Parsing N-Best Lists of Handwritten Sentences

Matthias Zimmermann*, Jean-Cédric Chappelier** and Horst Bunke*
*Department of Computer Science, University of Bern
Neubrückstrasse 10, CH-3012 Bern, Switzerland
{zimmerma,bunke@iam.unibe.ch}
**Artificial Intelligence Laboratory, Swiss Federal Institute of Technology
INR (Ecublens), CH-1015 Lausanne, Switzerland
{Jean-Cedric.Chappelier@epfl.ch}

Abstract

This paper investigates the application of a probabilistic parser for natural language on the list of the N-best sentences produced by an off-line recognition system for cursive handwritten sentences. For the generation of the N-best sentence list an HMM-based recognizer including a bigram language model is used. The parsing of the sentences is achieved by a bottom-up chart parser for stochastic context-free grammars which produces the parse tree of the input sentence as well as the word tags. From a collection of corpora we extract the linguistic resources to build the lexicon, a word bigram model and the stochastic context-free grammar. Results from experiments indicate an increase of the word and sentence recognition rate when using the proposed combination scheme.

Keywords: handwritten sentence recognition, natural language parsing, N-best list reordering.

1 Introduction

In the field of off-line handwriting recognition we observe a tendency to address problems of increasing complexity. High recognition rates have been published for the recognition of isolated digits [6] or characters [17]. The achieved performance for numeral string recognition [12] or isolated word recognition [11] is already significantly lower. If the task complexity increases further, as in the case of the recognition of handwritten addresses [13] or bank cheques [4], task specific knowledge like the relation between zip code and city name, or between courtesy amount and legal amount becomes essential.

For general text recognition task specific information can be found in the linguistic domain. So far the successful application of n-gram language models supporting the recognition of handwritten text lines has been reported [14, 18]. Unfortunately the power of n-gram language models is limited by the fact that they are unable to capture long distance relationships between words. For $n > 3$ it is difficult to accumulate sufficient amounts of text for an accurate estimation of the probability for less frequently observed word sequences.

In this paper we investigate the application of a bottom-up chart parser for stochastic context-free grammars as a postprocessing step in a hidden Markov model (HMM) based recognition system for handwritten sentences. With the proposed combination scheme we address two objectives. The first objective is to improve the word and sentence recognition rate. The second is to produce additional information in the form of grammatical word tags and a parse tree which can be very valuable in the context of semantic information retrieval or text understanding. To the knowledge of the authors it is the first time that linguistic information in the form of a stochastic context-free grammar has been applied in the field of handwriting recognition. Related works in the literature have been only found in the domain of machine printed text recognition [7] and speech recognition [10, 2]. For on-line handwritten sentence recognition [16] reported the use of a stochastic linear grammar in initial experiments.

The next section presents the resources involved, namely, handwritten sentence images serving as input, and the linguistic knowledge sources. Section 3 provides a description of the methodology. The HMM-based recognition module is described, some features of the parsing algorithm are highlighted and the combination scheme using recognition scores and parse probabilities is introduced. Experiments and results are pro-

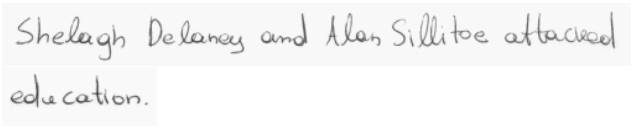

Figure 1. A sample sentence from the IAM Database.

vided in Section 4 and conclusions are drawn in Section 5.

2 Resources

2.1 Handwritten sentences

The goal of the work presented in this paper is to recognize complete handwritten sentences. We therefore assume that a segmentation of the given input text into individual sentences has already been accomplished. Each sentence consists of one or more consecutive lines (or fragments of lines) of handwritten text depending on the writing style, the writing size, the length of the input sentence and its position within a page of handwritten text. An example of a handwritten sentence image is given in Fig. 1.

