Applications of Natural Language to Information Systems

Proceedings of the Second International Workshop
June 26-28, 1996, Amsterdam, The Netherlands

Edited by

R.P. van de Riet
Vrije Universiteit, Amsterdam, The Netherlands

J.F.M. Burg
Vrije Universiteit, Amsterdam, The Netherlands

and

A.J. van der Vos
Vrije Universiteit, Amsterdam, The Netherlands

1996

IOS Press
Ohmsha

ISBN 90 5199 273 4(IOS Press)
ISBN 4 274 90102 5 C3000 (Ohmsha)
Library of Congress Catalogue Card Number: 96-76771

Publisher
IOS Press
Van Diemenstraat 94
1013 CN Amsterdam
Netherlands

Distributor in the UK and Ireland
IOS Press/Lavis Marketing
73 Lime Walk
Headington
Oxford OX3 7AD
England

Distributor in the USA and Canada
IOS Press, Inc.
P.O. Box 10558
Burke, VA 22009-0558
USA

Distributor in Japan
Ohmsha, Ltd.
3-1 Kanda Nishiki - Cho
Chiyoda - Ku
Tokyo 101
Japan

Applications of Natural Language to Information Systems
R.P. van de Riet et al. (Eds.)
IOS Press, 1996

The Fact Extraction Using the Keyfact

Mi-Seon JUN, Se-Young PARK, Man-Soo KIM
Natural Language Processing Section
System Software Department
Electronics and Telecommunications Research Institute
TaeJon, Korea
{msjun,sypark,mskim}@com.etri.re.kr

Abstract. Many information retrieval systems retrieve relevant documents based on exact matching of keywords between a query and documents. We shall show how to extract a fact from a document using an extended concept of keyword, called keyfact which can contain syntactic patterns and semantic information. In second document ranking, predefined keyfact cluster set of the query terms is compared to each relevant document. Because relevant documents are such a small fraction of a collection, this method is different with query expansion retrieval scheme and substantially reduces the computational cost of the experiment.
Keywords : Document Ranking, Fact, Semantic Information, Syntactic Pattern

1 Introduction

Many commercial information retrieval(IR) systems retrieve relevant documents based on keyword matching between a query and documents. There are two problems in using the method. The first problem is that keywords are ambiguous, and this ambiguity is causative of retrieving irrelevant document semantically. Therefore lexical ambiguity has to be resolved. The second problem is that a document is treated as a irrelevant document in spite of a relevant document, for the document does not include the same keywords as query terms. So an original query has to be expanded to semantically related words. The main function of an IR system is to rank relevant documents which satisfies the user's information need. In most retrieval models, the system ranks documents according to their inner product similarity, depending upon keywords in a query. However users are generally not interested in retrieving documents with matching keywords, but with concepts that relevant words or information represent.

Facts are truths in some relevant world. These are the things we want to represent and to search information. One representation of facts is so common that it deserves special mention: natural language sentences. Generally nouns and compound nouns are taken as keywords. Nouns and compound nouns are the most important elements for representing the fact in a natural language sentence. However besides the keywords such as nouns and compound nouns, verbs and adjectives have an important role in a sentence. Using the

a sentence, there are many lexical ambiguities. Furthermore in the syntactic problem, there are too many inflected forms of adjectives and adverbs in Korean. Specially in Korean language, one keyword has 20~30 senses in a dictionary in the worst case. Polysemous words in a query and documents can reduce the precision of a search significantly. Therefore lexical ambiguity has to be resolved. To resolve lexical ambiguity of keywords, we need several information such as keywords, verbs and adjectives. We introduced an extended concept of a keyword, called *keyfact* which can be represented as verb/ adjective, and can contain syntactic patterns and semantic information. We can consider lexical ambiguity of noun and verb.

The verb "쓰다" is a typical Korean polysemy, and has twenty one translatable English verbs such as "write", "spend", "wear", and "adopt" in a Sisa Korean-English dictionary[14]. In another example, the noun "모자" is a typical Korean polysemy. The noun has English nouns such as "mother and child" and "hat". So in our keyfact concept noun and verb/adjective are not independent of each other. If the keyfact can contain syntactic patterns and semantic information such as "모자를 쓰다/wear a hat" and "펜으로 쓰다/write with a pen", then noun which occur with the verb "쓰다" may be thought of as a clue for disambiguating senses. In the same manner, ambiguity of noun is much the same.

