論文
5-1-8

한국어에서 실용적 언어분석 단위의 인식과 평가

(Recognition and Evaluation of Efficient Language Analysis Unit for Korean)

박인철(In-Chol Park)[1)]

요 약

본 논문에서는 인터넷에서 대부분 정보의 표현 형태인 언어의 자동화된 분석을 위한 접근 방법으로 언어학적 접근과 전산학적 처리의 관점을 살펴본다. 또한, 대용량의 자료에 대한 실용적인 정보 색인과 검색, 정보추출, 기계번역 등을 위해 형태소분석, 구문분석 및 의미분석의 각 단계에서 실용적인 분석의 단위를 살펴보고, 언어학에 기반한 형태론적 단위보다 구문적 최장 분석 단위를 제안한다. 그리고 대량의 문서에 대한 실험을 통해, 제안하는 언어분석의 단위가 언어처리 과정에서 발생하는 모호성을 축소하는 제약으로도 사용될 수 있음을 보인다.

ABSTRACT

In this paper, we observe the differences between linguistic and computational aspect in the automatic processing of languages which are dominant representation method for information in the Internet. For efficient information retrieval, information extraction and machine translation from the massive documents, we investigate analysis units for morphology analysis, syntactic analysis and semantic analysis. and propose the syntactic longest analysis unit rather than morphological unit based on linguistics. Also, by evaluating with massive documents, we show that the proposed analysis units can be used for the constraint which can reduce the ambiguity occurring in the language processing.

1) 정회원 : 호원대학교 컴퓨터학부 부교수

논문접수 : 2004. 1. 10.
심사완료 : 2004. 1. 17.

1. 서론

인터넷의 활성화로 정보의 유통이 활발해지면서 정보 및 지식의 표현 형태인 언어에 대한 관심이 증대되고 있다. 인터넷을 통한 대용량 정보의 검색, 가공, 번역을 위해서는 언어에대한 기계의 자동화된 해석이 반드시 필요하다. 전통적인 언어 분석 과정을 살펴보면 크게 3가지 단계로 나누어진다[19].

첫 번째로 형태소분석 단계는 여러 형태소들의 묶음이 표층 형태로 나타나는 하나의 어절로부터 의미를 갖는 최소 단위인 각 형태소를 분석해 내는 것으로 정의된다[7]. 예를 들어, "감기는"이라는 어절에 대한 여러 가지 형태소 분석 결과는 다음과 같다.

① 감기(동사) + 는(어미)
② 감기(명사) + 는(조사)
③ 감(동사) + 기(명사형전성어미) + 는(조사)

두 번째 단계에서는 형태소들 사이의 문법적 관계를 이용하여 문장의 구조를 분석하는 것이다. 즉, 형태소분석 결과와 문법 규칙 등을 바탕으로 문장의 구조를 자동으로 분석하는 것을 목적으로 한다. 세 번째는 의미분석 단계로 한 문장 또는 문서내의 문장들이 나타내고자 하는 의미를 파악하는 것이다.

이러한 각 단계의 전후에는 필요에 따라 형태소분석 전처리 및 품사 태깅, 구문분석 전처리 단계 등이 있을 수 있다. 또한, 각 단계에서는 언어의 복잡성과 모호성 때문에 분석 결과가 결정적이지 않은 경우가 많이 발생하며, 각 단계의 이러한 모호성은 다음 단계에 전파되어 최종 단계에서는 분석 시간이 기하급수적으로 증가하게 된다.

현재의 문서 처리 수준을 살펴보면, 대량의 문서에 대한 색인 및 검색을 목적으로 할 경우, 언어해석의 복잡도와 모호성, 분석할 자료의 양에 따른 문제 때문에 형태소분석 과정만을 통해 문서를 분석하고 있는 경우가 대부분이다[20]. 대부분의 시스템에서 사용하는 형태소분석 방법은 최소의 의미적 단계(일반적으로 사전에 표제어로 등록 가능한 단어)로 문서를 분해하고, 이를 이용하여 정보검색을 위한 색인어를 추출한다. 또한, 검색을 위한 문장의 분석에서도 간단한 형태소분석을 통해 명사나 동사와 같은 주요한 단어를 찾아내고 이를 키워드로 이용한다.

자동 번역과 같은 고급 문서 처리에서는 문서의 양이 비교적 적은 경우 품사 태깅과 구문분석, 의미분석과 같은 상세한 언어 분석을 통해 번역하고자 하는 대상 언어로 번역한다[19]. 이때 이용되는 언어 자원은 매우 크고 복잡하며, 정교한 분석을 원할 경우 기계적 처리 속도는 일반적으로 느린 편으로 알려져 있다. 이를 해결하기 위해 가능한 최소의 단위로 문서를 분해한 후 분석을 시작하는 언어학적 접근 방법에서 기계적 처리에 적합하고 분석 복잡도를 줄이도록 분석의 단위를 변경함으로써 분석의 효율을 얻기 위한 연구들이 각광을 받고 있다[8].

본 논문에서는 언어 분석의 각 단계에서, 언어처리의 복잡도를 극복하기 위해 언어학적 처리 단위를 넘어서는 분석을 시도하고 있는 여러 연구 방법들을 살펴보고, 구문분석 및 의미분석의 정확성과 효율성의 향상을 위해 형태소분석 단계나, 형태소 분석 후처리 과정에서 언어학에 기반한 형태론적 단위보다 구문적 최장 단위의 분석 방법을 제안한다. 이 방법은 형태소분석 단계에서 품사의 유한상태 오토마타를 통해 인식하거나, 형태소 분석 후처리 과정에서 기호의 구조적 특성 또는 어휘의 의미적 유사성에 기반한 오토마타를 통해 인식될 수 있어, 구문분석 등의 단계에서 처리되는 것보다 처리 속도면에서 매우 빠르면서, 언어 분석의 정확성을 높일 수 있다. 마지막으로 본 논문에서는 문서의 통계적 분석을 통해 이 분석 단위에 의한 언어처리가 효율적임을 검증해보도록 한다.

