자동요약의 주제어 추출을 위한 의미사전의 동적 확장

Dynamic Expansion of Semantic Dictionary for Topic Extraction in Automatic Summarization

추 교 남*, 우 요 섭*★
Kyo-Nam Choo*, Yo-Seob Woo*★

Abstract

This paper suggests the expansion methods of semantic dictionary, taking Korean semantic features account. These methods will be used to extract a practical topic word in the automatic summarization. The first is the method which is constructed the synonym dictionary for improving the performance of semantic-marker analysis. The second is the method which is extracted the probabilistic information from the subcategorization dictionary for resolving the syntactic and semantic ambiguity. The third is the method which is predicted the subcategorization patterns of the unregistered predicate, for the resolution of an affix-derived predicate.

요 약

본 논문에서는 자동문서요약 시스템에서 정확하고 실용적인 주제어 추출을 위하여 한국어의 의미론적 특성을 고려한 의미사전의 확장 방법론에 대하여 논하고자 한다. 첫째로 동의어 사전을 통하여 의미표지 분석의 정확도를 높이고자 한다. 둘째로 하위범주화사전에 가중치를 부여하여 구문과 의미 분석에서 가장 올바른 분석 결과를 결정하는 참조 정보로 활용하고자 한다. 셋째로 미등록 용언의 하위범주화패턴 예측을 통하여 한국어에서 접사 파생되는 용언에 대하여 원활한 의미 분석을 수행할 수 있도록 한다.

Key words : Automatic Summarization, Topic Extraction, Semantic Processing, Thesaurus, Subcategorization

Ⅰ. 서론

기존의 표층적인 분석 방법과 차별화되는 효율적인 자동문서요약을 위해서는 원문의 어휘 개념, 의미역 등 한국어의 의미론적 특성을 고려하여 주제어를 추출하는 과정이 필요하다. 의미 분석 기반의 주제어 추출을 위해서는 한국어 문장이 내포하고 있는 의미적 결속성을 파악해야 하며 본 논문에 앞서 선행 연구된 문맥 분석 기반의 자동문서요약에서는 한국어 특징을 고려한 명사 시소러스와 하위범주화사전을 구축하고 활용하였다[1].

본 논문의 선행 연구인 문맥 분석 기반의 자동문서요약에서 수행한 주제어 추출 방법에는 다음과 같은 개선점이 나타났다. 첫 번째는 하위범주화사전과 명사 시소러스의 의미코드 정합 실패에 원인이 있었다. 의미 정합에 실패한 내용을 분석한 결과, 의미 정합이 실패한 요인은 하위범주화사전에 해당 보어 성분의 의미코드가 부여되어 있지 않아 상하위 관계를 분석할 수 없었기 때문이었다. 의미코드는 명사 시소러스의 계층 정보를 나타내는 문자열로 접두어 정합이 가능하도록 설계되었다. 접두어 정합이 실패하면 상하위 관계가 없는 것으로 판별된다[2].

또 다른 실패 요인으로는 실험 말뭉치의 보어 성분이 하위범주화패턴과 의미코드의 상하위 관계를 이루어야 하나 명사 시소러스의 다른 계층에 위치하고 있

* 인천대학교 정보통신공학과
(Department of Information and Telecommunication Engineering)
** 인천대학교 정보통신공학과
(Department of Information and Telecommunication Engineering)
※ 이 논문은 인천대학교 2007년도 자체연구비 지원에 의하여 연구되었음.
接受日:2009年 5月 06日, 修正完了日: 2009年 6月 26日

기 때문인 것으로 나타났다. 명사 시소러스는 12만 단어의 방대한 어휘가 트리구조의 형태로 구성되어 있어 수작업을 통한 어휘의 추가, 변경, 제거가 용이하지 않고, 잘못 편집이 될 경우, 하위범주화사전과의 일관성에 큰 손실이 발생할 수 있다[3].

이를 해결하기 위해서는 대량의 의미표지가 부여된 말뭉치(Sense Tagged Corpus)를 구성하여 하위범주화사전에 학습시키는 방법이 있으나 현실적으로 이러한 말뭉치를 구축하기가 어렵기 때문에 본 논문에서는 별도의 동의어(Synonym) 사전을 이용하여 의미 정합도를 높이는 방법론을 제안하고자 한다.

