정보검색에서 어휘체인을 이용한 효과적인 색인어 추출 방안

(An Efficient Index Term Extraction Method in IR using Lexical Chains)

강 보 영 † 이 상 조 ††

(Bo Yeong Kang) (Sang Jo Lee)

요 약 정보 검색(Information Retrieval)이나 디지털 도서관(Digital Library)과 같은 분야에서 가장 중요한 요소는 사용자가 필요로 하는 정보를 찾아주는 것이다. 이를 위해서 사용자가 사용하는 장치는 사용자의 의도뿐만 아니라 문서의 내용 또한 잘 파악하여야 한다. 본 논문은 문서의 의미적인 내용을 파악하는데 도움을 주는 효과적인 키워드 추출 시스템을 제안한다. 제안된 시스템은 문서에서 추출된 명사들의 의미(sense)를 결정(disambiguation)하고, 의미가 결정된 명사로 어휘체인을 생성한다. 특정 척도를 이용하여 강한 체인을 선별하고, 몇 개의 강한 체인에서 키워드들을 추출한다. 문서에서 사용된 명사들의 실제 센스를 결정하는 단계에서 *semantic window*라는 개념을 제안한다. 이것은 주변 명사들과의 의미관계를 미리 살펴보고, 문서내의 명사들의 센스를 결정하는 것이다. 본 시스템의 성능을 검증하기 위하여, 주요 구(key phrase) 추출 시스템인 KEA의 성능과 비교 분석하였다. 본 시스템은 정보 검색과 디지털 도서관을 포함한 범용적인 도메인에서 유용하게 사용될 수 있을 것으로 판단된다.

키워드 : 색인어 추출, 어휘 체인, 정보 검색, 키워드 추출

Abstract In information retrieval or digital library, one of the most important factors is to find out the exact information which users need. In this paper, we present an efficient index term extraction method which makes it possible to guess the content of documents and get the information more exactly. To find out index terms in a document, we use lexical chains. Before generating lexical chains, we roughly disambiguate the senses of nouns in a document using specific concept, called semantic window. Semantic window is that we look ahead semantic relations of peripheral nouns and disambiguate the senses of nouns. After generating lexical chains with sense-disambiguated nouns, we find out strong chains by some metrics and extract index terms from a few strong chains. We evaluated our system, using results of a key phrase extraction system, KEA. This system works in general domains of documents including Information Retrieval and Digital Library.

Key words : Index Extraction, Lexical Chain, Information Retrieval, Keyword Extraction

1. 서 론

인터넷상에서 정보 검색 시스템은 사용자가 필요로 하는 정보를 찾기 위하여 문서의 내용뿐만 아니라 사용자의 의도를 파악하고, 주어진 질의에 적합한 문서를 검색해야 한다. 그러나 인터넷상의 방대한 문서들은 문서를 대표하는 색인어가 없는 경우가 많고, 대표 색인어를 가진 문서라도 그 문서의 의미적인 내용을 적합하게 표현하지 못하는 경우를 흔히 발견할 수 있다. 만약 문서의 의미적인 내용을 표현하는 좋은 지시자나 색인어가 있다면, 텍스트 마이닝, 문서 요약 또는 문서 클러스터링과 같은 다양한 연구 분야에 많은 도움을 줄 것으로 판단된다. 본 논문에서는 문서내의 의미적인 관계에 기반하여 문서의 내용을 추측하고 보다 정확한 정보를 찾는데 도움을 줄 수 있는 효과적인 색인어 추출 시스템을 제안하고자 한다.

색인어를 탐색하는 기법은 크게 추출 색인 기법(extraction indexing)과 할당 색인 기법(assignment indexing)이 있다. 추출 색인 기법은 문서로부터 직접 추출된 용어를 문서의 내용을 표현하는 색인어 집합으

† 학생회원 : 경북대학교 컴퓨터공학과
comeng99@hotmail.com
†† 종신회원 : 경북대학교 컴퓨터공학과 교수
sjlee@bh.kyungpllk.ac.kr
논문접수 : 2001년 8월 22일
심사완료 : 2002년 5월 28일

로 선택하는 방식이고, 할당 색인 기법은 통제 용어 사전(controlled term vocabulary)에 이미 정의된 색인 용어를 문서에 분배하는 방법이다[1][2]. 할당 색인 기법의 경우 문서들의 패턴을 파악하여 적합한 통제 용어 사전을 자동으로 구축하는데 노력과 비용이 많이 드는 어려움이 있다. 추출 색인 기법은 문서내의 단어들이 잠재적으로 모호한 경향을 가진다는 단점이 있으나[2], 문서의 내용을 표현할 용어들을 선택하는데 도움을 줄 수 있는 담화구조와 관련된 언어학적인 현상에 관한 연구들이 추출 색인 기법의 성능을 개선하는데 도움이 되고 있다[3][4][5][6].

본 논문은, 현재 쉽게 이용 가능한 통제 용어 사전이 없고 시소러스를 구축하는데 오랜 시간이 걸리기 때문에, 문서로부터 직접 색인어를 추출하는 추출 색인 기법을 사용하도록 한다. 추출 색인 기법의 본질적인 단점을 보완하기 위해, 색인어 탐색 과정에서 명사의 의미를 결정(sense disambiguation)하여 의미(sense)가 결정된 색인어를 추출할 수 있도록 하였고, 담화 구조와 관련된 연구 중 Barzilay와 Elhadad의 어휘 체인(lexical chains)에 관한 연구를 색인어 탐색에 적용한다. 또한 어휘 체인을 색인어 탐색에 적용하는데 있어서 명사 의미 결정 과정을 강화시킴으로써 Barzilay와 Elhadad의 연구를 확장하였다[7]. 우리는 문서에서 중요하게 사용된 단어들이 문서의 의미적인 내용을 반영할 수 있다고 가정하였다. 문서에서 중요한 명사를 찾아내기 위해서 반복, 동의어, 상위어, 하위어, 반의어 관계 등을 가지는 명사들을 연결함으로써 어휘 체인을 생성하고 색인어를 탐색해낸다. 어휘 체인 생성시에 높은 점수가 부여되는 명사들은 문서의 내용을 나타내는데 중요한 요소로 간주되고, 문서의 지시자나 색인어가 될 가능성이 높아진다. 그러나 문서내의 명사들은 의미(sense)가 모호하므로, 본 논문은 어휘 체인을 생성하기 전에 명사들의 의미를 결정하는 과정을 제안하고, 명사의 의미가 *semantic window*라는 개념을 이용하여 결정될 수 있음을 보일 것이다.

본 논문의 구성은 다음과 같다. 2절에서는 추출 색인 및 어휘 체인과 관련된 기존 연구를 살펴본다. 그리고, 3절에서는 본 논문에서 제안한 색인어 추출 기법을 자세히 기술하고, 4절에서는 제안한 시스템을 실험, 분석한다. 마지막으로 5절에서 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

일반적으로 자동 색인에서는 문서로부터 인덱스 용어를 추출하는 추출 색인 기법이 사용된다[2]. Moens은 문서로부터 인덱스 용어를 선택하는 과정의 대부분은 다음과 같은 여섯 단계로 이루어진다고 하였다.

