Multilingual Probing of Deep Pre-Trained Contextual Encoders

Vinit Ravishankar
Language Technology Group
Department of Informatics
University of Oslo
vinitr@ifi.uio.no

Memduh Gökırmak
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
memduhg@gmail.com

Lilja Øvrelid
Language Technology Group
Department of Informatics
University of Oslo
liljao@ifi.uio.no

Erik Velldal
Language Technology Group
Department of Informatics
University of Oslo
erikve@ifi.uio.no

Abstract

Encoders that generate representations based on context have, in recent years, benefited from adaptations that allow for pre-training on large text corpora. Earlier work on evaluating fixed-length sentence representations has included the use of 'probing' tasks, that use diagnostic classifiers to attempt to quantify the extent to which these encoders capture specific linguistic phenomena. The principle of probing has also resulted in extended evaluations that include relatively newer word-level pre-trained encoders. We build on probing tasks established in the literature and comprehensively evaluate and analyse – from a typological perspective amongst others – multilingual variants of existing encoders on probing datasets constructed for 6 non-English languages. Specifically, we probe each layer of a multiple monolingual RNN-based ELMo models, the transformer-based BERT's cased and uncased multilingual variants, and a variant of BERT that uses a cross-lingual modelling scheme (XLM).

1 Introduction

Recent trends in NLP have demonstrated the utility of pre-trained deep contextual representations in numerous downstream NLP tasks, where they have almost consistently resulted in significant performance improvements. Detailed evaluations have naturally followed: these have either been follow-up works to papers describing contextual representation systems, such as Peters et al. (2018b), or novel works evaluating a broad class of encoders on a broad variety of tasks (Perone et al., 2018). This paper is an example of the latter sort; we perform a comprehensive, large-scale evaluation of what linguistic phenomena these sequential encoders capture across a diverse set of languages. This has often been referred to in the literature as *probing*; we use this terminology throughout this work.

Briefly, our goals are to probe our encoders in a multilingual setting – i.e., we use a series of probing tasks to quantify what sort of linguistic information our encoders retain, and how this information varies across language, across encoder, and across task. As such, our experiments do not attempt to attain 'state-of-the-art' results; instead, we attempt to use a comparable experimental setting across each experiment, to quantify differences between settings rather than absolute results.

In Section 2, we describe prior work in multiple strands of research: specifically, on deep neural pre-training, on multilingualism in pre-training, and on evaluation. Section 3 describes both the linguistic features we probe our representations for, and how we generated our probing corpus. In Section 4, we describe and motivate our choice of encoders, as well as describe our infrastructural details. The bulk of our contribution is in Section 5, where we describe and analyse our results. Finally, we conclude with a discussion of the implications of these results and future work in Section 6.

2 Background

2.1 Deep pre-training

A watershed moment in NLP has been the recent innovation spree in deep pre-training; it has represented a considerable step up from shallow pre-training methods, that have been used in NLP since the introduction of contextual word embedding models such as word2vec (Mikolov et al., 2013). Whilst deep pre-training has been used in non-NLP, image-oriented tasks, where the standard paradigm is to pre-train deep convolutional networks on datasets like ImageNet (Russakovsky et al., 2014), and then fine-tune on task-specific data, their introduction to textual domains has been considerably slower, yet has been picking up rapidly in recent years.

An early paper in this theme was CoVe (McCann et al., 2017), that pre-trained contextual encoders on seq2seq machine translation models. Another earlier seminal work that addressed numerous technical issues with pre-training was Howard and Ruder's ULMFiT (2018). Not long after, the principle of deep pre-training saw widespread adoption with ELMo (Peters et al., 2018a), that consisted of several innovations over CoVe: critically, the use of an unsupervised (albeit structured) task – language modelling – for pre-training, and the use of a linear combination of all encoder layers, instead of just the top layer. Architecturally, ELMo used two-layer bidirectional LSTMs along with character-level convolutions, to model word probabilities given the history.

With deep pre-training having been established as a valid strategy in NLP, alternative models with different underlying architectures were proposed. The OpenAI GPT (Radford et al., 2018) was one such model; instead of LSTMs, it used the *de*coder of an attention-based transformer (Vaswani et al., 2017) as its underlying *en*coder – the justification being that using the transformer's *en*coder would lead to each token having access to succeeding tokens. The GPT also achieved (then) state-of-the-art results by plugging generated fixed-length vectors into downstream classifiers.

