Degraded Text Recognition Using Word Collocation

Tao Hong and Jonathan J. Hull

Center of Excellence for Document Analysis and Recognition
Department of Computer Science
State University of New York at Buffalo
Buffalo, New York
taohong@cs.buffalo.edu hull@cs.buffalo.edu

ABSTRACT

A relaxation-based algorithm is proposed that improves the performance of a text recognition technique by propagating the influence of word collocation statistics. Word collocation refers to the likelihood that two words co-occur within a fixed distance of one another. For example, in a story about water transportation, it is highly likely that the word "*river*" will occur within ten words on either side of the word "*boat*." The proposed algorithm receives groups of visually similar decisions (called neighborhoods) for words in a running text that are computed by a word recognition algorithm. The position of decisions within the neighborhoods are modified based on how often they co-occur with decisions in the neighborhoods of other nearby words. This process is iterated a number of times effectively propagating the influence of the collocation statistics across an input text. This improves on a strictly local analysis by allowing for strong collocations to reinforce weak (but related) collocations elsewhere. An experimental analysis is discussed in which the algorithm is applied to improving text recognition results that are less than 60 percent correct. The correct rate is effectively improved to 90 percent or better in all cases.

1 Introduction

The recognition of images of text is a difficult problem, especially when the images are degraded by noise such as that introduced by photocopying or facsimile transmission. Recently, methods for improving the quality of text recognition results have focused on the use of knowledge about the language in which the document is written([5, 6]). These techniques often post-process the results of a word recognition algorithm that provides various alternatives for the identity of each word that are called its *neighborhood*. The objective of the language model is to choose the alternatives for words that make sense in the context of the rest of the text.

Word collocation data is one source of information that has been investigated in computational linguistics and that has been proposed as a useful tool to post-process word recognition results([1, 2, 8]). Word collocation refers to the likelihood that two words co-occur within a fixed distance of one another. For example, it is highly likely that if the word "*boat*" occurs, the word "*river*" will also occur somewhere in the ten words on either side of "*boat*."

Previous work in using word collocation data to post-process word recognition results has

Initial Word Neighborhoods

position	1	2	3	4	5	6	7	8
top1	Places	snow	me	whale	Kong	Kong	it	!
top2	Please	slow	we	where	Hong	Hong	is	
top3	Pieces	show	mo	chore	Hung	Kung	Is	

Word Neighborhoods After Iteration 1

position	1	2	3	4	5	6	7	8
top1	Places	show	me	where	Hong	Kong	is	!
top2	Please	slow	we	whale	Kong	Hong	it	
top3	Pieces	snow	mo	chore	Hung	Kung	Is	

Word Neighborhoods After Iteration 2

position	1	2	3	4	5	6	7	8
top1	Please	show	me	where	Hong	Kong	is	!
top2	Places	slow	we	whale	Kong	Hong	it	
top3	Pieces	snow	mo	chore	Hung	Kung	Is	

Figure 1: An example of the relaxation process (the sentence to be recognized is "*Please show me where Hong Kong is !* "

3 Experiments and Analysis

The recognition data used in the experiments were generated from the Brown Corpus and Penn Treebank databases. These are large corpora that together contain over four million words of running text. The Brown corpus is divided into 500 samples of approximately 2000 words each([7]). The part of the Penn Treebank used here is the collection of the articles from the *Wall Street Journal* that contains three million words. We used the frequency of a word pair to measure its collocation strength. There are totally 1,200,000 word pairs after training. Several such word pairs are listed below as examples:

the	doctor	64
a	doctor	27
doctor	and	8
doctor	was	8
doctor	who	7
his	doctor	6
doctor	bills	4
ward	doctor	1

The word collocation data used in the experiments discussed below was calculated from the combined corpus, which is referred to as *WC**.

Neighborhoods were generated for each of the 70,000 unique words in the combined corpus by the following procedure. Digital images of the unique words were generated from their ASCII equivalents by first converting them to an 11 pt. Times Roman font in postscript with the Unix command *ditroff*. The postscript files were then transformed into raster images with the *ghostscript* system.

Neighborhoods were generated for each word by first calculating a feature vector for the word known as the stroke direction feature vector ([3]). The neighborhoods for each dictionary word were then calculated by computing the Euclidean distance between its feature vector and the feature vectors of all the other dictionary words and sorting the result. The ten words with the smallest distance values were stored with each dictionary word as its neighborhood.

To mimic the performance of a word recognition technique in the presence of noise, the neighborhoods were corrupted. An assumed correct rate in each position in the neighborhood was given. For example, the top choice might be 80 percent correct, the second choice 10 percent correct, and so on.

The noise model was applied to the text by calling a uniform random number generator for each word in the passage and scaling the result between zero and one. The correct rate distribution was then used to select the position in the neighborhood into which the correct word was moved. Thus, in the above example, 80 percent of the time the correct word would remain in the top position, 10 percent of the time it would be moved into the second position, and so on.