단계1. 문서 내에서 개별 어휘들을 찾아낸다.
단계2. 불용어 목록을 이용하여 문서 내용을 불충분하게 표현하는 기능어 등을 제거한다.
단계3. 남은 단어들을 어근 형태로 변형한다.
단계4. 인덱스 용어를 선택적으로 구(phrases)로 형성한다.
단계5. 단어 어근 혹은 구들을 시소러스 클래스 용어로 선택적으로 대치(replacement)한다.
단계6. 단어 어근, 시소러스 클래스 용어 혹은 구의 가중치를 계산한다.

추출 색인 기법에는 몇 가지 단점이 있다[2]. 텍스트내의 단어들이 잠재적으로 모호한 성격을 가진다는 점이다. 또한, 텍스트내의 단어들과 구들은 문서를 표현하는데 있어 너무 구체적이기 때문에, 문서내의 보다 일반적인 정보들을 탐색하는데 어려움이 있다.

따라서 이러한 추출 색인 기법의 단점을 보완하기 위하여 담화 구조 및 관련된 언어학적 현상에 기반한 연구들이 대두되고 있다[3][4]. Luhn은 문서 색인을 위해 담화 구조를 사용하려는 시도를 하였다[8]. Bookstein은 서로 연관 있는 단어들이 발생하는 경향 또한 용어를 선택하는데 유용하게 고려된다고 제안하였고[9], Liddy와 Myaeng, Burnett등은 인덱스 용어 선택과 가중치 부여는 문서 내의 제목, 요약, 첫번째 단락 등 구조적인 위치에 의해 결정될 수 있다고 하였다[10][11]. 또한 Salton은 다양한 주제에 따라 문서들의 구조적인 분해에 대한 연구가 또한 문서내에서 중요한 화제를 찾아내는데 유용할 수 있다고 제안하였다[12].

Morris와 Hirst는 어휘 체인을 찾음으로써 문서내의 담화구조를 결정할 수 있고, 어휘 체인이 어휘 모호성을 해결하는데 충분한 문맥을 제공한다고 제안했다[5][6]. Hallyday와 Hasan 또한 만약 문장들 간에 응집성이 있고 일관적으로 기술되었다면, 그 문서 내에 연속되는 문장들은 이전에 언급된 개념이나 그것들과 연관된 다른 개념들을 가리키는 경향이 있다고 하였다[13]. 이와 같은 측면에서 문서 내의 단어들은 응집력이 있는 체인을 형성하는 경향이 있음을 알 수 있다. 다시 말해, 체인내의 각 단어들은 동일 참조(identity of reference)와 같은 특별한 응집성 관계에 의해 선행 단어들과 연관되는 것이다. 예를 들면, 예문 (1)에서 보는 바와 같이 이탤릭체로 된 단어들은 동일 참조의 관계로 체인을 형성한다.

(1) The major potential complication of total joint replacement is *infection*. *It* may occur just in the

area of the wound or deep around the prosthesis. *It may occur during the hospital stay or after the* patient goes home…*Infections* in the wound area are generally treated with antibiotics.

또한, 체인에서의 단어들은 하위어, 부분어 관계나 일반적인 연상 관계(general association of ideas)와 같은 방법으로 다른 단어들과 연관된다. 다음 예문 (2)는 하위어 관계 체인을 보여주고 있고, 예문 (3)은 일반 연상 관계 체인을 보여준다.

(2) The major potential *complication* of total joint replacement is *infection.*

(3) The evening prior to admission, take a *shower* or bath, *scrubbing* yourself well. Rinse off all the soap.

Hasan은 어휘 체인을 문서의 응집성을 측정하는 도구로 개발하였고[14], 문서 내 단어들과 이러한 단어들을 연결하는 관계들로 구성된 선형 연결 리스트로 체인을 규정하였고, 일곱 가지의 어휘 관계 유형[1]과 여섯 가지 유형[2]의 체인들 사이의 관계를 정의하였다[15]. 이에 반해 Morris와 Hirst는 부분어, 동의어, 반의어, 반복, 연어, 계층어(taxonomy)로 여섯 가지 유형의 어휘 관계를 정의 하였고, 어휘 체인들은 서로 관계성을 가지지 않는다고 하였다[5][6]. 또한 Roget 시소러스를 사용하여 어휘 체인을 자동으로 만들어내는 알고리즘을 고안하였지만, 이용 가능한 온라인 시소러스가 없었기 때문에 구현하지는 못했다[15].

Al-Halimi와 Kazman 또한 어휘 체인을 사용한 문서의 색인 방법을 제안하였다[15]. 그들은 화제에 의해 색인을 하는 과정에 초점을 맞추었고, 기존의 일차원적 구조의 어휘 체인의 개념에서 이차원적 구조의 어휘 트리의 개념을 제안하였다. 또한, 임의의 문서를 자동 색인하는데 어휘 트리가 유용하게 사용될 수 있음을 강하게 시사하는 연구 결과를 제시하였다.

Barzilay와 Elhadad는 이러한 Morris와 Hirst, Hallyday와 Hasan, Hearst등의 연구에 기초하여 어휘 체인을 문서 요약에 적용하였다[7]. 문서의 내용을 표현하는 중요한 명사를 찾기 위하여, Hearst의 알고리즘을 사용하여 먼저 문서를 단락화하고, 단락화한 문서를 기반으로 하여 어휘 체인을 구성하였다. 명사 의미 결정은 컴포넌트(component)를 이용하여 어휘 체인을 생성하면서 동시에 수행하였다. 본 논문은 이러한 Barzilay와 Elhadad의 어휘 체인 구성 방식 중, 정보 추출이나 색인 과정에서 중요하게 고려되는 명사 의미 결정 과정을 보다 강화하였다. 또한, 어휘 체인 생성과 명사 의미 결정을 동시에 수행하는 대신, 먼저 명사의 의미를 결정한 후 어휘 체인을 구성하였다. 명사의 실제 의미로 어휘 체인을 구성하는 방식이, 틀린 의미로 어휘 체인을 구성함으로써 생길 수 있는 불필요한 계산 과정들을 제거하기 때문에 더욱 효율적이라고 가정하였기 때문이다.

3. 어휘 체인을 이용한 색인어 추출 기법

3.1 제안된 시스템 구조

시스템의 전체적인 구성은 그림 1과 같다. 제안된 방법은 색인어를 추출하고자 하는 문장으로부터 명사를 추출한다. 추출된 명사는 다양한 의미를 갖고 있으므로, 주변 단어들과의 관련성을 탐색하기 위해서 제안한 *semantic window* 개념을 사용하여 명사의 의미를 결정한다. 그런 후, 의미가 결정된 명사들로 어휘 체인을 생성하고 어휘 체인 내에 있는 각 명사들과 체인들에 점수를 부여한다. Barzilay와 Elhadad의 변형된 수식에 의해 여러 체인들 중 강한 체인을 찾아내고 강한 체인들 속에서 색인어를 추출한다. 언급한 각 단계에 대한 상세한 설명은 3.2절에서 기술하겠다. 제안된 접근법은 의미가 결정된 명사들로 어휘 체인을 구성함으로써 모호한 의미로 어휘 체인을 구성하는데 발생하는 불필요한 계산 과정들을 제거할 수 있으며, 의미가 결정된 명사를 색인어로 추출함으로써 색인 추출 기법의 단점을 보완할 수 있는 장점이 있다. 제안된 시스템에서는 명사 의미 결정 및 어휘 체인 구성, 점수 부여의 각 단계에 있어서 다음과 같은 사실들을 가정한다.

