the percentage improvement is low, the pivot CSLS becomes 0.79 — once again, the second highest in the pivot translation sample.

While these results are aggregate, we also examined data for individual translations. For example, Figure 7 shows how the CSLS when translating from French to Spanish varies based on our choice of pivot. In this case, choosing German as the pivot will result in the highest CSLS. In these graphs, choosing the pivot language to be the same as the target language (es) simply identifies the direct translation. In this case, the direct translation performs worse than pivoting through any of German, Italian, or Portuguese. However, pivoting through English will yield the worst CSLS, even when compared to the direct translation.

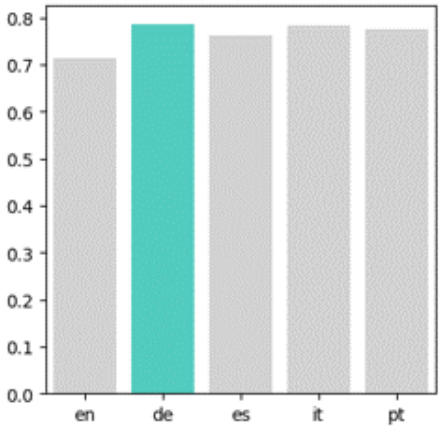


Figure 7. CSLS by pivot when translating from French to Spanish.

Further, Figure 8 shows how the CSLS when translating from English to French varies based on our choice of pivot. In this case, choosing Spanish as the pivot will result in the highest CSLS. However, the direct translation (fr as a pivot) yields the lowest CSLS. This implies that, for this specific language pair, any chosen pivot language will yield a better translation than the direct approach.

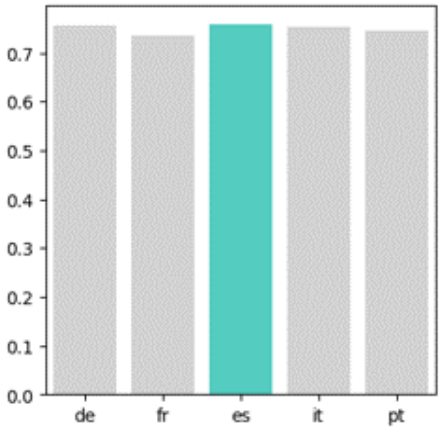


Figure 8. CSLS by pivot when translating from English to French.

The performance of all pivot languages for each of the language pairs is available in Appendices

A1 to A30.

Finally, we examined the effects of translating through a poorly chosen language. These results, shown in Figure 9, show that, for most language pairs, the direct translation performs the worst when compared to translating through any available pivot language. However, there are a few exceptions. Translating from French to German suffered by 0.1% when translating through Portuguese. Also, translations from French to Spanish and Spanish to French suffered by 6.1% and 5.4% respectively when translating through English. Further, translation to and from Portuguese sees massive detriment when translating through Italian and German, indicating that the embedding space of Portuguese clashes with the embedding spaces of those two languages. To conclude, although Lian et al.'s (2020) argument that translating through a poorly chosen pivot language degrades translation holds, its extent is much smaller than expected. In fact, for almost every language pair, direct translation performs the worst, indicating that pivot translation should be used for more accurate results.

src/tgt	en	de	fr	es	it	pt
en	-	direct	direct	direct	direct	de (49.7%)
de	direct	-	direct	direct	direct	it (47.7%)
fr	direct	pt (0.1%)	-	en (6.1%)	direct	it (49.6%)
es	direct	direct	en (5.4%)	-	direct	it (53.2%)
it	direct	direct	direct	direct	-	de (47.9%)
pt	de (45.1%)	it (11.7%)	it (11.0%)	it (20.2%)	de (6.1%)	-

Figure 9. The worst pivot language and percentage detriment under the direct translation for each language pair. "direct" indicates that the direct translation performed the worst.

To summarize, the results of our first experiment indicate the follow three general phenomena. First, translation from Portuguese to any language except English will always benefit greatly from English as a pivot language, whereas translation from Portuguese to English should only use the direct translation. Second, translation from any language to Portuguese should only use the direct

translation. Third, translation from any non-Portuguese language to any other non-Portuguese language will always benefit slightly from some third strategically chosen language as a pivot. The best pivot language for each language pair, as well as its percentage improvement, are highlighted in Figure 10.

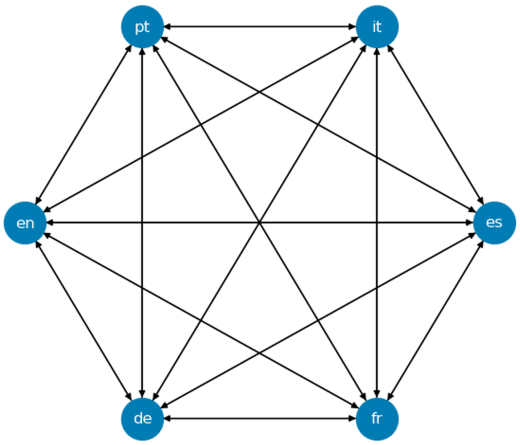
src/tgt	en	de	fr	es	it	pt
en	-	es 3.4%	es 3.2%	fr 2.5%	fr 4.5%	n.a. -
de	es 7.5%	-	it 3.9%	pt 3.0%	fr 3.9%	n.a. -
fr	es 5.9%	it 0.8%	-	de 3.3%	de 3.5%	n.a. -
es	fr 5.7%	pt 1.5%	de 4.0%	-	fr 3.6%	n.a. -
it	fr 6.6%	fr 2.4%	de 3.3%	fr 3.6%	-	n.a. -
pt	n.a. -	en 59.0%	en 70.2%	en 63.3%	en 75.5%	-

Figure 10. The chosen pivot language and percentage improvement over the direct translation for each language pair. "n.a." indicates that the direct translation performed best, and so no pivot was chosen.

4.2. Experiment 2

Inspired by foreign currency exchange arbitrage, our second experiment involved modeling language translation as a complete bidirectional graph and finding the shortest path between languages, as shown in Figure 11.