3.1 Testing Data

We randomly selected five articles from the Brown Corpus as the testing samples. They are *A06*, *G02*, *J42*, *N01* and *R07*. **A06** is a collection of six short articles from the *Newark Evening News*. *G02* is from an article "*Toward a Concept of National Responsibility* " from *The Yale Review*. **J42** is from a book "*The Political Foundation of International Law.*" *N01* is a chapter from an adventure fiction "*The Killer Marshal.*" *R07* is from a humor article "*Take It Off*" from *The Arizona Quarterly*. Each text has about 2000 words. There are totally 10,280 words in those testing samples. For each word in those texts, the top10 word candidate lists were generated. We simulated a word recognition algorithm based on different performance models. The performance models used here have top1 correct rates of 55%, 65%, 70%, 75%, 80%, 85%,

	Article					Average
	A06	G02	J42	N01	R07	
# of words	2040	2075	2078	2021	2066	2056
top1	73.6%	74.7%	75.1%	76.6%	76.1%	75.2%
top2	89.0%	88.7%	91.8%	89.8%	90.1%	89.9%
top3	94.6%	94.6%	95.7%	95.5%	95.8%	95.2%
top4	96.9%	97.1%	97.8%	97.7%	97.5%	97.4%
top5	98.0%	98.0%	98.7%	98.8%	98.5%	98.4%

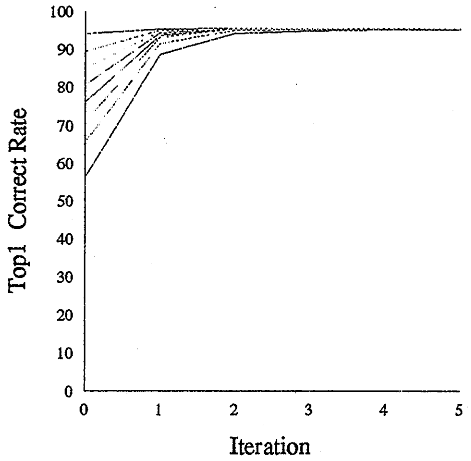

Figure 2: Result of relaxation using WC*

Initial	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5
56.2%	81.5%	86.0%	86.8%	87.1%	87.1%
65.5%	85.0%	87.9%	88.4%	88.5%	88.5%
71.0%	87.1%	89.2%	89.9%	89.4%	89.3%
76.0%	88.4%	89.9%	89.9%	90.0%	89.9%
80.6%	89.4%	90.4%	90.3%	90.4%	90.3%
85.0%	90.3%	90.8%	90.7%	90.8%	90.7%
89.5%	90.9%	91.2%	91.1%	91.2%	91.1%
94.1%	91.6%	91.6%	91.5%	91.6%	91.5%

Table 3: Correct percentage of top 1 by relaxation algorithm based on WC

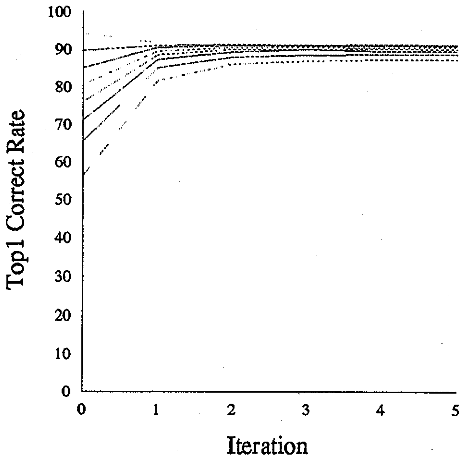

Figure 3: Result of relaxation using WC

4 Conclusions and Future Directions

In this paper a relaxation algorithm was described that used word collocation information to improve text recognition results. The experimental results were promising and showed that the correct rate at the top choice of a word recognition algorithm could be improved from 56 to 87 percent correct. The performance gap between $WC*$ and WC suggested that we should collect word collocation data from larger and more balanced English corpora. Analysis of the remaining errors showed that many of them could be corrected by using a larger window size and special strategies for processing proper nouns. Modifications of the ranking function will also be considered.

The relaxation algorithm currently works as one part of our degraded text recognition system. There are two types of linguistic constraints used in the system. One is local word collocation under statistical language modelling. Another is global structural constraints carried by English grammar. Visual global contextual information available inside a text page is also being considered for integration with the linguistic knowledge sources to further improve the performance of degraded text recognition.

References

[1] Henry S. Baird, Private communication about the use of word collocation to improve OCR results, February, 1989.

[2] Kenneth Ward Church and Patrick Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29, 1990.

[3] Tin Kam Ho, Jonathan J. Hull and Sargur N. Srihari, "A Word Shape Analysis Approach to Lexicon Based Word Recognition," in *Pattern Recognition Letters*, Vol. 13, pp. 821-826, 1992.

[4] Tao Hong and Jonathan J. Hull, "Text Recognition Enhancement with a Probabilistic Lattice Chart Parser," in *Proceedings of the Second International Conference on Document Analysis ICDAR-93*, Tsukuba, Japan, October 20-22, 1993.

[5] Jonathan J. Hull, Siamak Khoubyari and Tin Kam Ho, "Word Image Matching as a Technique for Degraded Text Recognition," in *Proceedings of 11th IAPR International Conference on Pattern Recognition*, The Hague, The Netherlands, pp. 665-668, 1992.

[6] Jonathan J. Hull, "A Hidden Markov Model for Language Syntax in Text Recognition," in *Proceedings of 11th IAPR International Conference on Pattern Recognition*, The Hague, The Netherlands, pp.124-127, 1992.

[7] H. Kucera and W. N. Francis, *Computational Analysis of Preset-day American English*, Brown University Press,1967.