Keyword Selection from Word Recognition Results using Definitional Overlap

Paul Filipski 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 14228
hull@cs.buffalo.edu

Abstract

A method is presented to locate a set of potential keywords for a given document in the output of a word recognition algorithm. A clustering step locates words of significant length that occur several times. A word recognition algorithm is applied to these words to generate groups of visually similar alternatives for each image. A simulated annealing algorithm is then used to choose a set of keywords that contains at most one representative from each neighborhood such that an inter-word compatibility measurement is maximized. The compatibility measure is based on the similarity of subject and the definitional overlap of two words as measured from a dictionary. Experimental results are presented that illustrate the ability of the technique to operate in the presence of noise.

1 Introduction

The selection of a set of keywords that characterize the subject of a passage of text is important for document classification and information retrieval as well as for the recognition of the remainder of the text. This is shown in experiments where the subject of a document is used to restrict the vocabulary of a word recognition algorithm [3]. Accurately determined keywords can also be used as "islands" to drive recognition of the surrounding text.

Keyword selection can be difficult in the presence of the noise that occurs when degraded images are processed by a word recognition algorithm. Such techniques typically produce *neighborhoods* or groups of possible choices for each word. The top choice may have a low correct rate when degraded images are input. However, the correct word decision may still be contained in the neighborhood [2].

There are several ways to choose keywords and determine the subject for a document. The *vector space model* chooses words that are frequent within a document but occur in a small number of documents in a collection [8]. Another method that characterizes the topic of a document is the frequent occurrence of words from the same subject [9]. The subject codes used in this work were extracted from the Longman Dictionary of Contemporary English (LDOCE). They are four-letter identifiers chosen from a set of 2567 possibilities that segment the dictionary into subsets of se-

Third Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, April 11-13, 1994.

mantically related word senses. For example, words tagged as MHZS such as *cross-section, extrapolate, percentile,* and *questionnaire* are related to statistics.

The LDOCE subject codes as well as a method of definitional overlap have been used to improve the performance of a word recognition algorithm [7]. Word decision alternatives were chosen that maximized the occurrence of subject codes within a sentence. Definitional overlap was used to choose decisions for words in a sentence that maximized the intersection between the vocabularies used in their dictionary definitions.

Definitional overlap has also been used to solve the problem of word sense disambiguation in which the objective is to choose the correct sense for the words in a sentence by maximizing the overlap between their definitions. A simulated annealing algorithm has been proposed for this task [1].

A problem with both applications of definitional overlap (word recognition and word sense disambiguation) is that there are no words in common between the definitions for many words. This makes it difficult to successfully apply the method to running text where it would be necessary to draw fine semantic distinctions. This suggests that definitional overlap may be most useful in locating words that are strongly related to one another such as keywords.

This paper proposes an algorithm for selecting keywords from the output of a word recognition method. A clustering technique is applied that locates word *images* that occur repeatedly in a document [4]. Potential keywords are identified as clusters that contain several instances of long words (*i.e.* words that contain more than three or four characters). The recognition neighborhoods for these words are then processed by a simulated annealing algorithm that uses an energy function based on both Longman dictionary subject codes as well as definitional overlap to measure the compatibility between the proposed keywords. At most one member from each recognition neighborhood is included in the set of keywords that minimizes the energy function. Non-keywords that were erroneously output by the clustering algorithm are not included.

The effect of the algorithm is to simultaneously identify several instances of the selected keywords with high reliability throughout a document. This overcomes recognition errors that may have placed the correct words in lower positions in their neighborhoods and provides word decisions with high confidence that can be used as seeds for further processing. A technique based on word collocation around the selected keywords is proposed to improve the recognition results for the other words in the document.

The rest of the paper discusses the proposed method. A statement of the algorithm is provided and explained. An example of the results calculated by the technique is presented and used to illustrate its advantages. Experimental results are discussed that demonstrate the ability of the technique to operate in the presence of noise. The keywords chosen by this method are compared to those chosen by the vector space model and it is shown that better than 90 percent of the keywords found from the clean ASCII text by the vector space model are located by the proposed method. Extensions of the technique that augment the lexicon with word collocation data are discussed.