Another system that represented a significant innovation was BERT (Devlin et al., 2018). BERT introduced a language modelling variant, dubbed *masked* language modelling, that allowed them to use transformer encoders as their underlying encoding mechanism.

2.2 Multilingual pre-training

Multilingual variants of pre-trained encoders that provide contextual representations for non-English languages have also been studied; there is, however, some diversity in precisely how they are generated.

Che et al. (2018) provide ELMo models (Fares et al., 2017) for 44 languages; all of these were trained on data provided as part of the CoNLL 2018 shared task on dependency parsing Universal Dependencies treebanks (Zeman et al., 2018). This makes 'multilingual' a bit of a misnomer: whilst this is the most obvious approach to multilingual *support*, these models are all *mono*lingual. This also leads to other issues downstream, such as a complete inability to deal with true multilingual phenomena like code-switching. Throughout this text, however, when not specifically referring to ELMo, our use of the term 'multilingual' is inclusive of ELMo's quasi-multilingualism.

This is contrasted with BERT's approach to (true) multilingualism, which trains a single model that can handle all languages. The authors use WordPiece, a variant of BPE (Sennrich et al., 2016), for tokenisation, using a 110K-size vocabulary, and proceed to train a single gigantic model; they perform exponentially smoothed weighting of their data to avoid biasing their model towards better-resourced languages.

Finally, XLM (Lample and Conneau, 2019) is another cross-lingual encoder based on BERT that implements a number of modifications. Along with BERT's masked language modeling or Cloze task-based modelling (Devlin et al., 2018; Taylor, 1953), XLM training uses another similar objective during training that the authors call translation language modeling. Here, two parallel sentences are concatenated and words masked in both source and target sentences words are predicted using context from both. The authors here also use their own implementation of BPE – FastBPE, for which they provide a vocabulary of around 120K entries. This vocabulary is shared across all of the languages and thus improves the alignment of embedded spaces, as shown in Lample et al. (2017).

2.3 On evaluation

Evaluation of contextual representations goes beyond merely deep representations; not too far in the past, work on evaluating shallow sentence representations was encouraged by the release of

the SentEval toolkit (Conneau and Kiela, 2018), which provided an easy-to-use framework that sentence representations could be 'plugged' into, for rapid downstream evaluation on numerous tasks: these include several classification tasks, textual entailment and similarity tasks, a paraphrase detection task, and caption/image retrieval tasks. Relevant to our paper is Conneau et al.'s (2018a) set of 'probing tasks', a variant on the theme of diagnostic classification (Hupkes et al., 2017; Belinkov et al., 2017; Adi et al., 2016; Shi et al., 2016), that would attempt to quantify precisely what sort of linguistic information was being retained by sentence representations. Based in part on Shi et al. (2016), Conneau et al. (2018a) focus on evaluating representations for English; they provide Spearman correlations between the performance of a particular representation mechanism on being probed for specific linguistic properties, and the downstream performance on a variety of NLP tasks. Along similar lines, and contemporaneously with this work, Liu et al. (2019) probe similar deep pre-trained to the ones we do, on a set of 'sixteen diverse probing tasks'. (Tenney et al., 2018) probe deep pre-trained encoders for sentence structure.

On a different note, Saphra and Lopez (2018) present a CCA-based method to compare representation learning dynamics across time and models, without explicitly requiring annotated corpora.

A visible limitation of the datasets provided by these probing tasks is that most of them were created with the idea of evaluating representations built for English language data. Within the realm of evaluating *multilingual* sentence representations, Conneau et al. (2018b) describe the XNLI dataset, a set of translations of the development and test portions of the multi-genre MultiNLI inference dataset (Williams et al., 2018). This, in a sense, is an extension of a predominantly monolingual task to the multilingual domain; the authors evaluate sentence representations derived by *mapping* non-English representations to an English representation space.

2.4 BERTology

Relevant to the probing theme of this paper is the sudden recent growth in papers studying precisely what is retained with the internal representations of pre-trained encoders like BERT. These include, for instance, analyses of BERT's attentions heads, such as Michel et al. (2019), where the authors prune heads, often reducing certain layers to single heads, without a significant drop in performance in certain scenarios. Clark et al. (2019) provide a per-head analysis and attempt to quantify what information each head retains; they discover that specific aspects of syntax are well-encoded per head, and find heads that correspond to certain linguistic properties, such as heads that attend to direct objects of verbs. Other papers provide analyses of BERT's layers, such as Tenney et al. (2019), who discover that BERT's layers roughly correspond to the notion of the classical 'NLP pipeline', with lower level tasks such as tagging lower down the layer hierarchy. Hewitt and Manning (2019) define a structural probe over BERT representations, that extracts notions of syntax that correspond strongly to linguistic notions of dependency syntax.

