Analyzing Linguistic Knowledge in Sequential Model of Sentence

Peng Qian Xipeng Qiu* Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{pqian11, xpqiu, xjhuang}@fudan.edu.cn

Abstract

Sentence modelling is a fundamental topic in computational linguistics. Recently, deep learning-based sequential models of sentence, such as recurrent neural network, have proved to be effective in dealing with the non-sequential properties of human language. However, little is known about how a recurrent neural network captures linguistic knowledge. Here we propose to correlate the neuron activation pattern of a LSTM language model with rich language features at sequential, lexical and compositional level. Qualitative visualization as well as quantitative analysis under multilingual perspective reveals the effectiveness of gate neurons and indicates that LSTM learns to allow different neurons selectively respond to linguistic knowledge at different levels. Cross-language evidence shows that the model captures different aspects of linguistic properties for different languages due to the variance of syntactic complexity. Additionally, we analyze the influence of modelling strategy on linguistic knowledge encoded implicitly in different sequential models.

1 Introduction

Sentence modelling is a central and fundamental topic in the study of language generation and comprehension. With the application of popular deep learning methods, researchers have found that recurrent neural network can successfully model the non-sequential linguistic properties with sequential data input (Vinyals et al., 2015; Zhou and Xu, 2015; Rocktäschel et al., 2015). However, due to the complexity of the neural networks and the lack of effective analytic methodology, little is known about how a sequential model of sentence, such as recurrent neural network, captures linguistic knowledge. This makes it hard to understand the underlying mechanism as well as the model's strength and weakness. Previous work (Li et al., 2016) has attempted to visualize neural models in NLP, but only focus on analyzing the hidden layer and sentiment representation rather than grammar knowledge.

Currently, there have been a few attempts (Yogatama et al., 2014; Köhn, 2015; Faruqui and Dyer, 2015) at understanding what is embedded in the word vectors or building linguistically interpretable embeddings. Few works focus on investigating the linguistic knowledge encoded in a sequential neural network model of a sentence, not to mention the comparison of model behaviours from a cross-language perspective. Our work, therefore, aims to shedding new insights into the following topics:

a) How well does a sequential neural model (e.g. language model) encodes linguistic knowledge of different levels?
b) How does modelling strategy (e.g. the optimization objective) influence the neuron's ability of capturing linguistic knowledge?
c) Does the sequential model behave similarly towards typologically diverse languages?

To tackle the questions above, we propose to visualize and analyze the neuron activation pattern

*Corresponding author.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 826–835,
Austin, Texas, November 1-5, 2016.

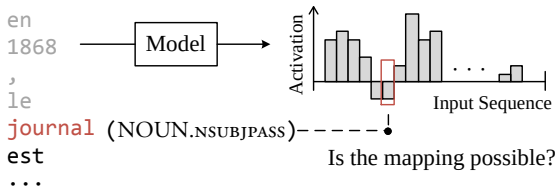


Figure 1: Experiment paradigm: correlating the dynamic activation pattern of the model neurons with linguistic features .

ID	(Non-)Linguistic Knowledge	Level
I	Sequence Length	Sequential
II	Gender / Definiteness Part-of-Speech	Lexical
III	Case / VerbForm / Mood Syntactic Role	Compositional

Table 1: List of the linguistic features to be correlated with model neuron behaviours.

so as to understand how a sequential neural model of sentence encodes linguistic properties of different level. By training vanilla LSTM language models with multilingual data and correlating the model neuron's activation with various linguistic features, we not only qualitatively show the activation pattern of a certain model neuron, but also quantify the selectivity of the neuron towards input language data or certain linguistic properties.

2 Methodology

2.1 A 'Brain' Metaphor of Artificial Model

Mitchell et al. (2008) correlates brain activities with linguistic stimuli under a popular brain-mapping paradigm. Since brain is a 'black box', researchers want to decode what is represented in a certain neuronal cluster of the brain at a certain time step. Here we propose that this paradigm can be applied to similar 'black-box' model, such as the neural network. This is what we call a 'brain' metaphor of the artificial model, as is visualized in Figure 1. We treat the neural network as a simplified 'brain'. We correlate the neuron behaviours with the input stimuli and design experiments to map the neuron activation to an explicit linguistic feature.

A sentence is, of course, a linear sequential arrangement of a cluster of words, but more than just a simple addition of words, as there exist complicated non-sequential syntactic relations. Thus, we consider three levels of features in the analysis of model behaviours, a) Sequential feature, a kind of superficial feature shared by any sequence data, b) Lexical feature, which is stable and almost independent of the sentence context, and c) Compositional feature, which is required for building the meaning of a sentence. Table 1 lists the details of the features involved in this paper.

