Analyzing Encoded Concepts in Transformer Language Models

Hassan Sajjad[◇] **Nadir Durrani**[◇] **Fahim Dalvi**[◇] **Firoj Alam**[◇] **Abdul Rafae Khan**[†] **Jia Xu**[†]
{hsajjad,ndurrani,faimaduddin,fialam}@hbku.edu.qa
[◇]Qatar Computing Research Institute, HBKU Research Complex, Qatar

{akhan4,jxu70}@stevens.edu
[†]School of Engineering and Science, Steven Institute of Technology, USA

Abstract

We propose a novel framework ConceptX, to analyze how latent concepts are encoded in representations learned within pre-trained language models. It uses clustering to discover the encoded concepts and explains them by aligning with a large set of human-defined concepts. Our analysis on seven transformer language models reveal interesting insights: i) the latent space within the learned representations overlap with different linguistic concepts to a varying degree, ii) the lower layers in the model are dominated by lexical concepts (e.g., affixation), whereas the core-linguistic concepts (e.g., morphological or syntactic relations) are better represented in the middle and higher layers, iii) some encoded concepts are multi-faceted and cannot be adequately explained using the existing human-defined concepts.[1]

1 Introduction

Contextualized word representations learned in deep neural network models (DDNs) capture rich concepts making them ubiquitous for transfer learning towards downstream NLP. Despite their revolution, the blackbox nature of the deep NLP models is a major bottle-neck for their large scale adaptability. Understanding the inner dynamics of these models is important to ensure fairness, robustness, reliability and control.

A plethora of research has been carried out to probe DNNs for the linguistic knowledge (e.g. morphology, syntactic and semantic roles) captured within the learned representations. A commonly used framework to gauge how well linguistic information can be extracted from these models is the *Probing Framework* (Hupkes et al., 2018), where they train an auxiliary classifier using representations as features to predict the property of interest. The performance of the classifier reflects the amount of knowledge learned within representations. To this end, the researchers have analyzed what knowledge is learned within the representations through relevant extrinsic phenomenon varying from word morphology (Vylomova et al., 2016; Belinkov et al., 2017a) to high level concepts such as syntactic structure (Blevins et al., 2018; Marvin and Linzen, 2018) and semantics (Qian et al., 2016; Reif et al., 2019; Belinkov et al., 2017b) or more generic properties (Adi et al., 2016; Rogers et al., 2020).

In this work, we approach the representation analysis from a different angle and present a novel framework ConceptX. In contrast to relying on the prediction capacity of the representations, we analyze the latent concepts learned within these representations and how knowledge is structured, using an unsupervised method. More specifically, we question: i) do the representations encode knowledge inline with linguistic properties such as word morphology and semantics? ii) which properties dominate the overall structure in these representations? iii) does the model learn any novel concepts beyond linguistic properties? Answers to these questions reveal how deep neural network models structure language information to learn a task.

Our inspiration to use the term *concept* comes from *"concept based explanation"* in computer vision (Kim et al., 2018; Ghorbani et al., 2019; Chen et al., 2020). Stock (2010) defined a concept as "a class containing certain objects as elements, where the objects have certain properties". We define an *encoded concept as a cluster of context-aware latent representations of words, where the representations are encoder layer outputs.*

Our framework clusters contextualized representations using agglomerative hierarchical clustering (Gowda and Krishna, 1978). The resulting clusters represent *encoded concepts*, captured within the learned representations (Please see Figure 1 for illustration). We then use a novel align-

[1]The code is available at https://github.com/hsajjad/ConceptX.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3082 - 3101
July 10-15, 2022

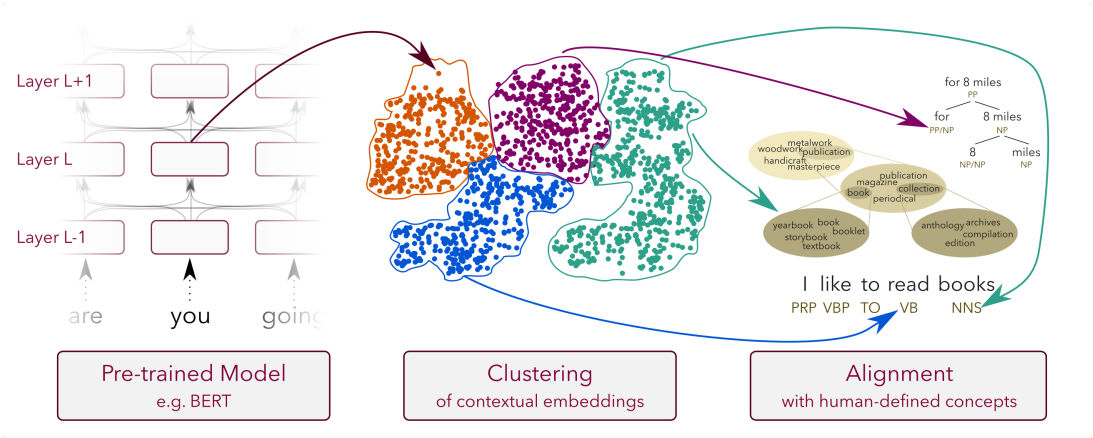


Figure 1: **ConceptX**: i) Extract representations from trained model, ii) Cluster the representations to obtain encoded concepts, iii) Align the concepts to human-defined concepts

ment function that measures the amount of overlap between *encoded concepts* and a range of predefined categories (that we call as *human-defined concepts* in this paper). We experimented with affixes, casing, morphological, syntactic, semantic, WordNet (Miller, 1995), and psycholinguistic concepts (LIWC Pennebaker et al. (2001)). The use of such a diverse set of human-defined concepts enables us to cover various abstractions of language. In Figure 3 we present a few examples of human-defined concepts that were aligned with the encoded concepts.

We carry out our study on seven pre-trained transformer models such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), with varying optimization functions, architectural details and training data. Some notable findings emerging from our analysis are as follows:

- Shallow concepts such as lexical ngrams or suffixes are predominantly captured in the lower layers of the network.
- WordNet and psycholinguistic-based concepts (LIWC) are also learned in the lower layers.
- Middle and higher layers encode concepts that capture core linguistic properties such as morphology, semantics and syntax.
- Roughly 50% of the encoded concepts adhere to our suite of human-defined linguistic concepts.
- The models learn novel concepts that are multi-faceted and cannot be adequately explained using the existing human-defined concepts.

Our contributions in this paper are as follow: i) We present `ConceptX`, a framework that interprets encoded concepts in the learned representation by measuring their alignment to the human-defined concepts. ii) We provide a qualitative and quantitative evidence of how knowledge is structured within deep NLP models with respect to a large suite of human-defined concepts.

2 Related Work

Most of the work done on interpretability in deep NLP addresses two questions in particular: *(i)* what linguistic (and non-linguistic) knowledge is learned within contextualized representations, *Concept Analysis* and *(ii)* how this information is utilized in the decision making process, *Attribution Analysis* (Sajjad et al., 2021). The former thrives on post-hoc decomposability, where we analyze representations to uncover linguistic phenomenon that are captured as the network is trained towards any NLP task (Adi et al., 2016; Conneau et al., 2018; Liu et al., 2019a; Tenney et al., 2019; Belinkov et al., 2020) and the latter characterize the role of model components and input features towards a specific prediction (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018). Our work falls into the former category.