The handwritten sentences have been taken from the IAM Database [15] which is based on sentences provided in the LOB Corpus of British English (LOB) [9]. For the experiments 1518 sentences written by 331 different persons were automatically extracted from the segmented IAM Database [19]. In [19] only the segmentation of the database into individual words is described. However, this segmentation yields the exact location of the bounding box of each word including a link into the ASCII ground truth of the corresponding sentence. Hence a procedure for the automatic extraction of complete sentences from an image can be easily derived from the method described in [19].

2.2 Linguistic resources

The linguistic resources used for the recognition of handwritten text and the parsing of the resulting sentences should be as consistent as possible. Since the handwritten sentences in the IAM Database are based on the LOB Corpus, the Lancaster Parsed Corpus (LPC) [3] was used to extract a suitable grammar[1].

The syntactic class, also called "Part-of-Speech tag" is attached to each word in the LPC. For example the word "education" is a singular common noun, tagged as NN, and the word "attacked" is tagged as VBD which stands for the past tense of a lexical verb. In order to better estimate the frequencies of the tagged words and to be able to produce a more robust bigram language model for the handwriting recognition module, the Tagged LOB Corpus (TLOB) [8] containing 1,000,000 tagged words was used as an additional linguistic resource. Principally the TLOB represents the same sentences as the LOB, but in many cases the surface form of a word differs from the LOB (e.g. the abbreviation of Mister in the LOB is the word "Mr.", but is "mr" in the TLOB[2]).

The following text provides an example of a sentence from the IAM Database, the corresponding tagged sentence from the TLOB, as well as a parse in the form of a bracketed sentence as provided by the LPC.

```
Shelagh Delaney and Alan Sillitoe attacked education .

Shelagh_NP Delaney_NP and_CC Alan_NP Sillitoe_NP
attacked_VBD education_NN ._.

[S [N& Shelagh_NP Delaney_NP [N+ and_CC Alan_NP
  Sillitoe_NP ] ] [V attacked_VBD ] [N education_NN ]
  ._. ]
```

In addition to the inconsistent surface forms of the words, differences between the tagset used by the TLOB and the LPC were detected. For the definition of the surface form of words we tried to be as close as possible to the LOB. For the tagset, the LPC was used as a starting point making small adaptations regarding the application of the tagset in the context of handwriting recognition. For example the tag DOX was introduced to replace the sequence "do_DO n't_XNOT" by the word "don't_DOX".

3 Methodology

3.1 The HMM-based recognition system

The recognition system used for the experiments is based on continuous density character HMMs [14]. Two modifications of the system described in [14] were applied. First, the number of states per character model was defined using Bakis length modeling [20]. The second modification is the introduction of multiple Gaussians per model state [5].

For the training an embedded model reestimating scheme (Baum-Welch algorithm) is applied on the preprocessed lines of text, while the Viterbi algorithm is used for text line recognition. During the Viterbi

[1] The LPC represents a subset of the LOB and contains parses for roughly 12,000 sentences including 145,000 words.

[2] The unification of linguistic resources is a laborious and time consuming task.

$\phi(s)$	Sentence candidates
20147	[...] and Alan Sillitoe attacked education
20116	[...] and Alan Sillitoe attacked education .
20058	[...] and Alan Sillitoe attacked Education

Table 1. The 3-best sentence candidates with their recognition scores.

$\psi(s)$	$\phi(s)$	$P_G(s)$	Sentence candidates
19854	20116	7.9e-30	[...] attacked education .
19838	20147	4.7e-35	[...] attacked education
19738	20058	2.8e-36	[...] attacked Education

Table 2. The 3-best sentence candidates after reordering of the N-best list.

decoding, word transition probabilities from a bigram language model M are used to improve the recognition rate of the system.

