Exploring Meaning Encoded in Random Character Sequences with Character-Aware Language Models

Anonymous ACL submission

Abstract

Natural language processing models learn word representations based on the distributional hypothesis, which asserts that word context (e.g., co-occurrence) correlates with semantic meaning. We propose that n-grams composed of random character sequences, or *garble*, provide a novel context for studying word meaning both within and beyond extant language. In particular, randomly-generated character n-grams lack semantic meaning but contain primitive information based on the distribution of characters they contain. By studying the embeddings of a large corpus of garble, extant language, and pseudowords using CharacterBERT, we identify an axis in the model's high-dimensional embedding space that separates these classes of n-grams. Furthermore, we show that this axis relates to structure within extant language, including word part of speech, morphology, and concreteness. Thus, in contrast to studies that are mainly limited to extant language, our work reveals that semantic meaning and primitive information are intrinsically linked.

1 Introduction

What primitive information do character sequences contain? Modern natural language processing is driven by the *distributional hypothesis* (Firth, 1957), which asserts that the context of a linguistic expression defines its meaning (Emerson, 2020). Because existing words—which represent an extremely small fraction of the space of possible character sequences—appear in context together, the distributional paradigm at this level is limited in its ability to study the meaning of and information encoded by arbitrary character level n-grams (wordforms). Furthermore, state-of-the-art computational language models operating within the distributional paradigm, such as BERT (Devlin et al., 2019), are mainly trained on extant words. We propose that character n-grams (i.e., sequences of alphabetic characters) outside the space of extant language provide new insights into the meaning of words, beyond that captured by word and subword-based distributional semantics alone. We explore this by studying the embeddings of randomly-generated character n-grams (referred to as *garble*), which contain primitive communicative information but are devoid of semantic meaning, using the CharacterBERT model (El Boukkouri et al., 2020). Such randomly-generated character n-grams are textual analogues of paralinguistic vocalizations.

Our analyses contribute to the growing understanding of BERTology (Rogers et al., 2020) by identifying a dimension, which we refer to as the *information axis*, that separates extant and garble n-grams. This finding is supported by a Markov model that produces a probabilistic information measure for character n-grams based on their statistical properties. Strikingly, this information dimension correlates with properties of extant language; for example, parts of speech separate along the information axis, and word concreteness varies along a roughly orthogonal dimension in our projection of CharacterBERT embedding space. Although the information axis we identify separates extant and randomly-generated n-grams very effectively, we demonstrate that these classes of n-grams mix into each other in detail, and that *pseudowords*—i.e., phonologically coherent character n-grams without lexical meaning—lie between the two in our CharacterBERT embeddings.

This paper is organized as follows. We first discuss concepts from computational linguistics, information theory, and linguistics relevant to our study. We then analyse CharacterBERT representations of extant and randomly-generated character sequences and how the relation between the two informs the structure of extant language, including morphology, part-of-speech, and word concreteness. Finally, we ground our information axis in a predictive Markov language model.

2 Modeling n-grams Beyond Extant Language

Models in computational linguistics often represent words in a high-dimensional embedding space

based on their co-occurrence patterns according to the distributional hypothesis (Landauer and Dumais, 1997; Mikolov et al., 2013). Embeddings that capture the semantic content of extant words are used for many natural language applications, including document or sentence classification (Kowsari et al., 2019), information retrieval and search (Mitra et al., 2018), language modelling and translation (Devlin et al., 2019), language generation (Brown et al., 2020), and more (Jurafsky and Martin, 2021). In these cases, vector operations performed on word embeddings are used for higher-level tasks such as search or classification.

Word embeddings have largely concerned themselves with extant language—that is, commonly used words which carry consistent meaning—and thus cannot represent character n-grams outside of this space. The few models that encompass *character* n-grams, which naturally include n-grams beyond extant words, often use RNNs (Mikolov et al., 2010) or encoder-decoder architectures (Sutskever et al., 2014) to represent character-level sequences. In parallel, the ubiquitous use of Transformer models has led to studies of their inner representations, weights, and *attention* mechanism (Rogers et al., 2020; Clark et al., 2019). While most Transformer models are trained using extant words and sub-words, largely focusing on their semantics and syntax; however, some recent models operate at the character level, such as CharacterBERT (El Boukkouri et al., 2020) and CharBERT (Ma et al., 2020). Strikingly, character-level models excel at character-level tasks (e.g., spelling correction; Xie et al. 2016; Chollampatt and Ng 2018) and perform comparably to word-level models at language-modelling tasks (Kim et al., 2016).