두 번째는 하위범주화사전에서 용언의 미등록에 원인이 있었다. 용언의 정합이 실패한 경우를 분석한 결과, 하위범주화패턴을 바르게 구성이 되어 있으나 사전에 용언이 미등록되어 있어 정합에 성공하지 못하였다. 특히, 미등록된 용언 중에는 본용언보다 접사가 붙어 파생된 용언이 많이 발생하였다. 이러한 접사 파생 용언의 특징은 본용언의 하위범주화패턴과 일치되는 경우가 대부분이었는데, 이는 곧 접사 파생 용언은 본용언과 의미 구조가 같고 동일한 상황 정보를 전달한다는 뜻이 된다.

대량의 말뭉치를 대상으로 분석을 수행한다면 용언 정합도가 낮게 나타날 수밖에 없다. 다양한 실험 대상에 대하여 하위범주화사전의 성능을 높이거나 일정한 수준을 유지하려면 하위범주화사전의 자동 확장 알고리즘의 개발이 필요하다. 즉, 미등록 본용언과 파생 용언에 의존하는 문형 패턴을 하위범주화사전에 등록된 하위범주화패턴과의 비교 분석을 수행하여 사전의 내용을 자동으로 확장할 수 있다면, 다양한 적용대상에 대하여 적응성을 갖춘 강건한 하위범주화사전의 구축이 가능하게 된다. 이에 본 논문에서는 하위범주화패턴을 동적으로 확장하는 방법론을 제안하고자 한다.

II. 본론

1. 의미표지 정합의 정확도 향상을 위한 동의어 사전

하위범주화사전과 명사 시소러스가 구성하는 선택제약에 의한 의미표지 정합 시, 실패율을 최소화하기 위하여 동의어 사전(Synonym Dictionary)을 구축하여 활용하고자 한다. 하위범주화사전과 명사 시소러스를 이용한 의미 정합이 실패하였다면, 이는 하위범주화사전의 패턴에 부여된 의미표지와 명사 시소러스에 등록된 의미표지 간 상하위 관계가 성립되지 않았음을 뜻한다.

이 의미표지는 의미코드가 부여되어 있어 상하위 분석 시에 문자열의 접두어 정합을 수행도록 구성되어 있는데 이때 문자열 정합이 이루어지지 않아, 상하위 관계를 분석할 수 없게 된다. 이러한 이유로는 명사 시소러스의 의미표지가 하위범주화사전의 용언 패턴에 미등록되어 있거나, 일부는 등록 오류로 잘못된 의미표지가 부여되었을 수 있다. 하위범주화사전의 구축에서 신문기사의 말뭉치로부터 용언과 어휘를 분석하여 반영하였고, 다른 어휘를 포괄할 수 있도록 범용적 패턴을 구성하였지만, 실제 어휘의 활용성을 모두 반영할 수는 없다. 또한, 명사 시소러스에 미등록된 어휘이거나 다른 어휘 계층에 존재할 경우에도 이러한 문제를 야기시킬 수 있다.

이를 해결하기 위해서는 미등록 어휘에 관한 의미표지 등록이나 오류 정정 등을 수작업으로 수행해야 하는 것이 일반적이지만 대량의 명사 시소러스나 하위범주화사전에 의미표지를 부여하고 수정하는 작업은 두 사전의 연동에 따른 성능에 부정적인 영향을 미칠 수 있기 때문에 신중하게 작업이 이루어져야 한다.

자동적인 방법으로는 의미표지가 부여된 말뭉치를 이용하여 하위범주화패턴을 학습시킨 후 수집된 어휘의 의미표지를 분석하여 의미표지의 등위를 조절하거나 등록, 삭제를 할 수 있다. 그러나 대량의 하위범주화사전과 명사 시소러스를 학습시킬 수 있는 의미표지가 부여된 말뭉치를 구축하는 것은 현실적으로 어려운 작업이다[4].

이에 본 논문에서는 명사 시소러스의 보조 사전으로 동의어 사전을 구축하여 의미표지의 상하위 관계 분석의 효율성을 높이고자 한다. 이 동의어 사전은 명사 시소러스에 대하여 버퍼의 역할을 하는 사전으로 명사 시소러스에 비하여 복잡한 계층구조를 갖지 않는다. 또한, 동의어 사전은 포괄적인 어휘 범주를 정의하고 이와 동일한 의미 속성을 갖는 어휘를 모아 놓은 외연(外延)적 사전이다.

