INTERSPEECH 2009 BRIGHTON

Improved Language Modelling Using Bag of Word Pairs

Langzhou Chen, KK Chin and Kate Knill

Toshiba Research Europe Limited, Cambridge Research Lab,
208, Cambridge Science Park, Milton Road, Cambridge, UK
langzhou.chen@crl.toshiba.co.uk

Abstract

The bag-of-words (BoW) method has been used widely in language modelling and information retrieval. A document is expressed as a group of words disregarding the grammar and the order of word information. A typical BoW method is latent semantic analysis (LSA), which maps the words and documents onto the vectors in LSA space. In this paper, the concept of BoW is extended to Bag-of-Word Pairs (BoWP), which expresses the document as a group of word pairs. Using word pairs as a unit, the system can capture more complex semantic information than BoW. Under the LSA framework, the BoWP system is shown to improve both perplexity and word error rate (WER) compared to a BoW system.

Index Terms: language model, speech recognition, LSA

1. Introduction

The bag-of-words (BoW) method is a widely used data representation method. Basically, it represents an object as a collection of independent items disregarding the relation and structures among the items, so that the data objects are easier to process or classify. It has been used in natural language processing (NLP) and information retrieval where the data object is the set of text documents and the items are an unordered collection of words which occur in the document. For automatic speech recognition (ASR), the BoW method has been successfully used to improve n-gram language models using semantic information from the training text [1]. A typical BoW method for language modelling is the LSA language model [1]. Using this method, a word document co-occurrence matrix is constructed and a LSA feature space generated where each word in vocabulary and each document in training corpus are projected as a feature vector in LSA feature space. The association between words and documents can be expressed as the distance of the feature vectors relevant to them.

In this work, an extension of the BoW method is presented to represent more complex and detailed semantic information. In some cases, a bunch of individual words are not enough to represent the accurate semantic information of the text. For example, for a document about "September 11", neither "September" nor "11" contains the true information of the original document. If this document is expressed as these two individual words, many unrelated documents such as "Oceans Eleven" or "black September" will be assumed to be close to "September 11". On the other hand, if the word pair is chosen to express the text information (the word pair "September 11" is considered as a single unit) all of the ambiguity mentioned above can be avoided. Based on this idea, the original BoW method is extended to bag-of-word pairs (BoWP) where text is expressed as a group of word pairs. These word pairs are unlimited in order and position, i.e. any 2 words in the same document can be selected as a word pair. Through this, the more detailed semantic information which is ignored by the BoW method can be kept by BoWP.

In the BoW method, the concept of a word can also be extended. In [2], the co-occurrences between word n-tuple and documents were investigated. They were used to reduce the ambiguity of unconstrained command and control. The word n-tuple used in [2] is the agglomeration of n successive words, while the BoWP method presented in this work is not limited to 2 successive words. Any 2 words in a document which have a strong semantic association can be collected as a word pair. For example, the word pair of 2 football teams, e.g. "Chelsea, Liverpool" can be used to express a football game very well, however, they are do not necessarily occur successively in the document.

Since BoWP is an extension of BoW, it can be used in many areas in which BoW is used, such as statistical language modelling (SLM) and IR. In this work, a BoWP based LSA method is presented. Similar to the BoW based LSA, a co-occurrence matrix of the word pairs and documents is constructed and singular value decomposition (SVD) carried out to map each document and word pair to a vector in LSA space. In the LSA feature space of BoWP, data selection is carried out to select a domain dependent training corpus, e.g. tourism. Some additional training data which contains task specific information is used as the query and the domain dependent corpus is selected from a general corpus. Then, the domain dependent corpus is used to train the LM for use by a ASR system in the corresponding domain.

A problem of BoWP is the combinatorial explosion of the word pairs. The word pair-document matrix may be too big to be calculated. In this paper, the problem is solved by building the topic dependent word pair-document matrix. The task is divided into different topics. For each topic, the word pairs are extracted and the word pair-document matrices are built separately. Mixture models are then used to combine the different topics together.

The rest of the paper is organised as follows. Section 2 briefly introduces the work of BoW and LSA based LM. In section 3, the method of BoWP is presented. Section 4 presents a framework for selecting the domain dependent corpus in LSA feature space. Finally, experimental results and the conclusions are given.