Character-level models are therefore an ideal tool for studying the information and meaning encoded in n-grams beyond the realm of extant language. Throughout this study, we use the CharacterBERT model to achieve this goal. CharacterBERT is uniquely suited for our study as it uses a CharacterCNN module (Peters et al., 2018) to produce single embeddings for any input token, built as a variant to BERT which relies on sub-word tokenization (El Boukkouri et al., 2020).

3 Primitive Information and Meaning Beyond Extant Language

Before presenting our results, we discuss general characteristics of the space beyond extant words; we reiterate that this space is missed by word and sub-word-based models. Due to CharacterBERT's use of English characters, we restrict our analysis to English character n-grams, and we study the properties of CharacterBERT embeddings including n-grams outside of extant language that contains lexicalized semantic meaning. By studying meaning encoded in n-grams that do not appear in consistent (or any) context in the model's training data, our framework goes beyond the traditional distributional hypothesis paradigm. In this way, we seek to understand core properties of information encoded in n-grams beyond their lexicalized semantics by simultaneously studying n-grams that contain different types of information.[1]

We use randomly-generated characters to create n-grams that contain primitive information but no semantic meaning. We adapt Marr's notion of primitive visual information for primitive textual information (Marr and Hildreth, 1980), and make the analogue between vision and language because information is substrate independent (Deutsch and Marletto, 2015). In our case, primitive textual information is lower-level communicative information which subsumes semantic meaning. Being textual, our randomly-generated n-grams are not bound by the constraints of human speech, and may be phonologically impossible.

In the following subsections, we provide three examples of language—distorted speech, paralanguage, pseudowords—which motivate our study of character-level embeddings for randomly-generated character n-grams. We then describe the complementary information encoded by word morphology.

3.1 Distorted Speech

In popular use "garble" refers to a language message that has been distorted (garbled), such as speech where semantic meaning is corrupted by phonological distortions. For example, the phrase "reading lamp" may become "eeling am" when garbled. Garbled speech contains lesser, or zero, semantic meaning compared to ungarbled speech, but the signal of speech media is nonetheless present as information, which according to Shannon (1951) may contain no meaning at all. Garbled speech satisfies the classical five-part definition of com-

[1] In analogy, the theory of ensemble perception in developmental psychology offers a framework to understand the human ability to understand the 'gist' of multiple objects at once (Sweeny et al., 2015).

munication provided by (Shannon, 2001); an *information source* (speaker) can *transmit* (verbalize) an informationally primitive message through the *channel* of speech media through the *receiver* (ears) to the *destination* (listener).

3.2 Paralanguage

Paralinguistic vocalizations are specifically identifiable sounds beyond the general characteristics of speech (Noth, 1990) and present another example of communication beyond lexicalized semantics. Paralinguistic vocalizations include *characterizers*, like moaning; and *segregates*, like "uh-huh" for affirmation. The border between such paralinguistic vocalizations and lexicalized interjections with defined semantic meanings is "fuzzy" (Noth, 1990).

3.3 Pseudowords

Pseudowords are phonologically possible character n-grams without lexical meaning. Wordlikeness judgments reveal that human distinction between pseudowords and phonologically impossible nonwords is gradational (Needle et al., 2020). As a unique informational class, pseudowords have been used in language neuronal activation studies (Price et al., 1996), infant lexical-semantic processing (Friedrich and Friederici, 2005), in poetry through nonsense (Ede, 1975), and in literary analyses (Lecercle, 2012). Pseudowords can also elicit consistent cognitive responses (Davis et al., 2019).

To consider pseudowords generatively, it is helpful to note that an alphabetic writing system covers not only ever word but every possible word in its language (Deutsch, 2011); pseudowords can thus be thought of as former possible-but-uninstantiated extant words—e.g., "cyberspace" was a pseudoword before the internet. We embed randomly generated pseudowords into our model to study their information content and relation to both extant words and randomly-generated n-grams.

3.4 Morphology

Morphology deals with the systems of natural language that create words and word forms from smaller units (Trost, 1992). Embedding spaces and the distributional hypothesis offer insights into the relationship between character combination, morphology and semantics. Notably, morphological irregularities complicate the statistics of global character-level findings in the embedding space, like through *suppletion*—where word forms change idiosynchratically e.g. *go*'s past tense is *went*, or *epenthesis*—where character are inserted under certain phonological conditions e.g. fox pluralizes as fox*e*s (Trost, 1992). As do the multiple 'correct' spellings of pseudowords under conventional phoneme-to-grapheme mapping (Needle et al., 2020). Distinctions between morphological phenomena can also be hard to define; the boundary between derivation and compounding is "fuzzy" (Trost, 1992).