2. 관련 연구

이 장에서는 한국어 및 일본어에서 언어를 구성하는 최소의 단위를 기반으로 문장을 분석하는 방법에서 벗어나 가장 최대의 분석 구조를 가지는 최장일치의 연구 방법에 대해 기술한다.

2.1 형태론적 최장일치

한국어는 명사나 동사와 같은 실질 형태소(내용어)에 조사나 어미와 같은 형식 형태소(기능어)가 결합된 교착어(agglutinative language)의 특징을 가지고 있다. 실질 형태소는 한 어절에서 중심이 되는 형태소로 구체적인 대상이나 동작을 나타내는 말로 체언, 수식언, 감탄사, 용언 등이 이에 해당한다. 형식 형태소는 실질 형태소에 결합되어 말과 말 사이의 관계를 형식적으로 표시하는 형태소로 조사, 어미, 접사 등이 이에 포함된다.

언어를 분석하는 관점에서 보면, 분석의 최소 단위인 형태소로 문장을 분석하는 것이 가장 중요한 첫 번째 단계가 된다. 그러나 한국어는 교착어 특성으로 인해 한 어절 내에는 많은 형태소들이 결합할 수 있으며, 이로 인해 많은 모호성이 발생한다. 예를 들어, "철수까지로밖에는"의 형태소분석 결과는 "철수/명사"+"까지/조사"+"로/조사"+"밖에/조사"+"는/조사"와 같으며, 이는 형태소 사이의 연결 정보를 어느 정도 이용하여 불필요한 모호성을 제거한 것이다. 즉, "밖에"는 "밖/명사"+"에/조사"로 분석될 수 있지만, "밖"의 왼쪽에 "로"가 조사이고, "밖"의 오른쪽 형태소인 "에"가 조사인 경우, "밖"은 반드시 조사이어야 한다. 이렇게 최장일치로 분석을 할 경우, 분석 대상의 왼쪽과 오른쪽 형태소가 분석을 위한 제약정보가 될 수 있기 때문이다. 따라서, 기계적 처리의 관점에서 "까지"+"로"+"밖에"+"는"이 하나의 조사로 묶이어, "까지로밖에는"이 하나의 단위로 분석되어도 문장의 큰 맥락을 벗어나지 않는다. 이와 같이 최장의 형태소들을 하나의 단위로 묶어서 형태소분석의 효율성을 얻기 위한 연구는 분석의 복잡도를 줄여 줄 뿐만 아니라, 분석의 정확도를 높일 수도 있다[16].

2.2 통사/의미론적 최장일치

2.1절에서는 같은 기능어끼리 분리하지 않고 결합하여 얻을 수 있는 장점에 대해 설명하였다. 이외에도 형태소분석 단계에서 내용어와 기능어가 결합하여 분리될 수 없는 단위로 인식해야만 하는 어절들이 형태소분석 단계에서 많이 발생한다. 이러한 형태소열은 구문분석 및 의미분석의 관점에서 보면, 하나의 단위로 인식되면 분석의 효율성이 높아진다. 그러나 대부분의 형태소분석기와 구문분석기를 구현할 때 각 단계의 분석 목적에 충실한 시스템을 만들기 때문에, 하나의 단위로 인식되는 형태소열에 대한 고려가 없었다.

예를 들어, 관용구(idiom)는 문장 "kick the bucket"와 같이 각 낱개의 단어 의미의 결합과 다른 특별한 의미를 지니는 단어들의 연결로서, 대개 문자 그대로 다른 언어로 번역하면 그 특별한 의미를 잃는 것을 말한다[13]. 즉, "kick the bucket"의 의미는 "kick"과 "bucket"의 의미의 합이 아니라 전혀 새로운 의미인 "die"를 나타낸다. 관용구를 구성 요소 사이의 의미적 긴밀도와 어휘화된 정도에 따라 분류한 연구[15]에 따르면, 어휘화된 정도가 가장 높고 내적구성에 있어서도 여러 가지 특성을 보이는 것을 숙어라 하고, 그 외의 것을 통사적으로 공기하는 통사적 연어(collocation), 형태론적으로 강한 결합관계에 있는 것을 형태적 연어로 분류하였다. 이에 의하면, "파리를 날리다"나 "비행기를 태우다"와 같은 문장은 숙어이고, "절대로 -아니다"나 "설마 -르까" 등은 통사적 연어, "-에 대해"나 "-ㄴ가 보다" 등은 형태적 연어로 구분할 수 있다. 이들 형태소를 하나의 단위로 연결하지 않고 문장을 분석하면, 중요한 의미 정보를 잃을 뿐만 아니라, 분석에서 복잡도만 증가하게 된다. 따라서, 이들은 정보검색에서 불필요한 불용어의 제거에 이용될 수 있을 뿐만 아니라, 개개 어휘의 의미 정보가 필요한 언어처리나 기계번역과 같은 응용 분야에서 하나의 단위로 인식되고 처리되어야 한다.[10]에서는 각 형태소에 품사를 부착하는 품사 태깅 단계에서 묶인말이라는 형태소 패턴을 사전에 저장하고 이를 이용하여 태깅을 시도하였다. 여기에서는 관용어 중에서 통사적으로 공기하는 것을 묶인말로 정의하였다. 이를 위해 강한 어순 제약을 가지며, 다른 요소의 삽입이 없이 공기하는 형태소열만을 묶인말로 제한하였으며, 어미-보조용언을 포함하여 150개의 묶인말과 200개의 어미-보조용언을 사전에 등록하였다.