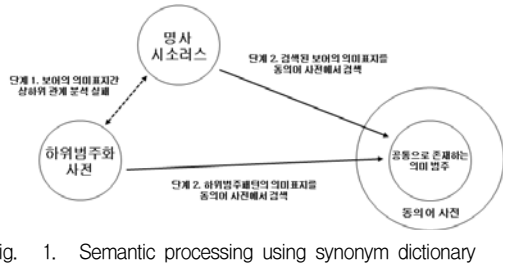

Fig. 1. Semantic processing using synonym dictionary
그림 1. 동의어 사전을 이용한 의미 처리

만약 하위범주화패턴과 명사 시소러스의 의미표지 간 상하위 관계 분석이 실패하였다면, 그림 1에 나타

낸 바와 같이 동의어 사전을 탐색하여 두 의미표지가 같은 의미 범주에 속하는지를 분석한다. 두 의미표지의 의미코드가 접두어 정합을 이루지 못하더라도, 동의어 사전에서 동일한 의미 범주에 속한다면 의미 정합이 이루어진 것으로 해석할 수 있다.

본 논문에서 구축한 동의어 사전은 43개의 포괄적인 의미 범주로 이루어져 있으며 이들 범주에 신문기사 말뭉치로부터 추출한 14,000여개의 어휘가 포함되어 있다. 43개의 포괄적인 범주는 기존의 동의어 사전에서 정의하고 있는 범주[5,6]와 명사 시소러스에서 정의된 최상위층의 정보, 그리고 신문기사에서 분류하고 있는 계층 정보를 활용하여 정의하였다. 표 1에 동의어 사전의 의미 범주를 나타내었다.

Table 1. Semantic domain of synonym dictionary
표 1. 동의어 사전의 의미 범주

동의어 사전의 의미 범주			
1.인간과 인간관계	12.주생활	23.운동	34.자연현상
2.가족과 친인척	13.밥과 금	24.나라 이름	35.동물
3.성과 결혼	14.언론과 출판	25.국가와 정치	36.식물
4.신체와 생리작용	15.정보와 통신	26.법과 질서	37.모양
5.병과 치료	16.교육	27.국방	38.빛과 색채
6.삶과 죽음	17.과학과 학문	28.사회와 사회활동	39.수와 수량
7.감각과 감각기관	18.종교와 믿음	29.경제와 경제활동	40.시간
8.생각과 감정	19.문명과 문화	30.직업과 직장	41.공간과 우주
9.성격과 태도	20.예술	31.산업	42.상태와 정도
10.의생활	21.취미	32.연료와 에너지	43.동작
11.식생활	22.놀이와 게임	33.도로와 교통	

동의어 사전은 의미표지가 부여된 말뭉치로부터 학습된 명사 시소러스 특징을 나타낸다. 명사 시소러스를 의미표지가 부여된 말뭉치로 학습을 시키면 활용도가 높은 층위로 어휘가 집중되며 이로 인하여 복잡한 계층 구조가 적용하고자 하는 범주에 맞도록 조절이 가능케 된다. 학습된 명사 시소러스의 이러한 범주 조절 기능을 외부 사전으로 옮겨 놓은 것이 동의어 사전이라 할 수 있다.

명사 시소러스의 계층을 조절하는 일은 복잡하고 어려운 문제이지만 동의어 사전은 간단한 범주 구조로 이루어져 있어 범주의 병합과 분리 등이 용이하다. 동의어 사전의 범주 구조가 현재는 43개이지만 경우에 따라서는 유사 범주로 병합하거나 분리하여 범주의 개수를 조절할 수 있다. 이러한 범주의 조절은 의미 분석의 정확도를 조절할 수 있는 중요한 요소가 된다. 또한, 명사 어휘의 추가도 명사 시소러스가 아닌 동의어 사전에서 용이하게 수행할 수 있고, 이때 활용 빈도가 높은 명사 어휘는 일괄적으로 명사 시소러스로 등록할 수 있다.