3 Corpora

3.1 Probing

Our data consists of training, development and test splits for 9 linguistic tasks, that can broadly be grouped into surface, syntactic and semantic tasks. These are the same as the ones described in Conneau et al. (2018a), with minor modifications. Due to the differences in corpus domain, we alter some of their word-frequency parameters. We also exclude the top constituent (**TopConst**) task; we noticed that Wikipedia tended to have far less diversity in sentence structure than the original Toronto Books corpus, due to the more encyclopaedic style of writing. A brief description of the tasks follows, although we urge the reader to refer to the original paper for more detailed descriptions.

1. Sentence length: In **SentLen**, sentences are divided into multiple bins based on their length; the job of the classifier is to predict the appropriate bin, creating a 6-way classification task.

2. Word count: In **WC**, we sample sentences that feature exactly one amongst a thousand mid-frequency words, and train the classifier to predict the word: this is the most 'difficult' task, in that it has the most possible classes.

3. Tree depth: The **TreeDepth** task simply asks the representation to predict the depth of the sentence's syntax tree. Unlike the original

paper, we use the depth the of the dependency tree instead of the constituency tree.

4. Bigram shift: In **BiShift**, for half the sentences in the dataset, the order of words in a randomly sampled bigram is reversed. The classifier learns to predict whether or not the sentence contains a reversal.

5. Subject number: The **SubjNum** task asks the classifier to predict the number of the subject of the head verb of the sentence. Only sentences with exactly one subject (annotated with the `nsubj` relation) attached to the root verb were considered.

6. Object number: **ObjNum**, similar to the subject number task, was annotated with the number of the direct object of the head verb (annotated with the `obj` relation).

7. Coordination inversion: In **CoordInv**, two main clauses joined by a coordinating conjunction have their orders reversed, with a probability of one in two. Only sentences with exactly two top-level conjuncts are considered.

8. (Semantic) odd man out: **SOMO**, one of the more difficult tasks in the collection, replaces a randomly sampled word with another word with comparable corpus bigram frequencies.

9. Tense prediction: The **Tense** prediction asks the classifier to predict the tense of the main verb: we compare the past and present tenses.

3.2 Data

Languages

Our choice of languages was motivated by three factors: i) the availability of a Wikipedia large enough to extract data from; ii) the availability of a reasonable dependency parsing model, and iii) typological diversity. The former, in particular, was a bit of a restriction, since not all sentences were valid candidates for extraction per task. Our final set of languages include an additional corpus for English, as well as French, German, Spanish, Russian, Turkish and Finnish. Whilst not nearly representative of the diversity of world languages, this selection includes morphologically agglutinative, fusional and (relatively) isolating languages, and it includes two scripts, Latin and Cyrillic. The languages also represent three families (Indo-European, Turkic and Uralic).

We build our probing datasets using the relevant language's Wikipedia dump as a corpus. Our motivation for doing so was that it a freely available corpus for numerous languages, large enough to extract the sizeable corpora that we need. Specifically, we use Wikipedia dumps (dated 2019-02-01), which we process using the WikiExtractor utility[1].

Preprocessing

We use the Punkt tokeniser (Kiss and Strunk, 2006) to segment our Wikipedia dumps into discrete sentences. For Russian, which lacked a Punkt tokenisation model, we used the UDPipe (Straka and Straková, 2017) toolkit to perform segmentation.

Having segmented our data, we used the Moses (Koehn et al., 2007) tokeniser for the appropriate language, falling back to English tokenisation when unavailable.

Next, we obtained dependency parses for our sentences, again using the UDPipe toolkit's pretrained models, trained on Universal Dependencies treebanks (Nivre et al., 2015). We then processed these dependency parsed corpora to extract the appropriate sentences; while in principle, each task was meant to have 120K sentences, with 100K/10K/10K training/validation/test splits, often, for the rarer linguistic phenomena, we ran out of source data, in particular with Turkish and Finnish, although to a smaller extent with Russian as well. In these situations, we ensured an equivalent split ratio.

Our use of non-gold-standard dependency parses implies inaccuracies that, in principle, would propagate to our training data. A valid counterargument, however, is that we do not rely on complete parse accuracies for all our tasks; several tasks do not require dependency or POS annotation, and the ones that do rely on a fixed subset of dependency relations, such as `nsubj` or `obj`. Having said that, we do acknowledge the divergences in parsing performance across language; unfortunately, given the substantial corpus sizes these experiments require, we could not use gold-standard parsed corpora.

