Joint N-gram Chinese Language Modeling with an Application to Chinese Word Segmentation

Xin He[1], Zhijian Ou[2], Jiasong Sun[3]
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[1]hexin09@mails.tsinghua.edu.cn, [2]ozj@tsinghua.edu.cn, [3]sun@thsp.ee.tsinghua.edu.cn

Abstract

The state-of-the-art language models (LMs) are n-gram models, which, for Chinese, are word-based n-grams. To construct Chinese word-based n-gram LMs, we need to have a lexicon and a Chinese word segmentation (CWS) step. However, there is no standard definition of a word in Chinese, and it is always possible to construct new words by combining multiple characters, which causes out-of-vocabulary (OOV) problems. These make lexicon definition and CWS being difficult and ill-defined, which deteriorates the quality of the Chinese LMs. Recently, conditional random fields (CRFs) have been shown to have the ability to perform robust and accurate CWS, especially in recalling OOV words. However they are in essence not Chinese language models, but conditional models of the position-of-character (POC) tag-sequence given the character-sequence. In this paper, we propose a new Chinese language model – joint n-gram, which incorporates the POC tags so that we escape from using a lexicon. It is a truly generative model of Chinese sentences. The effectiveness of the new LM is shown in terms of perplexities and CWS performances.

1. Introduction

Statistical language models (LMs) are core components for a large variety of language technology applications such as speech recognition, machine translation, handwriting recognition, and so on. The language modeling is essentially to estimate the distribution $p(\mathbf{x})$ over all possible sentences $\mathbf{x}$ in the target language. The state-of-the-art LMs are n-gram models, which, for Chinese, are word-based n-grams. Many studies attempt to improve over the n-gram language modeling from different perspectives. The main issue examined in this paper is to address the weakness in constructing Chinese word-based n-gram LMs.

There are two prerequisites in constructing Chinese word-based n-gram LMs. First, we need to define a lexicon. Second, we need a Chinese word segmentation (CWS) step to segment the text into word-sequences, since Chinese has no word boundaries marked by spaces. However, there is no standard definition of a word in Chinese, and there is only about a 75% agreement between native speakers as what is the "correct" segmentation [1]. Moreover, it is always possible to construct new words by combining multiple characters, which causes out-of-vocabulary problems. These make the two prerequisites – lexicon definition and CWS being difficult and ill-defined, which deteriorates the quality of the Chinese LMs.

Recently, conditional random fields (CRFs) have been shown to have the ability to perform robust and accurate CWS, especially in recalling OOV words [2]. Different from traditional lexicon-based CWS methods which relies on a predefined lexicon, CRF-based CWS methods are character-based, lexicon-free. By introducing the position-of-character (POC) tags, the CWS problem is solved as character-sequence tagging. However, the CRFs used in the CWS studies are in essence not Chinese language models, or say, not generative models $p(\mathbf{x})$ of Chinese sentences $\mathbf{x}$, but conditional models $p(\mathbf{y}|\mathbf{x})$ of POC tag-sequence $\mathbf{y}$ given character-sequence $\mathbf{x}$.

Motivated by the above observations, in this paper, we propose a new Chinese language model, which incorporates the POC tags so that we escape from using a lexicon. It is a truly generative model of Chinese sentences. The idea is to pair every character x_i in a sentence with its POC tag y_i, which defines a joint-state (x_i, y_i). We then model the joint-state sequence as a Markov source of order $n-1$, which we call a joint n-gram LM.

In the experiments, the joint n-gram Chinese LMs produce lower perplexities, when compared with both the n-gram LMs based on Chinese words and the n-

gram LMs based only on Chinese characters. Moreover, when applied to the CWS task, the joint n-gram Chinese LMs perform close to CRFs and better than the traditional word-based LMs. And the ability of the joint n-gram LM to recall OOV words is clearly demonstrated.

2. Conditional random fields

As we know, Chinese word segmentation can be formulated as a tagging problem [3]. Given a Chinese character-sequence, we can segment it by tagging each character with its position in a word, which we call a position-of-character (POC) tag. For example, the POC tag of "中" in "中国" is B, which means the beginning position, and "国" is E, which means the end position. In this paper, we use four different POC tags - B, M, E, S, which represents the beginning, middle, end of a word and a single-character word respectively.