2. Bag-of-words based LSA

LSA based language modelling has been well presented in [1]. It is a typical BoW method. Given a vocabulary with M words and a training corpus with N documents, a word document matrix $\boldsymbol{A}$ is constructed. Each document is associated with a column vector of dimension M and each

word is associated with a row vector of dimension *N*. Then SVD is carried out. Only keeping the *R* biggest singular values and associated singular vectors, the SVD of $\boldsymbol{A}$ can be expressed as:

$$\boldsymbol{A} \approx \widetilde{\boldsymbol{A}} = \boldsymbol{USV}^T \quad (1)$$

Where $\boldsymbol{S}$ is the diagonal matrix of singular values, $\boldsymbol{U}$ and $\boldsymbol{V}$ are the (M×R) left singular matrix and (N×R) right singular matrix respectively. $\widetilde{\boldsymbol{A}}$ is the *R*-rank best approximation of the original word document matrix $\boldsymbol{A}$.

In Eq. (1), the column vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$ define an orthonormal basis for the space of dimension *R* separately. Therefore, the word vector in $\boldsymbol{A}$ is projected onto the orthonormal basis of the column vector of matrix $\boldsymbol{V}$, meanwhile, the document vector in $\boldsymbol{A}$ is projected onto the orthonormal basis of the column vector of matrix $\boldsymbol{U}$. Based on this fact, each word in the word-document matrix can be expressed as a row vector of $\boldsymbol{US}$ and each document can be expressed as a row vector of $\boldsymbol{VS}$. The association between words and documents can be evaluated as the closeness of their feature vectors in the same LSA space. Given a method to map the distance of 2 vectors in LSA space to a probability value, semantic language modelling can be implemented in LSA feature space.

3. Bag-of-word pairs

In this work, the idea of BoW is extended to BoWP. A text is represented as an unordered collection of word pairs, disregarding the grammar and order information. The motivation of BoWP is exploring the semantic information that contains more complex word associations than a simple group of words.

3.1. Selection of word pairs

To construct the BoWP, the first problem that needs to be solved is selecting the word pairs. From the concept of BoWP, all the possible word pairs in the corpus should be selected. However, the number of possible word pairs is likely to be too large. Therefore the word pairs have to be pruned. To reduce the number of word pairs, the word pairs selected should each have a strong association together. For example, the average mutual information (AMI) has been proposed to select the trigger pairs in trigger model [3].

In this work, the word pairs are selected in the LSA framework. At first, the LSA feature space based on BoW is constructed. Then the word pairs are selected based on their distance in LSA space. Meanwhile, the words which are too general, e.g. {a, the} etc., are not allowed to appear in the word pairs. They are filtered out using the idf. Given a document *d*, the corresponding word pairs are selected as follows:

$$\begin{aligned} &WORDPAIR(d) = \\ &\{w_1, w_2 \mid w_1 \in d, w_2 \in d, idf(w_1) > \alpha, \\ &idf(w_2) > \alpha, \frac{\boldsymbol{u}(w_1)\boldsymbol{S}^2\boldsymbol{u}(w_2))}{\|\boldsymbol{u}(w_1)\boldsymbol{S}\| \cdot \|\boldsymbol{u}(w_2)\boldsymbol{S}\|} > \beta\} \end{aligned} \quad (2)$$

where α and β are empirical thresholds.

3.2. Word pair and document co-occurrence matrix

Similar to the BoW framework, the BoWP also starts from a co-occurrence matrix of the word pairs and the documents. Each column of the matrix represents a document and each row of the matrix is associated with a word pair, i.e.

$$\hat{\boldsymbol{A}} = \begin{bmatrix} a(w_1,w_1,d_1) & \cdots & a(w_1,w_1,d_n) & \cdots & a(w_1,w_1,d_N) \\ \vdots & \ddots & & & \\ a(w_i,w_j,d_1) & & a(w_i,w_j,d_n) & & a(w_i,w_j,d_N) \\ \vdots & & & \ddots & \\ a(w_M,w_M,d_1) & \cdots & a(w_M,w_M,d_n) & \cdots & a(w_M,w_M,d_N) \end{bmatrix} \quad (3)$$

In the word pair and document co-occurrence matrix, the cell $a(w_i, w_j, d_n)$ represents the normalised frequency of the word pair "$w_i w_j$" occurring in document *n*, i.e.

