인지과학, 제19권 제3호
Korean Journal of Cognitive Science
2008, Vol. 19, No. 3, 223~255.

문법성과 어휘 응집성 기반의 영어 작문 평가 시스템*

김 동 성[1] 김 상 철[2] 채 희 락[1†]

[1]한국외국어대학교 언어인지과학과
[2]한국외국어대학교 컴퓨터공학과

본 논문에서 우리는 문장의 문법성과 텍스트의 어휘 응집성 측정을 위주로 하는 영어 작문 자동 평가 시스템을 소개하려고 한다. 문법 검사를 위해서는 링크 파서를 사용하고 어휘 연쇄를 측정하기 위해서는 로제 시소러스를 사용한다. 자동 평가 시스템의 채점 신뢰도를 측정하기 위해서 자동 채점과 수동 채점의 결과를 통계적으로 비교한다. 카파 통계와 다국면 Rasch 모형에 따른 분석 결과 자동 채점은 수동 채점과 유사성이 크며 수동 채점과 비교해서 신뢰성에 특별한 문제가 없다는 결론을 내리게 된다. 본 연구의 가장 큰 의의는 다양한 종류의 기술과 도구를 바탕으로 신뢰할 만한 수준의 영작문 자동 평가 시스템을 개발했다는 것이다. 평가 대상이 문장 단위를 넘어 선 텍스트 단위이며, 단어나 문법 등의 형식적 측면만 검사하는 것이 아니라 내용적 측면도 평가한다.

주제어 : 문법성, 어휘 응집성, 자동 작문 평가 시스템, 카파 통계, 다국면 Rasch 모델, 파싱, 자연어 처리

---

* 본 논문은 세 저자가 공동으로 수행한 중소기업청의 2007년 산학연 공동 기술 개발 컨소시엄 사업 과제 "영어 작문 자동 평가 S/W 개발"에 기반을 두고 있다. 저자 중 김동성과 채희락은 한국외국어대학교 언어인지과학과에서 수행하고 있는 2단계 BK21 사업에 참여하고 있다. 김상철과 채희락은 2008년도 한국외국어대학교 교내 학술연구비의 지원을 받아 본 논문을 완성하였다.

† 교신저자: 채희락, 한국외대 언어인지과학과, 연구 세부 분야: 통사론, 인지과학
E-mail: hrchae@hufs.ac.kr

도 입

영어 작문을 평가하는 작업에는 많은 시간과 비용이 소요된다. 그래서 전산 프로그램에 대한 연구 및 개발이 최근 국내외에서 활발히 진행되고 있다(Burstein and Chodorow 1999; 진경애 2007; 김지은 · 이공주 2007). 영어 작문 자동 평가 시스템을 활용하는 것은 편리한 측면이 있지만 채점의 공정성과 타당성이 담보되지 않으면 시스템을 활용하는 것이 유용하지 않을 수도 있다. 따라서 시스템의 채점 방식이 올바르고 타당한지를 살펴 보는 것도 매우 중요한 일이다.

본 논문에서 우리는 영어 작문을 자동으로 평가하기 위한 시스템을 제안하려고 한다. 이 평가 시스템은 크게 자연언어 처리 기술을 활용한 평가 부분과 지식 체계를 활용한 글의 "응집성(cohesion)" 평가 부분으로 구성된다. 전자는 자연언어 처리 도구를 활용해서 평가 대상 글에서 활용된 어휘와 문법을 검사해서 오류의 정도를 매긴다. 후자는 "어휘 연쇄(lexical chain)"를 통해서 글의 응집성을 평가한다. 어휘 검사를 위해서는 "워드넷(WordNet)"의 어휘 목록을 활용하고 문법 검사를 위해서는 "링크 파서(link parser)"를 활용하였다. 그리고 어휘 연쇄를 측정하기 위해서는 "로제 시소러스(Roget's Thesaurus)"를 활용하였다.

영어 작문에 대한 평가는 객관식 선다형 문항에 대한 평가와는 여러 모로 차이가 있으며, 평가의 신뢰성은 다양한 방식으로 검증되어야 한다(이영식 1998; 지은림 1996). 평가의 신뢰도나 타당성을 확보하는 것은 자동 채점에서 매우 중요한 일이다. 채점의 신뢰도를 측정하기 위해서 본 연구에서는 자동 채점과 수동 채점을 통계적으로 비교 측정하였다. 수동 채점은 영어 독해에 능통한 대학 교수들에 의해 이루어 졌으며, ETS(Educational Testing Service)에서 제시한 에세이 채점 기준을 준수하였다(시사영어사 2006: 116).

수동 채점과 자동 채점 결과의 통계적 비교를 위한 방식은 크게 두 가지가 있다. 첫 번째는 두 가지 채점 결과가 서로 일치하는가를 비교하는 것이고, 두 번째는 전체 채점의 신뢰성에 비추어서 자동 채점이 합리적이고 공정한지 살펴보는 것이다. 첫 번째 과제를 위해서는 전문가 집단의 수동 채점 결과와 평가 시스템을 통한 자동 채점의 결과를 살펴 봐서 두 결과가 얼마만큼 일치하는지를 점검했다. 이를 위해 카파 통계를 통해서 집단 내의 일치도가 신뢰성이 있는지를 검사하였

다. 두 번째 과제를 위해서 평가 자체의 다국면적 요인을 측정하는 "다국면 Rasch 모형(Multi-facet Rasch Model)"에 따른 채점 신뢰도를 측정하였다. 이 모형에서는 평가의 주관성이나 오류에서 비롯된 평가자간의 편차를 분석해서 평가의 신뢰도를 파악하려고 하였으며, 자동 채점과 수동 채점의 신뢰도를 동일하게 측정하였다. 이런 방법으로 자동 채점의 수준을 수동 채점과 비교해서 자동 채점이 영어 작문 평가의 효과적인 측정 수단으로 활용될 수 있는지를 확인하고자 하였다.

본 연구의 가장 큰 의의는 문장 단위의 문법성과 텍스트 단위의 응집성 측정을 바탕으로 전문가 수준의 신뢰성을 갖춘 영작문 자동 평가 시스템을 개발했다는 것이다. 논문의 구성은 다음과 같다. 우선 2절에서는 영어 작문 평가 시스템에 대한 기존의 연구를 개관하려고 한다. 그리고 3절에서는 본 논문에서 제시하는 평가 시스템에서 철자 및 문법성과 응집성이 어떻게 측정되는지 그 방식을 소개할 것이다. 4절에서는 먼저 수동 채점 방식과 결과를 소개하려고 한다. 그리고 수동 채점과 자동 채점의 결과가 통계적으로 유사하며 자동 채점이 신뢰할 만한 수준이라는 것을 카파 통계와 다국면 Rasch 모델을 통해서 입증할 것이다.

기존 연구 개관

기존 연구들을 살펴 보면 영어 작문 자동 평가 방식은 크게 두 가지로 나눌 수 있다. 그 첫째는 NLP(Natural Language Processing) 기술이나 도구를 활용하는 방법이다. 예로서, "어휘 기반 계량 함수(lexically-based metrics)"를 이용하는 PEG(The Project Essay Grade) 시스템(Page and Peterson 1995), 문법과 철자 검색기를 이용하는 방법(Park et al. 1997), 담화 분석을 이용하는 E-Rator 시스템(Miltsakaki and Kukich 2004), "LG(link grammar)"를 이용한 방법(Londsale and Strong-Krause 2003) 등을 들 수 있다.

작문 평가 시스템으로 잘 알려진 PEG는 단어나 콤마의 수 등 형식적인 측면을 이용하기 때문에 문법적 오류와 의미적인 문제점을 포착하지 못한다. 또 하나의 대표적인 시스템인 E-Rator는 어휘나 문법 사항 위주의 채점을 하지만, 텍스트 "일관성(coherence)"을 이용하는 방법도 활용하고 있다. 텍스트 일관성이란 텍스트 요소들간에 서로 "논리적 모순 없이 연결되는지(make sense)"를 나타내는 지수이다. 여

기서 텍스트 요소란 문장이나 절을 말한다. 텍스트 일관성의 관계로는 "정교화(elaboration), 근거(support), 원인(cause), 예제(exemplification)" 등을 들 수 있지만, 이런 관계들에 대한 합의된 정의가 존재하고 않으며 더욱 이들 관계를 파악할 계산 방법은 아직 없다(Morris and Hirst 1991). Londsale and Strong-Krause (2003)은 LG 파서를 이용해서 텍스트의 문법적 오류 정도를 채점에 감안하는 방법을 제안하였지만, 이 역시 문법적인 오류만을 채점에 고려하는 한계가 있다.

이런 NLP 기술을 활용하는 방법은 아래의 두 번째 방법과 달리 다량의 샘플 텍스트를 가지고 시스템을 학습할 필요가 없는 장점이 있다. 그렇지만 NLP 기술의 한계로 인해서 어휘나 문법성만 평가하는 정도에 머물러 있으며 의미나 문맥적인 면을 효과적으로 고려하는 시스템은 아직 없다고 할 수 있다.