가정1. 주변 명사들과 가장 많은 관련성을 가진 명사의 의미가 그 명사의 실제 의미가 될 확률이 높다.

가정2. 명사의 의미가 주변 명사와 비슷한 관련성을

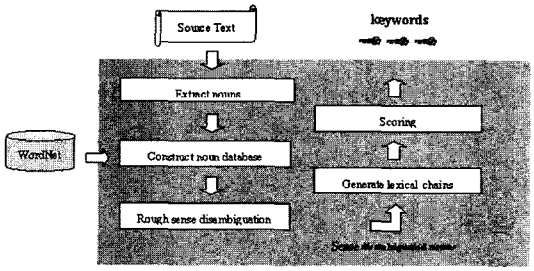


그림 1 제안된 색인어 추출 시스템의 전체적인 구성

1) synonym, meronym, repetition, taxonomy, antonym, co-taxonomy, and co-meronymy

2) epithet-thing, medium-process, process-phenomenon, actor-process, process-goal, and process-location of proces

가진다면, 보다 자주 사용되는 의미가 실제 의미가 된다.

가정3. 반복, 동의어, 반의어 관계의 점수가 상위어, 하위어 관계보다 어휘체인에서 중요한 명사를 탐색하는데 더 정보적(informative)이다

3.2 각 단계별 설계

3.2.1 명사 추출과 명사 의미 결정

제안된 시스템은 명사들만 사용하여 어휘 체인을 생성하기 때문에 문서로부터 명사를 추출하고, 각 명사의 동의어, 상위어, 하위어, 반의어를 워드넷(WordNet)에서 탐색하여 명사 데이타베이스를 구축한다. Princeton 대학은 워드넷 소프트웨어와 데이타 베이스뿐만 아니라 소스 코드를 연구자료로 제공한다. 먼저 입력된 명사에 대해 그에 해당하는 동의어, 상위어, 하위어 반의어들을 텍스트 파일에 출력으로 받을 수 있도록, 목적에 적합하게 워드넷 소스 코드를 수정하였다. 상위어는 2단계 위의 상위어까지 추출하도록 설계하였고, 하위어로는 5단계 아래의 하위어까지 추출하도록 하였다. 그런 후, 출력된 텍스트 파일을 사용하여 명사 데이타 베이스를 구축하였다. 그림 2는 워드넷을 사용하여 구축된 명사 데이타 베이스의 예를 나타낸다.

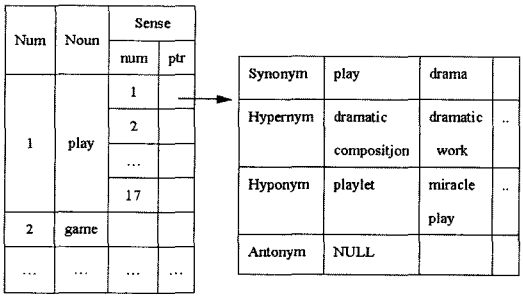


그림 2 명사 데이타베이스 예

추출된 명사들은 다양한 의미를 가진다. 예를 들면, 명사 *act*는 *a legal document, human activity, a subdivision of a play or opera* 등의 의미를 가진다. 이러한 다양한 의미들 중, 우리는 문서에서 그것이 실제로 사용되는 의미를 찾아내어야 한다. 이 때, Barzilay와 Elhadad는 컴포넌트를 이용하여 의미 결정과 어휘 체인 생성을 동시에 수행하였다[7]. 즉, 명사가 가진 모든 의미로 어휘 체인을 구성한 후, 그 중 가장 점수가 높은 어휘 체인에 연결된 의미를 그 명사의 의미로 채택하였다. 그러나 본 논문은 먼저 명사의 의미를 결정한 후 체인을 구성한다. 명사의 의미 모호성을 제거한 후 어휘 체인을 구성하는 것이, 모호한 의미로 어휘 체인을 구성함으로써 생길 수 있는 불필요한 계산 과정들을 제거하기 때문이다.

문서가 일관성이 있다면, 그 문서는 연관 있는 단어들로 구성되어 있다. 따라서, 주변 명사들과의 의미적인 관계를 찾아내고 명사의 의미를 결정하기 위하여 *semantic window* 개념을 제안한다. 의미 결정 과정에서 사용되는 의미 관계 유형은 반복, 동의어, 상위어, 하위어, 그리고 반의어이다. *Semantic window*의 개념은 다음 그림 3과 같다: 만약 명사 *run*의 센스 1이 윈도우 내에서 명사 *tally* 하나와 관련성을 가지고, 센스 2가 윈도우 내에서 두 명사, *trial*, *test*와 연관성을 가진다면, *run*의 의미는 센스 2가 될 것이다.

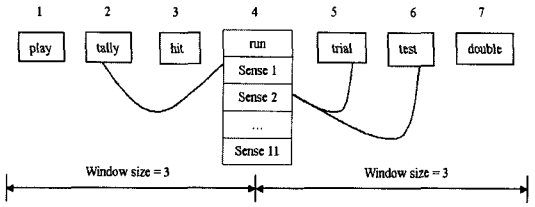


그림 3 Semantic window의 개념

그림 4는 의미 결정 과정을 보다 상세히 나타내고 있다. 그림 4는 *semantic window*를 사용하여 명사 *run*의 의미를 결정하는 과정으로, 윈도우 사이즈가 3일 때 *run* 주변 6 단어들과 어떤 관계를 가지는지 관찰하게 된다. 즉, 윈도우 사이즈 내에서 주변 단어들과 관련성이 가장 많은 의미를 찾아내는 것이다. 명사의 각 센스의 점수는 그 의미의 synset(synonym set)이 윈도우 사이즈 내에

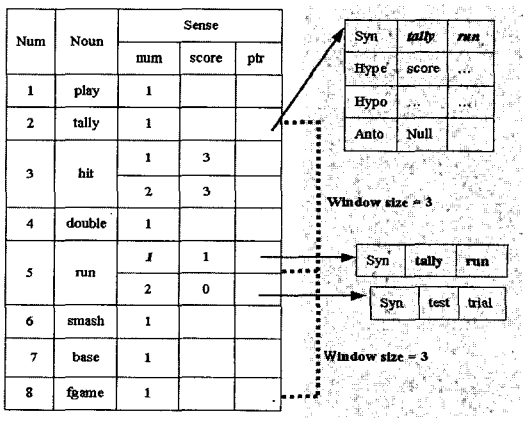


그림 4 명사 run의 의미 결정

서 관련성을 가지는 명사들의 수에 의존한다. *run*의 첫 번째 센스의 synset인 *tally*와 *run*은 두 번째 명사 *tally*와 관련성을 가진다. 따라서, 주변의 명사 하나와 연관성이 있으므로 점수 1을 얻는다. *Run*의 두 번째 센스의 synset인 *test*와 *trial*은 주변의 어떤 명사와도 관계를 가지지 않으므로 점수를 얻지 못한다. 따라서, 이 문맥에서 *run*의 의미는 센스 1이 된다. 만약, 센스들의 점수가 같다면 자주 쓰이는 센스를 그것의 의미로 설정한다. 예를 들면, 그림 3에서 세 번째 명사 *hit*의 센스들의 점수가 같을 때 센스 1을 *hit*의 의미로 선택한다.

