백과사전 영역에서 중심어주도패턴에 기반한 문장주제 할당 기법

(A Sentence Theme Allocation Scheme based on Head Driven Patterns in Encyclopedia Domain)

강 보 영 † 맹 성 현 ††

(Bo-Young Kang) (Sung-Hyon Myaeng)

요 약 기존의 주제 관련 연구들은 문서에 자주 등장하는 용어를 주제로 간주하는 등 문서에서 다루는 주제에 대한 정의가 모호하다. 또한 문서를 구성하는 기본 단위인 문장의 주제가 문서 요약 및 정보 추출 등의 연구 분야에 중요하게 활용될 수 있음에도 불구하고, 이에 대한 고려 없이 문서 전체의 주제를 추출하고 할당하는 연구가 대부분이다. 따라서 본 논문에서는 문장 단위의 주제 처리에 대한 기본 연구로서, 백과사전 영역에서 효과적인 중심어주도패턴에 기반한 문장주제 할당 기법을 제안하였다. 두산동아 백과사전 인물분야 2,381문서를 대상으로 성능을 분석해본 결과, 제안된 기법이 비교기준보다 향상된 성능을 보였으며, 특히 제안된 네 가지 중심어주도 패턴 중 술어를 기반으로 구성된 중심어주도패턴 유형 4가 학습집합에 대하여 평균 98.96%, 실험집합에 대하여 88.57%의 성능(F-score)으로 주제할당에 가장 효과적임을 알 수 있었다.

키워드 : 백과사전 도메인, 중심어 주도 패턴, 주제별 문장 분류, 문장 주제 할당

Abstract Since sentences are the basic propositional units of text, their themes would be helpful for various tasks that require knowledge about the semantic content of text. Despite the importance of determining the theme of a sentence, however, few studies have investigated the problem of automatically assigning the theme to a sentence. Therefore, we propose a sentence theme allocation scheme based on the head-driven patterns of sentences in encyclopedia. In a serious of experiments using Dusan Dong-A encyclopedia, the proposed method outperformed the baseline of the theme allocation performance. The head-driven pattern 4, which is reconfigured based on the predicate, showed superior performance in the theme allocation with the average F-score of 98.96% for the training data, and 88.57% for the test data.

Key words : head driven pattern, sentence theme allocation, sentence classification

1. 서 론

정보화의 시대로 사회가 변함에 따라 사회, 경제 전반에 걸쳐 정보의 역할은 날로 증대되어 가고 있다. 또한 인터넷, 대형 데이타베이스 및 지식 베이스의 출현과 같은 정보의 폭발적인 양적 증가로 현존하는 수많은 정보 시스템의 소장 정보를 가공, 보급하는데 있어서의 효율성의 문제가 제기되었다. 이러한 다양한 지식베이스 중 백과사전은 정보추출 및 검색, 온톨로지 구축 등의 연구 분야에 있어 유용한 지식베이스로 활용되고 있으므로[1, 2], 백과사전 도메인에서의 효율적인 지식베이스 관리는 보다 나은 성능의 정보 시스템 구현을 가능하게 할 것이다.

이러한 지식베이스를 활용하는 대부분의 시스템은 지식베이스로부터 추출된 용어(term)들을 활용하여 정보를 추출하는데, 해당 용어가 형태적으로 다양하게 발생하거나 관련 단어로 빈번히 출현하기 때문에 보다 정확한 정보처리에 있어 어려움을 겪고 있다[3]. 예를 들어 질의응답 시스템에 적용된 아래의 TREC-9 질의 예문을 살펴보자. 아래 질의는 추출된 용어를 기반으로 정보를 처리하였을 때 시스템이 마주치는 어려움을 묘사한다.

- 질의: What is the name of the inventor of Silly Putty? (TREC-9 질의 811번)

· 본 연구는 한국전자통신연구원(ETRI)의 지원에 의하여 연구되었음

† 비 회 원 : 한국정보통신대학교 공학부 박사후연구원
kby@icu.ac.kr
†† 종신회원 : 한국정보통신대학교 공학부 교수
myaeng@icu.ac.kr

논문접수 : 2004년 12월 22일
심사완료 : 2005년 3월 21일

• 정답: Silly Putty was a war baby. A *General Electric scientist* in New Haven, Conn., stumbled upon its formula while trying to make synthetic rubber during World War II.(WSJ910222-0177 중에서)

질의 1에서 "Silly Putty를 발명한 사람"에 대한 정답은 "General Electric scientist"이다. 질의응답 시스템은 질문의 용어 중 "inventor"와 정답 용어 중 "stumble upon"이 같은 의미로 사용되고 있음을 처리해야한다. 즉, "inventor"의 동사형이 "invent"이며 그것의 유의어 중 하나가 "stumble upon"으로 서로 유사한 의미의 용어로 처리해야 한다. 그러나 이러한 통사적인 카테고리 수정 혹은 형태론적인 변형은 현재 널리 활용되는 전자사전인 워드넷(WordNet)으로는 쉽게 처리하기 어렵다. 만약 예제 질의의 주제가 "발명, 연구"이며 정답 문장의 주제 또한 "발명, 연구, 발견" 등으로 분류되어 있다면 질의와 유사한 개념으로 정답 후보를 좁힘으로써 보다 효과적으로 정답을 추출할 수 있을 것이다.

이러한 기존의 용어 단위 접근의 어려움을 극복하기 위하여, 문서 혹은 단락의 주제 및 화제를 파악하여 정보처리를 수행하고자하는 접근들이 국내외에서 등장하였다[3-9]. 그러나 이러한 주제 관련 기존 연구들은 문서에 자주 등장하는 용어, 즉 키워드(key word)를 암시적으로 주제로 설정하는 등 주제(theme)에 대한 정의가 모호하다. 또한 전체 문서를 구성하는 문장의 의미적 내용이 문서 요약 및 정보 추출 등에 중요한 요소로서 활용될 수 있음에도 불구하고, 문서의 기본 구성 요소인 문장의 의미적 내용에 대한 고려 없이 문서 전체의 주제를 추출 및 할당하는 연구가 대부분이며[3-9], 문장 단위로 주제를 추출 및 할당하는 방법에 대한 연구는 거의 없다. 본 논문에서는 이러한 기존 연구들의 문제점들을 완화하고 문장 단위의 주제 처리에 대한 기본 연구로서, 백과사전 도메인에서 효과적인 중심어주도패턴 기반 문장주제 할당 기법을 제안하고 개발하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 주제 관련 기존 연구를 살펴보고 3장에서 제안된 중심어주도패턴 기반 주제 할당 시스템에 대하여 보다 자세히 설명한 후, 제안된 시스템의 성능을 4장에서 평가하고 분석한다. 마지막으로 5장에서 결론을 맺는다.