The word transition probabilities are incorporated in the following way

$$\phi(s_i) = \phi(s_{i-1}) + \log(L_M(w_i)) + \alpha \ \log(P(w_i|w_{i-1})) + \beta$$

where $\phi(s_i)$ denotes the log-likelihood score for the word sequence $s_i = (w_1, w_2, ...w_i)$ and $L_M(w_i)$ the likelihood score for the observation sequence representing w_i. The word transition probability from w_{i-1} to w_i is represented by $P(w_i|w_{i-1})$. Parameter α represents the grammar scale factor which weights the influence of the bigram model on the recognition result. Parameter β helps to control the word insertion rate of the recognizer.

For each handwritten line of text the N-best choices together with the corresponding recognition scores are stored. After the recognition of the text lines the N-best lists of a given sentence are concatenated and used to derive the N-best sentence candidates with the accumulated recognition scores. Tab. 1 provides an example of a 3-best list of sentence candidates for part of the text shown in Fig. 1 (the beginning of the sentences was the same for all candidates and has been truncated for printing). Note that the second candidate represents the correct transcription since it includes the full stop at the end of the sentence.

3.2 The bottom-up chart parser

For each sentence candidate in the N-best list the parser produces the most probable parse. The input for the parser is therefore a sequence of words provided by the handwriting recognition module.

The parser used for the experiments implements a bottom-up chart parsing algorithm for stochastic context-free grammars [1]. The probabilities of the lexicalized rules[3] were estimated from the observed frequencies of the tagged words provided in the TLOB, and the probabilities of the non-lexicalized rules were computed from the observed frequencies in the LPC. Here is an example of two non-lexicalized and two lexicalized rules:

```
S -> N& V N .
N& -> NP NP N+
NP  -> Shelag
NP -> Delany
```

The grammar G is then represented by the combined set of lexicalized and non-lexialized rules with attached rule probabilities learned from the bracketed sentences in the LPC. The start symbol S of the grammar has the syntactical meaning of an independent sentence. Given a sentence $s = (w_1, w_2, ...w_n)$ the parser finds the most probable parse and its probability $P_G(s)$.

3.3 The combination scheme

The combination of the recognition score with the parse probability is handled analogously to the incorporation of the transition probabilities from the bigram language model in the HMM. In contrast to the transition probabilities (which are merged with the recognition scores on the word level) the parse probabilities are merged on the sentence level. The sentence score $\psi(s)$ for a sentence $s = (w_1, w_2, ...w_n)$ is then defined in the following way

$$\psi(s) = \phi(s) + \gamma \ \log(P_G(s))$$

where $\phi(s)$ represents the recognition score from the handwriting recognition module and $P_G(s)$ the probability of the most probable parse for this sentence. Parameter γ weights the influence of the parse probability on the final sentence score. For $\gamma = 0$ the parse probability will not affect the sentence score at all. In case $\gamma > 0$ the sentence scores are influenced by the parse probabilities and a reordering of the N-best sentences may take place. Tab. 2 shows the reordered list using the example from Tab. 1 with $\gamma = 9$.

4 Experiments

For the experiments a multi-writer handwriting recognition system was trained on 4,200 lines of handwritten text. The lines contain over 30,200 words from

[3]Rules containing only a terminal symbol (word) on their right hand side are called lexicalized rules

which 84 character models were extracted (lower- and upper-case letters, digits and inter-punctuation symbols). For the number of states per character model Bakis length modeling was used and the system was trained using three Gaussians per model state. The optimization of the grammar scale factor α and the word insertion penalty β was made on a validation set containing 100 lines. Using the closed vocabulary assumption it was made sure that all words occurring in the validation and the test set were present in the lexicon.

In order to provide a realistic scenario for the experimental setup, all sentences of the test set were excluded from all linguistic resources used for the extraction of the grammar rules, the estimation of the rule probabilities, and the transition probabilities of the bigram language model. The underlying grammar contains 11,760 non-lexicalized rules which were extracted from 11,827 sentences of the LPC, and the number of lexicalized rules was restricted to the 10,000 most frequent ones using the counts of the TLOB corpus. In the set of lexicalized rules, 8,806 distinct word surface forms were found and used as lexicon for the extraction of the bigram language model. The transition probabilities were then estimated using the TLOB corpus.