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

4 Implementation

4.1 Encoders

We probe several popular pre-trained encoders (or, specifically, their multilingual variants). These include:

ELMo, monolingual We use Che et al.'s (2018) pre-trained monolingual ELMo models for each of our languages. Training was similar to the original English language ELMo, but allows for Unicode, and uses a sample softmax (Jean et al., 2014) to deal with large vocabularies. We probed four variants of each ELMo model - the character embeddings layer, the two LSTM layers, and an average of all three. For obtaining a fixed-length sentence representation, we use average pooling over the sequence of hidden states.

BERT We use the two multilingual variants - cased and uncased. Both variants have 12 layers, 768 hidden units, 12 heads and 110M parameters; the former includes 104 languages and fixes normalisation issues, whilst the latter includes 102 languages. For further classification, we use the first hidden state, represented by the `[CLS]` token.

XLM We probe only one variant of this encoder - i.e., the models fine-tuned on XNLI (Conneau et al., 2018b) data. Due to there being no XNLI data for Finnish, we do not probe our Finnish dataset with XLM. Unlike BERT, XLM uses 1024 hidden units and 8 heads.

Unfortunately, all our encoders did include Wikipedia dumps in their training data. Given that pretrained encoders tend to use as much easily accessible data as possible in pre-training, however, it is difficult to avoid using a completely unseen corpus for probing task extraction.

4.2 Implementation

Our probing procedure for each of our languages and encoders is relatively similar: we use a multilayer perceptron based classifier to assign the appropriate class label to each input sentence. During training, the encoders remain static, with all learning restricted to the classifier. In an attempt to avoid excessively complex classifiers, and to ensure consistency across tasks and languages, we use predetermined fixed hyperparameters – specifically, a sigmoid activation function, on top of a size 50 dense layer. We use a training batch size of 32, optimised using Adam (Kingma and Ba, 2014), and train for 10 epochs, allowing for early stopping.

We implement our system using the AllenNLP toolkit (Gardner et al., 2018), which crucially provides the ability to use the appropriate tokenisation schema, along with the appropriate vocabulary, for each encoder. Training and evaluation were carried out on NVIDIA RTX 2080 Ti GPUs, with 10GiB GPU memory.

5 Results

Due to our large experiment space, there are several dimensions along which our results can be analysed and discussed. For ease of analysis, all our figures are presented as heatmaps.

We have presented our results in two ways, for easy visualisation. The first of these is dividing them up by task, as in Figure 1. We present an alternative set of results for three of our encoders, in Figure 2.

5.1 Encoder

An observation that instantly stands out is the significant difference in performance on WC: consistently, across every language, all our transformer-based architectures see results very close to 0. Further, whilst not instantly visible in Figure 2, a quick look at Figure 1 shows that the same appears to hold (albeit to a lesser extent) for SentLen, TreeDepth and BiShift, all of which are either surface or syntactic phenomena. This appears to heavily imply that recurrent, sequential processing appears to retain lower level linguistic phenomena better than self-attentive mechanisms (that do not see the same drop in informativity for semantic tasks). This is perhaps a bit easy to justify with SentLen, which is a phenomenon that is directly proportional to recurrence depth.

The next phenomenon of interest is the difference between each of ELMo's layers. Interestingly, these do not appear to be as drastic as one would imagine, given the differences in performance on downstream tasks. The difference between raw word representations and actual contextual representations is fairly noticeable, particularly on the strongly syntactic BiShift. However, the differences between higher layers is rela-