Examining the direct translation scores between languages, we note the following three observations. First, observe the performance of Portuguese as a source language. It has extremely low CSLS scores when translating to German, French, Spanish, and Italian. Its CSLS score when translating to English, albeit low compared to other values, is relatively high when compared to other Portuguese-as-a-source values. Second, observe the performance of Portuguese as a target. Its scores are low when compared to other values, but only by a small amount. Third, translation between language pairs that exclude Portuguese have relatively higher CSLS scores between 0.71 and 0.77.



src/tgt	en	de	fr	es	it	pt
en	-	0.74	0.74	0.75	0.73	0.69
de	0.72	-	0.71	0.72	0.72	0.66
fr	0.72	0.73	-	0.76	0.75	0.68
es	0.73	0.73	0.76	-	0.77	0.73
it	0.71	0.73	0.75	0.77	-	0.68
pt	0.63	0.39	0.39	0.43	0.38	-

Figure 11. Visualization of language translation as a complete bidirectional graph; edge weights on the right represent the direct translation CSLS scores from the source to the target language.

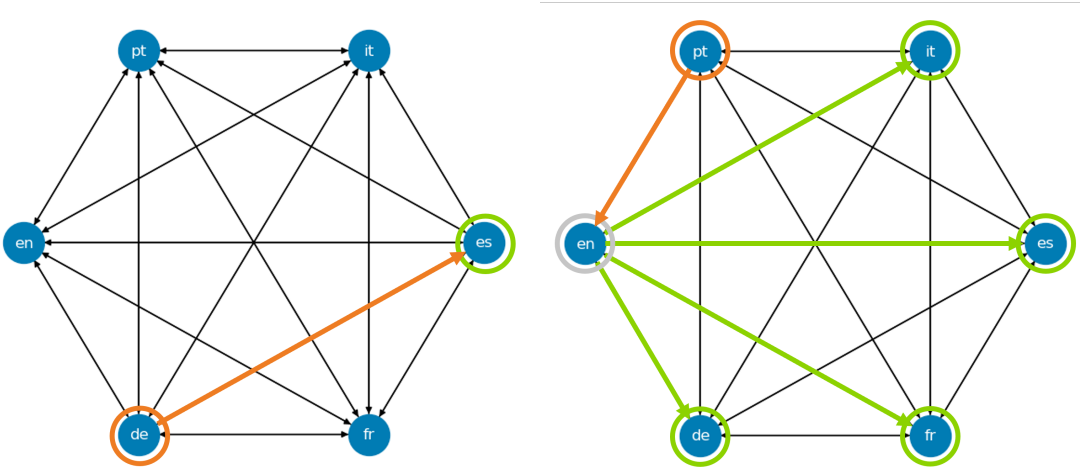


Figure 12. Sample findings from the Bellman-Ford algorithm; translation from German to Spanish is best when conducted directly, whereas translation from Portuguese to any other language will benefit from English as a pivot language.

This experiment, visualized in Figure 12, illuminates that the best way to translate between most language pairs is the direct translation. Given the multiplicative search implemented in the Bellman-Ford algorithm, this is congruent with our expectations; most translation scores are around 0.70, so moving through a pivot would reduce the score to around 0.49. In contrast, this is not true when Portuguese is the source. For example, direct translation from Portuguese to German would result in a score of 0.39. However, translation from Portuguese to English to German would result in a translation score of $0.63 * 0.74 = 0.47$. This yields a 20.5% improvement over the direct translation. The translation improvements from Portuguese to all other languages are demonstrated in Figure 13. This is unsurprising given how the direct CSLS of Portuguese to English is much higher than the direct CSLS scores of Portuguese to all other languages.

Target	Direct CSLS	Pivot CSLS (en)	% Improvement
de	0.39	0.47	20.5%
fr	0.39	0.47	20.5%
es	0.43	0.47	9.3%
it	0.38	0.46	21.1%
en	0.63	n.a.	n.a.

Figure 13. Translation from Portuguese to other languages modeled by the Bellman-Ford algorithm.

To summarize, the Bellman-Ford algorithm was able to identify the pivot language opportunities with the highest improvement in CSLS, but was not sensitive enough to detect all pivot language opportunities. This indicates that, although a reasonable proxy, machine translation cannot be perfectly modeled by foreign currency exchange rate arbitrage.

5. Discussion

The results of this study are heavily influenced by the poor performance of Portuguese; it is important to understand why this performance occurs. Although MUSE is the state-of-the-art translation engine, and Chen & Cardie's (2018) adaptation of MUSE to incorporate pivot languages outperforms the initial model's results, the model is still far from perfect. Despite being remarkably multilingual, MUSE falls short for certain languages (Cao et al., 2020). The results of this study indicate that Portuguese is one of these languages. Further, we see that over the course of ten years, machine translation results have drastically improved. Given the amount of technological innovation and availability of data, we expect results to continue improving drastically. As such, we expect machine translation algorithms to improve for all languages, especially languages that currently underperform significantly.

We also noted that, although the Bellman-Ford algorithm managed to identify the most extreme language paths, it was not sensitive enough to identify all pivot languages opportunities. In partricular, when CSLS scores are similar enough, the Bellman-Ford algorithm will recommend direct translation rather than translation through a pivot, which our first experiment showed can improve translation in almost all scenarios. Combining this with our expectation that, in the long run, CSLS scores of all language pairs will be relatively similar, it becomes evident that the Bellman-Ford algorithm is neither a suitable nor sustainable solution to identifying pivot language opportunities. As such, to detect all beneficial pivot language opportunities, machine translation should not be modeled after foreign currency exchange rate arbitrage.

Diving deeper, we investigate the symmetric properties of NMT. Referring to the table in Figure 11, it is evident that machine translation, although close, is not symmetric. This is because of the way translations are built. For example, translation from English to German involves taking the word embedding space of English, aligning it to that of German, and then learning a mapping between the two languages. In contrast, translation from German to English involves taking the word embedding space of German, aligning it to that of English, and then learning a mapping between the two languages. These are two different alignments and two different mappings. Given the probabilistic nature of neural networks and the slight differences of tasks, as expected, this technique does not yield symmetric results.