한국어와 유사한 활용을 하는 일본어에서도 이와 유사한 연구가 진행되었는데, [3]은 형

태소분석의 용이성 및 모호성 해소를 위해 용언의 활용형을 모두 하나의 엔트리로 사전에 등록하였다. 예를 들어, "책을 읽지 않았다/ほんおよみない"에서 "읽다"의 원형 "よむ"의 활용형 "よ"와 "-지 않-"을 의미하는 "みな"를 사전의 하나의 엔트리로 보았다. 즉, 활용형의 어미인 "み"와 후행하는 양상인 "ない"의 "な"를 결합하여 하나의 단어로 간주한 것이다. 이것은 다양한 활용형이 자주 나타나는 첨가어에서는 아주 많은 활용형을 사전에 등록해야 하며(예를 들어, "よみなさい"의 "みなさ"), 이러한 활용형은 기존의 단어와 중복됨으로써 다른 모호성을 야기할 수 있다[1].

이상의 연구들을 종합해 보면, 언어처리 기술을 실제 응용분야에 활용하기 위해서는 언어분석의 단위를 언어학적 관점보다는 컴퓨터를 이용한 언어처리의 관점에서 효율성을 중시한 언어 분석 단위와 이를 처리하는 방법이 반드시 필요함을 알 수 있다.

3. 효율적 언어 분석의 단위

이 장에서는 언어분석에서 각 단계별로 분석의 효율성을 높이기 위한 분석 단위를 살펴보고 이의 특징을 논한다. 이를 위해, 한국어 언어분석을 위한 형태소 한단위 및 어휘의 의미적 한단위 인식을 제안하고, 기존의 연구에서 이용되고 있는 구문 형태소[18]를 포함한 언어분석을 위한 한단위 인식에 대해 설명한다.

3.1 형태소 한단위

언어분석의 가장 첫 번째 단계인 형태소분석의 목적은 최소의 의미를 가지는 형태소 단위로 문장을 분해하는 것이다. 2장에서 설명한 바와 같이 한국어는 교착어 특성에 따라 조사나 어미 부분에서 여러 형태소가 결합하여 하나의 단위 형태소로 해석되는 경우가 많다. 앞에서 살펴본 "철수까지로밖에는"과 같이 "까지로밖에는"은 네 개의 조사가 결합하여 하나의 조사를 이루고 있다. [7]에 의하면, 단독으로 사용되는 조사가 60여개, 두개의 조사가 연속되어 결합한 경우가 143개, 세가지 조사의 결합이 45개, 네가지 조사의 결합이 12개로 분류되었는데, 이러한 조사를 조사 사전에 수록하고, 각 조사를 대표하는 대표조사를 해당 조사의 주요한 자질로 간주하고, 최장일치의 방법을 적용하면 긴 조사가 문장에 나타날수록 오히려 형태소분석의 결과가 줄어 들 것이다.

조사나 어미 이외에도 숫자, 수사 및 수 관형사와 같은 형태소는 주위의 여러 형태소와 결합하여 하나의 수사열로 처리되는 것이 구문적으로 훨씬 간단해질 뿐만 아니라 시간, 날짜, 혹은 수량을 나타내는 하나의 단일성을 가지기 때문에 의미적으로도 간단해진다. 예를 들어, "2001년 3월 24일", "10시 35분", "130kg", "삼만사천 칠백 오십원" 등을 구문분석을 위한 하나의 단위로 형태소분석 단계에서 분석하면, 구문분석의 계산 복잡도를 크게 줄일 수 있다. 이들을 인식하기 위한 규칙은 비교적 정규화되어 있어, 학습을 통해 쉽게 얻을 수 있다. 이를 통해 얻어진 규칙들의 예를 살펴보면 다음과 같다.

① 시간표현 : D+ 시간_U [D+ 시간_U]*
② 화폐표현 : D+ 화폐_U [D+ 화폐_U]*

또한, 숫자나 날짜 이외에 형태소분석에서 고려해야할 가장 큰 분야는 인터넷 환경에서 정보의 유통이 많아지면서 다양한 외국어 표현과 외국어와 숫자나 심볼이 결합된 형태의 형태소를 어떠한 관점에서 볼 것인가의 문제이다. 예를 들어, "CD-ROM과", "A/S를"과 같은 어절을 형태소분석하면 다음과 같은 결과가 나타나는 것이 일반적이다.

③ CD/외래어 + -/sym + ROM/외래어 + 과/조사
④ CD/외래어 + -/sym + ROM/외래어 + 과/명사
⑤ CD/외래어 + -/sym + ROM/외래어 + 고/동사 + 아/어미
⑥ A/외래어 + //sym + S/외래어 + 를/조사

그러나 "CD-ROM"이나 "A/S"를 하나의 단위로 보는 것이 바람직하지만 이러한 단어를

1) よる 명사로 "나머지/시대/밤"을 의미하고, 조사로 "-야/-여"를 나타낸다. "みな"는 대명사로 "모두/전부"를, 부사로 "몽땅"을 나타낸다. 또한, "い"는 명사로 "우물/위/뜻"을, 조사로 명령문에 쓰인다.

모두 사전에 등록하는 것은 불가능하다. 따라서, 영어나 심볼, 숫자가 결합하여 하나의 형태소 단위를 이루는 형태소열에 대한 조사가 이루어져야 하며, 본 논문에서는 이러한 형태소도 하나의 단위로 분석 처리하는 것이 효율적이라고 생각한다. 이는 정보검색과 같은 자연어 처리 응용 분야에서 검색 시스템의 성능 향상에도 많은 영향을 끼치기 때문이다. 이러한 형태소들을 분류해 보면 [표 1]과 같다[1]. 형태소분석 전에 이러한 패턴들을 인식하기 위한 오토마타를 이용하면 쉽게 이들을 그룹화 할 수 있다.

