Distributed Representation of Chinese Collocation

Bo Xia
Beijing Language and Culture University
blcuxiabo@126.com

Endong Xun
Beijing Language and Culture University
edxun@blcu.edu.cn

Abstract

We provide computational linguistics an unsupervised algorithm for *Collocation Representing* in vector space using segmented corpora. By collocation, we mean a directed association of head word and dependency word that constructed by fully random combine in a sentence. Our algorithm represents dependency words by dense vectors that are trained to predict contexts of the head word. We show that these vectors provide high performance for extracting collocation similarities in syntactic and semantic. The quality of work results performs effective on our test set, which is measured in a verb-verb phrase collocation prediction task.

1 Introduction

"You shall know a word by the company it keeps" (Firth, 1957). As an important role of language phenomenon, lexical collocations is still far from work in nature language processing such as parsing, word sense disambiguation, and topic modeling (Carlos Ramisch, 2013).

For example, a parser that lacks sufficient knowledge of verb-particle constructions might correctly assign *look up the tower* two interpretations ("glance up at the tower" vs. "consult a reference book about the tower"), but fail to treat the subtly different *look the tower up* as unambiguous ("consult a reference book" interpretation only) (Ivan A. Sag, 2002).

Composed of a head word and a dependency one, collocations are restricted lexical co-occurrences of two syntactically bound lexical elements (Kilgarriff, 2006).A large body of work, known as Multiword expressions (MWEs) made up of at least 2 words, has studied the expand phenomena of collocation which are sequence of words that acts as a single unit at some level of linguistic analysis (Calzolari et al., 2002).

Despite popularity of collocations and various labels, it is a limitation of collocation representation that many of collocations present an unusual structure (Iñaki Alegria, 2004), especially for languages with large vocabularies and many rare words. On the other hand, disambiguation of collocation in context is frequent in corpora (e.g., *bus stop*, as in *Does the bus stop here?* vs. *The bus stop is here.*) (Miriam R. L. Petruck, 2015). Since both of them are hard for structured collocation representation, the representation learning of collocations become our inspiration in the recent work of machine learning community.

In this paper, we study in depth one collocation phenomena: verb-verb phrase (In English, it can be compared to phenomena of infinitives and gerund as verb implement). We focus on distributed representations of collocations learned by neural networks, as it trained on corpora and huge amounts of Verb-verb phrase recognition with sparsity. We develop new model architectures that preserve the collocation regularities given contexts. We present collocations representation similarity goes beyond simple syntactic regularities. For measuring both syntactic and semantic regularities, we present experiments of predicting Verb-verb phrase recognition in test set that labeled by students of Applied Linguistics, and show that linguistic regularities can be learned with considerable accuracy. Moreover, we discuss how training time and accuracy depends on the amount of the corpora and collocations.

The remaining sections are organized as follows. Section 2 briefly discusses some previous works on methods for collocation representation. Section 3 describes the model architectures combined with ontology classes. Section 4 is devoted to the description and evaluation of the experiment by

means of c++ programming. Section 5 based on comprehensive comparisons among the three kinds of syntactic and semantic analysis. And, finally, section 6 outlines some conclusions and proposals for future work.

2 Related work

The major linguistic characteristics of collocation representation are: feature-based description of collocation and collections of real-world occurrences (Brigitte Krenn, 2000), including two aspects: (1) their composition, i.e. which the components of the MWLU are; and (2), what we call the surface realization, i.e. the order in which the components may occur in the text (Iñaki Alegria, 2004).

In morphosyntactic criteria, collocations are represented as instances in the lexical database using rules or labels to build up syntactic and semantic and structures. Collocations can be defined as associations of two lexical units in a specific syntactic relation (for instance adjective - noun, verb - noun (object), etc.). Recursively, a lexical unit can be a word or a collocation (Vasiliki-iFoufi, 2017). More precisely, a typology of verbal and nominal collocations has been defined in the encoding of Modern Greek (MG). This typology has been built on rich linguistic knowledge such as fixedness and morphosyntactic features (Aggeliki Fotopoulou, 2014).

Due to rich information and features of linguistic collocation representation, distributed vector representations of collocations allow us to mine collocation features statistically and consider word combinations as a unit during parsing if necessary.