Despite the asymmetric properties of unsupervised multilingual alignment, we see that the choice of pivot is mostly symmetric. In fact, let x and y represent two languages, $x \neq y$. Let $P(\cdot, \cdot)$ represent the best natural pivot language when the first argument is the source and the second argument is the target. Then, for $x \neq \text{pt}$ and $y \neq \text{pt}$, our results indicate that

$$P(x, y) = P(y, x) \tag{2}$$

However, $P(\text{pt}, y) = \text{en}$, which is congruent with the results of our first experiment: translation from Portuguese to any language will benefit from English as the chosen pivot language. Further, $P(x, \text{pt}) = \text{pt}$, which is also congruent with the results of our first experiment: the best translation from any language to Portuguese will always be direct.

6. Implications

6.1. Mathematical

The main contribution of this paper is that it challenges the transitivity of language translation. Given a set of languages, $\mathcal{L}$, one might expect the direct translation from ℓ_{src} to ℓ_{tgt} to perform best. Given intuition from the triangle inequality, this approach appears reasonable. However, this study shows that, in almost all cases, there exists some ℓ_{pvt} such that $\ell_{src} \to \ell_{pvt} \to \ell_{tgt}$ performs better than the direct translation. These results show that the triangle inequality does not hold in the case of neural machine translation.

Further, this study explores the symmetric properties of machine translation. Some might expect translation from ℓ_1 to ℓ_2 to be equally good as translation from ℓ_2 to ℓ_1. We show that, although the two translations will achieve similar performance, one direction will almost always

exceed the other, meaning that language translation is also not symmetric. However, choice of pivot language is symmetric in almost all cases. For example, the direct fr-en and en-fr CSLS scores are 0.72 and 0.74 respectively. This is because Chen & Cardie's (2020) UMWE algorithm aligns French to English in the former case and English to French in the latter. These are two different rotations, yielding two different performances. However, the best pivot language for fr-en is Spanish, and the best pivot language for en-fr is also Spanish. When using Spanish as a pivot language, both fr-es-en and en-es-fr obtain the exact same CSLS score of 0.76. This is because the UMWE algorithm aligns all languages to the same space, resulting in two copies of the exact same mapping.

Finally, the performance of the Bellman-Ford algorithm in our second experiment suggests that CSLS scores are not independent of each other; multiplying scores together will not yield the same score as when the pivot language is introduced in our first experiment. This results in a less sensitive model that can only identify extreme pivot language opportunities.

6.2. Linguistic

Given the history of translation, the results in this paper may feel counterintuitive. Human translators that are experts in a source and target language would always translate directly rather than passing through a pivot language. This study, however, provides evidence that pivot languages are useful in machine translation and will yield a better result than the direct translation almost every time. Further, this study shows that English is not the best pivot language to use in general. This contradicts the methods of many computational linguists who choose to pass through English due to its accessibility and availability of data.

6.3. Cultural

Languages are constantly changing. In 2004, Warnow et al. developed stochastic models to demonstrate how languages change over time and draw statistical inferences to elucidate linguistic history better than purely traditional means have been able to. Languages also die. This occurs when a language loses its last native speaker. Finally, languages can become extinct. This occurs when the language is no longer known, including by second-language speakers. NMT enables translation between languages without requiring knowledge of either language. This study shows that NMT can be improved through the use of pivot languages. As such, this study has long term cultural implications in language preservation; given adequate text data, dying languages can still

be translated to and from. The use of a pivot language can ameliorate this translation to improve linguistic and hence cultural preservation.

7. Limitations

The main limitation of this study is that it only considers six languages. Further, these languages are all of Indo-European descent and of Subject-Verb-Object typology. However, word embeddings contain a plethora of information about the word and are context-aware. For example, the embeddings contain context about the word's case, plurality, and role (noun, adjective, etc.). Further, embeddings' context awareness means that polysemy is handled well. For example, the same word can be represented by several different embeddings depending on the context in which it is used. Also, NMT algorithms have been shown to handle distant language pairs well. Finally, in this study, UMWE has shown that a pivot language is almost always better than the direct translation. Due to this, I would expect the methodologies used and results obtained in this study to be generalizable to many other languages.

There are, however, three main sources that would hinder the extension of this study to many other languages. They are summarized below, and visualized in Figure A31

First, the availability of bilingual dictionaries. The rapid development of cross-lingual word embedding methods has not been met with adequate progress in their evaluation (Glavas et al., 2019). Although NMT can learn high quality translations using only the vector-representations of words, a dictionary is needed to evaluate the final translation. Further, these dictionaries must all be generated in a consistent manner using bilingual lexicon induction to be comparable. Currently, there is a lack of appropriate dictionaries, which means that we have no way to evaluate the neural machine translation between many language pairs.

Second, training and evaluating models is time consuming. It took my computer on average seven hours to learn and evaluate the mapping of one language pair. As such, if we have n languages, the total computational time will be $7(n^2 - n)$. Therefore, if we had every dictionary we need, and were limited by the 157 languages that have high quality vector representations, we would require roughly 19 years to learn the optimal mapping for every language pair. Although this can be ameliorated through cloud computing solutions, not all researchers have access to this resource.

A final limitation is that NMT techniques only apply to written languages with available text

data on the internet. Languages spoken in poorer countries lack text on the internet and are deemed "low- resource". Although translation algorithms have shown good performance with low-resource languages, creating high quality word embeddings proves challenging, resulting in poorer translation quality. Further, languages that are primarily spoken, such as indigenous languages, cannot be used in NMT at this time.

8. Further Research

Primarily, this study should encourage the generation of bilingual dictionaries to scale up our experiments with more language pairs. With the ability to evaluate translation between many distant languages, we can determine which linguistic features in languages influence the choice of pivot languages. One can then ask more nuanced questions such as "a pivot language of which genealogical family would be best for translating Indo-European languages into Asiatic languages" or "a language of which typological structure would be the best pivot for translating SVO languages into SOV languages". This would result in even more impactful linguistic implications.