[표 1] 외래어 형태소 패턴의 예

표현 형식	예
E+ [-/&.] E+	CD-R, MBC-TV, IBM-PC, A/S, S/W, TCP/IP, ON/OFF, I/O, M&A, R&D, AT&T, Q&A, S&P, ftp.exe
E+ { [-/] E+ }+	IBM-PC/AT
E+ [-/.] D+	ISO-9001, UNOSOM-2, MIG-21, B-29, UXNET/700, OS/2, V/386, Mr.2
E+ D+	NO2 F16 BK21 Mpeg2
H ₩- E+	하이네트-p
D+ [.,] D+	14.6, 1,300, 192,000
E+ D E+	KBS1TV, KBS2FM, H2O
E+ D+ ₩/ E+ D+	JTC1/SC21

이러한 형태소 한단위 인식을 통해, 시간이나 외래어 및 신조어의 표현에서 복잡한 어휘적 분석을 피하고 적절한 의미적 단위로 인식함으로써 분석의 다음 단계인 품사 태깅이나 구문분석 등의 언어처리에서 효율성을 얻을 수 있다.

3.2 구문적 한단위

구문분석이란 형태소분석 결과를 바탕으로 이들 사이의 문법적 관계를 밝히는 과정이다. 형태소분석의 결과가 많을수록, 구문분석에서 모호성은 크게 증대된다. 따라서, 형태소분석 과정이나 그 후처리 과정에서 언어학적 분석의 틀을 깨지 않는 범위에서 구문분석에 필요한 문법적/의미적 단위로 분석 결과를 만들어야 한다.

본 논문에서는 구문 형태소를 여러 기능 형태소들이 결합하여 하나의 구문/의미적 단위를 형성하는 형태소 열로 정의한 논문[18]의 효용성을 살펴보고 이를 검토한다. 한국어에서 구문 형태소로 정의할 만한 기능 형태소의 결합으로 크게 두 가지가 있다. 하나는 기능 형태소들이 결합하여 하나의 양상 자질을 나타내는 경우이고, 다른 하나는 기능 형태소와 용언이 결합하여 하나의 심층격 조사역할을 하는 형태소열이다. 양상이란 어떤 사건이나 행동, 상태에 대한 화자의 태도를 표현한다. 예를 들면, 겸양, 추측, 피동, 소망, 가능, 부정, 진행, 시도, 완료 등이다. 이 양상은 구문분석 단계에서 구구조 규칙에 의해 용언에 대한 부가 자질의 형태로 표현될 수 있고, 의미분석 단계에서 여러 지식을 이용하여 인식된 구구조를 하나의 의미 자질로 나타낼 수 있다. 또한, 심층격 조사상당 형태소 열은 구구조 관점에서 볼 때, 주어진 문장의 정확한 문법적 구조를 파악하는데 도움이 되지만 의미분석을 어렵게 한다. 이런 양상 자질이나 조사 상당 어구를 구문 형태소 단위로 구문분석 전에 인식하면 형태소의 모호성 축소와 구문분석 과정을 단순화시킬 수 있으며, 보조용언 및 의사 조사들을 하나의 단위로 인식할 수 있다. 예를 들어, "먹을 수 있지 않을까?"와 같은 문장은 "먹다"라는 용언에 가능/추측 등의 양상 정보가 부가된 것으로 분석할 수 있다. 이를 간단히 "먹[가능, 추측, 평서]"라는 단위로 가정하고 구문분석을 진행함으로써, 분석의 모호성을 상당히 축소할 수 있는 것이다. [표 2]는 한국어에서 표현 가능한 양상과 이들의 의미를 정리한 것이다[11].

1) E : 알파벳, H : 한글, D : 숫자, \ : Escape

[표 2] 한국어 양상류의 종류와 의미

종류	조동사	대표어	영어 표현
태	피동	이, 히, 리, 기, 받, 당하, 어지, 되	be -ed
	사동	이, 히, 리, 기, 우, 시키, 게, 하	make, have, get
상	시발	어 돌	begin, start
	종료	어 버리	finish
	보유	어 놓	keep
	진행	고 있	be -ing
	완료	어 있	have -ed
양상	시행	어 보	try
	희망	고 싶	wish, want, hope
	강세	어 대	heavily, hard
	의도	려고 하	intend, plan
	가식	ㄴ 척하	pretend
	가능	ㄹ 수 있	can, be able to
	추측	ㄴ 것 같	seem as if
부정	부정	지 않, 지 못하, 지 말	not
양상	불가피	ㄹ 수밖에 없	can't help -ing
	불가능	ㄹ 수 없	cannot
	습관	곤 하	used to
	재귀	게 되	came to, become
	시인	기도 하	really
	당위	어야 하	must, have to
	규정	ㄴ 것이	
	원인	기 때문에	because
	예정	ㄹ 것이	will, may, might

또한, 조사의 최장일치가 조사들의 나열을 하나의 단위로 묶는다면, 조사와 불구 용언이 결합하여 하나의 조사로 묶이는 것을 한단위로 처리할 수 있다. 예를 들어, 조사가 "대하다", "인하다", "비하다"와 같은 의존명사에서 파생한 용언이나 "비롯하다"와 같은 일반 용언과 결합하여 문장 내에서 조사 상당 어구의 역할을 수행하는 경우가 많이 발견된다. 이는 의미분석의 관점에서 형태소들을 관찰하였을 때 명백해진다. 예를 들어, "철수로 인하여 싸움에 졌다"라는 문장에서 "-로 인해"는 원인격 조사 상당어구로 간주할 수 있다. 이러한 단어열을 하나의 구문적 단위로 처리할 경우, 구문분석과 의미분석에서 많은 모호성을 제거할 수 있다. 이런 경우를 영어 표현으로 변환하면 단 하나의 전치사로 대응된다. [18]에서는 이러한 형태소열을 의사 조사(pseudo particle)로 정의하고 이를 구문 형태소의 범주에 포함시켰으며, [표 3]은 구문 형태소들의 예를 나타내고 있다.

