추상적 의미 표상을 활용한 사진 자막 영작문 평가*

김동성
(이화여대)

Kim, Dong-Sung. (2016). English Caption Writing Assessment Using Abstract Meaning Representation. *The Linguistic Association of Korea Journal, 24*(4), 235-260. Since story-telling has been used in evaluating the development of language skills, English language proficiency test such as TOEIC includes a caption writing test. This paper investigates how linguistically motivated features are used for automatically scoring a picture-description writing test. Specifically, we design to build scoring models with features under the principles of relevancy, appropriateness, and task-detailed description. For the experiment, we gather the caption writing corpus upon several images. We statistically compare different performances among 9 statistical assessment factors, revealing that Abstract Meaning Representation (AMR) produces the best results in predicting human raters' scores. AMR shows the best performance in capturing the similar logico-semantic structure(s) among various sentential forms.

주제어(Key Words): 기계학습(Machine Learning), 자동 작문 채점(Automatic Writing Assessment), 자연언어처리(Natural Language Processing), 컴퓨터 언어 보조 학습(Computer-Assisted Language Learning)

1. 머리말

스토리텔링(story-telling)은 언어 능력 발전을 평가하는데 활용되는 것으로 알려져 있으며(Botvin and Sutton-Smith, 1977; McKeough and Malcolm, 2011; Sun and Nippold, 2012), TOEFL Junior Comprehensive Test나 TOEIC Writing Test에서 영

* 이 논문 또는 저서는 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2014S1A5A2A01011245). 논문에 대한 세세한 지적을 하신 세 분의 심사자들의 노고에 감사를 전한다. 논문의 모든 오류는 필자의 몫이다.

어 능력 평가 방안으로 활용되고 있다(Somasundaran et al. 2015). 평가에서 특정 상황이 묘사되어 있는 한 장이나 일련의 사진들을 대상으로 상황을 설명하는 자막 작문이 주어진다. 이를 통해서 사진 속 상황을 적절히 묘사하는 능력과 더불어 문법, 단어 선택까지 평가한다.

본 논문의 목표는 특정 사진에 대한 영작문 평가 요소로 어떠한 것이 적절한지에 대한 연구로서 자동 채점에서 활용되기 위한 가장 적절한 평가 요소를 찾아내는 것을 목표로 한다. 이를 위해서 특정 사진들에 대한 자막을 대량의 코퍼스 형태로 수집하고 정의된 평가 요소 중 어느 요소와 요소들의 결합이 가장 적합한지 밝히기 위해서 통계적인 방식을 활용한다. 간략하게 연구 내용을 설명하면 다음과 같다. 연구에서는 정해진 사진들에 대해 실험 참가자들에게 적합한 영어 자막을 기술하게 해서 대량의 문장을 수집했다. 이를 토대로 수동으로 모든 문장을 ETS TOEIC 자막 작문 채점 기준에 따라 채점했다. 채점된 점수와 미리 구성된 점수 대표 표본 문장을 대상으로 어떠한 평가 요소가 적절한지 비교하고, 특정 기계 학습 장치를 활용해서 어떠한 요소들이 채점을 가장 적절하게 예측하는지 살펴보았다.

Brew와 Leacock (2013)은 단문으로 구성된 작문은 긴 에세이 평가 방식과 다르며, 문법적 요소와 더불어 의미적 표현에 대한 분석이 더 필요하다고 주장했다. 자막의 경우에도 특정한 사진 속 상황을 묘사하기 위한 특정 표현이 있기 때문에 학습자의 자막은 이 표현과 관련성이 있고, 의미적으로 유사해야 한다. 따라서 문법적 구조와 더불어 특정 의미적 표현이 채점에 포함되어야 한다. 이러한 문법과 의미의 복합적 평가를 위한 방식으로 본 연구는 추상적 의미 표상(AMR: Abstract Meaning Representation)을 활용한다.

연구에서는 추상적 의미 표상과 더불어 9개의 복합적인 요소들을 포함시켜서 수동 채점과 비교했다. 개별 요소들이 수동 채점을 예측하는데 얼마나 정확하게 활용되는지를 통계적으로 살펴보고, 채점 요소들을 결합해서 얼마나 정확하게 채점 대상을 분류하는지 기계학습을 통해서 살펴보았다. 통계적 방식은 상관성 또는 선형성을 살펴보는 방식으로 개별 요소들 중 가장 적절하게 채점을 설명하는 요인을 찾고, 기계학습 방식은 어떠한 요인들로 채점을 구성해야 적합한지를 찾기 위해서 적용됐다.