CRFs are undirected linear-chain graphical models as shown in Figure 1 [4]. When used in the CWS task, they in essence are conditional models $p(\mathbf{y}|\mathbf{x})$ of POC tag-sequence $\mathbf{y}$ given character-sequence $\mathbf{x}$ [2]. In CRFs, given a character-sequence $\mathbf{x}=\{x_1x_2\cdots x_n\}$, we can obtain the most likely tag-sequence $\mathbf{y}=\{y_1y_2\cdots y_n\}$ by maximizing the conditional probability $p(\mathbf{y}|\mathbf{x})$.

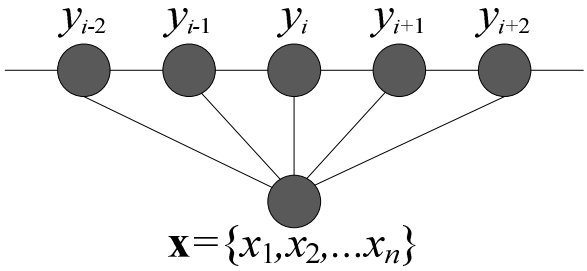

Figure 1. Graphical model representation of CRFs

Given a set of feature functions with their corresponding weights, we have $p(\mathbf{y}|\mathbf{x})$ as follows [2]:

$$p(\mathbf{y}|\mathbf{x})=\frac{\exp\sum_i\left(\sum_j\lambda_j f_j(y_{i-1},y_i,\mathbf{x},i)+\sum_k\mu_k f_k(y_i,\mathbf{x},i)\right)}{Z(\mathbf{x})} \quad (1)$$

Here $Z(\mathbf{x})$ is the normalization constant for the character-sequence $\mathbf{x}$. $f_j(y_{i-1},y_i,\mathbf{x},i)$ and $f_k(y_i,\mathbf{x},i)$ are the edge feature functions and the node feature functions respectively. λ_j and μ_k are the corresponding weights for the feature functions.

In practice, the feature functions are often organized into some classes, which are called feature templates, or simply say, features. In this paper, we use two sets of feature templates as shown in Table 1, which are used in different comparison experiments. In this table, we take the Chinese sentence "今天/天气/如何/？" ("what's the weather like today ?") and the corresponding POC tag-sequence "BEBEBES" as an instance to explain the meanings of these feature templates.

Table 1. Feature templates

Feature set-1	Instances	Feature set-2
$[x_i, y_i]$	[如, B]	$[x_i, y_i]$
	[气, B]	$[x_{i-1}, y_i]$
	[天, B]	$[x_{i-2}, y_i]$
	[何, B]	$[x_{i+1}, y_i]$
	[？, B]	$[x_{i+2}, y_i]$
$[x_{i-1}x_i, y_i]$	[气如, B]	$[x_{i-1}x_i, y_i]$
	[如何, B]	$[x_ix_{i+1}, y_i]$
	[气何, B]	$[x_{i-1}x_{i+1}, y_i]$
$[y_{i-1}, y_i]$	[E, B]	$[y_{i-1}, y_i]$

The feature set-1 in Table 1 has only three feature templates, in which the tag of the current character is only related to the current character and the previous tag. This is similar to character-based 2-gram model. The feature set-2 has nine feature templates. The CRF with feature set-1 is the basic CRF for comparison, while the CRF with feature set-2 is used to show the best performance of CRFs.

CRFs' ability to perform robust and accurate CWS is mainly due to their two virtues. First, CRFs are conditional models which are suited to the tagging problem. Second, CRF-based CWS introduces POC tags, which avoids the use of a lexicon and thus improves the recall rate of OOV. To develop Chinese language models, we need generative models of Chinese sentences. So we are motivated to incorporate the second virtue of CRFs to the traditional n-gram LMs to design a new Chinese language model, which is described in the following.

3. Joint n-gram model

The idea of the new model is to pair every character x_i in a sentence with its POC tag y_i, which defines a joint-state (x_i, y_i). We then model the joint-state sequence as a Markov source of order $n-1$, which we call a joint n-gram LM, as shown in Figure 2.