영어 작문 자동 평가 시스템에서 활용하고 있는 두 번째 방식은 정보 검색이나 패턴 인식에 사용하는 기술, 특히 텍스트 분류 기술이나 "군집화/클러스트링(clustering)" 기술을 활용하는 것이다. 이 방법의 핵심은 입력 문서의 특정한 패턴이나 특징을 파악하여 그 문서의 유형을 알아내는 방법으로 점수를 부여하는 것이다. 예로서, LSA(Latent Semantic Analysis)를 이용하는 IEA(The Intelligent Essay Assessor) 시스템(Landauer et al., 2003), PLSA(Probabilistic LSA)를 이용하는 것(Kakkonen and Sutinen 2004), "베이지안 확률(Bayesian probability)"을 이용하는 Rainbow 시스템(McCallum 1996), 하이브릿 방법의 CarmelTC 시스템(Rose, et al. 2003) 등을 들 수 있다.

대표적인 영어 채점 시스템의 하나인 IEA는 문서 검색이나 클러스터링에서 많이 사용하는 LSA 방법을 이용한다. 여기서는 문서가 가지는 어휘 패턴으로 해당 문서의 내용을 파악하게 되는데, 어휘 패턴으로는 문서의 내용상의 흐름을 파악하기가 어렵다. LSA를 확률 모델로 만든 PLSA도 LSA와 같이 어휘 패턴에 기반한 것으로 LSA의 한계를 넘지 못하고 있으며 채점 능력면에서도 LSA와 큰 차이가 없다(Kakkonen and Sutinen 2004). Rainbow는 베이지안 확률을 이용하는 것으로, 문서의 어휘 패턴이 주어질 때 해당 문서가 특정 부류에 포함된 확률을 계산한다. CarmelTC는 문장의 통사 분석 정보와 문서의 확률적 클래스 분류를 모두 활용하는 방법이다. 두 번째 방식에 의한 이들 시스템은 문장의 내용을 어휘 패턴 위주로 파악하기 때문에 문장간의 연결과 같은 문맥을 고려하는 데에 한계가 있다.

최근에 국내에서 개발된 영작문 자동 채점 시스템을 소개하고 있는 논문은 진경애(2007)와 김지은·이공주(2007)가 있다. 이들은 영어를 외국어로 배우는 한국의 중학생 영작문을 대상으로 문장 단위의 구문 오류 분석을 위주로 하는 시스템을 소개하고 있다. 이들은 한국어를 모국어로 하는 사람들이 영어 작문에서 주로 어떤 종류의 오류를 범하는지를 파악하여 이를 바탕으로 오류를 포함하고 있는 문장들도 처리할 수 있는 시스템을 개발했다. 문법적인 문장만 처리할 수 있는 시스템에서는 문장 분석 과정에서 비문법적인 요소를 만나면 어느 시점에서든 처리가 중단되기 때문에 오류의 심각성 정도를 알아 내기가 어렵다. 이런 측면에서 이 시스템의 의의를 찾을 수 있다. 그렇지만 이는 기본적으로 문장을 평가의 단위로 하고 있기 때문에 텍스트 차원의 일관성이나 응집성은 전혀 평가하지 못하고 있다. 또한 문장 차원의 분석도 학생이 제시한 답안과 교사가 제시한 모범 답안을 비교하여 그 유사성 정도에 따라 점수를 매기는 방식을 취하고 있기 때문에 자동화 정도도 높다고 할 수 없다.

평가 엔진 소개

이 소절에서는 본 논문의 평가 엔진을 소개한다. 이 시스템의 가장 큰 특징은 평가 대상이 문장 단위를 넘어 선 텍스트 단위이며 단어나 문법 등의 형식적 측면뿐만 아니라 응집성이란 내용적 측면도 평가한다는 것이다. 또한, 평가를 위해 모범 답안을 필요로 하지도 않는다. 평가 엔진은 크게 단어 철자 검사, 문장의 문법 검사 및 텍스트의 어휘 응집성 검사 부분으로 구성된다. 각각 서로 다른 언어 자원 및 시스템을 활용하였다.

철자 검사

철자 검사에서는 텍스트에서 사용된 단어들의 철자가 정확한지를 검사하게 된다. 일반적으로 전자 사전과 같은 형태의 언어 자원을 활용하는데, 본 연구에서 활용한 것은 워드넷(WordNet)이다. 워드넷은 영어 어휘들의 관계성을 정의한 사전으

로서 대략 206,941개의 단어들이 수록되어 있다. 워드넷에 수록된 개략적인 어휘 숫자를 품사별로 살펴보면 아래 표와 같다.

〈표 1〉 워드넷 사전의 어휘 개수

품사	개수
명사	146,312
동사	25,047
형용사	30,002
부사	5,580
합계	206,941

본 연구에서 철자 검사를 위해 활용한 것은 워드넷에 기재된 어휘들의 파생형이나 굴절형이다. 대략 20만여 개의 표제어가 실려 있으므로 하나의 표제어가 5개 정도의 활용형을 갖는다고 가정하면 전체 수록 어휘가 100만여 개 정도가 된다.[1] 워드넷에 수록된 이런 활용 어휘들을 대상으로 해당 어휘의 기본형이 있는지를 검사할 수 있다. 즉, 하나의 어형인 "토큰(token)"을 기본형인 "타입(type)"으로 검사하게 된다. 가령, 아래 (1a)와 같은 문장에서 likes는 기본형인 like로 전환하여 그 기본형이 사전에 있는지 검사하게 된다.

(1) a. She likes to play tennis.
b. She like to play tennis.

철자 검사에서 문제가 되는 것은 사전에 등록되어 있지 않은 "미등록어(unregistered words)"가 분석 대상이 될 경우이다. 사전에 등재되어 있지 않은 단어는 철자가 잘못된 것으로 처리되기 때문이다. 워드넷은 명사, 동사, 형용사, 부사의 네 가지 품사만을 다루고 있으므로 접속사와 전치사 등은 고려되지 않는다. 따라

1) 워드넷은 일반적으로 의미망으로서 활용이 되지만 온라인상에서 전자 사전으로도 사용되고 있다(http://www.wordnet-online.com).

서 자주 쓰이지만 사전에 없는 관계로 일상적으로 쓰이는 단어들도 잘못된 철자로 간주되는 오류가 발생할 수 있다. 미등록어는 신조어 혹은 사람 이름이나 지역 이름과 같은 "명칭(named entity)"이 대부분을 차지한다.[2] 이런 미등록어를 모두 사전에 수록한다는 것은 사실상 불가능하지만 본 연구에서는 문제를 최소화하기 위해 다음과 같은 접근법을 취하고 있다.

철자 검사를 위한 방식은 크게 세 단계로 구성되는데, 처음 두 단계는 미등록어 문제를 해결하기 위해서 설정되었다. 첫 번째는 (사전에 수록되어 있든 아니든 관계 없이) 자주 쓰이는 단어들의 목록을 미리 만들어서 이를 검사한다. 두 번째 단계는 사전에 등록이 되어 있지 않은 "두문자어(acronyms)"나 숫자가 쓰였는지 검사한다. 마지막 단계로 워드넷 사전을 활용해서 나머지 단어들의 철자를 검사하고 오철자 단어를 가려낸다. 각 단계에서 나온 결과를 종합하여 전체 채점을 하게 된다. 각각을 좀 더 자세히 설명하면 다음과 같다.

첫 번째 단계에서는 정보 검색 시스템 등에서 사용되는 "불용어 목록(stop list)"에 등록된 어휘들을 활용해서 검사한다(김영택 외 2001 참조). 이런 불용어 목록에는 의미가 빈약하여 정보 검색에서는 고려할 필요가 없지만 일상적으로는 많이 활용되는 단어들이 포함되어 있다. 가령 영어의 정관사나 부정관사인 the나 a는 일상적인 어휘들이지만 검색에 아무런 도움이 되지 않는다. 이런 단어들의 목록을 철자 검사에 활용하면 자주 활용되는 단어들을 (마지막 단계의) 사전을 통하지 않고 체크할 수 있으므로 철자 검사 과정이 간편해 진다. 또한 자주 쓰이는 단어지만 미등록어일 경우에도 검사가 가능해 진다. 본 연구의 철자 검사에서 활용한 불용어 목록은 980개의 단어로 구성되어 있는데, Oracle & ConText와 캔사스대학 및 오

2) 전산 정보 처리에서 "명칭 인식(named-entity recognition)"은 매우 중요한 분야 중 하나이다. 여러 어휘들 중 실제 검색어로 활용되는 어휘들은 주로 명칭과 연관된 어휘들이다. 정보 검색을 위해서는 텍스트에서 활용된 어휘들 중에서 정보 검색의 대상이 되는 명칭 어휘들을 산출하고, 이들 중에서 실제 검색의 대상이 되는 어휘들을 선출하여야 한다 (http://en.wikipedia.org/wiki/Named_entity_ recognition). 이 문제를 해결하기 위한 여러 알고리즘이 제시되었는데, 현재의 연구 방향은 코퍼스나 웹에서 수집된 데이터를 중심으로 학습하고 이를 통해서 분류하는 기계 학습(machine learning) 방식이 널리 활용되고 있다. 여러 기계 학습 방식 중 확률 기반의 HMM(Hidden Markov Model)이 제일 우수한 성능을 보인다고 한다(Zhuo and Su 2002).

