Advances in Natural Language Processing: Bridging Text and Knowledge via Grounding and Innovative Applications

Edoardo Barba and **Riccardo Orlando** and **Pere-Lluís Huguet Cabot**[*] and **Roberto Navigli**
Sapienza NLP Group, Sapienza University of Rome
{surname}@diag.uniroma1.it

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

Nowadays, in the era of massively pretrained Large Language Models (LLMs), the level of *understanding* that Natural Language Processing (NLP) solutions show is impressive. However, in the pursuit of achieving Artificial General Intelligence (AGI), relying solely on vast amounts of raw text for modeling language may present different limitations (Bender et al., 2021). As an example, the Zipfian distributional nature of linguistic phenomena (Kilgarriff, 2004) makes it very hard to model the long tail of rare cases, particularly in the context of less-resourced languages, prompting the exploration of alternative methodologies. An increasingly pertinent approach involves the integration of external knowledge bases and lexica to augment and refine the knowledge acquired during the language models' pre-training phase. By incorporating structured information from diverse sources, such as encyclopedic databases and lexical repositories, language models can potentially overcome the shortcomings associated with the long tail of rare linguistic cases. This shift towards a knowledge-enhanced paradigm not only addresses intra-linguistic challenges but also holds promise in facilitating inter-linguistic connections by providing a broader and common contextual understanding of linguistic phenomena across languages.

In this proposal, we present two distinct yet interconnected tasks, each integral to advancing the capabilities of language models. The first task revolves around the nuanced process of grounding textual content to reference knowledge bases, an essential step in strengthening language models' understanding and contextualization capabilities. By establishing a robust link between textual information and external knowledge repositories, we aim to enhance the models' comprehension of varied topics, fostering a deeper and more accurate representation of language.

At the same time, our proposal delves into the realm of cross-lingual modeling, with a specific focus on its application in Machine Translation. Addressing the intricacies of multilingual communication is a pivotal aspect of achieving comprehensive language understanding. We aim to investigate and contribute to the methodologies that enable language models to navigate seamlessly across languages, ensuring effective and contextually appropriate translation.

Engaging with these challenges within the UniDive working groups provides a unique opportunity to assess the current landscape of multilingual NLP comprehensively. By leveraging the collective expertise within the UniDive community, we seek not only to identify potential solutions but also to delineate promising research directions to propel the field forward.

2 Linking Raw Text to Structured Knowledge: Advancing NLP through Grounding

In the landscape of NLP, the fundamental act of linking raw text to structured knowledge stands as a linchpin, enriching language models with interpretability, control, and access to curated information. This so-called *grounding*, is exemplified by tasks such as Entity Linking, where systems have to connect the surface text of known entities (e.g., "Barack Obama") to a reference knowledge base (e.g., Wikipedia `https://en.wikipedia.org/wiki/Barack_Obama`), providing nuanced insights specific to the linked entity.

While grounding undoubtedly offers a plethora of advantages, the core challenge lies in effectively covering and structuring the vast expanse of global knowledge. For instance, many existing Entity Linking systems demand a predefined list of potential candidates for each entity to disambiguate, imposing significant constraints on their practical applicability.

Within the UniDive framework, we present ReLiK (Anonymous, 2023), an advanced Entity Link-

[*] Work carried out at Babelscape, Italy.

ing and Relation Extraction[1] system. Operating on a Retriever-Reader paradigm (Chen et al., 2017), ReLiK achieves state-of-the-art performances with a distinctive approach – it requires only entities or relations definitions[2] to link them to raw text seamlessly. This groundbreaking methodology not only streamlines the linking process during downstream applications but significantly elevates the model's generalization capabilities.

Moreover, the impressive inference speed of ReLiK, coupled with its generalization capabilities, not only facilitates leveraging EL and RE for downstream applications but also enables the efficient tagging of extensive datasets. Furthermore, ReLiK's capability to uncover novel lexicalizations for specific entities and enrich knowledge base relations between entities marks a transformative step towards continual evolution and enrichment of linguistic resources.

In summation, the journey of linking raw text to structured knowledge using systems like ReLiK addresses current challenges in grounding and charts a course toward advancing the synergy between language models and curated knowledge repositories within the UniDive paradigm.