2. 기존 연구

2.1 주제에 관한 연구 배경

하나의 텍스트에 대한 분석 및 연구는 텍스트를 구성하는 기본 단위인 문장에서 출발할 수밖에 없다. 따라서 대부분의 언어학자들은 주제에 관한 연구를 문장에서 출발하며 문단 및 문서로 그 범위를 확장하였다[10]. 주제를 보는 관점은 크게 문장 차원의 관점과 담화 차원의 관점으로 구분하여 볼 수 있다. 문장 차원의 주제 연구는 화제-평언(topic-comment) 구조와 관련하여 수행되었으며, 이 경우 화제(topic)는 문장에서 술어가 서술하고 있는 대상을 나타내고 평언(comment)은 화제에 대한 설명내용을 가리킨다. 또한 담화 차원의 관점에서는 담화의 핵심적 정보를 가리키는 중심내용(main idea)과 관련하여 주제 연구가 수행되었다[10].

문장 차원의 주제 연구는 화제(topic)와 주제(theme)를 유사한 개념으로 파악하는 연구가 초기에 지배적이었으나, 전체 담화를 분석하는 텍스트적 관점에서 주제를 파악하는 최근 연구의 경우 화제와 주제를 분리하는 경향이 두드러진다. 텍스트 관점에서 주제연구는 담화의 핵심적 정보를 가리키는 중심내용(main idea)과 관련하여 수행되었는데, 표 1은 Cunningham & Moore의 중심 내용 파악을 위한 요소의 부분과 그에 대한 정의를 제시한 것이다[10,11].

표 1 중심내용 파악을 위한 요소들(Cunningham and Moore[1])

	정 의
화제(topic)	글의 전체 내용을 포괄해 줄 수 있는 ***중심 화제***로서 보통 구
주제(theme)	전체 글이 함의하고 설명하는 세상이나 삶에 대한 ***일반화***

중심내용 파악을 위한 요소들 중, 화제는 주제 파악의 가장 기본적인 요소가 된다. 화제는 하나의 담화가 무엇에 관한 것인가를 표현하는 반면, "주제"는 텍스트가 무엇에 관한 것인가를 지시하는 화제뿐만 아니라 해당 화제를 설명하는 내용 즉, 평언까지 포함한다. 또한 표 1의 "주제"의 정의에서 "일반화"는 상위어로의 대치와 같은 텍스트의 표현적 일반화뿐만 아니라 작가의 의도, 목적 및 평가를 고려한 내용의 일반화까지 포함한다. 즉, "주제"는 화제 및 화제를 설명하는 내용 그리고 그것에 대한 표현 및 내용 일반화까지를 포함하는 개념인 것이다. 그러나 사실 전달에 목적을 두고 있는 설명적 텍스트에서의 "주제"는 내용일반화가 불필요하고 표현적 측면에서의 일반화만 수행된다[10]. 본 논문은 이러한 Cunningham & Moore의 주제에 대한 "일반화" 연구에 기초하여 본 연구에 활용될 문장 주제의 개념을 정의하고자 한다.

2.2 주제 단위 접근들

문서를 대상으로 주제를 탐색하여 적용하는 연구로는 대부분 문서에서 주제로 간주될 수 있는 용어 및 문장을 추출하는 방법이 주류를 이룬다. 문서로부터 주제를 추출하여 적용하는 연구로서, 박기림, 이시은 등의 연구

가 있는데, 박기림(2003)은 문서와 문서 사이의 주제 단위의 유사도를 이용하여 하이퍼링크의 가중치를 새롭게 부여함으로써 검색 성능을 향상시켰다[7]. 이시은(2003)은 의미 구역단위의 검색으로 여러 주제의 정보를 담고 있는 웹 페이지에 대한 검색 결과의 정확성을 높였으며[8], 정태진(2001)은 인터넷 검색엔진에서 주제단위로 검색하는 영역을 축소하여 특정 주제에 관련된 정보만을 찾아줌으로써 보다 적은 하이퍼링크를 방문하고도 적합한 문서를 검색하는 성능을 보였다[9].

이러한 주제 개념을 적용한 기존 연구들은 문서에서 활용하는 주제라는 용어가 무엇을 의미하는지에 대한 충분한 논의 없이 사용하고 있거나, 문서에 자주 등장하는 용어를 주제로 설정하는 등 주제에 대한 정의가 모호하다. 또한 문서로부터 주제로 간주될 수 있는 용어를 추출하는 연구가 대부분이며, 이러한 주제 추출 방법은 기존의 용어 추출 방법론에서 발생하는 동의어 및 어휘 의미 결정 문제 등에 있어 유사한 한계를 가진다. 이러한 한계를 완화하는 방법 중의 하나로 주제를 문서에 할당하는 기법이 있는데[3-9], 문장의 의미적 내용에 적합한 주제를 할당하는 방법은 거의 없다. 따라서 본 논문에서는 백과사전 도메인에서 효과적인 문장 패턴에 기반한 문장 주제 할당 기법을 제안하고 개발한다.

3. 중심어주도패턴 기반 문장주제 할당

본 논문에서는 주제에 관한 기존 연구들의 문제점들을 완화하고 문장 단위의 주제 처리에 대한 기본 연구로서, 백과사전 도메인에서 효과적인 중심어주도패턴 기반 문장 주제 할당 기법을 제안하였다. 제안된 기법은 다음의 단계적 절차에 의해 수행된다.

- 본 논문에 활용될 "문장주제"의 개념 정의
- "문장주제"의 개념에 따라 주제범주체계 구축
- 보다 효과적인 주제 할당을 위한 네 가지 중심어주도패턴 제안
- 제안된 문장 패턴에 기반하여 문장에 주제 할당

제안된 시스템의 전체적인 구성은 그림 1과 같다. 제안된 시스템은 먼저 주제할당에 사용될 주제 개념 및 주제범주체계를 설정하고, 구축된 주제범주체계 및 학습데이타로부터 추출된 중심어주도패턴을 활용하여 문장 패턴 학습을 수행한다. 그런 후 문장주제할당 모듈에서 입력 문장이 주어지면, 문장으로부터 중심어주도패턴을 추출한 후 학습지식을 활용하여 추출된 패턴에 주제를 할당하고, 마지막으로 패턴에 할당된 주제를 각 문장의 주제로 할당한다. 따라서 한 문장 내에 여러 개의 중심어주도패턴이 존재할 경우 해당 문장은 다(多)주제를 할당받을 수 있다.

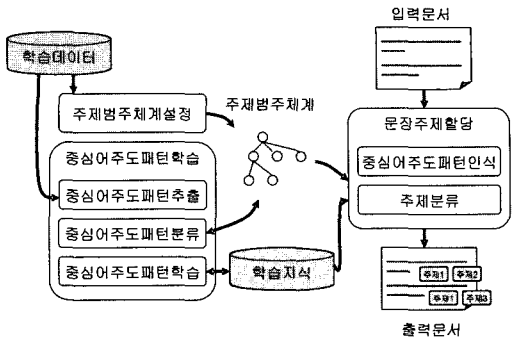
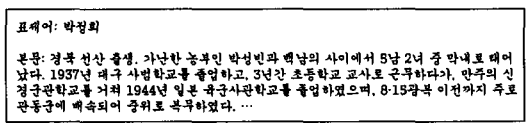


그림 1 전체 시스템 구조

3.1 백과사전 문서의 특성

본 절에서는 본 논문에서 다루고자하는 백과사전 문서 특성을 보다 자세히 살펴보도록 한다. 백과사전은 수록된 용어들에 대한 표제어 및 본문의 집합으로 구성되어 있고, 본문의 내용은 해당 표제어를 설명하는 내용들을 담고 있다. 즉, 백과사전에서 화제는 표제어로 고정되어 있으며, 본문은 해당 화제를 설명하는 내용인 평언들로 구성된다. 예를 들어 "박정희"에 대한 표제어 및 본문 내용 중의 일부분인 그림 2를 살펴보면, 본문에서 설명하고자 하는 대상 즉 화제는 표제어 "박정희"로 고정되어 있음을 알 수 있다. 또한 본문의 각 문장들은 화제 "박정희"를 설명하는 내용인 "경북선산 출생", "사범학교를 졸업하고"와 같은 평언들로 구성되어 있으며, 각 평언은 술어(predicate) 및 술어의 필수성분(argument)들로 구성되어 있다.