하이오주립대학에서 만들어낸 SMART를 합쳐서 만들었다. 이 목록에는 0에서 99까지의 숫자, a나 the와 같은 관사, 전치사, 접속사, 약어, 축약어(You'll, I'm,...)와 로마자 숫자(x, xi, xii,...) 등이 포함되어 있다.

두 번째 단계에서는 두문자어나 숫자로 된 어휘들이 있는지 검사한다. 두문자어는 작문을 하면서 자주 활용할 수 있는 것으로서, 가령 "Limited-English-Proficient"을 LEP와 같이 약어로 쓸 수 있다. 이런 경우에 사전에 등록되지 않아 잘못된 철자로 간주될 수 있기 때문에 특별히 따로 처리할 필요성이 있다. 또한 불용어 목록에는 0에서 99까지만의 숫자가 등록되어 있기 때문에 그 이상의 숫자를 처리해야 할 필요성이 있다. 철자 전체가 (100 이상의) 숫자로 되어 있으면 이를 오철자 어휘로 분류되지 않도록 하는 과정을 거친다.

마지막으로 위의 처리 과정을 거치고 남은 단어들이 워드넷 사전에 등록되어 있는지를 검사하고 그 결과를 산출한다. 결과물에는 사전에서 발견되지 않은 단어들이 따로 표시되어 있다. 예를 들어서 아래와 같이 <...> 안에 들어 있는 단어는 사전에 등재되어 있지 않은 것이다.

(2) Hello. Let me introduce myself. I am a <juniorl> at .....

위에서 <juniorl>는 철자에 오류가 있는 단어이다. <...>로 표시되지 않은 단어는 올바른 철자를 가진 것이며 이들의 개수를 전체 단어 개수로 나누어서 백분율로 점수를 산출한다. 철자 검사의 점수는 아래와 같이 산출한다[3].

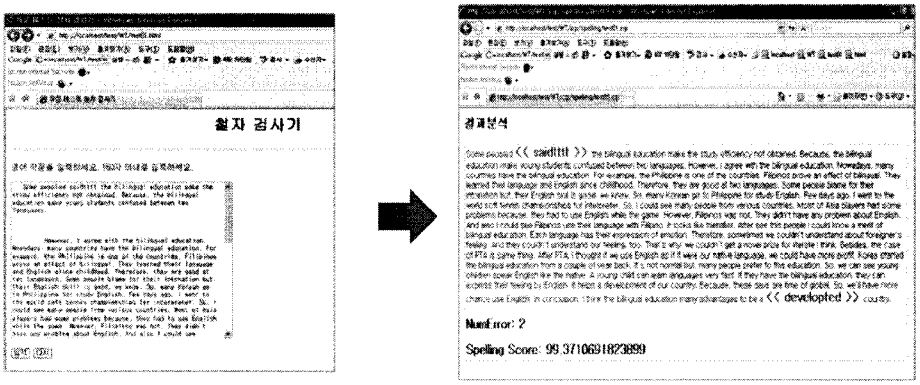
(3) 철자 점수 = (철자가 정확한 단어/전체 단어) x 100

위에서 언급했듯이, 철자 검사의 주요 문제점 중의 하나는 사람 이름이나 지역

3) 하나 이상의 철자가 틀린 경우 틀린 철자 수에 관계 없이 모두 철자가 틀린 단어로 간주된다. 그러나 학습자의 수준이나 어휘의 난이도에 따라서 달리 처리되어야 할 수도 있다. 예를 들어서, 어려운 단어의 경우 철자가 하나 틀린 것과 두 개 이상 틀린 것을 구별하여 채점하는 것이 더 합리적일 것이다. 그렇지만 본 연구에서는 이런 변수들을 고려하지 않고 모두 동일하게 처리하였다.

이름과 같은 명칭에 대한 인식이다. 워드넷에도 인명과 지명이 포함되어 있지만 백과사전들처럼 많이 포함되어 있지는 않다. 워드넷의 이런 측면 때문에 (인명이나 지명이 포함된 텍스트일 경우에) 채점 결과에 문제점이 없는지를 살펴볼 필요가 있었다. 실험을 통해서 여러 작문을 채점해 본 결과 지명이나 인명은 100점 만점에 2~5점 정도의 미미한 점수 차이밖에 유발하지 않는 것으로 관찰되었다. 이처럼 지명이나 인명과 같은 어휘들의 점수 기여도가 미미하였기 때문에 이들 어휘를 특별히 따로 처리하지는 않았다. 즉, 본 시스템에서는 워드넷에 수록되지 않은 어휘들을 모두 잘못된 철자로 취급하였다.

철자 검사 과정과 결과를 그림으로 살펴 보면 다음과 같다. 먼저, 입력된 텍스트에서 철자에 오류가 있는 단어가 발견되면 모두 표시가 된다. 다음으로, 그 발견 결과를 (3)에서 제시된 수식에 대입하여 계산한 결과가 나타난다.

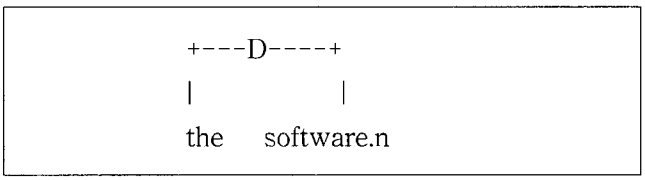


〔그림 1〕 철자 검사 과정/결과

문법 검사

본 연구에서는 파서를 활용해서 문법 검사를 자동으로 처리하였다. 파서는 문장의 문법성을 검사하는데, 올바른 문장과 잘못된 문장을 각각 파싱 성공과 실패로 나타낸다. 다시 말하면, 파싱에 성공한 문장은 문법적인 문장으로 간주되고 파싱에 실패한 문장은 비문법적인 문장으로 간주된다.[4)]

우리는 카네기멜론 공대의 Davy Temperley와 Daniel Sleator가 주도해서 만들어 낸 의존성 문법 이론 기반의 링크 파서를 활용하였다(Sleator and Temperley 1991).[5] 링크 파서는 기본적인 파싱과 더불어 여러 자연언어 처리 시스템에서 활용되고 있다.[6] 링크 파서는 의존 문법의 일종인 링크 문법을 활용한다. 다른 문법과의 차이성을 중심으로 간단히 소개하면 다음과 같다. 링크 파서는 단말 기호(terminal symbols)인 단어간의 문법적 관계에 따른 연결을 통해서 문장을 분석하는데, 단어들을 문법적 유형에 따라서 연결한다. 링크 파서를 활용해서 문장을 분석한 예를 보자. 아래의 그림에서 형태 D는 관형사와 명사를 연결하게 된다.



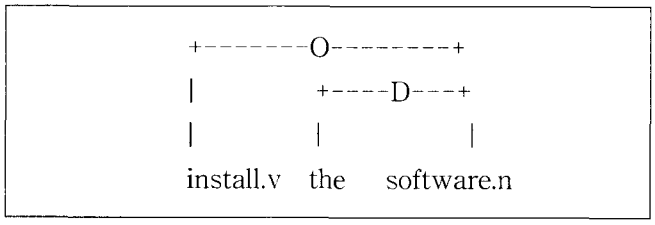
〔그림 2〕 관형사-명사 구구조 분석

직접 목적어가 동사에 연결되면 형태 O의 관계로 다음과 같이 분석된다.

4) 본 연구에서는, 어떤 문장이 비문법적인 것으로 판정되는 원인은 여러 가지 유형이 있지만, 일단 파싱에 실패하게 되면 그 오류 유형에 관계 없이 모두 동일한 정도의 비문법성을 가지는 것으로 처리된다. 그렇지만 문법성은 오류의 유형뿐만 아니라 문장의 복잡성과 작문자의 영어 수준 등도 고려해서 종합적으로 판단해야 할 정도성의 문제이다. 특히, 교육적인 측면에서는 작문자의 수준에 따른 채점이 되어야 그 효과가 증대되리라고 생각한다. 좀 더 효율적인 시스템이 되기 위해서는 이런 변수들이 모두 고려되고 비문법성도 양적으로 측정할 수 있어야 하겠지만 관련 연구는 다음으로 미루기로 한다.

