The Machine Translation Apple Does Not Fall Far from the Language Family Tree

Adam Deng
adamdeng@mit.edu

Cerine Hamida
chamida@mit.edu

Maria Christina Kalogera
marchka7@mit.edu

Abstract

Machine Translation (MT) models are not perfect: they often fail to capture information about syntactic, contextual, and morphological features of translation languages. Such insights can be potentially incorporated via language families (LFs). Our paper is an analysis of State-of-the-Art (SotA) MT models' accuracies with respect to whether languages of translation belong in the same LF, with the aim of identifying the properties and limits of MT systems and highlighting areas for improvement. Our analysis evaluates the accuracy of direct translations and investigates whether using pivot languages improves translation accuracy (TA).

Data is sourced from the Europarl corpus for 11 European languages representing three LFs and two isolates. Translations between all 110 language pairs are generated using DeepL and Google Translate (GT) and evaluated using 6 metrics measuring a diverse set of language properties. We report three sets of TA data: Raw metric scores, pairwise translations versus pivot translations, and DeepL versus GT. We plot our results through heatmaps and look for patterns within the data. We found that translations to and from English significantly outperform other translations, and the Romance LF performs particularly well. Furthermore, pivot translations have no significant benefit over pairwise translations with respect to TA.

Our current progress is a gateway for larger-scale research on this topic; we would next include more languages (especially non-European or low-resource), a more robust set of corpuses, and more translation systems.

1 Introduction

Despite major improvements in machine translation (MT), state-of-the-art models underperform because they use traditional accuracy metrics e.g. *n*-gram overlap and neglect to capture information about syntactic, contextual, and morphological features of translation languages. To improve translation, such insights can be incorporated via language families (LF), groupings of languages related through descent from a common ancestral language and that share features such as script, syntax and pronunciation. While translation accuracy (TA) and methodology has gained attention in recent years, there is little existing research on the relationship between TA and the correlation between languages (as defined by their belonging to the same family). In our paper, we analyze the relationship between language similarity with respect to LFs and TA. Furthermore, we evaluate the effectiveness of translation through pivot languages. Through such analysis, we can understand the limitations of current MT models and potentially suggest improvements therefor.

Pivot translating is a recent translation concept that aims to exploit the inherently uneven language understanding of models. If translation from language A to language B performs below expectation, the model might select a well-studied 'pivot' language, and define the aforementioned translation as first translating from language A to a pivot language, then from the pivot language to language B. The hypothesis for pivot translations is that they will perform significantly better than pairwise translations, which simply translate directly from language to language. The premier pivot

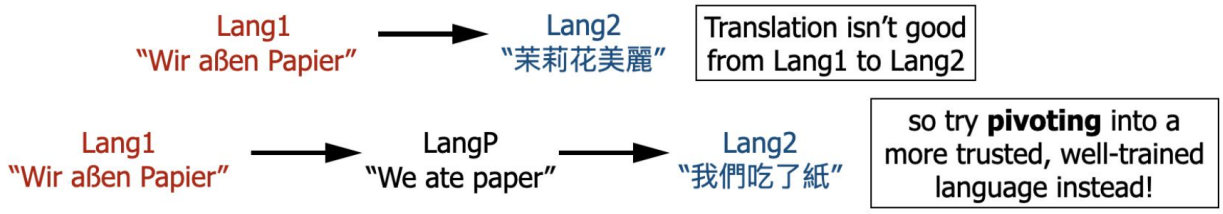

Figure 1: A depiction of pivot translation in action. English serves as the pivot language.

language is English, as it is by far the most well-studied language in linguistics and language research.

We find that there is a correlation between languages' belonging to the same family, and translation accuracy. For instance, translations between Romance languages perform particularly well. Translations to and from certain languages significantly outperform others: translations with English and the Romance LF languages as either source or target languages achieve higher accuracy scores. We also discover that pivot translations do not achieve significantly higher TA than pairwise translations.