$$a(w_i, w_j, d_n) = (1 - \varepsilon(w_i, w_j)) \cdot \frac{C(w_i, w_j, d_n)}{\sum_{x,y \in d_n} C(x, y, d_n)} \quad (4)$$

Where $C(x, y, d_n)$ is the frequency that the word pair (*x, y*) occurred in document d_n, $\varepsilon(w_i, w_j)$ is the normalised entropy for (w_i,w_j), i.e.

$$\varepsilon(w_i, w_j) = -\frac{1}{\log N} \sum_{n=1}^{N} \frac{C(w_i, w_j, d_n)}{\sum_n C(w_i, w_j, d_n)} \log \frac{C(w_i, w_j, d_n)}{\sum_n C(w_i, w_j, d_n)} \quad (5)$$

Eq. (3) shows a matrix with all the possible word pairs, i.e. the number of rows in the word pair and document matrix, $\widetilde{\boldsymbol{A}}$, is *M*×*M*. However, after word pair pruning using Eq. (2), the number of word pairs will be much smaller.

3.3. LSA feature space for BoWP

Given the word pair and document co-occurrence matrix $\hat{\boldsymbol{A}}$, a SVD operation is carried out to generate the LSA feature space of BoWP, i.e.

$$\hat{\boldsymbol{A}} \approx \hat{\widetilde{\boldsymbol{A}}}_P = \hat{\boldsymbol{U}}\hat{\boldsymbol{S}}\hat{\boldsymbol{V}} \quad (6)$$

In Eq, 6 $\hat{\boldsymbol{S}}$ is a *R*×*R* diagonal matrix, of the *R* biggest singular values. $\hat{\boldsymbol{U}}$ and $\hat{\boldsymbol{V}}$ are the left and right singular vectors respectively.

In the LSA feature space of the BoWP, each row vector of $\hat{\boldsymbol{U}}\hat{\boldsymbol{S}}$ represents a word pair in LSA space, i.e.

$$\overline{\boldsymbol{u}}(w_i, w_j) = \hat{\boldsymbol{u}}_{row(w_i,w_j)}\hat{\boldsymbol{S}} \quad (7)$$

Where $row(w_i, w_j)$ represents the row number of word pair ($w_i w_j$) in the word pair document co-occurrence matrix. Meanwhile, each document is still represented as a column vector of $\hat{\boldsymbol{S}}\hat{\boldsymbol{V}}$.

A new document d_{N+1}, can also be represented as a feature vector in BoWP LSA space, i.e.

$$\boldsymbol{v}(d_{N+1}) = \sum_{w_i \in d_{N+1}, w_j \in d_{N+1}} C(w_i, w_j, d_{N+1}) \cdot \hat{\boldsymbol{u}}_{row(w_i,w_j)} \quad (8)$$

4. Data selection based on LSA framework

In this work, the LM training data selection is carried out in LSA feature space. Basically using some in-domain training corpus as the query, a set of domain dependent training data is selected from a big general corpus based on the semantic information contained in the text. Then this selected domain dependent data is used to train a domain dependent LM for ASR. Since this framework can be used for both BoW and BoWP, the quality of the selected corpus can be used to

measure the semantic information that were captured by the different methods.

Given a new document d_{N+1} which contains task specific information, a subset of training documents $D(d_{N+1})$ which have similar semantic information can be extracted in LSA feature space. The subset of the documents $D(d_{N+1})$ can be generated by collecting all the documents that are close to $\boldsymbol{v}(d_{N+1})$ in LSA feature space, i.e.

$$D(d_{N+1}) = \{d_i \mid \cos(\boldsymbol{v}(d_{N+1}), \boldsymbol{v}_i \boldsymbol{S}) > \theta\} \quad (9)$$

Where $\boldsymbol{v}(d_{N+1})$ is the feature vector of d_{N+1} in LSA space and θ is an empirical threshold.

Using the selected corpus, the n-gram LM can be updated using marginal adaptation to generate the LM with the semantic information of the current task, i.e.:

$$P(w_i \mid w_{i-n+1}^{i-1}, d_{N+1}) = \frac{P(w_i \mid w_{i-n+1}^{i-1}) \frac{P(w_i \mid D(d_{N+1}))}{P(w_i)}}{\sum_w P(w \mid w_{i-n+1}^{i-1}) \frac{P(w \mid D(d_{N+1}))}{P(w)}} \quad (10)$$

5. Topic dependent BoWP

A major problem of BoWP is that the number of possible word pairs in the text is much bigger than the number of words. Because of the calculation cost, the BoWP may be too big to be used in a practical system. Although the number of word pairs can be pruned based on Eq. 2, heavy pruning of the word pairs will lead to significant loss of semantic information.