3.2.2 어휘 체인 생성 및 점수 부여

의미가 결정된 명사로 어휘 체인을 생성하고 어휘 체인 내의 각 체인과 명사에 점수를 부여하는 방법은 다음과 같다. 체인 내에서 명사들의 점수는 다른 명사들과의 관계에 의해 결정된다. 이 때, 체인 내에서 명사가 가지는 관계는 반복, 동의어, 상위어, 하위어, 반의어의 다섯 가지 유형이다. 만약 어떤 명사가 다른 명사들과 하나의 동의어 관계를 가진다면 동의어 관계에 해당하는 점수를 얻을 것이다. 유사한 방식으로, 반복, 반의어, 상위어 그리고 하위어 관계를 가지는 명사들도 해당 관계에 해당하는 점수를 부여받는다. 일반적으로, 반복 관계의 점수가 가장 높고, 동의어와 반의어 관계 점수가 상위어나 하위어 관계 점수보다 더 높다. 예를 들면, 그림 5에서 동의어와 반의어 관계 점수는 2이고 상위어와 하위어 점수는 1이라고 가정하자. 만약 명사 *run*이 *tally*와 하나의 동의어 관계를 가지고 *play*와 하나의 하위어 관계를 가진다면, *run*의 점수는 3이 될 것이다.

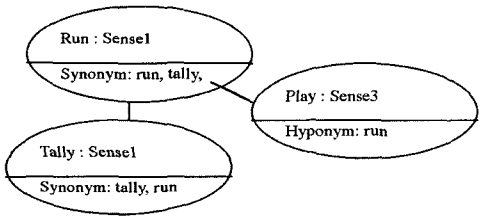


그림 5 체인에서 명사 run의 점수

위의 과정을 수식으로 표현하면 수식 (1)과 같다. 제안된 수식에서 $S_{NOUN}(N_i)$는 어휘 체인 내의 명사 N_i의 점수(score)이고, NR^k_{Ni}는 관계 k에 대해 명사 N_i가 가지는 어휘관계의 수(number of relations)를 나타낸다. 그리고 SR^k_{Ni}는 관계 k의 점수(score of relation)를 표현한다. 따라서 어휘체인 내의 명사 N_i의 점수는 주어진 다섯 가지의 관계에 대한 빈도와 점수의 곱으로 계산된다.

$$S_{NOUN}(N_i) = \sum_k (NR^k_{Ni} \times SR^k_{Ni}),$$
$$\text{where} k \in \{\text{indentity}, \text{synonym}, \text{hypernym}, \text{hyponym}, \text{antonym}\} \quad (1)$$

제안된 시스템에서는 명사에 점수를 부여할 때, 그 명사와 직접적인 연관을 가지고 있는 명사들만 고려한다. 예를 들면, 그림 6과 같이 명사 a의 점수는 명사 b와 c에 의해서만 결정되도록 한다. 왜냐하면, 명사들의 관계가 너무 다양하기 때문에, 점수를 매기기 어려운 경우들이 종종 발생하기 때문이다: 명사 a와 명사 b가 하위어 관계이고, 명사 b와 명사 d가 반의어 관계일 때, 명사 a와 명사 d의 관계를 결정할 수 없다. 또한, 명사 c와 d가 반의어 관계이고 e와 c가 상위어 관계일 때도, a와 e가 어떤 관계에 있는지 결정하기 쉽지 않다. 따라서, 만약 d와 e를 고려한다면 명사 a에 점수를 부여하는 것이 쉽지 않을 것이다.

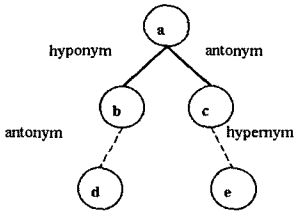


그림 6 점수 매김의 어려운 예

만약 어휘체인에서 명사들만 점수를 매기고 체인 그 자체에는 점수를 부여하지 않는다면, 그림 7과 같은 경우 명사 a와 b에 부여된 점수가 같으므로 두 명사는 동등하게 중요한 명사인 것으로 보인다. 그러나 명사 a보다는 명사 b가 문서의 내용을 표현하는데 보다 적합하다.

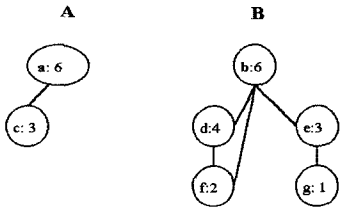


그림 7 명사 a와 b의 중요도

체인의 점수는 그 체인에 속해있는 명사들의 점수의 합과 명사의 수에 의존한다. 즉, 체인 C_x의 점수는 그 체인에 속해있는 명사 N_i들이 부여 받은 점수와 N_i수의 합에 의해 결정된다. 체인 내에 존재하는 명사의 수를 n이라 하면, 체인에 점수를 부여하는 식 $S_{CHAIN}(C_x)$는 아래 (2)와 같다.

3.2.3 강한 체인(Strong Chains)과 색인어 추출

많은 점수를 부여 받은 체인에는, 점수가 높은 단어나 문서에서 사용된 명사들과 연관을 많이 가지는 단어가 포함하고 있기 때문에, 점수를 많이 받은 체인이 점수를 적게 받은 체인보다 더 중요하다고 가정한다. 따라서 점수가 높은 체인들은 문서의 내용을 표현할 색인어를 포함하고 있을 가능성이 높다.

3.2.2절에서 제안된 수식을 이용하여 어휘 체인에서 체인과 명사들에 점수를 부여한 후, 강한 체인을 탐색한다. *강한 체인(Strong Chain)*이란 문서로부터 구성되는 여러 체인들 중 문서의 개념을 대표할 수 있는 체인으로서, 강한 체인을 선별하는 기준은 Barzilay와 Elhadad의 수식을 그대로 적용한다[7]. 단, 표준편차에 곱해주는 상수를 추출하고자 하는 색인어의 수에 따라 조절해야 하므로 상수 로 표현한다. 극소수의 색인어만 추출하고 싶은 경우 값은 높아질 것이고, 여러 개의 색인어를 추출하고 싶다면 C값은 낮아질 것이다. 따라서 전체 체인의 수를 m이라 하면, 체인 C_x가 강한 체인이 되기 위해서는 다음 수식 (3)을 만족하여야 한다.