2. 하위범주화패턴에 대한 가중치 부여

본 절에서는 하위범주화사전의 다양한 구성 정보에 대하여 통계치 분석을 통한 가중치를 부여하여, 의미분석에 있어 최적의 해를 결정하는데 활용하고자 한다. 의미 분석을 수행하는 많은 선행 연구에서는 언어 지식원으로 사용되는 사전을 배재하고, 통계적인 자동언어처리 방법론을 이용하여 용언의 패턴과 어휘의 의미를 분석하였다. 그러나 이러한 방법론은 응용분야에 따라, 형태소, 구문, 의미표지가 부여된 말뭉치를 통하여 통계 정보를 추출할 수 있는데, 주목할 만한 통계치를 얻기 위해서는 대용량의 훈련된 말뭉치가 필요하고, 이를 구축한다는 것은 현실적으로 많은 어려움이 따른다. 또한, 훈련된 말뭉치의 어휘적 균형성(Lexical Balance)을 보장받을 수 없다면 추출된 통계치는 더욱 신뢰성을 잃게 된다.

본 논문에서 활용하는 하위범주화사전은 반자동적인 방법으로 구축되었고 연구자의 튜닝(Tuning)을 통하여 신뢰할 수 있는 많은 양을 확보하였으므로 이 사전으로부터 조사와 의미표지, 의미역의 관계를 통계적으로 분석하여, 각 하위범주패턴에 대한 가중치 부여에 활용한다면 규칙기반의 방법론만을 적용할 경우보다 더욱 효율성 있는 정확도를 얻을 수 있으리라 판단된다.

Table 2. The probabilistic information of semantic marker
표 2. 의미표지의 통계 정보

Percentage(%)

의미표지	이/가	을/를	와/과	에/에서
사람	60.54	3.71	10.1	3.65
장소	3.2	10.3	1.4	50.1
물건	7.98	40.4	15.7	10.92
구체물	10.4	30.6	15.9	10.3
추상물	5.14	60.1	6.0	7.3
인간이외의 동물	25.1	30.7	20.1	10.3

우선 각 조사에 의존하여 나타나는 의미표지의 통계치를 분석하였다. 표 2에 나타낸 데이터는 하위범주화사전에 등록되지 않은 용언이고 해당 보어 성분이 명사 시소러스에 등록되어 있는 경우, 하위범주화사전에서 의미표지가 특정 조사에 의존되는 정보를 기준으로 해당 보어에 적절한 의미표지를 부여할 수 있는 기준으로 활용할 수 있다.

표 2의 통계치를 살펴보면 의미표지 '사람'은 주격조사 '이/가'에 절대적으로 의존하는 것으로 나타났으며 기타 다른 조사에서는 낮은 분포를 보였다. 의미표지 '물건', '구체물'의 의미표지에 대해서는 주목할

필요가 있다. 의미표지 '물건'은 '구체물'의 하위에 속하는 개념으로 구체물의 개념 속성을 상속받고 있다. 그러나 각 조사에 의해 의존되는 의미표지의 빈도는 다르게 나타남을 알 수 있었다.

하위범주화사전의 통계치를 기준으로 하여 의미표지를 결정하고자 할 때, 과연 어떤 의미표지를 부여하는 것이 적절한지에 대한 기준이 필요할 수 있다. 하위범주화사전의 설계 원칙 중 의미표지를 부여하는 기준을 참고하면, 명사 시소러스의 중간 층위면서 구체적인 의미를 나타낼 수 있는 의미표지를 부여하도록 되어 있다. 이 기준에 따라 보어 성분의 의미코드에서 가장 가까운 접두 의미코드를 가지고 있는 의미표지를 부여해야한다.

이제, 의미역의 통계치를 살펴보자. 의미역은 보어 성분이 문장 내에서 어떠한 역할을 수행하는가를 나타내는 것으로 어떤 용언이 지배하는지에 따라 같은 의미표지라도 의미역은 달라질 수 있다.

Table 3. The probabilistic information of semantic role
표 3. 의미역의 통계 정보

Percentage(%)

의미역	사람	구체물	물건	추상물	장소
AGT	46.1	10.5	5.3	20.7	3.2
CHD	10.4	15.7	0.4	1.2	0.4
EXS	5.3	0.5	-	-	-
ACC	10.1	30.2	30.4	10.1	20.1
LOC	0.7	10.7	10.1	6.7	50.34
COM	10	5.2	10.3	-	-

표 3에 나타낸 의미역의 통계치를 살펴보면 AGT나 CHD는 같은 의미표지 '사람'과 연관되어 나타날 때 AGT가 많이 부여되어 있고 CHD도 비율이 비교적 낮지 않음을 알 수 있다. 규칙적 방법으로 판별되지 못한 보어 성분을 의미표지 '사람'으로 결정할 경우, AGT를 제외한 나머지 의미역 정보들은 비교적 상호 연관성이 적고 조사에 의존되어 부여해도 무관할 정도의 통계치를 나타내고 있다.