In this work, the topic dependent BoWP is developed to solve the problem of the combinatorial explosion of the word pairs.

The in-domain data can be divided into many small topics. For example, for a tourism task various topics can be defined within the tourism domain, including booking tickets, sightseeing, hotel, restaurant, medicine etc.

For each topic, all the possible word pairs which occur in the in-domain data of this topic and satisfy Eq, 2 are extracted. These topic dependent word pairs can be used to build the word pair and document co-occurrence matrix as Eq. 3. Then, the LSA feature spaces of BoWP can be constructed for each topic. In the framework of this work, each topic has its own BoWP LSA feature space, and data selection is also carried out for each topic individually based on Eq. 9. Then mixture language models are adopted to combine the selected corpora of different topics through linear interpolation.

Since the number of word pairs for a particular topic is much smaller than the number of general word pairs, this method can solve the problem of combinatorial explosion of the word pairs without losing too much semantic information.

6. Experimental results

These experiments are based on a US English LVCSR system designed for the tourism domain which was divided into 86 topics manually. For each topic, the LSA feature space for BoWP was constructed individually as described in section 5.

The English Gigaword corpus [4] is used as the general LM training corpus. It contains about 1.86G words. An internal corpus was used for the in-domain LM training corpus which contains about 90k sentences. The in-domain corpora are used as the query to select the domain dependent training data from the general corpus.

The LVCSR acoustic model was a continuous HMM with 4.3k tied states and 14 Gaussians per state. A 33 dimensional feature vector with 10 MFCC cepstral features, log energy, and their first-order and second-order derivatives was used. The LVCSR LM vocabulary contained about 34k words.A trigram 1-pass viterbi decoding was used to obtain the best hypothesis.

In the first experiment, the comparison results of data selection based on BoW and BoWP are given. Perplexity is used to evaluate the performance of the semantic data selection. Two topics were investigated: restaurant and greeting. For each topic, the in-domain data of each topic was used as the query to select a topic dependent corpus. A trigram LM was then trained on the selected topic dependent corpus. The generated topic dependent LM was used to evaluate the perplexity of the query. A lower perplexity is expected to correspond to a better performance in data selection. Table 1 presents the perplexity results of this semantic data selection.

Topic	Size of query	Data selection method	Perplexity
restaurant	2.9k sentences	BoW	424.2
		BoWP	365.2
		BoW + BoWP	279.2
greeting	340 sentences	BoW	229.9
		BoWP	265.0
		BoW + BoWP	190.2

Table 1: Perplexity results of semantic data selection.

In Table 1, the restaurant topic contains about 2.9k sentences. This is long enough to extract plenty of word pairs to construct the feature vectors of BoWP. As can be seen, the resulting data selection based on BoWP is better than the BoW selection. On the other hand, when the topic dependent query is short, or the topic information is not strong such as for the greeting topic, there are insufficient word pairs to construct the BoWP feature vector so the data selection based on BoW is better. However, when the LMs of BoW and BoWP are interpolated, the perplexity can be reduced further in both cases, which is shown as "BoW + BoWP" in Table 1. This means that the BoWP selection results capture some information which is missed by BoW, and compensate the results of BoW.

To investigate the quality of the selected corpus over all the topics, the process described in section 5 was carried out on the BoW and BoWP methods individually. For each topic, the trigram LM was trained using the selected corpus.

Method	Size of selected corpus	Perplexity
General LM	1.8G words	283.8
BoW	254M words	240.4
BoWP	107M words	220.4

Table 2: Perplexity results for the mixture LMs with all the topics.

Then mixture models were constructed to combine the LMs of different topics by linear interpolation. These mixture LMs were used to evaluate the perplexity of the union of the queries of the different topics. In this experiment, to be consistent with topic dependent BoWP, the BoW based data selection is also carried out topic by topic and mixture LM built for each topic. The results are shown in Table 2.

The size of the selected corpus shown in Table 2 indicates the sum of the selected corpus over the 86 different topics. It shows that both BoWP and BoW based data selection get much better perplexity results than the general LM. Compared to the BoW method, the BoWP based data selection achieved the better perplexity result, while the size of selected corpus is much smaller. This result indicates that globally, the BoWP method captures more accurate semantic information than the BoW method.