이러한 평언을 구성하는 술어 및 술어의 필수 성분들의 분석을 통해 각 평언을 "출생, 교육, 역임" 등 몇 개의 담화 범주로 일반화할 수 있다. 즉, "1937년 대구사범학교를 졸업하고"는 "대구사범학교를 졸업하고"를 보고 "교육"으로, "3년간 초등학교 교사로 근무하다가"는 "교사로 근무하다가"를 보고 "역임"의 범주로 할당할 수 있다. 본 논문은 문장 내 평언들이 가지는 이러한 일반화된 담화 범주를 해당 문장의 주제라고 간주하며, 중심어주도패턴에 기초하여 문장에 주제를 효과적으로 할당하는 기법을 제안하였다.


표제어: 박정희

본문: 경북 선산 출생. 가난한 농부인 박성빈과 백남의 사이에서 5남 2녀 중 막내로 태어났다. 1937년 대구 사범학교를 졸업하고, 3년간 초등학교 교사로 근무하다가, 만주의 신경군관학교를 거쳐 1944년 일본 육군사관학교를 졸업하였으며, 8·15광복 이전까지 주로 관동군에 배속되어 중위로 복무하였다. …

그림 2 "박정희"에 관한 백과사전 표제어 및 본문

3.2 문장 주제

본 논문의 2.1절에서 설명하였듯이, Cunningham &

Moore는 주제를 "전체 글이 함의하고 설명하는 세상이나 삶에 대한 일반화"라고 정의하였으며, 텍스트의 표면적 측면뿐만 아니라 내용의 일반화까지를 포함하는 용어로 주장하였다. 따라서 주제는 화제 및 화제를 설명하는 내용 그리고 그것에 대한 표현 및 내용 일반화까지를 포함하는 개념인 것이다[10].

본 논문에서 다루고자하는 백과사전과 같은 설명적 텍스트의 경우 "주제"는 내용 일반화가 불필요하고 표현적 측면에서의 일반화만 수행된다. 또한, 백과사전에서 본문의 내용은 화제를 설명하는 평언들로 구성되어 있다. 결국 백과사전 본문에서 문장의 주제를 추정하는 것은 화제를 설명하는 내용인 평언을 표현적 측면에서 일반화시키는 과정으로 간주된다. 표현적 측면에서의 일반화는 하위어들을 상위어로 대체하거나 하나의 통일된 동의어로 표현하는 것이다.

정의 1(문장 주제): 문장 주제는 화제를 설명하는 평언의 일반화된 용어 T로 $<\alpha, \gamma, \Delta, \Phi>$에 의해 생성된다. 여기서 α는 화제를 설명하는 내용인 평언이고, γ는 평언 α로부터 구성된 중심어주도패턴이며, Δ는 평언의 일반화를 위하여 활용되는 주제범주체계이다. 또한 $\Phi: \gamma \rightarrow T$는 중심어주도패턴 γ를 주제 $T \in \Delta$로 대응하는 주제대응함수이다.

그림 3은 정의된 문장주제의 개념을 활용하여 입력문장이 주어지면 주제가 할당되는 과정에 대한 전체적인 흐름을 보여준다. 먼저 입력문장 S가 주어지면, 문장으로부터 평언 α가 추출되고 평언 α로부터 중심어주도패턴 γ이 구성된다. 중심어주도패턴 γ은 효과적으로 주제를 할당할 수 있도록 평언의 구조를 재구성한 것이다. 구성된 중심어주도패턴 γ을 입력으로 주제대응함수 Φ는 문장에 적합한 주제 T를 할당한다. 예를 들어, 그림 2에서 화제 "박정희"를 설명하는 내용 중 첫 번째 문장 "경북선산 출생"이 입력문장 S로 주어지면 해당문장에서 평언, <경북선산, 출생, null>이 추출된다. 추출된 평언 α로부터 "경북선산=<출생, null>" 혹은 "출생=<경북선산, null>"등의 중심어주도패턴 γ으로 재구성하고, 재구성된 중심어주도패턴 γ을 주제대응함수 Φ을 활용하여 주제 T, "출생"으로 할당한다. 따라서 입력문장 "경북선산 출생"은 "출생"의 주제를 할당받는다.

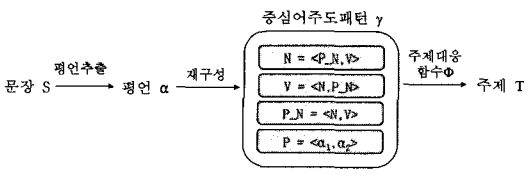


그림 3 문장주제 할당 기법

3.3 주제범주체계 구축

이전 절에서 문장 주제는 평언의 일반화와 관련이 됨을 시사하였으며, 본 절에서는 이러한 평언의 일반화를 위하여 활용되는 주제범주체계를 설정하는 과정에 대해서 설명한다. 본 논문에서는 문서에 자주 등장하는 용어가 중요한 개념 및 범주를 표현할 수 있다고 가정하고, 용어 빈도가 높은 단어를 대상으로 주제범주체계를 설정하였다.

평언 범주의 계층구조 설정 기준에 있어서는 François의 연구 및 WordNet의 계층적 개념 범주체계 연구를 참고하였다[10,12-14]. François는 각 개인이 삶을 통해 마주하게 되는 다양한 상황들을 동사의 의미적 특성들을 활용하여, "action, act, activity, event, process, state"의 여섯 가지 중 하나로 구별할 수 있다는 연구를 보고하였다[13,14]. 동사의 의미적 특성과 관련된 문장에 대한 개념적 분류 연구 외에도, 서술성 명사의 개념 분류로서 불어의 경우 "상태, 행위, 사건"의 담화 영역에서 발생하는 것으로 알려져 있다[10]. 이러한 관점은 모든 명사들은 "action, event, state"등을 포함한 25가지 최상위 범주에서 발생한다는 워드넷[12]의 명사에 대한 최상위 범주 구성과 중복된다. 따라서 동사 및 명사들이 발생하는 담화영역을 모두 포괄하는 워드넷의 범주 구성을 평언의 일반화를 위한 범주 기준으로 활용한다.

