Detecting Fine-Grained Semantic Divergences to Improve Translation Understanding Across Languages

Eleftheria Briakou ebriakou@umd.com
Department of Computer Science, University of Maryland, College Park, MD

Abstract

In this thesis, we focus on detecting fine-grained semantic divergences—*subtle meaning differences in sentences that overlap in content*—to improve machine and human translation understanding.

EN ... *The Maple Leaf Forever* ***served for many years*** *as a Canadian* ***national anthem****...*
FR ... *The Maple Leaf Forever (en) qui est un* ***chant patriotique*** *pro canadien...*
... The Maple Leaf Forever which is a Canadian patriotic song...

1 Introduction

A widespread hypothesis adopted by machine translation research is that a source text and its (human) translation—*or* parallel text—are equivalent in meaning. In principle, this hypothesis drives the way we think about our models when designing our training losses and our evaluation metrics and protocols. Yet when humans translate, they make lexical decisions influenced by cultural and situational aspects of language that break the hypothesis of meaning equivalence in nuanced ways (Hirst, 1995; Zhai et al., 2019). Consider the English and French sentences above drawn from WikiMatrix (Schwenk et al., 2021), a corpus that is routinely used to train translation systems and is perceived as highly parallel. While they share important content, highlighted words convey meaning missing from the other language (i.e., *served for many years*) or content reflecting fine-grained semantic divergences between concepts that, although related, are not equivalent (i.e., *national anthem* vs. *patriotic song*).

Regardless of why such subtle divergences exist in parallel texts, we hypothesize that they matter for machine translation systems—as they yield challenging training samples—and for humans—who might benefit from a nuanced understanding of the source. Thus, in this line of work, we argue that quantifying fine-grained divergences is crucial to **improve both *machine* and *human* translation understanding across languages**.

In what follows, we start by introducing methods for detecting fine-grained divergences in the wild (Briakou and Carpuat, 2020). As we will see, such methods lay the foundation for studying their connection to machine translation models (Briakou and Carpuat, 2021, 2022; Briakou et al., 2022) and human evaluations pipelines (Briakou et al., 2023).

2 Detecting Fine-grained Divergences

In our first piece of work, we start our exploration by asking: ***How frequent are semantic divergences in parallel texts?*** Our goal is to address challenges in detection of fine-grained divergences within bitexts in two settings: *human annotation* and *automatic prediction* (Briakou and Carpuat, 2020).

Starting with human annotation, we contribute the Rationalized English-French Semantic Divergences corpus, based on a novel divergence annotation protocol that exploits rationales to improve annotator agreement. Annotations on the collected dataset reveal that semantic divergences are surprisingly frequent, comprising 40% of samples in a cor-

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

pus consisting of Wikipedia-mined translations, and are perceived as highly parallel.

After establishing that divergences exist, we explore computational methods for detecting them at scale, crucially, without assuming access to gold supervision. To that end, we introduce a contrastive loss designed to make a multilingual language model sensitive to subtle cross-lingual differences between linguistically motivated synthetic samples. Despite being trained only on synthetic samples, we show that our model detects fine-grained divergences accurately, outperforming a strong sentence-level similarity model (Schwenk and Douze, 2017).

3 Improving Machine Translation

Equipped with the tools that allow us to study divergences at scale we now ask: ***How do fine-grained divergences impact Neural Machine Translation?*** We contribute a controlled empirical analysis on several aspect of NMT models that are exposed to different types and amounts of divergences at training time. Our findings reveal that small divergences hurt translation accuracy and confidence of NMT models, and crucially are one of the root causes that lead to neural text degeneration, i.e., translation outputs that are incoherent or get stuck in repetitive loops (Briakou and Carpuat, 2021).