[표 3] 의사 조사형 구문 형태소의 예

(-에) 달해, (-에) 따라, (-에) 비해, (-에) 의해, (-에) 처해, (-에) 한해, (-에) 반해, (-와) 같이, (-와) 견주어, (-와) 관련하여, (-와) 달리, (-와) 함께, (-와) 더불어, (-를) 비롯해, (-를) 통해, (-를) 향해, (-를) 두고, (-를) 맞아, (-을) 가지고 (-로) 말미암아, (-로) 미루어, (-와) 마찬가지로, (-와) 반대로, (-와) 별도로, (-에) 있어,

3.3 어휘의 의미적 한단위

[17]에서는 한글, 한자, 영어, 기호의 결합된 형태로 한 어절 또는 여러 어절에 걸쳐 구성되면서 하나의 뜻으로 사용되는 단어를 의미적 한 단어로 정의하였다. 최근의 정보추출관련 연구에서는 [표 4]와 같이 인명, 지명, 기관명 및 숫자 표현에서 단어들의 결합을 하나의 단위로 인식하거나 상품명이나 행사명과 같은 인공물을 하나 단위로 묶어서 인식하기 위한 연구가 진행되고 있다[5, 6].

[표 4] 의미적 한 단어의 예

단 위	예
행정구역	서울시 영등포구 여의도동 32번지
장소	쉐라톤 워커힐호텔 제이든 가든
성명	루이 15세, 빌 클린턴
단체명칭	전국대학생대표자협의회, 119구급대
행사명칭	제1회동아시아대회, APEC정상회담
직업(직위)	FIFA회장, 아-태재단이사장
사람	PLO의장
제품명	IPAQ 3850, hp deskjet 960c
기타의미	벙커C유, 헬기, 7.4 공동성명

이들을 일반적으로 개체명이라 하고, 개체명의 인식을 위해서는 개체유형의 자질을 학습

해야 하는데, 영어는 개체명이 대문자로 시작하는 특징 등을 가지고 있어, HMM을 이용한 비교사 학습을 통해 인식하는 연구가 있었다[1]. 반면에, 한국어는 철자특징이 부족하여, 규칙기반의 접근 방법이 비교적 많이 연구되어 왔다[9, 12]. 예를 들어, 지명의 인식을 위해서는 "-시", "-군", "-동"과 같은 행정구역을 인식하는 단서단어를 이용하고, 단체나 기관의 이름을 인식하기 위해서는 "협의회", "연합회", "-원"과 같은 단서단어를 구축해야 한다. 규칙 기반의 접근은 수작업에 의한 규칙작성으로 인해 많은 노력이 들기 때문에, 초기에 소규모의 규칙을 이용하고, DL-CoTrain[2]에 의해 비교사 학습으로 점진적인 확장을 하는 방법이 적은 노력으로 쉽게 성능을 얻을 수 있다. [14]는 한국어에 적합한 개체명 인식 방법을 제안하고 있는데, 그 예를 살펴보면, "최근 소설가 최인호씨는 그의 작품인 …"에서 개체명의 문자열인 '최인호'를 대상으로 철자정보인 '최인호', '최', '인호'를 추출하고, 개체명의 왼쪽 문맥인 '소설가'나 오른쪽 문맥인 '씨'를 찾을 수 있다. 이를 바탕으로, 이러한 철자정보와 문맥정보를 번갈아 적용하여 개체명을 인식한다. 또한, "소설가 황석영 선생의 신작인…"와 같은 문장에서 '소설가'로부터 '황석영'을 인명으로 판단하고, '황석영'으로부터 '선생'을 또다시 문맥정보로 유추하여 반복적인 작업을 통해 규칙을 확장하고 점점 더 많은 개체명을 인식해 나갈 수 있다.

이와 같이 개체명과 같은 어휘의 의미적 한단위 인식은 명사구인식 및 명사구의 수식범위로 발생하는 구문분석의 복잡도를 줄이면서, 의미분석에서 개체명의 의미를 밝히기 위한 노력을 피할 수 있어 언어분석의 정확성 및 효율성을 높일 수 있다.

[표 5] 상위 빈도수의 인접한 형태소 열

바이그램	트라이그램	4-그램
고/ecx 있/px	고/ecx 있/px 다/ef	고/ecx 있/px 습니다/ef ./sf
에/jca 대하/pvg	고/ecx 있/px 습니다/ef	고/ecx 있/px 다/ef ./sf
지/ecx 않/px	하/xsv 고/ecx 있/px	하/xsv 고/ecx 있/px 다/ef
ㄴ/etm 것/nbn	적/xsn 이/jp ㄴ/etm	것/nbn 이/jp 다/ef ./sf
ㄹ/etm 것/nbn	에/jca 대하/pvg ㄴ/etm	하/xsv 기/etn 로/jca 하/pvg
는/etm 것/nbn	고/ecx 있/px 는/etm	기/etn 로/jca 하/pvg 었/ep
에/jca 따르/pvg	기/etn 로/jca 하/pvg	하/xsv 고/ecx 있/px 습니다/ef
기/etn 위하/pvg	ㄹ/etm 것/nbn 으로/jca	ㄹ/etm 것/nbn 으로/jca 보이/pvg
어/ecx 주/px	ㄴ/etm 것/nbn 으로/jca	하/xsv 고/ecx 있/px 는/etm
게/ecx 되/px	에/jca 따르/pvg 아/ecs	하/xsv ㄹ/etm 예정/ncn 이/jp
어/ecx 있/px	에/jca 대하/pvg 어/ecs	이/npd 에/jca 따르/pvg 아/ecs
어/ecx 지/px	ㄹ/etm 것/nbn 이/jp	ㄹ/etm 것/nbn 이/jp 라/ef
기/etn 때문/nbn	기/etn 위하/pvg 어/ecs	되/xsv 고/ecx 있/px 습니다/ef
어야/ecx 하/px	것/nbn 이/jp 다/ef	고/ecx 있/px 는/etm 것/nbn
어/ecx 오/px	ㄹ/etm 예정/ncn 이/jp	하/xsv 기/etn 위하/pvg 어/ecs
지/ecx 못하/px	하/xsv 기/etn 위하/pvg	는/etm 것/nbn 이/jp 다/ef
을/jco 위하/pvg	하/xsv ㄹ/etm 것/nbn	기/etn 때문/nbn 이/jp 다/ef
ㄹ/etm 예정/ncn	것/nbn 으로/jca 보이/pvg	기/etn 로/jca 하/pvg 았/ep
을/jco 통하/pvg	ㄴ/etm 것/nbn 이/jp	어/ecx 있/px 습니다/ef ./sf
어/ecx 보/px	기/etn 때문/nbn 에/jca	에/jca 대하/pvg 어서/ecs 는/jxc