주제범주로 설정될 후보 용어를 선정하는 기준은 온톨로지 구축 시 일반적으로 적용되는 용어 빈도(Term frequency)에 기초한 기법을 활용하였다[2]. 즉, 대상 도메인에서 추출한 모든 명사 및 동사의 용어 빈도를 계산한 후, 상위 30%의 용어 빈도를 차지하는 명사 및 동사를 주제범주체계 설정을 위한 후보 용어로 선정하였다. 선택된 후보용어로 주제범주체계를 설정하는 과정은 그림 4와 같다. 그림 4에서 상자로 둘러싸여진 용어들은 주제범주체계를 위해 추출된 후보용어들이며, 그 외의 용어들은 최상위 범주를 포함한 워드넷 용어이다. 먼저 추출된 용어들로부터 일반화된 용어를 찾기 위하여 워드넷을 참조하여 일반화관계를 가지는 용어들을 트리구조로 연결한다. 일반화관계는 동의어, 상/하위어, 부분어 관계들이다. 그런 후, 구축된 트리구조를 워드넷의 각 최상위 용어와 연결한다. 예를 들면, "사망, 죽, 살해, 안치"로 구성된 트리구조는 워드넷 최상위 용어 중 "사건" 용어와 연결된다. 본 논문은 구성된 계층 구조 중, 워드넷 최상위 용어와 구성된 트리구조의 최상위 용어를 주제범주체계로 구성한다. 즉, 예제 그림의 "사건, 사망, 죽, 살해, 안치.."로 구성된 계층구조에서 워드넷 최상위 용어인 "사건"과 후보용어로 구성된 트리구조의 최상위 용어 "사망"을 주제체계를 구성하는 범주로 선정된다. 제안된 기준에 의해 설정된 백과사전 인물분야 주제범

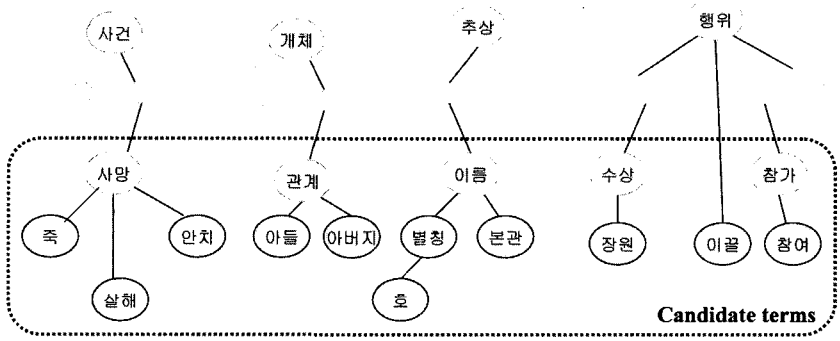


그림 4 주제범주 체계 구축 과정

주체계 구성은 표 2와 같다. 주제범주체계는 대분류 4개, 중분류 25개 구성으로 이루어지며 총 29개 범주로 나누어진다.

3.4 중심어주도패턴

문장의 주제를 탐색하기 위해서는 화제를 설명하는 평언의 내용을 살펴보아야 한다. 그림 2의 예제 문서에서 살펴보았듯이 백과사전 문서의 본문은 화제를 설명하는 평언들로 구성되며, 각 평언은 술어 및 술어가 필요로 하는 필수성분들로 구성되어 매우 짧고 간결한 경우가 대부분이다. 한국어의 술어는 동사 및 서술성 명사(predicative noun)로 나누는데[15], 서술성 명사란 "공부하다"와 같은 술어에서 기능동사 "하다"가 수반하는 용어 "공부"와 같은 명사를 지칭하는 표현이다.

표 2 인물 분야 주제범주 체계

대분류	중분류
추상(Abstraction)	이름/작풍/부정/관계
행위(Act)	참가/설립/조직/기록/발견/계승 연구/주장/수상/역임/교육/이동 죄/싸움/평가/거주
사건(Event)	발생/변화/사망/출생
개체(Entity)	작품

결국 백과사전 본문의 경우 이러한 술어 및 술어 주변 용어를 살펴보면 화제를 설명하는 내용인 평언을 쉽게 파악할 수 있으며, 이러한 평언으로부터 구성되는 중심어주도패턴에 기반하여 주제를 밝히고자 한다. 본 논문에서는 평언을 동사, 서술성명사, 주변 명사 및 개체명 리스트로 다음과 같이 정의한다.

정의 2(평언): 평언 α는 화제를 설명하는 내용으로 동사, 서술성명사, 주변 명사(N)로 성된 리스트로 정의된다.

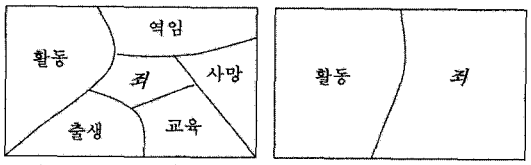
$$\alpha = <N, P\_N , V>$$

여기에서, N은 일반명사 및 개체명(named entity)을, P_N은 서술성 명사(predicative noun), V는 기능동사를 포함한 동사(verb)이다. 여기서, P_N은 기능동사 "하다, 되다, 당하다"등의 경우 적용되어 추출되는 서술성 명사이고, N은 동사를 중심으로 주변 1개 일반 명사 및 개체명까지 동사의 논항으로 추출된다.

평언 중 N은 동사를 중심으로 주변 1개까지 동사의 논항으로 추출하였으나, "하다, 되다"와 같은 기능동사의 경우 동사 바로 앞에 서술성 명사를 동반하여 해당 서술성 명사가 문장의 술어로서 활용되는 현상을 보이므로, 기능동사 바로 앞에 서술성 명사를 포함하여 주변 1개 명사 및 개체명까지 술어의 논항으로 추출한다. 즉, "영향/nn+을 받/v+았+다"에서 일반 동사 "받"의 경우 <영향, null, 받>으로 평언이 추출되며, 예문"<LOC: 하버드대학/nn>+에서 공부/nc+하/xsv+였으며"의 경우, 기능동사 "하"와 바로 앞의 서술성 명사 "공부"까지 포함한 <LOC, 공부, 하>를 평언으로 추출한다.

추출된 평언의 주제를 결정하는데 있어 본 논문은 주제 할당을 위하여 보다 효과적인 평언구성방식이 있다고 간주한다. 예를 들어 아래 그림 5는 예제 평언 <형, 확정, 되>를 이용하여 주제를 결정하는 다양한 방법을 묘사한다. 그림 5의 (a)는 평언의 구성요소 중에는 주제결정에 주도적 역할을 하는 어휘가 없고, 모든 구성요소가 주제결정에 있어 동일한 중요도를 가지는 경우를 설명한다. 따라서 예제 평언은 재구성되지 않고 그대로 주제 결정에 활용된다. 즉, 추출된 모든 평언 α가 가지는 다양한 주제 "활동, 역임, 죄, 출생, 교육, 사망"에서 평언의 구체적인 구성요소 "형, 확정, 되"를 각각 살펴보고 "죄"의 주제로 평언을 할당하는 것이다. 이 경우 "죄"의 주제 결정을 위해 활용되는 다양한 주제후보 "활동, 역임"등은 각 평언별로 주제가 태깅된 학습집합으로부터 유도될 수 있다.