Previous studies have explored visualization methods to analyze the learned representations (Karpathy et al., 2015; Kádár et al., 2017), attention heads (Clark et al., 2019; Vig, 2019), language compositionality (Li et al., 2016) etc. A more commonly used framework analyzes representations by correlating parts of the neural network with linguistic properties, by training a classifier to predict a

feature of interest (Adi et al., 2016; Belinkov et al., 2017a; Conneau et al., 2018). Several researchers used probing classifiers for investigating the contextualized representations learned from a variety of neural language models on a variety of character- (Durrani et al., 2019), word- (Liu et al., 2019a) or sub-sentence level (Tenney et al., 2019) linguistic tasks. Rather than analyzing the representations as a whole, several researchers also explored identifying salient neurons within the model that capture different properties (Dalvi et al., 2019a; Durrani et al., 2020; Suau et al., 2020; Mu and Andreas, 2020) or are salient for the model irrespective of the property (Bau et al., 2019; Wu et al., 2020).

Our work is inline with (Michael et al., 2020; Dalvi et al., 2022), who analyzed latent concepts learned in pre-trained models. Michael et al. (2020) used a binary classification task to induce latent concepts relevant to a task and showed the presence of linguistically motivated and novel concepts in the representation. However, different from them, we analyze representations in an unsupervised fashion. Dalvi et al. (2022) used human-in-the-loop to analyze latent spaces in BERT. Our framework uses human-defined concepts to automatically generate explanations for the latent concepts. This enabled us to scale our study to many transformer models.

In a similar work, Mamou et al. (2020) applied manifold analysis technique to understand the amount of information stored about object categories per unit. Our approach does away from the methodological limitations of probing framework such as complexity of the probes, effect of randomness etc (Belinkov, 2021). However, it is important to mention that the two frameworks are orthogonal and complement each other.

3 Methodology

A vector representation in the neural network model is composed of feature attributes of the input words. We group the encoded vector representations using a clustering approach discussed below. The underlying clusters, that we term as the *encoded concepts*, are then matched with the human-defined concepts using an alignment function. Formally, consider a Neural Network (NN) model $\mathbb{M}$ with L encoder layers $\{l_1, l_2, ...l_l, ..., l_L\}$, with H hidden nodes per layer. An input sentence consisting of M words $w_1, w_2, ...w_i, ..., w_M$ is fed into a NN. For each input word i, we compute the node output (after applying the activation functions) $y_h^l(w_i)$ of every hidden node $h \in \{1, ..., H\}$ in each layer l, where $\overrightarrow{y}^l(w_i)$ is the vector representation composing the outputs of all hidden nodes in layer l for w_i. Our goal is to cluster representations $\overrightarrow{y}^l$, from a large training data to obtain *encoded concepts*. We then align these with various human-defined concepts to obtain an explanation of them to build an understanding of how these concepts are represented across the network.

3.1 Clustering

We use agglomerative hierarchical clustering (Gowda and Krishna, 1978), which we found to be effective for this task. It assigns each word to a separate cluster and then iteratively combines them based on Ward's minimum variance criterion that minimizes intra-cluster variance. Distance between two representations is calculated with the squared Euclidean distance. The algorithm terminates when the required K clusters (aka encoded concepts) are formed, where K is a hyperparameter. Each encoded concept represents a latent relationship between the words present in the cluster. Appendix C presents the algorithm.

3.2 Alignment

Now we define the alignment function between the encoded and human-defined concepts. Consider a human-defined concept as z, where a function $z(w) = z$ denotes that z is the human-defined concept of word w. For example, parts-of-speech is a human-defined concept and each tag such as noun, verb etc. represents a class/label within the concept, e.g. $z(sea) = noun$. Similarly, suffix is a human-defined concept with various suffixes representing a class, e.g. $z(bigger) = er$. A reverse function of z is a one-to-many function that outputs a set of unique words with the given human-defined concept, i.e., $z^{-1}(z) = \{w_1, w_2, \ldots, w_J\}$, like $z^{-1}(noun) = \{sea, tree, \ldots\}$, where J is the total number of words with the human-defined concept of z. Following this notation, an encoded concept is indicated as c, where $c(w) = c$ is a function of applying encoded concept on w, and its reverse function outputs a set of unique words with the encoded concept of c, i.e., $c^{-1}(c) = \{w_1, w_2, \ldots, w_I\}$, where I is the set size.

To align the encoded concepts with the human-defined concepts, we auto-annotate the input data that we used to get the clusters, with the human-defined concepts. We call our encoded concept (c)

to be θ-aligned (Λ_θ) with a human-defined concept (z) as follows:

$$\Lambda_\theta(z,c) = \begin{cases} 1, & \text{if } \frac{\sum_{w' \in z^{-1}} \sum_{w \in c^{-1}} \delta(w,w')}{J} \geq \theta \\ 0, & \text{otherwise,} \end{cases}$$

where Kronecker function $\delta(w, w')$ is defined as

$$\delta(w,w') = \begin{cases} 1, & \text{if } w = w' \\ 0, & \text{otherwise} \end{cases}$$

We compute c and $\Lambda_\theta(z, c)$ for the encoder output from each layer l of a neural network. To compute a network-wise alignment, we simply average θ-agreement over layers.

4 Experimental Setup

4.1 Dataset

We used a subset of WMT News 2018[2] (359M tokens) dataset. We randomly selected 250k sentences from the dataset (≈5M tokens) to train our clustering model. We discarded words with a frequency of less than 10 and selected maximum 10 occurrences of a word type.[3] The final dataset consists of 25k word types with 10 contexts per word.

4.2 Pre-trained Models

We carried out our analysis on various 12-layered transformer models such as BERT-cased (BERT-c, Devlin et al., 2019), BERT-uncased (BERT-uc), RoBERTa (Liu et al., 2019b), XLNet (Yang et al., 2019) and ALBERT (Lan et al., 2019). We also analyzed multilingual models such as multilingual-bert-cased (mBERT) and XLM-RoBERTa (XLM-R, Conneau et al., 2020) where the embedding space is shared across many languages. This choice of models is motivated from interesting differences in their architectural designs, training data settings (cased vs. un-cased) and multilinguality.

4.3 Clustering and Alignment

We extract contextualized representation of words by performing a forward pass over the network using the NeuroX toolkit (Dalvi et al., 2019b). We cluster representations in every layer into K groups. To find an optimum value of K, we experimented with the ELbow (Thorndike, 1953) and Silhouette (Rousseeuw, 1987) methods. However, we did not observe reliable results (see Appendix C). Therefore, we empirically selected $K = 1000$ based on finding a decent balance between many small clusters (over-clustering) and a few large clusters (under-clustering). We found that our results are not sensitive to this parameter and generalize for different cluster settings (See Section 5.4). For the alignment between encoded and human-defined concepts, we use $\theta = 90\%$ i.e., we consider an encoded concept and a human-defined concept to be aligned, if they have at least 90% match.

4.4 Human-defined concepts

We experiment with the various **Human-defined concepts**, which we categorize into four groups:

- *Lexical Concepts:* Ngrams, Affixes, Casing, First and the Last Word (in a sentence)
- *Morphology and Semantics:* POS tags (Marcus et al., 1993) and SEM tags (Abzianidze et al., 2017)
- *Syntactic:* Chunking tags (Tjong Kim Sang and Buchholz, 2000) and CCG super-tags (Hockenmaier, 2006)
- *Linguistic Ontologies:* WordNet (Miller, 1995) and LIWC (Pennebaker et al., 2001)

At various places in this paper, we also refer to Morphology, Semantics and Syntactic concepts as core-linguistic concepts. We trained BERT-based classifiers using gold-annotated training data and standard splits for each core-linguistic concepts and auto-labelled the selected news dataset using these.[4]

5 Analysis

In this section, we analyze the encoded concepts by aligning them with the human-defined concepts.