그러나 더욱 정확한 의미역을 부여하기 위해서는 각 의미역이 구성하는 하위범주화패턴에 대한 정보를 참조할 필요성이 있다. 예를 들어, 주격조사에 의존하는 AGT는 대부분이 ACC, LOC와 같은 의미역을 동반하고 나타남을 하위범주화사전에서 분석할 수 있다. 이와 같은 정보를 체계적으로 정리하기 위하여 하위범주화사전에서 나타나는 하위범주패턴의 통계치를 분석하였다.

표 4에서 하위범주화패턴의 통계치를 살펴보면, 대부분의 문장은 AGT와 ACC나 LOC을 수반하고 CHD는 사전 내에서는 낮은 비율로 구성되어 있으나 필수적으로 COM, GOL, LOC 등이 동반되어 나타나게 된다. 즉, CHD로 분석되는 보어 성분은 다음에 RCP나 GOL이 ACC나 PTH등의 의미역이 동반될 확률보다 크다. 지금까지 논하였던 하위범주화사전의 통계치는 인간의 언어 활동을 심층적인 면에서 살펴볼 때, 선호도가 작용하고 있음을 알 수 있다. 이 가중치 정보는 의미 분석에서 여러 분석 후보 등이 발생할 경우나 미등록 용언의 하위범주화패턴을 예측하는 경우에 활용될 수 있다.

Table 4. The probabilistic information of subcategorization patterns
표 4. 하위범주화 패턴의 통계 정보

Percentage(%)

	문형 패턴	패턴의 빈도
1	[이 AGT]	9.8
2	[이 AGT] [에서 LOC]	8.7
3	[이 EXS] [에/에서 LOC]	0.5
4	[이 AGT] [에(대하여) MGL]	0.45
5	[이 AGT] [(때문)에/(으)로 CSE]	0.8
6	[이 CHD] [에 GOL]	1.3
7	[이 CHD] [에/에 LOC]	2.5
8	[이 CHD] [(으)로 ELM]	1.1
9	[이 AGT] [에(게)서 SRC] [(으)로(하여) PTH] [에/(으)로 GOL]	10.34
10	[이 TRS] [이/(으)로 TRR]	0.1
11	[이 CHD] [이 CNT]	1.18
12	[이 AGT] [을 ACC]	30.1
13	[이 AGT] [을 ACC] [(으)로 INS]	10.1
14	[이 AGT] [을 ACC] [(으)로 MEN]	5.3
15	[이 AGT] [(기)를 CNT]	5.8
16	[이 AGT] [을 CNT] [에 MGL]	3.2
17	[이 AGT] [을 RNG]	8.6
18	[이 AGT] [을 SRC]	7.5
19	[이 AGT] [을 PSS]	3.4
20	[이 AGT] [을 COM]	2.1
21	[이 AGT] [와/과 COM]	9.7
22	[이 CHD] [을 COM]	2.7
23	[이 CHD] [와/과 COM]	1.8
24	[이 AGT] [을 ACC] [와/과 COM]	9.3

3. 하위범주화패턴에 대한 가중치 부여

본 절에서는 하위범주화사전에 미등록되어 있는 용언에 대하여 의미 정합을 수행하는 알고리즘을 제안하고자 한다. 하위범주화사전을 문장의 구문과 의미 분석에 활용하기 위해서는 제일 먼저 사전에 해당 용언이 존재하는지를 판별해야 한다. 즉, 사전에 등록되

어 있지 않은 용언에 대해서는 분석이 불가능하다. 용언이 사전에 미등록되어 있을 경우에서도 원활한 의미처리가 가능하도록 하기 위해서는 용언에 대한 하위범주화패턴의 예측은 반드시 필요하다.

하위범주화패턴의 예측을 통하여 효과를 얻고자 하는 미등록 용언의 주요한 대상은 바로 접사('-하-')가 붙어 있는 파생 용언이다. 접사 '-하-'는 명사와 결합하여 용언의 역할을 수행하는 것으로 결합될 수 있는 명사가 한정되어 있는 것이 특징이다. 접사 '-하-'가 붙어 있는 파생 용언은 의미 구조를 공유하는 본용언이 존재할 가능성이 많기 때문에 하위범주패턴을 예측하여 해당 본용언과 의미 구조를 공유하는 방법으로 어휘의 의미적 중의성 해소와 의미역 분석을 수행하고자 한다. 접사 '-하-'가 붙어 있는 파생 용언뿐만 아니라 본용언도 하위범주화사전에 미등록 되어 있다면 이와 같은 방법으로 의미 처리를 수행할 수 있다.

