Investigating Language Universal and Specific Properties in Word Embeddings

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

Recently, many NLP tasks have benefited from distributed word representation. However, it remains unknown whether embedding models are really immune to the typological diversity of languages, despite the language-independent architecture. Here we investigate three representative models on a large set of language samples by mapping dense embedding to sparse linguistic property space. Experiment results reveal the language universal and specific properties encoded in various word representation. Additionally, strong evidence supports the utility of word form, especially for inflectional languages.

1 Introduction

Word representation is a core issue in natural language processing. Context-based word representation, which is inspired by Harris (1954), has achieved huge successes in many NLP applications. Despite its popularity, character-based approach also comes out as an equal competitor (Santos and Zadrozny, 2014; Kim et al., 2016; Ling et al., 2015b; Ling et al., 2015a; Faruqui et al., 2016; Ballesteros et al., 2015) . Moreover, questions arise when we consider what these models could capture from linguistic cues under the perspective of cross-language typological diversity, as is argued by Bender (2009).

Despite previous efforts in empirically interpreting word embedding and exploring the intrinsic/extrinsic factors in learning process (Andreas and Klein, 2014; Lai et al., 2015; Köhn, 2015; Melamud et al., 2016), it remains unknown whether embedding models are really immune to the structural variance of languages.

Current research has gaps for understanding model behaviours towards language typological diversity as well as the utility of context and form for different languages. Thus, we select three representative types of models and design a series of experiments to reveal the universals and specifics of various word representations on decoding linguistic properties. Our work contributes to shedding new insights into the following topics:

a) How do typological differences of language structure influence a word embedding model? Does a model behave similarly towards phylogenetically-related languages?

b) Is word form a more efficient predictor of a certain grammatical function than word context for specific languages?

c) How do the neurons of a model respond to linguistic features? Can we explain the utility of context and form by analyzing neuron activation pattern?

2 Experiment Design

To study the proposed questions above, we design four series of experiments to comprehensively compare context-based and character-based word representations on different languages, covering syntactic, morphological and semantic properties. The basic paradigm is to decode interpretable linguistic features from a target collection of word representations. We hypothesize that there exists a linear/nonlinear map between a word representation x and a high-level sparse feature vector y if the word vector implicitly encode sufficient information[1]. Figure1 visualizes how a

*Corresponding author.

[1]Our experiment results show that nonlinear mapping model significantly works better than linear map for all languages. Only nonlinear mapping accuracies are mentioned in the following sections due to the space limit.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1478–1488,
Berlin, Germany, August 7-12, 2016.

word embedding is mapped to different linguistic attribute vectors. For example, the Czech word *dětem* means children in English. Its grammatical gender is female. It is in the plural form and should be used in dative case. These are all important properties of a word. The word embedding of *dětem* is mapped to different sparse representation of these lexical properties respectively.

Listed in Table1 is the outline of the experiments.

ID	Attribute	Category
I	Part-of-Speech	Syntax
II	Dependency Relation	
III	Gender / Number / Case Animacy / Definite / Person Tense / Aspect / Mood / Voice PronType / VerbForm	Morphology
IV	Sentiment Score	Semantics

Table 1: Outline of Experiment Design.

For linear map, we train a matrix Θ that maps word embedding $\mathbf{x}$ to a sparse feature vector $\mathbf{y}$ with the least L_2 error. For nonlinear map, we train a neural network (MLP) with 4 hidden layers via back propagation. Their dimensions are 50, 80, 80, and 50 in order. For each linguistic feature of each language, a mapping model is trained on the randomly-selected 90% of the words with the target feature and tested over the remaining 10%. Details about the construction of the linguistic feature vectors will be mentioned in the specific section of a certain experiment.

For syntactic and morphological features, we construct the corresponding feature vectors of a word from the Universal Dependencies Treebank (Joakim Nivre and Zhu, 2015) and the Chinese Treebank (CTB 7.0) (Xue et al., 2010). For a certain word w with a certain linguistic attribute a (e.g. POS), w may be annotated with one or different labels (e.g. NOUN, VERB, etc) from the possible label set of a in the whole treebank. We calculate the normalized label frequency distribution $\vec{\mathbf{y}}_w^a$ from the manual annotation of the corpus as the representation of the linguistic attribute a for the word w in each language.

For word sentiment feature, we use the manually annotated data collected by Dodds et al. (2015). The data contains emotion scores for a list of words in several languages. In our experiment, the original score scale in Dodds et al. (2015) is transformed into the interval $[0, 1]$.

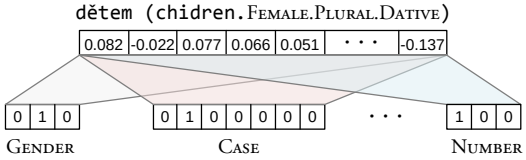


Figure 1: Visualizing experiment paradigm. The dense representation of a Czech word *dětem* is mapped to different sparse representation of the lexical properties respectively.