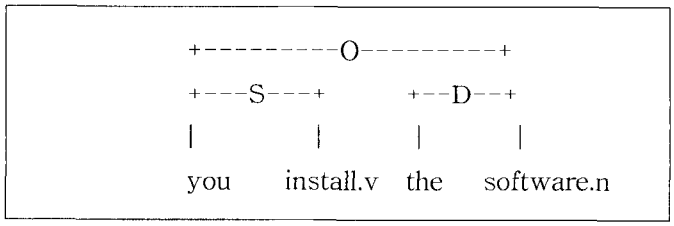
5) 다음의 사이트를 방문하면 링크 문법과 파서에 대한 정보를 얻을 수 있으며 파서 자체도 구할 수 있다(http://www.link.cs.cmu.edu/link/).

6) 링크 파서는 또한 AbiWord라는 무료 워드프로세서 프로그램에서 문법적 내용을 검사하는 기능으로 활용되고 있다 (http://www.nl.abisource.org/). 또한 언어 교육에서 이를 활용하기도 하였다(Brehony and Ryan 1994).



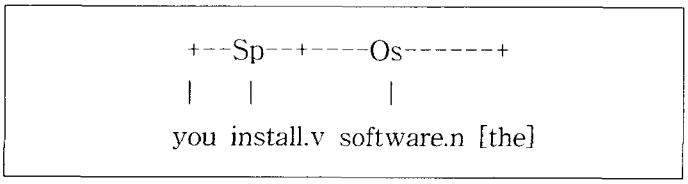
〔그림 3〕 동사-목적어 분석

마지막으로, 주어가 더해지면 주어 관계인 S와 목적어 관계 O를 연결하여 아래와 같이 분석한다.



〔그림 4〕 주어-동사구 분석

문법적인 문장은 관계가 서로 중첩되지 않고, 문장의 모든 단어가 연결되어야 하며, 문장에서 사용된 모든 관계성이 올바르게 연결된다. 만약 비문법적인 문장이 입력되면, 관계성이 올바르게 연결되지 않으며 연결성이 파악되지 않는다. 따라서 파싱에 실패한 어휘들은 연결되지 않은 것으로 표시된다. 아래 그림을 살펴 보자.



〔그림 5〕 비문법적인 문장

위의 그림은 문법적으로 올바르지 않은 'You install software the'가 입력되었을 경우에 그 처리 결과를 보여 준다. 그림에서 the는 연결되지 않고 떨어져 있는 단어로 나타내는데, 파싱에 성공하지 못했고 관계성이 정립되지 않았다는 것을 나타난다. 이런 방식으로 비문법성을 야기하는 문제의 단어를 표시한다. 문법 검사에서는 비문법성을 유발하는 부분을 표시해서 작문을 한 사람에게 피드백을 줄 수 있다. 출력은 "펜 트리뱅크(Penn Treebank)" 형태로도 가능하다. 아래 그림은 트리뱅크 형태의 출력을 보여 준다.

```
(S (NP The quick brown fox)
   (VP jumped
      (PP over
         (NP the lazy dog))))
```

〔그림 6〕 트리뱅크 형태의 출력

링크 파서는 다양한 문법적 내용을 처리할 수 있으며, 언어학적으로 의미 있는 내용을 정리하면 다음과 같다[7].

(4) a. 명사와 관련된 제약 조건
b. 관형사와 대명사의 오류
c. 수의 일치
d. 시간 표현
e. 동사와 연관된 문법성
f. 전치사와 연관된 문법성
g. 형용사와 연관된 문법성
h. 부사와 연관된 문법성
i. 질문 및 관계절 오류

7) 링크 파서가 처리할 수 있는 문법 현상은 매우 다양하다. 다음 사이트를 방문하면 그 내용을 확인할 수 있다(http://www.link.cs.cmu.edu/link/batch.html).

j. there/it/this의 특별 용법
k. 비교 구문
l. so...that/such...that과 같은 특수 구문
m. 하위절 연결
n. 등위절 연결
o. 등위절과 연관된 복수 표지
p. 동명사 구문
q. 특수 주어 (부정사, 구, 간접 의문문)

파싱의 성공률을 높이기 위해서 먼저 텍스트를 전처리할 필요성이 있다[8]. 예를 들어, 여러 구두점 중에서 [;], ['], [&]과 같은 기호는 파싱에 문제를 야기한다. 이를 처리하기 위해서 [&]는 and로 대체하고 나머지 기호들은 삭제하였다.

파서는 사전을 내장하고 있으며 그 속에 등록된 모든 단어들에 대한 문법적 지식을 가지고 있다. 가령 like라는 동사는 주어와 목적어가 필요한 타동사라는 정보가 시스템에 기재되어 있다. 문제는 파서 사전에 등재되어 있지 않은 단어들인데, 해당 문법 정보가 없어서 파싱 실패로 판별될 수 있다. Sutcliffe and McElligott (1994)는 링크 파서를 활용한 테스트에서 등재 어휘의 수에 따라 정확률을 약 40% 이상까지 상승시킬 수 있다고 하였다. 본 연구에서는 파서 사전의 등재 어휘 수를 늘리기 위해 워드넷 사전의 일부 어휘를 링크 파서 사전에 추가하였다. 추가된 어휘는 인명이나 지명과 같은 고유 명사를 포함하고 있으며 대략 1만 5천여 개 정도이다.[9]

---

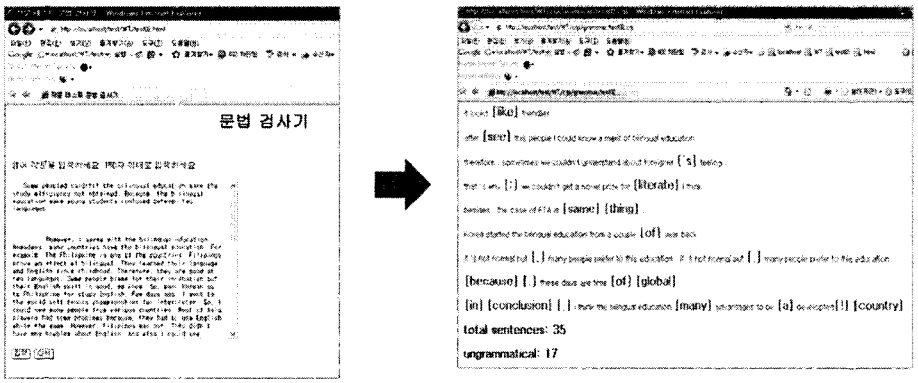
8) Sutcliffe and McElligott (1994)에서는 링크 파서의 초기 버전을 연구하면서 여러 문제점을 거론한다. 현재 버전 4.0까지 출시되었는데, 초기 버전의 거의 모든 문제점은 해결되었다. 그렇지만 구두점(punctuation mark)과 연관된 문제점은 여전히 남아 있다.

9) 링크 파서에 적합한 문법 정보를 담고 있는 추가 어휘들은 Elliot Turner에 의해서 개발된 것이다(http://www.eturner.net/linkgrammar-wn). 어휘 추가 후의 실제 테스트 결과는 추가 이전보다 대략 1-2% 정도의 정확도 향상을 가져 왔다. 이렇게 소폭으로밖에 향상하지 않은 이유는 전체 실험 양이 매우 적었기 때문일 것이다. 또한 (일반적인 어휘들은 링크 파서에 이미 내장되어 있으며) 추가된 단어는 높은 수준의 어휘이지만 평가 대상이 된 코퍼스는 일반적인 내용이어서 추가 어휘가 크게 활용되지 못한 연유도 있으리라고 생각한다.

일반적으로 문법 검사 결과는 문법적으로 올바른 문장의 개수를 세어서 전체 문장과의 백분율로 계산한다(Sutcliffe, Koch and McElligott 1994). 공식은 아래와 같다.

(5) 문법 점수 = (올바른 문장/전체 문장) x 100

실제 프로그램상의 문법 검사 과정과 처리 결과를 보면 아래 그림과 같다. 먼저 비문법적인 문장을 선별해 내고, 그 다음으로 (5)의 공식을 적용해서 문법 점수를 계산한다.



〔그림 7〕 문법 검사 과정/결과

어휘 응집성 검사

응집성은 텍스트의 구성 요소들이 동일한 대상이나 주제에 대한 이야기를 하고 있는지에 대한 정도성을 나타내는 개념으로 글의 짜임새를 나타내는 한 척도가 된다. 어휘 응집성은 단어들간의 의미적 관련성에 의해 발생하는 응집성을 말한다(Morris and Hirst 1991: 21). 동일한 어휘나 동일한 의미적 자질을 가진 어휘들이 나열되면 기본적으로 텍스트의 응집성이 커지게 된다. 예를 들어서, 탈 것의 의미로 bus, car, vehicle을 나열하면 어휘들의 구성에서 탈 것이라는 주제로 텍스트의 응집

성이 생기게 된다. 따라서 어휘들이 연쇄적으로 연결되면 응집성이 더 좋은 것으로 판별된다. 이런 연결성을 "어휘 연쇄(lexical chains)"라고 한다(Morris and Hirst 1991: 23). 어휘 연쇄는 문법적 구조와 상관 없이 텍스트의 응집성을 판별하는 데에 활용되는데, 하나의 개념적 구성으로도 구조화될 수 있다. 아래의 예에서처럼 어휘 연쇄가 place라는 하나의 개념 구성을 이루는 요소로 활용될 수 있다. 즉, Seoul, capital, city, inhabitant는 하나의 개념인 place를 설명하기에 적절한 연쇄를 만들 수 있다.