3 Model Architectures

The main observation from the previous section was that most of the complexity is caused by the massive linguistic features. While this is what makes linguistics so attractive, we decide to explore neural network models that might not be able to represent collocation as precisely as rules and labels, but can possibly be trained on much more data efficiently.

The model architecture for learning distributed representations of collocations in this study consists of three stages: target probability, architectures and parameter training, adapted from that of Mikolov et al. (2013) where it was found that neural network language model can be successfully trained on top of continuous vectors.

3.1 Collocation prediction target

Because each <HEAD, DEP> acts as a single unit in linguistic analysis, the task is to predict a collocation include head word and dependency word given the other words in a context as follows:

$$P(<HEAD, DEP> \mid Context) \quad (1)$$

So despite the fact that the <HEAD, DEP> collocations vectors are initialized randomly, <HEAD, DEP> can eventually capture syntactic and semantic features as a production of the target probability task.

However, by mapping each collocation to distinct vector representation, the performance must get worse as millions of VP collocation data are available.

3.2 Continuous Skip-gram Model

In this section, we consider each DEP word representation as a bag of HEAD word contexts, which means DEP word embedding takes collocation features into account by predicting contexts of HEAD word. Its construction gives our algorithm the potential to overcome the weaknesses of popularity mentioned above.

More precisely, we use each DEP word as an input to a log-linear classifier with continuous projection layer, and predict words within a certain range before and after the HEAD word, computed as:

$$P(Context(HEAD) \mid DEP) \quad (2)$$

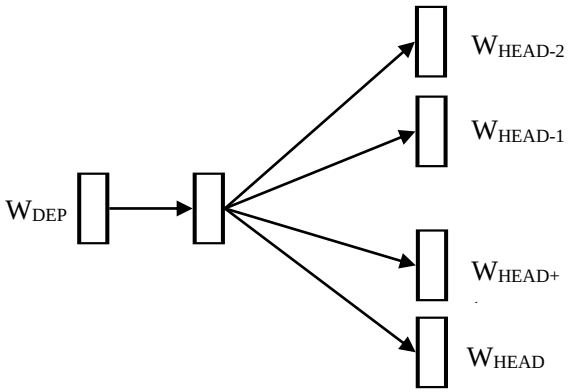

Figure 1. Continuous Skip-gram architecture. The DEP word predicts surrounding words given the HEAD word of current collocation.

At prediction time, using pretreatment collocation list for identification instead of fully random combine is an optional step, which focus on pretreatment collocation training. In order to bound the speed of character-level match, we use bit bool array memory the existence of mapping from the word to a collocation. Moreover, the error caused by online brute force solution are resolved by training the vector under large corpora.

3.3 Ontology modified hierarchical softmax

Popular models that learn representations ignore the morphology feature of words (Tomas Mikolov, 2017) and sematic feature of word ontology with hierarchal softmax.

In our work, the structure of the hierarchal softmax is a binary tree consisted of Huffman part and ontology part, where ontology comes from HowNet (Zhengdong Dong, 2000). Method of encoding Huffman code and ontology code for the hierarchy is the same with Mikolov et al. (2013c).

We modified original hierarchical softmax to two part: ontology part built by binarization hierarchy of HowNet taxonomy entity, while the other word part instead of ontology is organized in forms of Huffman tree, pictured as follows:

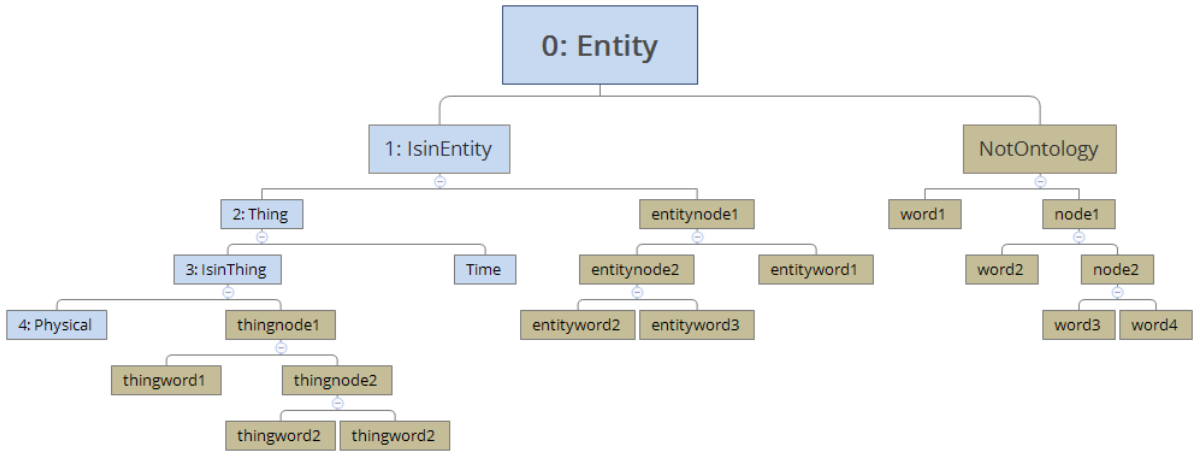

Figure 2. Ontology hierarchy.

More formally, given a large training corpus represented as a sequence of words w_1, ..., w_t, the objective of the skip-gram model is to maximize the following log-likelihood:

$$\sum_{C_t \in C} \sum_{\theta \in PathOf(DEP)} logP\left(d_\theta \left| V(C_t), I(\theta)\right.\right) \quad (3)$$

where the context Ct is the set of indices of words surrounding HEAD word. θ is a non-leaf-node index in path of Huffman code, d_θ is a Boolean answer of the binary decision. In particular, we have matrix V and I, corresponding, respectively, to vectors of leaf-node and non-leaf-node.

In conclusion, the probability of observing a context word Ct given <HEAD, DEP> will be parameterized using the aforementioned vectors and model. Vectors of leaf-node[HEAD] represent convergent meaning of all collocations start with HEAD word. Given <HEAD, DEP> and contexts, collocation intensity expressed as following:

$$\sum_{C_t \in C} \sum_{\theta \in PathOf(DEP)} P\left(d_\theta \left| V(C_t), I(\theta)\right.\right) \quad (4)$$

4 Experimental setup

The corpus used is segmentation version of the people's daily corpora consisting of news and serial novels, which contains 100,000,000 sentence, with 3,377,634,987 Chinese tokens of roughly 100,906 words.

Ontology include 305 classes, maximum depth of hierarchy is 8 which named "{system|制度}", maximum width of hierarchy 14 which named "{Natural Thing|天然物}".

There are 420,828 *verb-verb phrase collocations* of 828 headword verbs in the pretreatment list in total, performing acceptable, extracted by students major in linguistic instead of employing a dependency parser, using hundreds of linguistic rules considering part of speech, syllable length,

border and punctuation feature, specific adverbs and word classification of the Grammatical Knowledge-base of Contemporary Chinese (YU Shi-wen, 1996).

Averaged sentence length of *test set* is 14.79 measured by word. Test sentences are randomly taken from the people's daily corpora, labeled *arg0*, *arg1*, *modify*, *parallel*, *clause*, *verb-relation* and *belonging* relations by 7 students major in linguistic.

1. Initialization of hierarchical softmax with ontology
2. Read pretreatment collocation data
3. Read corpora
4. Online identification of <HEAD, DEP>
5. Predicting Contexts of HEAD by DEP
6. Updating leaf node matrix V and inter node

Figure 3. Training algorithm.

For u in Context(HEAD){
  e = 0
  For path in Tree(DEP){
    q = $\sigma(leafNode(u)^T InterNode(path))$
    g = $\eta(1 - path.direction\text{-}q)$
    e += $g^* InterNode(path)$
    $InterNode(path)$ += $InterNode(path) + g^* leafNode(u)$
  }
  $leafNode(u)$ += $leafNode(u) + e$
}

Figure 4. Updating embedding algorithm.

5 Result analysis

After the training converges, words with similar collocation meaning maps to a similar position in the vector space. All tasks use the same dictionary of 100,906 words in corpora with minimal frequency boundary of five. Comparison from involving collocation-pretreatment or not is evaluated in our model.

It takes about 4 days to train on one computer. Embedding file is 300MB without collocation-pretreatment and 130MB within collocation-pretreatment, trained from same corpora of 16.9GB.

