On the role of resources in the age of large language models

Simon Dobnik
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
name.surname@gu.se

Abstract

We evaluate the role of expert-based domain knowledge and resources in relation to the recent developments in natural language processing that focus on training large models by referring to our work on under-resourced scenarios which we believe also informs work on training "well-resourced" languages and domains.

1 Introduction

In the recent years large language models based on transformers that are trained end to end and automatically capture the structure of language have achieved remarkable performance (Devlin et al., 2018; Brown et al., 2020). While showing impressive performance on several tasks, several questions have been raised in relation to their training and usage. Do they really capture natural language semantics by learning from the linguistic form alone not being grounded in the world (Searle, 1980; Harnad, 1990; Bender and Koller, 2020)? What semantics do they learn? Large language models require a lot of data to train and to do that an approach in natural language processing has been to utilise (sometime indiscriminately) all the data that is available. However, access to the data is heavily biased to the data that can be found online, e.g. Wikipedia, or data that can be collected with crowd-sourcing platforms. Such selection of data on which the models are trained does not represent all possible contexts of language use or groups of society producing language which results in undesired and exaggerated thematic (Agrawal et al., 2017) and social bias (Bender et al., 2021) in the models. In addition to access to large datasets of text (and images), training such models is also costly in terms of time and available computational resources, both factors which are only available to a few world languages where English is over-represented.

On the other hand, curating of datasets in terms of collecting high-quality data and their annotation with linguistically-motivated annotation schemes has a long tradition in natural language processing. Transformer models learn linguistic structure end-to-end and systems using automatically learned contextualised embedding surpass models with expert-engineered features which raises a question whether all the years of hard expert work is superfluous. But can we be really sure that the models really have learned useful linguistic structure (Conneau et al., 2018)? Is that structure the same what we expect (Dobnik et al., 2018)? Since annotation of resources is directly connected with linguistics, which focuses on understanding of differences between languages and therefore explores a variety of world languages, the annotation work provides a good cross-linguistic coverage but frequently datasets have a limited coverage of examples and may not be large enough for training machine learning models. Another benefit of a close relation of this approach to linguistics is that the annotation categories are motivated by our (expert) understanding of how these languages work so the resulting representations are well-motivated and interpretable.

In this presentation we evaluate the previous questions about the role of data and resources for modern natural language processing in the light of our experience with building resources for under-resourced language from ground up. We highlight the idea that in such scenarios both kinds of resources are useful and in fact shows that they have complementary weaknesses and strengths. It follows that modern and future natural language processing must be informed by expert domain knowledge about language and linguistics as without these we are not able to evaluate the data that these models are utilising nor interpret what semantics or bias the models might have captured nor we can improve the models in a motivated way

either indirectly (by neural architecture choice) or directly (by injection of labels).

2 The need for data

Training of large language models requires a lot of data that spans over different contexts of language use and social groups in order to capture (some kind of) knowledge of language for natural language generation and interpretation and to avoid unwanted social and contextual bias. However, as discussed in the previous section even for well-resourced language models such as English it is still not clear whether this has been achieved as data selection and coverage of thematic and social contexts that are used in the training data has not yet been (to our knowledge) systematically evaluated. Equally, approaching the same problem from the engineering perspective it has been impossible to collect enough data or build a model large enough to test whether such an endeavour is theoretically and practically possible at all (Villalobos et al., 2022).

This need for data and its limitations becomes much more evident when we examine the under-resourced scenarios that we looked at. Arabic natural language processing is an interesting case. Modern Standard Arabic (MSA) is a standardised form of Arabic used in printed media and news and is supported well in terms of natural language models and resources. However, there are also several local varieties spoken over a large geographical span. In addition, Arabic may be also spoken (and written in social media) and code-switched with several other varieties and even different languages. Some of these have received more attention than others in NLP. For example, there has been a good support for the Egyptian variety but very little support for Algerian and the individual varieties in the Levantine area. Another interesting aspect of Arabic linguistic landscape is that it differs between regions/countries in what situation contexts different varieties are used, what other varieties are present in these contexts and how similar these varieties are.

Speakers/writers in Algeria (Adouane and Dobnik, 2017) use social media where varieties that were typically spoken in personal everyday communication are now written with Arabic script on a limited phone keyboard. There is no standard spelling for these varieties and the practical limitation of using different keyboards introduce high level of variation in the way these varieties are written by different users in different contexts on different social media. A further level of of variation is added when these varieties are code-switched with MSA and other languages, in case of Algerian with Berber, French and English, all written in the same script. Hence, one of the first tasks to tackle the bootstrapping of resources for Algerian was to build a code-switching detector based on a limited expert-annotated corpus using probabilistic (HMM) and bi-gram feature classification models.