The test set contained 67 sentences written on 170 text lines by 52 individuals. Using the extracted grammar directly on the transcriptions of the test sentences the parser failed only once to produce a parse tree. For the optimization of the parse scale factor γ, six-fold cross validation was applied on the 67 test sentences. For each partitioning there were 11 sentences in the test set and 56 sentences in the training set. On the training set the parse scale factor γ was optimized in the range from 0 to 15 [4]. For all partitionings $\gamma = 9$ produced the highest word recognition rates on the training set. Tab. 3 provides the averaged sentence and word recognition rates for the test sets using the handwriting recognizer alone and in combination with the parser. Since the top word recognition rates are close to each other, the student t-test was applied. It was found that the improvement is significant on the 95% level. The low sentence recognition rate can be explained by the fact that a sentences in the test set contains 17 words in average, and a sentence was considered to be correctly recognized only if the final recognition result matched the transcription of the sentence image exactly. The following two changes to the current system could further increase the relative improvement obtained by the use of a parser: First, a recognition system with a higher word recognition rate could produce less sentence candidates which are grammatically acceptable but contain many missrecognized words. Second, a less ambiguous grammar could help to rule out sentence candidates like 'what a struck you.' which can be parsed using the grammar extracted from the LPC.

[4] In the case of $\gamma = 0$, the parser is not used at all and for $\gamma = 15$ it has been found that the influence of the parser was too strong.

System	Sent. Top	Sent. 5-Best	Word Top	Word 5-Best
Recognizer	4.6%	10.6%	67.47%	71.36%
Recognizer and Parser	6.0%	10.6%	68.32%	71.47%

Table 3. Average sentence and word recognition rates for the top choice and the 5-best list.

From the correctly recognized words, 86% have been assigned the correct grammatical tag by the parser. It is remarkable that for the correctly recognized sentences 100% of the assigned tags were correct. For six of the seven correctly recognized sentences a parse tree was available in the Lancaster Parsed Corpus. In five cases the parse tree matched exactly the parse provided by the corpus. In one case the generated parse matched the provided parse tree except for a prepositional phrase `P` which was attached to the corresponding noun phrase `N` instead of the top level symbol `S`.

5 Conclusions

This paper presents the combination of an HMM-based handwriting recognition system and a bottom-up chart parser for stochastic context-free grammars. For the combination a sequential coupling of the HMM recognizer and a parsing module is used, where the parser is fed with the N-best candidate sentences provided by the handwriting recognition module. The handwriting recognition module was trained for a multi-writer task including material from over 300 individuals and using a lexicon of more than 8,000 words. The grammar for the parsing module was extracted from bracketed sentences resulting in more than 11,000 non-lexicalized rules and 10,000 lexicalized rules representing the lexicon of the handwriting recognition module.

Experiments using this combination scheme produced an increase in both the sentence and the word recognition rate. At the same time the correct grammatical tag was assigned to 86% of the correctly recognized words. It is remarkable that for correctly recognized sentences the correct tag rate was 100% and the quality of the generated parses was very high. From

this observation it can be concluded that a better handwriting recognition module will further increase the correct tag rate. The parse trees and grammatical word tags produced by the proposed method can provide valuable information for further processing the transcriptions of the handwritten input in the context of semantic information retrieval and text understanding systems.

Future work includes the improvement of the baseline recognition system using more training material, a higher number of Gaussians, and the use of more sentences in the training and test set. Further steps will involve the direct parsing of the recognition lattices produced by the handwriting recognition module without a previous extraction of N-best sentences.

References

[1] J.-C. Chappelier and M. Rajman. A generalized CYK algorithm for parsing stochastic CFG. *Actes de TAPD*, pages 133–137, 1998.