2 Algorithm Description

The proposed algorithm is illustrated in Figure 1. A segmented input document is provided to an algorithm that locates clusters of equivalent word images. This technique has been demonstrated to have high accuracy for degraded document images and the cluster centers (average of the

images in the cluster) have been shown to improve the performance of text recognition algorithms [4]. The number of images in each cluster and their length are used to locate potential keywords.

A word recognition algorithm is applied to the cluster centers. The neighborhoods of words from a given list that are visually similar to each potential keyword are then determined. There are several different types of word recognition algorithm that could be used, e.g., wholistic or segment-recognize-postprocess. The algorithm proposed in this paper makes no assumption about the type of word recognition technique used. The only assumption is that the visual characteristics of an input word are the only features used to calculate the neighborhoods.

A simulated annealing algorithm is then applied to the neighborhoods to search for the set of words, containing at most one entry from each neighborhood, that are strongly related to each other. The relationship between words depends on data extracted from the Longman Dictionary of Contemporary English (LDOCE) and a technique for measuring the semantic similarity of two words known as definitional overlap.

2.1 The Longman Dictionary

The LDOCE is a dictionary designed for non-native English speakers that defines 34,329 different words using 71,416 different senses. Each definition contains one or more senses and each sense contains a textual description as well as a part-of-speech and a subject code. The textual descriptions are composed from a base vocabulary of about 2000 words. The 14 parts of speech used in the LDOCE are N, V, ADJ, ADV, PREP, DETERMINER, PRON, INTERJ, PREFIX, CONJ, COMB, PREDETERMINER, INDEFINITE, DEFINITE.

The subject codes in the LDOCE provide semantic groupings for words. The first two characters of a subject code provide a general classification for the word, and the last two a more specific categorization. For example, the code GBZB refers to bookmaking and contains words such as *bookmaker* and *tote.* The related code GBZR includes horse-racing terms that deal with gambling such as *derby, jockey* and *paddock.* Both of these codes are more specific versions of gambling, GB--, which includes the words *casino, craps* and *gamble.*

Also, a large number of words in the dictionary cannot be classified into a single semantic group. These words are assigned the null subject code, ----. Some examples are *accomplish, fear* and *typical.*

2.2 Definitional Overlap

A computer-readable form of the LDOCE provides the data for computing definitional overlap. The overlap is defined as the portion of the textual description that is in common between the senses for two different words. For example, one sense for the word nationalism is "desire by a racial group to form an independent country." This sense has the POS tag N and the subject code PL--. One sense for the word sovereignty is "the quality of being an independent self-governing country," also with POS tag N and subject code PL--. The definitional overlap for these senses is "independent country," the two words that appear in both definitions.

A score is assigned to the words in the overlap based on how frequently the words occur in other definitions. For each word, w, we compute N_w, the number of senses whose definition contains w. Then, the score for each word is $1 - N_w/N_{\max}$, where $N_{\max}$ is the number of sense definitions in which the most frequently occuring word appears. This assigns high weights to words that discriminate different senses in much