2 Related Works

Review of Existing Literature We can divide past related works into Statistical Machine Translation (SMT) and Neural Machine Translation. The former is the old and somewhat outdated approach, while the latter is the modern approach to translation.

In SMT, Bayesian channel models are used to predict conditional probabilities of potential text translations given source text. This method cannot simultaneously capture grammaticality of the generated sentence and assess TA (Alekseyenko et al., 2012). Other SotA models use similarity-distance algorithms as a TA metric; it was found that the effect of distance is correlated with the ability to distinguish translations from a given source language from non-translated using text phonetic and lexical predictors (Gooskens, 2007). Such techniques are effective for determining the orthographic and lexical similarity of languages.

More recently, Neural Machine Translation models like seq2seq models are examples of Conditional Encoder-Decoder Language Models. In the first major exploration into Neural Machine Translation, Britz et al (2017) used the decoder to predict the next word of the target sentence y conditioned on the source sentence x. Neural Translation models also have better performance and employ better use of contextual domains and morphological phrase similarities than Statistical Machine Translation. They are also more efficient: to be optimized end-to-end, a single neural network does not require the users to individually optimize the subcomponents. They also do not require feature engineering and are more scalable, as the same method is implemented for all language pairs. However, neural methods are less interpretable than statistical ones, as users cannot easily specify lexical and grammatical rules/guidelines for translation (Ruder et al., 2017).

To the best of our knowledge, none of the existing research takes into account how incorporating properties of languages in the same family can improve the quality of translations.

3 Methods

3.1 Overview of Methodology

Figure 2 shows our full methodology for this project. From a corpus we select 100 common sentences in 11 languages representing language families and isolates. We then perform pairwise and pivot translations across the 110 language pairs

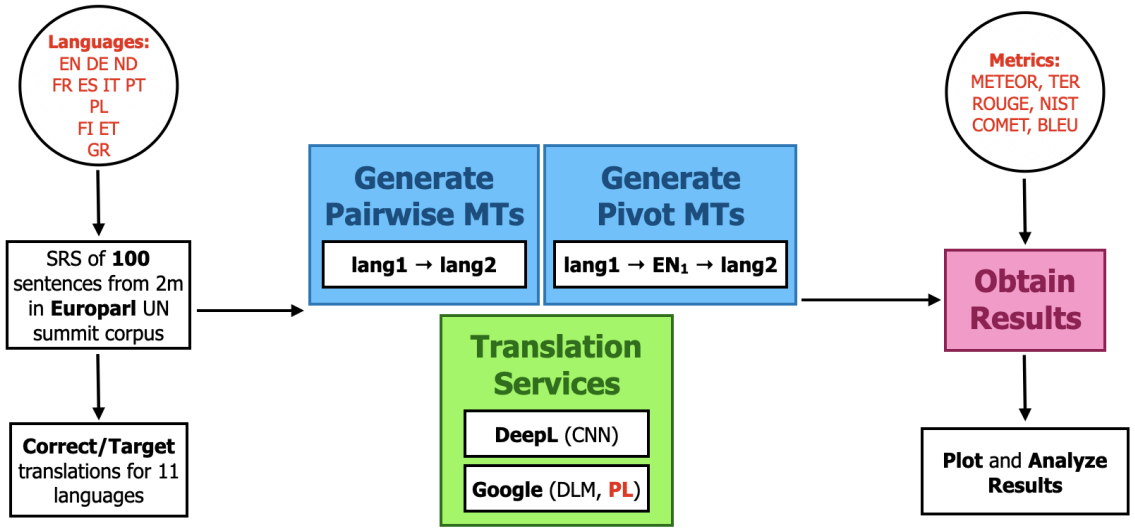

Figure 2: A pipeline diagram of our project.

using DeepL and GT. For each pair, we compute the TA between the source and target sentences using 6 different metrics that measure a variety of features, from n-gram overlap to syntactical similarity. After running these experiments, we visualize our results through three different types of heatmaps and analyze patterns.

