Towards a Computationally Relevant Typology for Polyglot/Multilingual NLP

Ada Wan
University of Zurich
Zurich, Switzerland
ada.wan@uzh.ch
adawan919@gmail.com

Abstract

The purpose of this writing is to facilitate discussion on possible future directions for typology for polyglot/multilingual NLP. Although there is a shared objective between linguistics and polyglot NLP of accounting for all languages (or as many as possible) using one framework, there are many differences that we need to take into consideration. We identify some of these differences and their underlying reasons, and provide some suggestions as to how we can build a computationally relevant typology that could be useful for both the science and engineering of language.

1 Introduction

Computational Linguistics (CL) is a discipline that has both a scientific and an engineering side. The engineering side of CL is often referred to as natural language processing (NLP) and is "largely concerned with building computational tools that do useful things with language, e.g., machine translation, summarisation, question-answering, etc." (Johnson, 2011). CL emerged as "the scientific study of language from a computational perspective"[1] about half a century ago. In its earlier decades, it employed heuristics that were motivated by the science of language (Linguistics (Lx)), but as the mainstream syntactic theories in Lx became more obscure, the field leaned towards more empirical methods which also happened to have been successful in modeling NLP data – hence we witnessed in NLP the "statistical turn" in the 1990s, much machine learning (ML) from the 2000s[2], and the "Deep Learning (DL) tsunami" in the 2010s (Manning, 2015). Yet, despite how methods in the science and engineering of language may have diverged, the objective of Lx, esp. of generative Lx, still has much in common with polyglot/multilingual NLP in that they both aim to treat crosslinguistic commonalities and differences within a universal framework (cf. e.g. Prince and Smolensky (2008) and Tsvetkov et al. (2016)). But since the history of NLP has witnessed linguistic methods underperforming less linguistically motivated methods, we will identify some differences in practices and values below in hopes of enabling better understanding and progress in our interdisciplinary field.

2 Clarification on Some Differences

2.1 Disparity between humans and machines – in representation and evaluation

Although there is a wide variety of what Lx has to offer, it is primarily the science of language from a perspective that is interpretable by humans. NLP algorithms can certainly, as they have, operate on units that are human-interpretable (e.g. tokens on word-level), but they do not have to. Processing on the level of sub-characters (for logographic languages), characters, bytes, or byte pairs has also shown to be effective. But subword units other than morphemes (smallest meaningful units) and phonemes (units of distinctive sound) are not always human-interpretable. Humans tend to seek meaning when evaluating and are very often biased by the categories that are prevalent in their native languages (and/or lan-

[1] from the website of the Association of Computational Linguistics https://www.aclweb.org/portal/, on which CL is even defined as a science (though unclear if it is to be differentiated from "engineeering")

[2] "dates" here rounded up to the nearest decade – surely ML existed before then but it is its application in NLP tasks that is of relevance here. Such dating is supported by the quote "[t]he phenomenal success of machine learning in engineering natural language applications has led to a curious situation: Natural language processing practitioners who were trained in the last 15 to 20 years may have established a quite successful career in this area with only a haphazard knowledge of the science of natural languages" from Dyer (2015).

guages with which they are familiar). Not only is there a gap on what humans can or cannot evaluate, their linguistic misjudgments, however insightful they may be from a psycholinguistic point of view, can be counter-productive to the evaluation of computational processes in NLP unless our goal is to model their psycholinguistic process.

With DL, not only is it easier to model on levels with finer granularity, but also in more of a massively language-agnostic, joint fashion, but if we keep insisting on evaluating on classes and terms that are interpretable to humans only, we could miss out on pitfalls that are exclusively computational. For example, if Chinese can be modeled successfully like English using ASCII characters via Pinyin but unsuccessfully using its logographic character set, that says nothing about their phylogenetic/geographical/typological[3] relation. Only when we include notions/concepts like "encoding" or "representation level" in a computationally relevant typology, would we have a means to address this dimension properly.

2.2 Phonemes vs. orthography

Orthography has been under-represented in the tradition of Lx and linguistic typology (Sproat, 2016). One rationale of this practice has to do with the study of language as a universal phenomenon being the primary focus of Lx. Since language is also present in communities and situations in which there are no writing systems, orthography was considered as an arbitrary artifact that would be less telling when it comes to the study of linguistic nature. On the other hand, the default processing format for NLP has been based on orthography. This format 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. The advantages of a phonemic representation are that it could be informative in cognate identification (as it filters out many orthographic idiosyncracies) and it can also serve as a common alphabet representing many languages with one encoder/decoder in DL (assuming such data is available and the transcriptions complete – also for languages with traditionally less studied phonological phenomena). The advantages of orthography are: (i) data are more readily available, (ii) it enables us to decipher graphic and non-vocal elements (including spaces as silence) easily, (iii) it is a medium by which compositionality in ways that are analogous to morphemic analyses or hypernym-hyponym relationship mining for "isolating"[4] logographic languages such as Chinese can be studied, and (iv) it is more accessible to those who are deaf. On the downside, for languages that do have a large character-level vocabulary (that surfaces in the data), the softmax bottleneck (Yang et al., 2017) could impact performance adversely.