Table 5. Syntactic dependency rules
표 5. 구문적 의존 규칙

구 분	설 명	
N1+N2	N2 (N1의 지배소)	① 관용어구 분석
		② 복합어 분석
		③ 소유격 분석
	N1 (N1의 지배소)	후위에서 전자로의 의존관계만 고려
NU1+N1	NU1과 N1의 연어 또는 수식관계 분석	
DT+N1	DT와 N1의 연어 또는 수식관계 분석	
AD+DT	AD와 DT의 활용 또는 수식관계 분석	
AD+V	AD와 V의 활용 또는 수식관계 분석	
V1+V2	V1가 중심어	① 관용어구 분석
		② V2가 보조사(VX)의 역할 분석
	V2가 중심어	① 관용어구 분석
		② V1이 절로서 V2에 의존되는지를 분석

본 논문에서는 접사 '-되-'가 붙어 있는 파생 용언에 대해서는 고려하지 않고 있는데, 접사 '-되-'는 모든 명사에 붙어 용언으로 파생될 수 있고, 하위범주화패턴과는 다른 문체적 특성을 지니기 때문에 분석이 어렵다. 일반적인 의미 분석에서는 분석하고자 하는 용언이 하위범주화사전에 등록되어 있어야 의미 우선 제약을 통하여 그 성능을 기대할 수 있다. 미등록 용언에 대해서는 의미 우선 제약을 수행할 수 없기 때문에, 반드시 구문적 의존 관계를 규명해야만 한다. 본 논문에서는 하위범주화패턴과 최적의 구문적 의존 관계를 이루는 경험론적(Heuristic) 규칙을 표 5와 같이 정의하였다[7].

미등록 용언의 구문적 의존 관계를 통하여 하위범주화패턴을 구성하면 하위범주화사전에서 유사한 패턴을 탐색한다. 미등록 용언의 하위범주화패턴과 유사한 용언의 하위범주화패턴이 나타나면 이 미등록 용언은 하위범주화사전에 등록된 용언과 동일한 하위범주화 구조를 지닌 것으로 보고 의미표지와 의미역을 부여한다. 다음 예 (1)을 살펴보자.

선생님이 학교로 가다. (1)
선생님이 학교로 출근하다.

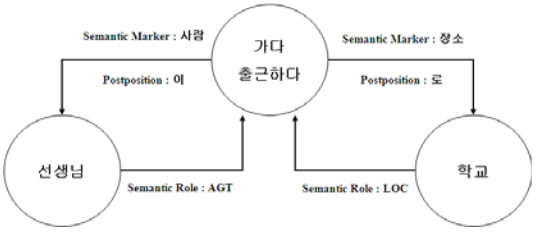

Fig. 2. The identical subcategorization structure of predicate '가다', '출근하다'
그림 2. 용언 '가다', '출근하다'의 동일한 하위범주 구조

그림 2에서 나타낸 바와 같이 '가다'와 '출근하다' 모두 같은 하위범주 구조로 이루어져 있다. 화용론적인 면에서 문맥은 다를 수는 있지만, 문장의 의미구조는 일치함을 보여주고 있다. 이는 '출근하다'가 '가다'의 의미적 속성을 나타낸다고 정의할 수 있다[8]. 만약, 미등록 용언의 하위범주화패턴이 다수 발견되었다면 2장 2절의 가중치 정보를 활용하여 최적의 하위범주화패턴을 예측할 수 있다.

4. 실험 결과

의미사전을 확장하는 것은 두 가지로 나누어 진행하였다. 첫 번째 말뭉치와의 정합 실험에서 획득한 정보를 하위범주화사전에 추가하는 것이고 두 번째는 빈도가 높은 용언들을 사전에 추가하였으나 모든 한국어 용언을 포함하고 있는 것이 아니기 때문에 하위범주화사전에 빠져 있는 새로운 용언이나 패턴을 추가하는 작업이 이에 해당한다. 말뭉치의 용언 중 하위범주화사전에 존재하지 않은 경우, 용언이 존재는 하나 알맞은 패턴이 없는 경우, 패턴이 있는데 의미정합에는 실패한 경우 하위범주화사전에 추가시키는 과정을 수행하였다.

