Ugarit: Translation Alignment Technologies for Under-resourced Languages

Workshop presented at DH2022 Tokyo

Chiara Palladino[1], Tariq Yousef[2], Farnoosh Shamsian[2], and Nadia Kanagawa[1]

[1] Furman University, USA
[2] Leipzig University, Germany

In this workshop[3], participants are going to learn the fundamentals of Translation Alignment and learn to use UGARIT[4], an online environment targeted at the creation of manually aligned datasets in different languages. The goal of the workshop is to introduce participants to an important topic in Digital Humanities, and to expand our community and available datasets by targeting East Asian languages and Japanese in particular.

Translation alignment is one of the most important tasks in Natural Language Processing. It is defined as the comparison of two or more texts in different languages, also called parallel texts or parallel corpora [5,9], by means of automated or semi-automated methods. The result often takes the form of a list of pairs of items, which can be words, sentences, or larger text chunks like paragraphs or documents. The aligned pairs may be one-to-one (one word in the source text corresponds to one word in the translation), but often align as one-to-many, many-tomany, or many-to-one. Each word correspondence may be complete or perfect (with complete overlap between two words), but also possible or incomplete (partial overlap, or both words being a translation of each other only in certain contexts [4]).

There are numerous methods for automated translation alignment: the most popular ones, such as statistical machine translation, are based on various levels of manually aligned training data [2], although new models are being proposed, such as neural machine translation [1]. However, the alignment of texts in different languages is an exceptionally complex task, especially when considering word-level alignment. It is often difficult to find perfect correspondences across languages that express ideas through different morphosyntactic constructs, with variations in word order, sentence length. In addition, it is notoriously difficult to establish correspondences within wordplays, metaphors, or allusions. For these reasons, manually aligned word pairs are extremely important to establish gold standards, as sources of training data to implement machine translation methods, and for many other purposes, including text mining and creation of dynamic lexica [4,8].

Some modern languages, like English, German, and Chinese, have an impressive infrastructure for managing parallel corpora. However, that is not the case for historical and generally under-resourced languages, such as Classical Arabic,

[3] https://dh2022.adho.org/workshops-and-tutorials/wt-11
[4] https://ugarit.ialigner.com/

Persian, Latin, Ancient Greek, Gaelic, Cherokee, Georgian, and even for many languages of East Asia, including Japanese, Korean, and Sanskrit. UGARIT is a web-based environment designed to support the needs of these languages, providing a framework for creating and using manually aligned corpora. During the workshop, we will introduce the tool and illustrate the many ways in which parallel corpora aligned with UGARIT are currently used: these will include pedagogy and language learning, interlinguistic and translation analysis, dynamic visualization, data mining, dynamic lexica, and training of machine translation models [3,6,7,10]. We will invite the participants to test the tool on their own corpus or with a prepared dataset, to try firsthand the work of translation alignment, and to visualize and investigate the results.

UGARIT currently supports most East Asian languages and alphabets, but there are very few aligned datasets currently available. With this workshop, we want to specifically target the creation of new parallel corpora in Japanese, Chinese and Korean, and gather more feedback and requests from this part of the Digital Humanities community.

Instructors:

Tariq Yousef is a research associate at Leipzig University, working on Computational Linguistics, Textual Alignment, and Data Visualization. He is the Lead developer of Ugarit. Contact: tariq.yosef@uni-leipzig.de .

Chiara Palladino is Assistant Professor of Classics at Furman University. As project partner in Ugarit, she uses the tool in teaching and research and has led multiple workshops and seminars on translation alignment. Her main interest lies in language learning processes with translation alignment. Contact: chiara.palladino@furman.edu .

Farnoosh Shamsian is a PhD candidate at Leipzig University. As a project partner in Ugarit, she uses the tool in teaching and research and has led multiple workshops and seminars on translation alignment. Her main interest lies in digital pedagogy and teaching Greek through digital annotations. Contact: shamsian@informatik.uni-leipzig.de .

Nadia Kanagawa is James B. Duke Assistant Professor of Asian Studies and History at Furman University. She is a Ugarit user who often works with and translates classical Japanese texts in her research on immigrants in the early Japanese state. Contact: nkanagawa@furman.edu .

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational linguistics **16**(2), 79–85 (1990)
3. Foradi, M.: Confronting complexity of babel in a global and digital age. what can you produce and what can you learn when aligning a translation to a language that you have not studied? In: DH2019: Digital Humanities Conference. pp. 9–12 (2019)
4. Graca, J., Pardal, J.P., Coheur, L., Caseiro, D.: Building a golden collection of parallel multi-language word alignment. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (2008)
5. Kay, M., Röscheisen, M., et al.: Text-translation alignment. Computational linguistics **19**(1), 121–142 (1994)
6. Palladino, C.: Reading texts in digital environments: Applications of translation alignment for classical language learning. J. Interact. Technol. Pedagog **18**, 724–731 (2020)
7. Palladino, C., Foradi, M., Yousef, T.: Translation alignment for historical language learning: a case study. Digital Humanities Quarterly **15**(3) (2021)
8. Véronis, J.: From the rosetta stone to the information society. In: Parallel text processing, pp. 1–24. Springer (2000)
9. Véronis, J.: Parallel Text Processing: Alignment and use of translation corpora, vol. 13. Springer Science & Business Media (2000)
10. Yousef, T., Janicke, S.: A survey of text alignment visualization. IEEE transactions on visualization and computer graphics **27**(2), 1149–1159 (2020)