2.2 Model Description

Since the goal is to understand the internal neurons' behaviour and how the behaviour patterns can be interpreted as a way to encode dynamic linguistic knowledge, we choose the most fundamental sequential sentence models as the research objects. We do not consider tree-structured model, as it explicitly involves linguistic structure in model architecture. We focus on word-based language model and compare it to two other counterparts in this paper.

Word-based Language Model Word-based language model (Mikolov et al., 2010) predicts the incoming word given the history context.

Character-based Language Model Instead of predicting the next word, character-based language model (Hermans and Schrauwen, 2013) predicts the incoming character given the history character sequence.

Task-specific Model A common task-specific model takes word sequence as input, but only predicts the category (e.g. sentiment) of the sentence after all the words are processed. In this paper, we consider a sequential model utilized for sentiment analysis task.

All the three sequential models are built on recurrent neural network with LSTM unit (Hochreiter and Schmidhuber, 1997). LSTM unit has a memory cell c and three gates: input gate i, output gate o and forget gate f, in addition to the hidden layer h of a vanilla RNN. The states of LSTM are updated as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o), \quad (4)$$

$$h_t = o_t \odot \tanh(c_t), \quad (5)$$

where x_t is the input vector at the current time step, σ denotes the logistic sigmoid function and $\odot$ denotes elementwise multiplication.

The dimension of the embeddings and the LSTM unit is 64. All three models use pretrained word embedding from Polyglot multilingual embeddings (Al-Rfou et al., 2013) trained with C&W (Collobert et al., 2011) model on Wikipedia. We train a lot of word-based and character-based LSTM language models with multilingual data from the Universal Treebank 1.2 (Joakim Nivre and Zhu, 2015), as well as a task-specific sentiment model on Stanford Sentiment Treebank (Socher et al., 2013b). We separate the training and testing data according to 90%/10% principle. We stop training when the loss of the test data does not decrease.

Regarding the analysis of the model behaviours, we collect the internal neuron activation of the hidden layer, three gates, and memory cell for all the data in the treebank/sentiment corpus[1]. For the sake of notation, we refer the hidden layer, input gate, output gate, forget gate and memory cell as h, i, f, o, c for three models, word-based language model (WL), character-based language model (CL) and task-specific model for sentiment analysis (SA). We mark the index of the neuron in the superscript and the meta information about the model in the subscript.

3 Qualitative Analysis

3.1 Sequential Feature

Karpathy et al. (2015) finds that some memory neurons of the character language model are selective towards the length of the input sequence. Similar patterns are also observed among the memory neuron activation pattern of the word-level language model as is shown in Figure 2, where deep purple color indicate strong negative activation and deep green color indicate strong positive activation. Moreover, we compute the correlation between the input sequence length and the activation pattern of every single neuron of h, i, f, o, c. Quantitative results in Figure 3 reveal that none of the hidden layer or gate neurons are strongly correlated with this sequential feature.

[1]The analyses cover languages such as English (en), German (de), Latin (la), Ancient Greek (grc), Bulgarian (bg), Spanish (es), Portuguese (pt), Italian (it), French (fr), Dutch (nl), Norwegian (no), Hindi (hi), Slovenian (sl), Hungarian (hu), Indonesian (id) and Chinese (zh).

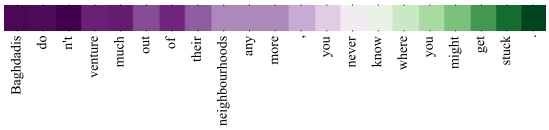


Figure 2: Memory neuron $c^{21}_{en,WL}$ that are sensitive to the length of the input word sequence.

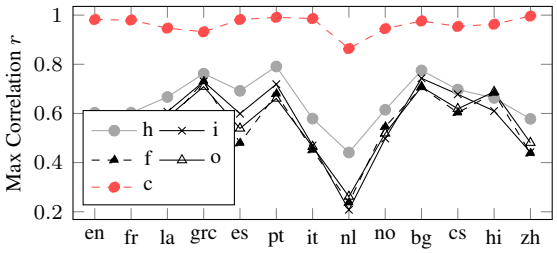
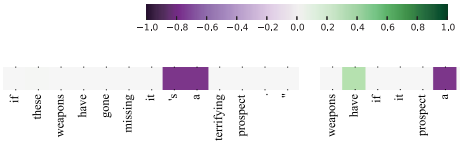


Figure 3: Comparison of neurons on correlating with the length of input sequence. Only the best correlation results are reported.