The literature generally divides lexical ambiguity into two types: syntactic and semantic[2]. Syntactic ambiguity refers to differences in syntactic category. Semantic ambiguity refers to differences in meanings. A number of approaches have been taken to word sense disambiguation. Lesk uses the Oxford Advanced Learners Dictionary[3] and Weiss uses word co-occurrences[4]. Many researchers used statistic information, semantic information, and both of the information as a knowledge for query expansion. Stiles and Lesk used statistic information of term association from documents. Salton experiments only synonym in the SMART system. Fox experiments five semantic category in the SMART system, and need humane intervention for selecting related words[8]. Above researches did not propose the problem of lexical ambiguity or did not considered automatic lexical disambiguation for query expansion. The ambiguity in a query must be resolved when the query is analyzed. Ambiguous words are not able to effectively expand before the ambiguity is not resolved. Query expansion enhances recall by adding some relevant documents excluded from exact matching. But it degrades precision. In order to improve precision, first the ambiguous word in a query is resolved by using knowledge base, when the query is analyzed. Second keyword concept which is defined as noun or compound noun need to be extend, and verb or adjective must be considered as indexing word. We resolve ambiguous query terms, and then expand unambiguous query terms. There are a wide choice of words to add to a query vector. One can add only the synonyms, or synonyms plus all descendants, or synonym plus parents and all descendants, or synonyms plus directly related words, etc. and any number of child links may be traverse. Expansion by synonyms plus and directly related word is benefit[1]. So we choose the parameter, and expanded queries are consisted with a special relationship FT(Fact Term) as well as semantic relationships using in general thesaurus.

In the following sections, we (1) describe the construction of a keyfact network for ranking the documents; (2) present a visualization of the keyfact network for keyfact retrieval; (3) show how the keyfact network can be used to rank documents and extract a fact; (4) evaluate ranking method to improve retrieval performance; and (5) make suggestions for future research.

2 Keyfact and Keyfact Cluster

The noun is the most important element for explaining the fact. Next we consider the compound noun which is composed of several nouns. The syntactic categories of the next complicated fact are noun phrases. A noun phrase consists of a noun and its modifiers that can be represented as inflected forms of verb and adjective. Korean verb and adjective have much more inflected forms as compared to English and French. The most simple fact(sentence) can be represented by noun(subject) and verb/adjective(predicate). So the keyfact is not independent of case slots, and contains syntactic patterns and semantic information. In this paper, we collected keyfacts and keyfact clusters in Gemong Korean encyclopedia for improved retrieval performance. The encyclopedia has two characteristics. First, it has syntactic characteristic composed of a title word and it's explanation part. Second, it has semantic characteristic that most words in the explanation part are semantically related with the title word. The encyclopedia is good to easily collect words and it's semantically related words of a word. We thought that the encyclopedia is a proper collection for construction of semantic information. The keyfacts extracted from a text can be represented in several forms. The forms have to be designed for easy matching. The keyfact weight is calculated in the same formula for calculating keyword weights based on the keyword frequency. When a user gives a query which contains some keyfacts as well as keywords, our system extracts the keyfacts from the query and tries to match the keyfacts which were extracted from documents. We use an exact matching method and when it fails, a partial matching method will be used. With keyfacts co-occurred with a keyword, we can use these keyfacts for disambiguating the keyword and ranking documents.

A cluster is defined as a set of co-occurring keyfact terms. Co-occurring terms are usually relevant to each other and are sometimes synonyms. For keyfact clustering, we considered the sense definition of the noun in the Gemong Korean encyclopedia. In the simple automatic indexing method, raw terms are analyzed using a stemming algorithm and stop words are removed using a stop list. The stop words are usually prepositions, postpositions, determinants and those that appear too frequently to discriminate any documents. And then it finds an identical inflected verb and gets a basic form of the inflected verb by referencing verb dictionary. Our verb dictionary consists of two parts which are inflected verb form and basic verb form. One basic form can have many inflected form. Therefore inflected form is compared with the input text and basic form is used for disambiguating and ranking. Figure 1 shows one basic form can have many inflected forms. We assigned semantic relationships using in general thesaurus such as BT(Broader Term), NT(Narrow Term), RT(Related

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.