The shared Europarl corpus among our 11 languages is our control variable. There are five independent variables: translation source language, translation target language, metric (one of the six), translation system (DeepL or GT), and type of translation (pairwise or pivot). Technically speaking, 'type of translation' is arguably not a true independent variable, as GT is known to employ pivoting; however, it is meaningful to assess whether GT's pivot results correspond exactly to GT's 'pairwise' results.

3.2 Data

Our data is extracted from the "Europarl: A Parallel Corpus for Statistical Machine Translation" dataset, a corpus of spoken words from the proceedings of the European Parliament provided in 21 European languages. We select 11 languages therefrom: French, Spanish, Italian, Portuguese (Romance LF); English, Dutch, German (Germanic LF); Finish, Estonian (Uralic LF); Greek (language isolate) and Polish (Slavic; language isolate for this paper). From this dataset, we randomly select 100 out of the over 600,000 sentences common to all the chosen languages.

3.3 Translation Systems and Computational Resources

We perform our experiments using two translation models: DeepL and GT. DeepL, a newer, proprietary model based in Germany, uses a CNN rather than the traditional RNN and is regarded as a pioneer for the neural network approach to MT. DeepL includes smaller 'fix' mechanisms to make up for the clear deficiencies of a CNN backbone. It is known for being particularly 'human-like' in its translations. On the other hand, the popular GT has been built on neural network-like structures since 2016 to remedy overly-literal translations, and is hailed as a SotA standard for MT. It includes over 100 translation languages, compared with DeepL's 29, and offers full webpage translations. GT is also known to employ pivot translations, while DeepL does not.

For our research, we use Jupyter Notebook extensively, with occasional usage of Google Colaboratory, a cloud-based Python Notebook. For GT translations, we use the Translation API from Google Cloud under Prof. Yoon Kim's group account. For DeepL translations, we leverage DeepL's free API. Calculating the accuracies with

our metrics include adapting the use of several Python libraries: nltk for the BLEU, NIST and METEOR scores; torchmetrics for the TER and ROUGE scores; and comet for the COMET score. Throughout the project pipeline, multiple standard Python tools and modules—matplotlib, numpy, and pandas—are also imported. Our exclusive file format is .npy, which offers nearly 50 times faster speed than a .txt equivalent.

3.4 Metrics

We evaluate our results using the BLEU, NIST, TER, ROUGE-L, METEOR and COMET scoring tools. BLEU is an accuracy metric based on n-gram overlap and brevity penalty. NIST is BLEU with some alterations: Whereas BLEU simply calculates *n*-gram precision, adding equal weight to each one, NIST gives greater weight to a rarer *n*-gram. The TER metric evaluates the number of actions required to change a translated segment into one of the reference translations (NB: a lower TER score means a higher performing model). ROUGE-L measures the longest common subsequence between the model-generated and target translations. We compute the *f*-measure scores, which average the ROUGE-L recall (how much of the reference summary the system summary is recovering or capturing) and precision (how much of the system summary was in fact relevant). The METEOR metric is based on the harmonic mean of unigram precision and recall. It also includes measures of stemming and synonymy matching, along with exact word matching. Finally, COMET is a neural framework for training multilingual machine translation evaluation models. In our case, we used a COMET model pre-trained on Direct Assessments and MQM ratings, which are translation assessments created by human translators. The COMET metric should help evaluate how "natural" a translation is, taking into account language constructs that most traditional metrics cannot consider, such as Function Words (FW), Non-verbal Agreement (NVA) and Verb Tense/Aspect/Mood (VT) agreement. All in all, we consider 6 diverse metrics that, together, create sufficient robustness for a comprehensive analysis of language translation.

3.5 Preprocessing

Our Europarl corpus data is obtained as approximately 200MB-large text files. We convert them into a more useful dictionary representation and save them as .npy files. For the technical, granular details on our approach beyond that provided in this paper, please see our GitHub code.

