Machine에 의한 자연 언어 이해의 효과성 및 탄력성 증대를 위한 자연 언어 이해 기법과 분류 기법의 연결적 통합 사용에 대한 연구

-A study of improving the flexibility and effectiveness of natural anguage understanding considering natural language classification methodologies-

이 원 부

동국대학교 경상대학 정보관리학과

Abstract

This study seeks a way of dealing with unformatted natural language considering fuzzy set theory. The goal of the study is to establish a framework of an effective language understanding system that is linked to language classification system.

This study has found that language understanding is strongly influenced by the language classification. The understanding of language may be improved by the former classification of the language. This study shows that the precision of language classification depends upon the way of how the language is classified in advance. In this study, a fuzzy logic was used to improve the precision of language classification. It was considered that the fuzzy logic might be able to distinctively classify natural language texts into pretinent homogenious groups where contents of the language were identical. Accordingly, in the study, it was expected that classification of language were precisely classified by the fuzzy logic. An experimental system was designed to evaluate the performane of a natural language understanding system that was connected to a fuzzy language classification system. Finally, the experiment suggests that a successful language understanding should require an real time interaction between mem and machine fuzzy provious language classification.

1. 서론

현대 정보 사회의 가장 큰 현상적인 특징은 "정보의 폭발성"이다. 과학 기술 연구 및 경영 활동의 각 분야에 있어서 주어진 목표의 성공적인 달성 여부는, 곧 이 폭발적으로 증가하는 정보의 효과적인 처리 능력에 달려있다고 해도 과언이 아니게 되었다. 이 폭발적으로 제공되는 정보들의원초적인 전달 media(매체)는 자연 언어(natural language)로 된 문서(document or text)이다. 학술 단체의 전문 학술지나 기술 잡지 또는 특허 공보나 정부의 공문이 전형적인 자연 언어로 된 문서들의 예로서 들 수 있다. 이러한 원초적(인위적으로 가공및 조작이 되지않은)으로 발생된 본래의 내용을 간직하고 있는 문서로 된 정보의 효과적인 저장 및 검색 사용 능력은 바로 오늘날에 있어 고도화된 각 부문별 정보 활동에 있어서 궁극적인 성공요소가 되는 것이다.

현대 정보 사회에 있어 1970 년대 이래로 가속화된 컴퓨터 technology 의

눈부신 발전은 정보 사용자로 하여금 더욱 더 많은 량의 문서정보를 더욱 작은 저장 매체(ex. 광디스크)에 효과적으로 수록케하여 사용자의 정보 관리적 활동을 더욱 효율적으로 하게 만들었다. 특히 그간 도서관학 분야에서, 수집되는 문서정보(text정보)들을 정보내용에 따라 사전적으로 분류 및 정의하여 일정 장소(저장매체)에 저장한 후, 주어진 정보 검색의 주제 및 검색 절차(search procedure)에 따라 사후적으로 사용자들의 기호에 맞게 정보의 인출 및 수정을 해주는 여러가지 기법들이 눈부시게 개발 되었다. 하지만 유입되는 정보의 초기 분류 및 색인 과정은 값비싼 인간 사용자의 질적이고 노동 집약적인 노력을 계속적으로 요구하고 있으며, 사후적인 정보검색의 실행이나 사용 수정을 위한 제반 기법들도 아직 완벽하게 사용자의 기호에 맞으면서 효과적인 것은 아직 개발되지 못하고 있다.

더우기 인출된 수많은 특정 주제 관련 제반 문서정보들의 컴퓨터를 이용한 기계적인 정리, 요약 및 이해(summary 또는 understanding)는 아직 요원한 실정으로써 사용자들의 효과적이고 신속한 대량적인 문서정보의 자동화 취급 능력에 대한 수요를 충족시키지 못하고 있는 실정이다. 위와 같은 문서정보 관리상의 애로를 극복하기 위해, 부분적으로는 이미 도서관학 이외의 여러 각 관련 분야에서 적지않는 연구가 있어왔다. 예로서 효과적인 정보의 초기 분류 및 색인을 하기 위해 정보관리학, 전산학 및 언어학 분야에서 인공지능을 이용한 기계에 의한 자연언어 분류기법(Buell 1982 ; Doris 1980 ; Gibbs & Laszlo 1980 ; Lancaster & Fayen 1973)들이 계속 연구되어 왔으며 이와는 별도로 기계에 의한 자연언어 문장의 이해를 위한 제반연구도 인지학, 전산학 및 언어학 분야에서 공동으로 연구되어 왔다(Dejong 1977 & 1982 ; Schmucker 1984 ; Robert 1968 ; Green , Chamsky & Laughery 1963).

일반적으로 자연언어의 분류란 수많은 자연 언어 정보(document정보)중에서 사용자의 특정 검색 주제에 맞는 자연언어 문장들을 신속하게 효율적으로 골라내는 것을 의미하는 것이고, 자연언어의 이해(natural language understanding)란 특정 자연 언어 정보문장들을 컴퓨터로 하여금 판독(읽게)해서 그 정보의 내용을 인간에 의해 미리 주어진 이해를 위한 룰에 따라 컴퓨터가 스스로 요약, 정리하게 하는 것을 의미한다. 따라서 이 두 분야에서의 연결적인 연구는 궁극적으로 우리 사용자들의 문서정보 사용 능력을 고양시켜 인간의 간섭이 없는 자동화된 대량 정보의 지능적이고 효율적인 관리를 가능케 할 것으로 기대된다. 즉 필요 자연 언어정보의 분류로부터 그 내용의 궁극적 이해까지 인간정보 사용자의 노동(手作業)이 필요없는 일관적인 문서정보 관리의 완전 기계화(자동화)가 가능해 질 수 있게 된다.

국내외적으로 고찰해볼 때 현재의 이 두 분야에서의 연구는 제각기 개별적으로 이루어지고 있어 상호 연계성이 전혀없는 실정이다.