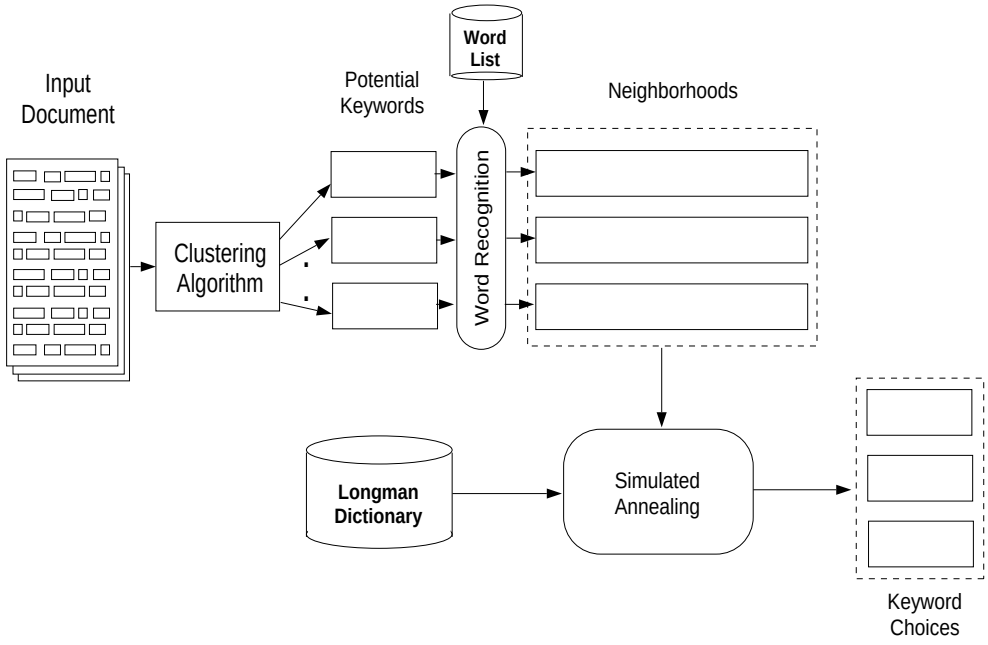

Figure 1: Diagram of algorithm.

the same way that the vector space model assigns high weights to words that discriminate between different documents.

The composite score for two words w_1 and w_2 are the senses $s_{1,i}$ and $s_{2,j}$ that maximize equation 1 in Figure 2.

In this equation, the POS weighting function ($POSwt$) returns the weight for the part of speech for either $s_{1,i}$ or $s_{2,j}$ that has the smallest value in Table 1. This has the effect of preferring matches between senses that are both nouns (because keywords are often nouns) and inhibiting matches between senses that are other parts of speech. The subject weighting function ($SUBwt$) shown in Table 1 assigns high values to senses with identical subject codes and lower values based on various degrees of mis-match.

In the above example, the word independent has a score of 0.999 and country has a score of 0.985. The subject codes are identical, and both words are nouns, so the overlap score for nationalism and sovereignty is 1.984.

2.3 Simulated Annealing

Each of the potential keywords in an input document can be thought of as a variable whose domain of legal values is the recognition neighborhood for that word. The problem of choosing at most one keyword from each neighborhood is then a search for the set of words that minimize an energy function.

The energy function E for a given configuration C (one choice for each potential keyword) is calculated as shown in equation 2 in Figure 2. The sum of the composite scores for each pair of words x, y in C is calculated. This is multiplied by a weighting factor determined by the probability of correctness p_x assigned to word x by the recognizer as well as a scaling factor s. The effect of s is to linearly scale the recognition probabilities between $(1 - \frac{1}{s})$ and 1.

The energy function provides a measure of the semantic similarity of the given choices for each word. A small value for this function indicates that the current choices form a coherent set of words, all related in meaning. A large value implies that the choices are less well-related.

Simulated annealing works by repeatedly considering random modifications to the current configuration (different assignments for the identity of each potential keyword). A new configuration is accepted probabilistically based on its effect on E. A parameter T represents the "temperature" of the system. This parameter controls the degree of fluctuation allowed in the energy from one iteration to the next.

If a proposed random modification results in a lower energy, it is accepted with probability one. Otherwise, the probability of accepting the change is given by $e^{-\frac{\Delta E}{T}}$. Allowing the search to accept a change that increases the system's energy prevents simulated annealing from becoming stuck in a local minimum.

The algorithm reduces the value of T according to a predetermined schedule. Early values of T are high so it is more likely that a non-decreasing change will be accepted. As T is decreased, it becomes less likely that an energy-increasing change will be accepted. This controlled reduction of the acceptable energy change locates a global minimum rather than a local one. In our experiments T was initially set to 5.0 and reduced by 5 percent with each iteration.

