6L−8

Automatic Keyword Assignment in English Documents
"The Intelligent Document Manager"

Phyllis Anwyl, Mitsuhisa Kanaya, Tetsuya Morita
Ricoh Research and Development Center

Introduction

Subsequent to the development of the 知的ワープロファイリングシステム for Japanese document management [1], an English version, the Intelligent Document Manager has been developed. In conjunction with this, an automatic keyword assignment algorithm for English documents has been designed.

The algorithm performs two major functions – 1) a full-text, syntactic analysis of the target document to extract a list of compact noun phrases; and 2) a statistically-based evaluation of those noun phrases to find the most suitable keyword candidates.

The *compact noun phrase extraction* locally analyzes parts-of-speech patterns to find short clauses (less than about seven words) of adjectives, nouns and a subset of connectors (conjunctions and prepositions).

The *keyword candidate evaluation* uses word and phrase occurrence frequencies from the document in a straightforward, efficient formula.

Smooth system presentation due to two other modules in the package – our proprietary matrix thesaurus [2] and the user interface – allow the system to interact efficiently with the user during keyword registration and document retrieval.

Contrasting Systems

Many commercially available retrieval systems rely on well-trained human readers to assign keywords to a given document, allowing only a small set of words and phrases which have been pre-selected. To deal with a high volume of information as well as reduce costs associated with reader-dependent systems, automatic assignment of keywords is obviously preferable. Many automatic systems also rely on pre-defined word lists from which keywords are assigned, if they occur in the target text. Other commercial systems base retrieval on phoneme analysis and co-occurrence.

In the system under study, we have opted to retain a more natural set of keywords, which can be easily tailored by the user, if desired. In addition, we have enhanced the system with our matrix thesaurus, the Dynamic Keyword Connection [2] to allow for variability of keyword phrases amongst documents and increase the effectiveness of the retrieval.

Description

Previous developments at Ricoh with linguistic analyses of English made available morphologic and syntactic parsers, and their associated dictionaries and grammars, for use in the Intelligent Document Manager. Work at Ricoh on keyword assignment [1] from Japanese documents created the framework for our approach. Thus, the algorithms for both Japanese and English keywords consist of two parts – one to identify useful noun phrases within the document, and one to score the suitability of the noun phrases as keywords. However, reflecting the peculiarities of each language, the implementations are entirely different.

COMPACT NOUN PHRASE EXTRACTION

The extraction algorithm is broadly based on ideas similar to those underlying studies to locally disambiguate parts of speech.[3] In the CLAWS process [4], disambiguation, based on collocational probabilities, uses collocational rules. These syntax-based context frame rules delimit which combinations of parts of speech may occur, not necessarily in a particular functional phrase, but rather in a limited run of ambiguous parts of speech. Context frame rules were also used to develop the TAGGIT system [3].

It is this idea of context frame rules which we have adapted for use in our system. These context frame rules resemble the very lowest level of typical context-free grammar (CFG) rules – rules for the nodes one level above terminals on a parse tree. Thus, it would seem feasible to simply use the lowest level of noun phrase CFG rules from an existing English parser. However, examination of such rules revealed that the depth of the noun phrase branch grew quickly and the resulting noun phrases were too complex to serve as keyword phrases. Thus, we created a special set of context frame rules specific to only those noun phrases resembling keyword phrases.

We have termed these *compact noun phrases*. They are, indeed, noun phrases, however, they are not the robust, lengthy noun phrases as universally defined in English grammar. For example, examine the following two fragments:

> ... recently elected as the President of Czechoslovakia, Vaclav Havel, internationally recognized as a leading figure in the world of arts and letters ...

> ... recent research and development projects, undertaken by Japanese semiconductor makers...

Although long, the strings of 22 and 10 words each constitute a single noun phrase. However, from the standpoint of document retrieval they contain several keyword candidates, each only a few words long.

To find these phrases, we have created the compact noun phrase rules. These rules were determined by parsing and lexical analysis of about 6500 multi-word entries in special-purpose dictionaries. As a result, rather than defining a noun phrase to consist of:

[[article adverb adjective] noun
[conjunction relative-clause]]

(repeated in any length combination), we found keyword-like noun phrases to be less than about seven words and to contain only nouns, adjectives and some connectors (a subset of conjunctions and prepositions which varies according to phrase length). These features are described in a simple *transition matrix* (Fig 1). This matrix indicates which parts of speech (POS) may serve as the first word (START), last word (FINAL), and body (CONTINUE) of a compact noun phrase.

The basic process is as follows (Fig 2). The parts of speech in the text are scanned. If a word is found which can serve as the START in a noun phrase, a compact noun phrase candidate is initialized. Words are sequentially added to the candidate as long as they are permitted by the categories (CONTINUE, FINAL) in the transition matrix.

Once a phrase has been identified, further occurrences of the same phrase are also taken into account. Occurrences of shorter phrases within longer phrases are also taken into account.

POS		START	CONTINUE	FINAL
noun		●	●	●
adjective		●	●	
conjunction	*and*		●	
prep - *usual*	*of*		●	
- *rare*	*on, for*		▲	

Figure 1 – Transition Matrix

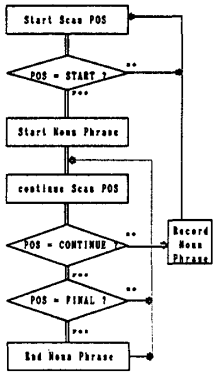

Figure 2 – Process flow

KEYWORD CANDIDATE EVALUATION

After the compact noun phrases are extracted from the target document, each is scored, to indicate its suitability as a keyword. The phrases are then ranked according to their relative scores.

To keep the scoring simple and fast, it uses only two, readily available, values – 1) *phrase count* – the number of occurrences of the phrase within the document and 2) *word count* – the number of occurrences of each noun or adjective which appears in a compact noun phrase.

These values were combined with an empirically determined *weight* to yield the following formula for the *phrase score*:

$$(\sum_{i=0}^{n-1} word\_count_i) + weight * (phrase\_count)$$

Although the weight was not theoretically derived, it can be explained by the fact that it effectively provides a balance between the the two occurrence frequencies, phrase count and word count, and insures that frequently- occurring, shorter phrases take precedence over infrequent, long phrases, which may contain many frequently-occurring nouns and adjectives.

Conclusion

Building upon the foundation of our Japanese document filing system, an automatic keyword assignment algorithm for English has been developed. Detailed results of evaluative tests are too lengthy to discuss here, however, the algorithm has proven quite effective in actual use.

References:

1. 森田哲也、小川、小林：ファジイ文書検索システム（1）～実験システムと評価～情報処理学会第３９回全国大会、pp。1067-8, 1989
2. 小川泰嗣、森田、小林：ファジイ文書検索システム（2）～キーワードコネクションマトリックスの学習方式～情報処理学会第３９回全国大会、pp。1067-8, 1989
3. DeRose, Steven J. "Grammatical Category Disambiguation by Statistical Optimization", Computational Linguistics, 14:1, 1988.
4. Garside, Roger, et al (eds). The Computational Analysis of English: A Corpus-based Approach, London: Longman, 1987.