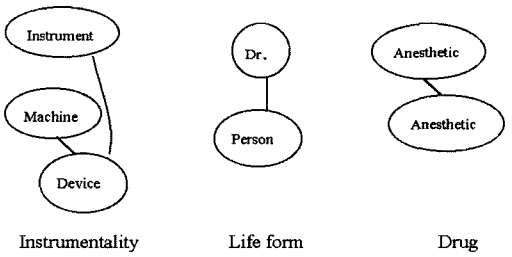
$$S_{CHAIN}(C_x) > M + C \cdot \sqrt{\frac{1}{M}\sum_{i=1}^{m}(S_{CHAIN}(C_i) - M)^2} \quad (3)$$

$$where,\ M = \frac{1}{m}\sum_{i=1}^{m} S_{CHAIN}(C_i)$$

강한 체인으로부터 색인어를 추출할 때, 가장 강한 체인에서만 색인어들을 추출할 것인지, 그렇지 않으면 몇 개의 강한 체인으로부터 다양한 색인어들을 추출해야 하는지 결정해야 한다. 본 논문은 후자의 방식을 채택한다. 문서는 단지 하나의 개념으로만 이루어진 것이 아니라 다양한 개념으로 이루어져 있기 때문에, 다양한 개념들로부터 색인어를 추출하는 것이 효율적이기 때문이다. 체인은 관련된 명사들의 연결로 이루어져 있기 때문에 하나의 체인은 하나의 개념을 표현하고 다양한 체인은 다양한 개념을 표현하는 것으로 간주될 수 있다. 예를 들면, 아래 단락에서 다음과 같은 체인이 생성된다.

Dr. Kenny has invented an anesthetic machine. This device controls the rate at which an anesthetic is pumped into the blood. But the cost of this instrument is too expensive, so only few persons can purchase it.

위의 단락은 *instrumentality*, *life form*, *drug*의 세 개념을 표현하는 세 체인으로 구성될 수 있다. 만약 우리가 *instrumentality*의 개념을 표현하는 가장 강한 체인 하나만으로부터 색인어를 추출한다면, 같은 개념을 나타내는 색인어들만을 추출하게 된다. 그러나, 위 단락은 단 하나의 개념으로만 이루어진 것이 아니라 여러 개의 개념으로 이루어져 있다. 따라서 *instrumentality*, *drug*와 같은 다양한 개념으로부터 색인어를 추출하는 것이 효율적이다.



3.3 시스템 동작 과정

Barzilay와 Elhadad에서 논의 되었던 다음의 짧은 예제 문장을 이용하여 제안된 시스템의 동작 원리를 보다 상세히 살펴보도록 한다.

(1) Dr. Kenny has invented an anesthetic machine. This device controls the rate at which an anesthetic is pumped into the blood. But the cost of this instrument is too expensive, so only few people can purchase it.

(2) Dr. Kenny has invented an anesthetic machine. The doctor spent two years on this research.

첫 예문은 *machine*에 두 번째 예문은 *doctor*에 초점을 맞추어 설명하므로 두 문장에는 의미적인 차이가 있다. 단순히 단어 빈도만 관찰하는 기존의 시스템들은 단어의 출현 빈도만 고려하므로 두 문장의 차이를 발견하지 못하지만, 제안한 시스템은 두 문장의 차이점을 식별해 낼 수 있다. 먼저 각 문장으로부터 다음과 같이 명사를 추출한다.

(1) 첫번째 예문의 명사들

[1] Dr. [2] Kenny [3] anesthetic [4] machine
[5] device [6]rate [7] anesthetic [8] blood
[9] cost [10] instrument

(2) 두 번째 예문의 명사들

[1] Dr. [2] Kenny [3] anesthetic [4] machine
[5] doctor [6] years [7]research.

윈도우를 이용하여 명사의 의미를 결정한 결과는 아래와 같다. *Kenny*의 센스가 0인 이유는 *Kenny*가 워드넷에 수록되지 않은 미등록어이기 때문에, 의미를 결정할 수 없기 때문이다. 의미가 결정된 명사들을 이용하여 그림 8, 9와 같이 어휘 체인을 생성한다.

(1) 첫번째 예문의 명사 의미 결정

[1] Dr. : 1 [2] Kenny:0 [3] anesthetic:1
[4] machine:1 [5] device:1 [6] rate:1
[7] anesthetic:1 [8] blood:1 [9] cost:1
[10] instrument:1

(2) 두 번째 예문의 명사 의미 결정

[1] Dr.:1 [2] Kenny:0 [3] anesthetic:1
[4] machine:1 [5] doctor:1 [6] years:1
[7] research.:1

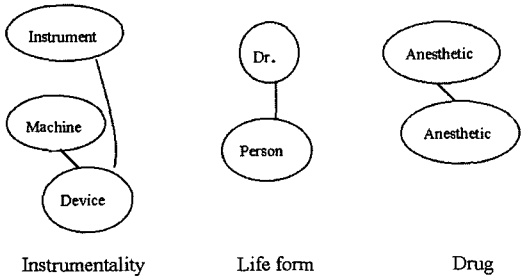


그림 8 첫 번째 예문의 어휘 체인

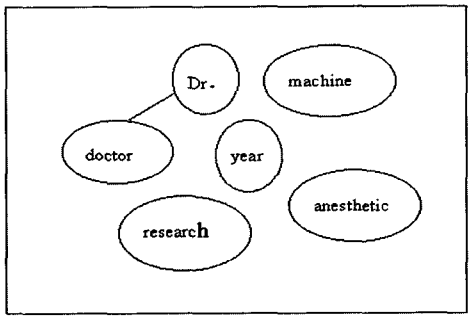
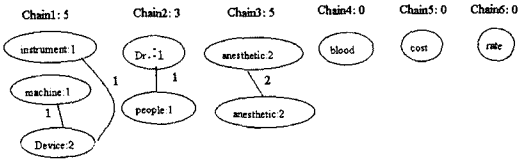


그림 9 두 번째 예문의 어휘 체인

첫번째 예문의 각 체인에 아래와 같이 점수를 매기고, 강한 체인 조건을 만족시키는 몇 개의 강한 체인으로부터 색인어를 추출한다: 동의어와 반복 관계의 점수는 2이고 상위어, 하위어 관계의 점수는 1이다. 강한 체인 공식에서 상수 C의 값은 0.6으로 하였을 때, 첫번째 예문에서 강한 체인을 선별할 기준 점수는 약 3.58이었다. 따라서 체인 1과 체인 3이 강한 체인으로 선택된다.

위 예문에서, 적은 수의 체인에서 몇 개의 강한 체인을 선별해야 하므로 상수 C의 값을 0.6으로 낮추었다. 만약 상수 C를 Barzilay와 Elhadad와 같이 2로 유지한다면, 어떤 강한 체인도 얻어낼 수 없다. 3.3절에서 언급



하였듯이, 상수 C의 값은 문장의 길이나 필요로 하는 색인어의 개수에 따라 적응시켜야 한다. 그 결과 첫번째 예문의 색인어로는 *anesthetic*과 *machine*이 추출되고, 두 번째 예문에서는 *doctor* 혹은 *Dr.*가 색인어로 추출된다.