4 Character-Level Language Models for Information Analysis

As described above, state-of-the-art language models serve as a tool to study meaning as it emerges though the distributional hypothesis paradigm. Existing work on the analysis of Transformers and BERT-based models have explored themes we are interested in, such as semantics (Ethayarajh, 2019), syntax (Goldberg, 2019), morphology (Hofmann et al., 2020, 2021), and the structure of language (Jawahar et al., 2019). However, all of this work has limited itself to the focus of extant words, largely due to the word and sub-word-based nature of these models.

We study the structure of the largely unexplored character n-gram space which includes extant language, pseudoword and garble character n-grams, seen through the representations created by CharacterBERT, as follows. To explore how the character n-gram space is structured in the context of character based distributional semantics, we embed 40,000 extant English words, 40,000 randomly-generated character n-grams, and 20,000 pseudowords. We choose the 40,000 most used English words that have been annotated for concreteness/abstractness ratings (Brysbaert et al., 2014). Randomly-generated character n-grams are forced to have a string length distributions that matches the corpus of extant words we analyze. To generate pseudowords, we use a pseudoword generator.[2]

The CharacterBERT (El Boukkouri et al., 2020) general model has been trained on nearly 40 GB of Reddit data using character sequences. We leverage this model to create representations of character n-grams that may not have been seen in the training data. This allows us to use the resulting 512 dimensional embeddings for exploration via visualisation, topology modelling via distances and projections, and classification error analysis.

[2] http://soybomb.com/tricks/words/

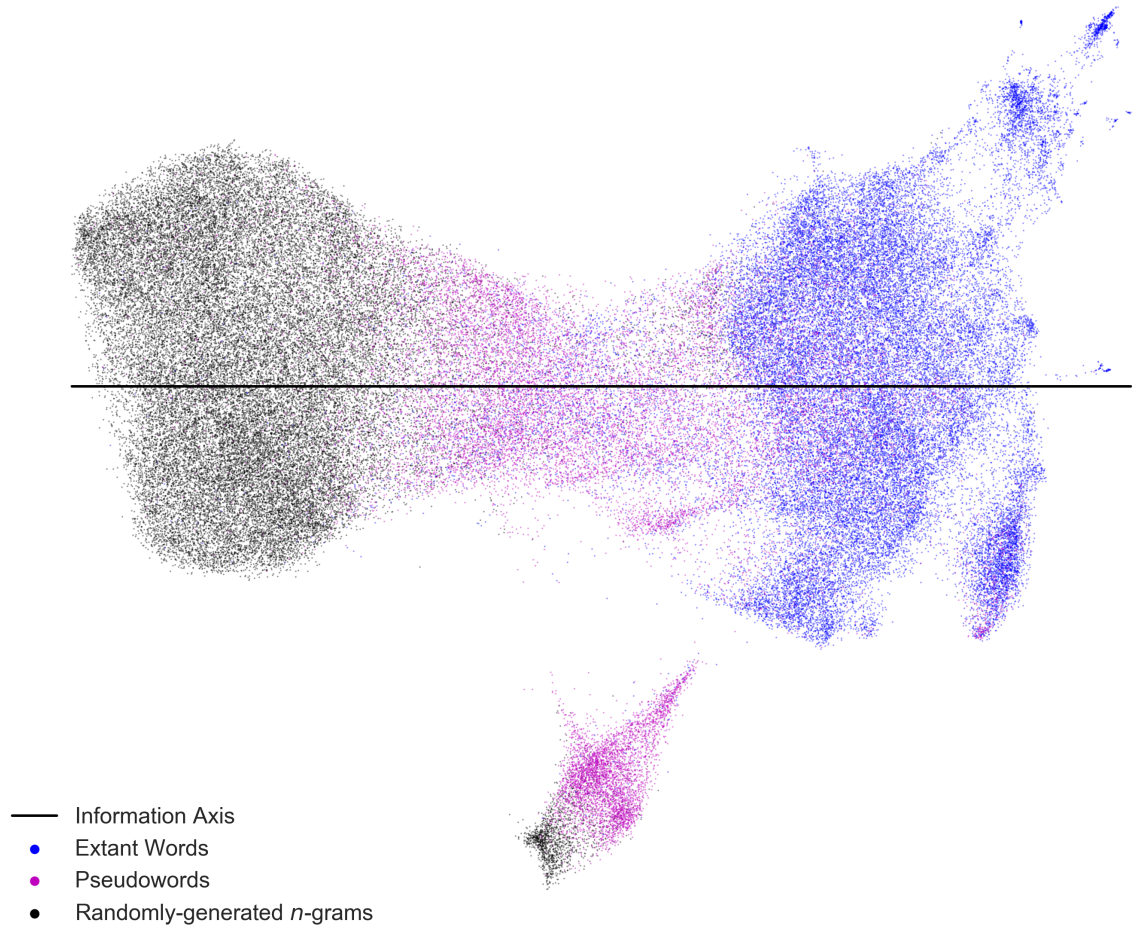

Figure 1: UMAP projection of CharacterBERT embeddings for extant words (blue), pseudowords (magenta), and randomly-generated character n-grams (black). The solid black line shows the information axis that we define in this work. The bottom-most cluster of random and pseudoword character n-grams is comprised of character n-grams ending in "s", and the top-most clusters of extant words are comprised of compound words.