4. 실험 및 평가

2장 및 3장에서 살펴본 언어분석의 단위가 효율적인지 검증하기 위해 본 논문에서는 실제 사용되는 문서를 대량으로 분석하여 이를 검증하였다. 3장에서 살펴 본 구문 형태소를 찾기 위해 국어정보베이스[4]의 품사 태깅된 말뭉치에서 추출한 어휘화된 바이그램, 트라이그램, 4-그램을 대상으로 상위 빈도를 나타내는 형태소열 중 하나 이상의 어절에 걸쳐 있으며, 내용어가 포함되지 않고 기능어만으로 구성된 형태소열을 추출하였다. [표 5]에서 "때문", "예정"과 같은 단어는 내용어이지만 출현 빈도가 높고 자립성이 약하기 때문에 포함시켰다. 이들을 대상으로 강한 결합성 여부와 내용어에 어떠한 기능적, 의미적 정보를 부가하고 있는지를 기준으로 분류하였으며, [표 5]는 약 20만여 어절에서 얻어진 바이그램, 트라이그램, 4-그램 중 상위 20개를 보여주고 있다.[1)]

본 논문에서는 이중 어휘를 제외하고 인접한 품사들을 다시 출현 빈도수로 정렬하여 인접한 품사들 사이의 강한 관계를 바탕으로 분류하여 구문 형태소들을 정리하였다. 이들 강한 결합을 나타내는 품사열중 의미 있는 형태소 열은 "ecx px", "etm nbn", "etm ncn" 등이다. 형태소 및 구문, 의미적 한단위 분석을 위해 유한상태 오토마타(Finite State Automata) 형태로 구현하였으며, 언어분석 결과의 한단위 인식을 위한 알고리즘은 [표 6]과 같다.

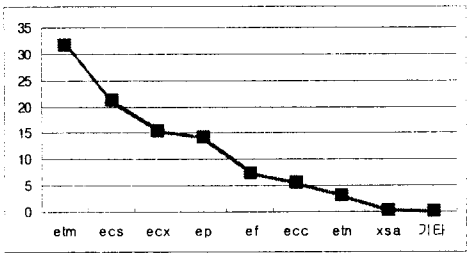
[표 6] 의미적 한단위 인식 알고리즘

```
grouping_morphemes() {
   for each phrase
   while (phrase) {
      while (check_type(phrase)) {
      // if phrase is mixed data type
          if ((E+ [-/&.] E+) ||
              (E+ { [-/] E+ }+) ||
              (E+ [-/.] D+) ||
              (E+ D+) ||
              (H ₩- E+) ||
              (D+ [.,] D+) ||
              (E+ D E+)) {
              grouping_symbols() ;
          }
          if (phrase is composed of hangul)
{
              processing_longest_prefix_suffix()
;
              processing_derived_affix() ;
              processing_auxiliary_verb() ;
              processing_bounded_noun() ;
              processing_pseudo_particle() ;

processing_verb_endings_noun_particles() ;
          }
          processing_symbol_number_english()
;
      }
   }
   for each phrase
   while (phrase) {
      lookup_named_entity_dic() ;
      detect_context_of named_entity() ;
      recognizing_named_entity() ;
   }
}
```

3장에서 설명한 구문 형태소 단위의 인식 과정에서 발생할 수 있는 문제점을 알아보기 위해 품사 태깅된 20만 어절에서 구문 형태소가 발생할 수 있는 형태소 환경에서 각 품사의 전이 확률을 조사하였다. [그림 1]은 용언에 후행하는 품사들의 분포 비율을 보여주고 있다.

1) 이하에서 설명하는 품사는 KIBS의 분류 기준을 따랐다.



[그림 1] 용언에 후행하는 품사의 종류

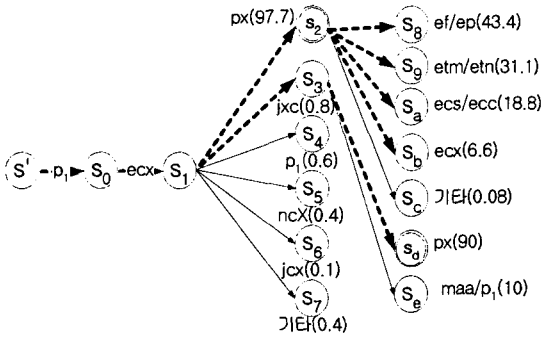
이중 많은 부분을 차지하는 것이 "p_1 etm"과 "p_1 ecx"이며, "ecx"나 "ep", "ef", "ecc", "etn"과 같이 용언과 함께 나타나는 어미가 주류를 이루고 있음을 알 수 있다. 이중 기타는 "jxc", "ncn", "xsn", "pvg", "jca", "nbn", "jp", "jcm" 등을 나타내며 전체의 약 0.4%를 차지하는데, 이는 모두 태깅 오류나 어절이 비문법적인 경우를 나타내고 있다. 예를 들어, "앞서/pvg+지나가/pvg"나 "토막내/pvg+가/pvg"나 "알리/pvg+어/jxc", "잇/pvg+은/jxc", "열/pvg+것/nbn", "열/pvg+수/nbn", "굴러가/pvg+도/jxc"와 같은 경우이다. 따라서, 구문 형태소 단위의 처리는 이러한 오류를 자동으로 걸러낼 수 있다. 이는 분석 과정의 효율성 뿐만아니라, 언어적 모호성이나 오류를 제약하는 역할을 할 수 있음을 의미한다. [표 7]은 "p_1 ecx" 품사열 다음에 나타나는 품사에 대한 발생률을 표시한 것이다[1]. 즉, 용언(p)과 보조적 연결어미(ecx) 다음에 나타나는 품사는 보조용언(px)이나 드물게 보조사(jxc)가 나타남을 알 수 있다.