그러나 그림 5의 (b)는 (a)와는 달리, 예제 평언 <형, 확정, 되>의 구성요소 중에서 평언의 주제 결정에 전적



(a) a : <형, 확정, 되> (b) Head : 확정, Tail : <형,되>

그림 5 평언을 이용한 다양한 주제 결정 방식

으로 영향을 미치는 어휘가 있다고 간주하고, 이러한 어휘가 가지는 주제 범주가 바로 평언의 주제로 결정되는 방식이다. 따라서 평언은 이러한 주제 결정에 영향을 미치는 어휘를 중심으로 재구성된다. 이 경우 평언의 주제 결정에 결정적인 영향을 미치는 어휘를 중심어(Head)라 하고, 이러한 중심어의 주제를 결정하기 위하여 활용되는 추가적인 어휘 정보를 주변어(Neighbor)라 부른다. 즉, (b)의 경우는 예제 평언에서 평언의 주제 결정에 전적으로 영향을 미치는 중심어가 술어성명사 "확정"이라고 보고, 해당 중심어가 가지는 "죄, 활동"과 같은 주제 범주 중, 주변어 <형, 되>를 보고 중심어 "확정"의 주제를 "죄"로 결정하고 예제 평언의 주제로도 "죄"를 할당한다.

본 논문은 위의 예제 그림에서 설명한 주제 결정 방식 중, 평언의 주제를 결정하는데 결정적인 역할을 하는 중심어가 있다고 가정하고, 이러한 중심어 어휘를 고려한 평언구성이 주제 결정에 있어 보다 효과적이라고 간주한다. 따라서 추출된 평언을 중심어를 고려하여 재구성하는데, 이러한 평언구성방식을 다음 정의 3과 같이 중심어주도패턴이라 정의한다.

정의 3(중심어주도패턴): 중심어주도패턴은 보다 효과적인 주제 결정을 위하여 평언 α로부터 중심어를 고려하여 재구성된 패턴으로 다음과 같이 네 가지 유형으로 구성될 수 있다.

(중심어주도패턴1) V = <N, P_N>
(중심어주도패턴2) N = <V, P_N>
(중심어주도패턴3) P_N = <N, P_N>
(중심어주도패턴4) P = $<\alpha_1, \alpha_2>$

중심어주도패턴4에서 술어 P는 평언에 술어성명사가 존재하면 술어성명사가 P가 되고 그렇지 않으면 동사가 P가 된다. 또한 α_1, α_2는 각각 평언에서 술어 P를 제외한 구성소이다.

추출된 평언으로부터 중심어주도패턴1로 재구성하는 것은 평언의 구성요소 중 주제 결정에 전적으로 영향을 미치는 중심어를 동사(V)로, 추가정보인 주변어를 주변명사(N) 및 술어성 명사(P_N)라고 간주하는 경우이다. 중심어주도패턴2와 3도 중심어주도패턴1과 같은 방식으로 설명될 수 있으며, 중심어주도패턴4는 평언의 구성요소 중 중심어를 술어(P)로 주변어를 술어를 제외한 구성요소(a_i)라고 보고, 술어의 주제를 결정함으로써 평언의 주제를 결정하고자 하는 경우이다.

표 3은 두 예제 평언 <학교, 공부, 하>와 <전투, null, 죽>을 정의된 네 가지 중심어주도패턴으로 재구성하는 방법을 설명한다. 예를 들어 표 3의 (1)은 평언에서 주제 결정에 결정적 역할을 하는 중심어를 동사라고 간주하는 중심어주도패턴1을 나타낸다. 즉, 평언 <학교, 공부, 하>에서 중심어는 동사 "하"이고 "하"의 주제를 결정하기 위하여 추가적인 정보로 사용되는 주변어는 <학교, 공부>이다. 본 연구에서는 평언으로부터 구성되는 제안된 네 가지 중심어주도패턴 중 문장에 주제를 할당하는 데 있어 보다 효과적인 패턴을 실험을 통해 선택한다.

표 3 재구성된 중심어주도패턴 예제

중심어주도패턴	Head =	<Neighbor>	
(1)중심어주도패턴1	하	학교,	공부
	죽	전투,	null
(2)중심어주도패턴2	학교	공부,	하
	전투	null,	죽
(3)중심어주도패턴3	공부	학교,	하
	null	전투,	죽
(4)중심어주도패턴4	공부	학교,	하
	죽	전투,	null

3.5 주제대응함수

본 절에서는 3.4절에서 제안된 각 중심어주도패턴을 해당 주제로 대응하는 주제대응함수에 대해 설명한다. 주제대응함수는 평언으로부터 구성된 중심어주도패턴을 보다 일반화된 용어인 주제로 대응하는 함수로서, 각 패턴에 따라 다음과 같이 네 가지 함수로 나눌 수 있다. 대응 기법으로는 Naive Bayesian 알고리즘을 적용한다.

문장 주제는 화제를 설명하는 평언의 일반화된 용어 T로 $<\alpha, \gamma, \Delta, \Phi>$에 의해 생성된다. 여기서 α는 화제를 설명하는 내용인 평언이고, γ는 평언 α로부터 구성된 중심어주도패턴이며, Δ는 평언의 일반화를 위하여 활용되는 주제범주체계이다. 또한 $\Phi: \gamma \rightarrow$T는 중심어주도패턴 γ를 주제 T$\in\Delta$로 대응하는 주제대응함수이다.

(주제대응함수1) Φ_1: V=<N, P_N> → T (N, V, P_N $\in\alpha$ 이고 T $\in\Delta$)
(주제대응함수2) Φ_2: N=<V, P_N> → T (N, V, P_N $\in\alpha$ 이고 T $\in\Delta$)
(주제대응함수3) Φ_3: P_N=<N, V> → T (N, V, P_N $\in\alpha$ 이고 T $\in\Delta$)
(주제대응함수4) Φ_4: P = $<\alpha_1, \alpha_2>$ → T (P, α_1, α_2 $\in\alpha$ 이고 주제 T $\in\Delta$)

학습 및 분류 기법으로 적용되는 Naive Bayesian 알고리즘에서 주제대응함수의 동작과정을 중심어주도패턴1

을 예를 들어 자세히 설명하면 다음과 같다. 평언 α의 각 구성소, N, P_N, V에 대한 값이 각각 주어지면, 해당 평언은 α = <v_k, n_k, pn_k>로 표현될 수 있다. 주제대응함수1은 평언 α의 주제를 결정하는 것을 동사 v_k의 주제를 결정하는 과정으로 간주하고, 동사 v_k의 주제는 주변어 <n_k, pn_k>를 활용하여 결정한다. 따라서 평언 <v_k, n_k, pn_k>은 중심어주도패턴1 v_k=<n_k, pn_k>로 재구성된다. 주제대응함수1에서 평언 α의 주제를 결정하는 과정은 동사 v_k의 주제를 결정하는 과정으로 다음 수식 (1)과 같다.

$$\begin{aligned} T &= argmax_{T_i \in \Delta} P(T_i | v_k) \\ &= argmax_{T_i \in \Delta} \frac{P(n_k, pn_k | T) P(T)}{P(n_k, pn_k)} \\ &= argmax_{T_i \in \Delta} P(n_k, pn_k | T_i) P(T_i) \end{aligned} \quad (1)$$

평언 α의 범주 T는, v_k의 주제가 T_i일 확률 $P(T_i)$와 v_k의 주제가 T_i일 때 용어 ne_k, p_nn_k가 각각 발생할 확률 $P(n_k, pn_k|T_i)$을 구하여 가장 높은 값을 가지는 주제를 평언의 주제(T)로 결정한다. 이러한 수식유도과정은 평언을 주제로 할당하는 다양한 주제대응함수2~함수4에 대해서도 같은 방식으로 유도될 수 있다.

