Sentence Boundary Detection for Handwritten Text Recognition

Matthias Zimmermann

International Computer Science Institute (ICSI), Berkeley, USA
zimmerma@icsi.berkeley.edu

Abstract

Sentence segmentation is an essential first step for many natural language processing applications such as parsing, summarization, topic segmentation, etc.

This work compares two machine learning approaches and their combination for the detection of sentence boundaries in the output of handwritten text recognition systems:

- hidden-event language models (**HELM**) [1]
- maximum entropy (**MaxEnt**) [2]

The proposed approach is validated by experiment using a state-of-the-art recognition system [3] on the IAM Database [4].

Task

For each position in a word stream the system has to decide if a sentence boundary token <s> has to be inserted after the current word.

1 *The summonses say they are " likely to persevere in such unlawful conduct . "* <s> *They ...*
2 *" It comes at a bad time , " said Ormston .* <s> *" A singularly bad time ...-*

The above example illustrates why this can be a challenging task even when the output of the recognizer would be 100% correct.

The presence of recognition errors significantly increases the task complexity as punctuation marks may be omitted or inserted in incorrect locations.

This work was supported by the Swiss National Science Foundation through the research network IM2.

Method

For the **HELM** approach [1] a statistical word *n*-gram LM is first trained on all available texts including sentence boundary tokens <s> that are treated as normal words. In the decoding step the HELM estimates a posterior probability of having a sentence boundary token between any two words of the text to be segmented.

In the second approach posterior probabilities are estimated by a **MaxEnt** classifyer based on a set of locally obtained binary features:

...	12	tomorrow	.	Mr.	Michael	...
	w_{i-2}	w_{i-1}	w_i	w_{i+1}	w_{i+2}	

Feature class	Features at Position *i*
Word	w_{i-2}(12), w_{i-1}(tomorrow), w_i(.), w_{i+1}(Mr.), w_{i+2}(Michael)
Bigram	b_{i-1}(12_tomorrow), b_i(tomorrow_.), b_{i+1}(._Mr.), b_{i+2}(Mr._Michael)
Captial	c_l(0a), c_r(AA), c_3(a.A), c_5(0a.AA)
Line Break	l(none)

For the **combination scheme** the MaxEnt posteriors are integrated as an additional knowledge source into the decoding step of the HELM according to [1].

Evaluation of the segmentation systems is based on the following performance metrics:

Reference	w.w.w|w|w.w|w.w.w.w|
System	w.w|w.w.w.w|w.w.w.w|
Counts	...f.m.m...c.....c

NIST: 75%, (1 false alarm + 2 misses) / 4
Recall: 50%, 2 correct boundaries out of 4
Precision: 67%, 2 correct boundaries out of 3
F-Measure: 57%, 2 * 0.5 * 0.67 / (0.5 + 0.67)

Experiments

The development and test set of [3] are used here as cross-validation sets:

Name	Sentences	Words	WER (Word Error Rate)
Set 1	200	3995	29.5%
Set 2	200	3936	23.5%

Effect of the MaxEnt Feature classes (Set 1):

Features	NIST	F-Measure
Word	60.5%	71.3%
W + Bigram	54.5%	72.3%
W + B + Capital	41.5%	74.8%
W + B + C + Line Break	35.0%	80.1%

Evaluation of System Performance (Sets 1+2):

System	NIST	F-Msr.	Recall	Precision
HELM	43.0%	77.6%	74.5%	81.0%
MaxEnt	41.3%	76.9%	69.3%	86.9%
Combination	40.5%	79.8%	80.0%	79.6%

Conclusion

Sentence boundary detection for handwritten text recognition can reach accuracy rates of 80% (F-Measure) for recognition noisy recognizer output (WER 26%).

[1] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G.Tür , "Prosody-based segmentation of speech into sentences and topics", Speech Communication, vol. 32(1-2), pp. 127-154, 2000

[2] J. C. Reynar and A. Ratnparkhi, "A maximum entropy approach to identifying sentence boundaries", 5th Int. Conference in on Applied Natural Language Processing, pp. 16-19, Washington, USA, 1997

[3] M. Zimmermann and H. Bunke, "Optimizing the integration of statistical language models in HMM based offline handwritten text recognition", 17th Int. Conference on Pattern Recognition, vol. 2, pp. 541-544, Cambridge, England, 2004

[4] 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, vol 5, pp.39-46, 2002