4.1 Identifying the Information Axis

To guide our exploration of the high-dimensional topology of the resulting embeddings, we use the UMAP dimensionality reduction technique (McInnes et al., 2018). UMAP creates a low-dimensional embedding by searching for a low-dimensional projection of the data that has the closest possible equivalent fuzzy topological structure as the original representations, thereby preserving both local and global structure.

We use the UMAP embeddings to extract an *information axis* that captures most variance among extant and randomly-generated n-grams. To assign n-grams an "information axis score," we minmax-normalize the UMAP coordinates along this axis. Thus, our information axis establishes a link between extant language and garble, thereby connecting semantic meaning and primitive information. Figure 1 shows how CharacterBERT embeddings of extant, pseudoword, and randomly-generated character n-grams arrange themselves in this space.

4.2 Statistical Properties of n-grams Along the Information Axis

We perform several statistical tests to differentiate between categories of character n-grams along the information axis. First, Table 1 lists the median and standard deviation of minmax-normalized position along the information axis, demonstrating that extant words, pseudowords, and garble are clearly separated.

Next, we use the Kolmogorov-Smirnov (KS; Massey Jr 1951) two-sample test to assess differences between the information axis distributions of our n-gram classes. All of the KS tests very significantly indicate differences between types of character n-gram and parts of speech along the information axis ($p \ll 0.001$). Furthermore, the KS statistic score is 0.94 for (extant, random), 0.83 for (extant, pseudoword), and 0.70 for (pseudoword, ran-

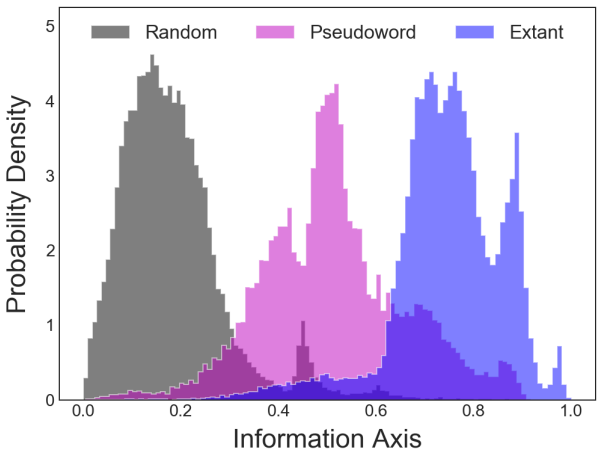

Figure 2: Probability density of CharacterBERT embeddings for extant words (blue), pseudowords (magenta), and randomly-generated character n-grams (black) as a function of minmax-normalized position along the information axis shown in Figure 1.

Character n-gram type	information axis Position
Extant	0.75 ± 0.12
Noun	0.74 ± 0.12
Verb	0.72 ± 0.09
Adjective	0.76 ± 0.11
Adverb	0.87 ± 0.09
Pseudoword	0.50 ± 0.15
Random	0.17 ± 0.11

Table 1: Median and standard deviation of minmax-normalized position along the information axis shown in Figure 1, for extant words (including parts of speech), pseudowords, and randomly-generated n-grams.

dom), indicating that extant and random n-grams differ most significantly along the information axis (consistent with Figures 1–2).

4.3 Hyperplane Classifier

The visualisation of the character n-grams suggests that a hyperplane classifier is suitable for separating extant words and garble. We use a support vector machine (Cortes and Vapnik, 1995) trained on half of our 40,000 commonly-used extant words and half of our computer-generated garble to classify unseen extant, garble and pseudoword character n-grams. We use this method to explore the information in the high-dimensional embeddings, and to observe which words cross over a so-called 'river' of meaning, located at 0.5 of the information axis, its midpoint.

The classifier achieves an accuracy of 98.9% on unseen extant language and garble character n-grams, suggesting we can learn about the embeddings through error analysis. In particular, we found similarities among extant words classified as garble. 74% (270/363) were compound or derivative words, similar to many extant language terms that lie near the midpoint of the information axis. 19% (69/363) were foreign words like "hibachi", or dialect words like "doohickey." The garble classification errors—garble classified as extant language—were in small part due to our randomization method inadvertently creating extant language mislabelled as garble, accounting for $\sim 10\%$ of the 377 errors we identify.