4. 성능분석

본 논문에서 제안한 기법은 문장에서 추출한 중심어 주도패턴에 주제를 할당한 후 해당 문장에 주제를 할당하는 시스템으로 제안된 시스템의 성능은 중심어주도패턴을 해당 주제로 정확히 분류하는 성능에 의존한다. 따라서 본 장에서는 추출된 중심어주도패턴에 대한 주제 할당 성능을 실험 및 평가한다.

시스템 성능 평가를 위해 도입되는 척도는 수식 (2)~수식 (4)와 같이 정확율(Precision)과 재현율(Recall), 그리고 F-score이다. 정확율(P)은 시스템에 의해 추출된 정답에 대한 실제 정답의 비율이고 재현율(R)은 실제 정답에 대한 시스템에 의해 추출된 정답의 비율이다.

	'예'가 정답	'아니오'가 정답
'예'로 추출	a	b
'아니오'로 추출	c	d

$$P = a/(a+c) \quad (2)$$

$$R = a/(a+b) \quad (3)$$

$$F-score = \frac{2PR}{P+R} \quad (4)$$

사용된 실험 문서는 두산동아 백과사전 인물 분야 2,381 문서로 시대별로는 고대, 중세, 현대, 직업별로는 정치, 과학, 의학, 군사 등에 고르게 분포되어 있다. 인물분야 2,381 문서를 단문 분할, 형태소 분석, 개체명 태깅 및 정답 유형 태깅한 후, 9,987개의 평언을 추출하여 각 중심어주도패턴별 학습 및 분류 성능 검증을 위한 실험 집합으로 구성하였다. 형태소 분석의 경우 약 97%의 정확도를 보이고, 개체명 태깅은 약 F-Score 70%의 성능을 보인다.

4.1 비교기준(Baseline)

문장의 주제 할당에 대한 기존 연구 및 비교대상으로 활용 가능한 시스템이 거의 없으므로, 본 절은 본 논문에서 사용하고자하는 비교기준에 대하여 먼저 설명한다. 본 논문은 3.4절 그림 5에서, 추출한 평언을 주제로 대응하는 다양한 방법 중, 평언을 중심어주도패턴으로 재구성하지 않고 그대로 학습 및 분류에 활용하는 (a)기법이 평언을 중심어주도패턴으로 재구성하여 주제 할당을 수행하는 제안된 기법의 효과 및 성능에 대한 비교대상이 될 수 있다고 간주하였다. 따라서 평언을 그대로 문장 주제 학습 및 할당에 적용하는 기법의 성능을 기준 성능으로 활용한다.

비교대상 기법의 주제 결정 과정을 설명하면 다음 수식 (5)와 같다. 평언 α의 각 구성요소에 대한 값이 각각 a_1, a_2, a_3로 주어지면, 해당 평언은 α=<a_1, a_2, a_3>로 표현될 수 있다. 따라서 평언 α의 주제 T는 평언 α가 주제 T_i일 확률 $P(T_i)$와 주제 T_i에서 각 구성요소 a_j가 발생할 확률 $P(a_j|T_i)$을 구하여 가장 높은 값을 가지는 주제를 평언의 주제(T)로 결정한다.

$$\begin{aligned} T &= argmax_{T_i \in \Delta} P(T_i | \alpha) \\ &= argmax_{T_i \in \Delta} P(a_1, a_2, a_3 | T_i) P(T_i) \end{aligned} \quad (5)$$

비교대상 기법은 인물문서로부터 추출된 9,987개 평언을 중심어주도패턴으로 재구성하지 않고, 5차 교차검증(5-fold cross validation)을 수행한 후 F-score 마이크로 평균(micro average)을 계산하였다. 교차 검증에 사용된 학습 및 실험 집합은 각각 학습 집합 8,110개, 실험집합 1,996개의 평언으로 구성되어있다. 학습 집합에 대하여 학습을 수행한 후 주제 할당을 한 결과 F-score 평균 41.9%의 성능을 보였으며, 학습 집합으로 학습을 수행한 후 실험집합에 대하여 주제 할당을 한 결과 F-score 평균 37.4%의 성능을 보였다.

4.2 중심어주도패턴에 기반한 주제 할당

본 절은 본 논문에서 제안한 네 가지 중심어주도패턴에 대한 학습 및 주제 할당 성능을 기준 성능과 비교 설명한다. 각 중심어주도패턴에 대한 학습 및 분류 성능은 그림 6에 제시되어 있다. 그림 6은 각 패턴에 대하여 5차 교차검증(5-fold cross validation)을 수행한 후 F-score를 마이크로 평균(micro average)한 결과이다. 그림 6의 결과로부터 제안된 네 가지 중심어주도패턴 중 술어를 평언의 중심어로 간주한 중심어주도패턴4의 성능이 다른 중심어주도패턴1-3의 성능에 비해 월등히 우수함을 알 수 있다. 중심어주도패턴4는 학습 집합에

대하여 학습을 수행한 후 주제 할당을 한 결과 98.86%의 성능을 보였으며, 학습 집합으로 학습을 수행한 후 실험집합에 대하여 주제 할당을 한 결과 88.57%의 성능을 보였다.

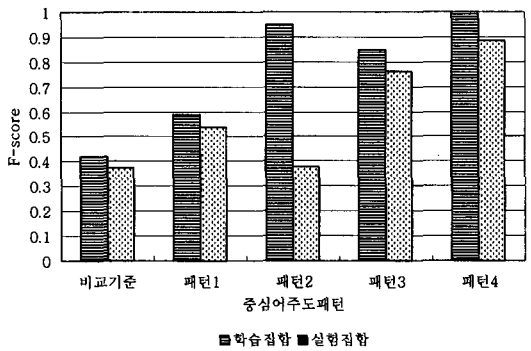


그림 6 각 문장 패턴의 학습 및 분류 성능(그림texture 수정)

표 4는 기준 성능에 대한 제안된 시스템의 성능 향상 정도를 제시한다. 표 4에서 알 수 있듯이 중심어를 중심으로 중심어주도패턴으로 재구성한 제안된 기법이 평언을 그대로 주제 분류에 적용한 기준 성능보다 모든 제안된 중심어주도패턴에서 향상된 성능을 보임을 알 수 있다. 특히, 중심어주도패턴4는 기준성능보다 학습 집합에서 약 136%, 실험 집합에서 약 137% 향상된 성능을 보였다. 일련의 중심어주도패턴에 대한 실험으로부터 제안된 기법이 기준성능보다 향상된 성능을 보임을 알 수 있었고, 특히 제안된 기법의 중심어주도패턴들 중 술어가 가지는 주제 영역을 중심으로 주제를 할당하는 패턴4가 월등히 우수한 성능을 보임을 알 수 있다.