5.1 Nearest neighbors of collocation using our representations

We first evaluate the quality of our representations in 100 dimensions on the task of collocation similarity or relatedness, by listing its 10 nearest neighbors arbitrary using the Euclidean metric of matrix LeafNode. The performance of embedding obtained in the collocation lookup-table was extremely good, both for two comparison tasks, as shown in Table 1:

主张	主张	后空翻	后空翻
唯名论	诉诸武力	孟加拉虎	接特卡切夫
黑名	一中一台	白虎	直体
自决权	国统	国母	屈体
自决	国统会	聽	王克楠
一中一台	分治	明矾	黄华东
以强凌弱	台独	国父	21 13
中间线	代表权	饰物	9.725
兼爱	强权政治	單	前空翻
强加于人	联合公报	然否	段青
当事国	罔顾	天打雷劈	范红斌

Table 1: Comparisons of collocation neighbors with first columns for using pretreated collocation, second columns for not.

We observe that both for using or not collocation pretreatment, our collocation neighbors are reasonable, as it provides satisfactory performance across part-of-speech. Pretreated collocation experiments of "后空翻" performed worse than no pretreatment involved informed us that arbitrary online collocation identification coverage is not as good as collocation extraction stage.

5.2 Examples of the collocation possibility

In this section, we describe an evaluation of the collocation relationship obtained with our neural network language model.

For an overview of collocation relationship possibility given context, we evaluate arbitrary collocation <HEAD, DEP> possibility of two words in a sentence among $sentence\_length^2$ times, using following computational formula:

$$\sum_{C_{t1} \in S} \sum_{C_{t2} \in S} logP(< C_{t1},\ C_{t2} > \mid S) \quad (5)$$

$P(< C_{t1}, C_{t2} > |Sentence)$ is same as training formula.

For each pair, the higher the probability that they are collocated, the tighter their relationship is. The Illustration of collocation relationship is shown as following:

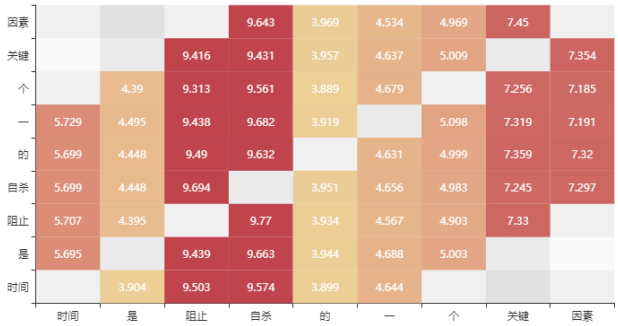

Figure 5. “时间 是 阻止 自杀 的 一 个 关键 因素”(Time is a key factor in preventing suicide.)

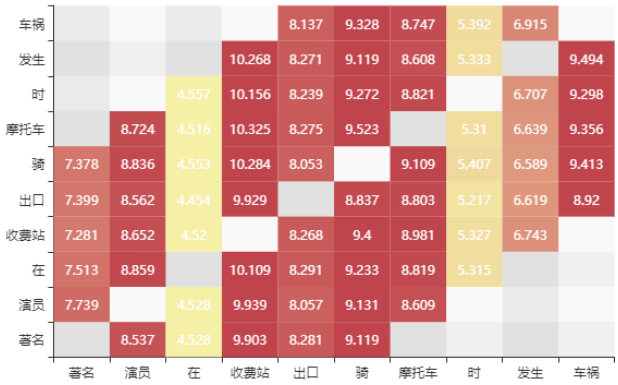

Figure 6. “著名 演员 在 收费站 出口 骑 摩托车 时 发生 车祸”(The famous actor had an accident when he was riding a motorcycle at the toll gate exit.)

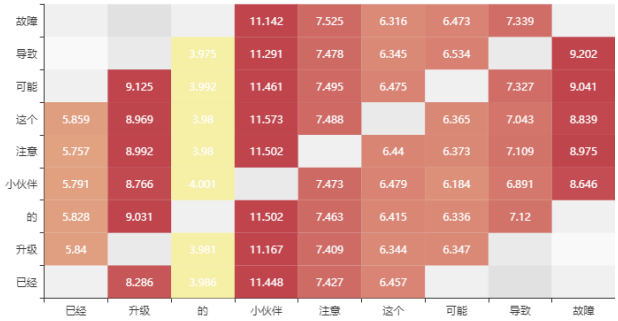

Figure 7. “已经 升级 的 小伙伴 注意 这个 可能 导致 故障” (People who have upgraded note that this may cause the machine to breakdown)

Moreover, the semantics and the semantic compositionality of a collocation full integration into our models, for example, such features are assigned to “著名 演员”(the famous actor), “收费站 出口 骑 摩托车”(riding a motorcycle at the toll gate exit).