(6) Seoul -> capital -> city -> inhabitant -> place

작문에 있어서 텍스트의 응집성은 매우 중요한 요소인데, 이런 응집성의 정도를 판별할 수 있는 주요 방법 중의 하나가 어휘 연쇄를 측정하는 것이다.[10] 의미적으로 관련이 있는 단어들을 모아 놓은 사전을 활용해서 관련 어휘들이 어떤 연쇄에 있는지를 측정하는 것이다. 이런 목적을 위해서 여러 언어 자원 중에서 개념별 분류 어휘집인 시소러스(thesaurus)를 활용하며, 주로 Peter Mark Roget에 의해서 작성된 로제 시소러스를 많이 활용해 오고 있다(Morris and Hirst 1991; Jarmasz and Szpakowicz 2003). 로제 시소러스에는 단어들을 분류하는 개념들이 계층적으로 구성되어 있기 때문에 동일한 계층에 들어가는 어휘들은 하나의 개념 집합에 포함되는 어휘군으로 분류할 수 있다.

로제 시소러스에는 어휘들이 개념들의 계층적 구조에 따라 배열되어 있는데, 이는 온톨로지의 구성과 흡사하다. 각각의 개념 항목에 명사, 동사, 형용사와 같이 품사별로 어휘들을 나열하였다. 하나의 개념 아래에 나열된 어휘들은 그 개념에 대한 하위어들이고, 하나의 개념 안에 나열된 어휘들은 의미적으로 서로 밀접한 연관성을 갖고 있다. 따라서 텍스트내 각각의 어휘 연쇄를 이루는 단어들은 하나

10) 물론, 어휘 연쇄만으로 텍스트의 응집성을 정확하게 측정할 수는 없다고 본다. 어휘 응집성 이외의 다른 요소들도 텍스트 응집성을 결정하는 데에 작용할 것이며, 어휘 연쇄만으로는 어휘 응집성조차 정확하게 측정할 수 있다고 보기는 어렵기 때문이다. 그렇지만 어휘 연쇄가 어휘 응집성과 더 나아가 텍스트 응집성을 결정하는 주요 요인이라는 것에는 이론의 여지가 없을 것 같다.

의 개념에 포함된다. 텍스트의 응집성은 이러한 개념들이 얼마나 응집되어 있는지에 따라 결정되므로 어휘 연쇄들은 텍스트의 응집성을 나타내는 척도가 된다.

로제 시소러스의 구성을 간단하게 살펴보면 다음과 같다. 큰 단위의 개념은 Class로 분류되며, Class는 다시 I, II, III, ...과 같은 세부 구성 단위로 이루어 진다. 그리고 이들은 다시 개별 개념 요소인 1, 2, 3, ...의 구성으로 나뉘며 이들은 하나의 단락(paragraph)으로 구성된다. 동일 단락에 들어 있는 단어들은 하나의 개념을 표현하기 때문에 모두 동질적인 단어들이다. 아래 그림은 life라는 단어가 [Matter - Organic Matter - Vitality]로 분류되어 있는 것을 나타낸다. 이 경우에 어휘 연쇄를 측정하는 방식은 다음과 같다. 만약 하나의 텍스트에서 life, living, vitality와 같은 단어가 연쇄적으로 활용되었다면, 이것은 Vitality의 407번 단락의 개념을 나타낸다고 할 수 있다.

```
Class 1
 :
Class 4 Matter
        I
        :
        III Organic Matter
            A
            B Vitality
                407 life, living, vitality,....
```

〔그림 8〕 로제 시소러스 구성의 예

개별 텍스트의 응집성은 해당 텍스트에서 출현하는 단어들이 같은 구성에 속하는지를 검사함으로써 측정할 수 있다. 검사 방식은 다음과 같다. 먼저 전체 텍스트에서 후보 어휘 집합을 추출한다. 그리고 추출한 어휘로 로제 시소러스를 통해 적절한 어휘 연쇄들을 만들어 낸다. 아래와 같은 텍스트가 주어졌을 때 어휘 연쇄를

구성하는 방식을 단계적으로 살펴보자.

(7) There is no **doubt** that, **globalisation** is such a new **event** in the **world**. **English** became **common language** for any kind of **international communication** such as **business**. **Learning English** is the most **important** thing for **immigrants** and their **family** to **successfully survive** in **English spoken countries**. However **Limited-English-Proficient** (LEP) **students** have **problem** with **learning actual knowledge** in **education** in **school**. Therefore such **country** like **United States** of **America promoted bilingual education system** to the **public schools**.

먼저 후보 어휘 집합을 추출한다. 이 때 고려되어야 하는 것은 해당 어휘가 의미적으로 불필요한 어휘인지 아닌지를 판단하는 것이다. 이전 소절에서 논의한 바와 같이 불용어 목록에 포함된 어휘들은 정보 검색에서 쓸모 없는 어휘들이므로 어휘 연쇄에서도 불용어 목록에 해당하는 어휘들을 고려하지 않는다. 위 텍스트에서 굵은 글씨의 어휘는 모두 고려의 대상이다. 다음으로, 추출 어휘가 구성하는 어휘 연쇄들을 만들어 낸다. (7)을 통해 추출한 단어들을 하나씩 검토하면서 각 단어가 로제 시소러스의 어떤 개념에 속하는지 찾아 내어 해당 단어를 그 개념에 속하는 것으로 배치한다. 후보군의 모든 어휘를 대상으로 이런 과정을 거치면 아래와 같은 개념 목록과 각 개념에 속하는 단어 목록이 만들어 진다.

(8) a. UNCERTAINTY: doubt
b. LANGUAGE: English, language, English, English, bilingual
c. BELIEF: doubt, school, schools
d. COMMUNICATION: communication, spoken, learning, knowledge
e. MANKIND: public, common

이런 방식으로 우리가 원하는 어휘 연쇄의 목록이 만들어 진다. 한 가지 유의할 점은 단어들은 기본적으로 중의적이거나 다의적일 수 있으므로 하나의 단어가 여

러 의미군에 포함될 수 있다는 것이다. 그렇기 때문에 하나의 단어가 동시에 여러 개념 연쇄에 포함될 수 있다.

각각의 어휘 연쇄가 길면 길수록 텍스트는 하나의 주제로 응집되어 있다고 볼 수 있다. 따라서 어휘 연쇄를 활용해서 텍스트의 응집성을 연산하기 위해서는 각각의 연쇄의 길이나 어휘 구성을 측정하면 된다. 로제 시소러스를 활용한 어휘 연쇄의 수량적 측정은 Barzilay and Elhadad (1997)에 의해서 다음과 같이 제안되었다[11]. 아래 공식에서 |LC|는 어휘 연쇄의 길이를 나타낸다. t는 어휘 연쇄 LC에 포함된 단어이며, freq(t)는 t가 그 어휘 연쇄에 몇 번 출현하는지 표시하는 빈도이다.

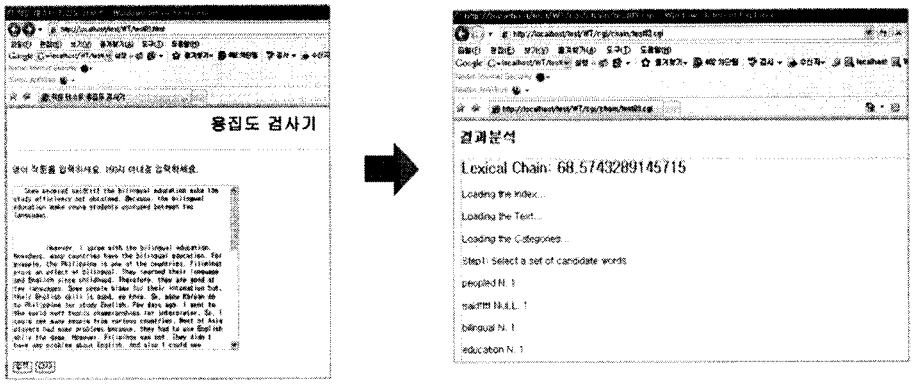
$$(9)\ score(LC) = (1 - \frac{|LC|}{\sum_{t \in LC} freq(t)}) \times \sum_{t \in LC} freq(t)$$

본 시스템에서도 위의 공식이 활용되었으며, 철자 검사와 문법 검사가 모두 백분율로 환산되었으므로 (9)의 계산 값도 백분율로 전환하였다. 백분율로 환산하기 위해서는 100을 곱해야 하나 위의 공식을 통해 산출된 점수가 100을 곱하기에는 값이 너무 크기 때문에 일률적으로 10을 곱해서 점수를 산출하였다.

(10) 어휘 연쇄 (응집성) 점수 = score(*LC*) x 10

프로그램을 통해서 응집성 검사를 한 과정과 결과를 그림으로 나타내면 다음과 같다.