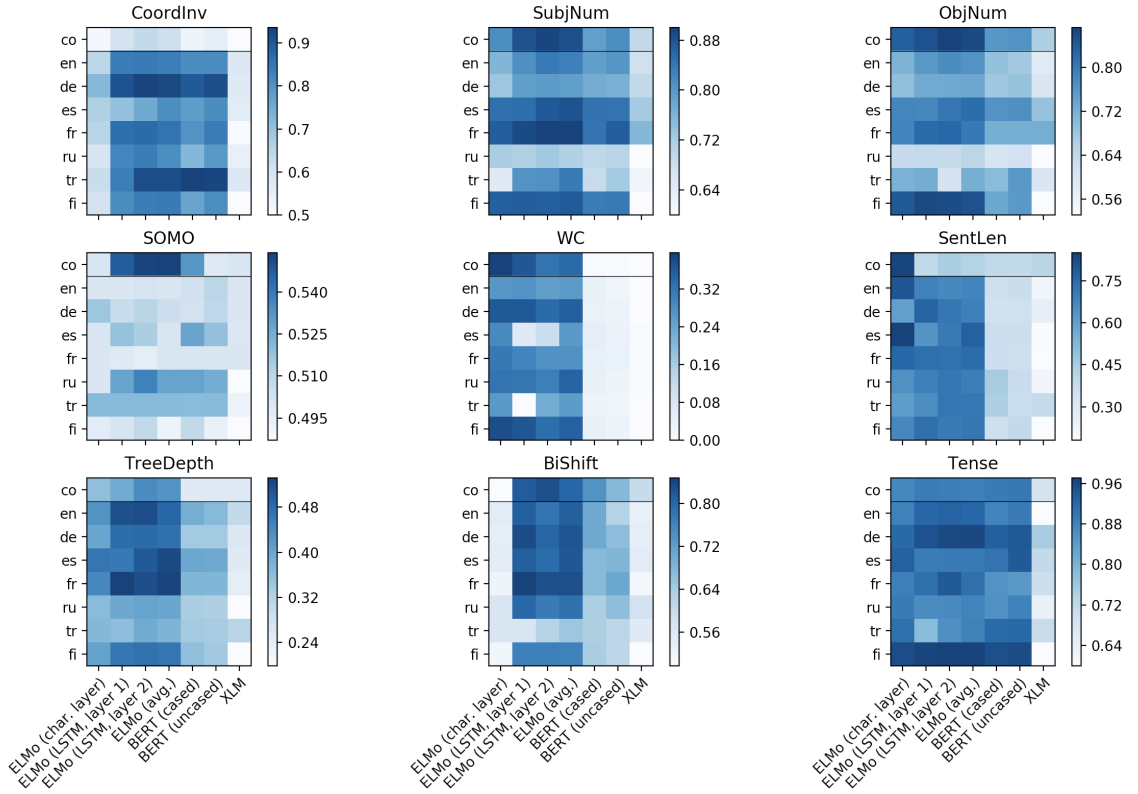

Figure 1: Detailed results per task, per language per encoder. Each task's result heatmap has its own scale. All results mentioned in this paper refer to classification accuracies in [0.0, 1.0]. Henceforth, 'co' refers to probing results on Conneau et al.'s (2018a) original corpus.

tively murkier, and whilst the average of the three does appear to represent some phenomena better (such as CoordInv), it isn't clear that this difference is meaningful. Notably, SentLen appears to be poorly represented in higher layers, which ties in with other analyses of ELMo (Peters et al., 2018b), that imply that higher layers are likelier to learn more semantic features.

BERT's cased variant appears to retain information slightly better than the uncased one, which is in line with the authors' descriptions of their own models.

Finally, and perhaps most interestingly, we turn our attention towards XLM. Despite being based on BERT (and indeed showing similar *patterns* in performance), XLM appears to perform a lot worse than all our other encoders on virtually every task. It is not immediately clear why: however, given that this drop in performance is visible in every language, our conjecture is that due to the translation-based modelling employed by XLM, the encoder does indeed succeed at learning language-independent representations, or 'universal' representations. However, this universality comes at a cost: in an attempt to adequately represent a variety of typologically diverse languages, XLM appears to lose its ability to retain *specific* linguistic phenomena pertaining to specific languages; in a sense, it is incapable of building a representation for a language that adequately captures a specific phenomenon in that language *and no other*. This follows intuitively from the method used training on the TLM objective: the authors concatenate aligned parallel sentences and predict masked words in the source *and* the target sentence, using context from both sentences at the same time to predict each masked word. This is likely to have had a detrimental effect on XLM's ability to retain characteristics specific to each language. In Figure 3, we show the relative performance of BERT and XLM per probing task. There is a clear trend towards BERT's enhanced retention of linguistic features being less prominent for the more semantic tasks, which fits our hypothesis, as semantics are likelier to hold cross-linguistically.