Joint n-gram LMs are n-gram models based on joint-states. In contrast, the traditional n-gram LMs are either based on Chinese words or based only on Chinese characters. The word-based n-gram LMs suffer from OOV problem. The n-gram LMs based only on characters ignore the information of POC tags. By this analysis, joint n-gram models appear to be a better choice for Chinese LMs.

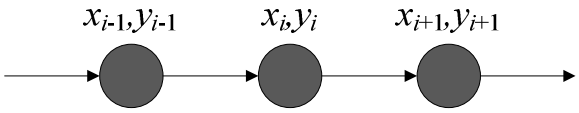

Figure 2. Graphical model representation of a joint bigram model

The joint n-gram LM can be used for CWS. Given a character-sequence $\mathbf{x} = \{x_1 x_2 \cdots x_n\}$, we can obtain the most likely joint-state sequence $(\mathbf{x}, \mathbf{y}) = \{(x_1, y_1)(x_2, y_2) \cdots (x_n, y_n)\}$ using the Viterbi algorithm [5]. Then, we can extract the corresponding optimal POC tag-sequence $\mathbf{y} = \{y_1 y_2 \cdots y_n\}$, which gives the result for CWS.

4. Experimental results

Two sets of experiments are conducted to evaluate the effectiveness of the new LMs in terms of perplexities and CWS performances.

4.1. Perplexity

In our experiments, we use the simplified Chinese corpus of Peking University (PKU) in SIGHAN 2005 as our training and testing corpus[1], which has been manually segmented into Chinese words. Information about this corpus is shown in Table 2.

Table 2. Information about the PKU Corpus (# denotes "the number of")

	#Word types	#Word tokens	#Character types	#Character tokens
Train	55,303	1,109,947	4,698	1,826,448
Test	13,147	104,372	2,932	172,733

A standard measure of the quality of a LM is the perplexity of the LM (i.e. the log probability it assigns to some held-out data set) [6]. Thus we compare the perplexities of different LMs, including the joint n-gram models, the character-based n-gram models and the word-based n-gram models. The main conclusions are as follows, which can be drawn from the detailed results in Table 3. Note that different LMs may model linguistic units at different levels - character-level or word-level. For meaningful comparisons, we report in Table 3 both the character-level perplexities and the original word-level perplexities for word-based n-grams.

Table 3. Perplexities of various LMs for the PKU corpus. The numbers in the brackets are the character-level perplexities for word-based n-grams, i.e. being normalized by the number of characters.

Model	logprob	ppl
Joint2gram	-334936	82.69
Joint3gram	-312875	61.82
Char2gram	-352488	104.218
Char3gram	-313977	62.73
Word2gram	-318442	765.17 (66.62)
Word3garm	-311570	662.91 (60.73)

(1) The perplexities of the joint n-gram models are always lower than those of the character-based n-gram models of the same order.

(2) The character-based 3-gram model produces lower perplexities than the word-based 2-gram model. Considering that on average there are 1.5 characters per word in Chinese [7], this comparison is appropriate.

(3) It can be observed that the perplexity reduction from char3gram to joint3gram is not as large as that from char2gram to joint2gram. Note that the size of the state-space of joint n-gram is increased by incorporating POC tags. The high-order joint n-gram may have data sparseness problem, which could hurt the performance of the high-order joint n-gram. So we conduct another perplexity experiment by using a larger training corpus – the segmented corpus of National Language Committee (NLC)[2]. Information about this corpus is shown in Table 4.

[1] http://www.sighan.org/bakeoff2005/

[2] http://www.china-language.gov.cn/

Table 4. Information about the NLC corpus (# denotes "the number of")

	#Word types	#Word tokens	#Character types	#Character tokens
Train	88,905	31,267,367	8,380	45,763,646
Test	13,809	77,927	3,515	113,911

As shown in Table 4, the size of NLC corpus is almost 30 times as that of PKU corpus, which could ameliorate the data sparseness problem. The new experiment result is shown in Table 5. It can be seen that the perplexity reduction from char3gram to joint3gram becomes larger.

In summary, the joint n-gram Chinese LMs produce lower perplexities, when compared with both the n-gram LMs based on Chinese words and the n-gram LMs based only on Chinese characters.

