On the Intra- and Inter-linguistic Challenges of Multilingual Silver-Data Creation and Disambiguation Biases

Edoardo Barba[1], **Niccolò Campolungo**[1], **Simone Tedeschi**[1,2] and **Roberto Navigli**[1]
[1]Sapienza NLP Group, Sapienza University of Rome, [2]Babelscape, Italy
{barba, tedeschi, navigli}@diag.uniroma1.it,
campolungo@di.uniroma1.it

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

With the advent of Large Language Models (LLMs), the Natural Language Processing (NLP) field made tremendous progress both in terms of model performances and notoriety. However, many of the challenges related to language diversity are yet to be solved and still affect the NLP community.

From an inter-linguistic perspective, NLP models, and more in general, research efforts in the field, tend to be heavily biased towards a small number of well-resourced languages, such as English, Chinese, and Spanish, with the vast majority of the other languages being underrepresented and understudied. This general trend resulted in lower-quality NLP models for mid- and low-resource languages, limiting their applicability to a relatively small audience. Multilingual NLP recently gained increasing traction, but the lack of resources is still difficult to overcome. Indeed, human annotations are expensive and time-consuming but, at the same time, indispensable in most tasks.

On the intra-linguistic front, NLP models often struggle to understand and incorporate many linguistic phenomena within a single language, including high-resource languages. For example, idiomatic expressions, euphemisms, metaphors, irony, and sarcasm are often not captured by NLP models, resulting in incorrect interpretations and misunderstandings in real-world applications. In addition, the complexity of languages and their cultural context also means that NLP models are not always equipped to handle diverse forms of expression, such as slang or regional dialects.

In this proposal, we discuss three tasks that, in our opinion, depict the status of both inter-linguistic and intra-linguistic fronts, especially for what regards their open challenges. The first two deal with the paucity of data in multilingual settings, while the third one with the need of robust benchmarks to discover inter-linguistic phenomena. We believe that delving into these challenges in the context of the UniDive working groups will be a unique occasion to properly assess and analyze the current state of multilingual NLP, as well as to identify potential solutions and research directions.

2 Multilingual Silver-Data Creation

The lack of high-quality training data is a pervasive problem affecting NLP tasks both at semantic and pragmatic level. This data is scarce or unavailable in most languages, preventing NLP tools from achieving universality. Inspired by the success of recent silver-data creation strategies, here we present consolidated works that try to address data paucity in the context of Named Entity Recognition (NER) and Idiom Identification. We focus on these two tasks as they both deal with Multiword Expressions (MWEs)[1], hence particularly relevant for the various UniDive working groups.

Multilingual Named Entity Recognition The NER task aims at identifying and classifying named entities (e.g., person, location, organization) in unstructured text. It is often used as a primary step in a wide range of NLP applications such as information extraction (Finkel et al., 2005), question answering (Babych and Hartley, 2003), and entity linking (Martins et al., 2019), inter alia. Nevertheless, named entities exhibit unique and idiosyncratic features, such as having multiple surface forms (e.g., *Barack Obama* vs. *Mr. Obama*), and being ambiguous (e.g., *Apple* can either refer to the fruit or to the company), thus making NER especially challenging. As in many other tasks, due to their complexity, large amounts of training data are required to let systems capture the features needed to achieve high performances. In the last decade, several studies have tried to address data scarcity by automatically generating NER training data (Nothman et al., 2013; Al-Rfou et al., 2015; Tsai et al., 2016; Pan et al., 2017; Tedeschi et al., 2021; Tedeschi and Navigli, 2022). Recently, Tedeschi et al. (2021) introduced WIKINEURAL

[1]Named Entities can be composed by a single word, but many of them are multi-words.

and demonstrated that by jointly exploiting the reliability of symbolic approaches and the language modeling capabilities of Transformer-based architectures, silver-data creation strategies can produce accurate annotations with quality comparable to that of manually-created ones. While WIKINEURAL showed promising results in 9 languages, there is still important work to scale NER to low-resource and endangered languages, that could be done in the context of WP 1 and 3, including devising and assessing a truly language-agnostic approach that overcomes the language-specific features hampering the current state of the art.