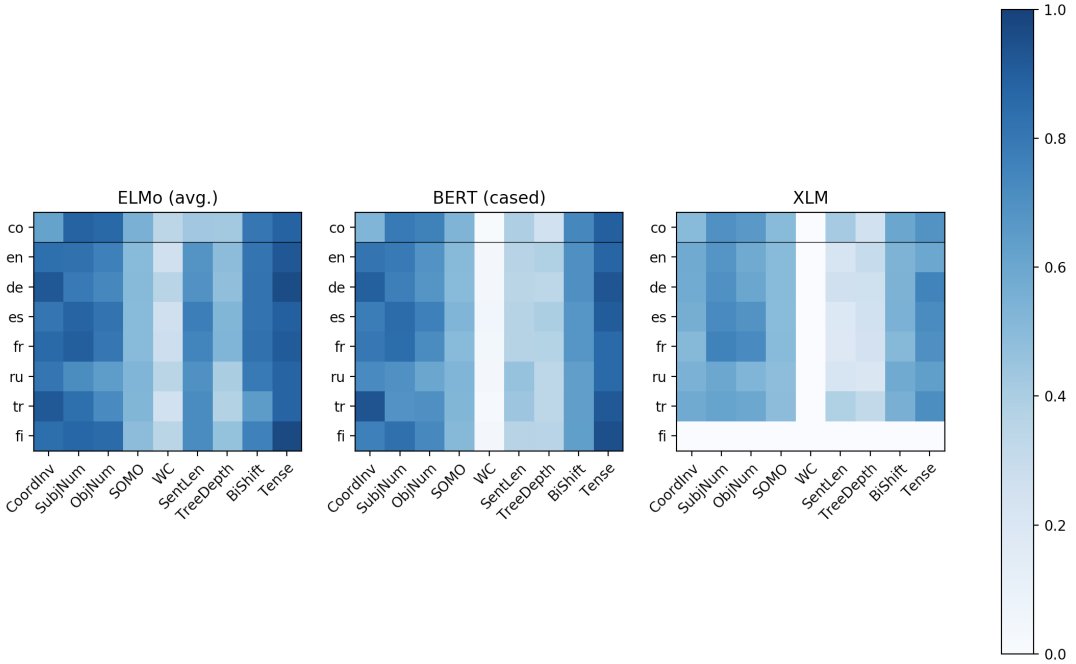

Figure 2: Results for select encoders, per language per task. All results use the same scale, [0.0, 1.0].

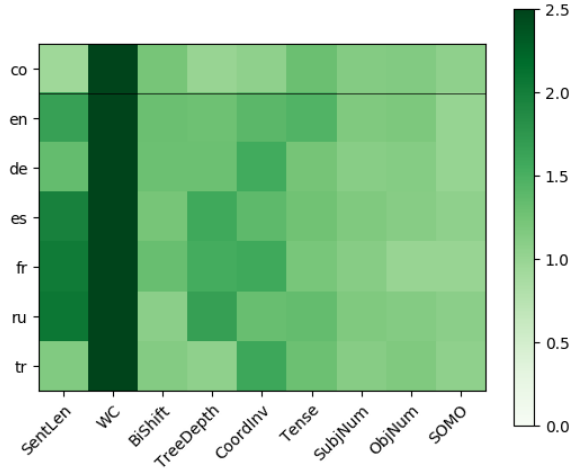

Figure 3: BERT (cased) scores divided by the corresponding XLM scores. Tasks are ordered, from surface to syntax to semantic level tasks.

A point to be made here is that despite SubjNum, ObjNum and Tense being classified as semantic tasks, it isn't clear that they are truly being probed for semantic information: all three phenomena tend to be visible with morphological marking. This gives us an alternative justification for XLM's relative improvement in retention: XLM is likely capable of storing each language's individual morphological information in different internal subspaces de, as each language is likely to reflect morphology purely orthographically, and in mutually exclusive ways.

Our observations on the differences between encoders are also easily visible in Figure 4, where multiple 'belts' of varying performances emerge.

5.2 Language

To motivate one of the main focuses of this paper – our analysis of our results along linguistic lines – we present Figure 5, which displays what one might call the net 'informativity' of an encoder, i.e. an average of how much information each encoder retains averaged over tasks. The most noticeable effect here is the drop in informativity for Russian and Turkish. While this is perhaps understandable for Turkish – which has smaller probing corpora, and a less reliable Wikipedia than the other languages – Russian's opaqueness cannot be

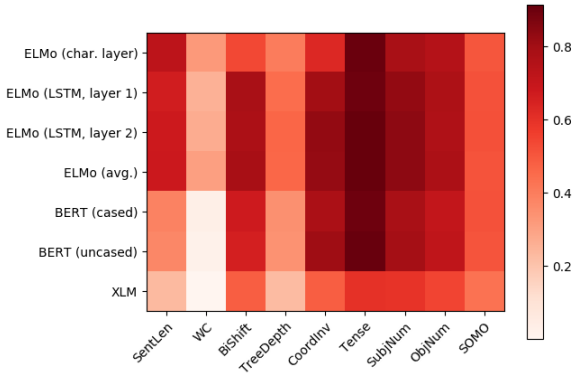

Figure 4: Linguistic information retained per encoder, per task; scores are averaged over language.

