Does Syntactic Knowledge in Multilingual Language Models Transfer Across Languages?

Prajit Dhar **Arianna Bisazza**
Leiden Institute of Advanced Computer Science
Leiden University, The Netherlands
{p.dhar,a.bisazza}@liacs.leidenuniv.nl

Abstract

Recent work has shown that neural models can be successfully trained on multiple languages simultaneously. We investigate whether such models learn to share and exploit common syntactic knowledge among the languages on which they are trained. This extended abstract presents our preliminary results.

1 Introduction

Recent work has shown that state-of-the-art neural models of language and translation can be successfully trained on multiple languages simultaneously without changing the model architecture (Östling and Tiedemann, 2017; Johnson et al., 2017). In some cases this leads to improved performance compared to models only trained on a specific language, suggesting that multilingual models learn to share useful knowledge cross-lingually through their learned representations. While a large body of research exists on the multilingual mind, the mechanisms explaining knowledge sharing in computational multilingual models remain largely unknown: What kind of knowledge is shared among languages? Do multilingual models mostly benefit from a better modeling of lexical entries or do they also learn to share more abstract linguistic categories?

We focus on the case of language models (LM) trained on two languages, one of which (L1) is over-resourced with respect to the other (L2), and investigate whether the *syntactic* knowledge learned for L1 is transferred to L2. To this end we use the long-distance agreement benchmark recently introduced by Gulordava et al. (2018).

2 Background

The recent advances in neural networks have opened the way to the design of architecturally simple multilingual models for various NLP tasks, such as language modeling or next word prediction (Tsvetkov et al., 2016; Östling and Tiedemann, 2017; Malaviya et al., 2017; Tiedemann, 2018), translation (Dong et al., 2015; Zoph et al., 2016; Firat et al., 2016; Johnson et al., 2017), morphological reinflection (Kann et al., 2017) and more (Bjerva, 2017). A practical benefit of training models multilingually is to transfer knowledge from high-resource languages to low-resource ones and improve task performance in the latter. Here we aim at understanding *how* linguistic knowledge is transferred among languages, specifically at the syntactic level, which to our knowledge has not been studied so far.

Assessing the syntactic abilities of monolingual neural LMs trained without explicit supervision has been the focus of several recent studies: Linzen et al. (2016) analyzed the performance of LSTM LMs at an English subject-verb agreement task, while Gulordava et al. (2018) extended the analysis to various long-range agreement patterns in different languages. The latter study found that state-of-the-art LMs trained on a standard log-likelihood objective capture non-trivial patterns of syntactic agreement and can approach the performance levels of humans, even when tested on syntactically well-formed but meaningless (*nonce*) sentences.

Cross-language interaction during language production and comprehension by *human* subjects has been widely studied in the fields of bilingualism and second language acquisition (Kellerman and Sharwood Smith; Odlin, 1989; Jarvis and Pavlenko, 2008) under the terms of *language transfer* or *cross-linguistic influence*. Numerous studies have shown that both the lexicons and the grammars of different languages are not stored independently but together in the mind of bilinguals and second-language learners, leading to observ-

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 374–377
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

able lexical and syntactic transfer effects (Kootstra et al., 2012). For instance, through a cross-lingual syntactic priming experiment, Hartsuiker et al. (2004) showed that bilinguals recently exposed to a given syntactic construction (passive voice) in their L1 tend to reuse the same construction in their L2.

While the neural networks in this study are not designed to be plausible models of the human mind learning and processing multiple languages, we believe there is interesting potential at the intersection of these research fields.

3 Experiment

We consider the scenario where L1 is over-resourced compared to L2 and train our bilingual models by *joint training* on a mixed L1/L2 corpus so that supervision is provided simultaneously in the two languages (Östling and Tiedemann, 2017; Johnson et al., 2017). We leave the evaluation of pre-training (or transfer learning) methods (Zoph et al., 2016; Nguyen and Chiang, 2017) to future work.

The monolingual LM is trained on a small L2 corpus (LM_{L2}). The bilingual LM is trained on a shuffled mix of the same small L2 corpus and a large L1 corpus, where L2 is oversampled to approximately match the amount of L1 sentences (LM_{L1+L2}). See Table 1 for the actual training sizes. For our preliminary experiments we have chosen French as the helper language (L1) and Italian as the target language (L2). Since French and Italian share many morphosyntactic patterns, accuracy on the Italian agreement tasks is expected to benefit from adding French sentences to the training data *if* syntactic transfer occurs.