이 결과는 첫번째 예문은 기계 그 자체를 묘사하고 있고, 두 번째 예문은 의사에 초점을 맞추고 있는 예문1과 예문 2의 차이점을 명확하게 보이고 있다. 통계적인 방식을 주로 사용하는 기존의 색인어 추출 시스템은 이러한 차이점을 찾아내지 못한다. 만약 우리가 복합명사나 워드넷에 등록되어있지 않은 미등록어등을 처리할 수 있다면, 첫번째 예문의 색인어로는 *anesthetic machine*이 그리고 두 번째 예문의 색인어로는 *Dr. Kenny* 혹은 *doctor*가 추출될 것이다.

4. 실험 및 평가

지금까지 색인어(keyword) 추출 시스템에 관한 정량적인 성능평가에 관한 연구는 거의 보고되지 않았다. 따라서, 중요어구(key phrase)추출 시스템인 KEA의 성능과 제안된 시스템의 결과를 비교함으로써 시스템 성능 평가를 수행하였다[16][17].

4.1 실험 환경

명사 의미 결정을 위한 실험에 있어서, 태깅된 Brown Corpus 문서 8건을 사용하였다. 각 문서의 내용이 너무 방대하기 때문에, 문서에서 첫번째 70에서 130 명사까지만 입력 문서로 하여 명사 의미 결정을 수행하였다. 그런 후 결정된 의미가 옳은지를 관찰하기 위하여 사람이 직접 의미 태깅을 한 SemCor 문서의 명사들과 비교함으로써 제안된 명사 의미 결정 모듈의 성능을 평가하였다.

색인어 추출 실험을 위해서는 온라인 ACM 논문 아카이브로부터 10개 논문의 10개 초록이 추출되었고, 10개 초록의 내용은 컴퓨터 과학 분야와 관련이 있었다. 그리고, 경북대학교 컴퓨터 공학과와 한국 과학 기술원에 재학중인 대학원생들로 구성된 다섯 명의 피실험자(subject)들이 초록으로부터 색인어를 추출하였다. 그런 후, 피실험자들이 추출한 색인어와 제안된 시스템이 추

출한 색인어를 비교함으로써 제안한 시스템을 분석하고, 중요어구 추출 시스템인 KEA와 제안된 시스템의 성능을 비교하였다. KEA는 중요 후보 어구를 탐색한 후, 각 후보들에 대한 특징 값을 계산한다. 이 때, 중요 어구 후보를 예측하기 위해 KEA는 기계학습 알고리즘을 사용한다. 기계학습 알고리즘은 훈련 문서를 사용하여 예측 모델을 만들고, 그 모델을 새로운 문서에서 중요 어구를 찾아내는데 사용된다. 제안된 시스템과 KEA는 몇 가지 면에서 차이점이 있다: KEA는 중요어구(key phrase) 추출 시스템이고 제안한 시스템은 색인어(key word) 추출 시스템이다. KEA는 모든 품사의 단어들을 추출하고 특정 도메인에서 동작하도록 설계되었지만, 제안된 시스템은 명사인 단어들만 추출하며 일반적인 도메인에서 작동하도록 고안되었다. 또한, 제안한 시스템의 실험에 사용된 데이타 컬렉션이 KEA에서 사용한 데이타 컬렉션과 다르다.

4.2 방법 평가(Method Evaluation)

4.2.1 명사 의미 결정

의미 결정을 위한 실험에서 다음 표 1과 같은 결과를 얻을 수 있었다. 의미 윈도우 크기가 5 일 때 가장 좋은 결과를 얻었다.

표 1 명사 의미 결정의 결과

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Ave-rage
win = 5	.75	.75	.73	.66	.70	.70	.68	.54	.69
win = 10	.77	.73	.64	.63	.67	.63	.68	.48	.65
win = 15	.69	.69	.58	.61	.69	.58	.66	.43	.62

만약 관찰되는 명사 주변에 과도한 수의 명사를 고려하게 되면, 명사의 의미가 혼동될 수 있다. 다시 말하면, 위의 결과는 관찰 명사 주변의 5 명사만 살펴보았을 때 선택된 센스는 옳은 경향이 강하지만, 명사 주변에 15명사를 고려하여 그 명사의 의미를 선택한다면, 선택된 의미는 비교적 틀린 경향이 강한 것으로 해석될 수 있다. 이것은 일정 개수 이상의 주

변 명사를 고려하게 되면, 상대적으로 관련이 적은 다른 문단에 있는 명사들의 의미에 의해 관찰되는 명사의 의미가 혼동될 가능성이 크기 때문이라고 추측한다. 예를 들면, 그림 10에서 *machine*의 의미는 윈도우 크기가 1 일 때 *efficient person*이고, 윈도우 크기가 1보다 커지면 *device*가 된다. 윈도우 크기를 확장 시켰을 때, *machine*의 의미는 두 번째 문단의 *device, instrument, robot*과 같은 명사에 의해 영향을 받아서 변하게 되는 것이다. 이러한 현상이 윈도우 크기가 클 때 명사 의미 결정의 성능을 저하시킨다고 예측된다. 더 나아가, 최적의 윈도우 크기는 텍스트 장르, 문단 길이 등 여러 요소에 의해 영향을 받을 수 있으므로, 최적의 윈도우 크기 결정에 대해서는 향후 연구가 필요하다. 본 논문에서는 실험결과를 기반으로 최적의 성능을 보이는 윈도우 크기로 명사 의미 결정을 하도록 한다.

Dr. Kang is a machine. The doctor works very hard and is called work worm. She usually works in her laboratory in Taegu.

She invented a marvelous *instrument* when she was young. The *device* could move along the road without any difficulty for itself. So many companies showed interest to her intelligent *robot*.

그림 10 machine의 의미 결정

4.2.2 시스템 분석

우리는 피실험자에게 문서의 내용을 가장 잘 표현하는 몇 개 단어에 1을 표시하고, 그렇지 않은 단어에는 0을 표시하도록 요청하였다. 예를 들면, 초록 3의 결과는 표 2와 같다: 표 2에서 *ambiguity*와 *agent*가 색인어로 선택되고, 그 때 퍼센트 일치(percent agreement)는 100%가 된다[18].

표 2 초록 3에서 subject들에 의해 선택된 중요 명사

Num	Words	Judges				
		1	2	3	4	5
1	ambiguity	1	1	1	1	1
2	agent	1	1	1	1	1
3	system	1	0	1	1	1
4	pragmatics	0	1	0	0	0
5	TELL	0	1	0	1	1
6	Kripke	0	0	1	1	1
7	structure	0	0	1	1	1
8	paper	0	0	0	0	0
9	pragmatics	0	0	0	0	0
...	...	0	0	0	0	0
24	sentences	0	0	0	0	0
25	disjunctions	0	0	0	0	0

우리는 Gale에 의해 정의된 척도인 퍼센트 일치를 사용하여 색인어에 대한 피실험자들 사이의 일치 정도를 측정하였다[18]. *퍼센트 일치(percent agreement)*는 관찰된 일치에 대한 가능한 일치의 비율로 나타내어진다. 표 3은 선택된 색인어의 개수에 따른 피실험자들 사

이의 퍼센트 일치를 보이고 있다. 만약 각 문서에 대해 2개의 색인어를 추출한다면 피실험자들 사이의 평균적인 일치정도는 93.6%이다.