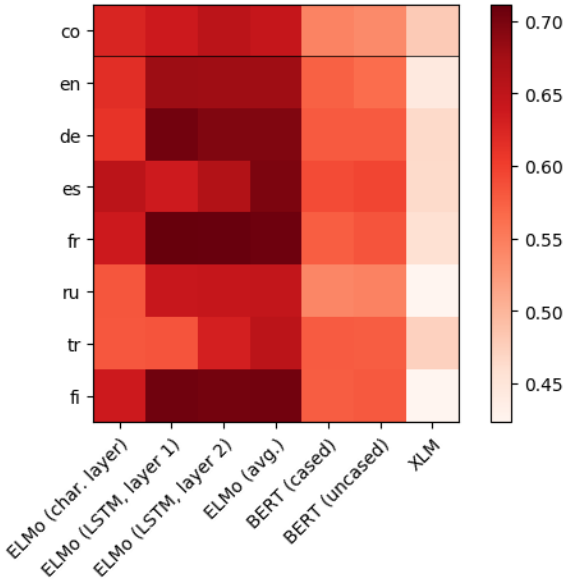

Figure 5: Net encoder 'informativity' per language; results averaged over all tasks.

as easily explained away, particularly when contrasted with Finnish, which tends to have fewer resources.

We further introduce Figure 6, which displays the averaged results of three systems – ELMo's multilayer variant, BERT's cased variant and (absent for Finnish) XLM. Most linguistic differences appear to be clustered in the semantic part of this heatmap. There are numerous possible factors that could explain these divergences, not the least of which is the actual probing corpus itself: however, we attempt to provide a justification, from a typological perspective, for some of these results.

When averaged across encoders, the Tense task stands out as fairly easy to probe for all languages. It thus seems that information about verbal temporal properties is retained in the sentence representation. For the tasks of subject and object number, however, we observe clear differences between the languages. Here, French and Spanish appear to be somewhat easier to probe than other languages. We hypothesise that this is due to both languages marking nominal number, not just with verb agreement, but also with plural articles, resulting in representations that are more informative regarding number. Contrast this with English and German, which either do not have plural articles, or have plural articles that morphologically overlap with non-plural forms, or with Russian, that tends to avoid articles in general.

Other interesting observations are German's relative ability at retaining information on CoordInv and Tense, as well as Finnish's extraordinarily high performance on Tense. Further, SentLen appears to be retained better, counter-intuitively, in Russian, Turkish and Finnish; a brief look at Figure 1 shows that, interestingly, this is likely due to BERT.

Finally, we note that our results do not seem to indicate that English is somehow better represented in our multilingual systems, nor does it appear to perform significantly better than other languages in general, indicating that none of our models are 'learning' English first and then adapting to other languages.

5.3 Task

From a monolingual perspective, most of what needs to be said regarding the choice of probing tasks has already been said in the original (Conneau et al., 2018a). There are however several differences, induced both by our modifications to the original framework, and by our corpus's multilingualism.

The first of these is the apparently consistent differences in performance on certain tasks which include, amongst others, CoordInv, where our variant appears to be more easily retained than the original. This can be explained away by minor issues we faced during implementation, using dependency trees instead of constituency trees. Due to more complicated representation of conjuncts in UD-style dependency trees, some of our sentences had issues with using the appropriate casing after swapping conjuncts, as well as ensuring consistent punctuation. While we attempted to avoid these by

writing filtering rules, these were imperfect, and it is likely that stray punctuation and the like might have informed our representations about the conjuncts being swapped, in some instances.

Another task with minor differences is our implementation of SOMO; we attribute this to not being able to accurately reproduce Conneau et al.'s (2018a) modified corpus-frequency range (40-400) to adequately fit all our corpora.

We note that there do not appear to be significant differences in the TreeDepth task, despite our using dependency trees instead of constituency, and despite our tree depth/sentence length decorrelation procedure being markedly simpler.