[실험 1] 하위범주화사전에 존재하지 않은 용언에 대한 확장

구 분	대상 용언의 수	중복을 제외한 용언의 수	추가된 패턴의 수
하위범주화사전에 존재하지 않는 용언	1492개	887개	123개

[실험 2] 하위범주화사전에 존재하는 용언의 확장

구 분	대상 용언 수	중복을 제외한 용언의 수	추가된 패턴의 수
패턴은 있으나 정합되는 명사의 개념이 모두 없어 실패했던 경우	13560개	3165개	845개
패턴은 있으나 모든 하위범주의 개념과 완벽히 일치하지 않는 경우	22037개	5165개	1213개

[실험 3] 실험 말뭉치에서 용언의 유효 범위

전체 말뭉치 용언	하위범주화사전과 정합된 용언	정합도
20,372개	15,521개	76%

[실험 4] 의미표지와 구문 정합이 이루어진 용언 정보

용언 15,521개 중

의미표지 정합이 이루어진 용언	14,023개	정합도 90%
구문 정합이 이루어진 용언	13,833개	정합도 89%

[실험 5] 보어 성분의 의미적 중의성 해소와 의미역 분석

가. 의미표지가 정합된 보어

	전체 보어	단일 의미 보어	다의성 보어
의미표지 정합이 이루어진 용언 14,032개	9,965개	4,875개	5,090개

나. 9,965개의 보어 중 의미표지의 정확도

실험 내용	정확히 분석된 보어 / 대상 영역 (개)	정확도
의미표지 정합	9,064 / 9,965	91%
구문 정합	8,917 / 9,965	89%
의미역 정합	9,021 / 9,965	90%

다. 5,090개의 보어 중 의미표지의 정확도

실험 내용	정확히 분석된 보어 / 대상 영역 (개)	정확도
의미표지 정합	4,512 / 5,090	88%
구문 정합	4,475 / 5,090	88%
의미역 정합	4,435 / 5,090	87%

본 논문에서는 의미사전의 확장을 통하여 향상된 의미 분석의 성능을 평가하였다. 의미사전의 확장전 성능 평가에 이용하였던 동일한 실험 말뭉치를 대상으로 의미사전의 확장 방법을 실험하였는데 먼저 용언의 정합도는 76%로 의미사전의 확장 방법을 사용하지 않았던 용언 정합도 52%보다 약 24% 정도 향상되었다. 이는 접사 '-하-'에 의해 파생되었고 하위범주화사전에 미등록된 용언에 대한 하위범주화패턴 예측을 통하여 얻은 결과이다.

용언 정합이 이루어진 용언 15,521개 중 의미표지 정합이 이루어진 용언 14,023개에 의존하는 보어 성분 9,965개에 대하여 의미표지와 의미역 정합의 정확도를 측정하였다. 보어 성분 9,965개에 대하여 의미표지의 정확도 91%, 구문의 정확도 89%, 의미역의 정확도 90%를 얻을 수 있었다. 이 결과는 의미사전의 확장 방법을 적용하지 않은 의미 분석 결과와 비교하여 의미표지 4%, 구문 3%, 의미역 3%의 성능이 향상되었음을 보여주고 있다. 특히, 의미사전의 확장 방법을 통한 다의성 보어 성분의 의미 분석의 성능을 살펴보면, 의미표지 88%, 구문 88%, 의미역 87%로 의미사전의 확장 방법을 적용하지 않은 다의성 보어의 의미 분석 성능과 비교하여 의미표지 6%, 구문 11%, 의미역 79%가 향상되었다.

분석 실패의 원인을 살펴본 결과, 대부분의 경우 보어 성분이 명사 시소러스에 등록되지 않아, 의미표지를 분석할 수 없었는데, 본 논문에서는 미등록 명사에 대해서는 별도의 해결 방안은 고려하지 않고 기본적으로 명사 시소러스에 등록하는 것을 원칙으로

하고 있다. 본 논문에서 활용한 명사 시소러스는 개념 분류가 세부적으로 나뉘어져 있다. 만약, 미등록 명사의 개념 예측을 수행하여 임의로 명사 시소러스에 등록시킬 경우, 그 신뢰도가 보장되지 못한다면 명사 시소러스의 개념 분류 체계를 지속적으로 관리하기 어렵기 때문에 본 논문에서는 분석에 실패한 미등록 명사를 수집하여 그 중요도를 판별하여 일괄적으로 명사 시소러스에 등록하고자 한다.

