Returning the L in NLP: Why Language (Variety) Matters and How to Embrace it in Our Models

Barbara Plank
Computer Science Department
IT University of Copenhagen

Abstract

NLP's success today is driven by advances in modeling together with huge amounts of unlabeled data to train language models. However, for many application scenarios like low-resource languages, non-standard data and dialects we do not have access to labeled resources and even unlabeled data might be scarce. Moreover, evaluation today largely focuses on standard splits, yet language varies along many dimensions [3]. What is more is that for almost every NLP task, the existence of a single perceived gold answer is at best an idealization.

In this talk, I will emphasize the importance of language variation in inputs and outputs and its impact on NLP. I will outline ways on how to go about it. This includes recent work on how to transfer models to low-resource languages and language variants [5, 6], the use of incidental (or fortuitous) learning signals such as genre for dependency parsing [2] and learning beyond a single ground truth [1, 3, 4].

Biography. Barbara Plank is Professor in the Computer Science Department at ITU (IT University of Copenhagen). She is also the Head of the Master in Data Science Program. She received her PhD in Computational Linguistics from the University of Groningen. Her research interests focus on Natural Language Processing, in particular transfer learning and adaptations, learning from beyond the text, and in general learning under limited supervision and fortuitous data sources. She (co)-organised several workshops and international conferences, amongst which the PEOPLES workshop (since 2016) and the first European NLP Summit (EurNLP 2019). Barbara was general chair of the 22nd Northern Computational Linguistics conference (NoDaLiDa 2019) and workshop chair for ACL in 2019. Barbara is member of the advisory board of the European Association for Computational Linguistics (EACL) and vice-president of the Northern European Association for Language Technology (NEALT).

References

[1] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, 2021.

[2] Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. Genre as Weak Supervision for Cross-lingual Dependency Parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, 2021.

[3] Barbara Plank. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of KONVENS 2016, Ruhr-University Bochum*. Bochumer Linguistische Arbeitsberichte, 2016.

[4] Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, 2014.

[5] Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, 2020.

[6] Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. From Masked Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-shot Spoken Language Understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, 2021.