5.1 Overall Alignment

First we present to what extent the encoded concepts in the entire network align with the human-defined concepts. We compute the overall score as the percentage of the aligned encoded concepts to the human-defined concepts across layers using the function described in Section 3.2. We

[2] http://data.statmt.org/news-crawl/en/

[3] Our motivation to select a small subset of data and limiting the number of tokens is as follows: clustering a large number of high-dimensional vectors is computationally and memory intensive, for example 200k vectors (of size 768 each) require around 400GB of CPU memory. Applying transformations (e.g., PCA) to reduce dimensionality may result in loss of information and therefore undesirable. We wanted to stay true to the original embeddding space.

[4] Please see Appendix B for details.

	BERT-c	BERT-uc	mBERT	XLM-R	RoBERTa	ALBERT	XLNet
Overall alignment	47.2%	50.4%	66.0%	72.4%	50.1%	51.6%	43.6%

Table 1: Coverage of human-defined concepts across all clusters of a given model

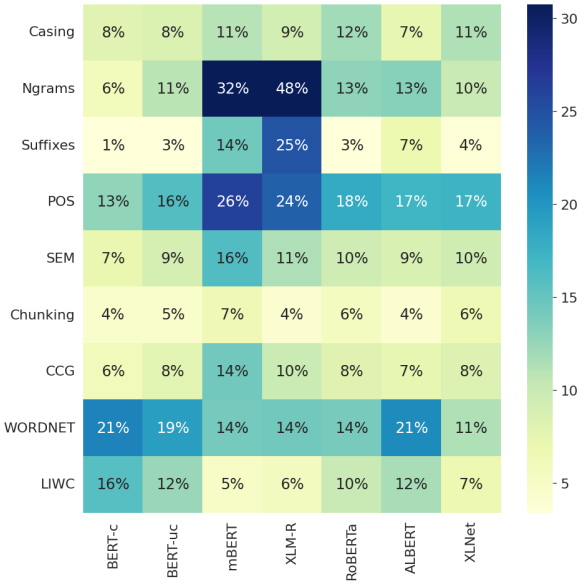


Figure 2: Average Alignment (%) between encoded concepts and human-defined concepts

found an overall match of at least 43.6% in XLNet and at most 72.4% in XLM-R (See Table 1). Interestingly, the multilingual models (mBERT and XLM-R) found substantially higher match than the monolingual models. The inclusion of multiple languages during training causes the model to learn more linguistic properties. Note that the extent of alignment with the human-defined concept may not necessarily correlate with its overall performance. For example XLNet performs outperforms BERT on the GLUE tasks, but aligns less with the human-defined concepts compared to BERT in our results. A similar observation was made by Belinkov et al. (2020) who also found that the translation quality of an NMT model may not correlate with the amount of linguistic knowledge learned in the representation. Various factors such as: architectural design, training data, objective function, initialization, etc, play a role in training a pre-trained model. More controlled experiments are needed to understand the relationship of each factor on the performance of the model and on the linguistic learning of the model.

We further investigated per concept[5] alignment to understand which human-defined concepts are better represented within the encoded concepts. Figure 2 presents the results.

Lexical Concepts Pre-trained models encode varying amount of lexical concepts such as casing, ngrams and suffixes. We found between 7-11% encoded concepts that align with the casing concept (title case or upper case). We observed that most of these encoded concepts consist of named entities, which were grouped together based on semantics.

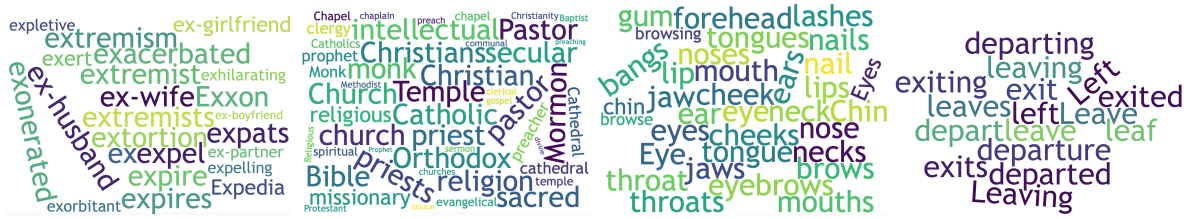
Comparing suffixes and ngrams While affixes often have linguistic connotation (e.g., the prefix *anti* negates the meaning of the stem and the suffix *ies* is used for pluralization), the ngram units that become part of the vocabulary as an artifact of statistical segmentation (e.g., using BPE (Sennrich et al., 2016) or Word-piece (Schuster and Nakajima, 2012)) often lack any linguistic meaning. However, models learn to encode such information. We found a match ranging from 1% (BERT-cased) up to 25% (XLM-R) when comparing encoded concepts with the suffix concept. A similar pattern is observed in the case of the ngram concept (which is a superset of the suffix concept) where a staggering 48% matches were found. Figure 6a shows an ngram cluster found in layer 2 of BERT-c.[6]

Morphology and Semantics We found that the encoded concepts based on word morphology (POS) consistently showed a higher match across all models in comparison to the other abstract concepts, aligning a quarter of the encoded concepts in the case of mBERT. The alignment with semantic concepts is relatively lower, with at most 16% match across models. This reflects that while the models learn both linguistic properties, morphological ontology is relatively preferred compared to the semantic hierarchy.

Syntactic These concepts capture grammatical orientation of a word, for example Chunking:B-NP is a syntactic concept describing words in the beginning of a noun phrase. CCG:PP/NP is a concept

[5]The first word, last word and prefix concepts showed less less than 1% alignment with the encoded concepts. We do not present their results in the interest of space.

[6]Appendix A shows more examples of the ngram, suffix, LIWC and WordNet clusters.



(a) Ngram:ex (b) LIWC:religion (c) LIWC: Bio (d) WordNet:Motion

Figure 3: Examples of BERT-c encoded concepts aligned with the human-defined concepts

in CCG super tagging, describing words that takes a noun phrase on the right and outputs a preposition phrase for example "[in[the US]]". We found relatively fewer matches, a maximum of 7% and 14% matching encoded concepts for Chunking and CCG concepts respectively. The low matches for syntactic concepts suggest that the models do not encode the same syntactic hierarchy suggested by these human-defined syntactic tasks.

Linguistic Ontologies Comparing the encoded concepts with static linguistic ontologies, we found WordNet concepts to be the second most aligned concept (11-21%) with the human-defined concepts. LIWC also shows a relatively higher alignment compared to the other human-defined concepts in a few models (e.g., BERT-c). However, this observation is not consistent across models and we found a range between 5-16% matches. These results present an interesting case where several models prefer the distinction of lexical ontology over abstract linguistic concepts such as morphology. Figure 3 shows examples of encoded concepts aligned with WordNet and LIWC. We see that these concepts are built based on a semantic relationship e.g., the clusters in Figure 3b, 3c and 3d group words based on religious, facial anatomy, and specific motion-related vocabulary respectively.

Comparing Models The results of multilingual models (mBERT, XLM-R) are intriguing given that their encoded concepts are dominated by ngram-based concepts and POS concepts, and their relatively lesser alignment with the linguistic ontologies. On the contrary, several monolingual models (BERT-c, ALBERT) showed a better match with linguistic ontologies specially WordNet.

The higher number of matches to the ngram (and suffix) concepts in the multilingual models is due to the difference in subword segmentation. The subword models in XLM-R and mBERT are optimized for multiple languages, resulting in a vocabulary consisting of a large number of small ngram units. This causes the multilingual models to aggressively segment the input sequence, compared to the monolingual models[7] and resulted in highly dominated ngram-based encoded concepts, especially in the lower layers. This may also explain the relatively lower match that multilingual models exhibit to the linguistic ontologies. We discuss this further in the context of layer-wise analysis in Section 5.2.

Comparing BERT cased vs. uncased, interestingly BERT-uc consistently showed higher matches for the core-linguistic concepts (See Figure 2). We speculate that in the absence of casing information, BERT-uc is forced to learn more linguistic concepts, whereas BERT-c leverages the explicit casing information to capture more semantically motivated concepts based on linguistic ontologies.