표 4 기준성능에 대한 제안된 기법의 성능향상(F-score)

	학습 집합	실험 집합
비교기준	0.4192	0.3743
중심어주도패턴1	0.5875 (+40.12%)	0.5375 (+43.60%)
중심어주도패턴2	0.9491 (+126.36%)	0.3802 (+1.58%)
중심어주도패턴3	0.8511 (+102.98%)	0.7634 (+103.95%)
중심어주도패턴4	0.9886 (+135.77%)	0.8857 (+136.61%)

4.3 주제할당에 대한 오류 분석

본 절은 교차 검증에 사용된 다섯 개의 학습 및 실험 집합 중 하나의 집합을 대상으로 비교기준 기법 및 각 중심어주도패턴에 의한 주제 할당 오류를 분석하였다. 각 기법에 대하여 주제 할당 오류를 분석한 결과 다음과 같은 세 가지 원인에 의해 오류가 발생함을 관찰할 수 있었다. 주제 할당 오류에 대한 첫 번째 원인으로는 특정 주제가 편중되어 할당됨으로써 주제를 잘 못 할당하는 경우가 발생하는 것으로 분석되었다. 두 번째 원인으로는 같은 학습집합으로 학습하더라도 실험을 위한 학습 데이타양을 더 많이 필요로 하는 기법의 경우, 학습 데이타 부족에 의하여 오류가 발생하는 것으로 관찰되었다. 마지막으로 중심어가 가지는 주제 후보가 다양할수록 주제 할당에서 보다 정확한 주제를 선택하는데 어려움이 있는 것으로 분석되었다.

오류를 발생시키는 각 원인에 대하여 보다 자세히 설명하면 다음과 같다. 먼저 특정 주제가 패턴에 편중되어 할당됨으로써 발생한 오류는 비교기준 기법 및 중심어주도패턴1, 3에서 대부분 관찰되었다. 표 5에서 알 수 있듯이, 비교기준 기법의 경우 실험집합의 1,976개 평언에 주제를 할당한 결과 약 52.8%인 1043개의 평언에서 할당 오류를 보였으며, 그 중 39.4%인 778개의 평언이 "활동"으로 할당되어 발생한 오류이었다. 동사를 중심어로 간주하는 패턴1과 술어성명사를 중심어로 간주하는 패턴3 또한 한 주제로 편중되어 주제가 할당되는 현상을 보였다. 이러한 현상은 학습집합을 비교기준 기법 및 중심어주도패턴1, 3으로 재구성하여 학습에 적용하였을 때 학습집합이 특정 주제에 편중되어 구성되기 때문으로 추측된다. 표 5의 결과로부터, 특정주제가 패턴에 편중되어 할당됨으로써 비교기준 및 중심어주도패턴1, 3이 중심어주도패턴4 보다 표 4에서 낮은 성능을 보임을 알 수 있다.

표 5 특정주제 할당 오류(실험집합)

	할당오류	특정주제할당
비교기준	0.5282	0.3941 (활동)
중심어주도패턴1	0.4161	0.3692 (활동,역임)
중심어주도패턴3	0.1913	0.1213 (이동)

실험을 위한 학습집합이 부족하여 발생하는 주제 할당 오류는 일반명사를 중심어로 간주하는 중심어주도패턴2의 경우 주로 관찰되었다. 즉, 일반적으로 동사나 서술성 명사에 비해 일반 명사가 다양하게 발생하므로, 일반명사를 중심어로 간주하는 패턴2는 동사나 서술성명사를 중심어로 간주하는 패턴1에 비해 각 명사 당 활용할 수 있는 학습 집합이 부족한 것이다. 또한 학습에 사용되지 않은 다양한 일반명사가 출현하는 것도 성능저하의 한 요인이 되었다. 실제 패턴2로 구성된 실험집합에 주제를 할당한 결과 60.9%의 주제할당 오류가 있었으며 그중 27%는 학습에 수행되지 않은 명사가 출현하여 주제 할당에 실패하였다.

주제 할당 오류에 대한 마지막 원인으로 중심어가 가지는 주제 후보가 다양할수록 보다 정확한 주제를 선택

하는데 어려움이 있는 것으로 분석되었다. 본 논문에서 문장의 주제를 결정하는 기법은 그림 5에서 이미 설명하였듯이 중심어가 가지는 다양한 주제후보 중 하나의 정답 주제를 선택하는 방식으로, 각 중심어에 대한 주제후보 개수가 정답 추출에 영향을 줄 수 있다. 표 6은 각 평언 혹은 각 중심어에 대한 평균 주제후보 개수 및 주제후보로부터의 주제 추출 정확율을 제시한다. 주제 추출 정확율은 시스템에 의해 추출된 주제후보 개수에 대한 정답 주제 개수의 비율로, 각 중심어 당 추출된 주제후보 개수가 낮고 정답 주제를 보다 정확히 추출할수록 좋은 성능을 보인다. 즉, 비교기준 기법의 경우 각 평언 당 평균 주제후보 개수는 29개이며, 추출된 주제후보로부터 정답 주제가 추출될 정확율은 1.6%로 매우 낮다. 그러나 중심어주도패턴4의 경우 각 중심어 당 평균 주제 후보 개수는 약 1.5개이며 추출된 주제후보로부터 정답 주제가 추출될 정확율은 84.15%로 다른 패턴들에 비해 매우 높다. 표 6의 결과로부터 알 수 있듯이, 중심어주도패턴4가 적은 수의 주제후보로부터 보다 정확히 주제를 추출함으로써 표 4의 주제할당에서 높은 성능을 보임을 알 수 있다.

표 6 평균 주제후보 개수 및 주제 추출 정확율

	평균개수	주제추출 정확율
비교기준	29/α	0.0163
중심어주도패턴1	14.83/V	0.3316
중심어주도패턴2	3.45/N	0.1177
중심어주도패턴3	13.34/P_N	0.4614
중심어주도패턴4	1.47/P	0.8415

비교기준 기법 및 각 중심어주도패턴에 대한 일련의 오류 분석을 통하여, 같은 평언 집합으로부터 중심어주도패턴을 재구성할 경우 패턴4가 보다 적은 수의 주제후보를 비교적 정확하게 추출할 수 있으며, 각 주제 범주에 대한 학습 집합을 균등하게 분포시킴으로서 비교기준 기법 및 다른 패턴에 비해 보다 나은 성능을 보이는 것으로 추측된다.

5. 결 론

본 논문에서는 백과사전을 도메인으로 하는 시스템에서 성능 향상에 중요한 역할을 할 수 있는 중심어주도패턴에 기반한 문장주제 할당 기법을 제안하였고, 문장에 주제를 할당할 수 있는 네 가지 형태의 중심어주도패턴을 제안하였다. 백과사전 인물분야 2381 문서에서 추출된 9,987 평언을 대상으로 각 중심어주도패턴의 성능을 분석한 결과, 제안된 패턴 중 술어 기반의 중심어주도패턴4가 학습 집합에 있어 98.96% 및 실험집합에 있어 88.57%로 가장 효과적인 주제할당 성능을 보임을 알 수 있었다. 제안된 문장주제 할당 기법은 정보검색이나 질의-응답 시스템뿐 아니라, 요약, 담화분석 등 자연어처리시스템의 성능향상에 중요한 역할을 할 것으로 기대된다.