11) Medelyan(2007)에서는 Barzilay and Elhadad(1997)에서 제안한 방식을 활용해서 작문을 채점하는 실험을 하였으며, 기계에 의한 채점과 사람의 채점이 서로 일치하는 결과를 보이고 있는 것으로 되어 있다.



〔그림 9〕 응집도 검사 과정/결과

평가 시스템의 신뢰도

수동 채점과 자동 채점

실험을 위해서 다섯 가지의 주제로 작성된 작문 코퍼스를 수집하였다. 개별 작문의 분량은 대략 A4 용지 1장 정도이며, 대학교 3-4학년 수준의 학생으로 구성된 영어 작문 수업을 통해서 수집되었다. 코퍼스는 각각의 주제당 5개씩 모두 25개의 문서로 구성된다.[12] 수집된 코퍼스는 상, 중, 하의 수준을 주제당 대략 1:2:2 정도의 비율로 유지하였다.

채점은 전문가들에 의한 수동 채점과 평가 엔진에 의한 자동 채점으로 나누어 실시했다. 수동 채점은 어휘 부문, 문법 부문 및 어휘 응집성을 포함한 구성 부문의 세 영역으로 나누어 각각 개별적으로 점수를 부여하는 방식으로 실시했다. 평

12) 작문의 주제는 TESOL의 작문에서 많이 거론되거나 TOEFL 작문 시험의 주제로 출제될 만한 것이다. 작문의 다섯 주제는 다음과 같다: 1. 자신을 소개하는 글을 작성하시오. 2. 캠퍼스를 묘사하는 글을 작성하시오. 3. 갖고 싶은 직업과 그 이유를 설명하시오. 4. 인터넷은 세기의 발명인가 아니면 해악을 주는 것인가에 따라서 다른 논증을 하시오. 5. 영어를 이중 언어로 하는 것을 찬성하는지 반대하는지를 기술하시오.

가 엔진이 측정한 것은 이전 절에서 기술한 내용과 동일하다. 수동 채점의 기준으로는 ETS(Educational Testing Service)에서 제시한 에세이 채점 기준을 따랐다(시사영어사 2006: 116). 그 기준은 부록 1의 표와 같다.

수동 채점의 전문가 집단은 미국에서 석박사 과정의 공부를 한 사람들로 영어 해독에 전혀 문제가 없는 두 사람으로 구성되었다. 그러나 그 전에 영작문을 공동으로 채점한 경험은 없었다. 두 명은 서로 상의 없이 독립적으로 채점을 하였지만 채점 전에 채점의 기준 및 형식에 대해서는 서로 논의를 하였다. 그리고 채점의 공정성을 위해서 학생의 수준이나 주제의 난이도에 대해서도 사전에 논의하였다. 학생의 수준과 주제의 난이도를 채점에 반영하는 것은 전체 채점의 공정한 분포를 결정하는 데에 도움이 된다.

수동 채점은 크게 어휘와 문법 및 응집성/구성을 대상으로 해서 각각 5점 만점으로 실시하였다. 채점의 기준은 부록의 채점 가이드를 바탕으로 하였다. 전체 채점에서는 각 항목의 중요성 정도에 따라 비중을 달리해야 하는데, 본 연구에서는 항목별 비중을 아래와 같이 책정하였다.[13]

(11) 전체 점수 = 0.1 x 어휘 + 0.4 x 문법 + 0.6 x 응집성

수동 채점과 자동 채점의 평균과 분산은 다음과 같다.[14] [15]

13) 채점 공식 (11)에서 0.1, 0.4, 0.6의 계수는 직관적/자의적으로 정해진 것이 아니라 (12)에서 활용한 카파 통계를 기반으로 산출하였다. 여러 수치를 대입해서 카파 통계를 적용하고, 수동 채점이 자동 채점과 일치하는지의 여부를 검사해서 그 비율을 최적화하였다. 다시 말하면, 0.1, 0.2, 0.3 또는 0.1, 0.2, 0.1과 같은 여러 조합의 계수를 대입한 결과를 수동 채점과의 일치도를 계산해서 이를 최적화한 계수를 선정한 것이다.

14) 어휘 자동 채점의 결과는 25개 모두 5점으로 일치하여서 분산이 0이다.

15) 합산 점수의 평균은 표에 나오는 채점의 평균을 수식 (11)에 대입해서 나온 결과가 아니고, 개별 실험 텍스트 25개를 각각 (11)에 대입해서 산출한 결과를 평균한 것이다.

〈표 2〉 채점 결과의 평균과 분산

		자동 채점	수동 채점 1	수동 채점 2
어휘	평균	5	3.7	3.7
	분산	0	0.9	0.9
문법	평균	2.8	2.9	3.6
	분산	1.0	1.0	0.9
응집력	평균	2.5	3.1	3.5
	분산	0.8	0.8	0.9
합산 점수	평균	2.8	2.9	3.6
	분산	0.5	0.5	0.8

위의 표에서 살펴보면, 각각의 채점 점수는 다르게 나타나나, 대부분의 경우에 분산은 다르지 않게 나타난다. 이는 개별 채점의 평균은 다를 수 있으나 채점 점수의 분포가 서로 같게 나타남을 말해 준다. 예를 들어서, 문법의 경우에 자동 채점과 수동 채점자 1 및 수동 채점자 2의 결과가 문법의 경우에 평균이 각각 2.8, 2.9, 3.6으로 다르나, 분산은 1.0과 0.9로 거의 동일하다.

채점의 일치도 및 신뢰도

시스템에 의한 자동 채점과 2명의 전문가에 의한 수동 채점의 내용을 바탕으로 채점 결과의 유용성을 판단함에 있어서는 각각의 채점이 서로 일치하는가와 전체 채점이 신뢰성이 있는가가 중요하다. 채점 결과가 서로 일치하는지는 동일한 채점 결과를 산출하는가의 문제이다. 전체 채점이 신뢰성이 있는지는 채점이 합리적인지와 전반적으로 채점의 결과가 공정성이 있는지와 관련된 문제이다. 본 연구에서는 두 문제에 대해서 각각 다른 통계적 평가를 적용 하였다.

첫 번째로 채점자간의 일치도 문제를 다루기 위해 본 연구에서는 카파 통계를 활용하였다. 카파 통계는 일치도 측정의 실험에서 일치하는 수준이 어느 정도의 통계적 분포를 갖는가를 측정하며 그 값으로 일치하는 수준이 타당한지를 검사할

수 있다. 카파 통계를 공식으로 표현하면 아래와 같다.[16]

$$(12)\ \kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

여기서 P(A)는 실제의 "관측 값(observed agreement)"을 나타내고 P(E)는 우연에 의한 "예상/기대 값(agreement by chance)"을 나타낸다. 카파 통계 지수의 분포는 완전한 일치인 1에서 우연에 의한 일치 수준인 0사이에 존재하는데, 카파 값의 변화에 따라서 일치하는 수준이 차이가 난다. Landis and Koch(1977)에 따르면, 카파 통계의 일치도의 범위에 따른 일치 수준은 다음과 같다.

(13) Kappa 통계의 일치도 범위와 일치 수준

0.00	거의 일치하지 않음 (Poor)
0.01-0.20	약간 일치함 (Slight)
0.21-0.40	보통 정도의 일치를 보임 (Fair)
0.41-0.60	상당한 정도의 일치를 보임 (Moderate)
0.61-0.80	실질적으로 일치함 (Substantial)
0.81-1.00	거의 완벽하게 일치함 (Almost perfect)

본 연구에서는 자동 및 수동 채점의 일치도를 카파 통계로 측정하였으며, 측정 결과 수치는 0.41이었다. 이는 위의 일치도 범위에 따르면 "상당한 정도"를 나타낸다.

다음으로 전체 채점의 신뢰도를 측정하기 위해서 다국면 Rasch 모델을 활용하였다. 이 모델은 다중 환경의 변인들이 통계적 모델에 어떻게 반영되는가의 문제를 다룬다. 작문의 경우에는 수행자의 능력과 시험의 난이도, 채점자의 엄정성이 채점에서 중요한 요소가 된다. 이런 경우에 전체 변인을 모두 고려하면 전체 채점의

16) "Kappa Statistics: an index which compares the agreement against that which might be expected by chance." (http://www.dmi.columbia.edu/homepages/chuangj/kappa/)

신뢰성을 측정할 수 있다. 작문 평가의 경우에는 여러 연구에서 이 모델을 채택해서 신뢰성을 측정했다(이영식 1998; Engelhard 1992).

다국면 Rasch 모델의 영어 작문 측정 수식을 간단히 표현하면 다음과 같다(이영식 1998: 182).