[표 7] <p_1 ecx>에 후행하는 품사

품 사	발 생 률
px	97.6445
jxc	0.83135
pvX	0.63341
ncX	0.35629
jcX	0.11876
기타	0.41568

구문 형태소의 정의에 따르면, <p_1 ecx> 다음에는 <px>나 <jxc>의 품사를 가지는 형태소가 나타나야 한다. 실제 코퍼스의 분석 결과는 <pvX>나 <ncX>, <jcX>와 이외의 몇 가지 품사를 가지는 형태소가 나타나고 있다. <p_1 ecx pvX>가 나타나는 경우는 "흐르/pvg+어/ecx 가/pvg"나 "몰리/pvg+어/ecx 오/pvg"와 같은 경우인데, 이들은 하나의 용언으로 사전에 등록되는 경우가 일반적이며, 그렇지 않은 경우, "흐르/pvg+어/ecs 가/pvg"와 같이 형태소분석이 되어야 올바른 경우이다. 또한, <p_1 ecx ncX>나 <p_1 ecx jcX>와 기타 다양한 형태소가 발생하는 경우는 "놓/pvg+고/ecx 갈등/ncn"과 같은 경우와 "믿/pvg+지/ecx+를/jco", "내/pvg+고/ecx+야/ecx"등의 경우인데, 이들은 명백히 형태소분석 오류이다. 따라서, 용언과 보조적 연결어미를 따르는 형태소의 품사는 "px"와 "jxc"만이 존재한다. 그리고, <p_1 ecx px>로 인식된 형태소열 내에서는 약 4.5%의 오류가 발생하고 있는데, 이들 오류의 유형은 크게 두 가지 이유로 나눌 수 있었다. 첫째는, 하나의 단어로 굳어져 사용되는 단어에서 발생했다. 즉, "따라 가다", "되돌아 가다", "늘어 나다", "들어 가다", "옮겨 가다"와 같이 하나의 사전 목록으로 존재하는 단어를 "따르/pvg+어/ecx 가/px+다/ef"로 분석한 오류이다. 또한, 하나의 단어 형태로 자주 쓰이는 용언도 이와 유사한 형태를 보이고 있다. "사/pvg+아/ecx 먹/px"나 "쓰/pvg+어/ecx 먹/px"이 이와 같은 유형인데, 이는 형태소분석용 사전이 정교하지 못해서 생기는 오류로 볼 수 있다. 둘째는, "아/ecx 가/px"나 "아/ecx 오/px"와 같은 환경과 "-고/ecx 가/px"나 "-고/ecx 오/px"를 혼동한 경우이다. 예를 들어, "데리/pvg+고/ecx 가/px+다/ef", "들/pvg+고/ecx 오/px+다/ef" 등이 이에 해당한다. 이는 구문 형태소를 인식할 때 부분적으로 어휘를 사용함으로써 보다 정확하게 구문 형태소를 인식할 수 있음을 보여준다.

1) 발생률은 P(<품사> | <p_1 ecx>)일 확률



[그림 2] <p_1 ecx>에 후행하는 4-gram 품사열

[그림 2]는 <p_1 ecx px>나 <p_1 ecx jxc> 이후에 나타나는 품사들의 전이확률을 보여주고 있다. 이중 점선으로 표시한 부분이 구문 형태소의 인식 범위에 포함되는 것으로 오토마타의 정확성을 보여주고 있다. 이중 품사에 나타난 숫자는 전이확률을 백분율로 표시한 것이다. 즉, 오토마타에서 $\delta(s_1, px) = s_2$이며, 이때 P(<px>|<p_1 ecx>)=0.977임을 나타내고 있다. 실선은 실제 코퍼스의 분석에서 나타나는 전이과정이지만 분석 오류인 경우를 나타내고 있다. 이는 실제로 차지하는 비중이 매우 낮다. [표 8]은 이러한 오류의 예를 나타내고 있다.

[표 8] <p_1 ecx px> 환경에서 나타나는 오류의 예

오류인 분석 결과	정확한 분석 결과
뜻하/p_1+지/ecx+는/jxc 않/paa	뜻하/p_1+지/ecx+는/jxc 않/px
막가/p_1+아/ecx+도/jxc 되/mag	막가/p_1+아/ecx+도/jxc 되/px
걷/p_1+어/ecx	걷/p_1+어/ecx
붙이/px+ㄴ/jxc	붙이/px+ㄴ/etm
몰/p_1+고/ecx 올/px	몰/p_1+고/ecx 올/px
파장/jxc	파장/jxc
나타/p_1+아/ecx	나타/p_1+아/ecx
나/px+,/sp	나/px+아/ecc+,/sp

[표 9]는 양상 형태소들간의 비율을 보여주고 있는데, 한국어에서는 대부분 "진행", "부정", "상태", "피동", "봉사"와 같은 양상의 표현에 집중되어 있음을 알 수 있었다.