The garble misclassified as extant language mostly contained phonologically impossible elements, though some were pseudowords.

When pseudowords were forcibly classified into extant or garble character n-grams, more pseudowords were classified as extant language than garble (7106 as garble to 12894 as extant). Labelling affirms these intuitions, with pseudowords like "flought" looking intuitively familiar. Given CharacterBERT's massive Reddit training data, typos and localized language may account for the classifier's tendency to classify pseudowords as extant language. Also, our embedding space only uses the 40,000 most common English words out of 208,000 distinct lexicalized lemma words (Brysbaert et al., 2016), which if included may impact spatial structure.

5 Structure of Extant Words along the Information Axis

We use this section to discuss the structure of language across the information axis derived from our low-dimensional UMAP space. We structure our analysis across this axis as it organises the relative structure of extant words vs. randomly-generated character n-grams, while also distinguishing internal structure within the extant word space.

5.1 Extant vs. Pseudowords vs. Garble

At the scale of global structure, the information axis highlights that extant words are separated from randomly-generated character n-grams (Figure 1). We note the midpoint of the two character n-gram classes at 0.5 on our information axis. Pseudowords populate the region near the midpoint of the information axis, and also overlap with both extant English and garble character n-grams (Figure 2). There is no distinct boundary between

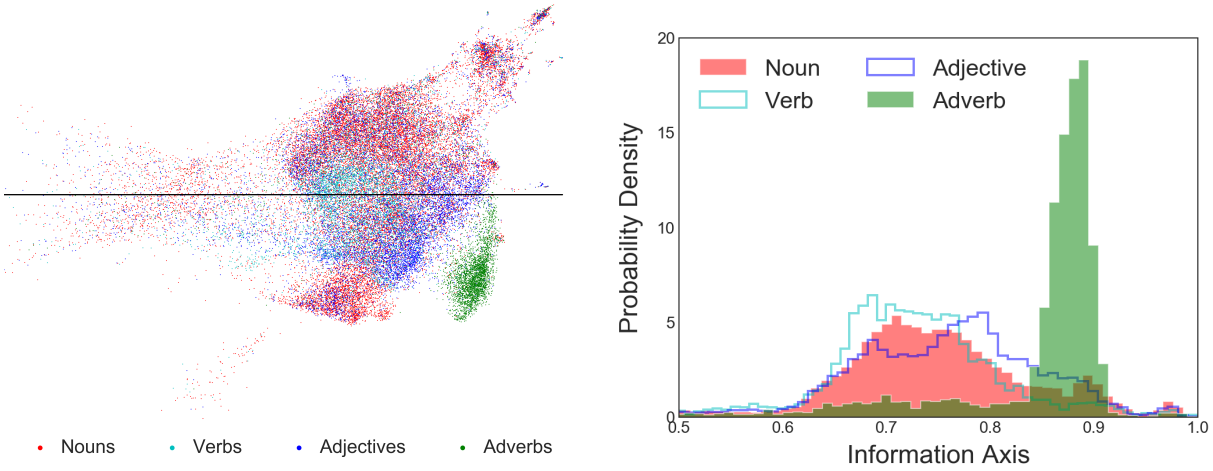

Figure 3: *Left panel*: UMAP projection of CharacterBERT embeddings for extant words split by part of speech into nouns (red), verbs (cyan), adjectives (blue), and adverbs (green). *Right panel*: Probability density of extant words, split by part of speech, as a function of minmax-normalized position along the information axis shown in Figure 1.

the three classes of n-grams, which is consistent with both morphological descriptions of compound and derivational words and descriptions of paralanguage as "fuzzy". This global structure—and the structure internal to extant language (Figure 3)—goes beyond the distributional hypothesis by including n-grams that do not appear in consistent (or any) contexts. Pseudowords lie between extant and garble character n-grams, but there is no distinct boundary between pseudowords and the other classes of n-grams.

Extant language and garble regions have different internal structure (Figure 1). The garble region has comparatively less structure than the extant language region, though there is some internal structure, notably a cluster of character n-grams ending in character "s" separated from the main garble region.