(a) $h^{30}_{en,WL}$ + raw sent (b) $h^{30}_{en,WL}$ + shuffled sent

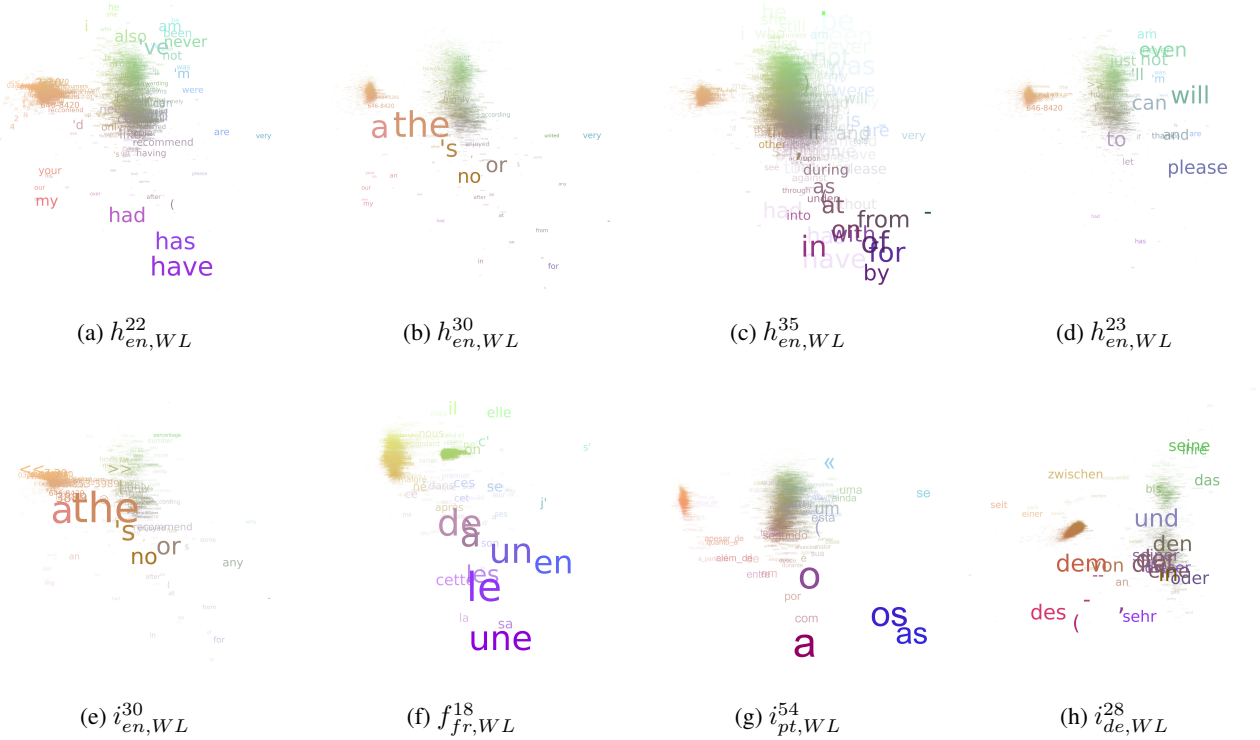
Figure 4: Visualising the activation of a neuron towards raw English sentences and sentences with shuffled word order.

3.2 Lexical Feature

For a inner neuron of the model, we can get the activation of a certain neuron in a certain model component towards a certain input word. A model neuron may be most active towards the words of some category instead of other words.

We notice that some neurons (e.g. Neuron $h^{30}_{en,WL}$) strongly activate towards functional words such as the determiners 'a', as is visualized in Figure 4. This activation can be observed even when we feed the model with a abnormal English sentence with shuffled word order.

Since it is not easy to go through all the neuron activation pattern, we design a visualization method to vividly show how a neuron selectively respond to



(a) $h^{22}_{en,WL}$ (b) $h^{30}_{en,WL}$ (c) $h^{35}_{en,WL}$ (d) $h^{23}_{en,WL}$

(e) $i^{30}_{en,WL}$ (f) $f^{18}_{fr,WL}$ (g) $i^{54}_{pt,WL}$ (h) $i^{28}_{de,WL}$

Figure 5: Visualising the lexical space responded by a certain neuron of the word-level language model trained on different languages.

certain words, inspired by the work in Cukur et al. (2013) and Huth et al. (2016).

