Parsing N-Best Lists of Handwritten Sentences

Matthias Zimmermann, Jean-Cédric Chappelier*, and Horst Bunke

zimmerma@iam.unibe.ch

Department of Computer Science, University of Bern

Neubrückstrasse 10, CH-3012 Bern, Switzerland

*Artificial Intelligence Laboratory

Swiss Federal Institute of Technology

INR (Ecublens), CH-1015 Lausanne, Switzerland

Organization

1. Introduction
2. Resources
3. Recognition of Handwritten Sentences
4. Parsing Sentences
5. Combination Scheme
6. Experiments and Results
7. Conclusions and Future Work

Introduction

The work presented in this talk addresses the task of recognition of general handwritten text.

- Offline recognition
- Large variation of writing styles and writing instruments
- Segmentation problem
- Lack of task-specific knowledge
- Large vocabulary

Introduction

The proposed parsing of n-best lists supports the following goals:

- Improvement of sentence and word recognition rate
- Production of parse trees
- Assignment of word tags (grammatical categories)

Resources

Handwritten sentence images

- Segmented IAM database
- Automatic extraction of sentences
- Generation of corresponding transcriptions

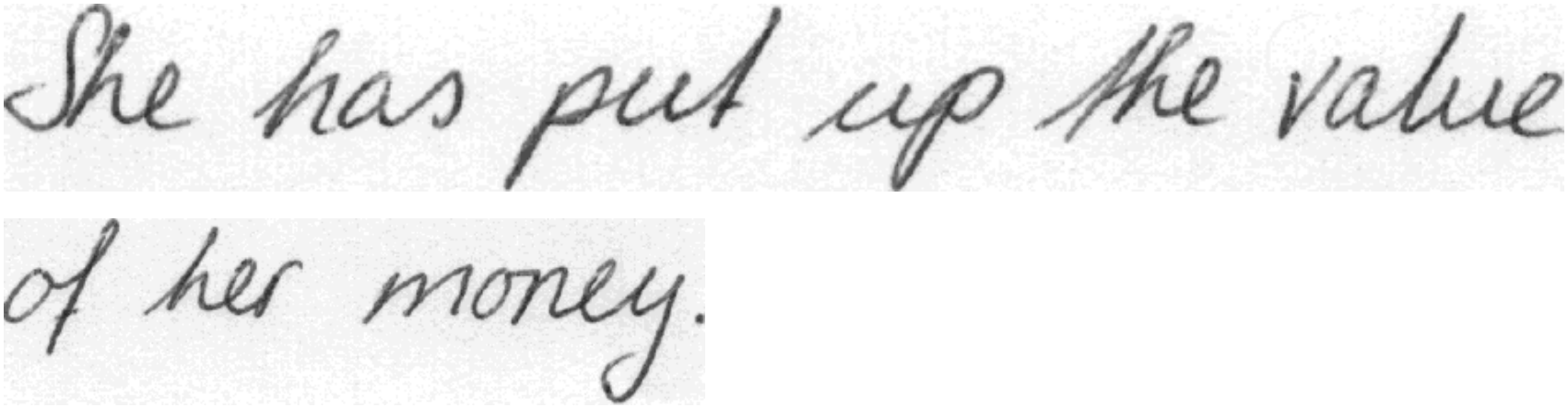
Linguistic resources

- Tagged LOB corpus for bigram models
- LPC for the stochastic context-free grammar, SCFG

IAM Database

- More than 1,500 scanned pages of handwritten text
- Material from over 600 different writers
  - 95,000 correctly segmented words
  - 13,000 lines of handwritten text
  - Over 5,000 complete sentences
- Covering a vocabulary of over 12,000 words

A Handwritten Sentence



She|has|put|up|the|value|of|her|money|.