결과적으로 이 두 분야에의 그간의 개별적 연구는, 두 분야간의 상호 연계성의 단절에 기인한 비현실적인 연구 가정(assumption) 설정 및 기술개발의 비효율성을 피할 수 없었다. 따라서 본 연구는 이 두 분야(자연언어분류 및 이해)를 각기 개별적으로 연구 분석하는 시각 보다는 상호 보완적인 연결적 방법으로 고찰하여 궁극적으로 문서정보 관리의 완전 자동화를 위한 통합화 방안을 모색함으로써 두 분야의 그간의 문제점들 (비현실적인 연구가정의 도입 및 개발 기술의 제약성등)을 극복할 수 있는 방법을 고찰하는 것이다. 특히 두 분야중에서, 자연 언어 이해 분야의 연구의 문제점 파악 및 기술개발의 애로사항들을 중점적으로 고찰하여 궁극적으로 자연 언어의 기계적 이해를 위한 효과적이고 효율적인 방법을 개발해 보고자 한다. 이를 위해 먼저 이 두분야의 그간의 연구 개요 및 현황을 간단히 훑어 보기로 한다.

2. 자연 언어 분류(Natural language classification)기법의 개요 및 문제점 파악

'자연 언어의 분류' 라는 개념은 현재 '문서 검색' 또는 '정보 인출(imformation retrieval)'이라는 개념과 거의 동일하게 쓰인다 (Lancaster 1986 ; Mathies & Watson 1973). 본래 "자연 언어 분류"란, 일정한 keyword(또는 index word) 들에 의해 사전적으로 이미 색인된 자연 언어로 된 전체 문서정보들 중에서, 사용자가 찾아보고자하는(사용자의 검색 주제에 맞는)문서정보들을 사용자가 제공하는 keyword 들에 의해 효과적으로 (필요한 문서정보는 하나도 빠뜨리지 않고) 그리고 효율적 (불필요한 정보는 전부빼고)으로 골라내는 것을 말한다. 일반적으로 모든 자연언어 문서정보들을 특정한 keyword 로 사전적으로 분류한다는 것은 개개 분류자들의 경험, 숙련도 또는 주제에 대한 정통성(familiarity)에 따라 상당히 자의적이고 가변적이다. 더구나 일반적인 자연언어 분류 방법에서는 모든 문서정보들이 사전적으로 특정 어휘들에 의해 반강제적으로 색인 되어 있기 때문에, 사후에 사용자들이 색인에 사용된 특정어휘 이외의 다른어휘를 사용하고자 할 시에는 목표 문서정보에의 접근및 인출이 불가능하게 된다. 따라서 일반 사용자들은 기존 index에 사용된 특정 어휘군이나 내부적인 검색절차에 미리 어느정도 정통해야만 소기의 목적을 달성할 수 있게 된다.

이러한 일반적인 자연언어 분류방법의 불편한 점을 고려하여, 사전적인 색인과정 없이 사용자들이 그들의 고유한 검색주제에 대한 keywords들을 직접 사용할 수 있게하는 real_time적인 문서 검색 방법이 개발되었다(Lancaster & Fayen 1973). 좀더 기술적으로 고찰해보면, 이 real_time적인 문서검색방법은, 컴퓨터로 하여금 전체 자연언어로 된 문서정보들을 사용자들이 제공한 키워드(user index ward)들을 중심으로 매 검색시마다 읽게(scanning)하여 검색주제에 맞는 문서들을 즉각적으로 골라내는 방법이다.

이 방법은 사전적인 index를 필요로 하는 일반적 방법에 비해, 사용자들의 검색주제에 상당히 탄력적인 반응을 보일 수 있을뿐 아니라 비숙련 사용자들에 요구되어지는, 색인에 사용된 복잡한 keyword들에 대한 사전지식을 전혀 필요로 하지 않게된다. 따라서 검색 시스템의 사용의 편리성이나 시스템에 대한 사용자들의 user_friendliness가 높아 지게 된다. 반면에 이 방법은 문서검색시간의 지연을 피할수 없다. 왜냐하면 매 검색시마다 자연언어로 된 문서정보(text)들이 컴퓨터에 의해 하나도 빠짐없이 전체적으로(entirely) 읽혀져야 하기 때문이다. 하지만 컴퓨터 technology (ex. 光파일 문서처리 system등)의 급격한 발전추세에 비추어보면 이러한 검색시간의 지연문제도 멀지않아 극복될 것으로 보인다.

자연언어 분류시스템이란, 특정 keyword들을 사용하여 일정검색 주제에 맞는 자연언어로 된 문서들을 사용자의 검색의도에 맞게 인출할수있게 하는것을 그 목적으로 한다. 일반적으로 자연언어 검색(또는 분류)시스템의 효과적인 구축및 효율적인 사용을 위해서는, 1)사전적으로 적절한 문서 색인용 어휘선정, 2)사용자들의 주어진 주제검색에 사용되는 사후적인 어휘의 적절한 선택능력이 요구 되어진다. 하지만 효과적인 인덱스 어휘선정과 적절한 검색주제용 어휘구사는 정보검색에 대한 고도로 숙련된 사용자들의 전문성을 요하고 또한 경우에 따라서는 그 숙련된 전문성의 신뢰도에도 차이(variation)가 있을수있다. 더구나 대개의 경우 정보검색에 대한 초보자들인 개개의 사용자들로 하여금 그들의 정보검색에 대한 숙련도를 불문하고 일정주제에 대한 적절한 검색용 어휘 선택요구는 자연언어 검색(분류) system 사용에 본질적으로 내재된(inherent) 어려움이 아닐수 없다.

3.자연언어 이해(natural language understanding)기법의 개요및 문제점 분석

'자연언어 이해'란 인간을 대신하여 컴퓨터로 하여금 자연 언어로된 문장이나 정보들을 스스로 독해, 정리 및 요약을 할 수 있게하는 것을 뜻한다.

그간 자연언어 이해에 대한 연구는 1960년대 이래로 인공 지능 학자 및 언어 학자들을 중심으로 활발히 진행 되어왔다. 하지만 컴퓨터로 하여금 지능적인 인간을 흉내내어 일반적인 자연언어를 이해시키는 것은 엄청나게 어려운 일이 아닐 수 없다. 왜냐하면 자연언어란 컴퓨터 언어에 비해 그 구사방법이나 표현방법 또는 함축적인 표현 방법등이 무궁무진하게 다양하기 때문이다. 일반적으로 컴퓨터 언어에서는, 중요한 명령문들의 문법적 구조가 통일되어있어서 어느 누가 사용하더라도 그 문법적 이해 과정이 아주 기계적(monotonous)이다. 반면 자연언어 문장의 사용에 있어서는 언어의 사용이나 형태에 아무제약이 없기 때문에 통일적 기본문형 이라는 것이 없이, 개개의 사용자의 구사언어가 바로 합법적인 이해 대상언어가 되기 때문에 그 문법적 이해가 아주 가변적이고 난해하다. 더구나 문법적 이해가 일단되었다 하더라도 그 자연언어가 담고있는 본질적인 내용들의 의미론적 이해는 더욱 복잡하고 어려운 작업이다. 따라서 아직까지는 이 자연언어 이해 분야에서의 신기술 및 이론 개발은 다른분야에 비해서 그 진보의 속도가 상당히 더딘 편이다.

