Towards a Computationally-Relevant Linguistic Typology for Polyglot/Multilingual NLP

Anonymous ACL submission

Abstract

This is a position paper/squib on methodology for issues involving typology in polyglot/multilingual NLP. It is intended to provide a supplementary view on "typology" that is even more marginalized than that on "linguistic typology" in NLP.

1 Introduction

There is a recent, yet long-due, surge of focus on multilinguality and typological diversity in NLP[1] – fortuitously at a time when the advancement of neural methods helped establish relative uniformity on the algorithm front that we can be afforded the opportunity to focus on data and its representation at a scale that was not possible before when there was more variety in both algorithms and data. Inspired by this subject of "typology" in NLP, we take a fresh look at the question "what science underlies natural language engineering?"(Wintner, 2009) and identify some gaps in our current NLP practice. In addition, we propose making typology in computation[2] a focus that can serve the more scientifically-minded among language engineers.

2 (Multilingual) NLP and linguistic typology

2.1 Text/Orthography

Sproat (2016) comments on how both the objectives and interests of NLP and of linguistic typology can be orthogonal to each other. And with regard to the under-representation of orthography in linguistic typology, this cannot be more true.

The reason why there is a disregard for orthography in contemporary linguistics has to do with the focus of linguistics being "language" as a universal phenomenon. The claim was that since language is also present in communities and situations in which there are no writing systems, orthography ought to be considered as an artifact that would be less telling when it comes to the study of the nature of the human mind. A secondary argument for the marginalization of orthography in linguistics is also an attempt to break away from the descriptive nature of the philological past of "the study of language".

On the other hand, the default processing format for NLP has been based on orthography. This form for automatic processing – which stemmed from the mere reason of convenience in the field's pioneer days – has nonetheless provided us with a better means to study text messages, emojis, Braille, poems by E. E. Cummings, and subtitles with sign languages. While we do not dispute the possibility and benefits of including a more sophisticated phonological system, esp. that which mirrors articulatory and phonetic (near-)universals, we must be aware of the fact of how much there needs to be done to supplement information that is thus far not in our "database" such as WALS (Dryer and Haspelmath, 2013), e.g. on tones and grapheme-phoneme relationship (both of these are under-represented in traditional analyses).

2.2 On the disparity between science for human and for machines

Language engineering for and by humans as well as machines is understood as NLP. The science of language from the perspective that is primarily interpretable by humans is linguistics – the basic units can be, e.g. a phoneme, a morpheme etc..

[1] For a more comprehensive survey and review, please refer to Ponti et al. (2018) and O'Horan et al. (2016).

[2] as opposed to "computational typology": as any "computational x" may only imply the study of x with computational means or the implementation/extension of concepts in x, but not necessarily of concepts that have thus far been overlooked or considered beyond x – the purpose of this paper is to fill in the gap of what is missing amidst this already underappreciated field that addresses linguistic diversity.

NLP can certainly, as it has, operate on such units and terms, but it does not have too. In fact, the basic units of computation are bits and bytes. To what extent have we reconciled with and accepted a less anthropocentric view of a science for NLP?

In the First Workshop on Computational Approaches to Code Switching (Solorio et al., 2014), only 4 out of 7 teams made use of character encoding information when it could have been one of the easiest but accurate cue to differentiate between Mandarin and English. Encoding formats are also a hyperparameter that one can manipulate in (the optimization of) byte-based processing. But to what extent would these be considered relevant for linguistic typology? Should the next generation of students in NLP be oblivious to these?

And a more important point that follows is – what if the "types" that are computationally relevant in text processing do not align with our human-based typological classification? If our only fallback science is linguistic typology in human view and that we only examine languages of different "types" under such categorization scheme, how can we ever expect new typological differences that are related to machine processing be discovered and handled? E.g. if the effects of crosslinguistic variation can be minimized through a shared/common vocabulary irrespective of genealogical distance, how should these be evaluated? We believe a knowledge system should be established in its own right that addresses computationally relevant concepts and issues for handling diverse languages or language data. It should also note how these types from a computational perspective relate to the "canonical" types in linguistic typology.

2.3 Language independent != linguistically naive IFF we have both performance AND evaluation in mind for all varieties

Bender (2009) mentions how language-agnostic papers have often been (erroneously) assumed to be language-independent. Despite how many recent papers claim "typological diversity" through the coverage of a broader range of languages, their coverage does not entail effectiveness and comparable performance across all languages. Languages are often cherry-picked to exclusively highlight the strengths of a model. A dearth of qualitative discussion with respect to how linguistic idiosyncrasies have been handled is still the norm in most papers. And even if they did address an expectation along the lines of typological relatedness, they all fall back upon a notion of typological concept that is human-centric as opposed to a generalization and analysis that is more profound and empirical on a level that is both typology and computation relevant.

But perhaps a more appalling fact is that we as a field lack a comprehensive and systematic knowledge base of what kind of algorithm works well with what languages. For instance, if n-gram word-level models are not considered sufficient for morphologically complex languages, do character-level n-grams fare better? If so, what order of n-grams fare well with which languages, variants, genres, or data types? If a certain language-independent heuristic has been leveraged, what are the potential difficulties that were expected and overcome, and how have linguistic variations been handled? With distributed representations, many linguistic properties become even more "inherent" and implicitly embedded than before when explicit modeling of linguistic concepts was the norm. As expert multilingual/polyglot NLP practitioners, we need to be able to have a better grasp on this.

3 Conclusion

We have brought up a few points that would help build a more solid scientific foundation for language-independent NLP. We believe by creating a direction in polyglot/multilingual NLP that is more comprehensive and encompassing would help make typology in NLP more relevant for all NLP practitioners, not just the ones who happen to be also interested in modeling particular traditional typological concepts.

May it be linguistic typology / computational typology (for humans) and typology in computation (for machines) – both of these directions should be appreciated and supported. Yet, it is imperative to know when, how, and why they need to be made distinct from each other in NLP. Cross-fertilization between the two areas is expected to be fruitful, but their relevance and validity should be independent of each other. By stating the areas that we need to be aware of when building a knowledge system, i.e. science, to support the practice of NLP, we are also making room for evaluation standards that would be appropriate for each classification system.

References

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Helen O'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *arXiv preprint arXiv:1807.00914*.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Richard Sproat. 2016. Language typology in speech and language technology. *Linguistic Typology*, 20(3):635–644.

Shuly Wintner. 2009. Last words: What science underlies natural language engineering? *American Journal of Computational Linguistics*, 35(4):641–644.