5.3 Verb-verb phrase recognition

We further denote verb-verb phrase recognition as a question, which the goal is to select a DEP word that is the most coherent with the given HEAD in rest of the sentence. We start from the requirement that DEP word must be verb in vocabulary, the relationship between words can be captured in a window (window-size = 5), using *arg1* relation in sentences as *test set.*

For each sentence, we show the comparison of whether the collocation-pretreatment are used in model training, maximal verbs in the sentence, and influence of collocation directional feature. The results are presented in Table2.

As expected, the DEP word recognition becomes harder with larger verb amount in sentence. Compared to collocation-pretreatment involved, training alone performed as optimal choice. Influence of collocation directional feature depends on using of collocation-pretreatment.

We observe that the optimal choice of collocation model comes with keeping directional feature and collocation-pretreatment away, which shows that our architecture indeed represents linguistic regularities of a <HEAD, DEP>.

6 Conclusion and future work

In this paper, we described an unsupervised learning algorithm to learn <HEAD, DEP> *collocation representations* by taking ontology hierarchy into account. It indicates an idea of *restructure corpora* for unsupervised algorithm, which the DEP vectors are learned to predict surrounding words of HEAD.

Our experiments on collocation distance extraction and verb-verb phrase recognition show that the method is competitive. It can be expected that several linguistic analysis applications such as MWEs identification, target-polarity word extraction, and SMT can benefit from the model architectures described in this paper.

collocation-pretreatment	maximal verbs	if bi-directional	nCorrect	nTotal	Accuracy
TRUE	2	TRUE	129	161	**0.80124**
TRUE	3	TRUE	78	109	0.71559
TRUE	4	TRUE	38	59	0.64406
TRUE	2	FALSE	127	161	0.78881
TRUE	3	FALSE	76	109	0.69724
TRUE	4	FALSE	36	59	0.61016
FALSE	2	TRUE	143	185	0.77297
FALSE	3	TRUE	89	127	0.70078
FALSE	4	TRUE	42	70	0.60000
FALSE	2	FALSE	156	185	**0.84324**
FALSE	3	FALSE	103	127	0.81102
FALSE	4	FALSE	54	70	0.77142

Table 2: Illustration of Verb-verb phrase recognition. Accuracy means DEP word recognition accuracy.

In the future, it would be interesting to extend our techniques from second order to multiple order relations. Note that by partitioning sentence based on boundary model can disambiguate fully random combine. We will open source the implementation of our model, in order to facilitate comparison of future work on collocation representations.

References

Ramisch, C., Villavicencio, A., and Kordoni, V. 2013. Introduction to the special issue on multiword expressions: From theory to practice and use. ACM.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Dan, F. 2001. Multiword Expressions: A Pain in the Neck for NLP. International Conference on Computational Linguistics and Intelligent Text Processing Springer-Verlag.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., and Macleod, C., et al. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. International Conference on Language Resources & Evaluation.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents.

Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. 2004. Representation and treatment of multiword expressions in Basque. The Workshop on Multiword Expressions: Integrating Processing Association for Computational Linguistics.

Mikolov T, Chen K, Corrado G, et al. 2013. Efficient Estimation of Word Representations in Vector Space. Computer Science.

Benson M. 1989. The Structure of the Collocational Dictionary. International Journal of Lexicography.

俞士汶, 朱学锋, 王惠,等. 1996. 现代汉语语法信息词典规格说明书[J]. 中文信息学报.

Krenn B. 2000. Collocation Mining: Exploiting Corpora for Collocation, Identification and Representation. VDE-Verlag GmbH.

Caseli H D M, Villavicencio A, Machado A, et al. 2009. Statistically-driven alignment-based multiword expression identification for technical domains.

Zarrieß S, Kuhn J. 2010. Exploiting Translational Correspondences for Pattern-Independent MWE Identification.

Wakaki H, Fujii H, and Suzuki M, et al. 2009. Abbreviation generation for Japanese multi-word expressions.

Zhengdong Dong, Qiang Dong. 2000. Introduction to hownet.

Kordoni, Valia. Beyond Words: Deep Learning for Multiword Expressions and Collocations. Proceedings of ACL 2017.