An interesting area of research would be to analyze why certain language pairs perform poorly in MUSE and hence in UMWE. This study illuminates Portuguese as one of the problematic languages. There have also been user-reported issues when translating from Tagalog. To the best of our knowledge, no research has been conducted on why MUSE fails with these languages, or even what other languages it fails on. Future research could analyze the embedding spaces of these problematic languages and the constructed alignments and mappings used in translation to determine where and why the translation fails.

Finally, it would be beneficial to replicate this study as translation between problematic languages improves. Currently, our results are influenced by the poor performance of Portuguese, which arises due to algorithmic inefficiency. Once this inefficiency is addressed, we could obtain a more true understanding of machine translation between language pairs to continuously update our understanding of pivot languages.

9. Conclusion

This study leverages neural machine translation to align word embeddings of source and target languages to the space of the pivot language. Then, through two experiments, we determine which pivot (if any) will yield the best CSLS score for a fixed source and target. From our first experiment

using UMWE, we draw three conclusions. First, translation from Portuguese to any other language benefits 58-75% from using English as a pivot language. Second, translation from any language to Portuguese does not benefit from any pivot language and favours the direct translation. Third, translation between any pair of languages that does not involve Portuguese benefits 1-7% from the use of a pivot language. The stark improvements for Portuguese are explained by the MUSE algorithm, which UMWE is based on, having poorer performance with the Portuguese language. Our second experiment, using the Bellman-Ford algorithm, argues that translation should be direct for all language pairs except for those where Portuguese is the source. By passing through English, we can achieve a 9-21% improvement in score. However, since the Bellman-Ford algorithm was not sensitive enough to detect all pivot language opportunities, we argue that it should not be used in this case. The results of this study contradict commonly held beliefs about machine translation, including transitive and symmetric properties. Further, we show that English, the most widely used pivot language in industry, performs much worse than other potential pivot languages. Our results should encourage the generation of more evaluation dictionaries to scale up our experiments to distant language pairs. Then, we could infer which linguistic properties make some pivot languages perform better than others.

10. References

[1] Abney, S.P. (2002). "Statistical Methods and Linguistics". *MIT Press*. DOI: 10.7551/mitpress/1507.003.0003

[2] Artetxe, M., Labaka, G., Agirre, E. (2017). "Learning bilingual word embeddings with (almost) no bilingual data". *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. DOI: 10.18653/v1/P17-1042

[3] Brownnlee, J. (2019). "A Gentle Introduction to Neural Machine Translation". *Deep Learning for Natural Language Processing*. Retrieved from: https://machinelearningmastery.com/introduction-neural-machine-translation/

[4] Cao, S., Kitaev, N. & Klein, D. (2020). "Multilingual Alignment of Contextual Word Representations". *International Conference on Learning Representations*. Retrieved from: https://doi.org/10.48550/arXiv.2002.03518

[5] Chen, X. & Cardie, C. (2018). "Unsupervised Multilingual Word Embeddings". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. DOI: 10.18653/v1/D18-1024

[6] Conneau, A., Lample, G., Ranzato, M.A., et al. (2018). "Word Translation Without Parallel Data". *International Conference on Learning Representations*. Retrieved from: https://arxiv.org/pdf/1710.04087.pdf

[7] Dorrel, D. & Henderson, J.P. (2020). "Classification and distribution of languages". *LibreTexts Social Sciences*. Retrieved from: https://socialsci.libretexts.org/

[8] Garg, A., & Agarwal, M. (2018). "Machine Translation: A Literature Review". Retrieved from: https://arxiv.org/abs/1901.01122

[9] Glavas, G., Litschko, R., et al. (2019). "How to (Properly) Evaluate Cross-Lingual Word Embeddings". *Association for Computational Linguistics*. DOI: 10.18653/v1/P19-1070

[10] Grave, E., Bojanowski, P., Guta, P., et al. (2018). "Learning Word Vectors for 157 Languages". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Retrieved from: https://aclanthology.org/L18-1550

[11] Lian, X., Jain, K., et al. (2020). "Unsupervised Multilingual Alignment using Wasserstein Barycenter". *International Joint Conferences on Artificial Intelligence Organization*. Retrieved from: https://doi.org/10.24963/ijcai.2020/512

[12] Mikolov, T., Le, Q.V. & Sutskever, I. (2013). "Exploiting Similarities among Languages for Machine Translation". *Computing Research Repository.* Retrieved from: https://arxiv.org/pdf/1309.4168.pdf

[13] Piperski, A. (2014). "An application of graph theory to linguistic complexity". *Yearbook of the Poznań Linguistic Meeting.* DOI: 10.1515/yplm-2015-0005.

[14] Sennrich, R. & Haddow, B. (2016). "Linguistic Input Features Improve Neural Machine Translation". *Proceedings of the First Conference on Machine Translation.* DOI: 10.18653/v1/W16-2209.

[15] Smith, S.L., Turban, D.H.P., et al. (2017). "Offline Bilingual Word Vectors, Orthogonal Transformations, and the Inverted Softmax". *International Conference on Learning Representations.* Retrieved from: https://openreview.net/pdf?id=r1Aab85gg

[16] Warnow, T., Evans, S.N., et al. (2004). "Stochastic Models of Language Evolution and an Application to the Indo-European Family of Languages". *Berkeley Department of Statistics.* Retrieved from: https://www.stat.berkeley.edu/users/evans/659.pdf

[17] Zellig S. Harris. (1954). "Distributional Structure". *WORD.* DOI: 10.1080/00437956.1954.11659520

11. Appendices

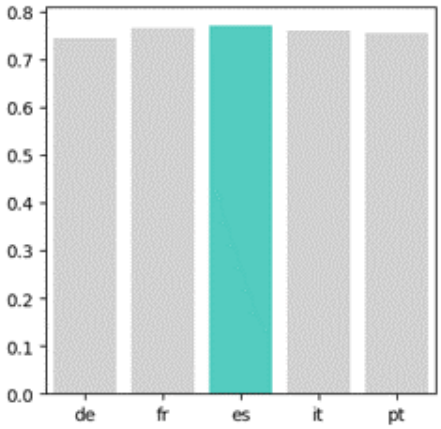


Figure A1. CSLS by pivot when translating from English to German.