3 Embedding Model Description

Faced with three questions proposed before, we select the following models from various candidates, as they are popular, representative and based on either word context or purely word form.

Type I C&W Model (referred as CW in short), which aims to estimate the joint probability of a word sequence (Collobert et al., 2011). In this paper, C&W word vectors are all from the released version of the polyglot multilingual embeddings (Al-Rfou et al., 2013) trained on Wikipedia.

Type II Skip-gram[2] (referred as SG in short), which aims to predict the context words based on the target word. We use word2vec (Mikolov et al., 2013) to train SG on multilingual Wikipediea provided by (Al-Rfou et al., 2013).

Type III Character-based LSTM autoencoder (referred as AE in short), which takes the character sequence of a word as the input and reconstruct the input character sequence. It takes the advantage of pure word form instead of the context. The hidden layer vector of the model is used as a representation of the word. In this way, we are able to quantify the utility of pure word form by evaluating the representation generated from the character-based LSTM autoencoder on different decoding tasks. We trained one-hidden layer AE with the words covered in CW for each language independently.

To ensure a fair comparison, all the word vectors have the same dimension 64. CW and SG are trained with a common 5-word window size.

4 Results

4.1 Part-of-Speech

In experiment I, we decode Part-of-Speech, the most basic syntactic feature, from word embed-

[2]SG results for some languages are missed due to the lack of the corpus data or special preprocessing.

ISO	Language	$\|V\|$	CW	SG	AE	
ar	Arabic	24967	0.712	0.658	0.648	I
ga	Irish	3164	0.826	–	0.697	
zh	Chinese	30496	0.780	0.721	n/a	II
fa	Persian	11471	0.895	0.827	0.746	III
la	Latin	6678	0.746	–	0.707	
hi	Hindi	12703	0.858	0.799	0.592	IV
ta	Tamil	1940	0.768	–	0.541	
eu	Basque	11212	0.857	–	0.711	
et	Estonian	2166	0.862	0.765	0.530	V
fi	Finnish	26086	0.910	0.818	0.715	
hu	Hungarian	6105	0.912	0.831	0.674	
de	German	29899	0.916	0.902	0.74	VI
fr	French	29445	0.905	0.889	0.759	VII
pt	Portuguese	17715	0.927	0.903	0.746	
he	Hebrew	22754	0.911	–	0.680	
ru	Russian	55416	0.959	0.913	0.906	
hr	Croatian	12581	0.926	0.862	0.790	
da	Danish	10705	0.913	0.913	0.666	
sv	Swedish	8408	0.938	0.888	0.670	
no	Norwegian	18709	0.926	0.861	0.704	
sl	Slovenian	19514	0.919	0.820	0.756	
cs	Czech	55789	0.949	0.883	0.853	
ro	Romanian	3170	0.858	0.814	0.618	
en	English	15116	0.857	0.839	0.659	
id	Indonesian	15635	0.852	0.819	0.801	
it	Italian	21184	0.902	0.880	0.700	VIII
es	Spanish	33696	0.906	0.883	0.75	
el	Greek	8499	0.937	0.879	0.801	
pl	Polish	18062	0.941	0.842	0.800	
bg	Bulgarian	17079	0.920	0.852	0.741	

Table 2: Model comparison on decoding POS, along with WALS word-order features. Type I: VS+VO+Pre+NR. II: SV+VO+Pre+RN. III: SV+OV+Pre+NR. IV: SV+OV+Post+RN/Co. V: SV+OV+Post+NR. VI: SV+ND+Pre+NR. VII: SV+VO+Pre+NR. VIII: ND+VO+Pre+NR.

ding. To construct the POS vector for each word, we calculate the normalized POS-tag frequency distribution from the manual annotation of the Universal Dependencies (Version 1.2) (De Marneffe et al., 2014) and Chinese Treebank (CTB 7.0) (Xue et al., 2010) for each language.

We evaluate the predicted results by judging whether the most probable POS tag of a word predicted by the model equals to the most probable correct POS tag of the word. Formally, for a set of words W in a language, the correct tag of the i^{th} word W_i is $y^a_{W_i}$ and the predicted tag is $\hat{y}^a_{W_i}$. The accuracy is computed as:

$$acc = \frac{1}{|W|}\sum_{i}^{|W|}\Delta(\hat{y}^a_{W_i}, y^a_{W_i}) \quad (1)$$

$$\Delta(\hat{y}^a_{W_i}, y^a_{W_i}) = \begin{cases} 1 & \hat{y}^a_{W_i} = y^a_{W_i} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

It is obvious that context-based representation (CW and SG) performs better than character-based representation (AE). We, however, notice that AE peforms nearly as well as the context-based embedding on Russian, Czech and Indonesian.