3.6 Generating Translations

We use pairwise translations as our baseline translation system, in which the probabilities of source and target sentences are used to find the best translation. Given a vector *s* decomposed into sequences of words, we translate each of the phrases of source language to the target using phrase translation distributions. Given a vector of source phrases s, we predict the best target translation t using EM algorithm:

$$t_{pred} = argmax_t p(t|s) = argmax_t p(s|t)p(t)$$

Using the phrase translation distributions we translate sequences of words:

$$t_{pred} = argmax_t \sum_{m=1}^{M} \lambda_m h_m(t, s)$$

with h the feature function and λ the weight for each feature function. Finally, the model maximizes a log-likelihood combination of feature functions to choose the best translation.

We implement the DeepL translation software, which uses syntactic tree banks to train CNNs to predict the highest scoring target translation on pairwise source target corpora.

Our second methodology uses pivot translations. We develop another model, the pivot model, with improves word alignments inside phrase pairs. We are given a source phrase *p* from the source language that is connected to the pivot phrase EN in English, and then phrase EN is

connected with target phrase s which is in the target language.

$$P(S|P) = \sum_{EN} P(S, EN|P) = \sum_{EN} P(S|EN, P)P(EN|P)$$
$$= argmax_{EN} P(S|EN)P(EN|P)$$

For implementing the models over k pivot languages, k pivot models are then estimated using linear interpolation. We use these methods for estimating both the phrase translation probability and the lexical weight respectively, which are coefficients α and β.

$$P(s|t) = \sum_{i=1}^{k} \alpha_i P_i(s|t)$$

$$P(s|t, \alpha) = \sum_{i=1}^{k} \beta_i P_i(s|t, \alpha)$$

3.7 Visualizing Results

After translation accuracies are calculated, we use matplotlib to create three sets of heatmaps. The first is a heatmap of TA for each pair given a metric, translation service (DeepL or GT) and translation method (pairwise or pivot). The second is a heatmap of the differentials between pairwise and pivot TA for each pair for a given metric and translation service. The third is a heatmap of the differentials between DeepL and Google Translate TA for each pair for a given metric and translation method. For a 11 by 11 'matrix' of translation pairs, with rows representing start language and columns representing target language, accuracies or differentials thereof are generated. Green was the main color, with darker shades indicating better accuracy (which for all metrics means more positive numbers, except TER, for which the opposite holds).

4 Results

We break our results discussion into three sections: Raw metric scores; Pairwise vs pivot scores; and DeepL vs GT. As a reminder, translations are conducted from row to column language, e.g. a translation from English to Dutch is row 2, column 1 (with 1-based indexing).

4.1 Raw Metric Scores

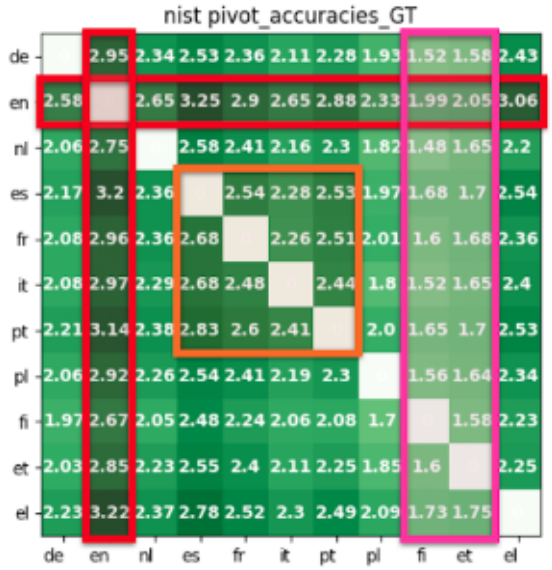

Figure 3: GT pivot translation accuracies with NIST.

Across the board with all metrics, we see that translations to and from English significantly outperform translations without English as neither the source nor target language; in Figure 3, this is the two bands outlined in red. This pattern is observed also in the 'negative metric' TER (Translation Error Rate), in which smaller numbers indicate better performance.