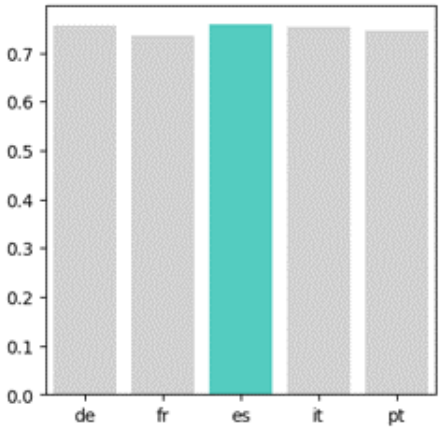


Figure A2. CSLS by pivot when translating from English to French.

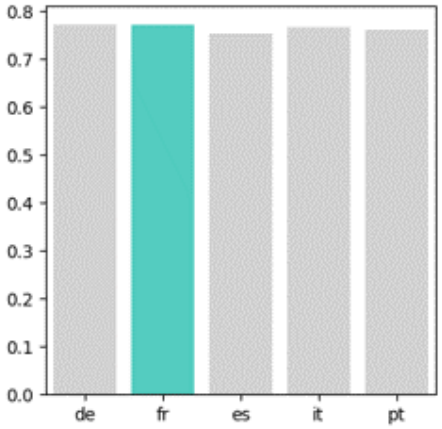


Figure A3. CSLS by pivot when translating from English to Spanish.

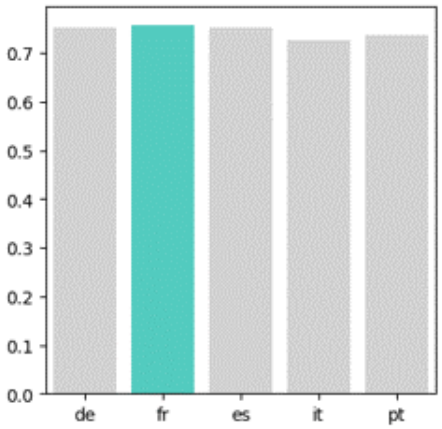


Figure A4. CSLS by pivot when translating from English to Italian.

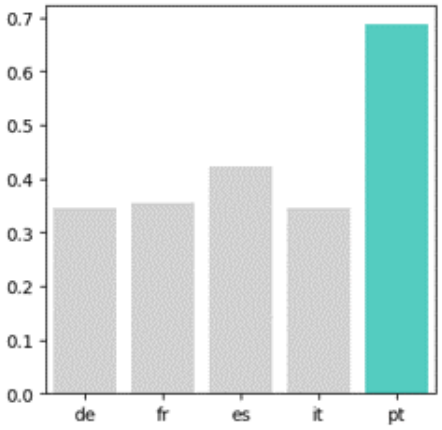


Figure A5. CSLS by pivot when translating from English to Portuguese.

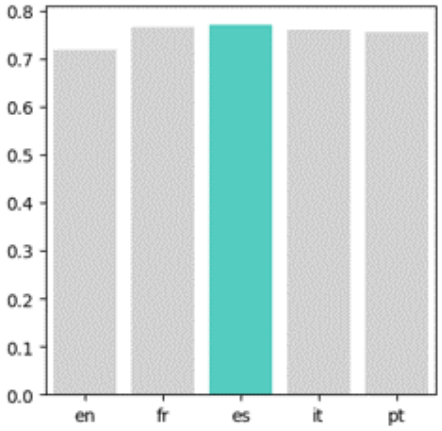


Figure A6. CSLS by pivot when translating from German to English.

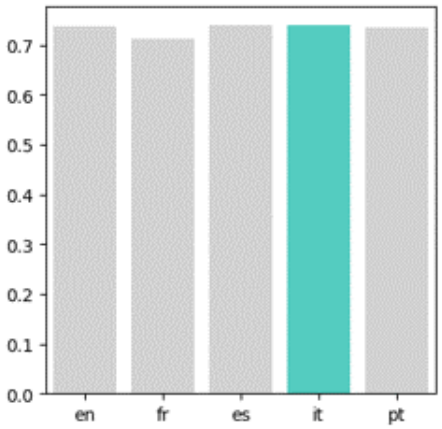


Figure A7. CSLS by pivot when translating from German to French.

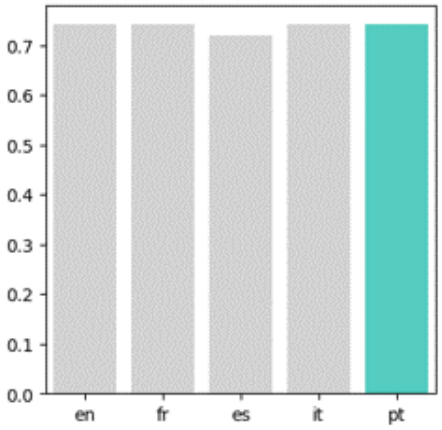


Figure A8. CSLS by pivot when translating from German to Spanish.

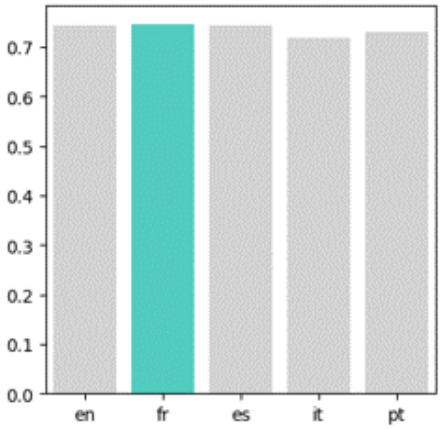


Figure A9. CSLS by pivot when translating from German to Italian.