Multilingual Idiom Identification Another language phenomenon involving MWEs is that of idiomatic expressions, i.e. word compounds whose meanings cannot be derived by compositionally interpreting their components, e.g., *break the ice* or *kick the bucket*. Although the automatic identification and understanding of idioms are essential for a wide range of NLP tasks, such as question answering (Jhamtani et al., 2021; Mishra and Jain, 2016) and machine translation (Anastasiou, 2010), they are still largely under-explored. Indeed, the majority of the past idiom extraction strategies focus on specific syntactic constructions (e.g., verb/noun or verb/particle idioms) and on a limited number of languages (Cook et al., 2007; Muzny and Zettlemoyer, 2013; Verma and Vuppuluri, 2015; Senaldi et al., 2019), thus strongly limiting the applicability of idiom identification systems in real-world scenarios.

Recently, to alleviate the above-mentioned issues, Tedeschi et al. (2022) introduced a new multilingual framework, ID10M, that enables the automatic generation of training data for the idiom identification task. Specifically, they first exploited Wiktionary for extracting idiomatic and literal meanings of potentially-idiomatic expressions (PIEs). Then, they used Wikimatrix to retrieve usage examples of the extracted PIEs, and, finally, labeled the retrieved examples through a Transformer-based dual-encoder architecture designed to capture the semantic idiosyncrasy property of idiomatic expressions. As mentioned in the previous paragraph, silver-data creation strategies constitute a valuable opportunity to address the data paucity in languages other than English and might be deepened in the context of Work Packages 1 and 3. Idiomatic expressions, as well as other figurative phenomena, should instead be further studied in the context of Work Package 4 for promoting intra-linguistic diversity in NLP systems' modeling capabilities.

3 Inter-linguistic Disambiguation Bias

When moving to the inter-linguistic front, a key issue that affects mature models is the lexical-semantic disambiguation bias. A case in point here is Machine Translation (MT), which – while being an important and crucial task *per se* – also depicts the status of many other tasks on which the high performance on standard benchmarks fails to represent the actual model generalization capabilities. Indeed, although state-of-the-art MT systems, both commercial and non-commercial, achieve impressive scores on standard benchmarks (Kocmi et al., 2022; Freitag et al., 2022), they still fall short when dealing with the long tail of word meanings. This is due to the Zipfian nature of word sense distributions, whose modeling has proven to be a hard endeavor.

In the past few years, several works have been put forward in order to detect these translation issues. Unfortunately, such works either: i) focus on one (or few) language pairs, ii) exhibit low sense coverage, or iii) rely on completely automatically built vocabularies (Rios Gonzales et al., 2017; Raganato et al., 2019; Emelin et al., 2020). Consequently, the analyses produced by these benchmarks only paint a partial picture of a model's ability to handle the long tail of infrequent senses. This issue is addressed by Campolungo et al. (2022), who introduced DIBIMT, the first fully manually-annotated benchmark for lexical-semantic biases in Machine Translation, which focuses on infrequent senses and provides wide sense coverage in 5 languages; the resulting analyses show that MT models, especially open-source ones, are still far from correctly handling infrequent senses, and are generally biased towards translating into more frequent senses of a polysemous word. Unfortunately, because in the presence of a pair of low-resource languages systems are likely to perform poorly even on frequent senses, DIBIMT was built as an English-centric benchmark. In the context of WP 1 and 4, it would be key to increase language coverage when translating not only from/to English but also from/to other languages. An interesting direction to pursue within UniDive is to equip open-source MT models with the ability to overcome this pervasive lexical-semantic bias issue.

Acknowledgements

This document closely follows the ACL Rolling Review template, which itself builds upon many previous contributions.

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme, the PerLIR project (Personal Linguistic resources in Information Retrieval) funded by the MIUR Progetti di ricerca di Rilevante Interesse Nazionale programme (PRIN 2017), and the PNRR MUR project PE0000013-FAIR.

This work has been carried out while Simone Tedeschi was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. POLYGLOT-NER: massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 586–594. SIAM.

Dimitra Anastasiou. 2010. *Idiom treatment experiments in machine translation*. Cambridge Scholars Publishing.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653, Online. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondrej Bojař, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. Association for Computational Linguistics.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2019. Joint learning of named entity recognition and entity linking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 190–196, Florence, Italy. Association for Computational Linguistics.

Amit Mishra and Sanjay Kumar Jain. 2016. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Marco Silvio Giuseppe Senaldi, Yuri Bizzoni, and Alessandro Lenci. 2019. What do neural networks actually learn, when they learn to identify idioms? *Proceedings of the Society for Computation in Linguistics*, 2(1):310–313.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228, Berlin, Germany. Association for Computational Linguistics.

Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 681–687, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.