3 Grounding Impact on Downstream Applications

Grounding text to reference knowledge bases lies at the core of many AI problems, such as Information Retrieval (Hasibi et al., 2016; Xiong et al., 2017), Automatic Text Summarization (Amplayo et al., 2018; Dong et al., 2022), Language Modeling (Yamada et al., 2020; Liu et al., 2020), and Automatic Text Reasoning (Ji et al., 2022), inter alia. One particularly interesting and recent application is mitigating the disambiguation bias within Machine Translation (MT) (Campolungo et al., 2022). Lexical ambiguity, a longstanding challenge in MT, is illustrated by instances such as the sentence "He poured a shot of whiskey". The polysemous word "shot", in this context, signifies a *small quantity*, suggesting a possible translation into Italian as "Versò un goccio di whiskey". However, several MT systems (open and commercial ones) propose an alternative translation like "Versò uno sparo di whiskey", where the noun "sparo" unexpectedly means *gunshot*.

In this proposal, we introduce WSP-NMT (Iyer et al., 2023), an innovative approach that leverages grounding towards BabelNet (Navigli and Ponzetto, 2012), a wide-coverage lexical-semantic knowledge resource that helps to scale tasks and applications to hundreds of languages, to enhance models' resilience against the disambiguation bias through a pre-training noisification strategy. Specifically, before the fine-tuning of an MT model, during a pre-training phase, it is tasked to reconstruct input sentences that undergo noise injection by substituting words with their translations from multiple languages. For example, given the sentence "The cat is on the table", a possible noisification would be "The gatto in on the mesa", where the world "cat" is substituted with its Italian translation "gatto", and "table" with it Spanish translation "mesa". This innovative approach enhances the latent disambiguation capabilities of the final machine translation systems and significantly improves the overall translation quality. The novelty with respect to previous approaches (Pan et al., 2021) lies in the creation of the pre-training dataset. Indeed, the dataset is created by leveraging a Word Sense Disambiguation system (Barba et al., 2021), i.e., a system that connects words in context to their specific meaning itemized in reference knowledge bases, which in this case is BabelNet. In this way, leveraging BabelNet, we can substitute words with their translations in other languages without incurring in the lexical ambiguity problem. Remarkably, these improvements are observed in both mid- and low-resource settings, highlighting the effectiveness of external knowledge bases in bridging language gaps, even in low-resource language scenarios. In essence, the application of grounding methodologies, exemplified by WSP-NMT, not only addresses disambiguation bias in MT but also demonstrates the potential of external knowledge bases to enhance language understanding and translation capabilities across diverse linguistic landscapes.

In conclusion, we firmly believe that these innovative approaches, showcased within the UniDive framework, demonstrate the potential for enhanced language understanding and cross-lingual applications. For a deeper exploration of their methodologies, we encourage interested readers to refer to the original papers for comprehensive details.

[1] The task of extracting relation between entities, such as capital-of between Room and Italy.

[2] A definition can be any textual description. For example, the entity "Barack Obama" can be defined with the opening of its Wikipedia Page: *Barack Obama is an American politician who served as the 44th president of the United States.*

4 Acknowledgements

Edoardo Barba gratefully acknowledges the PNRR MUR project PE0000013-FAIR / PE01-FAIR-SPOKE-5-DIAG CUP B53C22003980006, which currently funds his postdoctoral research fellowship (RTDA) in full.

References

Reinald Kim Amplayo, Seonjae Lim, and Seung-won Hwang. 2018. Entity commonsense representation for neural abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 697–707, New Orleans, Louisiana. Association for Computational Linguistics.

Anonymous. 2023. Relik: Retrieve, read and link: Fast and accurate entity linking and relation extraction on an academic budget. In *Submitted to The Twelfth International Conference on Learning Representations*. Under review.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. Exploiting entity linking in queries for entity retrieval. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016*, pages 209–218. ACM.

Vivek Iyer, Edoardo Barba, Alexandra Birch, Jeff Z. Pan, and Roberto Navigli. 2023. Code-switching with word senses for pretraining in neural machine translation.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Syst.*, 33(2):494–514.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *Text, Speech and Dialogue*, pages 103–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2901–2908.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *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 244–258, Online. Association for Computational Linguistics.

Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. 2017. Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 763–772. ACM.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.