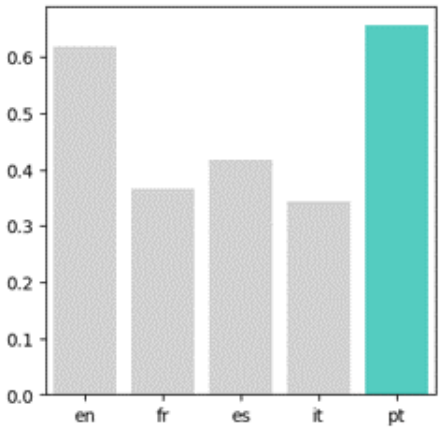


Figure A10. CSLS by pivot when translating from German to Portuguese.

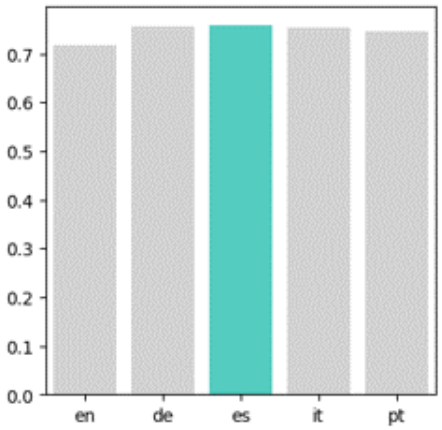


Figure A11. CSLS by pivot when translating from French to English.

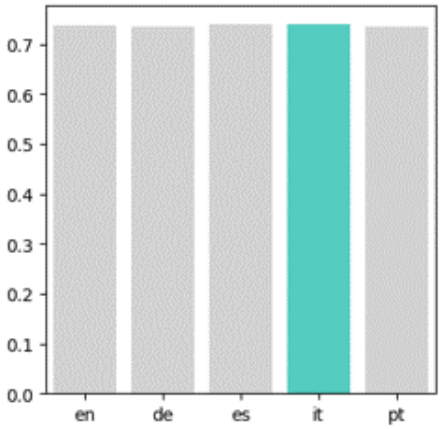


Figure A12. CSLS by pivot when translating from French to German.

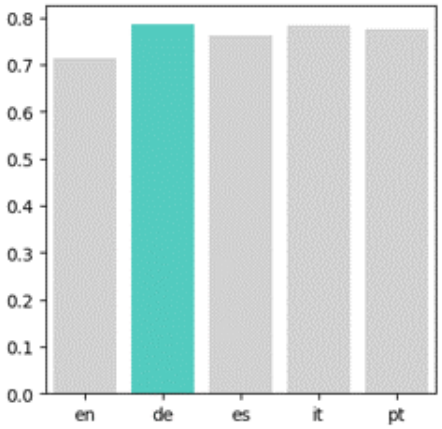


Figure A13. CSLS by pivot when translating from French to Spanish.

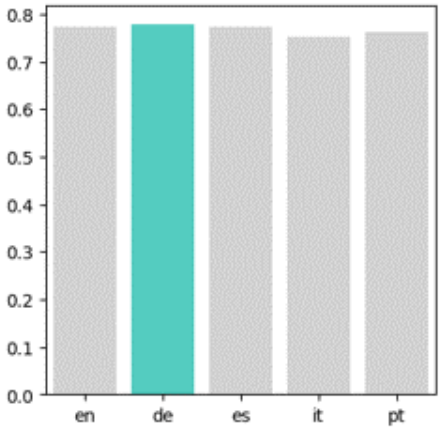


Figure A14. CSLS by pivot when translating from French to Italian.

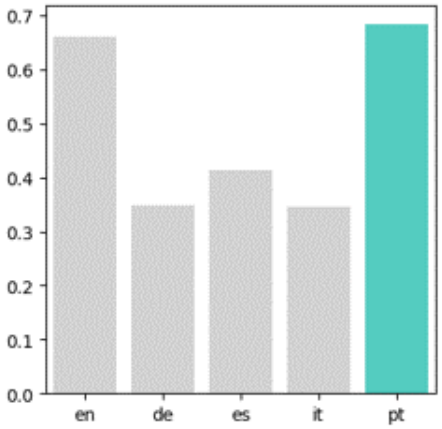


Figure A15. CSLS by pivot when translating from French to Portuguese.

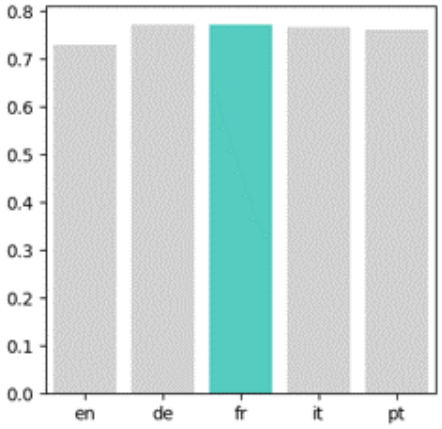


Figure A16. CSLS by pivot when translating from Spanish to English.

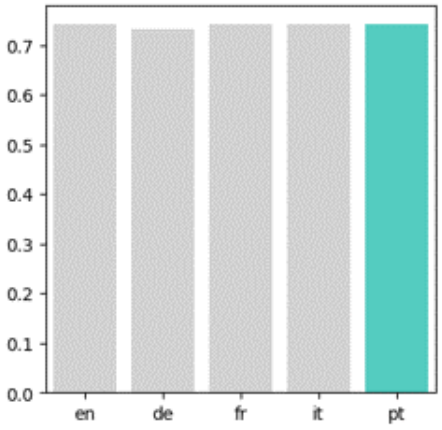


Figure A17. CSLS by pivot when translating from Spanish to German.