[2] J.-C. Chappelier, M. Rajman, R. Aragüés, and A. Rozenknop. Lattice parsing for speech recognition. In *6e Conf. sur le Traitement Automatic du Langage Naturel (TALN99)*, pages 95–104, 1999.

[3] R. Garside, G. Leech, and T. Váradi. *Manual of Information for the Lancaster Parsed Corpus*. Norwegian Computing Center for the Humanities, Bergen, 1995.

[4] N. Gorski, V. Anisimov, E. Augustin, D. Price, and J.-C. Simon. A2iA check reader: A family of bank check recognition systems. In *5th Int. Conf. on Document Analysis and Recognition*, pages 523–526, Bangalore, India, 1999.

[5] S. Guenter and H. Bunke. Optimizing the number of states, training iterations, gaussians in an HMM-based handwritten word reconizer. In *7th Int. Conf. on Document Analysis and Recognition, Edinbourgh, Scotland*, volume 1, pages 472–476, 2003.

[6] T. Ha and H. Bunke. Off-line handwritten numeral recognition by perturbation method. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):535–539, 1997.

[7] T. Hong and J. Hull. Text recognition enhancement with a probabilistic lattice chart parser. In *Int. Conf. on Document Analysis and Recognition*, pages 222–225, Tsukuba, Japan, 1993.

[8] S. Johansson, E. Atwell, R. Garside, and G. Leech. *The Tagged LOB Corpus, Users's Manual*. Norwegian Computing Center for the Humanities, Bergen, Norway, 1986.

[9] S. Johansson, G. Leech, and H. Goodluck. *Manual of Information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital Computers*. Department of English, University of Oslo, Oslo, 1978.

[10] D. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchman, and N. Morgan. Using a stochastic context-free grammar as a language model for speech recognition. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 189–192, Detroit MI, USA, 1995.

[11] A. L. Koerich, Y. Leydier, R. Sabourin, and C. Y. Suen. A hybrid large vocabulary handwritten word recognition system using neural networks and hidden Markov models. In *8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 99–104, Niagra-on-the-Lake, Canada, Aug. 2002.

[12] C.-L. Liu, H. Sako, and H. Fujisawa. Integrated segmentation and recognition of handwritten numerals: Comparison of classification algorithms. In *8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 369–374, Niagra-on-the-Lake, Canada, Aug. 2002.

[13] U. Mahadevan and S. N. Srihari. Parsing and recognition of city, state, and zipcodes in handwritten addresses. In *5th Int. Conf. on Document Analysis and Recognition*, pages 325–328, Bangalore, India, 1999.

[14] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15:65–90, 2001.

[15] U.-V. Marti and H. Bunke. The IAM-database: an English sentence database for off-line handwriting recognition. *Int. Journal on Document Analysis and Recognition*, 5:39–46, 2002.

[16] R. Srihari, S. Ng, C. Baltus, and J. Kud. Use of language models in on-line sentence/ phrase recognition. In *3rd Int. Workshop on Frontiers in Handwriting Recognition*, pages 284–294, Buffalo NY, USA, 1993.

[17] S. Uchida and H. Sakoe. An off-line character recognition method employing model-dependent pattern normalization by an elastic membrane model. In *5th Int. Conf. on Document Analysis and Recognition*, pages 499–502, Bangalore, India, 1999.

[18] A. Vinciarelli, S. Bengio, and H. Bunke. Offline recognition of large vocabulary cursive handwritten text. In *7th Int. Conf. on Document Analysis and Recognition, Edinbourgh, Scotland*, volume 2, pages 1101–1105, 2003.

[19] M. Zimmermann and H. Bunke. Automatic segmentation of the IAM off-line handwritten English text database. In *16th Int. Conf. on Pattern Recognition*, volume 4, pages 35–39, Quebec, Canada, Aug. 2002.

[20] M. Zimmermann and H. Bunke. Hidden Markov model length optimization for handwriting recognition systems. In *8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 369–374, Niagra-on-the-Lake, Canada, Aug. 2002.