The higher matches in multilingual models in comparison to the monolingual models, and BERT-uncased in comparison to BERT-cased suggest that the training complexity is one factor that plays a role in a model's ability to learn linguistic nuances. For example, multilingual models need to optimize many languages, which is a harder task compared to learning one language. Similarly, the absence of capitalization in training data makes the learning task relatively harder for BERT-uc compared to BERT-c models, thus resulting in higher matches for BERT-uc. We speculate that the harder the training task, the more language nuances are learned by a model. Belinkov et al. (2020) made a similar observation, where they showed that the linguistic knowledge learned within the encoder-decoder representations in NMT models correlates with complexity of a language-pair involved in the task.

5.2 Layer-wise Alignment

Now we study the alignment of human-defined concepts across layers to understand how concepts

[7] In our dataset, mBERT has 13% more words after subword segmentation compared to BERT-c.

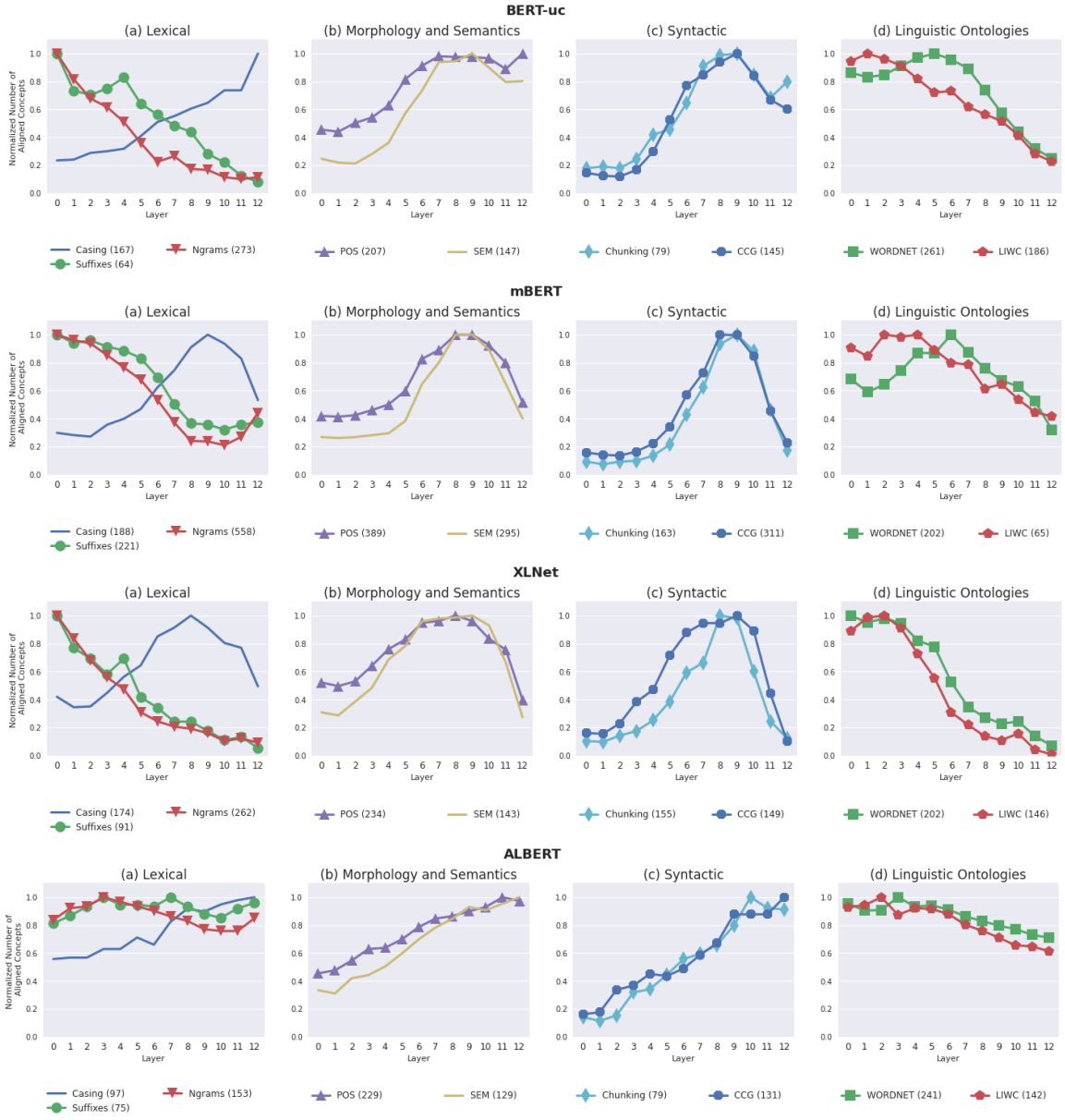


Figure 4: Layer-wise concept alignment. Y-axis is the normalized number of aligned concepts. The number within brackets of each human-defined concept, e.g. Casing (166), shows the maximum layer-wise match

evolve in the network. Figure 4 shows results for selected models.[8] The y-axis is the normalized number of aligned concepts across layers.

Overall Trend We observed mostly consistent patterns across models except for ALBERT, which we will discuss later in this section. We found that the shallow concepts (such as ngram and suffixes) and the linguistic ontologies (LIWC and WORDNET) are better represented in the initial layers and exhibit a downward trend in the higher layers of the network. On the contrary the core linguistic concepts (POS, Chunking, etc.) are better represented in the higher layers (layer 8-10). The last layers do not show any consistently dominating human-defined concepts considered in this work. We can generalize on these trends and hypothesize on how encoded concepts evolve in the network: the initial layers of the pretrained models, group words based on their lexical and semantic similarities where the former is an artifact of subword segmentation. With the inclusion of context and abstraction in the higher layers, these groups evolve into linguistic manifolds. The encoded concepts in the last layers are influenced by the objective function and learn concepts relevant to the task. Durrani et al. (2021) also made similar observation

[8] See Figure 10 in the Appendix for complete results.

when analyzing linguistic concepts in pre-trained models that are fine-tuned towards different GLUE tasks.

Concept-wise Trend In the following, we discuss different concepts in detail. As we mentioned earlier, the high presence of ngram and suffix concepts in the lower layers is due to subword segmentation. At the higher layers, the models start encoding abstract concepts, therefore get better alignment with the core linguistic concepts. Casing shows an exception to other lexical concepts and has similar trend to POS and SEM. Upon investigating we observed that the words appearing in these clusters have a hybrid connotation. For example, more than 98% of the encoded concepts that match with Casing are named entities, which explains the trend. The syntactic concepts observe peak in the higher-middle layers and a downward trend towards the end. These findings resonate with the earlier work on interpreting neural network representations for BERT. For example Liu et al. (2019a) also showed that probes trained with layers 7-8 give the highest accuracy when trained towards predicting the tasks of Chunking and CCG tagging. Although here, we are targeting a slightly different question i.e. how the latent concepts are encoded within the representations and how they evolve from input to output layers of the network.

We observed a downward trend in linguistic ontologies (WordNet, LIWC) as we go from lower layers to higher layers as opposed to the core linguistic concepts (POS, CCG, etc.). This is because of the context independent nature of these concepts as opposed to the core-linguistic concepts which are annotated based on the context. The embedding layer is non-contextualized, thus shows a high match with linguistic ontologies. With the availability of context in contextualized layers, the encoded concepts evolve into context-aware groups, resulting in higher matches with core-linguistic concepts.