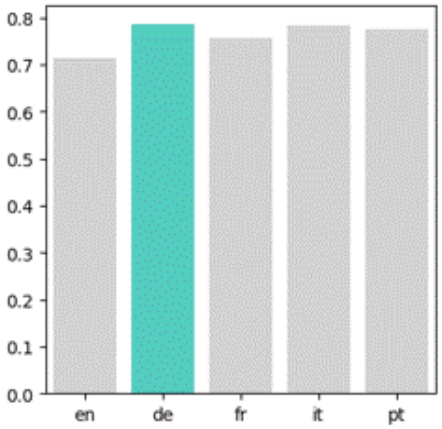


Figure A18. CSLS by pivot when translating from Spanish to French.

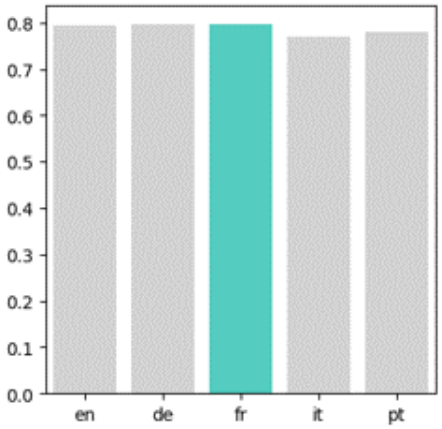


Figure A19. CSLS by pivot when translating from Spanish to Italian.

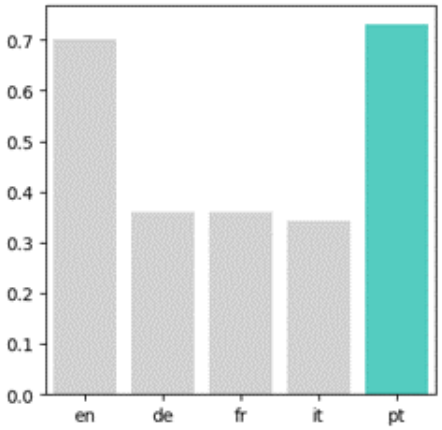


Figure A20. CSLS by pivot when translating from Spanish to Portuguese.

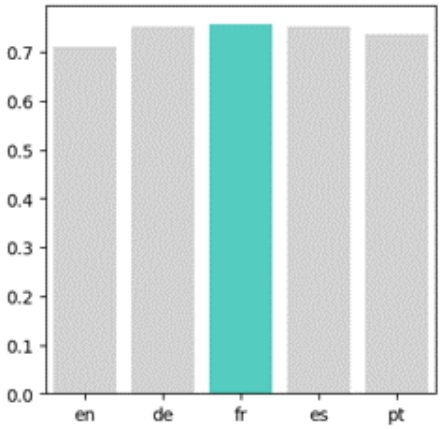


Figure A21. CSLS by pivot when translating from Italian to English.

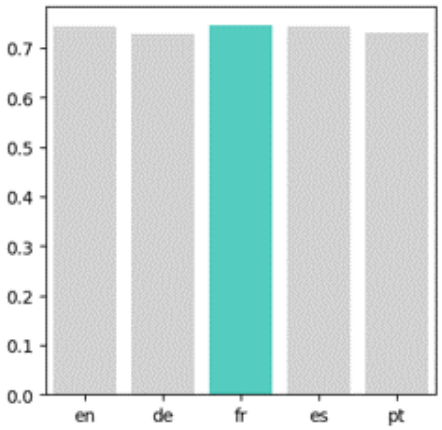


Figure A22. CSLS by pivot when translating from Italian to German.

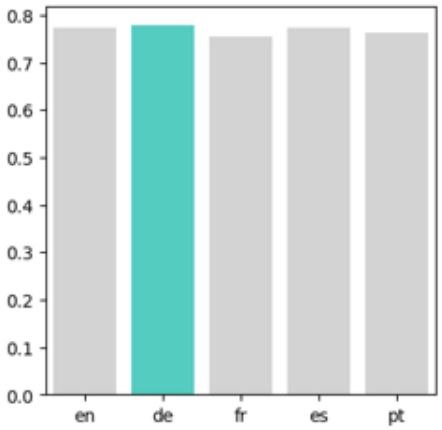


Figure A23. CSLS by pivot when translating from Italian to French.

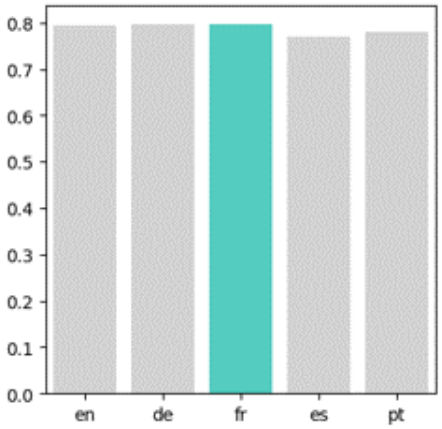


Figure A24. CSLS by pivot when translating from Italian to Spanish.

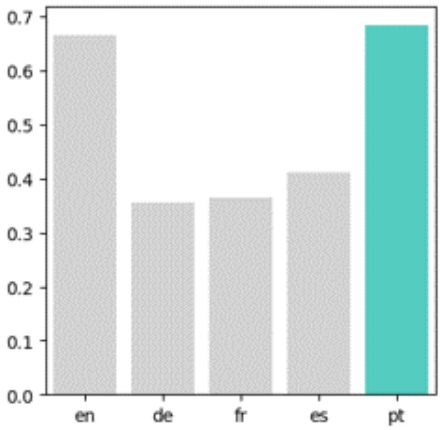


Figure A25. CSLS by pivot when translating from Italian to Portuguese.

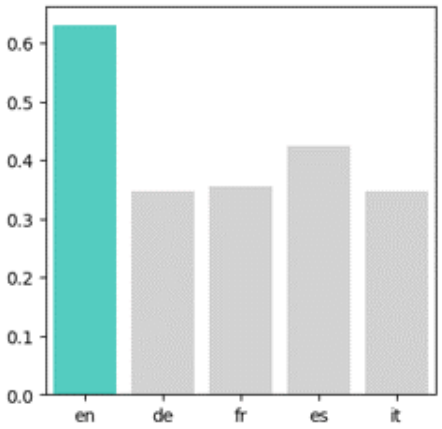


Figure A26. CSLS by pivot when translating from Portuguese to English.