2.4 Example

An image of a page from an article in the journal *Applied Psychology* was used to test the system. Each word image in this passage was input to a clustering algorithm that labeled it as a potential keyword based on its length and the number of times the word appears in the document [4]. The potential keywords were then passed

$$Score(w_1, w_2) = \max_{i,j} \left\{ POSwt(s_{1i}, s_{2j}) \cdot SUBwt(s_{1i}, s_{2j}) \cdot \sum_{w \in s_{1i} \cap s_{2j}} (1 - N_w/N_{max}) \right\} \quad (1)$$

$$E(C) = \left[\left(\sum_{(x,y) \in C} Score(x, y) \right) \cdot \prod_{x \in C} \left(1 - \frac{1 - p_x}{s} \right) \right]^{-1} \quad (2)$$

Figure 2: Equations describing the simulated annealing energy function.

Table 1: Weighting functions for POS tags and subject codes.

POS tags		*Subject Codes*	
noun	1.0	identical, non-null	1.0
adjective	0.5	identical, null	0.7
adverb	0.5	two characters agree	0.9
verb	0.1	different	0.2
other	0.0		

to a recognition algorithm that generated a neighborhood for each word based on its shape. The neighborhoods were then corrupted to simulate errors in recognition. A uniform random number generator was used to move the top choice for each word into a lower position in the neighborhood according to a fixed distribution. For example, the distribution $(0.8, 0.07, 0.05, 0.02, 0.02, 0.01, 0.01, 0.01, 0.005, 0.005)_{10}$ applied to a ten word neighborhood would retain the correct choice in the first position in 80 percent of all cases and move it to the second position seven percent of the time, the third position five percent, and so on.

Table 2 contains the word recognition neighborhoods for the potential keywords from this article. The simulated annealing search was set up to try 250 random variations for each word at each of 167 temperatures in a cooling schedule. The initial configuration was generated by selecting one random choice from each neighborhood. Table 3 shows the initial state of the algorithm, the state halfway through the search, and the final state. In this table, the number in parentheses after each word is the position of that choice in the word recognition neighborhood. This run of the algorithm used the word recognition simulator described above.

The example illustrates the ability to focus on a meaningful set of keywords after starting with a random set of choices, each of which could have been a keyword for some passage of text. In the final result, nine out of ten of the correct keywords are found. This is a significant improvement over the six keywords found by the recognition algorithm alone. The single error (*infidelity* substituted for *reliability*) was caused by a superficial similarity in definitions in the LDOCE. This can be repaired with by adding information from another lexical database.

After the search has completed, the strength of the match for each word is measured by adding the overlap scores for that

Table 2: Word recognition simulation for sample document.

reliability	*employee*	*scores*	*scales*	*delinquency*
faithfully	**employee**	**scores**	**scales**	bankruptcy
visibility	employer	scares	stares	incumbency
McNally	employed	scorer	exiles	**delinquency**
liability	ex-player	scenes	crates	headquarter
foldability	employees	mores	series	delinquents
reliability	earphones	comes	states	despondency
mid-July	implores	scams	scabs	instant-replay
wistfully	explores	sauces	aisles	soft-currency
infidelity	emptiness	serves	stores	introductory
artificially			males	infrequently

theft	*measure*	*sample*	*performance*	*correlations*
theft	reassure	**sample**	performances	**correlations**
their	**measure**	simple	persuasions	concoctions
draft	romance	temple	penetrations	translations
melt	masons	ample	pantomime	convictions
chefs	masseur	compile	perchlorate	circulations
there	concurs	cripple	**performance**	constrictors
thrift	kicking	supple	promotions	conclusions
these	massacre	sculpts	key-someone	interactions
cheat	despite	scripts	personal-care	connections
shaft	outside		predominate	contaminate

Table 3: Example run of system on sample document.