참 고 문 헌

[1] Kupiec, J., "MURAX: A Robust linguistic approach for question answering using an on-line encyclopedia," Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval, pp. 181-190, 1993.

[2] Maedche, A., Ontology Learning for the semantic web, Kluwer academic publishers, 2002.

[3] Moens, M-F., Automatic indexing and abstracting of document texts, pp. 103-132, Kluwer academic publishers, 2000.

[4] Lewis, D. D., Schapire, R. E., Callan, J. P. and Papka, R., "Training algorithms for linear text classifiers," Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, pp. 298-306, 1996.

[5] Schütze, H., Hull, D. A., and Pedersen, J. O., "A comparison of classifiers and document representations for the routing problem," Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval, pp. 229-237, 1995.

[6] Yang, Y., "An evaluation of statistical approaches to text categorization," Information retrieval, Vol. 1, pp. 69-70, 1998.

[7] 박기림, 장유진, 김민구, 박승규, "문서 내의 주제정보를 이용한 개선된 링크분석 알고리즘", 한국정보과학회 가을학술발표논문집(1), pp. 7-9, 2003.

[8] 이시은, 황인준, "의미 구역에 기반한 관련 웹 페이지 요약 기법", 한국정보과학회 봄학술발표논문집(B), pp. 597-599, 2003.

[9] 정태진, 장병탁, "강화학습을 이용한 웹 정보 검색", 한국정보과학회 가을학술발표논문집(II), pp. 94-96, 2001.

[10] 서혁, "담화의 구조와 주제구성에 관한 연구", 서울대학교 박사학위 논문, 2002.

[11] Cunningham, J. W. and Moore, D. M., "The Confused world of main idea," In J. F. Baumann.(ed.), Teaching main idea, N.Y.: IRA, 1986.

[12] Fellbaum, C. et al., WordNet: An electronic lexical database, pp. 23-46, The MIT press, 1998.

[13] François, J. and Denhiére, G., "Etude expérimentale de la validité cognitive d'un classement aspectuel et actanciel des prédications," verbum 3, pp. 117-138. 1992.

[14] Kékenbosch, C. and Bromberg, M., "Metacategories and sentence classification," Journal of pragmatics, Vol. 35, pp. 1-22, 2003.
[15] 이성헌, "전자사전구축을 위한 언어기술의 한 방법: 대상부류", 언어학, 제30권, pp. 185-206, 2001.

부 록

아래 표는 70여개의 평언으로부터 구성된 중심어주도 패턴4에 대하여 수동으로 할당한 정답주제와 제안된 시스템에 의해 할당된 주제를 제시한다.

강 보 영
1997년 경북대학교 컴퓨터공학과 졸업 1999년 경북대학교 대학원 영어영문학과(문학석사). 2002년 경북대학교 대학원 컴퓨터공학과(공학석사). 2004년 8월 경북대학교 대학원 컴퓨터공학과(공학박사). 2004년 9월~현재 한국정보통신대학교(ICU) 박사후연구원. 관심분야는 자연어처리, 정보검색, 질의응답, 문서분류, 의료문서정보처리 등임. AIRS 2005 Organization Committee Member, Information Sciences, Knowledge and Information Systems 심사위원으로 활동

맹 성 현
1983년 미국 캘리포니아 주립대학 학사 1987년 미국 Southern Methodist University(SMU) 석사 및 박사. 미국 Temple University 조교수. Syracuse University 종신교수. 충남대학교 교수 역임. 현재 한국정보통신대학교(ICU) 공학부 교수 관심분야는 정보검색, 텍스트마이닝, 디지털도서관, 시맨틱 웹 등임. 2002년 ACM SIGIR Conference Program Committee Chair, AIRS 2004 Program Committee Chair, Information Processing & Management, Journal of Natural Language Processing, Journal of Computer Processing of Oriental Languages 편집위원 등으로 활동. Home page: http://ir.cnu.ac.kr.

번호	N	P_N	V	정답	할당
1	LOCATION	null	머물	거주	거주
2	재능	물려	받	계승	null
3	전통	이어	받	계승	계승
4	PERSON	null	사귀	관계	관계
5	인연	null	맺	관계	이동
6	와일드	교유	하	관계	관계
7	로댕	관련	되	관계	관계
8	혁명	지도	하	교육	교육
9	문학	공부	하	교육	교육
10	경제학과	졸업	하	교육	교육
11	왜성	발견	하	발견	발견
12	석굴	발견	하	발견	발견
13	準位	null	알아내	발견	발견
14	사건	발생	하	발생	발생
15	반란	null	일어나	발생	발생
16	산업혁명	시작	되	발생	발생
17	운동	null	일어나	발생	발생
18	건강	회복	되	변화	null
19	마르크스주의	전향	하	변화	변화
20	민중	변화	시키	변화	변화
21	LOCATION	null	죽	사망	사망
22	南昌	null	죽	사망	사망
23	충도	병사	하	사망	사망
24	파리	요절	하	사망	사망
25	연구소	설립	하	설립	설립
25	아학	설치	하	설립	null
27	군단	설치	되	설립	설립
28	반란	토벌	하	싸움	싸움
29	이집트	원정	하	싸움	싸움
30	영국	패	하	싸움	싸움
31	장관	취임	하	역임	역임
32	단장	선임	되	역임	역임
33	총판	임명	되	역임	역임
34	금상	수상	하	수상	수상
35	회	입상	하	수상	수상

번호	N	P_N	V	정답	할당
36	함수	고안	하	연구	연구
37	문화	연구	하	연구	연구
38	연구	모두	하	연구	연구
39	복사	관찰	하	연구	연구
40	연구	null	하	연구	연구
41	耆老所	null	들어가	이동	이동
42	이후	귀국	하	이동	이동
43	이름	null	붙	이름	null
44	본명	null	따르	이름	이름
45	별명	null	얻	이름	이름
46	책	발간	되	작품	작품
47	명곡	편곡	하	작품	작품
48	작품	각색	하	작품	작품
49	가요	작곡	하	작품	작품
50	기법	사용	하	작풍	활동
51	미학	null	담	작풍	작풍
52	건설	묘사	하	작풍	작풍
53	학파	형성	하	조직	조직
54	집단	조직	하	조직	조직
55	기구	조직	하	조직	조직
56	부상	체포	되	죄	죄
57	형	선고	받	죄	죄
58	분원	복역	하	죄	죄
59	차례	투옥	되	죄	죄
60	혁명	참가	하	참가	참가
61	전투	참가	하	참가	참가
62	대회	출석	하	참가	참가
63	운동	참여	하	참가	참가
64	타운	출생	하	출생	출생
65	경상북도	null	태어나	출생	출생
66	역할	인정	받	평가	평가
67	존경	null	받	평가	평가
68	간사	활약	하	활동	활동
69	감독	데뷔	하	활동	활동
70	재건	null	힘쓰	활동	활동