좀더 기술적으로 자연언어 이해과정 및 제반 관련 기법들을 고찰해보자면 다음과 같다. 자연언어로 된 문서를 총체적으로 이해하기 위해서는, 먼저 컴퓨터가 전체적인 자연언어 정보의 개개의 단위(sentence)구조들을 문법적으로 파악하여 정보내의 단위구조들의 문법적인 타당성을 파악하여야 한다. 이 과정을 syntactic analysis라고 부른다. 일단 개개의 자연 언어 정보단위들의 문법적 구조파악이 끝나면 이것을 그 정보들이 표현하고있는 개별적인 내용적 의미를 파악할 수 있게하는 의미론적 해석으로 연결시켜 주어야 한다. 이 과정을 semantic analysis라고 부른다. 마지막으로 각개의 단위적인 자연 언어 정보의 의미론적 이해들은 전체적으로 연결 통합되어 하나의 통합된 자연언어 정보(text)의 내용으로 요약 및 정리(pragmatic analysis) 되어진다.

자연언어 이해에 있어서, 자연언어 정보단위(sentence)들의 개별적인 (언어학적이고 어문학적) 이해도 쉽지 않지만 개개 정보 단위들의 유기적 연관성 파악을 통한 전체적인 문장 이해(understanding)는 더욱 극난하다. 그 이유로서는 1)단위문장들의 유기적 연관관계에 대한 일관된 정형성(formality)을 포착하기가 거의 불가능하고, 2)그 정형성을 포착 했다 하더라도 자연언어의 문장이 내포하는 내용및 서술 형태에 따라 그 정형성의 종류는 수없이 많아지게 된다. 따라서 이러한 어려운 점등을 고려하여, 현재 개발되거나 연구중인 대부분의 자연언어의 이해 기법들은 극히 제한된 가정을 설정하고 있다.

그 가정이란 이해 대상 자연언어(document)는 1)일정 내용에 따라 이미 선별이 되어있어 정형성의 인지가 정확히 된 상태이고, 2)그 내용의 서술 형태나 사용 어휘들이 이해를 하기에 비교적 단순하거나(monotonous) 기계적 이라는 점이다. 현재의 대표적인 기계에의한 자연언어 이해 기법의 하나인 frame식 자연언어 이해기법을 이용한 신문기사(news story)이해 방법을 예로 들어보자. 이 방법은 자연언어(natural language)로 쓰여진 news story들을 이해하기 위해서는, 먼저 news story들이 내용에 따라 1)특정주제(정치,경제,사회,문화 등등)별로 이미 분류가 되어있다는 가정으로 부터 출발하고 있고, 2)또한 그 news story들의 전형적인 서술 형식(예를들면 6W1H원칙)이나 사용 어휘들이 주제 내용에 따라 사전적으로 쉽게 인지가 될 수 있다는 것을 전제로 하고 있다.

자연언어(document)의 내용및 서술형태의 정형성을 고려하여 현재까지 개발된 자연언어 이해를 위한 대표적 technology들을 소개해보면 frame식, script식, semantic_net식 및 goal_objective방식 등이 있다.

궁극적으로 현재의 개발된 자연언어 이해기법들의 대부분은 그 대상 자연언어들의 이해의 범위에 대한 주제 선택이 사전에 어떻게 효과적으로 좁혀져 있는지의 여부와 대상 자연언어의 적절한 정형성의 사전적인 인지에 성공여부가 달려 있다는 가정을 설정하고 있다. 하지만 이러한 가정은 매우 비 현실적

이다. 왜냐하면 1)기계에 의한 자연 언어의 이해를 위해서 자연 언어들이 주제별로 미리 사전적으로 분류가 되어야 한다는(특히 사람에 의해)가정은, 사전에 적절하게 미리 분류가 안된 자연 언어들은 사후적으로 그 내용을 이해하기 위한 대상이 전혀 될 수 없다는 점이다. 즉 사전적인 자연 언어 분류를 사후적인 자연언어 이해과정(적절어휘 파악과 서술의 정형성의 선택)의 전제조건으로 한다면 그 이해의 폭이 몹시 제한되지 않을수 없다. 또한 2)사전적으로 자연언어 문장들이 분류가 되는 시점에서는, 어느정도 그 문장의 내용들에 대한 이해도 동시에 수반되기 때문에 사후적인 자연 언어 분류 과정과 연결성이 없는 사후의 독립적인 자연 언어 이해 시도는 작업의 중복된 감을 피할 수 없다는 점이다.

4. 자연 언어 분류 과정 (natural language classifcation)과 이해과정 (natural language understanding)의 통합에 대한 논리적 model설정

자연언어 이해 과정에 있어 가장 어려운 부분은 상기에서 언급했듯이, 자연 언어로된 문장들의 서술 형태의 정형화(formality)의 인식이다. 일단 정형화의 인식이 정확히 된 후라야만 그 정형화에 맞는 적절한 이해 기법을 사용할 수 있다. 예를들어 news story type 으로 정형화된 문장을 이해가기 위해서는 6 W 1 H 원칙을 이용한 신문기사적 이해 기법이 이용이 된다든가 또는 일상적으로 우리가 쉽게 추리할 수 있는 일련의 연속된 동작을 서술한 문장들을 이해하기 위해서는 script 를 이용한 순서적 문장 이해 기법이 이용이 될 수 있다.