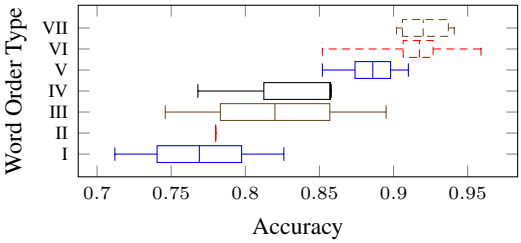


Figure 2: Interaction between CW performances on decoding POS tag and WALS word order features.

It turns out that these languages employ affix markers to indicate the POS category of a word. For example, in Indonesian, co-occurrence of the prefix *'me-'* and the suffix *'-kan'* in the word form means that this word is a verb.

Besides, we explore the relationship between CW performances on decoding POS tags and the word order typology of different languages, since CW is sensitive to word order. We classify the languages into 8 types, based on the basic word order features (Order of Subject and Verb; Order of Object and Verb; Order of Noun and Adposition; Order of Noun and Relative clause) from the World Atlas of Language Structures (Dryer and Haspelmath, 2013). Figure 2 shows that CW performs similar in this experiment for languages of the same word order type, indicating an implicit interaction between typological diversity and model performance.

4.2 Dependency Relation

In this section, we will get into the details of Experiment II: decoding dependency relation from word representation. Dependency relation refers to how a word is syntactically related to other words in a sentence. It is the label annotated on the arc of the dependency tree.

We compute the normalized frequency distribution of dependency relations for each word in the Universal Dependency Treebank and Chinese Treebank (CTB 7.0) (Xue et al., 2010). The distribution of dependency relations is the probabilistic distribution of different arc types, such as subject, object, nmod, etc. Evaluation is similar to that in Section 4.1.

We can see from Figure 3 that the overall performance is worse than that in Experiment I, as dependency analysis is more difficult than POS induction. CW achieves the best performance. It

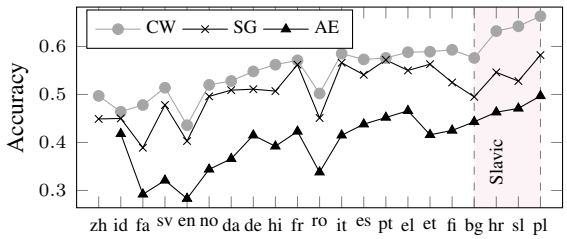


Figure 3: Comparison of models on decoding DEPENDENCY RELATION.

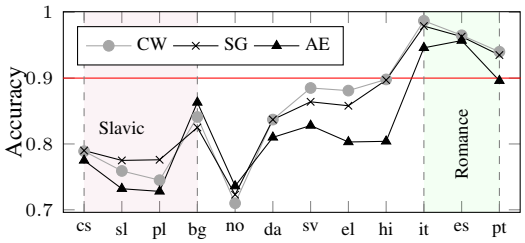


Figure 4: Comparison of models on decoding GENDER.

is also interesting to see that all the embeddings work slightly better on Slavic languages.

4.3 Morphological Features

Experiment III aims to decode morphological information from various word representation. Evaluation is similar to that in Section 4.1. Morphological information refers to the explicit marker of the grammatical functions. We consider 12 morphological features, as is shown in Table 1. They can be split into 5 nominal features (GENDER, NUMBER, CASE, ANIMACY, DEFINITENESS) and 7 verbal features (PERSON, TENSE, ASPECT, MOOD, VOICE, PRONOUNTYPE, VERBFORM).

Gender is a very special feature for western languages. It is partially based on semantics, such as biological sex. In most of the languages with gender features, there are agreements between the noun and the determiners. This could be a good indicator for context-based model. On the other hand, gender is also expressed as an inflectional feature via declension or umlaut, especially for adjectives and verbs. Therefore, we can see from Figure 4 that the AE also achieves some good results without using context information.

From a typological perspective, we found that all the embeddings work well on decoding word gender of Romance languages (Italian, Spanish and Portuguese) but worst on Slavic languages (e.g. Czech, Slovenian). This is probably

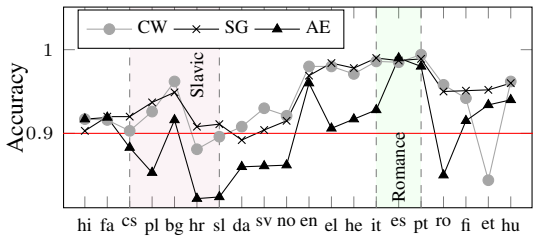


Figure 5: Model comparison on decoding NUMBER.