Comparing Models While the overall trend is consistent among BERT-uc, mBERT and XLNet (and other studied models – Figure 10 in Appendix), the models somewhat differ in the last layers: see the large drop in core-linguistic concepts such as POS and Chunking for XLNet and mBERT in comparison to BERT. This suggests that BERT retains much of the core-linguistic information at the last layers. Durrani et al. (2020) observed a similar pattern in their study, where they showed BERT to retain linguistic information deeper in the model as opposed to XLNet where it was more localized and predominantly preserved earlier in the network.

While the overall layer-wise trends of multilingual models look similar to some monolingual models (mBERT vs. XLNet in Fig 4b,c), the former's absolute layer-wise matches (numbers inside the brackets in Figure 4 e.g. Casing (166)) are generally substantially higher than the monolingual counterparts. For example, the POS and SEM matches of mBERT are 38.9% and 30% respectively which are 18% and 15% higher than BERT-uc. On the contrary, the number of matches with linguistic ontologies is often lower for multilingual models (mBERT LIWC alignment of 65 vs. BERT-uc alignment of 186). We hypothesize that the variety of training languages in terms of their morphological and syntactic structure has caused the multilingual models to learn more core-linguistic concepts in order to optimize the training task. Although, the knowledge captured within linguistic ontologies is essential, it may not be as critical to the training of the model as the linguistic concepts.

ALBERT showed a very different trend from the other models. Note that ALBERT shares parameters across layers while the other models have separate parameters for every layer. This explains the ALBERT results where we see relatively less variation across layers. More interestingly, the encoded concepts in the last layers of ALBERT showed presence of all human-defined concepts considered here (see the relatively smaller drop of ALBERT alignment curves in Figure 4).

5.3 Unaligned Concepts

In Table 1 we observed that at least 27.6% (in XLM-R) and up to 56.4% (in XLNet) encoded concepts did not align with the human-defined concepts. *What concepts do these unaligned clusters contain?* In an effort to answer this question, we analyzed these clusters and observed that many of them were **compositional concepts** that involves more than one fine-grained categories of the human defined concepts. Figure 5a shows an example of the unaligned concept which partly aligns with a semantic category (SEM:geopolitical entity) and a morphological category (POS:adjective). Similarly, Figure 5b is a verbs related to cognitive processes and Figure 5c shows an unaligned cluster that is composed of different verb forms (past, present and gerunds). The alignment with multiple human-

(a) SEM:GPE+POS:JJ (b) POS:VB*+LIWC:cogmech (c) POS:VB*

Figure 5: Examples of unaligned encoded concepts: (a) combination of geopolitical entities and their related adjectives, (b,c) different forms of verb with specific semantics

defined concepts can be used to generate explanations for these unaligned concepts. For example, Figure 5a can be aligned as a mix of geopolitical entities and adjectives. We also quantitatively verified the number of unaligned encoded concepts that can be explained using composition of different concepts (See Appendix E: Table 9) and found that a majority of the clusters can be explained using a combination of three pre-defined concepts..

Moreover, note that encoded concepts are often multifacet i.e., they represent more than one relationship. For example, the encoded concept in Figure 5c consists of different forms of verbs but at the same time, these verbs are semantically similar. The semantic relationship present here is not adequately captured using the human-defined concepts used in this work. These are the *novel concepts* that require richer annotations or human-in-the-loop setup to generate adequate explanations.

5.4 Generalization of Results

Do the results generalize over different dataset selection and using different number of clusters? We ran experiments using different split of the news dataset for several models, and also performed alignment using different values of K, the number of clusters. The results are consistent across the board. Please see Appendix F for details.

6 Conclusion

We presented `ConceptX`, a novel framework for analyzing the encoded concepts within deep NLP models. Our method uses unsupervised clustering to discover latent concepts within the contextualized representations and then aligned these concepts with a suite of human-defined concepts to generate explanations for them. Our results illuminate how DNNs structure language information. A few notable findings are: i) lower layers capture shallow linguistic concepts, ii) whereas the abstract linguistic concepts such as morphology and semantics are preserved higher in the network, iii) the extent of alignment varies across different models and different human-defined concepts, iv) we found that novel explanations and an improved coverage of concepts can be achieved via compositionality.

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. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pages 242–247, Valencia, Spain.

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

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *CoRR*, abs/2102.12452.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 45(1):1–57.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia. Association for Computational Linguistics.

Zhi Chen, Yijie Bei, and Cynthia Rudin. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, D. Anthony Bau, and James Glass. 2019a. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI, Oral presentation)*.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*.

Fahim Dalvi, Avery Nortonsmith, D. Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019b. Neurox: A toolkit for analyzing individual neurons in neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. 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 1504–1516, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32:9277–9286.

K Chidananda Gowda and G Krishna. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2):105–112.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Julia Hockenmaier. 2006. Creating a CCGbank and a wide-coverage CCG lexicon for German. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, ACL '06, pages 505–512, Sydney, Australia.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure.

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv:1909.11942*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

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.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv:1907.11692*.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and Sueyeon Chung. 2020. Emergence of separable manifolds in deep language representations. In *International Conference on Machine Learning*, pages 6713–6723. PMLR.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *CoRR*, abs/2006.14032.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Investigating Language Universal and Specific Properties in Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488, Berlin, Germany. Association for Computational Linguistics.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Peter Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi, and Nadir Durrani. 2021. Fine-grained interpretation and causation analysis in deep NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 5–10, Online. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

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.

Wolfgang G Stock. 2010. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, 61(10):1951–1969.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2020. Finding experts in transformer models. *CoRR*, abs/2005.07647.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Robert L. Thorndike. 1953. Who belongs in the family. *Psychometrika*, pages 267–276.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. 2016. Word Representation Models for Morphologically Rich Languages in Neural Machine Translation. *arXiv preprint arXiv:1606.04217*.

John Wu, Hassan Belinkov, Yonatan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity Analysis of Contextual Word Representation Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Seattle. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Appendix

A Human-defined concept labels

A.1 Lexical Concepts:

Ngrams, Affixes, Casing, First and the Last Word.

A.2 Morphology and Semantics:

POS tags: We used the Penn Treebank POS tags discussed in (Marcus et al., 1993), which consists of 36 POS tags and 12 other tags (i.e., punctuation and currency symbols). In Table 2, we provide POS tags and their description.

SEM tags: (Abzianidze et al., 2017) consists of 73 sem-tags grouped into 13 meta-tags. In Table 3, we provide a detailed information of the tagset, and in Table 5, we provide fine and coarse tags mapping.

A.3 Syntactic:

Chunking tags: For Chunking we used the tagset discussed in (Tjong Kim Sang and Buchholz, 2000), which consists of 11 tags as follows: NP (Noun phrase), VP (Verb phrase), PP (Prepositional phrase), ADVP (Adverb phrase), SBAR (Subordinate phrase), ADJP (Adjective phrase), PRT (Particles), CONJP (Conjunction), INTJ (Interjection), LST (List marker), UCP (Unlike coordinate phrase). For the annotation, chunks are represented using IOB format, which results in 22 tags in the dataset as reported in Table 4.

CCG super-tags Hockenmaier (2006) developed, CCGbank, a dataset with Combinatory Categorial Grammar (CCG) derivations and dependency structures from the Penn Treebank. CCG is a lexicalized grammar formalism, which is expressive and efficiently parseable. It consists of 1272 tags.

A.4 Linguistic Ontologies:

WordNet: (Miller, 1995) consists of 26 lexicographic senses for nouns, 2 for adjectives, and 1 for adverbs. Each of them represent a supersense and a hierarchy can be formed from hypernym to hyponym.

LIWC: Over the past few decades, Pennebaker et al. (Pennebaker et al., 2001) have designed psycholinguistic concepts using high frequency words. These word categories are mostly used to study gender, age, personality, and health to estimate the correlation between these attributes and word usage. It is a knowledge-based system where words are mapped different high level concepts.