Out of the three language families studied, the Romance LF (outlined in orange in Figure 3) is particularly convincing, with high intra-LF translation accuracies, appearing visually as a 'darkened square' near the middle of the heatmaps. In other words, intra-Romance translations perform better than non-intra-Romance language translations. For the other language families, more languages would have to be used to generate convincing patterns (e.g. including Danish in the Germanic family; the Uralic LF might be extended from Finnish and Estonian to include the distant relative Hungarian).

Systematic TA deficiencies emerge when assessing translations to specific languages using specific metrics. As seen in Figure 3 in pink outline, translations to Finnish and Estonian are of significantly lower accuracy than translations to all other languages. For translations to Greek assessed with ROUGE, performance is numerically only half as good as all other pairwise or pivot translations.

4.2 Pairwise vs Pivot Scoring

With GT, which automatically pivots, one obtains a nearly uniformly ±0.0 accuracy differential between pivot and pairwise accuracy (see Figure 4; note that the shading is almost the same color). As for DeepL, there are slight but statistically insignificant improvements produced by pivot scoring, usually less than +0.02 for each metric (see figure 4). One has evidence that pivot language translations do not significantly improve DeepL.

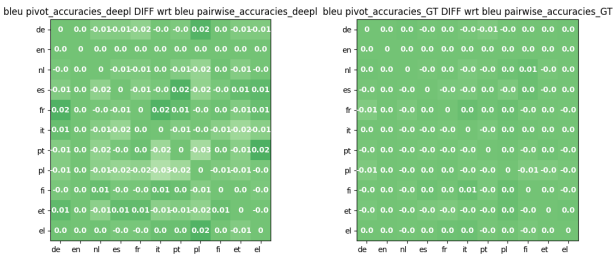

Figure 4: Pivot vs pairwise accuracy differential in GT using BLEU (left); in DeepL (right). The highest-magnitude differential was 0.02.

4.3 DeepL vs GT

DeepL generally outperforms GT across the board, especially in pivot accuracies through NIST (see Figure 5), with TA to Romance languages particularly noteworthy, at +0.20 on average. We find that DeepL almost always performs slightly better than GT across all configurations of independent variables. This is particularly evident using the NIST and COMET metrics. We also notice that the greatest differential occurs with translations to Romance languages, while translations to Dutch and Polish offer the least improvement when switching from GT to DeepL.

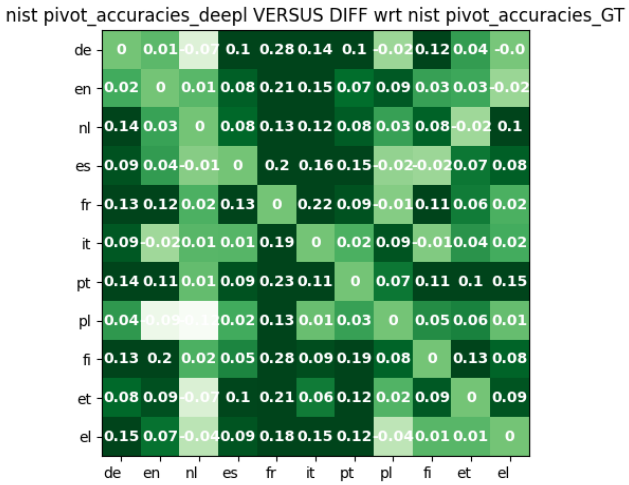

Figure 5: DeepL vs GT performance on pivot accuracies with NIST. Pairwise TA has a similar graph. Deeper color indicates better performance by DeepL over GT.

5 Discussion

5.1 Raw Metric Scores

For future experiments, to more conclusively analyze the impact of LF translations vs non-LF translations, more LFs will be consulted, and existing LFs will be expanded. For instance, the Germanic LF can be expanded to include Danish, while the Finno-Ugric LF can be extended to include Hungarian, a distant relative of Finnish and Estonian.