	Language	$\|V\|$	C&W	SG	AE	# Case
Analy.	Danish	372	0.947	0.946	1.000	3
	Swedish	5893	0.995	0.990	0.981	2
	Bulgarian	104	0.636	0.546	0.818	4
Agglut.	Finnish	21094	0.868	0.871	**0.908**	15
	Hungarian	4536	0.852		**0.901**	22
	Tamil	1144	0.896	–	**0.835**	7
	Basque	8020	0.761	–	**0.857**	15
Fusional	Hindi	10682	0.712	0.704	0.646	7
	Czech	38666	0.788	0.776	0.663	7
	Polish	13715	0.828	0.785	0.636	7
	Slovenian	15150	0.796	0.768	0.617	6
	Croatian	9945	0.807	0.789	0.628	7
	Greek	5790	0.841	0.851	0.774	5
	Latin	4773	0.674	–	0.636	7

Table 3: Model comparison on decoding CASE.

because that Romance languages employ regular rules to judge the gender of a word. However, Slavic languages have other nonlinear fusional morphological features that are not easy to tackle.

Number refers to the linguistic abstraction of objects' quantities. It is an inflectional feature of noun and other parts of speech (adjective, verb) that have agreement with noun. The basic value can be singular, dual or plural. We can see from Figure 5 that SG, CW and AE all perform well. AE performs almost as well as CW and SG on English, Spanish and Portuguese.

Case is one of the most significant features. Gender and number are indexical morphemes, which means that there is a phrase in the sentence that necessarily agrees with the target item. Case, on the contrary, is a relational morpheme, according to (Croft, 2002). Case reflects the semantic role of a noun, relative to the pivot verb. All the languages studied in this paper, more or less, employ word inflection to explicitly express the specific case role. The model performances are listed in Table 3.

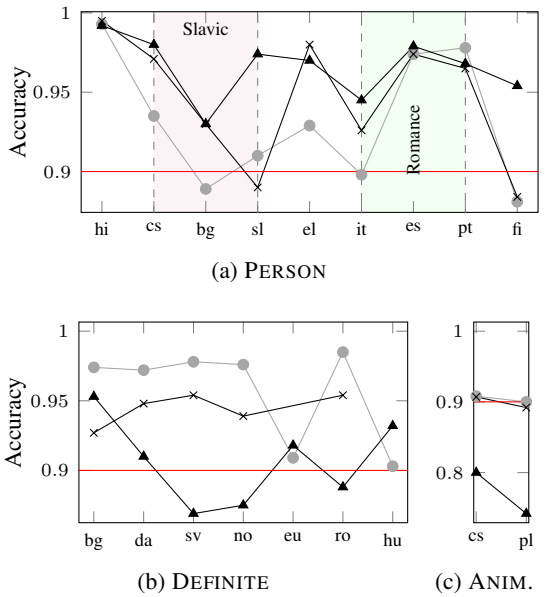
We notice some important inter-language differences. Swedish has only two cases, nominal and genitive. The form of genitive case is very simple. Adding an *s* to the coda of a noun will change it to genitive case. Thus, we can see that character-based encoding performs well

on Swedish. Since genitive case usually means possession, we also notice that context-based distributed representation also performs well in decoding case information from Swedish words.

By classifying these languages into different morphological types in Table 3, we find that word vectors of highly inflected fusional languages (e.g. Czech) performs worse than agglutinative languages (e.g. Finnish). This is typically reflected in AE, as agglutinative languages simply concatenate the case marker with the nominative form of a noun. The morphological transformation of agglutinative languages is linear and simple. Besides, the case system of the analytic languages has been largely simplified due to historical change. Therefore, all the embeddings perform well on analytic languages. This evidence supports that morphological complexity is positively correlated with the quality of word embedding.