#	Tag	Description
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	to
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb
37	#	Pound sign
38	$	Dollar sign
39	.	Sentence-final punctuation
40	,	Comma
41	:	Colon, semi-colon
42	(	Left bracket character
43	)	Right bracket character
44	"	Straight double quote
45	’	Left open single quote
46	"	Left open double quote
47	’	Right close single quote
48	"	Right close double quote

Table 2: Penn Treebank POS tags.

B BERT-based Sequence Tagger

We trained a BERT-based sequence tagger to auto-annotate our training data. We used standard splits for training, development and test data for the 4 linguistic tasks (POS, SEM, Chunking and CCG super tagging) that we used to carry out our analysis on. The splits to preprocess the data are avail-

ANA (anaphoric)		**MOD (modality)**	
PRO	anaphoric & deictic pronouns: he, she, I, him	NOT	negation: not, no, neither, without
DEF	definite: the, loIT, derDE	NEC	necessity: must, should, have to
HAS	possessive pronoun: my, her	POS	possibility: might, could, perhaps, alleged, can
REF	reflexive & reciprocal pron.: herself, each other	**DSC (discourse)**	
EMP	emphasizing pronouns: himself	SUB	subordinate relations: that, while, because
ACT (speech act)		COO	coordinate relations: so, {,}, {;}, and
GRE	greeting & parting: hi, bye	APP	appositional relations: {,}, which, {(}, —
ITJ	interjections, exclamations: alas, ah	BUT	contrast: but, yet
HES	hesitation: err	**NAM (named entity)**	
QUE	interrogative: who, which, ?	PER	person: Axl Rose, Sherlock Holmes
ATT (attribute)		GPE	geo-political entity: Paris, Japan
QUC	concrete quantity: two, six million, twice	GPO	geo-political origin: Parisian, French
QUV	vague quantity: millions, many, enough	GEO	geographical location: Alps, Nile
COL	colour: red, crimson, light blue, chestnut brown	ORG	organization: IKEA, EU
IST	intersective: open, vegetarian, quickly	ART	artifact: iOS 7
SST	subsective: skillful surgeon, tall kid	HAP	happening: Eurovision 2017
PRI	privative: former, fake	UOM	unit of measurement: meter, $, %, degree Celsius
DEG	degree: 2 meters tall, 20 years old	CTC	contact information: 112, info@mail.com
INT	intensifier: very, much, too, rather	URL	URL: http://pmb.let.rug.nl
REL	relation: in, on, 's, of, after	LIT	literal use of names: his name is John
SCO	score: 3-0, grade A	NTH	other names: table 1a, equation (1)
COM (comparative)		**EVE (events)**	
EQU	equative: as tall as John, whales are mammals	EXS	untensed simple: to walk, is eaten, destruction
MOR	comparative positive: better, more	ENS	present simple: we walk, he walks
LES	comparative negative: less, worse	EPS	past simple: ate, went
TOP	superlative positive: most, mostly	EXG	untensed progressive: is running
BOT	superlative negative: worst, least	EXT	untensed perfect: has eaten
ORD	ordinal: 1st, 3rd, third	**TNS (tense & aspect)**	
UNE (unnamed entity)		NOW	present tense: is skiing, do ski, has skied, now
CON	concept: dog, person	PST	past tense: was baked, had gone, did go
ROL	role: student, brother, prof., victim	FUT	future tense: will, shall
GRP	group: John {,} Mary and Sam gathered, a group of people	PRG	progressive: has been being treated, aan hetNL
DXS (deixis)		PFT	perfect: has been going/done
DXP	place deixis: here, this, above	**TIM (temporal entity)**	
DXT	temporal deixis: just, later, tomorrow	DAT	full date: 27.04.2017, 27/04/17
DXD	discourse deixis: latter, former, above	DOM	day of month: 27th December
LOG (logical)		YOC	year of century: 2017
ALT	alternative & repetitions: another, different, again	DOW	day of week: Thursday
XCL	exclusive: only, just	MOY	month of year: April
NIL	empty semantics: {.}, to, of	DEC	decade: 80s, 1990s
DIS	disjunction & exist. quantif.: a, some, any, or	CLO	clocktime: 8:45 pm, 10 o'clock, noon
IMP	implication: if, when, unless		
AND	conjunction & univ. quantif.: every, and, who, any		

Table 3: Semantic tags.

Task	Train	Dev	Test	Tags	F1
POS	36557	1802	1963	48	96.69
SEM	36928	5301	10600	73	96.22
Chunking	8881	1843	2011	22	96.91
CCG	39101	1908	2404	1272	94.90

Table 4: Data statistics (number of sentences) on training, development and test sets using in the experiments and the number of tags to be predicted

able through git repository[9] released with Liu et al. (2019a). See Table 4 for statistics and classifier accuracy.

[9] https://github.com/nelson-liu/contextual-repr-analysis

C Clustering details

Algorithm 1 assigns each word to a separate cluster and then iteratively combines them based on Ward's minimum variance criterion that minimizes intra-cluster variance. Distance between two vector representations is calculated with the squared Euclidean distance.

Algorithm 1 Clustering Procedure

Input: $\overrightarrow{y}^l$: word representation of words
Parameter: K: the total number of clusters

1: **for** each word w_i **do**
2: assign w_i to cluster c_i
3: **end for**
4: **while** number of clusters $\neq K$ **do**
5: **for** each cluster pair c_i,$c_{i'}$ **do**
6: $d_{i,i'}$ = inner-cluster difference of combined cluster c_i and $c_{i'}$
7: **end for**
8: c_j,$c_{j'}$ = cluster pair with minimum value of d
9: merge clusters c_j and $c_{j'}$
10: **end while**

C.1 Selection of the number of Clusters

The Elbow curve did not show any optimum clustering point, with the increase in number of clusters the distortion score kept decreasing, resulting in over-clustering (a large number of clusters consisted of less than 5 words). The over-clustering resulted in high but wrong alignment scores e.g. consider a two word cluster having words "good" and "great". The cluster will have a successful match with "adjective" since more than 90% of the words in the cluster are adjectives. In this way, a lot of small clusters will have a successful match with many human-defined concepts and the resulting alignment scores will be high. On the other hand, Silhouette resulted in under-clustering, giving the best score at number of clusters = 10. We handled this empirically by trying several values for the number of clusters i.e., 200 to 1600 with step size 200. We selected 1000 to find a good balance with over and under clustering. We understand that this may not be the best optimal point. We presented the results of 600 and 1000 clusters to show that our findings are not sensitive to the number of clusters parameter.

D Coarse vs. Fine-grained Categories

D.1 Coarse vs. Fine-grained Categories

Our analysis of compositional concepts showed that several fine-grained concepts could be combined to explain an unaligned concept. For example, by combining verb categories of POS to one coarse verb category, we can align the encoded concept present in Figure 5c. To probe this more formally, we collapsed POS and SEM fine-grained concepts into coarser categories (27 POS tags and 15 SEM tags). We then recomputed the alignment with the encoded concepts. For most of the models, the alignment doubled compared to the fine-grained categorizes with at least 39% and at most 53% percent match for POS. This reflects that in several cases, models learn the coarse language hierarchy. We further questioned *how many encoded concepts can be explained using coarse human-defined concepts.* Compared to Table 1, the matches increased by at most 17 points in the case of BERT-uc. The XLM-R showed the highest matching percentage of 81%. The higher alignment suggests that most of the encoded concepts learned by pre-trained models can be explained using human-defined concepts. (See Appendix D for detailed results).

D.2 Corase POS and SEM labels

Tables 5 and 6 present results for our mapping of fine-grained SEM and POS tags into coarser categories.