(14) $log(p_{nijmk}/p_{nijk-1}) = B_n - D_i - C_j - F_k$

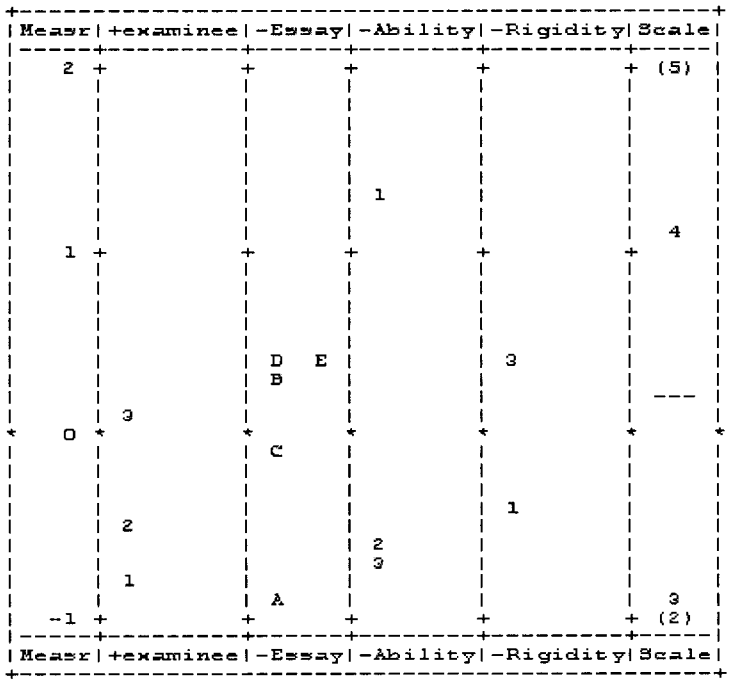
이 수식이 측정하는 것은 다음과 같다. 채점자 j가 학생 n에게 작문 주제 m에서 k로 평가하는 확률은 k-1로 평가될 확률과의 로그 비율의 통계적 분포로 결정된다. 이러한 통계적 분포는 피험자의 작문 능력(B_n)에서 작문 주제의 난이도인 D_i를 빼고, 다시 채점자의 엄정성이나 공정성인 C_j를 빼고, 다시 점수의 상승도인 F_k를 뺀 것이다.[17] 이 모델은 여러 국면을 고려한 통계적인 모델로 피험자의 능력, 문제의 난이도, 채점자의 공정성과 가능한 점수의 상승을 고려해서 영작문을 채점하는 데에 활용된다. 특히 Facets 프로그램은 이미 상용화된 것으로 대입 논술고사의 채점의 객관성 연구에도 적용한 바가 있다(지은림 1996).[18]

여기서는 자동 채점과 수동 채점을 모두 고려해서 측정하였다. 즉, 자동 채점이 수동 채점에 견주어서 어떠한 측면을 갖는지를 고려하였다. 다국면 Rasch 모델의 전체적 평가는 아래 그림에 나타난 바와 같다.

---

17) 이영식(1998: 182)에서는 다국면 Rasch 모델을 확률적 모델로 표현하고 있으나, 이 모델은 통계에 대한 모델이다. 즉, 변인들의 전체 시스템에서 분포에 대한 측정인 통계적 모델이지 출현되는 빈도를 전체 빈도에 비추어 보는 확률적 모델은 아니다. 따라서 본 논문에서는 통계적 모델로 설명한다.

18) Facets 프로그램은 http://www.winsteps.com/facets.htm에서 무료로 다운로드 받아 사용할 수 있다.



〔그림 10〕 다국면적 결과 분석 요약

위 그림에서 +examinee 열의 1은 자동 채점, 2와 3은 각각 수동 채점 1과 2를 나타낸다. -Essay 열에서 A, B, C, D, E는 각주 12에서 논의된 에세이의 주제에 따른 분류이다. -Ability 열은 개별 학생이 작문 수업에서의 학점을 토대로 작문 능력을 상, 중, 하로 1에서 3까지 책정한 것을 나타낸다. -Rigidity 열에서는 채점자의 엄밀성을 1에서 3으로 책정한 것을 나타낸다.

첫 번째 열은 단위가 logit으로 기준점이 0이며 최하치가 -1이고 최상치가 +2이다. 두 번째 열은 채점자별 점수의 전체 분포를 나타내는데, 다섯 번째 열의 엄밀성과 비교해 보았을 때, 3번 채점자가 가장 엄밀하게 채점한 것을 알 수 있다. 세 번째와 네 번째 열은 주제의 난이도 정도와 작문 참여자의 작문 능력의 정도를 통계적으로 측정한 것인데, 어떤 것도 2.0 이상이나 -1 이하인 것이 없어서 최하치나 최상치를 벗어난 결과가 없다. 따라서 정해진 분포 안에 채점 결과가 모두 분포함

을 알 수 있다. 이것은 자동 채점과 수동 채점 사이에 차이점을 발견할 수 없고 채점의 공정성도 수동 채점과 동일하게 유지되고 있다는 것을 보여 준다.

채점자의 분포도를 살펴 보면 다음과 같다.

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N examinee
75	26	2.9	3.03	-.81	.31	1.09	.4	1.07	.3	.91	1 1
78	26	3.0	3.15	-.52	.30	.59	-1.6	.59	-1.6	1.47	2 2
96	26	3.7	3.48	.14	.28	1.23	.9	1.33	1.2	.56	3 3
83.0	26.0	3.2	3.22	-.40	.30	.97	-.1	.99	.0		Mean (Count: 3)
9.3	.0	.4	.19	.40	.02	.27	1.1	.30	1.2		S.D. (Population)
11.4	.0	.4	.23	.49	.02	.34	1.4	.37	1.5		S.D. (Sample)

〔그림 11〕 채점자의 분포도

위 채점자 분포도에서 자동 및 수동 채점자들의 분포는 "외적합 표준화 값 (Outfit ZStd)"이 최상 값인 2나 최하 값인 -1을 넘지 않는다. 이영식 (1998: 189)에서 논의한 바와 같이 최상 값이나 최하 값의 범주를 넘는 값은 부적합한 것으로 취급된다. 따라서 전체 채점자들에게 엄밀성 및 학생의 작문 능력과 주제의 난이도에 따른 공정성이 유지되고 있음을 알 수 있다.

결론적으로, 채점의 결과는 전반적으로 자동 채점과 수동 채점간에 커다란 차이점을 보이지 않았다. 오히려 수동 채점간의 차이가 자동-수동 채점간의 차이보다 더 컸다. 그러므로 자동 채점의 결과는 수동 채점의 결과와 유사성이 크며 자동 채점의 신뢰성에 커다란 영향을 미치지 않는다는 결론을 내릴 수 있다.

결 론

본 논문에서 우리는 자연어 처리 기술과 지식 처리 기술을 활용한 영어 작문 자동 평가 시스템을 소개했다. 전자는 어휘와 문법을 검사하기 위해 활용되었으며 후자는 어휘 연쇄를 통한 글의 응집성을 평가하기 위해 활용되었다. 어휘 검사를 위해서는 워드넷 사전을 사용하고 문법 검사를 위해서는 링크 파서를 사용하였다.

그리고 어휘 연쇄를 측정하기 위해서는 로제 시소러스를 사용하였다. 이처럼 본 시스템은 다양한 종류의 관련 기술이나 도구를 활용하여 만들어 졌다.

자동 평가 시스템의 유용성은 그 신뢰도에 달려 있다. 채점의 신뢰도를 측정하기 위해서 본 연구에서는 자동 채점과 수동 채점을 통계적으로 비교했다. 카파 통계를 통해서 두 유형의 채점 결과가 서로 일치하는지, 집단 내의 일치도가 신뢰성이 있는지를 살펴 보았다. 그리고 다국면 Rasch 모형에 따라 전체 채점의 신뢰성에 비추어 자동 채점 결과가 합리성과 공정성이 있다고 볼 수 있는지도 검토해 보았다. 그 결과 자동 채점은 수동 채점과 유사성이 크며 수동 채점과 비교해서 신뢰성에 특별한 문제가 없다는 결론을 내릴 수가 있었다. 다시 말하면, 본 논문에서 제안한 자동 채점 시스템이 전문가들 수준의 평가 능력을 가지고 있다는 결론을 내릴 수가 있었다.

본 연구의 가장 큰 의의는 다양한 종류의 기술과 도구를 바탕으로 신뢰할 만한 수준의 영작문 자동 평가 시스템을 개발했다는 것이다. 평가 대상이 문장 단위를 넘어 선 텍스트 단위이며, 단어나 문법 등의 형식적 측면만 검사하는 것이 아니라 내용적 측면도 평가한다. 또한, 평가를 위해 모범 답안을 준비해야 하는 것도 아니기 때문에 그 만큼 자동화 정도가 높다고 할 수 있다. 그렇지만 오류 유형별로 나타나는 문제성의 정도를 채점에 반영하지 못한 것은 앞으로 보완이 되어야 할 것이다. 문법적으로 잘못된 문장이라고 해서 모두 동일한 정도의 문제성을 가지고 있는 것은 아니기 때문이다.

참고문헌

김영택 외 (2001). 자연언어 처리, 생능출판사.