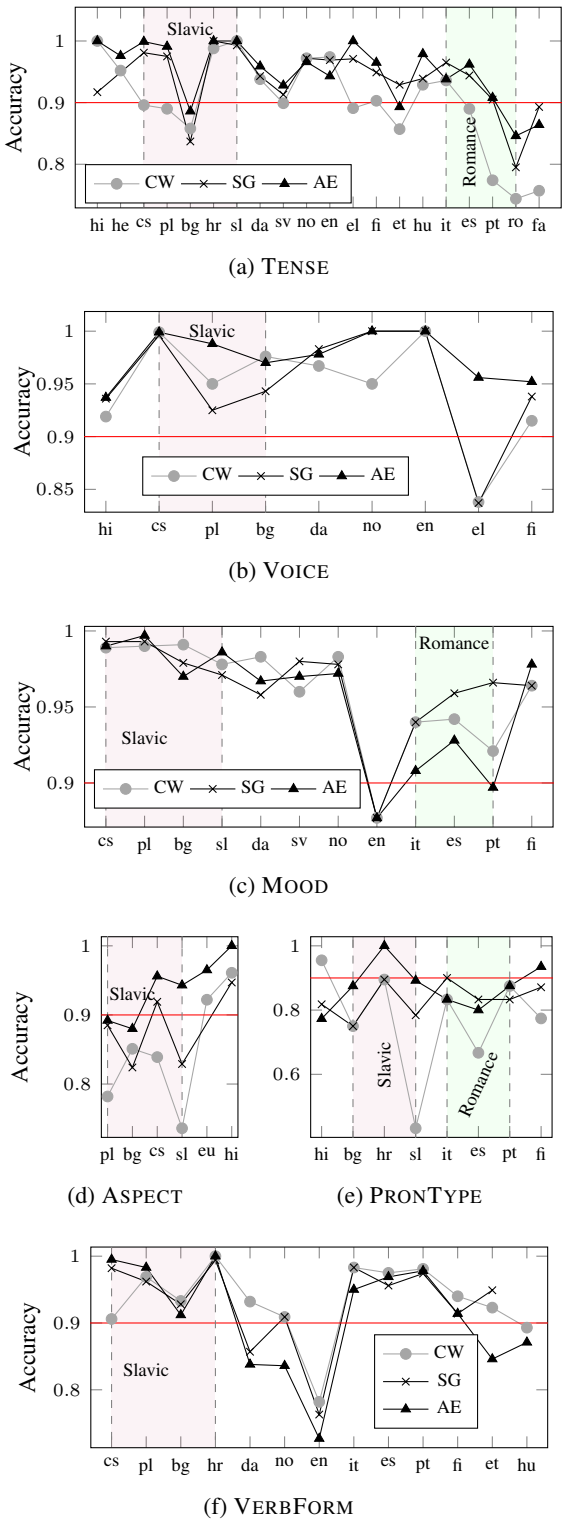
Besides, for fusional languages, using distributed representation and context information would largely increase the performance. This, in turn, indicates that cases are a special semantic relations distributed in the words around the target noun. Although a case is not explicitly agreed with other components in an utterance, the word category might serve as a good indicator, such as preposition and verb.

Animacy is a special nominal feature in a few languages, which is used to discriminate alive and animate objects from inanimate nouns. Generally, it is based on the lexical semantic feature. As is shown in Figure 6, it is easier to decode animacy from the context-based representations than character-based representation.



(a) PERSON (b) DEFINITE (c) ANIM.

Figure 6: Model comparison on decoding PERSON, DEFINITENESS and ANIMACY.



(a) TENSE (b) VOICE (c) MOOD (d) ASPECT (e) PRONTYPE (f) VERBFORM

Figure 7: Model comparison on TENSE, VOICE, MOOD, ASPECT, PRONTYPE and VERBFORM

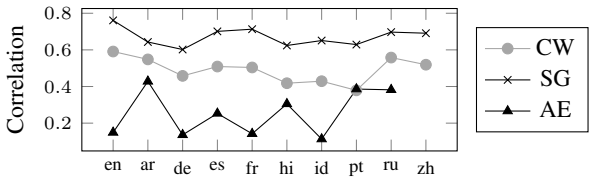


Figure 8: Model comparison on EMOTION.

From Figure 6, 7, we can see that all the three models give quite perfect performance on decoding person, definiteness, tense, voice, mood, aspect, pronoun type and verb form.

Overall, character-based representation is most effective for Slavic languages on decoding verbal morphological features but not nominal features. The result is vice versa for Romance languages, which is not as morphologically complex as Slavic. It is worth noticing that models behave differently on Bulgarian, an analytic language, although Bulgarian belongs to Slavic language from the phylogenetic perspective. We think that this is because many morphological features in Bulgarian have been simplified or weakened.

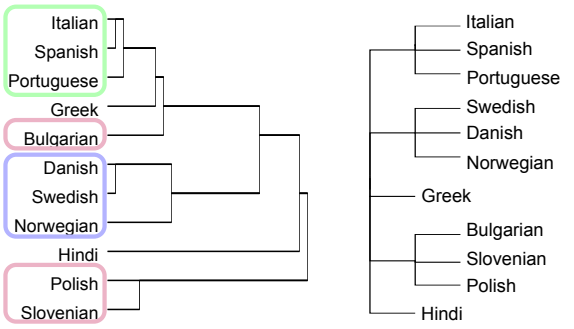
4.4 Emotion Score

In Experiment IV, we use the manually annotated data collected by Dodds et al. (2015). The data contains emotion scores for a list of words in several languages. In our experiment, the original score scale is transformed into the interval $[0, 1]$. A nonlinear map is trained to regress the representation of a word (CW, SG, AE) to its emotion score.