Suppose the vocabulary of a language spans a default lexical space. An internal neuron of the artificial model modulates this linguistic space via showing selective activation pattern towards certain words. we carry out PCA on all the word embedding in the language model and extract the most prominent 3 principal components as the bases. Then we project the word vectors onto these three basis to get a new representation of the words in a low dimensional space. We draw all the words on a plane, where the location of each word is determined by the first two components and the text color is determined by three main components as RGB value. To visualize how a target neuron x respond to this lexical space, we modify the appearance of the word by scaling the size of the word text against the product of the log-transformed word frequency and the absolute value of the mean activation, and setting the degree of transparency of the text against the relative positive/negative activation strength among all the existed activation value of a target neuron x. In this way, large font size and low degree of transparency of a word w indicate that the target neuron x frequently activates towards the word w. This means that we can interpret a neuron's selectivity towards the lexical space just by looking for large and explicit words on the visualized two-dimensional space.

Figure 5 visualizes the lexical space responded by four hidden layer neuron of the English language model, as well as four gate neurons of different languages respectively. We can see that words with the similar grammatical functions are located near each other. Besides, it is interesting to see that some hidden layer neurons activate selectively towards determiner words, pronoun, preposition or auxiliary verbs. This phenomena have also been observed on gate neurons. For example, forget gate neuron $f^{18}_{fr,WL}$ activates towards the determiners in French. Input gate neuron $i^{54}_{pt,WL}$ activates towards the determiners in Portuguese. Notice that not all of the inner neurons show interpretable activation pattern towards the lexical space.

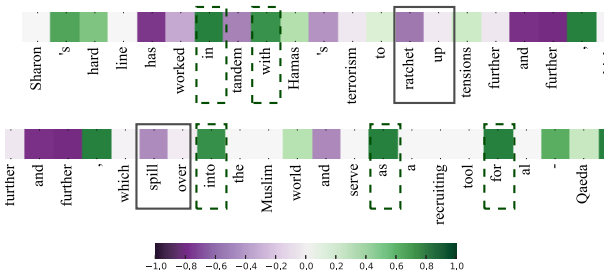


Figure 6: Visualising the Neuron $h^{35}_{en,WL}$ neuron activation towards verb-preposition composition.

3.3 Compositional Feature

To validate whether the internal neuron of the model can discriminate the local composition and long-distance composition, we choose the preposition as the object for observation.

In English, preposition can be combined with the previous verb to form a compound verbal phrase, such as 'check it *in*','give him *up*', 'find *out* what it will take'. This function of the preposition is annotated as the compound particle in the Universal Dependency Treebank. Another function of the preposition is to serve as the case marker, such as the preposition in the phrase 'lives *in* the central area', 'Performances will be performed *on* a project basis'. Given that these two functions of the preposition are not explicitly discriminated in the word form, the language model should tell the difference between the prepositions served as the compound particle and the prepositions served as the case marker if it indeed has the ability to handle word meaning composition.

For the hidden layer, we notice that hidden layer neuron $h^{35}_{en,WL}$ is sensitive to the function of the preposition. It only activates when the possible preposition does not form a composition with the former verb, as is vividly shown in Figure 6. The prepositions marked by dashed box serve as case marker while those in solid box form a phrase with previous verb. The activation pattern are obviously different. Similar pattern is also found in the gate neurons.

4 Quantitative Analysis

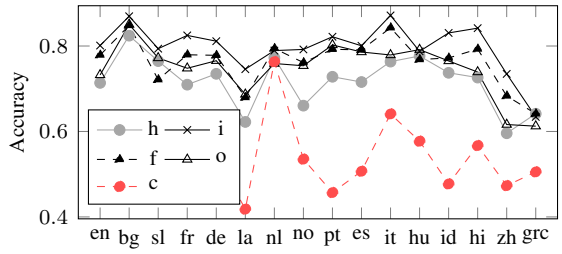
4.1 Decoding Lexical/Compositional Feature

Visualization only provides us with an intuitive idea of what a single neuron is encoding when processing language data. In this section, we employ a mapping paradigm to quantitatively reveal the linguistic knowledge distributed in the model components.