Data and training details: We train our LMs on French and Italian Wikipedia articles extracted using the WikiExtractor tool.[1] For each language, we maintain a vocabulary of the 50k most frequent tokens, and replace the remaining tokens by <unk>. For the bilingual LM, all words are prepended with a language tag so that vocabularies are completely disjoint. Their union (100K types) is used to train the model. This is the least optimistic scenario for linguistic transfer but also the most controlled one. In future experiments we plan to study how transfer is affected by varying degrees of vocabulary overlap.

[1] https://github.com/attardi/wikiextractor

Following the setup of Gulordava et al. (2018), we train 2-layer LSTM models with embedding and hidden layers of 650 dimensions for 40 epochs. The trained models are evaluated on the Italian section of the syntactic benchmark provided by Gulordava et al. (2018), which includes various non-trivial number agreement constructions.[2] Note that all models are trained on a regular corpus likelihood objective and do not receive any specific supervision for the syntactic tasks.

4 Results and Conclusions

Table 1 shows the results of our preliminary experiments. The unigram baseline simply picks, for each sentence, the most frequent word form between singular or plural. As an upper-bound we report the agreement accuracy obtained by a monolingual model trained on a large L2 corpus.

Table 1: Accuracy on the Italian agreement set by the unigram baseline, monolingual and bilingual LMs.

Model	Training (#tok)	Agreement_{IT} Orig.	Nonce
Unigram	—	54.9	54.5
LSTM_{IT}	10M_{IT}	80.7	79.9
LSTM_{FR+IT}	$80\text{M}_{FR} + 8\times10\text{M}_{IT}$	82.4	77.5
LSTM_{IT} (large)	80M_{IT}	88.2	82.6

The effect of mixing the small Italian corpus with the large French one does not appear to be major. Agreement accuracy increases slightly in the original sentences, where the model is free to rely on collocational cues, but decreases slightly in the nonce sentences, where the model must rely on pure grammatical knowledge. Thus there is currently no evidence that syntactic transfer occurs in our setup. A possible explanation is that the bilingual model has to fit the knowledge from two language systems into the same number of hidden layer parameters and this may cancel out the benefits of being exposed to a more diverse set of sentences. In fact, the bilingual model achieves a considerably worse perplexity than the monolingual one (69.9 vs 55.62) on an Italian-only held-out set. For comparison, Östling and Tiedemann (2017) observed slightly better perplexities when mixing a small number of related languages, however

[2] For more details on the benchmark and LM configurations refer to https://github.com/facebookresearch/colorlessgreenRNNs

their setup was considerably different (character-level LSTM with highly overlapping vocabulary).

This is work in progress. We are currently looking for a bilingual LM configuration that will result in better target language perplexity and, possibly, better agreement accuracy. We also plan to extend the evaluation to other, less related, language pairs and different multilingual training techniques. Finally, we plan to examine whether lexical syntactic categories (POS) are represented in a shared space among the two languages.

Acknowledgments

This research was partly funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646. The experiments were conducted on the DAS computing system (Bal et al., 2016).

References

Henri Bal, Dick Epema, Cees de Laat, Rob van Nieuwpoort, John Romein, Frank Seinstra, Cees Snoek, and Harry Wijshoff. 2016. A medium-scale distributed system for computer science research: Infrastructure for the long term. *Computer*, 49(5):54–63.

Johannes Bjerva. 2017. One model to rule them all: Multitask and multilingual modelling for lexical analysis. *CoRR*, abs/1711.01100.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Robert J. Hartsuiker, Martin J. Pickering, and Eline Veltkamp. 2004. Is syntax separate or shared between languages?: Cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological Science*, 15(6):409–414.

Scott Jarvis and Anna Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. Routledge.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1993–2003, Vancouver, Canada. Association for Computational Linguistics.

Eric Kellerman and ed. Sharwood Smith, Michael. *Crosslinguistic influence in second language acquisition*. Pergamon.

Gerrit Jan Kootstra, Janet G. Van Hell, and Ton Dijkstra. 2012. Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *Bilingualism: Language and Cognition*, 15(4):797819.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge Applied Linguistics. Cambridge University Press.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. *CoRR*, abs/1802.00273.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.