정형화 인식의 어려움 이외에 자연 언어 이해 과정의 또 하나의 어려운 점은 각 문장들이 일정한 정형의 틀로 인식이 되었다 하더라도 그 분장들의 개별적인 정형화의 정도 (homogeniety)는 다를 수 있다는 것이다. 예를 들면 어떤 동일한 주제에 따라 분류되어 있는 문장들이라도 그들의 1)서술 형태 및 2)주제에 대한 내용의 관련성 (relevance)들은 서로 다를 수 있다. 즉 정형화의 순수성(homogeniety)이 서로 다르다는 것이다. 이 순수성의 variation 은 궁극적으로 전반적인 문장들의 이해의 과정을 더욱 어렵게 만들게 된다.

따라서 효과적으로 자연 언어를 이해하기 위해서는 먼저 1)대상 자연 언어(natural language texts)들을 내용 및 서술 형태에 따라 homogenious 한 group 으로 정밀 분류를 한 후 2)각 분류에 알맞는 적절한 자연 언어 이해 기법들을 선택 하여야 한다. 정밀 분류가된 문장들은 그 만큼 문장 상호간의 내용상의 homogeniety 에 대한 variation 이 없어지게 되어 적절한 자연 언어 이해 기법의 선택이 수월하게 되고 그 효과성이 높아지게 된다.

여기서 문장들의 분류라 함은 real-time 적인 분류로서, 자연 언어 문장들의 사전적이고 자의적인 index 과정을 필요로 않는다. 즉 사용자들이 그들의 개별적 취향에 따라 그들의 고유한 분류 단어들로 대상 자연 언어 문장들을 real time적으로 정밀 분류한 후 그 분류에 사용된 어휘 및 내용을 고려하여 기존의 자연 언어 이해 기법을 선택하거나 또는 새로운 이해 기법의 정형을 사용자 스스로 설정하여 궁극적인 자연 언어의 이해 과정을 자동화 하는 것이다.

자연 언어들(texts)을 homogenious 한 group 으로 정밀 분류를 하기 위해서는 먼저 일정한 검색 및 이해 주제에 대한 개개의 대상 자연 언어들의 주제에 대한 내용상의 관련성(relevance)을 먼저 구별할 수 있어야 한다. 이를 위해서는 fuzzy set(Zadeh 1975) 이론을 이용한 자연 언어의 상대적 분류 검색 방법이 사용되어질 수 있다 (Baldwin 1979a, 1979b & 1984; Bochvar 1939; Gupta, Saridis & Gaines 1977; Zenner 1985; Yager 1980). Fuzzy 이론은 개개의 (individual) 대상 자연 언어 문장들을 일정 주제에 대한 관련성에 따라 등급을 매겨 유사한 등급의 문장들끼리 정밀 분류를 가능케

해준다.

Fuzzy 이론을 사용하면 검색 및 이해 주제에 대해 모든 문장들의 상대적인 관련성들이 등급이 정해지므로 사용자들은 이 등급을 이용하여 대상 자연 언어 문장들을 그들의 기호에 맞게 아주 정밀하게 상대적분류 (classification)를 할 수 있게된다. 이러한 상대적 분류는 1)궁극적으로 자연 언어 문장들의 서술 형식이나 사용 어휘들의 variation 을 줄일 수 있게 되고 또한 2)사용자의 개인적인 기호 및 요구에 맞는 효과적인 자연언어 이해 기법을 선별적으로 채택할 수 있게 한다.

상기내용의 실증적 검정을 위해 본 연구에서는 자연 언어 분류와 자연 언어 이해 기법을 통합하여 다음과 같은 실험적인 문서 정보 관리 system 모델을 개발하였다.

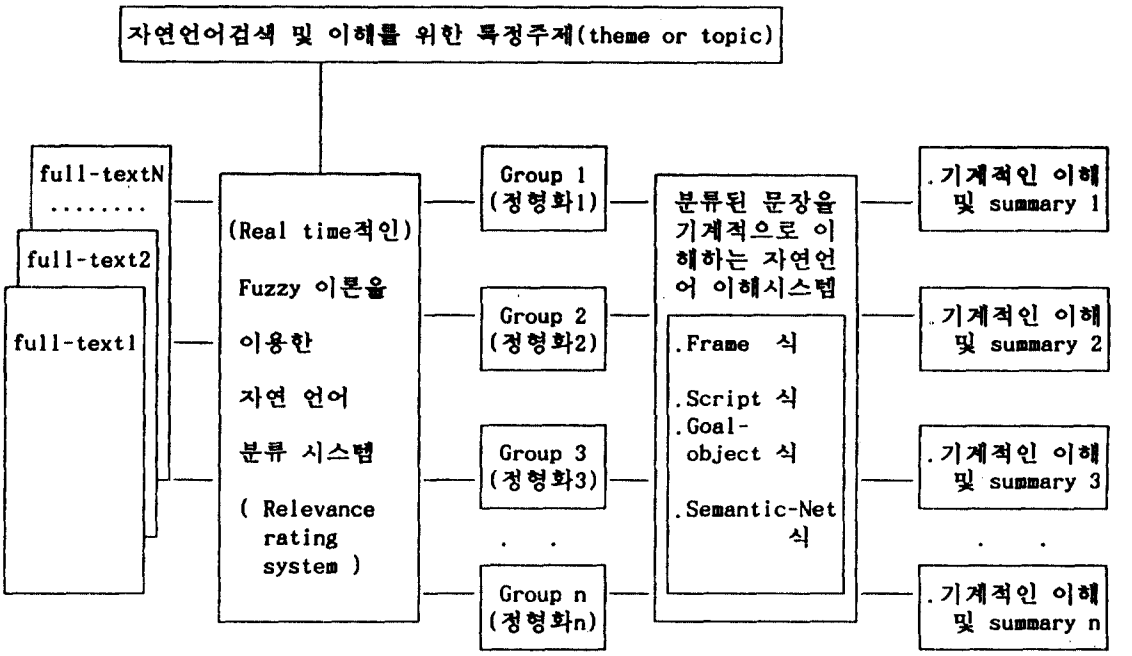

Figure 1 자연언어 분류시스템과 자연언어 이해 시스템의 논리적 통합 모델

5. 통합 시스템의 검색 및 이해 기능에 대한 실험

본 연구에서는 Fuzzy 이론을 이용한 real-time적 자연 언어 text 분류 system을 사용하여 아래와 같은 개요로 정보 검색에 대한 하나의 실험이 실행되었다(Lee 1975). 먼저 실험 대상 text들로써는 AP통신에서 제공되는 516개의 news story(Exhibit 1)들이 무작위로 추출되어 '주식 투자 관계', '예산 수립 관계' 및 '일반 경제관계' 순서로 각각 25, 35 및 50개가 수작업으로 선정되고 나머지는 일반 training text story 군으로 분류되었다. 또한 실험 대상 user들로서는 각 검색 topic당 3명 (i.e., 전문가, 평균적 사용자 및 초보자) 씩을 선정하여 총 15명의 Query(Exhibit 2)를 대상으로 실험이 시행 되었다. 본 Fuzzy 문서 분류 system은 IBM PC AT를 주 기종으로 Compu Serve communication S/W 를 이용하여 AP news story database 에 연결되었다. 또한 실제적으로 자연언어 문장들의 scanning 및 검색을 위한 프로