6 Discussion

6.1 Implications

Having elaborated our results, it becomes crucial to contextualise their importance. 'Probing' an encoder, or more correctly, using diagnostic classifiers to attempt to *quantify* what information an encoder stores, appears to be a reasonable approach to *qualifying* this information. However, there has been some critique of this approach. To paraphrase Saphra and Lopez (2018), the architecture of a diagnostic classifier does affect the performance of a probing task; further, lower layers of encoders may represent information in ways designed to be picked up on by their own higher layers; this might prove difficult for simple classifiers to truly probe.

This is an excellent critique of the principle using *absolute* probing performance, or *absolute* numbers representing performance on an abstract insight task, as a yardstick. Critically, this work is focussed, both practically and in principle, on elucidating *relative* results, in a wide space of languages and encoders. The relative underparameterisation of the classifier and the use of one constant set of hyperparameters across experiments is an attempt to minimise the *relative* interference of the classifier. i.e., our goal is to keep the classifier's interference – its lens – as consistent as possible.

6.2 Future work

One potential strand of research relates directly to the tasks themselves: our choice of tasks was fairly restrictive, and does not include many tasks that are truly *semantic*, which does not provide us with enough information to draw conclusions similar to Liu et al. (2019), which is that pretrained models encode stronger syntax than semantics. An obvious goal, therefore, is the more careful design of tasks, particularly within a multilingual context: the tasks proposed by Liu et al. (2019) and Tenney et al. (2018) are not strictly easy to motivate cross-linguistically due to the burden of annotation. This could include more semantic-level probing by means of existing cross-lingual semantic resources, such as the Parallel Meaning Bank (Abzianidze et al., 2017).

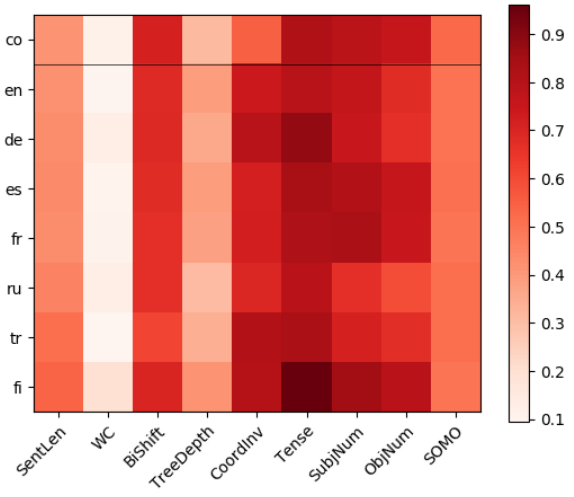

Figure 6: Linguistic insight per language per task, averaged over one variant of every encoder: multilayer ELMo, cased BERT, and XLM (bar Finnish).

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. *arXiv:1702.03964 [cs]*. ArXiv: 1702.03964.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *arXiv:1608.04207 [cs]*. ArXiv: 1608.04207.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng,

and Ting Liu. 2018. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. *arXiv:1807.03121 [cs]*. ArXiv: 1807.03121.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. *arXiv:1803.05449 [cs]*. ArXiv: 1803.05449.

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

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv:1809.05053 [cs]*. ArXiv: 1809.05053.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. *arXiv:1803.07640 [cs]*. ArXiv: 1803.07640.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *arXiv:1801.06146 [cs, stat]*. ArXiv: 1801.06146.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2017. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *arXiv:1711.10203 [cs]*. ArXiv: 1711.10203.

Sbastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On Using Very Large Target Vocabulary for Neural Machine Translation. *arXiv:1412.2007 [cs]*. ArXiv: 1412.2007.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. ArXiv: 1412.6980.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *arXiv:1901.07291 [cs]*. ArXiv: 1901.07291.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. page 22.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. *arXiv:1708.00107 [cs]*. ArXiv: 1708.00107.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *arXiv preprint arXiv:1905.10650*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan

Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2015. Universal dependencies 1.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Christian S. Perone, Roberto Silveira, and Thomas S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv:1806.06259 [cs]*. ArXiv: 1806.06259.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. *arXiv:1802.05365 [cs]*. ArXiv: 1802.05365.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting Contextual Word Embeddings: Architecture and Representation. *arXiv:1808.08949 [cs]*. ArXiv: 1808.08949.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. page 12.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575 [cs]*. ArXiv: 1409.0575.

Naomi Saphra and Adam Lopez. 2018. Understanding Learning Dynamics Of Language Models with SVCCA. *arXiv:1811.00225 [cs]*. ArXiv: 1811.00225.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Wilson L Taylor. 1953. cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.