2.3 Data and approaches

Traditionally, linguists are concerned with distinctive, qualitative features to describe, characterize, and formalize language(s). This gives rise to their preference for data that help them discern linguistic phenomena. Hence the data they collect can be thematic in nature or tend to have high concentration of distinctive, yet often rare, phenomena of interest to themselves as specialists. This is the manner in which data were collected for the typological database WALS (Dryer and Haspelmath, 2013). Foundational assumptions as well as availability of experts affected the distribution of material, which is also gauged to exhibit balance in the collection of languages over language families and geographical areas (Malaviya et al., 2017). Therefore, researchers using the database for data-driven research need to be mindful of this.

Corpus Lx can be corpus-based or corpus-driven (Tognini-Bonelli, 2001). The former method strives to use data to corroborate or refute linguistic theories or hypotheses, whereas the latter uses data as the sole empirical basis which determines the analysis without prior assumptions and expectations. In order to identify limits and blind spots of any theoretical framework, or to discover unprecedented insights into linguistic patterns, data-driven methods are necessary. Corpus linguists, who are often the ones who curate datasets in NLP, need to be aware of any potential conflict of interest. For most applications in NLP, data should cover more breadth, for linguistic studies, more depth. This may entail collecting two different sets of data.

[3] classes generally considered in typology in NLP (e.g. in Littell et al. (2017))

[4] Chinese is considered to have a low morpheme to word ratio – that is a claim that is based on an English-centric concept of "word" as a conceptual unit and on a practice that marginalizes orthography.

3 Constructive Suggestions

We believe the development of a typological science that is computationally relevant and beyond the phylogenetic, geographic, and linguistic dimensions would be helpful for polyglot/multilingual NLP. The advancement of neural methods has helped establish relative uniformity on the algorithm front that we can be afforded the opportunity to focus on data and its representation in a manner that was not possible before when there was more variety in both algorithms and data. We hypothesize that there is systematicity in the way language data can be classified and that there is a way to justify successful performance crosslinguistically. Our field lacks a comprehensive and systematic knowledge base of what kind of algorithms (architectures/optimizers/training regimes) works well with what languages in what tasks and sizes.

Lin et al. (2019) incorporates "data-dependent features" (dataset size, type-token ration (TTR), word overlap and subword overlap) and "dataset-independent features" (geographic, genetic, inventory, syntactic, phonological, and featural distance – for information that is external of the dataset at hand) to determine the success of crosslingual transfer. Hence, we can think of:

1. creating and maintaining a **comprehensive knowledge system** involving experimental results of the world's languages/variants, documenting results of ablation studies with values for each feature as noted above and, in addition: data representation (e.g. on the level of byte, character, phoneme, BPE, word, phrase, or sentence, and if byte, its encoding format), data genre, data source, and other hyperparameters such as architecture type and size, number of parameters, training time, maximum sequence length, and batch size. Instead of testing for tasks on a one-off basis, we would perform consistent ablation studies on each dataset and subportions thereof with each of these features. Ideally, the datasets would be parallel corpora to ensure fair comparison.

2. studying how all the features in (1) are correlated, and how linguistic variants can be grouped;

3. academic "crowdsourcing": hosting a non-competitive shared tasks to populate this database as ablation studies at this scale would require an enormous amount of computational resources and time. Teams will have identical specifications for data, data handling, algorithm, and hyperparameter values and will cooperate in compiling this massive database.

4 Conclusion

We have identified some aspects that have thus far been neglected in the traditional typological studies of language. We believe that creating a computationally relevant and comprehensive typological science for polyglot/multilingual NLP that includes features of data, size, algorithms, and tasks would be more useful for language engineers and insightful for language scientists.

Acknowledgments

We would like to thank the three anonymous reviewers for their valuable feedback. We would also like to thank the organizers' kind permission to extend our extended abstract to its current format and length. We are grateful, esp. for reviewer #2's objection/question "who has done typology *well*?" – reminding us that we need to be constructive in our criticisms of a direction less explored so to encourage real progress.

References

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

Chris Dyer. 2015. Book reviews: Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax by emily m. bender. *Computational Linguistics*, 41(1):153–155.

Mark Johnson. 2011. How relevant is linguistics to computational linguistics. *Linguistic Issues in Language Technology*, 6(7).

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational*

Linguistics: Volume 2, Short Papers, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Christopher D. Manning. 2015. Last words: Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

A. Prince and P. Smolensky. 2008. *Optimality Theory: Constraint Interaction in Generative Grammar*. Wiley.

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

E. Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. Studies in corpus linguistics. J. Benjamins.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. 2017. Breaking the softmax bottleneck: A high-rank rnn language model. *arXiv preprint arXiv:1711.03953*.