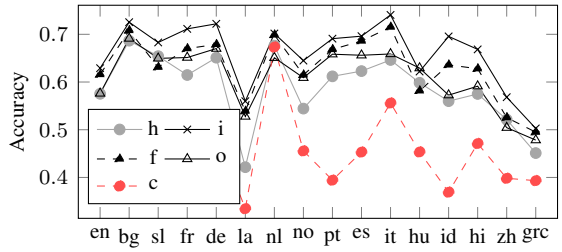
Instead of looking at one single neuron, here we use the whole 64 neurons of each model component as a 64-dimensional vector h, i, f, o, c respectively. The basic method is to decode interpretable linguistic features from target neuron clusters, which has been used in (Köhn, 2015; Qian et al., 2016). We hypothesize that there exists a map between a neuron cluster activation vector x and a high-level sparse linguistic feature vector y if the neuron cluster's activation pattern implicitly encode sufficient information about a certain lexical or compositional feature.

Hence we design a series of experiments to map the hidden layer, three gates, and memory cell vector activated by a target input word w in a sentence to the corresponding linguistic features of the word w, which are annotated in the Universal Dependency Treebank. Our experiments cover POS TAG, SYNTACTIC ROLE, GENDER, CASE, DEFINITENESS, VERB FORM and MOOD. These linguistic features are all represented as a one-hot vector. The mapping model is a simple softmax layer, with the activation vector as the input and the sparse vector as the output. For each linguistic feature of each language, a mapping model is trained on the randomly-selected 90% of all the word tokens and evaluated over the remaining 10%. Notice that GENDER, CASE, DEFINITENESS, VERB FORM, and MOOD only apply to certain word categories. We give a default 'N/A' tag to the words without these annotations so that all the word can be used for training. The evaluation result is only computed from the words with the features. This requires the mapping model to not only recognize the differences between the sub-categories of a linguistic feature (e.g. CASE), but also discriminate the words that we are interested in from other unrelated words (e.g. words without CASE annotations). Accuracies for each model component h, i, f, o, c are reported in Figure 7 and 8.

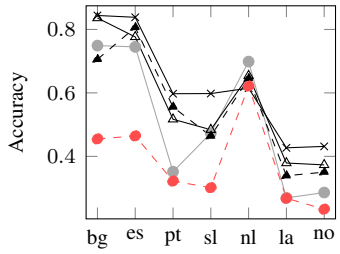
Comparing different model components, we notice that gate neurons except output gate are generally better than hidden layer and memory cell neurons on decoding linguistic knowledge. Input gate and



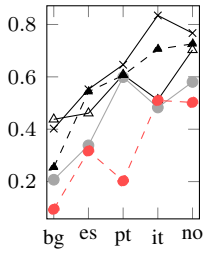
(a) POS TAG



(b) SYNTACTIC ROLE

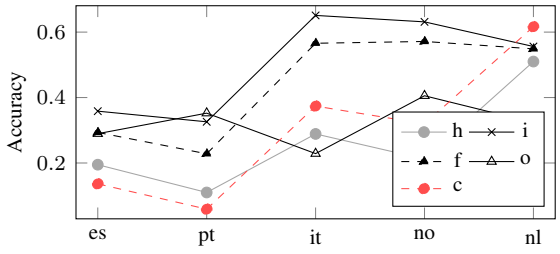


(c) CASE

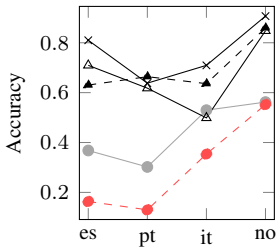


(d) GENDER

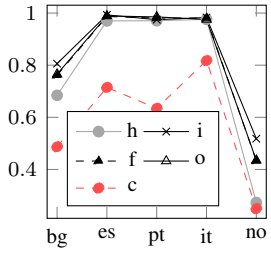
Figure 7: Comparison of neurons on decoding POS AG, SYNTACTIC ROLE, CASE, and GENDER.



(a) VERB FORM



(b) TENSE



(c) DEFINITENESS

Figure 8: Comparison of LSTM neurons on decoding VERB FORM, TENSE, and DEFINITENESS.

forget gate are the best, while memory cell is the worst. It shows that the gates of a recurrent language model are more sensitive to the grammar knowledge of the input words.

Comparing decoding results on different languages, we find that it is generally easier to decode POS TAG than SYNTACTIC ROLE for all the languages. One interesting thing is that the mapping model works better with Bulgarian, a slavic language, but worse on Norwegian on decoding CASE while the situation is opposite on decoding GENDER. It might be because that gender is a weakened grammatical feature in Bulgarian. Therefore, knowledge about GENDER may not be so important in building the grammatical structure of the Bulgarian language data.