Coarse	Fine-grained
ACT	QUE
ANA	DEF, DST, EMP, HAS, PRO, REF
ATT	INT, IST, QUA, REL, SCO
COM	COM, LES, MOR, TOP
DSC	APP, BUT, COO, SUB
DXS	PRX
EVE	EXG, EXS, EXT, EXV
LOG	ALT, AND, DIS, EXC, EXN, IMP, NIL, RLI
MOD	NEC, NOT, POS
NAM	ART, GPE, HAP, LOC, NAT, ORG, PER, UOM
TIM	DEC, DOM, DOW, MOY, TIM, YOC
TNS	EFS, ENG, ENS, ENT, EPG, EPS, EPT, ETG, ETV, FUT, NOW, PST
UNE	CON, ROL
UNK	UNK

Table 5: SEM: Coarse to Fine-grained mapping

D.3 Results

Table 7 presents the alignment results of using coarse POS and SEM concepts. We observed that

Coarse	Fine-grained
Adjective	JJ, JJR, JJS
Adverb	RB, RBS, WRB, RBR
Conjunction	CC
Determiner	DT, WDT
Noun	NN, NNS, NNP, NNPS
Number	CD
Preposition	IN, TO
Pronoun	PRP, PRP$, WP, WP$
Verb	VB, VBN, VBZ, VBG, VBP, VBD
No Changes	$, -LRB-, #, FW, -RRB-, LS, POS, "", EX SYM, „, :, RP, ., PDT, MD, UH,

Table 6: POS: Coarse to Fine-grained mapping

the alignment doubles in most of the cases which reflects that in several cases, models learn the coarse language hierarchy. However, they do not strictly adhere to fine-grained categories existed in human-defined concepts. We further extend the alignment of coarse POS and SEM categories to the overall alignment with the human-defined concepts. Table 8 presents the results. We see a match of up to 81% in the case of XLM-R. The high alignment suggests that many of the encoded concepts can be explained using coarse human-defined concepts.

	POS		SEM	
	Fine	Coarse	Fine	Coarse
BERT-cased	13%	42%	7%	15%
BERT-uncased	16%	43%	9%	18%
mBERT	26%	53%	16%	26%
XLM-RoBERTa	24%	47%	11%	21%
RoBERTa	18%	43%	10%	20%
ALBERT	17%	42%	9%	17%
XLNet	17%	39%	10%	18%

Table 7: Alignment of fine-grained human defined concepts compared to coarse categories

E Compositional Coverage

Table 9 shows the amount of coverage we obtain when aligning with the morphological concepts when allowing 90% of the words in the cluster to be from N concepts.

	BERT-c	**BERT-uc**	**mBERT**	**XLM-R**
Overall alignment	61.5%	63.6%	77.7%	81.0%
	RoBERTa	**ALBERT**	**XLNet**	
	62.9%	64.0%	55.3%	

Table 8: Coverage of human-defined concepts using coarse POS and SEM labels across all clusters from a given model

F Robustness of Methodology across Datasets and Settings

Figure 8 shows the layer-wise patterns using 600 clusters instead of 1000 as used in the main paper. We observe that the overall trends largely remain the same.

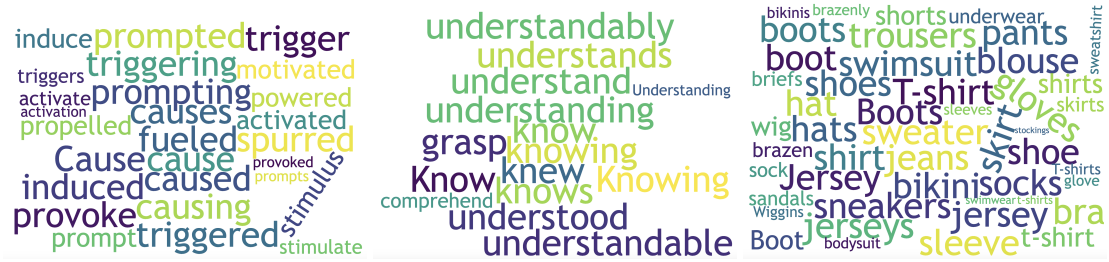
To further demonstrate the robustness of our method with respect to dataset, we sub-sampled another dataset from the News corpus with a different vocabulary by selecting words that appear between 2 to 10 times in the corpus. Note that the selection of vocabulary is due to the memory and computation limitations. Figure 9 shows the results using this selection of data. Compared to Figure 4, we can see that the overall patterns are largely similar and confirms the robustness of our findings. The slight difference in the patterns of WordNet and LIWC are due to the large selection of proper nouns in the second set of the data.

G Layer-wise results

Figure 10 present layer-wise results for all the understudied models.

(a) (b) (c)

Figure 6: Example clusters: (a) ngram:ace, (b) POS:CD, (c) Chunking:B-VP + Suffix:ed



(a) (b) (c)

Figure 7: Example clusters: (a) LIWC:cause, (b) WORDNET:verb.cognition, (c) WORDNET:noun.artifact

Concepts	BERT-c	BERT-uc	mBERT	XLM-R	RoBERTa	ALBERT	XLNet
1	13%	16%	26%	24%	18%	17%	17%
2	11%	12%	20%	23%	13%	13%	12%
3	14%	13%	14%	18%	11%	15%	9%
4	6%	6%	4%	4%	5%	5%	3%
5	2%	1%	1%	1%	2%	1%	1%
6	1%	0%	0%	0%	1%	1%	0%

Table 9: Percentage of alignment when an encoded concept is composed of N morphological concepts. As can be seen, most concepts are composed of either 1, 2 or 3 morphological concepts, showing that several concepts learned by these models are indeed compositional in nature.