The systematic translation gaps can be explained by either metrics' being mal-adjusted for certain languages, or the underperformance of translations for certain languages. This was especially obvious for ROUGE with Greek on both DeepL and GT (see Figure 6). As no other metric records such a gap, ROUGE's applicability to Greek is called into question. However, for translations to Finnish and Estonian, multiple metrics (NIST, TER, METEOR, and BLEU) indicate underperformance, while COMET indicates overperformance (see the pink outline at Figure 3). As almost all metrics suggest some sort of systematic performance imbalance, it is more likely that MT systems perform worse with Finnish and Estonian, as

opposed to this being the unreliability of the metrics.

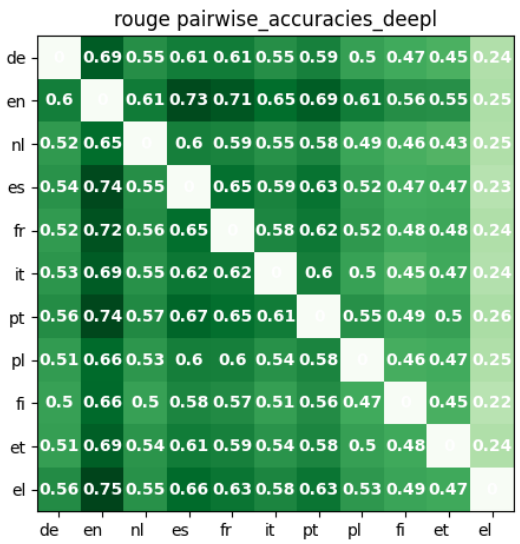

Figure 6: ROUGE DeepL pairwise TAs. Note the abysmal performance of translations to Greek. With respect to earlier analysis, the strong performance by English and the Romance LF can be observed.

To resolve the translation gap issue, one can manually inspect translations, e.g. one by one, selecting sentences and assessing translation quality to see what errors are observed.

Although it is tempting, we do not combine all 6 metrics together for a 'supermetric' for several reasons. The concept of one metric for translation accuracy is unattainable and will inevitably either leave out critical aspects of translations or dilute them to the point of meaninglessness. Furthermore, there is no rigorous or mathematically provable justification for any arbitrary weighting of the metrics.

5.2 Pairwise vs Pivot Scoring

Translation pivoting fails to significantly improve DeepL performance. It appears DeepL's neural network translation already accounts for the 'human-like' quality of translations, making translations between any two languages particularly understandable, ergo reasonably good. On the contrary, GT's translations are regarded as more granular and word-for-word, meaning pivoting would substantially improve otherwise cumbersome translation.

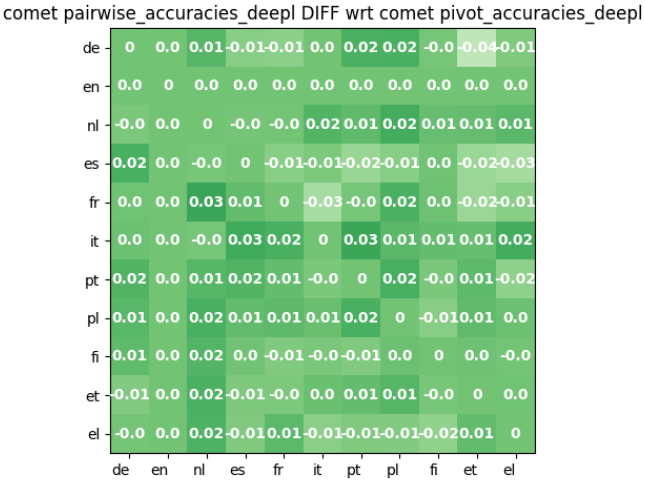

Figure 7: COMET DeepL pairwise vs pivot TAs. The highest improvement was only +0.03, recorded with French to Ducth and Italian to Portuguese.

However, pivot languages may be useful for translations involving low-resource and understudied languages, such as Native American languages (e.g. Cherokee) and African languages (e.g. the Bantu family). As DeepL does not offer such options (and is in general quite restricted, with under 30 languages offered at the time of writing), and GT struggles with such languages, it may be wise to pivot to English in translations.