On the other hand, Levantine dialects (Abu Kwaik et al., 2018b) are various closely related Arabic dialects that are spoken and written in social media but such context makes them hard to distinguish from each other as phonological form which underlies a lot of discriminating power is missing (Abu Kwaik et al., 2018a). Finally, Wolaytta (Gebreselassie and Dobnik, 2022), is one of several languages spoken in Ethiopia, belongs to the Omotic family of African languages which is different from Amharic, an official national language which belongs to the Semitic family of languages and for which most NLP resources exits. Wolaytta is mostly used in spoken form in personal communication and radio and has been standardised in the written form in school texts and religious textbooks. In terms of NLP resources, there is no social media but they are radio programmes , school textbooks, religious literature and a Wolaytta-English dictionary.

Comparing these cases we can see that there are large linguistic differences between these target varieties and the language used in the closest set of contexts for which NLP resources exist and also that we have limited records of of contexts in which they are used either be cause data is missing or because the variety is not used in those contexts. Consequently, building NLP resources had to rely on a large support from expert linguistic and social knowledge because the training examples were limited we relied on simple machine learning methods such as Bayesian classification which in conjunction with the expert knowledge gave satisfactory results.

3 The need for the right method

Different (i) contexts of language use, (ii) relation to the closest variety for which NLP resources exits, (iii) availability of data, (iv) avail-

ability of expert annotation required very different tools and approaches to build resources and NLP applications for these varieties. For example, using character and sub-word models and CNNs, weak supervision (bootstrapping from an existing labeller, self-training) (Adouane et al., 2018b; Abu Kwaik et al., 2020), injecting background knowledge from lexicon and pre-trained sub-word embeddings (Adouane et al., 2018a), pre-training (Abu Kwaik et al., 2022), text normalisation with alignment of tokens (Adouane et al., 2019b), data augmentation (Adouane et al., 2019a). It is often the case that a simple model works better than a more complex model, most likely because it is able to generalise better from a limited data (Abu Kwaik et al., 2019a,b). In sum, understanding language and its context is important even at the age of large language models to make an informed choice what model should be used when.

4 Are "well-resourced" languages also under-resourced?

We argued previously that there is still an open question whether language model have in fact reached understanding of language as they have not been exposed to all contexts of language use. Hence, we are facing with similar under-resourced scenarios also in cases of "well-resourced" languages where existing large language models are applied in contexts or tasks for which the model has not been initially trained on. Language is continuously changing and speakers/writers are creative, especially in social media (Noble et al., 2021). Hence, pre-trained language models may become quickly outdated.

Our work on generating spatial descriptions of images shows that since pre-training of visual features such as ResNet (He et al., 2016), Faster-RCNN (Ren et al., 2015) and CLIP (Radford et al., 2021) that are trained to identify objects affects what is model able to learn about predicting relations which are likely to be hallucinated from a language model, simply because the model has not been pre-trained in this way and until such features are explicitly identified (Ghanimifard and Dobnik, 2019). A significant body of work on language and vision has focused on generation of image descriptions that focus on a single sentence. Extending the task to multi-sentence generation requires application of different models (Ilinykh and Dobnik, 2020).

Adaption of pre-trained models from the image captioning domain on object classification (where objects are in the attention focus of the scene) to the domain of situated language (where a robot without a specific model of visual and thematic attention) is very different reveals that visual information in such cases is used quite differently than in an image captioning scenario (Ilinykh et al., 2022).

finally, a comparison of generated noun phrases in generated multi-sentence descriptions to human descriptions (Ilinykh and Dobnik, 2022) reveals a difference. Models are more general predictors than humans across the board and opt for more general descriptions of objects than humans. This is because they are trained on a single task, but also within this task they are biased to find a single generalisation following a training objective covering all of the examples equally, whereas in reality humans might use descriptions that are more general or more specific on a case-to-case bases. Since general descriptions are more frequent than the specific ones, they always win. Overall, it appears that a very fine grained knowledge of language data is required to capture all the contexts.

5 Conclusions

We might want to rethink how to train such models – having one large model is useful, but perhaps not the end of the NLP story. Expert-based knowledge is crucial for selecting data and evaluating contexts in which such model is trained. Expert-based resources are useful to make informed choices about the model architectures and to support training of end-to-end models by feature engineering and selection. This also includes application of pre-trained feature representations. Fine grained evaluation of such models is necessary in order to achieve the models fit to the previous requirements, with targeted positive and negative linguistic examples (beyond the level of granularity of a Touring test as implemented in the GLUE benchmarks (Wang et al., 2019)), which is one of our current efforts.

