Stochastic Language Models for Automatic Acquisition of Lexicons from Printed Bilingual Dictionaries

Song Mao and Tapas Kanungo
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
maosong,kanungo@almaden.ibm.com

Abstract

Electronic bilingual lexicons are crucial for machine translation, cross-lingual information retrieval and speech recognition. For low-density languages, however, the availability of electronic bilingual lexicons is questionable. One solution is to acquire electronic lexicons from printed bilingual dictionaries. While manual data entry is a possibility, automatic acquisition of lexicons from scanned images of bilingual dictionaries would expedite the prototyping process of cross-language systems. Printed dictionaries have a logical model that defines the syntax of the dictionary entries – i.e. order of the dictionary entry, its part of speech, its pronunciation and its definition. In this article we propose an algorithm to automatically extract bilingual dictionary entries based on stochastic language models. We demonstrate this algorithm on a printed Chinese-English dictionary. This work can be easily used for extracting information from other tabular structures like telephone books, catalogs, etc.

1 Introduction

Document image analysis (DIA) is the process of converting a document page image into a symbolic file. An important component of the any DIA system is the analysis of the logical structure of document page images. Document logical structure specifies logical labels and the logical relation among document components. For instance, in a scientific journal paper, *title* and *author* are logical labels and *title* is followed by *author*. Document logical structure analysis is a process of assigning logical labels and relations to the physically segmented regions. While numerous algorithms have been proposed for analyzing a document's physical structure, there is relatively less published work on document logical structure analysis.

Bilingual lexicon is an indispensable tool in natural language understanding, machine translation, cross-language information retrieval and speech recognition. In this paper, we describe an algorithm for parsing the logical structure of printed bilingual dictionaries using a stochastic language model. We discuss relevant literature in Section 2. In Section 3 we present a stochastic language model for representing the logical structure of a printed Chinese-English dictionary. In Section 4, we describe a protocol for training and evaluating our algorithm and in Section 5 we discuss the experimental results.

2 Survey

Language models have been used for efficiently modeling the syntactic knowledge in the given data in many applications besides linguistics. Formal languages represented by the grammatical rules have been used in pattern recognition [5]. In the computer vision area, language models such as context-free grammars have been used for recognizing visual activities [7]. N-grams and Hidden Markov Model (HMM) are very popular language models used in speech recognition [17]. Language models such as finite state automata have been used for document image decoding [11].

Document layout structure consists of document physical layout analysis and logical layout analysis. Document physical analysis method can be categorized into top-down approaches [13, 2], bottom-up approaches [14, 8] and hybrid approaches [16]. A survey of page segmentation algorithms can be found in [15]. Mao and Kanungo [12] evaluated three research physical segmentation algorithms and two commercial products. Rule-based system have been used for automatically labeling the document images [10]. The linguistic properties of document layout have been studied in [6, 19]. Conway [3] performed document layout recognition based on a set of page grammars. Krishnamoorthy et. al. used a set of block grammars for syntactic segmentation and labeling of document pages.

Efficient parsing algorithms have been proposed for generating syntactic structures based on grammatical rules.

Earley algorithm [1] is a top-down dynamic programming approach. Cocke-Younger-Kasami (CYK) algorithm [1] is a bottom-up dynamic programming approach. While deterministic language models can be useful, they are not useful when the input has sentences with grammatical errors. Stochastic parsing techniques can remove this difficulty by selecting a parsing tree with the highest probability. Stolcke [18] augmented Earley parser by a set of probability and efficiently find the optimal parsing result.

3 Document Model

We use a generative document model for modeling dictionary pages as shown in Figure 1(a) and Figure 1(b). This model contains a physical layout model, a logical structure model, a text generation model, a border noise model and a local noise model. In this paper, we focus on the logical structure recognition and the detailed description of the physical structure model and recognition algorithm can be found in [9]. Figure 1(c) shows a sample segmentation result of a dictionary page.

3.1 Document Logical Model

In the Chinese-English dictionary, there are four categories of objects: Pinyin (Chinese pronunciation), Chinese word, part of speech and the English definition. Within each lexicon item, the objects of different types are logically structured. We use context-free grammar (CFG) to model this logical structure of Chinese-English dictionary lexicon items. Let $G = (V_N, V_T, P_s, S)$ denote a context-free grammar and $V_N = \{S, P, C, T, U, G, D\}$ be a set of all nonterminal symbols, where P is a Pinyin phrase, C is a Chinese phrase, T is a translation phrase, U is a translation unit phrase, G is a part of speech phrase and D is a definition phrase. $V_T = \{a, b, c, d\}$ is a set of all terminal symbols, where a is Pinyin, b is Chinese, c is part of speech and d is English definition. Finally, let P_s be the finite set of stochastic production rules and S be the nonterminal start symbol. For the Chinese-English dictionary, we formally define the possible production set P as follows:

S $\rightarrow$ P C T [1.0]	T $\rightarrow$ U T [$1-\gamma$]
P $\rightarrow$ a [α]	U $\rightarrow$ G D [1.0]
P $\rightarrow$ a P [$1-\alpha$]	G $\rightarrow$ c [τ]
C $\rightarrow$ b [β]	G $\rightarrow$ c G [$1-\tau$]
C $\rightarrow$ b C [$1-\beta$]	D $\rightarrow$ d [δ]
T $\rightarrow$ U [γ]	D $\rightarrow$ d D [$1-\delta$]

where the numbers in the squared bracket are rule probabilities. The word split error from physical segmentation can be handled by the rules P $\rightarrow$ a P, C $\rightarrow$ b C, G $\rightarrow$ c G and D $\rightarrow$ d D. However, if two or more words from different categories are merged during physical segmentation procedure, the parser will generate erroneous results. Hence, it is important to train the physical model parameters to avoid merge error as much as possible. On the other hand, if split error becomes sever, the length of string of physically segmented words in lexicon item becomes longer and the parsing of it becomes slower.

We augment each production rule with a probability. The parsing result with the highest probability is retained and all other parsing results are discarded. In this paper, we use the Stolcke-Earley parser [18] to generate the most probable parsing result based on stochastic context-free grammars. The terminal symbol sequences are the inputs to the parser.

3.2 Modeling the Uncertainty in the Input Symbols

In our problem the terminal symbols are also probabilistic. That is, for each word-token we have probabilities that the token can be one of the four categories, namely, Pinyin, or a Chinese word, or a part of speech, or part of a definition. These probabilities are derived from the distribution of the height feature that is extracted from the word bounding box generated by the physical segmentation stage. Thus the input has a distribution on the four categories. Let x be a random variable corresponding to the feature extracted from the input and let C be a random variable whose value can be one of the four categories.

By studying the physical segmentation results as shown in Figure 1, we modeled the height feature distribution of Pinyin, Chinese and part of speech using Gaussian distributions. However, since English words can have ascender, descender or none of them, the height distribution has three peaks. Mixture density can be used to model this kind of distribution. In our case, we used the height histogram distribution obtained from a training dataset as the empirical probability density function. For a given x,

$$P(C = i|x) = \frac{f(x|C = i) \cdot P(C = i)}{\sum_j f(x|C = j) \cdot P(C = j)}.$$

The prior probability $P(C = i)$ can be easily estimated from the training dataset. Therefore for each input feature x, we can compute $P(C = i|x), i = 1, 2, 3, 4$ using above equations.

4 Experimental Protocol

Our dictionary image page test data set contained six scanned pages from a Chinese-English dictionary [4]. Each dictionary page contains about 90 lexicon items. We used one dictionary page with 88 lexicon items for estimating the input feature model parameters and stochastic context-free grammar production rule probabilities. We then evaluated our algorithm with the estimated model parameters on the

(a) (b) (c)

Figure 1. This figure shows a real dictionary page in (a), a synthetic dictionary page in (b) using estimated style parameters for (a), and a sample physically segmented real dictionary page in (c). Notice that one of the headword and the corresponding Chinese word have both been split into two tokens.

remaining five dictionary pages. Different OCR engines are selectively applied to the the subimages corresponding to logical components parsed by the stochastic parser. The script of the text was automatically classified by the algorithm. Finally, the logical structure of the file was generated.

The platform used for the implementation, algorithm training and testing was a Dell PC with a 355 MHz Pentium II CPU and 128 MB main memory running Linux 7.0 operating system.

5 Experimental Results and Discussions

We report four performance indices of our algorithm: parsing accuracy, parsing failure rate, parsing error rate and parsing miss rate. We report these numbers on individual pages of the dictionary pages and then report the average over all five test dictionary pages as shown in Table 1.

Parsing accuracy is defined as the ratio of number of correctly parsed lexicon items and the total number of lexicon items. Parsing failure rate is defined as the ratio of number of lexicon items that can not be parsed and the total number of lexicon items. Parsing error rate is defined as the ratio of number of lexicon items that are incorrectly parsed and the total number of lexicon items. Parsing miss rate is defined as the ratio of number of lexicon items that are not parsed and the total number of lexicon items.

Most failed parses are due to inaccurate feature extraction as shown in Figure 2(a) where the height of part of

Table 1. Performance Indices. The values are in percentage.