To evaluate the predicted results, we measure the Spearman correlation between the gold scores and predicted scores. The result in Figure 8 reveals a significantly strong correlation between the predicted emotion scores of SG and the real emotion scores. CW comes the second. For AE, it is hard to decode emotion just from the word form.

5 Contrastive Analysis

As we have mentioned before, Type I C&W model utilizes ordered context information to train the distributed word representation. Type II skip-gram model utilizes unordered context information. Type III character-based LSTM autoencoder model utilizes the grapheme information to represent a word. Towards the key questions that we raised at the very beginning of the paper, we propose our contrastive analysis based on the experiment results.



(a) Hierarchical tree based on model performances

(b) WALS Genus Tree

Figure 9: Comparison of the tree based on model performances and the WALS dendrogram manually constructed by linguists.

5.1 Typology vs. Phylogeny

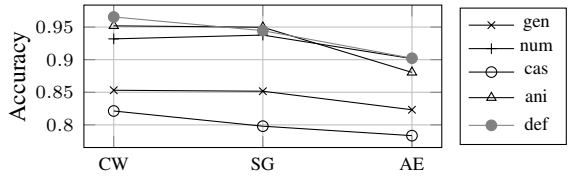
Experiment results have shown that word embedding models are influenced by the syntactic and morphological diversity more or less. Here we display how typological similarity and phylogenetic relation is revealed from the observed model performance variation. We hierarchically cluster languages according to the model performance on decoding syntactic and morphological features. The dendrogram of the languages in Figure 9 vividly shows that most of the phylogenetic-related languages are clustered together.

However, there is some interesting exceptions. Bulgarian does not form a primary cluster with other Slavic languages (e.g. Slovenian). We think that this is because Bulgarian is typologically dissimilar to Slavic language family. Therefore, Figure 9 reflects that language typology explains the model variation better than language phylogeny.

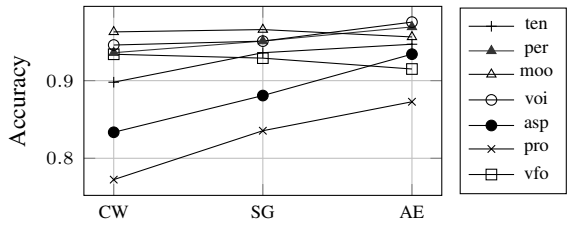
5.2 Form vs. Context

Here we discuss the effectiveness of word form and different types of word context.

Regarding the correlation between context type and language function, previous results show that SG performs worse than CW on decoding POS and dependency relation while SG performs better than CW on decoding emotion score. Since CW keeps word order of the context, this comparison suggests that word order information is vital to syntactic information, but it might also be a kind of noise for the word vectors to encode semantic information.



(a) Nominal features



(b) Verbal features

Figure 10: Overall performances of different models (averaged over languages) on decoding morphological features.

Lan.	Raw	Shuf.	Lan.	Raw	Shuf.
Russian	0.906	0.671	Slovenian	0.800	0.653

Table 4: Comparison of original and shuffled character-based word representation on decoding POS tag.

Source Language	**Arabic**		**Finnish**		
Target Language	fa	ud	en	shuf en	rand
Bigram type overlap.	0.176	0.761	0.891	0.864	0.648
Bigram token overlap.	0.689	0.881	0.999	0.993	0.650
Trigram type overlap.	0.523	0.522	0.665	0.449	0.078
Trigram token overlap.	0.526	0.585	0.978	0.796	0.078
Reconstruction Acc.	0.586	0.689	0.95	0.83	0.22

Table 5: Comparison of morpho-phonological knowledge transfer on different language pairs. The reconstruction accuracy is correlated with the overlapping proportion of grapheme patterns between source language and target language.

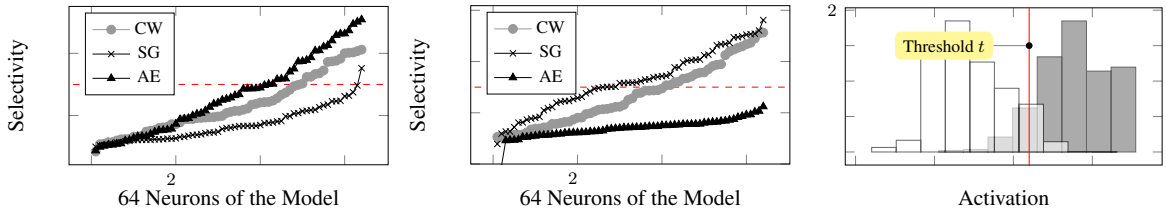
Regarding the correlation between form and language function, previous results on POS, dependency relation and emotion scores show the effectiveness of the word context. However, for morphological features, results in Table 10 indicate that context-based word representation works slightly better than character-based representation. Specifically, character-based embedding (AE) does outperform context-based embedding (CW, SG) on decoding verbal morphological features, even though AE does not access any context information. In other words, word form could be an explicit and discriminative cue for the model to decode the morphological feature of a word.