4.2 The Dynamics of Neuron Behaviour

Since sentence meaning is dynamically constructed by processing the input sequence in a word-by-word way, it is reasonable to hypothesize that the linguistic feature of an input word w won't sharply decay in the process. Naturally, we would like to ask whether it is possible to decode, or at least partially infer, a word's property from the neuron behaviours of its context words. Specifically, if the model process a verbal phrase '*spill over*' or '*in the garden*', will the property of the word '*spill*', '*in*' be combined with the following word and decodable from the model neuron activation behaviours towards the following word, or will the property of the word '*over*', '*the garden*' be primed by the previous word and decodable from the model neuron behaviours towards the previous word?

To quantitatively explore this question, we carry out a mapping experiment similar to the previous one. The difference is that here we map the hidden layer, three gates, and memory cell vector activated by a target input word w in a sentence to the corresponding linguistic features of the previous/following word $w_{-2/-1}/w_{+1/+2}$ in a 5-word window context. Results in Figure 9 shows that the linguistic feature POS TAG is partially primed or

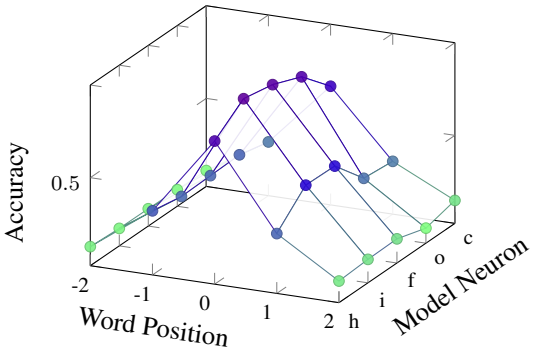


Figure 9: Neuron dynamics on decoding POS.

ISO	**Language**	f	i	o	h
en	English	0.316	-0.022	0.107	0.156
la	Latin	0.152	0.131	0.158	0.085
grc	Ancient Greek	0.293	0.248	0.166	0.274
pt	Portuguese	0.301	0.313	0.161	0.209
nl	Dutch	0.196	0.096	0.205	-0.134
no	Norwegian	0.335	0.057	0.269	0.033
bg	Bulgarian	0.324	0.280	0.071	-0.082

Table 2: Comparison of model components' correlation with tree structure stastistics.

kept in the context words in English. The longer distance, the less probability to decode it from the neuron activations. Still, the nearest context words w_{-1} and w_{+1} prime/keep the most relevant information of the target word w. Similar patterns are also found for other linguistic feature in other languages.

4.3 Correlation with Dependency Tree

Since the sequential model can modelling non-sequential input, we naturally want to know whether any component of the model is dynamically correlated with the statistics of tree structure. Inspired by the case study in Zhou and Xu (2015), we count the syntactic depth of each word in a sentence and compute the correlation between the depth sequence and the dynamics of the average activation of the model neurons in Table 2. We did not find strong correlation between the mean neuron activation dynamics with the syntactic tree depth. One possible explanation is that the language model only use the history information, while the depth of a word is computed in a relative global context.

5 Model Comparison

In this section, we would like to investigate whether different sentence modelling strategy and optimization objective affect the neuron's implicit encoding of linguistic knowledge, especially the grammatical properties.

5.1 Word vs. Character

It is obvious that word-based language model and character-based language model intend to model the language data at different granularity. Although both of them are effective, the latter is often criticized for an unreasonable modelling strategy.

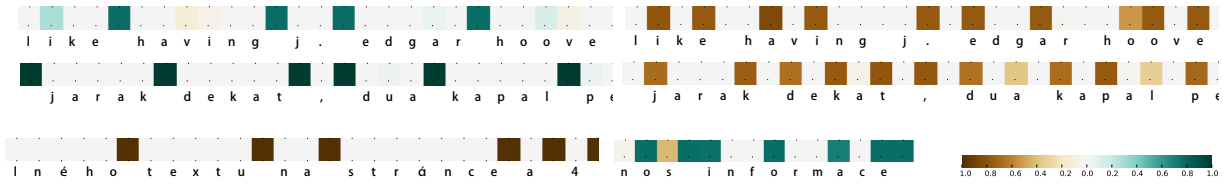
In addition to the findings in Karpathy et al. (2015), we see that some of the hidden layer neurons of the character-based language model seems to be sensitive to specific characters and character clusters, as is indicated from the visualization of the neuron activation pattern in Figure 10. We are surprised to find that some neuron of the hidden layer activates selectively towards white space character. This is interesting as it means that the model learns to detect word boundary, which is exactly an important linguistic feature.