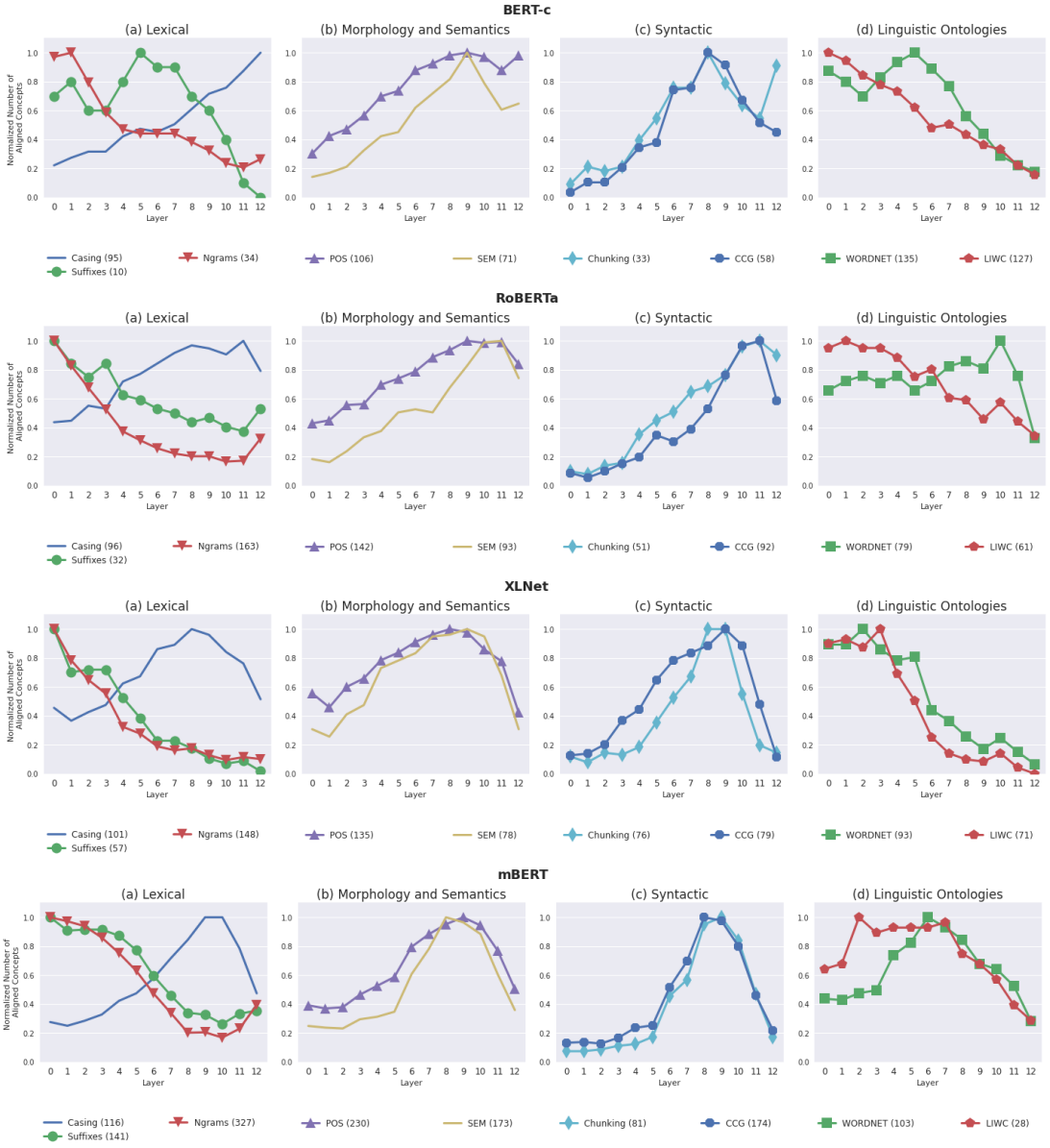


Figure 8: Layer-wise results using 600 clusters.

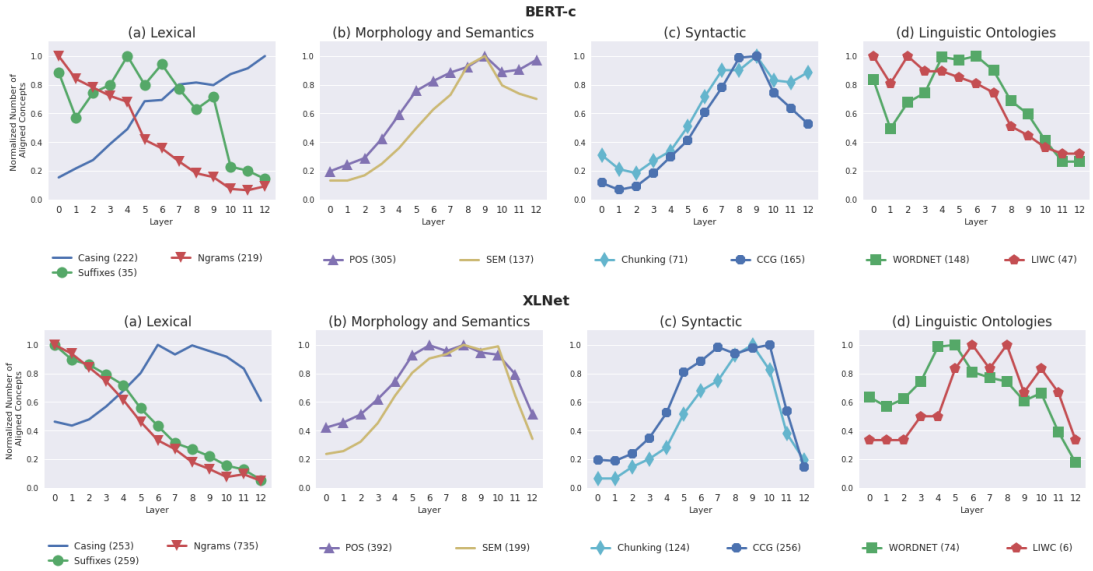


Figure 9: Layer-wise results on a separately sampled dataset.

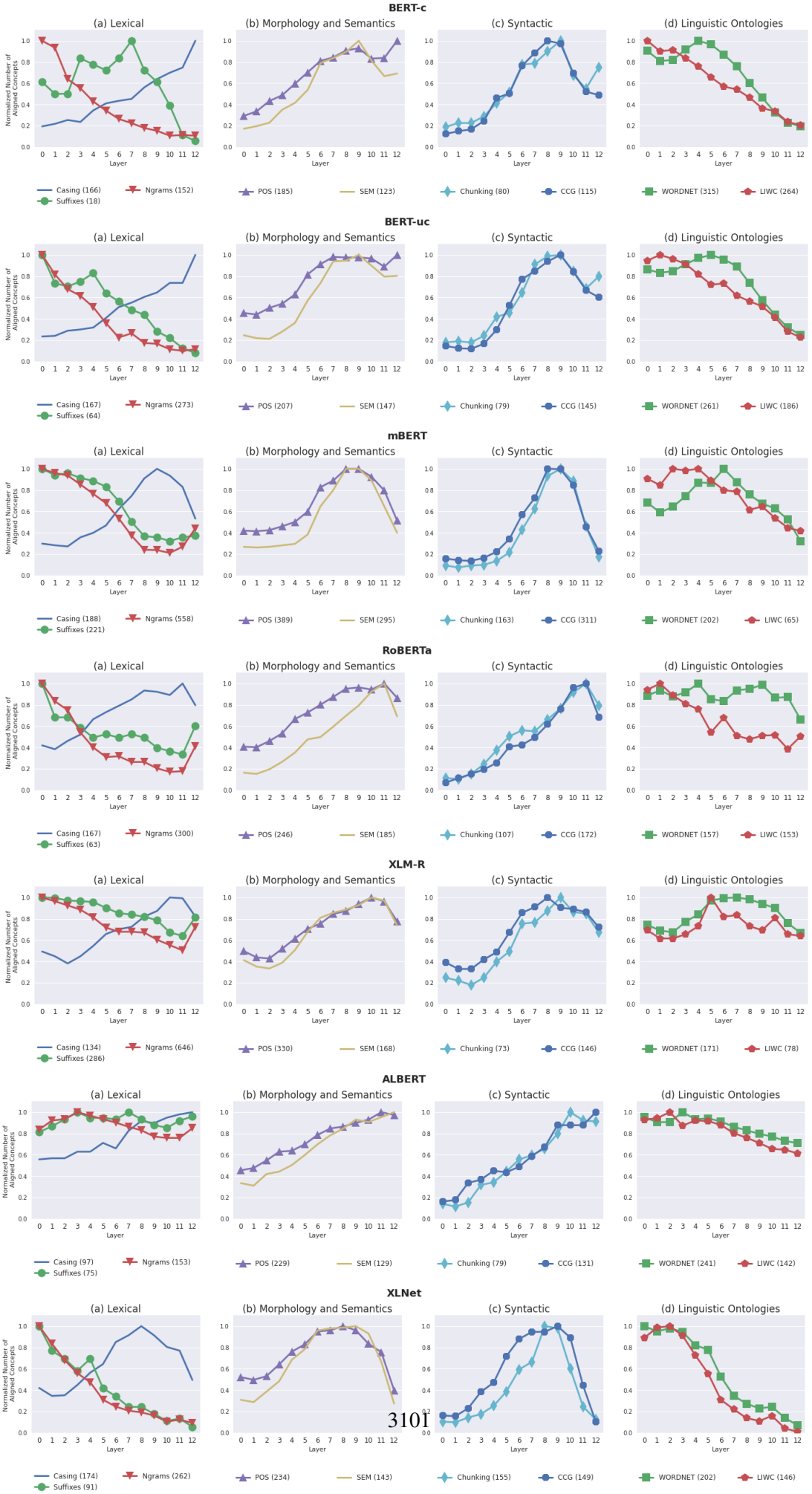


Figure 10: Layer wise results