그램은 PASCAL 과 C 를 이용하여 개발되었다. 본 연구의 주요한 실험 목적은 자연 언어 분류의 성과와 이에 영향을 미치는 제반 요소들과의 상관 관계를 밝혀보는 것이다. 실험 결과를 요약하여보면 다음과 같다.

	stock investment	budget	economy
θ = 0.9	15/25	20/35	25/50
θ = 0.7	20/25	30/35	35/50
θ = 0.5	25/25	34/35	45/50

Figure 2-1 Fuzzy검색방법을 이용한 정밀분류 검색결과(recall기준)

* y/x
  - y → 인출되어야할 texts의 수
  - x → 정확하게 인출된 text의 수

* θ : text 인출시 사용된 cut_off value
(인출을 위한 최소의 relevance value)

일반적으로 fuzzy검색방법에서는, 어떤 text의 주제에 대한 relevance value가 1이면 그 text의 내용이 주어진 검색주제에 완벽하게 맞아 떨어진 것을 뜻하고 relevance value가 0.5이면 검색 주제에 비교적 관계있는 것을 뜻한다. 또한 relevance value가 0이면 대상 text의 내용이 주어진 검색주제에 전혀 일치하지 않는것을 의미한다. 또한 cut_off value(θ)란, 사용자가 특정주제를 가지고 대상 자연언어 정보(text)들을 검색 및 인출할 때 대상 text들의 출력을 위한 최소한의 기준적 relevance value를 의미한다. 위의 예에서 θ=0.9라는 것은 인출된 자연언어 text의 relevance의 value가 모두 0.9 이상이라는 뜻이다. 위의 도표에서 나타난 바와같이 궁극적으로 fuzzy이론을 이용한 자연언어 검색방법에 있어서는, 사용자들은 cut_off value들을 사용자 스스로가 적절히 조정함으로써 대상 자연언어 text들을 비슷한 relevance value군들로 정밀분류 및 인출을 할 수 있게 되었다.

위의 도표에서 보면 낮은 θ value는 비교적 많은 수의 관련된 문서들을 인출시켜 주지만 반면에 인출된 문서들의 내용상의 variation은 크게된다. 일반적으로 낮은 θ value는 비교적 높은 자연언어 검색 성과를 보여주지만 반대로 자연 언어의 이해 과정은 어렵게 만들게 된다.

위의 도표의 검색 성과 평가는 실제로 인출된 문서의 양을 고려치 않고 인출되어야 할 문서의 수를 기준으로 한것이다. 하지만 경우에 따라서는 실제 인출이된 문서의 수량이 중요하게 고려 되어야 할 때가 있다. 즉 검색의 정밀성이 고려 되어야 할 때가 있다. 일반적으로 자연언어 text의 검색에 있어서, 검색 결과를 평가하는 기준은 2가지가 있다. 그 기준이란 검색의 1)효과성에 대한 평가 기준으로 "recall"(Blair & Maron 1985)이라는 비율이 있고 2)효율성에 관한 기준으로 "precision"(Blair & Maron 1985)이라는 비율(ratio)이 있다.

Recall은 어떤검색의 결과, 인출되어야 할 text들을 얼마나 하나도 빠뜨리지 않고 인출 했는지의 여부를 평가하는 비율이다. 이 recall비율은 실제로 정확하게 인출된 text의 수를 인출되어야 할 text의 총수로 나눈 것이다. 따라서 recall은 검색의 목적에 대한 달성도(효과성)를 나타내 준다. 하지만 이 recall에서는, 관련성이 없는 text들을 아무리 많이 인출 했더라도, 그 중에서 정확하게 인출된 text들만 인출 되어야 할 text의 총수로 나

누는 것이므로 검색의 정밀성(효율성)을 무시하게 된다. 따라서 이의 보충을 위해 검색의 정밀성 평가를 위한 Precision(정확성)이라는 비율이 있다. 이 Precision은 어떤 검색방법을 통해, 얼마나 불필요한 text들을 제외시켜 꼭 인출 되어야 할 text들만 찾아 냈는지의 여부를 평가해 준다. Precision의 비율은 실제로 정확하게 인출된 text의 수를 인출 되어진 text의 총수로 나눈 것이다. 상기 언급한 Figure 2-1은 recall을 적용한 검색 결과 평가이다. Precision을 적용한 검색 결과 평가는 다음과 같다.

	stock investment	budget	economy
θ = 0.9	15/17	20/23	25/29
θ = 0.7	20/27	30/41	35/45
θ = 0.5	25/40	34/52	45/70

Figure 2-2 Fuzzy검색방법을 이용한 정밀 검색(precision 기준)결과 도표

* y/x
- y → 인출되어야할 texts의 수
- x → 정확하게 인출된 text의 수

위 도표를 분석해보면, recall에서와 같이 정확하게 인출된 text의 수는 각 cut_off value에 따라 동일하지만 실제로 인출된 총 text의 인출건수는 다르다. 위 도표에서 보면 이미 언급한 바와 같이 낮은 θ value는 비교적 많은수의 필요 정보들을 인출시켜 주지만 그에 못지않게 불필요한 정보들을 대량으로 인출시켜 주고 있다. 결론적으로 낮은 θ value를 이용한 자연언어 분류는 문서들의 homogeneity간에 비교적 높은 variation이 있게되어 그 만큼 자연언어 이해의 과정이 어렵게 된다.