To prove that word form could provides informative and explicit cues for grammatical functions, we train another shuffled character-based word representation, which means that the autoencoder inputs shuffled letters and outputs the shuffled letters again. We use the hidden layer of the shuffled autoencoder as the representation for each word. The result in Table 4 shows that now the character-based model cannot perform as well as the original character-based autoencoder representation does, which again proves that the order of the word form is necessary for learning the grammatical function of a word.

Since many languages share similar phonographic writing systems, we naturally want to know whether the grapheme-phoneme knowledge from one language can be transferred to another language. We train an autoencoder purely on Finnish and directly test the trained model on memorizing raw English words, letter-shuffled English words and random letter sequences. Results in Table 5 indicate that the character autoencoder can successfully reconstruct raw English words instead of the letter-shuffled English words or random letter sequences. However, if we train an autoencoder purely on Arabic and then directly test the trained model on memorizing Urdu (ud) words or Persian (fa) words, the reconstruction accuracy is quite low, although Arabic, Persian and Urdu use the same Arabic writing system.

To explain the behaviour of AE, we calculate the correlation between the bigram character frequency in the words of the training language (e.g. Finnish) and the bigram character frequency in the words of the testing language (e.g. English). Table 5 reveals that phonological knowledge can be transferred if two languages share similar bigram and trigram character frequency distribution. For example, Finnish and English are both Indo-European language. Their writing system stores similar phonological structure. Arabic is a Semitic language. Persian is an Indo-European language. Their writing system stores different phonological structures respectively. This again proves that character-based LSTM autoencoder does 'memorize' the grapheme or phoneme clusters of a words. Morpho-phonological knowledge can be transferred among typologically-related languages.

Additionally, we are surprised to find that using the English word representations encoded



(a) me-- kan (b) Country names (c) Single neuron

Figure 11: Visualising the Neuron activation pattern for different word embedding models

by AE model trained on Finnish can increase the accuracy of English AE embedding in Experiment I (up to 0.7076), compared with the original accuracy 0.6587. This is probably due to the shared knowledge about the morphemes in the word form.

5.3 Neuronal Activation Pattern

Le (2011) found out that it is empirically possible to learn a 'Grandmother neuron'-like face detector from unlabelled data. In their experiment on unlabeled data, one neuron learns to activate specifically towards the pictures with cat faces instead of other pictures. Based on this finding, we hypothesize that there should exist selective neuron activation towards a linguistic feature trigger. The feature trigger can be a special consonant cluster, a specific suffix or the syntactic category of a word.

To quantitatively show the collective neuron behaviours and the individual neuron response towards different linguistic trigger, we compute the maximum probability that a neuron discriminates the words with trigger f from the words without trigger f. We defined this probability as the Degree of Selectivity p. For a given neuron $\mathbf{n}$ in a given model $\mathbf{M}$ towards linguistic trigger f, we try to find a threshold t that maximizes $p_{f,t}$,

$$c_{f,t} = \frac{N^+_{f,t}}{N_f}, c_{\neg f,t} = \frac{N^+_{\neg f,t}}{N_{\neg f}},$$

$$Selectivity = p_{f,t} = \frac{2 \times c_{f,t} \times c_{\neg f,t}}{c_{f,t} + c_{\neg f,t}}$$

where $N^+_{f,t}$ is the number of correctly discriminated words with linguistic feature f based on the threshold t. N_f is the real number of words with linguistic feature f. $N^+_{\neg f,t}$ is the number of correctly discriminated words without linguistic feature f based on the threshold t. $N_{\neg f}$ means the real number of words without linguistic feature f. $c_{f,t}$ / $c_{\neg f,t}$ is the accuracy for the neuron $\mathbf{n}$ of model $\mathbf{M}$ to detect the existence / nonexistence of the linguistic feature f. $p_{f,t}$ is the F-score of $c_{f,t}$ and $c_{\neg f,t}$, indicating the degree to which a certain neuron discriminates the words with/without a certain trigger f at a certain threshold t.

After calculating the selectivity of 64 neurons in an embedding model towards a linguistic trigger f, we sort the neurons according to the value of selectivity and draw the curve in Figure 11 for each model. The x-axis is the rank of the model neurons based on their selectivity towards a certain linguistic trigger. The y-axis is the selectivity of the corresponding neuron. The curve can tell us how many neurons selectively respond to trigger f to a certain degree. For example, we can see from Figure 11 that the max selectivity of the AE neurons reaches nearly 0.9. This means that one neuron of the AE model is especially sensitive to the prefix *'Me-'* and affix *'-an'*. It can detect the words with the prefix *'Me-'* and the affix *'-an'* just from its activation pattern.