Besides, some neuron activates selectively towards vowel/consonant characters in a phonographic language, such as English. This interesting phenomenon also indicates that the model implicitly captures the phonology system, since it can discriminate the vowel character clusters from the consonant character clusters. We also find these two detectors in other languages, such as Indonesian and Czech in Figure 10.

5.2 Word Prediction vs. Task-specific Model

We compare a word-based LSTM language model and a word-based LSTM sentiment model. Here, for a fair comparison, all the models are trained only on the Stanford Sentiment Treebank Dataset (Socher et al., 2013a). The results show that the neurons in these two models displays similar behaviours towards superficial sequential features, but totally different behaviours towards high-level linguistic features, such as semantic and syntactic knowledge.

Both some of the internal neurons of the memory cell in the language model and the sentiment model emerge to be sensitive to the length of the



(a) $h^{19}_{en,CL}$, $h^{33}_{id,CL}$, $h^{9}_{cs,CL}$ detect white space. (b) $h^{16}_{en,CL}$, $h^{7}_{id,CL}$, $h^{15}_{cs,CL}$ detect phonology.

Figure 10: Visualising the activation of hidden neurons of English, Indonesian and Czech language model.

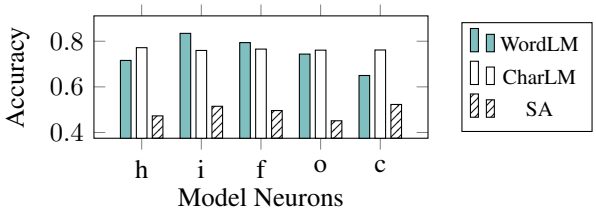


Figure 11: Comparison between the internal neurons of English word-based language model, character-based language model and sentiment model on decoding POS TAG.

input sequence, as we expected, since the sentence length is a non-linguistic features shared by all the sequential input. However, different optimization objectives force the models to represent and capture the linguistic properties of different aspects. The language model focus more on the syntactic aspect, as is visualized and quantified in previous sections. Neurons of the sentiment model tends to be sensitive only towards the sentiment aspect of the words, although the sentiment model use the similar LSTM unit, dimensionality and pretrained embedding. We apply the same visualization method in Section 3.2 to the 64 hidden layer neurons of the sentiment model and manually interpret the visualization results one by one. We did not see any strong activation pattern towards the functional words like those found in language model hidden layer neurons.

To quantify the differences of the linguistic knowledge encoded in different sentential model, we again use the previous feature-decoding experiment method. We compare the performance of the components in three models on decoding POS TAG from English data. Notice that we use Stanford POS Tagger (Kristina Toutanova and Singer, 2003) to automatically tag the sentences in the sentiment data. For the character-based language model, we use the neuron activation towards the end character of each words in the decoding experiment.

Results in Figure 11 shows that even a character-based language model can achieve pretty well on decoding the most important lexical features from the activation pattern of the internal neurons. This is a strong evidence that word-level feature detector can emerge from a pure character-based model. Sentiment model, on the contrary, fails to capture the grammatical knowledge, although we might think that a successful sentiment analysis model should be able to combines the grammar property of the words with the sentiment information. Current results indicate that for pure sequential model with vanilla LSTM units, the objective of the sentence modelling tasks will largely affect how the model acquires and encodes linguistic knowledge.

6 Related Works

Karpathy et al. (2015) explores the memory cell in character-based language model. Their visualization results show some interesting properties of the memory neurons in LSTM unit. However, their exploration on character-based model does not intend to correlate high-level linguistic knowledge, which are intuitively required for sequential modelling of a sentence.

Li et al. (2016) propose a method for visualizing RNN-based sentiment analysis models and word-based LSTM auto-encoder in NLP tasks. Li et al. (2015) investigates the necessity of tree structure for the modelling non-sequential properties of languages. Bowman et al. (2015) studies the LSTM's ability of capturing non-sequential tree structure. Despite the useful findings, these works make no attempts to investigate the internal states of the neurons for a better understanding of the model's power or weakness.

Our work not only provides qualitative visualization of model neurons' behaviours and detailed

quantitative investigation with multilingual evidence (16 for POS decoding experiment), but also reveal the influence of language syntactic complexity and modelling strategy on how well the internal neurons capture linguistic knowledge, which have been overlooked by previous work on interpreting neural network models.