III 결론

본 논문에서는 자동문서요약 시스템에서 정확하고 실용적인 주제어 추출을 위하여 한국어의 의미론적 특성을 고려한 의미사전의 확장 방법론에 대하여 다음과 같이 논하였다. 첫째로 동의어 사전을 통하여 의미표지 분석의 실패율을 줄였다. 둘째로 하위범주화사전에 가중치를 부여하여 의미 분석에서 가장 올바른 분석 결과를 결정하는 참조 정보로 활용하였다. 셋째로 미등록 용언의 하위범주화패턴 예측을 통하여 한국어에서 접사 '-하-'로 파생되는 용언에 대하여 원활한 의미 분석을 수행할 수 있도록 하였다.

앞으로 의미 분석을 통하여 더 높은 성능의 자동문서요약의 주제어를 추출하기 위해서는 다음과 같은 연구가 보완되어야 한다. 첫째로 하위범주화사전의 의미역 체계를 확장하고 이를 계층화하여 문장의 각 어휘성분에 대한 의미역을 세부적으로 분석해야만 한다. 그리고 하위범주화사전에 용언의 다양한 용례를 추가하여 문장 분석 및 생성에 응용할 수 있도록 구성하여야 한다. 각 용언의 하위범주의 의미표지 성분에도 다양한 제약 조건을 부착하여 정밀한 의미 분석에 활용해야 한다.

둘째로 명사 시소러스에는 현재 범용적인 명사 성분이 거의 포함되어 있다. 그러나 오히려 방대한 어휘 체계가 정확한 문맥 분석에 비효율적으로 작용할 수 있다. 그리고 적용 범주에 필수적 어휘 성분이 미등록어인 경우, 반드시 등록시켜야 문맥 분석 효과를 기대할 수 있다. 이에 명사 시소러스의 각 의미표지에 연관된 하위범주패턴과 용례 정보를 지속적으로 추가하고 학습시켜, 방대한 의미표지 체계를 정리할 수 있는 연구가 필요하다.

참고문헌

[1] 추교남, 우요섭, "문맥과 공통 주제의 의미 분석을 통한 다중 문서의 자동 요약." 한국정보기술학회, 제5권-2호, pp.89-103, 2007

[2] 우요섭, 양승현, 김영섬 등, "시소러스와 용언 패턴을 이용한 의미역 부착 한국어 하위범주화 사전의 구축." 한국정보과학회, 제6권-3호, pp.364-372, 2000

[3] 박현재, 우요섭, "의미 정보를 이용한 이단계 단문 분할." 한국정보처리학회, 제7권-9호. pp.2876-2884, 2000

[4] K. N. Choo., Y. S. Woo. and S. H. Kang, "Automatic Extension of Korean Predicate-based Subcategorization Dictionary from Sense Tagged Corpora." Springer, Lecture Notes in Computer Science 3045, pp.585-592, 2004

[5] 신문기사 종합 시소러스, 한국언론연구원, 2000.

[6] 신현숙 등, 현대 한국어 학습사전, 한국문화사, 2000

[7] K. N. Choo., Y. S. Woo. and H. K. Min, "Icon Language-based Auxiliary Communication System Interface for Language Disorders." Springer, Lecture Notes in Computer Science 3665, pp.93-101, 2005

[8] 추교남, 멀티미디어 XML 문서에 대한 의미 분석 기반의 지능적 자동 요약, 인천대학교 박사학위 논문, 2007

저 자 소 개

추 교 남 (정회원)

1997년: 인천대학교 정보통신공학과 (공학사)
1999년: 인천대학교 대학원 정보통신공학과 (공학석사)
2007년: 인천대학교 대학원 정보통신공학과 (공학박사)
2009년 ~ 현재: 인천대학교 정보기술교육원 초빙교수
<주관심분야> 한국어정보처리, 시멘틱 웹, 인공지능

우 요 섭 (정회원)

1986년: 한양대학교 전자통신공학과 (공학사)
1988년: 한양대학교 대학원 전자통신공학과 (공학석사)
1992년: 한양대학교 대학원 전자통신공학과 (공학박사)
1992년 ~ 현재 인천대학교 정보통신공학과 교수
<주관심분야> 한국어정보처리, 시멘틱 웹, 인공지능