[표 9] 구문 형태소에서 양상의 분포

어휘1	어휘2	빈도수	발생률	양상의미
고	있	1339	28.4	진행
지/지는	않	560	11.9	부정
아/어	있	342	7.3	상태
아/어	지	324	6.9	피동
아/어	주	315	6.7	봉사
게	되	295	6.3	재귀
아/어	오	210	4.5	진행
지	못하	172	3.7	부정
어야/만	하	151	3.2	당위
아/어	보	120	2.5	시도
아/어	가	92	1.9	진행
어/어	내	85	1.8	종결
어	놓	75	1.6	보유
게	하	69	1.5	사동
아/어	나/나가	64	1.4	진행
아/어	버리	53	1.1	종결
어	드리	49	1.0	존경
아야	하	48	1.0	당위

이상의 실험을 통해서 알 수 있는 바와 같이 구문 형태소와 같은 의미적 한단위의 분석은 구문분석 및 의미분석 과정에서 오류나 모호성을 미리 축소할 수 있는 제약으로써 사용되어 분석의 정확도를 개선할 수 있었다. 그러나, 의미적 한단위 인식을 위한 방법이 유한상태 오토마타로 구현되었기 때문에 계산속도 측면에서는 기존의 방법에 비해 큰 차이가 없었다.

5. 결 론

본 논문에서는 언어처리를 위한 분석의 단위로, 언어학적 관점이 아닌 기계처리의 관점에서 연구된 내용들을 조사하고 언어처리의 주요한 단계인 형태소분석, 구문분석 및 의미분석에서 활용 가능한 분석단위를 제안하거나 살펴보았다. 형태소분석 단계에서 의미적 단위를 파악하여 가능한 최장의 형태소로 묶거나, 부가적 정보를 포함하는 형태소인 양상관련 형태소를 하나로 통합함으로써, 구문분석 및 의미분석의 효율성을 극대화 할 수 있는 몇 가지 방법을 설명하였다. 또한, 4장에서는 문장을 이해하는데 가장 중요한 구문분석을 효율적으로 진행하기 위한 구문 형태소 단위의 처리가 적합한지 대용

량의 실제 자료를 분석하여, 이의 실용성을 검증해 보았다. 이러한 분석 단위는 언어분석의 실용성측면에서 뿐만 아니라, 언어의 모호성을 축소하는 제약정보로도 활용될 수 있음을 관찰할 수 있었다. 향후 연구로는 구문 형태소 뿐만 아니라, 개체명 단위의 인식에 대한 명확한 분석 단위의 정의 및 정량적 평가를 통해 이의 유용성을 검증할 예정이다.

참고 문헌

[1] Bikel, D. M., Miller, S., Schwartz R., Weischedel, R., "Nymble: a High-Performance Learning Name-finder", In Proceedings of th 5th Conference On Applied Natural Language Processing, pp. 194-201, 1997

[2] Collins, M., Singer, Y., "Unsupervised Models for Named Entity Classification", EMNLP/VLC-99, pp. 189-196, 1999.

[3] Hisamitsu, Toru., Nitta, Yoshihiko., "An Efficient Treatment of Japanese Verb Inflection for Morphological Analysis", The 15th International Conference on Computational Linguistics, pp.194-200, 1994.

[4] KIBS : Korean Information Base System, http://kibs.kaist.ac.kr/

[5] MUC-7, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings /ne_task.html

[6] S. Sekine, H. Isahara, "IREX Project overview", Proceedings of the IREX workshop, 1999.

[7] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 대학원 컴퓨터공학과 박사학위 논문, 1993.

[8] 김영길, 양성일, 서영애, 김창현, 홍문표, 최승권, "한영 자동번역을 위한 한국어 구문 분석 전처리", 28회 한국정보과학회 추계학술대회, pp. 0175 ~ 0177, 2001

[9] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유명사의 추출과 분류", 한국정보과학회 가을 학술발표논문집, Vol. 27, No.2, pp. 170-172, 2000.

[10] 박혜준, 윤준태, 송만석, "말뭉치 품사 꼬리달기 시스템", 21회 정보과학회 춘계 학술발표 논문집, pp.829-832, 1994.

[11] 안동언, 조정미, 김길창, "영한 기계번역의 한국어 생성 시스템에서 조동사의 생성", 5회 한글 및 한국어 정보처리 학술대회, pp.533-544, 1993.

[12] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", 한글 및 한국어 정보처리 학술대회, pp. 292-299, 2000.

[13] 이정민, 배영남, 언어학사전, 박영사, 1993.

[14] 이현숙, 정의석, 황이규, 윤보현, "Cotraining 학습을 이용한 한국어 개체명 인식", 제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호, 2002.

[15] 이희자, "현대 국어 관용구의 결합 관계 고찰", 6회 한글 및 한국어 정보처리 학술대회, pp.333-352, 1994.

[16] 최재혁, 이상조, "양방향 최장일치법에 의한 한국어 형태소분석기에서의 사전횟수 감소방안", 한국정보과학회 논문지, 제 20권, 10호, pp. 1497-1507, 1993.

[17] 허윤영, 권혁철, "의미적 한 단어 유형 분석 및 형태소 분석 기법", 제 6회 한글 및 한국어 정보처리, pp.128-131, 1994.

[18] 황이규, 이현영, 이용석, "형태소 및 구문 모호성 축소를 위한 구문단위 형태소의 이용", 한국 정보과학회 논문지(B), Vol. 27, No. 7, pp. 784-793, 2000.

[19] 김영택 외, 자연언어처리, 생능출판사, 2001

[20] 김명철 외, 최신정보검색론, 홍릉과학출판사, 2001

박인철

1980 - 1984 전북대학교 전산통계학과 졸업(이학사)
1984 - 1986 전북대학교 전산통계학과 대학원 졸업(이학석사)
1992 - 1998 전북대학교 전산통계학과 대학원 졸업(이학박사)
1992 - 현재 호원대학교 컴퓨터학부 부교수로 재직 중
관심분야 : 한국어정보처리, 정보검색, 에이젼트, 시멘틱웹