7 Conclusion

In this work, we analyze the linguistic knowledge implicitly encoded in the sequential model of sentence. Through the visualization and quantification of the correlation between the neuron activation behaviour of different model components and linguistic features, we summarize that:

- Model neurons encode linguistic features at different level. Gate neurons encode more linguistic knowledge than memory cell neurons.
- Low-level sequential features are shared across models while high-level linguistic knowledge (lexical/compositional feature) are better captured by language model instead of task-specified model on sentiment analysis.
- Multilingual evidence indicates that the model are sensitive to the syntactic complexity of the language. It would also be a promising direction to incorporate the factor of language typological diversity when designing advanced general sequential model for languages other than English.
- Word-level feature detector can emerge from a pure character-based model, due to the utility of character composition.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was partially funded by National Natural Science Foundation of China (No. 61532011 and 61672162), the National High Technology Research and Development Program of China (No. 2015AA015408).

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Samuel R. Bowman, Christopher D. Manning, and Christopher Potts. 2015. Tree-structured composition in neural networks without tree-structured architectures. *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches.*

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Tolga Cukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. 2013. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–70.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. *arXiv preprint arXiv:1506.05230*.

M. Hermans and B. Schrauwen. 2013. Training and analysing deep recurrent neural networks. *Advances in Neural Information Processing Systems*, pages 190–198.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Alexander G. Huth, Wendy A. De Heer, Thomas L. Griffiths, Frdric E. Theunissen, and Jack L. Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*.

Maria Jesus Aranzabe Masayuki Asahara Aitziber Atutxa Miguel Ballesteros John Bauer Kepa Bengoetxea Riyaz Ahmad Bhat Cristina Bosco Sam Bowman Giuseppe G. A. Celano Miriam Connor Marie-Catherine de Marneffe Arantza Diaz de Ilarraza Kaja Dobrovoljc Timothy Dozat Tomaž Erjavec Richárd Farkas Jennifer Foster Daniel Galbraith Filip Ginter Iakes Goenaga Koldo Gojenola Yoav Goldberg Berta Gonzales Bruno Guillaume Jan Hajič Dag Haug Radu Ion Elena Irimia Anders Johannsen Hiroshi Kanayama Jenna Kanerva Simon Krek Veronika Laippala Alessandro Lenci Nikola Ljubešić Teresa Lynn Christopher Manning Ctlina Mrnduc David Mareček Héctor Martínez Alonso Jan Mašek Yuji Matsumoto Ryan McDonald Anna Missilä Verginica Mititelu Yusuke Miyao Simonetta Montemagni Shunsuke Mori Hanna Nurmi Petya Osenova Lilja Øvrelid Elena Pascual Marco Passarotti Cenel-Augusto Perez Slav Petrov Jussi Piitulainen Barbara Plank Martin Popel Prokopis Prokopidis Sampo Pyysalo Loganathan Ramasamy Rudolf Rosa Shadi Saleh Sebastian Schuster Wolfgang Seeker Mojgan Seraji Natalia Silveira Maria Simi Radu Simionescu Katalin Simkó Kiril Simov Aaron

Smith Jan Štěpánek Alane Suhr Zsolt Szántó Takaaki Tanaka Reut Tsarfaty Sumire Uematsu Larraitz Uria Viktor Varga Veronika Vincze Zdeněk Žabokrtský Daniel Zeman Joakim Nivre, Željko Agić and Hanzhi Zhu. 2015. Universal dependencies 1.2. In *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague*.

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

Arne Köhn. 2015. Whats in an embedding? analyzing word embeddings through multilingual evaluation.

Christopher Manning Kristina Toutanova, Dan Klein and Yoram Singer. 2003. Part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*.

Jiwei Li, Minh Thang Luong, Jurafsky Dan, and Eudard Hovy. 2015. When are tree structures necessary for deep learning of representations? *Proceedings of EMNLP*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of NAACL*.

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

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Investigating language universal and specific properties in word embeddings. In *Proceedings of ACL*.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, and Andrew Ng andmChristopher Potts. 2013a. Parsing with compositional vector grammars. In *EMNLP*.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the EMNLP*.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2755–2763.

Dani Yogatama, Manaal Faruqui, Chris Dyer, and Noah A Smith. 2014. Learning word representations with hierarchical sparse coding. *arXiv preprint arXiv:1406.2035*.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.