위 두 도표를 종합해보면 cut_off value를 높게 잡을수록 recall value는 떨어지지만 precision value는 올라간다. 즉 recall과 precision은 서로 반비례적인 관계를 가지고 있다. 따라서 상기 실험의 결과는 검색의 효과성을 높이기 위해서는 검색의 정밀성을 희생해야 하고 반대로 검색의 정밀성을 고양하기 위해서는 효과성을 희생해야 한다는 것을 보여주고있다. 실제로 recall과 precision은 서로 상호 보완적으로 사용이 되어진다. 예를들면 필요문서 정보를 하나도 빠뜨리지 않고 인출시키고자 할 때에는 recall을 중요시 해야하고 반대로 몇개의 관련된 문서정보의 인출로도 만족 할 시는 precision을 더욱 중요하게 고려해야 한다. 하지만 앞에서도 언급했듯이 자연언어의 이해과정으로 연결되는 효과적인 검색을 위해서는 precision이 더욱 중요하게 고려 되어야 한다고 볼 수 있다.

본 연구의 실험을 통해보면 (recall과 precision을 불문하고), 非fuzzy 검색방법에비해 fuzzy 검색방법은 text들의 특정주제에 대한 relevance value를 이용하여 검색대상 text들을 효과적으로 구별분류를 해주고있다. 반면非 fuzzy 검색방법은 검색 대상 text들의 주제에 대한 관련성(relevance)을 'yes(θ=1.0)'나 'no(θ=0.0)'로만 파악을 하기때문에 검색대상 text들의 relevance value에 따른 상대적 구별을 못하게 된다.

	검 색 방 법			
	Fuzzy		Non_Fuzzy	
	recall	precision	recall	precision
0.9	0.63	0.88	0.93	0.41
0.7	0.87	0.75	0.93	0.41
0.5	0.95	0.40	0.93	0.41

figure 3 fuzzy와 非 fuzzy검색 방법의 평균적인 검색성과 비교

앞에서 언급한대로, relevance value에 따라 검색대상 text들을 정밀분류한다는것은, 그만큼 검색대상 text들의 내용이나 서술형식 또는 사용어휘들의 다양성(variation)을 줄이게 되어(검색대상 text들의 정형성을 높이게 되어) 기계에 의한 자연언어의 이해성과가 궁극적으로 높아지게 된다. Fuzzy검색방법은 사용자들이 그들의 고유한 cut_off value를 사용하여 인출되는 검색 text들의 인출양들을 조절하여 사용자 스스로가 검색의 정밀성이나 효율성을 높일수 있게 되고, 또한 이 사용자 중심의 자연언어 검색 과정은 곧바로 탄력적인 자연언어 이해과정에 연결되어질 수 있다.

본 연구에서는 자연언어 분류시스템과 탄력적으로 연결 운용되어질 수 있는 pilot적인 자연언어 이해 시스템이 개발되었다. 이 pilot system을 이용한 실험 결과, 중요한 지표적 실험분석 결과가 얻어져 장래 궁극적인 통합 system의 완성에 대한 하나의 방향이 제시되었다. 먼저 pilot system의 architecture를 소개하면 다음과 같다.

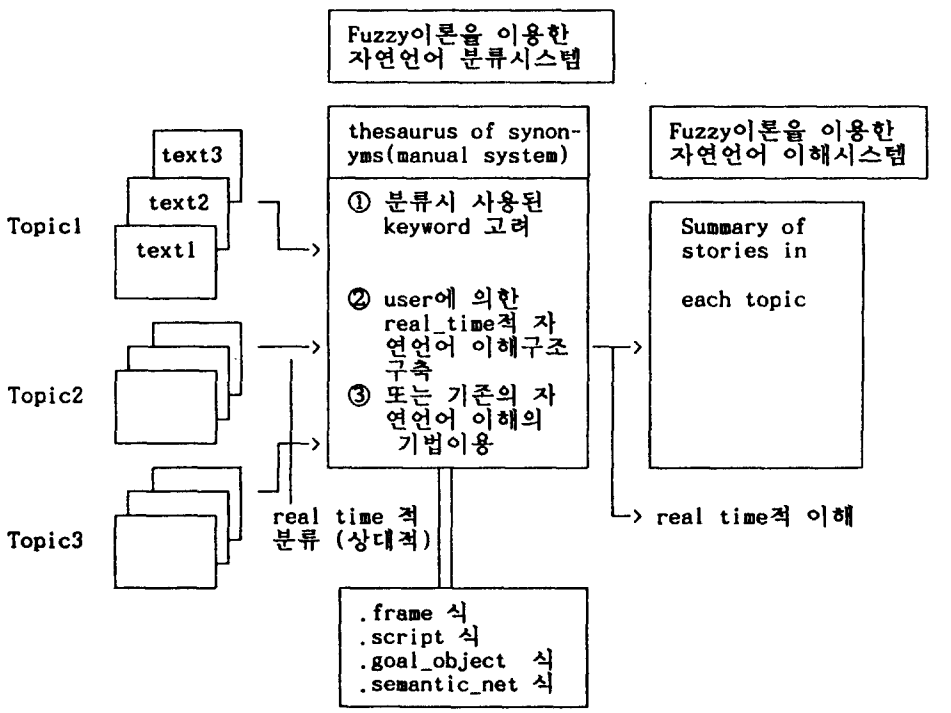

Figure 4. 자연언어이해 시스템의 구조도

상기 자연언어 이해를 위한 pilot system은 앞에서 언급한 실험에서 사용된 AP 통신에서 제공된 stock 투자관계, 예산(budget)관계 및 일반 경제관련 news story들을 재 실험대상으로 하였다. 자연언어 이해를 위한methodolgy로서는 사용자가 제공한 keyword들을 중심으로 한 "word_to_word sementic interpretation" 기법이 사용 되었다. 상기 pilot시스템의 실험대상 문서(story)들에 대한 real_time적인 이해 성과(understanding performance)는 手作業으로 정밀분석되어 다음과 같은 결과가 얻어졌다.

topic / θ	stock	budget	economy	average
θ = 0.9	15/15 1.0	17/20 .85	20/25 0.8	0.88
θ = 0.7	16/20 0.8	19/30 .63	22/35 .63	0.69
θ = 0.5	18/25 .72	19/34 .56	23/45 .51	0.60

Figure 5 pilot 시스템의 자연언어 이해 능력에대한 평가표

* y/x
- y → 자연언어 분류를통해 주제관련 text로 확인및 인출된 text의 수
- x → 만족하게 요약이 되어진 text 수

위 도표에서 살펴보면 θ가 높을수록 실험대상 text들의 이해도(keyword들의 interpretation을 통한 전체 내용에 대한 추리적 이해도)는 높아졌다. 이는 θ value는 실험대상 text들의 내용에 따른 분류의 폭을 점점 좁게 만들어 궁극적으로 text들 내용의 homogeneity를 높여주어 이해의 정도를 높여준다는 것을 의미한다. 따라서 상기 실험은 사전적인 분류 검색시 high cut_off value를 사용하면 궁극적으로 사후적 이해의 성과도 그만큼 높아지게 될 수 있다는 것을 보여주고 있다.