5.2 Parts of Speech and Morphology

In our UMAP projection, detailed structure emerges for extant words split by part-of-speech (Figure 3). In particular KS statistics between all part-of-speech pairs significantly indicate that their distributions differ along the information axis. Furthermore, KS statistic values are 0.12 for (noun, verb), 0.11 for (noun, adjective), 0.64 for (noun, adverb), 0.22 for (verb, adjective), 0.72 for (verb, adverb), and 0.64 for (adjective, adverb). This suggests that adverbs are most cleanly separated from other parts of speech along the information axis (consistent with Figure 3), which may indicate that morphemes/prefixes/suffixes have important effects in embedding space. A detailed investigation is beyond the scope of this paper and may require analyses through alternative heuristics such as pseudomorphology and lexical neighborhood density (Needle et al., 2020).

Many extant words near the midpoint of the information axis are, or may be, compound words; the boundary between derivative and compound words is thought to be fuzzy because many derivational suffixes developed from words frequently used in compounding (Trost, 1992). Both derivative and compound words populate other spaces of the extant language region, but conflicting definitions hamper straightforward statistical analysis.

Morphological traits such as adjectival suffixes $-ness$, $-ism$, $-ility$, and $-able$, or the adverbial suffix $-ly$ correlate to mapping, but the boundaries for morphological classes are not distinct. Garble ending in "s" occupies a closer region to extant language, arguably due to the semantic associations of ending in "s" derived from the suffix $-s$. Note, the morphological heuristics of affixation applies to lexicalized words but not garble. Pseudowords ending in "s" share that region of garble ending in "s", however, such seemingly plural pseudowords tend closer to extant language, reflecting the notion that wordform similarity increases with semantic similarity (Dautriche et al., 2017). Given the fuzziness of morphology and the opaqueness of English spelling (Needle et al., 2020), pseudowords ending in "s" may or may not be due to affixation.

5.3 Concreteness/Abstractness

The internal positioning of different parts of speech within the extant language space of our low-

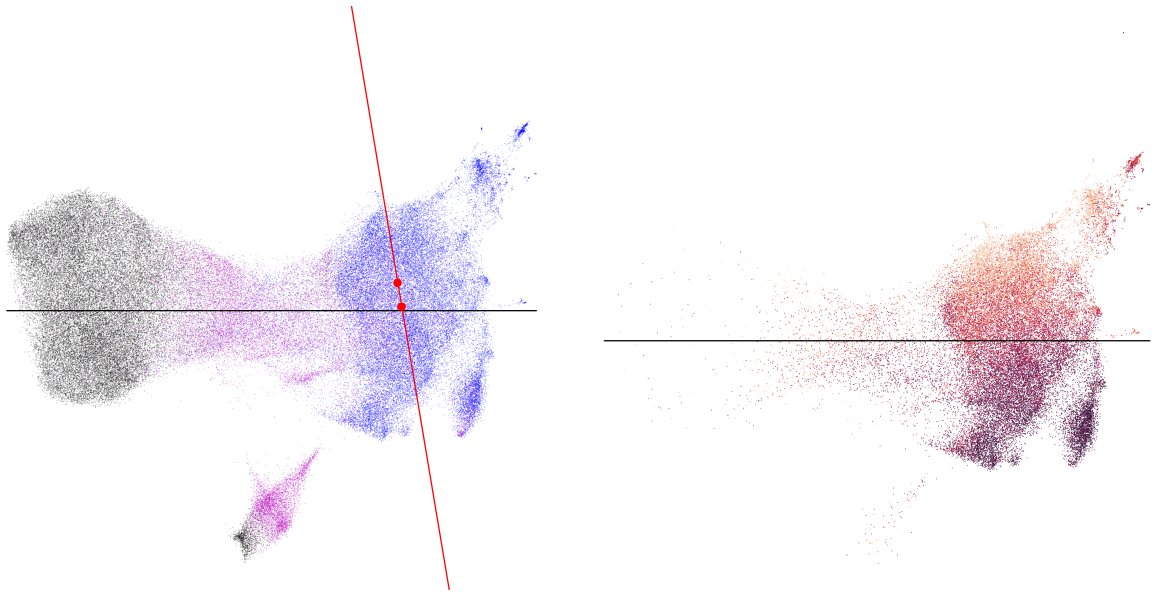

Figure 4: *Left panel*: UMAP projection of CharacterBERT embeddings for extant words (blue), pseudowords (magenta), and randomly-generated character n-grams (black). The solid black line shows the information axis that we define in this work, and the red line shows the axis that captures variability in word concreteness, computed by connecting the unweighted average UMAP position for extant words with that weighted by minmax-normalized concreteness (red dots). *Right panel*: UMAP of only extant words, colored by minmax-normalized concreteness, with lighter colors indicating more concrete words.

dimensional space suggests that the representations also capture notions of concreteness (e.g nouns) and abstractness (e.g adverbs) which we explore by projecting concreteness scores from the (Brysbaert et al., 2014) study. We calculate the center of extant UMAP coordinates with no weighting and with weighting by minmax-normalized concreteness and used those points to define a *Concreteness Axis* (Figure 4 left panel). It captures the visual intuition (Figure 4 right panel) that concreteness is roughly orthogonal to our information axis. The bootstrap-resampled angle distribution between information and concreteness axes is 86.6 ± 1.2 degrees.