In Table 3, ASR results of different data selection methods are given. The test set for the ASR experiments contains 1000 sentences in the tourism domain. The LMs based on the different data selection methods can be combined by linear interpolation, which is indicated as "+" in Table 3. All the data selection work was based on the gigaword corpus, the LMs make use of the in-domain information, but do not use the in-domain data directly.

LM	WER (%)	SER (%)
General	43.7	74.4
BoW	40.0	71.7
BoWP	39.5	72.0
BoW + BoWP	37.7	70.0

Table 3: Effect of LM training corpus data selection on recognition performance.

As the domain mismatch between the gigaword corpus and tourism is large, the general LM achieved poor results. The results in Table 3 show that a LM based on semantic data selection can improve the recognition rate significantly. If the BoW and BoWP method are used separately, they achieve comparable results. However interpolating the LM with BoW data selection and the LM with BoWP data selection can achieve better results than the LM with BoW data selection only. This means that the BoWP captures additional, more complex, semantic information that completements the BoW. Without extra training data, combining the BoW and BoWP method together, the WER of the test data was reduced from 43.7% to 37.7%.

The experiments in Table 3 compare the performance of the different data selection methods, the in-domain data was not used to train the LM directly. In this work, the size of in-domain data is about 90K sentences. It is large enough to train a reliable in-domain LM. Table 4 shows the results of experiments interpolating the in-domain LM with each of the data selection LMs above.

LM	WER (%)	SER (%)
In-domain only	27.7	59.9
In-domain + BoW	26.8	58.8
In-domain + BoWP	26.8	58.4
In-domain + BoW + BoWP	26.6	58.2

Table 4: Effect of data selection LMs on the ASR performance of in-domain LMs.

Table 4 shows that when the in-domain data was used to train the LM directly, the ASR performance was improved significantly. It also shows that when the in-domain LM is used, the differences in the ASR performance between different data selection methods became very small. The main reason is that the in-domain data was used for both LM training and data selection and the domain information overlapped.

7. Conclusions

In this work, methods of data selection based on semantic information are investigated. The bag-of-words (BoW) model was extended to the bag-of-word pairs (BoWP). The BoWP inherits the advantages of the BoW which models the global semantic information of the whole documents. Meanwhile, it can capture more complex semantic information beyond the group of individual words. A framework of topic dependent BoWP was presented to solve the combinatorial explosion of the word pairs. Perplexity experiments indicate that BoWP methods can select smaller but more accurate domain dependent data than the BoW method from a general corpus.

ASR experiments showed that using a data selection method to extract the LM training data with task specific information, can yield a significant improvement in the recognition performance over a general LM. If the BoW method and BoWP method are used individually, the ASR performance based on BoWP data selection is only slightly better than BoW method. However, when the 2 methods are combined by linear interpolation, an evident gain was achieved compared to using BoW only. This means that the information in the BoWP complements that of the BoW.

However, when the in-domain data is used to train the LM directly, the gain from data selection becomes small, with little difference between the methods. This is due to overlap of information. A potential way to more efficiently get improvements using the BoWP method is unsupervised LM adaptation. In the framework of unsupervised adaptation, the adaptation process is carried out online and the BoWP can be used to capture the semantic information in the ASR hypotheses directly.

In this work, the BoWP method is implemented in a framework of LSA. However, we believe the BoWP can be extended to other BoW methods, such as PLSA [5], LDA [6], etc.

8. REFERENCES

[1] J. R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling," *Proc. IEEE*, vol. 88, no. 8, pp. 1279-1296, 2000.

[2] J. R. Bellegarda and K.E.A. Sliverman, "Toward Unconstrained Command and Control: Data-driven Semantic Inference," in *Proc. ICSLP*, Beijing, 2000.

[3] R. Rosenfield, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer Speech and Language*, Vol. 10, pp.187-228, 1996.

[4] David Graff, "English Gigaword", Linguistic Data Consortium, Philadelphia, 2003.

[5] D. Gildea and T. Hofmann, "Topic-based Language Models Using EM," in *Proc. EUROSPEECH*, Budapest, 1999,

[6] David M. Blei, Andrew Y. Ng and Michael I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, pp. 993–1022, 2003.