Tagged LOB Corpus

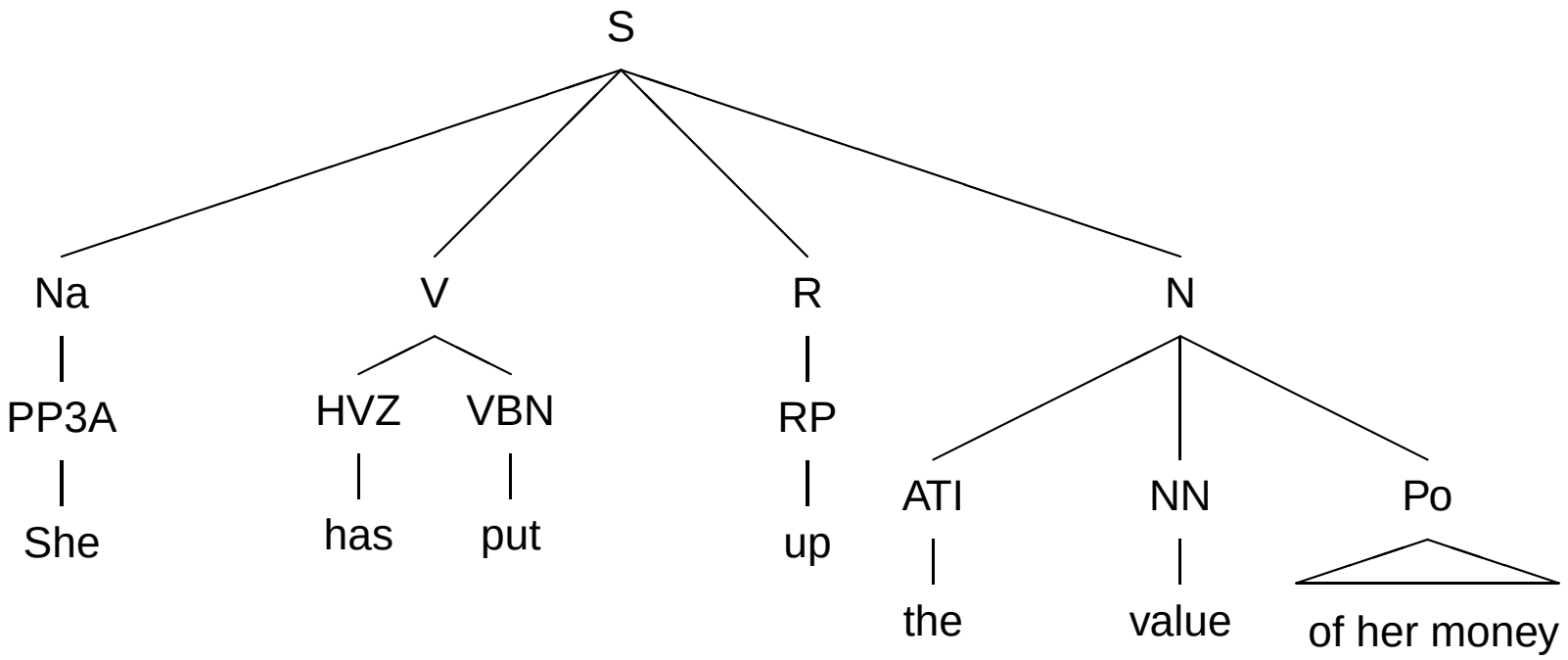
- Covers the LOB corpus
  - 1,000,000 word collection of British English text printed in 1961
  - 500 text samples of 2,000 words each
  - 15 text categories including press, general fiction, etc.
- Explicit segmentation of the text into words
- A grammatical tag is assigned to each word
- 167 tags are defined

Lancaster Parsed Corpus

- Covers a subset of 12,000 sentences from the LOB
- Parse trees in the form of bracketed sentences
- Only few long sentences
  - Average length 11.4 words
  - 19 words for the complete LOB

A Bracketed Sentence

$[S$ $[Na$ $[PP3A$ She $]][V$ $[HVZ$ has $][VBN$ put $]][R$ $[RP$ up $]][N$ $[ATI$ the $][NN$ value $][Po$ $[INO$ of $][N$ $[PP\$$ her $][NN$ money $]]]]]



Recognition of Handwritten Sentences

- Image normalization (skew, slant, position and height)
- Feature extraction (a sliding window)
- HMM for each character
  - Linear topology
  - Bakis modeling
  - Multi Gaussian
- Bigram language model
- Viterbi decoding

N-Best List

Rank	Score	Candidate sentence
1	23923.6	She has put up the value other money .
2	23921.8	She has put up the value of her money .
3	23890.3	She had put up the value other money .
4	23888.4	She had put up the value of her money .
5	23854.3	She has put up the value at her money .
...	...	...

back

Parsing Sentences

- Parsing of SCFG
- Bottom-up chart parser based on CYK
- Early like elements
- Most probable parse
- Probability of most probable parse

Example Productions

Non-lexicalized rules

$$
\begin{array}{lcl}
\text{S} & \rightarrow & \text{Na V R N} \\
\text{Na} & \rightarrow & \text{PP3A} \\
\text{V} & \rightarrow & \text{HVZ VBN}
\end{array}
$$

Lexicalized rules

$$
\begin{array}{lcl}
\text{PP3A} & \rightarrow & \text{She} \\
\text{HVZ} & \rightarrow & \text{has} \\
\text{VBN} & \rightarrow & \text{put}
\end{array}
$$

Parse Probabilities

Rank	Parse Prb.	Candidate sentence
1	7.691e-23	She has put up the value other money .
2	4.629e-20	She has put up the value of her money .
3	2.631e-22	She had put up the value other money .
4	1.584e-19	She had put up the value of her money .
5	1.125e-21	She has put up the value at her money .
...	...	...

back

Combination Scheme

Recognition scores and parse probabilities are combined to a sentence score using the parse scale factor γ

$$recognition\ score + \gamma \log(parse\ probability)$$

the sentence score is used to reorder the n-best list

System	Top choice
Baseline	She has put up the value other money .
W. Parsing	She has put up the value of her money .

Experiments and Results

- 4,000 lines of text (30,000 words) to train the handwriting recognizer
- A closed vocabulary of 8,800 words
- The test set contains 67 sentences written by 52 individuals
- The SCFG contains 11,760 non-lexicalized productions
- The parse scale factor was optimized using six-fold cross-validation on the 67 sentences

Results

System	Sentence Rate	Word Rate
Baseline	4.6%	67.47%
With Parsing	6.0%	68.32%

- The improvement of the word recognition rate is significant on the 95% level
- For the correct sentences the tagging rate was 100%

Conclusions and Future Work

- A combination of a HMM based recognition system and a probabilistic bottom-up chart parser has been investigated
- A SCFG for general English text has been extracted from bracketed sentences
- By the incorporation of parse probabilities a significant improvement of the word recognition rate has been achieved
- A very high tagging rate for correct sentences was observed

Future Work

- Future work includes a a better training of the handwriting recognition module, the use of validation sets and larger test sets
- The probabilistic parsing of larger N-best lists will be investigated

Latest Results