References

Kathrein Abu Kwaik, Stergios Chatzikyriakidis, and Simon Dobnik. 2019a. Can modern standard Arabic approaches be used for Arabic dialects? Sentiment analysis as a case study. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics (WACL-3)*,

pages 40–50, Cardiff, United Kingdom. Association for Computational Linguistics.

Kathrein Abu Kwaik, Stergios Chatzikyriakidis, and Simon Dobnik. 2022. Pre-trained models or feature engineering: The case of dialectal Arabic. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection (OSACT) at LREC 2022*, pages 41–50, Marseille, France. European Language Resources Association.

Kathrein Abu Kwaik, Stergios Chatzikyriakidis, Simon Dobnik, Motaz Saad, and Richard Johansson. 2020. An Arabic tweets sentiment analysis dataset (ATSAD) using distant supervision and self training. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools with a Shared Task on Offensive Language Detection (OSACT4-2020) at Language Resources and Evaluation Conference (LREC 2020)*, pages 1–8, Marseille, France. European Language Resources Association (ELRA).

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018a. A lexical distance study of Arabic dialects. *Procedia Computer Science 142: Proceedings of the 4th International Conference on Arabic Computational Linguistics (ACLing 2018)*, 142:2–13.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018b. Shami: a corpus of Levantine Arabic dialects. In *Proceedings of LREC 2018, 11th International Conference on Language Resources and Evaluation*, pages 1–8, Phoenix Seagaia Conference Center, Miyazaki, Japan. European Language Resources Association (ELRA).

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2019b. LSTM-CNN deep learning model for sentiment analysis of dialectal Arabic. In *Proceedings of ICALP'19: The 7th International Conference on Arabic Language Processing*, Communications in Computer and Information Science (CCIS), pages 1–14, Nancy, France. Springer.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2018a. Improving neural network performance by injecting background knowledge: Detecting code-switching and borrowing in Algerian texts. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching at 56th Annual Meeting of the Association for Computational Linguistics (ACL-2018)*, pages 20–28, Melbourne, Australia. Association for Computational Linguistics.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2019a. Neural models for detecting binary semantic textual similarity for Algerian and MSA. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop WANLP 2019 at ACL-2019*, pages 78–87, Florence, Italy. Association for Computational Linguistics.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2019b. Normalising non-standardised orthography in Algerian code-switched user-generated data. In *Proceedings of The 5th Workshop on Noisy User-generated Text (W-NUT) at EMNLP 2019*, pages 1–10, Hong Kong. Ritter, Alan and Xu, Wei and Baldwin, Tim and Rahimi, Afshin.

Wafia Adouane and Simon Dobnik. 2017. Identification of languages in Algerian Arabic multilingual documents. In *Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP)*, pages 1–8, Valenica, Spain. The European Chapter of the Association for Computational Linguistics (EACL), Association for Computational Linguistics.

Wafia Adouane, Simon Dobnik, Jean-Philippe Bernardy, and Nasredine Semmar. 2018b. A comparison of character neural language model and bootstrapping for language identification in multilingual noisy texts. In *Proceedings of the Second Workshop on Subword and Character Level Models in NLP (SCLeM) at 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 1–10, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. *arXiv*, arXiv:1712.00377 [cs.CV]:1–15.

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, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario

Amodei. 2020. Language models are few-shot learners. *arXiv*, arXiv:2005.14165 [cs.CL]:1–75.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv*, arXiv:1805.01070 [cs.CL].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*, arXiv:1810.04805 [cs.CL]:1–14.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Tewodros Gebreselassie and Simon Dobnik. 2022. Wolaytta word embeddings. Technical report, manuscript, Centre for Linguistic Theory and Studies in Probability (CLASP), Gothenburg, Sweden.

Mehdi Ghanimifard and Simon Dobnik. 2019. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG-2019)*, pages 1–15, Tokyo, Japan. Association for Computational Linguistics.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1–3):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Nikolai Ilinykh and Simon Dobnik. 2020. When an image tells a story: The role of visual and semantic information for generating paragraph descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2022. Do decoding algorithms capture discourse structure in multimodal tasks? A case study of image paragraph generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 480–493, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nikolai Ilinykh, Yasmeen Emampoor, and Simon Dobnik. 2022. Look and answer the question: On the role of vision in embodied question answering. In *Proceedings of the 15th International Conference on Natural Language Generation (INLG)*, Colby College, Waterville, ME, USA.

Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. Semantic shift in social networks. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv*, arXiv:2211.04325 [cs.LG].

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.