Table 5. Perplexities of various LMs for the NLC corpus

Model	logprob	ppl
Joint2gram	-225334	82.65
Joint3gram	-199258	49.59
char2gram	-238329	106.61
char3gram	-202296	52.63

4.2. Chinese word segmentation

In this set of experiments, we compare the CWS performances of the joint n-gram models, the CRFs and word-based n-gram model on the PKU corpus. We follow the evaluation metrics in Bakeoff [8]. We have F-measure $F = 2RP/(R+P)$, where R and P are the recall rate and precision of the segmentation. Roov and Riv represent the recall rate for OOV words and in-vocabulary words. The main conclusions are as follows, which can be drawn from the detailed results in Table 6.

(1) The F-measure of the joint 2-gram model is close to that of the CRF-1 model (i.e. the CRF with the feature set-1 defined in Section 2). Note that both the CRF-1 model and the joint 2-gram model exploit character-level 2-order information. This is a fair comparison.

(2) The F-measure of the CRF-2 model becomes slightly better than the joint 3-gram model. Note that both the CRF-2 model and the joint 3-gram model exploit character-level 3-order information. This is a fair comparison. The slightly better performance of the CRF-2 model may be attributed to the CRFs' virtue of being conditional models.

Table 6. CWS performances on the PKU corpus.

Model	R	P	F	Roov	Riv
CRF-1	0.93	0.937	0.933	0.606	0.941
CRF-2	0.944	0.954	0.949	0.682	0.953
Joint2gram	0.934	0.938	0.936	0.441	0.952
Joint3gram	0.938	0.942	0.94	0.443	0.956
word2gram	0.949	0.913	0.931	0.016	0.982
word3gram	0.949	0.913	0.931	0.016	0.982

(3) Both the CRFs and the joint n-gram models show their capability to recall OOV words. The Roov rates for these two models are much higher than the word-based n-gram models. These superior performances of the two models could be due to the fact that both models incorporate POC tags.

In summary, when applied to the CWS task, the joint n-gram Chinese LMs perform close to CRFs and better than the traditional word-based LMs. Moreover, the ability of the joint n-gram LM to recall OOV words is clearly demonstrated.

5. Conclusion

In this paper, we mainly address the weakness in constructing Chinese word-based n-gram LMs. Lexicon definition and CWS are two prerequisites in constructing Chinese word-based n-gram LMs. However, there is no standard definition of a word in Chinese, and it is always possible to construct new words by combining multiple characters, which causes OOV problems. These make lexicon definition and CWS being difficult and ill-defined tasks, which deteriorates the quality of the Chinese LMs. In this paper, we propose a new Chinese language model – the joint n-gram LM. It incorporates the POC tags so that we escape from using a lexicon – like CRFs. And it is a truly generative model of Chinese sentences – unlike CRFs which are in essence conditional models. The effectiveness of the new LMs is shown in terms of perplexities and CWS performances.

6. Acknowledgments

This work is supported by National Natural Science Foundation of China (61075020).

7. References

[1] R. Sproat, C. Shih, W. Gale, and N. Chang, “A stochastic finite-state word-segmentation algorithm for Chinese”, Computational Linguistics, 1996, pp. 22(3):377-404.
[2] F. Peng, F. Feng, and A. McCallum, “Chinese segmentation and new word detection using conditional random fields”, COLING 2004, Geneva, Switzerland, 2004, pp. 562–568.
[3] N. Xue, “Chinese word segmentation as character tagging”, International Journal of Computational Linguistics and Chinese Language Processing, 2003, pp. 8(1):29-48.
[4] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: probabilistic models for segmenting and labeling sequence data”, ICML ’01: Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, 2001, pp. 282–289.
[5] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, 1989, pp. 77(2): 257-286.
[6] S. F. Chen and J. Goodman, “Empirical study of smoothing techniques for language modeling”, Computer Speech and Language, 1999, pp. 359-394.
[7] J. Luo, L. Lamel, and J. L. Gauvain, “Modeling characters versus words for mandarin speech recognition”, ICASSP 2009, Taipei, Taiwan, 2009, pp. 4325-4328.
[8] T. Emerson, “The second international Chinese word segmentation bakeoff”, Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, 2005, pp. 123-133