This suggests that among the many latent features that structure the CharacterBERT representations, our information axis measure and word concreteness are approximately orthogonal to each other in projection. We leave a detailed investigation of this finding, including its relation to the visual information (Brysbaert et al., 2016) carried by concrete and abstract words, to future work.

5.4 Markov Chain Model

We also create a language model using the Prediction by Partial Matching (PPM) variable order Markov model (VOMM) to estimate the probability of each of these character n-grams (Begleiter et al., 2004). The model calculates the *logpdf* for each character n-gram in which more commonly occurring character n-grams have a lower score, and less commonly occurring character n-grams receive a high score. The model is trained on extant words, then used to score all of the extant, pseudowords and garble character n-grams. We use this score to capture the likelihood of character n-grams in our character sequence space (Figure 5).

These Markov model values correlate with our information axis measure. In particular, the Spearman correlation coefficient between information axis and Markov chain information content is 0.4 (highly significant) for randomly-generated n-grams, and 0.007 (not significant) for extant words. Thus, for random character n-grams, our information axis measure is correlated with statistical properties of the character n-grams from the Markov model (see the left panel of Figure 5). However, our information axis measure more clearly separates the classes of n-grams, thus going beyond purely statistical information (see the right panel of Figure 5). This suggests that the CharacterBERT model learns information beyond character-level statistical information, even for n-grams that never explicitly appear in the training data.

6 Discussion and Conclusion

Using the CharacterBERT model, we embedded a large corpus of character level n-grams outside of extant language to study how the primitive information they contain relates to the semantic information carried by extant language. The key findings of this paper are:

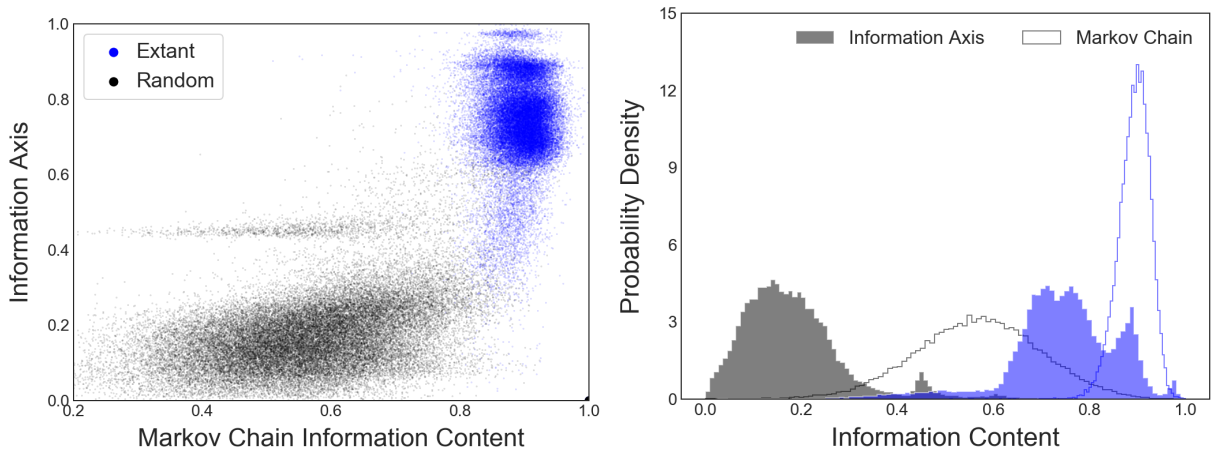

Figure 5: *Left panel*: Minmax-normalized position along the information axis shown in Figure 1 vs. minmax-normalized information content from our Markov Chain model, for extant words (blue) and randomly-generated character n-grams (black). *Right panel*: Probability density of minmax-normalized information content measures from our UMAP projection (filled histograms) and Markov Chain model (unfilled histograms).

1. Extant words and randomly-generated character n-grams are separated along a particular axis in our UMAP projection of CharacterBERT embedding space (Figures 1–2);

2. Pseudowords lie between extant and randomly-generated n-grams along this axis, but there is no distinct boundary between the classes of n-grams (Figures 1–2);

3. The structure of CharacterBERT embeddings of extant language, including based on part-of-speech and morphology, is correlated with the information axis (Figure 3);

4. Word concreteness varies along a dimension that is roughly orthogonal to the information axis in our UMAP projection (Figure 4);

5. Separation between extant and randomly-generated n-grams captured by CharacterBERT is correlated with and more coherent than that based purely on the statistical properties of n-grams (Figure 5).