	Initial state	*Halfway*	*Final state*
Temperature	5.000	0.067	0.001
Energy	5.456	0.870	0.346
reliability	infidelity(9)	infidelity(9)	infidelity(9)
employee	implores(7)	employee(1)	employee(1)
scores	serves(10)	scores(1)	scores(1)
scales	series(5)	scales(1)	scales(1)
delinquency	delinquency(3)	delinquents(5)	delinquency(3)
theft	chefs(5)	chefs(5)	theft(1)
measure	kicking(7)	massacre(8)	measure(2)
sample	temple(3)	sample(1)	sample(1)
performance	persuasions(2)	performance(6)	performance(6)
correlations	translations(3)	convictions(4)	correlations(1)

word combined with the other words in the final configuration. For the example run shown in Table 3, the average match value is 1.76. The maximum is 3.06 for *performance*, and the minimum is 0.20 for *sample*. This implies that the word *sample* contributes less to the overall system, and could be rejected as a non-keyword.

3 Experiments

The algorithm was implemented as described and simulated in the following manner. Images of passages from the Brown Corpus were generated in an 11 pt. Times Roman font with a postscript to bitmap conversion algorithm. The Brown Corpus is a collection of 500 ASCII text passages of about 2000 words each that are designed to represent edited American English [5]. Documents from the Brown Corpus that were used to test the algorithm are listed in Table 4.

The word recognition simulator was applied to the potential keywords found in G02 and various levels of noise were introduced. The simulated annealing algorithm was run multiple times on G02 with the word recognition performance set to different levels. The results of these runs are shown in Table 5. Each column of this table presents the results of the algorithm when the simulated word recognition algorithm performs at the given level. For example, when word recognition performance is set at 80 percent, the keyword selection algorithm makes only one error. As the word recognition results become degraded, the keyword selection algorithm still performs acceptably, making only four errors when the top choice of the word recognition algorithm is correct 30 percent of the time.

It is interesting to observe that two of the errors made by the system, choosing *amnesty* for *concept* and *protocols* for *principle*, result in choices that are still very well-related with the other input words.

The performance of the keyword selection system has been compared to the performance of the vector space model run on an ASCII version of the documents. For G02, the vector space model identifies seven keywords: *sovereignty, responsibility, nationalism, principle, nations, concept* and *national*. The proposed algorithm correctly chooses all seven of these when given word recognition input that is 80 percent correct in its top choice. Also, when the performance of word recognition is lowered to 30 percent correct, the algorithm still correctly identifies five of the seven keywords. For J61, the vector space model identifies five keywords: *frieze, fresco, cleaning, plaster* and *sketches*. The keyword selection algorithm chose four of these.

3.1 Future Directions

The use of the keywords located by the proposed technique in an island-driving approach to improve recognition will be fully explored. A preliminary experimental investigation was conducted of the potential improvement possible in recognition performance by application of word collocation data. Collocations for each of the chosen keywords shown in Table 3 were collected from the entire Brown Corpus using various window sizes. The nouns and the adjectives that occurred in the window around each instance of a keyword were saved.

The ASCII for the Applied Psychology article was filtered to remove stop words. This reduced the 3267 words in the original file to 1856 non-stop words. The collocations from the Brown Corpus were matched against this data and the number of words in common were determined. This illustrates the upper bound possible in detecting and correcting word recognition errors. The results shown in Table 6 illustrate the performance that can be achieved when the noun collocations are used alone or when

Table 4: The Brown Corpus samples used in experiments.

sample	title, author, source, year
G02	*Toward a Concept of National Responsibility*, by A.S. Miller, The Yale Review, 1961
J61	*Completing and Restoring the Capitol Frescos*, by A. Cox, Museum News, 1961

Table 5: Performance as word recognition degrades.