5.3 Summary of Findings

Our paper yielded insights on translation through LFs and pivot languages. We found that the Romance LF performs better than average, that pairwise translations involving English perform exceptionally well, and that certain languages' translations underperform or metrics are unfit for accuracy measurement. We also found that pivoting in DeepL had no significant improvement on translation accuracy, and that DeepL translations achieve better TA than GT.

6 Impact Statement

Our work—which is fundamentally analysis-driven—examines part of the puzzle of TA.

The natural continuation of this project is expansion of scope. In its current form, the project is restricted by translation languages (our 11 are all Indo-European and mostly well-studied) and translation systems (just GT and DeepL). Thus, we suggest the following avenues upon which to expand our research. First, we can be more encompassing with our LFs, involving more languages per family. We should test non-European languages, especially Chinese and Arabic, though finding a suitable corpus to serve as the control variable will be difficult, as it is very hard to find a database with e.g. English, Chinese, and Arabic sentences, let alone for all our languages. Furthermore, more translation systems can be employed to increase our confidence for consensus, such as Yandex Translate and Microsoft Translate. Doing so could provide us more statistical confidence in our assessment of metric impropriety versus MT deficiency; in other words, if we see an unusual pattern, we can more easily infer if it is an unsuitable metric or simply MT model inaccuracies if we had more MT models.

The greatest societal impact of our work would be its natural extension into low-resource languages, particularly Native American (e.g. Cherokee) and African LFs (e.g. Bantu). By analyzing deficiencies in MT with respect to such languages, we can turn those insights into MT model improvements to make translation models more robust and accurate, so that they represent a broader and more complete spectrum of human linguistics. We posit that, by understanding the mechanisms and logic of diverse LFs, especially understudied ones, we can improve TA for languages in general. By including low-resource languages in our greater analysis, we would also preserve their heritage, a critical issue as such languages are in danger of extinction.

Once exploration of the above problem is reasonably complete, we can also pursue a 'reverse experiment', as hinted in our project proposal. For that project line, when a translation is performed from an unknown source language to a known target language, we can analyze the metrics of translation to attempt to deduce the source language (this is in similar spirit to guessing an ESL speaker's country of origin). Such a project is interesting because there is little existing research on this question. A potential idea for analysis is to use K-means clustering on existing ‘training’ language pairs for which both the source and target languages are known. This project also has the potential to advance understanding of low-resource languages, as including them in such research creates links between well-understood/researched languages like English and Spanish and these low-resource languages. Furthermore, evaluation and analysis of this reverse experiment can potentially identify flaws in current MT systems and lead to suggestions on improvements to such systems, just as was the extended/eventual purpose of our paper as well.

Acknowledgements

Thank you to the 6.8611 course staff for answering our questions regarding NLP research. A special thanks to Prof. Yoon Kim for providing us access to the GT API and offering advice and directions for LT research; Dr. Michael Maune for insights on linguistics and writing; and Saaketh Vedantam, our Teaching Assistant.

Footnotes and References

[1] Arabic is not a language isolate, as Hebrew and Amharic are in the same (Semitic) family, but in our research, it is the lone Semitic language.

[2] Obtained from a page on DeepL's justification, https://www.deepl.com/en/why-deepl-pro.

Alexander V. Alekseyenko, Quentin D. Atkinson, Remco Bouckaert, Alexei J. Drummond, Michael Dunn, Russell D. Gray, Simon J. Greenhill, Philippe Lemey, and Marc A. Suchard. 2012. *Mapping the*

Origins and Expansion of the Indo-European Language Family. Science, 337(6097):957–960.

Charlotte Gooskens. 2007. *The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages*. Journal of Multilingual and Multicultural Development, 28(6):445.

Francois Barbançon, Steven N. Evans, Luay Nakhleh, Don Ringe, and Tandy Warnow. 2013. *An Experimental Study Comparing Linguistic Phylogenetic Reconstruction Methods.* Diachronica, 30(2): 143 – 170.

Britz, Anna Goldie, Minh-Thang Luong, Quoc Le. 2017. *Massive Exploration of Neural Machine Translation Architectures"*, https://arxiv.org/abs/1703.03906.