Cut_off value의 사용이외에 위도표에서 중요하게 인지 되어야 할 또하나의 사실은, 검색대상 text의 homogeneity뿐만 아니라 검색주제 자체의 homogeniety에 따라서도 이해의 성과가 달라진다는 점이다. 위의 도표에서 보면 'economy'라는 검색 주제는 'stock investment'라는 주제에 비해 관련 text들의 서술형태나 사용 vocabulary들이 비교적 다양하다. 즉 economy라는 검색주제는 stock investment라는 검색주제보다 그 homogeniety가 낮다는 말이다. 따라서 검색주제의 homogeneity가 낮을수록 machine에 의한 자연언어 이해의 성과가 낮아지게 된다. 이러한 낮은 homogeneity를 극복하기 위해서는 관련text들을 더욱 더 정밀하게 구분해야 할 필요성이 있고 따라서 높은 cut_off value의 사용이 강력히 권장되어 진다.

하지만 경우에 따라서는 사용자의 검색 및 이해 목적에 따라, 좀 더 나은 자연언어의 기계적인 이해를 위하여 high cut_off value가 항상 사용되어야 한다는 것은 옳지 않을 때가 있다. 왜냐하면 때에 따라서는 low cut_off value가 필연적으로 사용되어야 할 때가 있기 때문이다. 예를들면, 군사작전시 특정지역의 지형및 날씨에 관한 자연적 조건에 대한 자연언어 text들을 검색하려고 한다면 이 경우에는 지형이나 날씨에 관련된 text들을 주제에 대한 relevance의 경중을 불문하고 하나도 빠짐없이 인출해야 한다. 이러한 경우에는 비교적 검색의 정밀성을 잃더라도 정보인출의 높은 효과성이 더욱 중요하게 고려 되어야 한다. 즉 low cut_off value의 사용이 필수적이게 되고, low cut_off value는 결론적으로 낮은 homogeneity를 유발하게 된다. 따라서 이 경우에는 이해대상 text들의 낮은 homogeneity에서 야기되는 기계

적인 자연언어의 이해혼란은 어느정도 까지는 피할수 없게 된다.

6. 결론

본 연구에서는, 기계에 의한 자연언어 text의 이해(natural language understanding) 증진을 위해 자연언어 분류(natural language classification) 기법과 자연언어 이해 기법과의 통합적 연결성을 중점적으로 연구했다. 현재 및 종전에 개발된 대부분의 natural language understanding 기법들은 자연언어 검색및 분류 방법들과의 연결성을 전혀 고려하지 않고 language understanding 자체의 기법 연구에만 몰두해 왔다.

예를 들면 자연언어 이해과정의 전제 조건으로서, 1)자연언어 text들은 사전적으로 공히 일정 검색주제들에 관해 먼저 분류가 되어 있어야 한다는 것과 사용자들은 2)대부분의 경우 주어진 분류 text의 내용에 따라 사전적으로 선택된 이해기법의 테두리안에서만 수동적으로 자연언어 이해를 시도해야 한다는등의 비현실적인 가정을 설정하고 있다. 따라서 현재 개발중이거나 연구되는 대부분의 자연언어 이해기법들은 이해대상자연언어 text들의 내용및 서술형태의 다양성에 따른 이해과정의 variation을 탄력적으로 그리고 real time적으로 처리할 수가 없다.

따라서 본 연구에서는 상기 언급된 자연언어 기법들의 내재된 문제점을 극복하는 방편으로서, 사전적인 자연언어(text)의 classification 과정과 분류된 text들의 정리 요약을 위한 사후적인 자연언어(text)의 understanding 과정을 real_time적으로 연결시킬수있는 논리적 model을 고찰하여 보았다. 본 연구에서 이 논리적 model은 사용자위주의 자연언어 이해구조로써 PASCAL과 C를 이용하여 프로그램화 되어, AP통신에서 제공하는 자연언어로 된 texts(news story)들을 대상으로 하여 실제적인 자연언어 검색및 이해과정이 실험적으로 평가되었다. 실험의 결과 다음과 같은 주목할만한 결론이 도출되었다.

첫째 기계에의한 자연언어 이해 과정은, 대상 text들의 내용또는 서술형식 그리고 사용어휘(vocabulary)의 정형성(formality)에 영향을 받는다. 이해대상 text들의 내용의 정형성이 높을수록 기계에 의한 자연언어 이해과정은 효과적으로 수행이 될 수 있다.

둘째 자연언어 text의 정형성을 높이기 위해서는, 이해대상 자연언어(text)들이 내용에 따라 사전적으로 정밀하게 구별 되어야 한다. 이를 위하여, 특정주제에 관해 대상 text들의 상대적인 관련성(relevance)를 인지할수 있는 fuzzy theory를 이용한 자연언어 분류 검색방법이 사용될 수있다. Fuzzy 검색및 분류방법은 일정주제에 대한 각 text들의 relevance value에따라 전체 text들을 등급을 매겨 group화 시키게 되므로 궁극적으로 자연언어 분류및 검색의 정밀화를 달성할 수 있게 된다.

셋째 검색및 이해 대상 topic(주제)들의 homogeneity가 궁극적으로 자연언어 text들의 검색 및 이해에 영향을 미친다. 주제의 homogeneity란, 일정주제에 관한 text들에서 사용되어지는 vocabulary들이나 전개 내용 그리고 문장 서술 형태들의 다양성(variation)을 의미한다. 따라서 이해 주제(topic)의 homogeneity가 높을수록, 이해대상 text들의 문장 구성 형식이나 내용이 다양해지므로 fuzzy classifiacation을 이용한 구별적이고 정밀한 text들의 분류가 더욱 중요하게 요구되어진다.