Correct Choice	*80%*	*65%*	*50%*	*30%*
concept	concept	contrary	amnesty	amnesty
national	national	national	national	national
nationalism	nationalism	nationalism	nationalism	nationalism
nations	nations	nations	nations	nations
political	political	political	political	political
principle	principle	probable	protocols	protocols
responsibility	responsibility	responsibility	inscrutability	responsibility
sovereignty	sovereignty	sovereignty	sovereignty	sovereignty
which	whom	which	whom	watch
world	world	would	world	could

the nouns plus the adjectives are used. The nouns matched up to 30 percent of the non-stop words when a window size of 100 words was used. Performance improved to a match rate of 37 percent when nouns as well as adjectives were utilized. Thus better than one-third of the non-stop words could be successfully recognized by this procedure.

The tradeoff in performance while increasing the window size will be explored. It may be advantageous to use a small window size to locate results that are highly reliable and then recursively apply the word collocation data with those small windows in the manner of a spreading activation model.

Enhancements of the word pair scoring mechanism will also be explored. The determination of a semantic distance between two words is an important aspect of the Wordnet system [6]. This database will be utilized to improve performance in keyword selection and reliability in recognition improvement.

Future tests of the system will use the results of a fully tuned word recognition algorithm applied to real document images. Complete articles from a document image database will be scanned and truthed. Noise will then be introduced in these images and the word image clustering procedure applied. The images of potential keywords will be given to a word recognition algorithm and the resulting neighborhoods will be the input to the simulated annealing search described in this paper.

4 Conclusions

In this paper an algorithm was presented for selecting keywords from the image of a document. The word images extracted from the document were clustered and a limited set of potential keywords were chosen. The images of only those words were recognized and the resulting word recognition neighborhoods were filtered using definitional overlap and a simulated annealing

Table 6: Word collocation statistics for test article.

Window size	Nouns *no. uniq. collocs.*	*overlap*	*pct.*	Nouns and adjectives *no. uniq. collocs.*	*overlap*	*pct.*
4	156	50	3	229	68	4
10	489	172	9	652	221	12
20	919	291	16	1220	393	21
30	1267	359	19	1676	439	24
40	1551	419	23	2059	508	27
100	2693	550	30	3571	682	37

algorithm. The output is a set of keywords for the document.

The advantages of this method include its bypassing of the need to recognize all the ASCII in the document before choosing keywords. Also, the image data for the keywords are merged to generate improved prototypes before recognition. This improves recognition performance and provides better keyword selection than if each word was recognized separately. The ability of the simulated annealing algorithm to choose keywords from the word recognition neighborhoods was demonstrated and the tolerance of the method to recognition errors was discussed.

Acknowledgments Siamak Khoubyari and Tao Hong made significant contributions to this work.

References

[1] L. Guthrie, J. Guthrie, Y. Wilks, J. Cowie, D. Farwell, B. Slator, and R. Bruce. Machine-tractable dictionaries and large-scale sense resolution in text. In *Second Annual Symposium on Document Analysis and Information Retrieval*, pages 17–45, Las Vegas, Nevada, April 1993.

[2] T. K. Ho, J. J. Hull, and S. N. Srihari. A computational model for recognition of multifont word images. *Machine Vision and Applications, special issue on Document Image Analysis*, 5(3):157–168, Summer 1992.

[3] J. J. Hull and Y. Li. Word recognition result interpretation using the vector space model for information retrieval. In *Second Annual Symposium on Document Analysis and Information Retrieval*, pages 147–155, Las Vegas, Nevada, April 1993.

[4] S. Khoubyari and J. J. Hull. Keyword location in noisy document images. In *Second Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, April 1993.

[5] H. Kucera and W. N. Francis. *Computational analysis of present-day American English*. Brown University Press, Providence, Rhode Island, 1967.

[6] G. A. Miller. Wordnet: An online lexical database. *International Journal of Lexicography*, 3(4):235–312, 1990.

[7] T. G. Rose and L. J. Evett. Semantic analysis for large vocabulary cursive script recognition. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 236–239, Tsukuba Science City, Japan, October 1993.

[8] G. Salton. *Automatic text processing*. Addison Wesley, 1988.

[9] D. E. Walker and R. A. Amsler. The use of machine-readable dictionaries in sublanguage analysis. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.