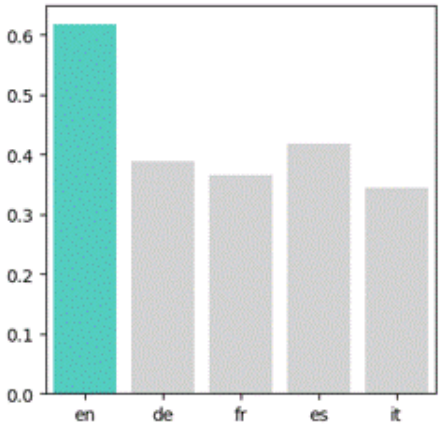


Figure A27. CSLS by pivot when translating from Portuguese to German.

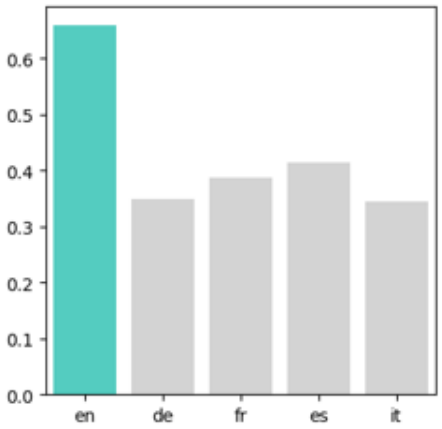


Figure A28. CSLS by pivot when translating from Portuguese to French.

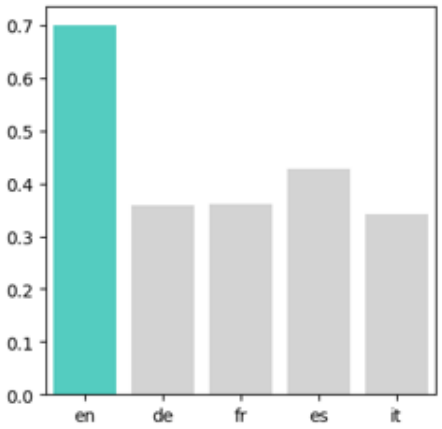


Figure A29. CSLS by pivot when translating from Portuguese to Spanish.

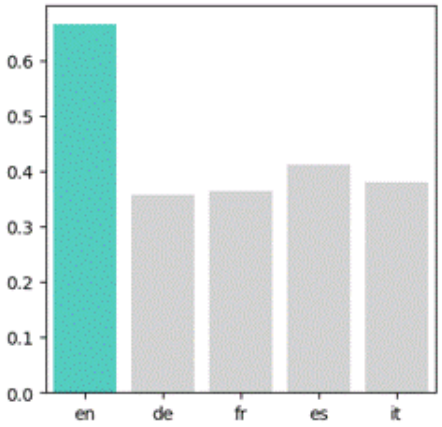


Figure A30. CSLS by pivot when translating from Portuguese to Italian.

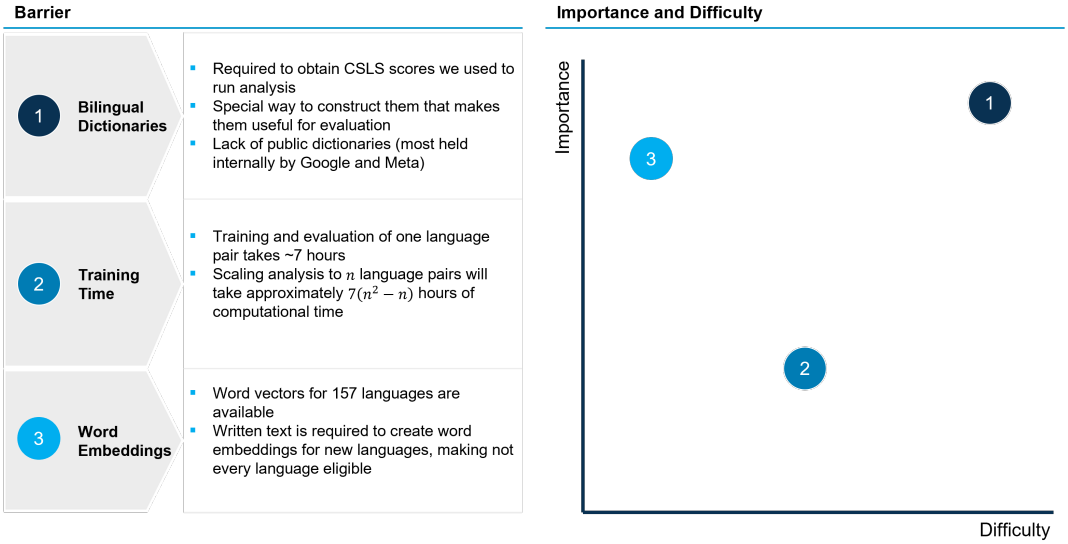


Figure A31. Importance and difficulty to surmount of different barriers hindering the extension of this study.

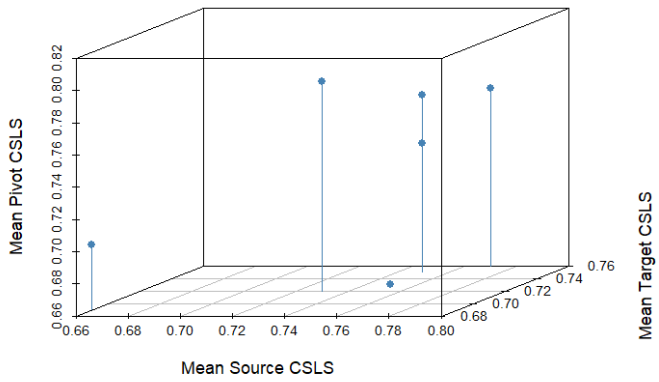


Figure A32. A prototype of how a language's performance as a source and target may affect the extent to which it improves translation when used as a pivot.