It is also interesting to see from Figure 11 that neurons of AE respond more selectively to morphological triggers than those of the word-based model. For example, almost 30% of the AE neurons fall in the selectivity level [0.7, 1] towards the verb marker, namely prefix *'Me-'* and affix *'-an'*, in Indonesian. Context-based model also shows some selectivity towards this morphological triggers. For SG model, the max selectivity of the model neurons is only just above 0.7.

On the contrary, the context-based distributed models showed strong selective activation towards country names in Indonesian. However, the selectivity of all the AE neurons is below 0.7 towards these semantically-related words.

Similar patterns are found also in other languages. We conclude that the character-based model captures much morphological information

/ syntactic marker than semantic information. The popular word-based model captures both semantic information and syntactic information, although the latter is not displayed as explicitly as the former.

6 Related works

There have been a lot of research on interpreting or relating word embedding with linguistic features. Yogatama et al. (2014) projects word embedding into a sparse vector. They found some linguistically interpretable dimensions. Faruqui and Dyer (2015) use linguistic features to build word vector. Their results show that these representation of word meaning can also achieve good performance in the analogy and similarity tasks. These work can be regarded as the foreshadowing of our experiment paradigm that mapping dense vector to a sparse linguistic property space.

Besides, a lot of study focus on empirical comparison of different word embedding model. Melamud et al. (2016) investigates the influence of context type and vector dimension on word embedding. Their main finding is that concatenating two different types of embeddings can still improve performance even if the utility of dimensionality has run out. Andreas and Klein (2014) assess the potential syntactic information encoded in word embeddings by directly apply word embeddings to parser and they concluded that embeddings add redundant information to what the conventional parser has already extracted. Tsvetkov et al. (2015) propose a method to evaluate word embeddings through the alignment of distributional vectors and linguistic word vectors. However, the method still lacks a direct and comprehensive investigation of the utility of form, context and language typological diversity. This is exactly our novelty and contribution.

It is worth noticing that Köhn (2015) evaluates multilingual word embedding and compares skip-gram, language model and other competitive embedding models. They show that dependency-based skip-gram embedding is effective, even at low dimension. Although Köhn (2015) work involves different languages, they focus on the similarity among multilingual embeddings with only 7 languages. Our work, however, not only provides a comprehensive investigation with massive language samples (30 for Experiment I) and nonlinear mapping models, but also reveal the utility of pure word form and novelly point out the cross-language differences in word representation, which have been overlooked by huge amount of monolingual/bilingual research on well-studied languages.

7 Conclusion

In this paper, we quantify the utility of word form and the effect of language typological diversity in learning word representations. Cross-language perspective and novel analysis of neuron behaviours provide us with new evidence about the typological universal and specific revealed in word embedding. We summarize from our experiments on a massive set of languages that:

- Language typological diversity, especially the specific word order type and morphological complexity, does influence how linguistic information is encoded in word embedding.
- It is plausible (and sometimes even better) to decode grammatical function just from the word form, for certain inflectional languages.
- Quantification of neuron activation pattern reveals different characteristics of the context-based model and the character-based counterpart.

Therefore, we think that it is necessary to maximize both the utility of word form and the advantage of the context for a better word representation. It would also be a promising direction to incorporate the factor of language typological diversity when designing advanced word representation model for languages other than English.

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, 61473092, and 61472088), 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.

Jacob Andreas and Dan Klein. 2014. How much do word embeddings encode about syntax? In *Proceedings of ACL*, pages 822–827.

Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2015. Improved transition-based parsing by modeling characters instead of words with LSTMs. In *Proceedings of EMNLP*.

Emily M Bender. 2009. Linguistically naïve!= language independent: why nlp needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32. Association for Computational Linguistics.

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.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology. `http://wals.info/`.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of ACL*.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of NAACL*.

Zellig S. Harris. 1954. Distributional structure. *Synthese Language Library*, 10:146–162.

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*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Proceedings of AAAI*.

Arne Köhn. 2015. Whats in an embedding? analyzing word embeddings through multilingual evaluation. In *Poceedings of EMNLP*.

Siwei Lai, Kang Liu, Liheng Xu, and Jun Zhao. 2015. How to generate a good word embedding? *arXiv preprint arXiv:1507.05523*.

Q. V. Le. 2011. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595 – 8598.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015a. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of EMNLP*.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015b. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*.

Nianwen Xue, Zixin Jiang, Xiuhong Zhong, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. 2010. Chinese treebank 7.0. *Linguistic Data Consortium, Philadelphia*.

Dani Yogatama, Manaal Faruqui, Chris Dyer, and Noah A Smith. 2014. Learning word representations with hierarchical sparse coding. In *Proceedings of ICML*.