넷째 본 연구에서는 자연언어 분류시, 높은 cut_off value는 비교적 높은 homogeneity를 보여주었다. 즉 fuzzy이론을 이용한 자연언어 분류시에 사용된 높은 θ value는 분류대상 자연언어들의 높은 formality를 보장하게 되어 궁극적으로 자연언어의 기계적 이해의 성과를 높여 주었다.

다섯째 자연언어 이해과정에 있어, 우선 사용자의 목적에 따라 적절한 검색 및 평가 기준(recall 이나 precision)이 선정되고 이 검색 기준에 따라 먼저 대상 text들이 효과적(recall)으로 또는 효율적(precision)으로 선정이

된 후 사후적으로 적절한 자연언어 기법의 선택및 적용이 권장된다.

결론적으로 보면, 기계에의한 자연언어 분류및 이해 과정을 성공적으로 달성하는 데에 있어서는 사용자들이 자연언어 분류 기법과 이해 기법을 real time적으로 동시연결하여 궁극적인 text들의 정형성 확보가 가장 중요한 전략적 point가 된다. 이를 위해 fuzzy 이론을 이용한 자연언어 분류기법의 사용이 요구되어진다. 또한 fuzzy이론 이외에, 본연구에서는 아직 가설로써만 설정되어 있지만(실험적 뒷받침은 되지 못하고 있지만), text에 사용되는 vocabulary들의 관련 synonym을 고려 한다면 궁극적으로 자연언어 text들의 정형성 향상에 큰 도움이 될것으로 고려된다. 그 이유로써는 자연언어 text의 정형성에 영향을 미치는 요인으로서, 무엇보다도 text의 내용표현을 위해 사용되어지는 vocabulary들이 제일중요하고 그 vocabulary들은 관련된 synonym에 의해 쉽게 정형화가 될수있기 때문이다. synonym을 사용한 vocabulary들의 정형성의 향상은 궁극적으로 전체 text들의 정형성의 향상에 연결이 될 수 있다.

본연구에서 도출된 실험적 결론들은 test sample의 수적인 제약성에 기인된 통계적인 검증의 취약성을 가지고 있다. 앞으로 sample의 증대를 통한 통계적 확인 검증절차가 필요하다. 또한 본연구에서는 자연언어 이해 system의 시스템적인 불완전성으로 인해 pilot test의 선행이 불가피 했으나 향후 시스템의 개발완료에 따른 main test의 실행이 따르게 된다.

REFERENCES

J. Baldwin(1979a),"Fuzzy Logic And Its Application To Fuzzy Reasoning," *Advances in Fuzzy Set Theory And Applications*, Amsterdam, North-Holland, pp93-116

J. Baldwin(1979b),"A New Approach To Approximate Reasoning Using A Fuzzy Logic,"*Fuzzy Sets And Systems*, Volume 2, No. 4, pp309-325

J.Baldwin(1984),"Fuzzy Relational Inference Language(FRIL)," Fuzzy Sets And Systems, Volume 14, No. 2, Amsterdam, North-holland, pp155-174

D.Blair and M.Maron(1985)"An Evaluuation of Retrieval Effectiveness For A Full Text Document Retrieval System,"*ACM Communication*, March,Volume28,No. 3,pp289-299

D.Bochvar(1939),"On A Three-Valued Logical Calculus And Its Application To The Analiysis Of Contradictions," *Journal Of Symbolic Logic*, Volume 4,1939, pp98-99

D. Buell(1982),"An Analysis Of Some Fuzzy Subset Applications To Information Retrieval Systems," *Fuzzy Sets and Systems*, Volume 7, No. 1,pp 35-42

G.Dejong(1977),"Skimming Newspapers Stories By Computer," *Proceedings of the 5th International Joint Conference On Artificial Intelligence*, Volume 1,MIT Press, Cambridge, MA, p16

G.Dejong(1982),"An Overview Of The FRUMP System," in Strategies For Natural Language Processing, W.Lehnet and M.Ringle, Lawrence Erlbaum and Associates, Hillsdale, New Jersey, pp28-45

M.Doris(1980),"To Improve Searching, Check Search Results," *Online* 4,July, pp32-47

I.Durham, D.Lamb and J.Saxe(1983),"Spelling Correction In User Interfaces,"*ACM Communications*,October,1983,Volume 26,No 10,pp 196-209

M.Gibbs and G.Laszlo(1980),"Document Ordering Through Lockheed's DIALOG An SDC's ORBIT-A User's Guide," *Online* 4, October, pp26-31

B.Green, C.Chomsky and K.Laughery(1963),BASEBALL:An Automatic Question Answerer,Computers and Thought(Eds. E. Feigenbaum and J. Feldman),McGraw-Hill, New York, pp207-216
M. Gupta, G.Saridis and B.Gaines(1977),Fuzzy Automate And Decision Process, North-Holland, New York

F. Lancaster(1986),Vocabulary Control For Information Retrieval:2nd Edition, Information Retrieval Press, Arlington, VA

F.Lancaster and E.Fayen (1973),Information Retrieval ON-Line Melville Publishing Company, L.A.

B.Lee(1989) "Consideration of A Quert Methodology to Identify Natural Language Texts That Correspond to Specitied Topics," Ph.D thesis, University of Cincinnati, PP 199-228

W.Lee(1975), Experimental Design And Analysis, W.H.Freeman and Company, San Fransisco, p5

M.Mathies and P.Watson(1973),Computer-Based Reference Service, American Library Association, pp66-80

A.Robert(1968),Information In The Language Sciences, American Elsevier Publishing Company, Inc.,Reading, MA

K.Schmucker(1984),Fuzzy Sets, Natural Language Computations And Risk Analysis, Computer Science Press, Rockville, Maryland, pp5-19

L. Zadeh (1975) " The Concept of a Linguistic Variable And its Application To Approximate Reaconing, " *Information Science*, Volume8, pp199-249

R.Yager(1980)"A Logical on-line Bibliographic Searching : An Application of Fuzzu Sets, "*IEEE Transactions on Systems, Man and Cybernetics*, Volume SMC-3, NO.3, pp28-44

R.Zenner(1985),"A New Approach To Information Retrieval Systema Using Fuzzy Expressions," *Fuzzy Sets and Systems*, Volume 17, No. 1, North-holland, Amsterdam, pp9-22