Bo Xia

Session 5B (room TU101): Semantics

2018年11月23日

Distributed Representation of Chinese Collocation

Let’s begin.

Hello, ladies and gentlemen. I’d like to start by introducing myself, my name is xiabo, studying for a master's degree at Beijing Language and Culture University

I plan to speak about collocation representation, what I would like to do today is to illustrate an unsupervised algorithm for Collocation Representing in vector space using segmented corpora.

In the first part I will introduce collocation and collocation representation.

In the next section I give a basic definitions and formula of collocation representation.

In part three, I am going to show experiment detail and result.

In the last part I would like to give some practical examples.

There will be plenty of time at the end of my speech for a discussion

Now let us turn to point one, collocation.

“You shall know a word by the company it keeps”

As an important role of language phenomenon, lexical collocations is still far from work in nature language processing such as word sense disambiguation, parsing and topic modeling.

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)

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.

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.

On the other hand, If there is one thing I’d like to get across to you today it is that sparse data is killing ML, which is the bottleneck in development of NLP application. For example, Sparse data can be low frequency of n-grams, or lack of sentence structure, or border information of MWEs or collocations.

By collocation, we mean a directed association of head word and dependency word that constructed by fully random combine in a sentence.

Generate period

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).

Recognition period

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)

To overcome words sparsity, distributed representation is used to …, to overcome collocation sparsity, emmmm.

Why we have to represent collocation?

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.

We provide computational linguistics an unsupervised algorithm for Collocation Representing

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.

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 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.

“You shall know a word by the company it keeps” (Firth, 1957).

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.

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 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.

I have divided my presentation into Y parts

In the first part I give a few basic definitions.

I plan to speak about TITLE AND SUBJECT

What I would like to do today is

to explain

to illustrate

to give my background information on

to outline

to have a look at

What I want my listeners to get out of my speech is

If there is one thing I’d like to get across to you today it is that

I have divided my presentation into Y parts

In the first part I give a few basic definitions.

In the next section I will explain

In part three, I am going to show

In the last part I would like to give a practical example..

There will be plenty of time at the end of my speech for a discussion

Now let us turn to point one

Let us now move on to the second part, which is, as I said earlier

There are three things we have to consider: one, two, three

Now let us look at the first aspect which is

First of all,

Content

Quantity

Sequencing your ideas

Keeping the audience’s attention

Signposting where you are

That’s all I would like to say about subject of part A, and now let us turn to

Now that we’ve seen … let us turn to …

Now let’s take an example

An example of this can be found

To illustrate this

Let’s see this through an example

For example,

For instance,

Let me rephrase that,

In other words

Another way of saying the same thing is

That is to say

What is very significant is

What is important to remember

I’d like to emphasize the fact that

I’d like to stress the importance of

What I tried to bring out

What we need to focus on

To summarize

To sum up,

Let me summarize by saying

So that concludes my overview

In conclusion

Briefly said

In short,

What I’ve tried to show in this part

To recap what we’ve seen so far

As I already said earlier

As we saw in part one

To repeat what I’ve said already

We will see this a little later on

This will be the subject of part 3

We will go into more detail on that later

I quote the word of

In the word of

According to

Here I’d like to quote

As Mr. X says in his book

There is a famous quotation that goes

As you all may well know

It is generally accepted that

As you are probably aware of

I’d like to summarize/sum up

At this stage I would like to run through/over the main points

So, as we have seen today

As I have tried to explain this morning BT finds itself in

In conclusion I would like to say that

My final comments concern

I would like to finish by reminding everyone that

I’d be happy to answer any questions

If there are any questions please feel free to ask

Thank you very much for your attention and if there are any suggestions or comments

Let’s look at current distribution of the market, as you can see

I’m going to show you now the most recent figures available

My next slide concerns the method by which