김지은 · 이공주 (2007). 중학생 영작문 실력 향상을 위한 자동 문법 채점 시스템 구축, 한국콘텐츠학회 논문지 7.5, 36-46.

시사영어사 (2006). TOEIC Writing Test 공식 문제집, 시사영어사.

이영식 (1998). 영어작문 평가의 채점 신뢰도에 대한 분석, 영어교육 53.1, 179-200.

지은림 (1996). many-facet Rasch 모형을 적용한 대입 논술고사 채점의 객관성 연구,

교육평가연구 9.2, 5-22.

진경애 (2007). 영작문 자동 채점 시스템 개발 연구, 영어어문교육13.1, 235-259.

Barzilay, R. and M. Elhadad (1997). Using Lexical Chains for Text Summarization, *Proceedings of the ACL Intelligent Scalable Text Summarization Workshop.*

Tom Brehony & Kevin Ryan (1994). Francophone Stylistic Grammar Checking (FSGC) Using Link Grammars, *Computer Assisted Language Learning* 7.3, 257-269.

Burstein, J., and M. Chodorow (1999). Automated Essay Scoring for Nonnative English Speakers, *Proceedings of the ACL99 Workshop on Computer-mediated Language Assessment and Evaluation of Natural Language Processing.*

Engelhard, G. (1992). The Measurement of Writing Ability with a Many-faceted Rasch Model, *Applied Measurement in Education* 5.3, 171-191.

Jarmasz, M. and S. Szpakowicz (2003). Not as Easy as It Seems: Automating the Construction of Lexical Chains Using Roget's Thesaurus, *Proceedings of the 16th Canadian Conference on Artificial Intelligence.*

Kakkonen, T. and E. Sutinen (2004). Automatic Assessment of the Content of Essays Based on Course Materials, *Proceedings of the 2nd International Conference on Information Technology: Research and Education.*

Landauer, T., D. Laham, and P. Foltz (2003). Automatic Scoring and Annotation of Essays with the Intelligent Essay Assessor, In Mark D. Shermis and Jill C. Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Lawrence Erlbaum.

Landis, J. and G. Koch (1977). The Measurement of Observer Agreement for Categorical Data, *Biometrics* 33.1, 159-174.

Londsale, D. and D. Strong-Krause (2003). Automatic Rating of ESL essays, *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing* Volume 2.

McCallum, A. (1996). Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, http://www.cs.cmu.edu/mccallum/bow.

McCallum, A. and K. Nigam (1998). A Comparison of Event Models for Naive Bayes Text Classification, *Proceedings of the AAAI-98 Workshop on Learning for Text Classification.*

Medelyan, O. (2007). Computing Lexical Chains with Graph Clustering, *Proceedings of the ACL 2007 Student Research Workshop*.

Miltsakaki, E. and K. Kukich (2004). Evaluation of Text Coherence for Electronic Essay Scoring Systems, *Natural Language Engineering* 10.1, 25-55.

Morris, J. and G. Hirst (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics* 17.1, 21-48.

Page, E. and N. Peterson (1995). The Computer Moves into Essay Grading: Updating the Ancient Test, *Phi Delta Kappa* 76, 561-565.

Park, J. C., M. Palmer, and G. Washburn (1997). An English Grammar Checker as a Writing Aid for Students of English as a Second Language, *Proceedings of the Conference of Applied Natural Language Processing* (ANLP).

Rose, C., A. Roque, D. Bhembe, K. Vanlehn (2003). A Hybrid Text Classification Approach for Analysis of Student Essays, *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing* Volume 2.

Sleator, D. and D. Temperley (1991). Parsing English with a Link Grammar, Carnegie Mellon University Computer Science Technical Report CMU-CS-91-196.

Sutcliffe, R. and McElligott (1994). Using the Link Parser of Sleator and Temperley to Analyse a Software Manual Corpus, In R. Sutcliffe, H. Koch and A. McElligott, eds., *Industrial Parsing of Software Manuals*. Rodopi B.V.

Zhou, G. and J. Su (2002). Named Entity Recognition Using an HMM-based Chunk Tagger, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

1차 원고 접수: 2008. 9. 4.
2차 원고 접수: 2008. 9. 18.
최종 게재 승인: 2008. 9. 20.

(*Abstract*)

An English Essay Scoring System Based on Grammaticality and Lexical Cohesion

Dong-Sung Kim[1] Sangchul Kim[2] Hee-Rahk Chae[1]

[1]Dept. of Linguistics and Cognitive Science, Hankuk University of Foreign Studies

[2]Dept. of Computer Science, Hankuk University of Foreign Studies

In this paper, we introduce an automatic system of scoring English essays. The system is comprised of three main components: a spelling checker, a grammar checker and a lexical cohesion checker. We have used such resources as WordNet, Link Grammar/Parser and Roget's thesaurus for these components. The usefulness of an automatic scoring system depends on its reliability. To measure reliability, we compared the results of automatic scoring with those of manual scoring, on the basis of the Kappa statistics and the Multi-facet Rasch Model. The statistical data obtained from the comparison showed that the scoring system is as reliable as professional human graders. This system deals with textual units rather than sentential units and checks not only formal properties of a text but also its contents.

Keywords : Grammaticality, Lexical Cohesion, Automatic Essay Scoring System, Kappa Statistics, Multi-facet Rasch Model, Parsing, Natural Language Processing

부 록

본 부록에서는 수동 채점시에 활용한 ETS의 작문 채점 가이드와 학생들의 영작문 채점 결과와 텍스트의 일부를 제시한다.

1. ETS 작문 채점 가이드

	점수	채점 가이드
평가 지침	5점	● 효과적으로 주제에 답했고 과제도 해냈다. ● 구성과 내용 전개가 좋다. ● 통일성과 일관성이 있다. ● 사소한 문법/어휘 오류는 있지만 고른 언어 구사력을 보여 준다.
	4점	● 몇몇 부분이 충분히 상세하지 않을 수 있지만 주제에 잘 답하고 과제도 잘 해냈다. ● 구성과 내용 전개가 대체로 좋다. ● 대체로 통일성과 일관성이 있다. ● 사소한 오류는 있지만 언어 사용은 능숙한 편이다.
	3점	● 주제에 답하고 과제를 수행하였다, ● 모호한 경우는 있으나 통일성, 연속성, 일관성을 보여 준다. ● 명확하지 못한 문장이 간혹 있다. ● 한정적이지만 올바른 어휘와 구문을 사용한다.
	2점	● 내용 전개가 부족하다. ● 구성 또는 생각이 부족하다. ● 부적절한 진술과 설명을 포함한다. ● 구조와 어휘에 오류가 많다.
	1점	● 결점으로 인해서 응답이 심각하게 훼손되었다. ● 상세한 내용이 거의 없다. ● 심각하고 빈번하게 문법 및 어휘 오류가 발견된다.
	0점	● 주제와 연관성이 없거나 아무런 의미가 없는 문자들의 조합이다.

2. 작문 채점 결과 및 해당 텍스트 일부

여기서 제시되는 견본은 작문 능력이 상, 중, 하인 학생의 영작문 1개씩이다. 지면상의 제약과 개인 정보임을 고려해서 텍스트의 일부분만 제시한다.

		견본1	견본2	견본3
수동채점 1	철자	5	3	3
	문법	4	3	2
	응집	4	3	2
수동채점 2	철자	4	3	3
	문법	4	3	2
	응집	4	3	2
자동채점	철자	5	5	5
	문법	4	4	2
	응집	5	3	2

<견본 1>

The cost of bilingual program is very expensive and the outcome of the program is not very promising. In addition if the program doesn't work out well for the children who are in the program, they might not be able to speak well in both languages they learned, which means they might not have a mother tongue. But I think it is a kind of process to move on a better stage. Plus, it's just a small piece of side effect in learning foreign languages. Through bilingual programs, children can acquire other languages and cultures more easily. And this program can at least give children a chance to be familiar with the languages. Children have the ability to learn more easily if the language acquisition starts early enough. I learned that if children who are younger than four years old start bilingual education, they can catch the language information more readily and can eventually become bilinguals.

<견본 2>

This building is 5-stored building that has great view of myoung su lake so many students take picture here. Next, there is linguistic building that is commonly used by langustic students. This biggest building is where you could take course of different languages. In the linguistic buildings, there is cafe that you only speak English, called English-Zone, and information center. Also there is food store and cafeteria known for its tastes so it is always crowded with students. There is terrace back of the cafeteria where students could cheat-chat to one another.

<견본 3>

How much one can revitalize the nation's economy is important, of course, but there are another equally important considerations, neglect of which may produce frustration in later years. If I would become a hero who do one thing as my pleases, I would have a immoral candidate defeat. And I will help a honest politician to win a presidential election. It will make our society just and fair. The ability about economy is needed as a president of one nation. But, the ability about economy or specialty matters less than do honesty. How much one can revitalize the nation's economy is important, of course, but there are another equally important considerations, neglect of which may produce frustration in later years.