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

Adam Deng, Cerine Hamida, Maria Christina Kalogera

Despite major improvements in machine translation, state-of-the-art models often use traditional accuracy metrics such as *n*-gram overlap, failing to capture information about syntactic, contextual and morphological features of translation languages. Such insights can be incorporated via language families, groupings of languages related through descent from a common ancestral language and that share features such as script, syntax and pronunciation. While translation accuracy and methodology has gained attention in recent years, there is little existing research on the relationship between translation accuracy and the correlation between languages, as defined by their belonging to the same family. In our paper, we explore the relationship between language similarity with respect to language families and translation accuracy by translating two languages both ways. We advance that understanding such relationships can provide machine translation researchers with insights on how to improve model accuracies.

Review of Existing Literature We can divide past related works into Statistical Machine Translation and Neural Machine Translation. Historically, noisy Bayesian channel models were used to predict conditional probabilities of translated text given source text. But this method cannot simultaneously capture the grammaticality of the generated sentence and assess the correctness of translation (Alekseyenko et al., 2012). Other state-of-the-art models use similarity-distance algorithms to measure the distance between languages on the translation product and concluded that the effect of distance is directly 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. To obtain more accurate results, Barbançon et al. (2013) investigated etymons[1] and cognates, identifying them, computing distances between related words with frequency support from the corpus, and then measuring the overall degrees of similarity between pairs of languages.

More recently, Neural Machine Translation models such as sequence-to-sequence 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).

None of the existing research takes into account how incorporating properties of languages in the same family can improve the quality of translations.

Research Overview We quantify the relationship between translation accuracy and language relatedness by computing correlations. This allows us to assess reliability of translations, as well as obtain insights into the nature of linguistic similarity between languages. Such insights can then be used in translation models to improve performance—in particular, causation for translation errors can be evaluated. The project involves selecting several languages and listing a large number of selected sentences in each of those languages. Then, a state-of-the-art translation system will translate a certain sentence in language A into language B, and vice versa. The machine translation will be assessed against the target (correct) translation. Based on translation accuracy, results can be plotted.

Parameters of Study The 13 languages we study are English, German, Dutch, Spanish, French, Italian, Russian, Ukrainian, Czech, Mandarin, Arabic, Finnish, and Greek. These languages are all represented by most, if not all, of the translation services; they are also well-represented by corpora. These constitute 3 language families of 3 languages each. There are 4 language isolates[2]. We will utilize the test set corpora provided by the Association for Computational Linguistics workshops on Statistical Machine Translation which is divided into language pairs e.g. Chinese-English in both directions (https://www.statmt.org/wmt14/index.html; this is the largest test set of its kind, and, as it was produced during a summit, contains a large amount of reference data. For training, we employ the Europarl corpora (https://www.statmt.org/europarl), which is closest in format to the desired sentence-sentence format and requires the least preprocessing.

Multiple translation services—broadly considered among the best—will be used. DeepL will serve as the primary translator; it is described as "captur[ing] even the slightest nuances and reproduc[ing] them in translation"[3]. Google Translate and Microsoft Translator are also employed. The three translators use different methodologies, useful for translation corroboration and avoidance of systematic biases.

To evaluate the accuracy of translations, multiple existing **metrics** will be used: the standard 'core four' of BLEU, NIST, TER, and METEOR, in addition to Yisi, MEE, COMET21, and ARC. While BLEU, NIST, TER and METEOR are mainly based on n-gram similarity, Yisi, MEE, COMET21, and ARC also measure morphological and semantic similarity and naturalness. The combination of these metrics should make our analysis less prone to systematic errors.

To perform training, testing, and assessment of results, we will use Google Colab (Python Notebook) for shared work, and locally-hosted Jupyter Notebook for individual experimentation. Python modules, like matplotlib, numpy, and pandas, will be used for plotting results and calculating correlations.

Benefits to Machine Translation By understanding translation errors, and which language-specific constructs caused them, we can identify improvements for translation. For instance, if the same pattern of pronoun errors was made across all languages, we would flag it in our paper, and researchers could study this direction further. Additionally, we can better assess existing or propose new linguistic similarity metrics to examine translation quality. Indeed, using a robust set of metrics for training and evaluating models is not only important for quantifying model performance, but also for identifying areas of improvement. Further exploration in the relationship between language families and translation accuracy can also help improve the quality of translations for languages well-studied in the linguistics field but not by NLP research community — in particular, indigenous African and Southeast Asian languages—which allows for the building of translation systems that avoid systematic biases common in models constructed without training on such languages.

The success of this project does not depend on the success of the translations. Rather, the project is successful insofar as it provides new insights into translation accuracy, language similarity, and their relationship, and pushes the frontier of machine translation further.

Further research can also be performed with the reverse purpose; that is, to determine accuracy of existing language families, or create new language families. One experiment is measuring translation accuracies for languages A and B, in which at least one of A and B is an unclassified or classification-debated language. This can yield insight into how language families should be created, as well as how accurate existing families are, e.g. whether Japanese is an Altaic language, the strength of the category "Finno-Ugric" (Finnish and Hungarian together), or how Southeast Asian languages with similar scripts should be classified.

Footnotes and References

[1] ancestral words which are the root for more modern words

[2] 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.

[3] From 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.