These findings suggest that character-based Transformer models are largely able to explore the relation between extant words and randomly-generated character strings. In particular, character-level models capture complex structure in the space of words, pseudowords, and randomly-generated n-grams. These findings are consistent with work suggesting that character-level and morpheme-aware representations are rich in semantic meaning, even compared to word or sub-word models (Al-Rfou et al., 2019; El Boukkouri et al., 2020; Ma et al., 2020; Hofmann et al., 2020, 2021).

Our study is limited to extant words in English and randomly-generated character n-grams using the English alphabet. Given the unique impact of a specific language and alphabet on representation spaces, there is motivation to see whether the relationships we identify generalise to other languages and alphabets. Finally, we reiterate that our analysis was limited to the last embedding layer of the CharacterBERT model; future work may focus on weights in earlier layers, including attention mechanisms explored by other BERTology studies (Clark et al., 2019; Jawahar et al., 2019). By only analysing the final embedding layer, we study the 'psychology' of such character-level models; in analogy, much may be gained by studying the 'neuroscience' of such models encoded in their attention weights (Wang, 2020).

Our work may prompt avenues for future work with character-aware language models, such as the analyses of nonsense poetry like Lewis Caroll's "Jabberwocky," or the innovative and highly personalized lyrics of rap artists. Philological studies also may benefit from character-level models using broader n-gram spaces especially if dynamic analyses are employed, as may studies into lexicalization or pseudoword acceptability. Language acquisition studies which require the distinction of language from noise may also be aided by character-level models that perform well using information outside extant language.

References

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166.

Ron Begleiter, Ran El-Yaniv, and Golan Yona. 2004. On prediction using variable order markov models. *Journal of Artificial Intelligence Research*, 22:385–421.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in psychology*, 7:1116.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T Piantadosi. 2017. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, 41(8):2149–2169.

Charles Davis, Hannah Morrow, and Gary Lupyan. 2019. What does a horgous look like? nonsense words elicit meaningful drawings. *Cognitive Science*, 43.

David Deutsch. 2011. *The beginning of infinity: Explanations that transform the world*. Penguin UK.

David Deutsch and Chiara Marletto. 2015. Constructor theory of information. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2174):20140540.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Lisa Susan Ede. 1975. *The nonsense literature of Edward Lear and Lewis Carroll.* The Ohio State University.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915.

Guy Emerson. 2020. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7436–7453.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Manuela Friedrich and Angela D. Friederici. 2005. Phonotactic Knowledge and Lexical-Semantic Processing in One-year-olds: Brain Responses to Words and Nonsense Words in Picture Contexts. *Journal of Cognitive Neuroscience*, 17(11):1785–1802.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

V Hofmann, J Pierrehumbert, and H Schütze. 2020. Dagobert: generating derivational morphology with a pretrained language model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (and forerunners)(EMNLP)*. ACL Anthology.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves bert's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Daniel Jurafsky and James H Martin. 2021. Speech and language processing 3rd edition.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Jean-Jacques Lecercle. 2012. *Philosophy of nonsense: The intuitions of Victorian nonsense literature*. Routledge.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Charbert: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50.

David Marr and Ellen Hildreth. 1980. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Bhaskar Mitra, Nick Craswell, et al. 2018. *An introduction to neural information retrieval*. Now Foundations and Trends.

J Needle, J Pierrehumbert, and Jennifer B Hay. 2020. Phonological and morphological effects in the acceptability of pseudowords.

Winfried Noth. 1990. *Handbook of semiotics*. Indiana University Press.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Cathy J Price, RJS Wise, and RSJ Frackowiak. 1996. Demonstrating the implicit processing of visually presented words and pseudowords. *Cerebral cortex*, 6(1):62–70.

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.

Claude E Shannon. 1951. The redundancy of english. In *Cybernetics; Transactions of the 7th Conference, New York: Josiah Macy, Jr. Foundation*, pages 248–272.

Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Timothy D Sweeny, Nicole Wurnitsch, Alison Gopnik, and David Whitney. 2015. Ensemble perception of size in 4–5-year-old children. *Developmental science*, 18(4):556–568.

Harald Trost. 1992. Computational Morphology. In *Morphology and Computation*. The MIT Press.

Xin Wang. 2020. The curious case of developmental bertology: On sparsity, transfer learning, generalization and the brain. *arXiv preprint arXiv:2007.03774*.

Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y Ng. 2016. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*.