Drawing from those findings, a natural question arises: ***How can we mitigate the negative impact of divergences on* NMT*?*** To this aim, we explore two orthogonal strategies. Our first strategy intervenes in the training assumption of translation equivalence in parallel texts and aims to model divergences explicitly. Drawing from our prior work on automatically detecting divergences, we propose a divergent-aware framework—DIV-FACTORIZED—that incorporates token-level divergence signals into NMT training (Briakou and Carpuat, 2021).

Our second strategy proposes an orthogonal mitigation direction: instead of altering training to model divergences closely, we aim to automatically re-write divergent samples to yield more equivalent translations. In this direction, we introduce two approaches to solve this problem in the lack of supervised data. Our first approach—EQUIV SEM-DIV—relies on synthetic translations, i.e., translations generated by MT, that selectively replace divergent references under a semantic equivalence condition (Briakou and Carpuat, 2022). Our extensive evaluations on both intrinsic and extrinsic tasks for two medium-resource languages show that this approach is capable of revising divergences in parallel texts, given synthetic translations of sufficient quality. In our subsequent work, we address this problem in low-resource conditions via introducing—BITEXTEDIT—an editing-based model that, given a parallel text, edits one of the two references to generate a refined version of the original as necessary. Our editing model is trained solely on synthetic supervision via leveraging recent advances in bitext mining based on massively multilingual sentence embeddings (Artetxe and Schwenk, 2019) and is shown to utilize divergences more effectively in heterogeneous data scenarios (Briakou et al., 2022).

4 Assisting Humans to Detect Translation Differences in Meaning

After exploring how detecting semantic divergences helps us improve machine translation understanding, we finally turn to analyze ways of assisting *humans* in understanding and detecting translation differences. Although detecting divergences in parallel texts as a binary classification task, i.e., equivalence vs. divergence, is found to be sufficient for augmenting and improving NMT, we hypothesize that other task framings that shed more light on the nature of divergences are needed to improve human translation understanding. In this direction, our last piece of work asks: ***How can we explain semantic divergences in a human-interpretable fashion?***

To that end, we equip divergence detectors with the ability to indicate not just *whether* divergences exist but also tell us *where* the translation differences reside (Briakou et al., 2023). Drawing on social science studies, we introduce a method to extract contrastive phrasal highlights that explain the predictions of our divergent detectors by explicitly modeling the relationships between the contrasted texts. We contribute evidence that contrastive phrasal highlights match human-provided rationales of divergence better than standard highlighting approaches, and more importantly, they assist bilingual speakers in annotating fine-grained divergences, easing the need to ask for human rationales. Finally, we show that contrastive highlights could help humans detect critical errors due to local mistranslations in machine-translated texts.

References

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Briakou, E. and Carpuat, M. (2020). Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

Briakou, E. and Carpuat, M. (2021). Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7236–7249, Online. Association for Computational Linguistics.

Briakou, E. and Carpuat, M. (2022). Can synthetic translations improve bitext quality? In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4753–4766, Dublin, Ireland. Association for Computational Linguistics.

Briakou, E., Goyal, N., and Carpuat, M. (2023). Explaining with contrastive phrasal highlighting: A case study in assisting humans to detect translation differences. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11220–11237, Singapore. Association for Computational Linguistics.

Briakou, E., Wang, S., Zettlemoyer, L., and Ghazvininejad, M. (2022). BitextEdit: Automatic bitext editing for improved low-resource machine translation. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1469–1485, Seattle, United States. Association for Computational Linguistics.

Hirst, G. (1995). Near-synonymy and the structure of lexical knowledge. In *AAAI Symposium on Rep- resentation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity. pages 51– 56.*, pages 51–56.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Schwenk, H. and Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. In Blunsom, P., Bordes, A., Cho, K., Cohen, S., Dyer, C., Grefenstette, E., Hermann, K. M., Rimell, L., Weston, J., and Yih, S., editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Zhai, Y., Safari, P., Illouz, G., Allauzen, A., and Vilnat, A. (2019). Towards recognizing phrase translation processes: Experiments on english-french. *CoRR*, abs/1904.12213.