표 3 피실험자들 사이의 퍼센트 일치

	Percent agreement
Manually extracted keyword = 2	93.6%
Manually extracted keyword = 3	87.8%
Manually extracted keyword = 4	80%

피실험자들은 각 초록으로부터 2 , 3, 4개의 색인어를 각각 추출하였고, 제안된 시스템은 2개에서 6개까지의 색인어를 각각 추출하였다. 시스템이 필요로 하는 색인어의 개수에 따라 강한 체인의 수식에서 상수 값은 약간씩 변화되었다. 평가 척도로는 다음의 수식을 이용하였다.

$$Recall = \frac{a}{a+c}$$

$$Precision = \frac{a}{a+b}$$

$$Error = \frac{b+c}{a+b+c+d}$$

System	Judges	
	Keyword	Not keyword
Keyword	a	b
Not keyword	c	d

본 실험은 재현율과 정확도가 같은 중요도를 가진다고 가정하고, 하나의 값인 F-value로 측정한다. 재현율과 정확도가 높을수록 F-vlaue의 값이 높아진다.

$$F-value = \frac{2 \times P \times R}{P+R}$$

P: precision, R: recall

다음의 표들은 제안된 시스템에 의해 선택된 색인어의 수와 subject가 추출한 색인어의 수에 따른 결과를 보여준다. 예를 들면, 표 4에서 초록으로부터 subject가 추출한 색인어가 2개이고 제안된 시스템이 색인어로 4개의 명사를 추출하였을 때, 재현율은 0.59이고 정확도는 0.30이다.

표 4 피실험자들이 추출한 색인어가 2개일 때

Keywords \ Result	Recall	Precision	F-value
Extracted keyword = 2	.27	.27	.27
Extracted keyword = 3	.46	.30	.36
Extracted keyword = 4	.59	.30	.40
Extracted keyword = 5	.77	.20	.31
Extracted keyword = 6	.82	.26	.39

표 5 피실험자들이 추출한 색인어가 3개일 때

Keywords \ Result	Recall	Precision	F-value
Extracted keyword = 2	.24	.36	.28
Extracted keyword = 3	.39	.39	.39
Extracted keyword = 4	.55	.48	.51
Extracted keyword = 5	.67	.40	.50
Extracted keyword = 6	.76	.38	.50

표 6 피실험자들이 추출한 색인어가 4개일 때

Keywords \ Result	Recall	Precision	F-value
Extracted keyword = 2	.18	.36	.24
Extracted keyword = 3	.30	.39	.34
Extracted keyword = 4	.43	.43	.43
Extracted keyword = 5	.52	.42	.46
Extracted keyword = 6	.61	.41	.49

그림 11은 F-value와 시스템에 의해 추출된 색인어 수와의 관계를 나타낸다. 그림 11에서, x축은 제안된 시스템에 의해 추출된 색인어 수이고 y축은 F-vlaue이다. 피실험자가 추출한 색인어 수가 2개일 때 전체적인 시스템 성능은 상대적으로 불안정하다. 본 실험에서, 피실험자들이 추출한 색인어 수가 3개이고 제안된 시스템에 의해 추출된 색인어 수가 4개일 때 가장 좋은 F-value를 얻었다.

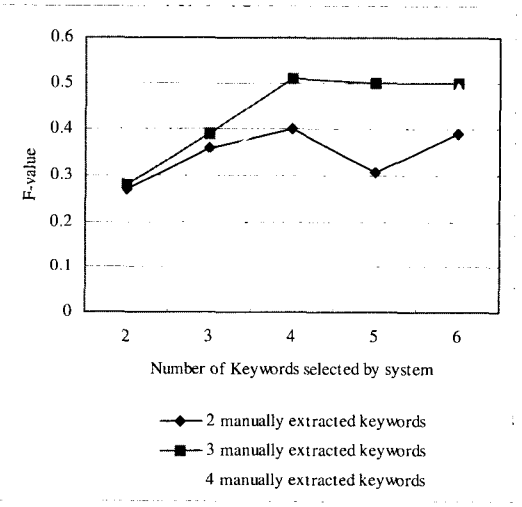


그림 11 색인어 개수에 따른 F-vlaue

제안된 시스템이 가장 좋은 결과를 주는 조건에서 시스템 정확도를 관찰하였다. 시스템 정확도는 다음 식과 같이 주어진다.

$$\text{Accuracy} = 1 - \text{Error}$$

아래 표는 시스템 정확도를 나타낸다. 이전에 보여진 색인어 추출에 대한 재현율과 정확도에 비해서, 상대적으로 높은 시스템 정확도를 얻을 수 있었다.

표 7 시스템 정확도

Text	A1	A2	A3	A4
Accuracy	.91	.81	.88	.82
Text	A5	A6	A7	A8
Accuracy	.87	.97	.77	.84
Text	A9	A10	A11	**Overall**
Accuracy	.88	.96	.79	.86

4.2.3 KEA와의 비교

KEA는 문서에서 자동으로 중요 어구를 추출하는 알고리즘으로 피실험자가 선택한 5개의 중요 어구에 대하여 KEA가 약 1에서 2개 정도를 맞춘다는 것을 알 수 있다[16][17]. 제안된 시스템은 그림 12와 같이 피실험자가 선택한 5개의 색인어에 대해서 2.5개를 맞추는 결과를 보였다. KEA와 비교하여보았을 때, 특정 도메인 훈련 셋(training set)등을 이용하지 않고도 정확한 색인어를 추출함을 알 수 있다.