Page No.	1	2	3	4	5	page average
parsing accuracy	91.11	89.80	96.70	93.10	90.63	92.21
parsing failure rate	0	2.04	1.10	1.15	7.29	2.32
parsing error rate	6.67	7.14	2.20	5.75	2.08	4.81
parsing miss rate	2.22	1.02	0	0	0	0.66

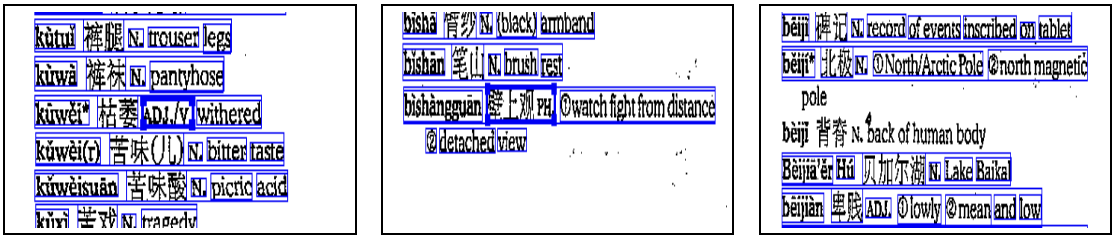
speech is too large; most incorrect parses are due to the fact that words of different category within a lexicon item got merged as shown in Figure 2(b). However, missed parses are due to incorrect physical segmentation result on the missed lexicon item as shown in Figure 2(c).

Currently our feature includes token height only. We can see that the parsing result is very good given the simple feature we extract from physically segmented tokens. However, the performance can be further improved if users extract better and more features from the segmented tokens. We plan to evaluate our algorithm on a much larger dataset and extract multi-dimensional features from physically segmented tokens.

References

[1] A. V. Aho and J. D. Ullman. *The Theory of Parsing, Translation, and Compiling*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[2] H. S. Baird, S. E. Jones, and S. J. Fortune. Image segmentation by shape-directed covers. In *Proceedings*



(a) (b) (c)

Figure 2. This figure shows an example of failed parsing (a), incorrect parsing (b) and missed parsing (c).

of International Conference on Pattern Recognition, pages 820–825, Atlantic City, NJ, June 1990.

[3] A. Conway. Page grammars and page parsing a syntatic approach to document layout recognition. In *Proceedings of International Conference on Document Analysis and Recognition*, pages 761–764, Tsukuba Science City, Japan, October 1993.

[4] J. DeFrancis, Y. Q. Bai, S. Z. Fang, V. H. Mair, R. M. Sanders, R. Y. D. Sun, and B. Y. Yi, editors. *ABC Chinese-English Dictionary*. Chinese Grand Dictionary, Shanghai, China, 1997.

[5] K. S. Fu. *Syntactic Methods in Pattern Recognition*. Academic Press, New York and London, 1974.

[6] M. Hurst. Layout and language: An efficient algorithm for detecting text blocks based on spatial and linguistic evidence. In *Proceedings of SPIE Conference on Document Recognition*, pages 56–67, San Jose, CA, January 2001.

[7] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:852–872, 2000.

[8] A. K. Jain and B. Yu. Document representation and its application to page decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:294–308, 1998.

[9] T. Kanungo and S. Mao. Stochastic language model for analyzing document physical layout. In *Proceedings of SPIE Conference on Document Recognition and Retrieval*, San Jose, CA, January 2002. Submitted.

[10] J. Kim, D. X. Le, and G. R. Thoma. Automated labeling in document images. In *Proceedings of SPIE Conference on Document Recognition*, pages 111–117, San Jose, CA, January 2001.

[11] G. E. Kopec and P. A. Chou. Document image decoding using markov source models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:602–617, 1994.

[12] S. Mao and T. Kanungo. Empirical performance evaluation methodology and its application to page segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:242–256, 2001.

[13] G. Nagy, S. Seth, and M. Viswanathan. A prototype document image analysis system for technical journals. *Computer*, 25:10–22, 1992.

[14] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:1162–1173, 1993.

[15] L. O'Gorman and R. Kasturi. *Document Image Analysis*. IEEE Computer Society Press, Los Alamitos, CA, 1995.

[16] T. Pavlidis and J. Zhou. Page segmentation and classification. *Graphical Models and Image Processing*, 54:484–496, 1992.

[17] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[18] A. Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21:165–201, 1995.

[19] L. Todoran, M. Aiello, M. Christof, and M. Worring. Logical structure detection for heterogeneous document classes. In *Proceedings of SPIE Conference on Document Recognition*, pages 99–110, San Jose, CA, January 2001.