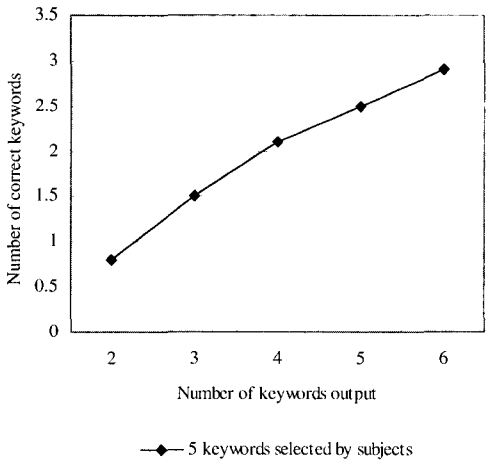


그림 12 피실험자들이 추출한 색인어가 5개일 때 시스템 성능

5. 결론 및 향후 연구

문서에서 색인어를 탐색하는데 있어서 문서의 의미적인 내용을 파악하여야 하는 중요함에도 불구하고, 단어나 문장들 사이의 의미적인 관계를 이용하는 시스템은 거의 없다. 본 논문은 문서의 내용을 이해하는 것을 색인어 탐색에서 중요한 요소라고 간주하여 어휘 체인을 사용하는 효과적인 색인어 추출 시스템을 제안하였다. 또한, 어휘 체인을 생성하기 전에 *semantic window*라는 개념을 이용하여 명사들의 의미를 대략적으로(roughly) 결정하는 방법을 제안하였다. 의미 결정 결과는 윈도우 크기가 5일 때, 평균 69%의 정확도를 보였다. 그러나 최적의 윈도우 크기를 결정하는 방안에 대하여는 향후 연구가 필요하다. 색인어 추출 성능에 관한 실험을 위하여 10개 논문에서 추출한 초록 10건을 사용하였다. 피실험자가 추출한 색인어가 3개이고 시스템이 추출한 색인어가 4개 일 때, 가장 좋은 F-value인 0.51의 성능을 보였다. 그때, 피실험자들 사이의 퍼센트 일치는 87.8%이었고 시스템 정확도는 0.86이었다. 중요어구 추출 시스템인 KEA와 비교하여보았을 때, KEA가 5개 색인어에서 평균 1.2개를 추출한 것에 비하여, 제안된 시스템은 약 2.5개를 맞추어 주목할 만한 결과를 제시하였다. KEA가 중요 어구를 추출하기 위하여 해당 도메인을 위한 훈련 셋을 사용하는 기계학습이 필요하였던 것에 비해, 제안된 시스템은 학습과 같은 전처리 과정을 거치지 않고도 주목할 만한 결과를 주었다는 점에서 이점이 있다.

그러나, 제안된 시스템에는 몇 가지 한계점이 있다. 본 시스템은 문서에 존재하는 명사들만 고려하여 색인어를 하기 때문에, 문서에 존재하지는 않지만 문서의 내용을 더 잘 표현하는 색인어를 추출할 수 없다는 것이다. 이 부분에 관하여서는 이미 추출된 색인어와 그것의 센스 정보를 이용하여 문서에 존재하지 않지만 문서의 내용을 표현해내는데 적합한 단어를 탐색하는 방법을 고려하고 있다. 예를 들면, 추출된 색인어가 *run*, *smash*, *hits*, *base*라면, 이 색인어들의 의미 정보를 이용하여 *baseball*과 같은 일반화된 색인어도 추출 가능해야 할 것이다. 또한 제안된 시스템은 워드넷을 사용하여 일반 도메인에서 동작하도록 설계되었다. 그러나, 실험에 사용된 문서들은 컴퓨터 과학 분야의 전문용어나 약어를 많이 포함하고 있었다. 이러한 요소들이 정확한 어휘체인을 구성하는데 방해로 작용하였다. 예를 들면, *information retrieval*이 *query*와 연관성을 가지는 것을 워드넷을 사용하여 찾아낼 수 없다. 워드넷은 일반 목적

으로 고안된 시소러스 이므로 특정 전문 분야들의 용어까지 수록하고 있지는 않기 때문이다. 마지막으로, 색인어들은 복합 명사이거나 중요 어구로 이루어져 있을 때 더 정보적이기도 하다. 예를 들면, *information integration*은 *information*, *integration*의 각각 떨어진 색인어로 취급되기보다 복합명사로 구성되어 있는 것이 더욱 정보적이다. 그러나 제안된 시스템은 복합명사를 아직 지원하지 못하고 있다. 따라서 본 논문에서 제시한 이러한 문제들을 해결함으로써 문서의 내용을 더욱 잘 표현해낼 수 있는 안정된 색인어 추출 시스템을 개발해 나갈 것이다

참 고 문 헌

[1] Lancaster, F.W., and Warner, A.J., *Information Retrieval Today*, Arlington, VA: Information Resources Press, 1993.

[2] Moens, M.-F., *Automatic Indexing and Abstracting of Document Texts*, Kluwer Academic Publishers, 2000.

[3] Hahn, U., "Making unerstanders out of parsers: semantically driven parsing as a key concept for realistic text understanding applications," International Journal of Intelligent Systems, Vol. 4, pp. 345-393, 1989.

[4] Lewis, D.D., and Sparck Jones, K., "Natural language processing for information retrieval," Communications of the ACM, Vol. 39, No.1, 92-101, 1996.

[5] Morris, J., and Hirst, G., "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," Computational Linguistics, Vol.17, No.1, pp. 21-43, 1991.

[6] Morris, J., "Lexical cohesion, the thesaurus, and the structure of text," Master's thesis, Department of Computer Science, University of Toronto, 1988.

[7] Barzilay, R. and Elhadad, M., "Using lexical chains for text summarization," In the Proceedings of the ACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.

[8] Luhn, H.P., "Statistical approach to mechanized encoding and searching of literary information," IBM Journal of Research and Development, Vol.1, No.4, pp.309-317, 1957.

[9] Bookstein, A., Klein, S.T., and Raita, T., "Clumping properties of content-bearing words," JASIS, Vol.49, No.2, pp. 102-114, 1998.

[10] Liddy, E.D., and Myaeng, S.H., "DR-LINK's: linguistic-comceptual approach to document and detection," The First Text REtreival Conference (TREC-1), pp. 113-129, 1993.

[11] Burnett, M., Fisher, C., and Jones, K., "InTEXT processing indexing in TREC-4," The Fourth Text REtrieval Conference (TREC-4), pp. 287-294, 1996.

[12] Salton, G., Singhal, A., Mitra, M. and Buckley, C., "Automatic text structuring and summarization," IP&M, Vol.33, No.2, 193-207, 1997.

[13] Halliday, M.A.K., and Hasan, R., *Cohesion in English*, London: Longman, 1976.

[14] Hasan, R., *Coherence and Cohesive Harmony*. In J. Flood (Ed.) Understanding Reading Comprehension, pp. 181-219, Newark, DE: IRA, 1984.

[15] Al-Halimi, R. and Kazman, R., *Temporal Indexing through Lexical Chaining*. In fellbaum, C., ed., wordNet: An Electronic Lexical Database and Some of its Applications, Cambridge, MA: The MIT Press, 1998.

[16] Frank, E., Paynter, G., Witten, I., Gutwin, C. and Nevill-Manning, C., "Domain-specific keyphrase extraction," In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan-Kaufmann, 668-673, 1999.

[17] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C. and Nevill-Manning, C.G., "KEA: Practical Automatic Keyphrase Extraction," In Proceedings of Digital Libraries (99: The fourth ACM Conference on Digital Libraries, pp. 254-255, 1999.

[18] Gale, W., Church, K., and Yarwsky, D., "Estimating upper and lower bounds on the performance of word-sense disambiguation programs," In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics(ACL-92), pp. 249-256, 1992

강 보 영

1997년 2월 경북대학교 컴퓨터공학과 졸업. 1999년 2월 경북대학교 영어영문학과(문학석사 : 의미론 전공). 2002년 2월 경북대학교 컴퓨터공학과(공학석사). 2002년 3월 ~ 현재 경북대학교 컴퓨터공학과 박사과정. 관심분야는 정보 검색, 문서 요약, 지식 베이스 구축 등

이 상 조

1974년 경북대학교 수학교육과(이학사). 1976년 한국 과학기술원(이학석사). 1994년 서울대학교 컴퓨터공학과(공학박사). 관심분야는 자연어 처리, 기계번역, 운영체제, 프로그래밍 언어, 데이터베이스