이 연구는 다음 세 가지 점에서 전산적 언어 처리를 활용한 언어 교육 분야인 컴퓨터를 활용한 언어 보조 학습 및 언어 교육 분야에서 활용될 수 있다. 첫째로, 현재 여러 자연어처리 기술을 활용한 자동 작문 채점이 널리 활용되고 있는데(Brew & Leacock, 2013), 자동 채점의 요소로 어떠한 요인들이 활용되어야 하는지에 대한 방향성이 제시될 것이다. 둘째로 실제 채점에서 어떠한 점이 중요한 요인으로서 채점 기준(scoring rubrics)을 설명하는데 활용될 것이다.

2. 사진 자막 채점에 대한 기존 연구

2.1. 채점과 연관된 연구

다음 그림 1을 기술하는 자막과 채점을 고려해 보자.

그림 1.

그림 1을 가장 적절히 기술하기 위해서는 {bear, stand, iceberg}등의 필수적인 어휘가 요구되고 "A bear stands on an iceberg."나 "The bear is standing on an iceberg."와 같은 문장이 적합한 기술들로 판단된다. 그러나 이미지와 연관된 자막은 일대일로 대응하는 것이 아니라, 일대다로 대응되므로 기술방법에 따라 전혀 다른 어구, 어휘, 구조를 선택할 수 있다. 그림 1에 대해서 (1a-j)의 다양한 기술들이 연구에서 자막 피실험자 실험에서 관찰되었다.

(1) a. A polar bear is standing on an iceberg.
b. A polar bear is on the ice.
c. The polar bear is standing on an ice cap.
d. A white bear is looking at me.
e. A white bear stands on the ice.
f. There is a polar bear on the ice.
g. A bear is standing on an ice cap.
h. It is a polar bear.
i. A polar bear is standing on top of an iceberg.
j. This is a polar bear.

(1a-j) 문장들을 채점해야 한다면, 이 문장들 중 어느 문장이 이미지를 가장 적절히 기술하고 있는지 판별해야 한다. 다시 말하면, 채점 척도를 1, 2, 3점으로 설정한다면, 개별 척도 점수에 가장 적합한 문장들을 선별해야 한다. 만약 "A bear stands on an iceberg."나 "The bear is standing on an iceberg."와 같은 문장이 적합한 기술들이라면, 이와 비슷한 문장들을 3점으로 판단한다. 이러한 작업을 판별하면서도 어휘에 대한 해석으로 ice cap이 iceberg와 같은 것인지 다른 것인지 판단해야 하며, "A polar bear"나 "The polar bear"가 같은 어구로 판단되는지 판단해야 한다. 또한 문법적으로 "There is"나 "This is"가 서로 다른 것인지도 구별해야 한다. 채점에서 가장 중요한 요소는 어떠한 문장이 개별 척도와 가장 적합하며, 이 문장의 의미-논리 구조와 가장 적합한 문장은 어떠한 것인지 구분해야 한다.

이러한 연유로 하나의 문장과 같은 문장인지 다른 문장인지 판별하는 것과 연관된 여러 연구들을 참조할 수 있다. 일반적인 문장의 유사도는 기준이 되는 문장과 비교가 되는 문장들 사이에 단어의 배열이 어떤지를 측정하게 된다. 따라서 같은 단어가 같은 순서로 사용되는 경우가 가장 유사성이 높은 경우가 된다. Li et al. (2004)는 특정 단어의 배열 순서를 기본으로 문장 간의 유사도를 측정해서 특정한 주어-동사-목적어 패턴을 검색하도록 했다. Achananuparp et al. (2004)는 TF-IDF (Term Frequency-Inversed Document Frequency)를 적용한 유사도 공식에 의해 측정했다. 이 연구에서 문장 간 단어에 기초한 유사도는 매우 높은 수준이지만, 단락과 같이 더 많은 문장들 간에 유사도는 낮은 수준이며, 문장 간 논리적 포함관계와 같은 의미적 표현은 더 낮은 수준이 된다. Liu와 Wang (2008)은 WordNet에서 관측되는 단어 간의 유사도를 벡터화해서 문장 간의 유사도를 측정했다. 이 방식은 문장 내에 위치한 모든 단어들을 하나의 벡터로만 간주했기 때문에 관련된 단어들의 구조적 문제를 고려하지 않았다. 따라서 구조화된 해석이 불가능하기 때문에 엄밀하게 측정된 의미 논리 구조를 측정한 것은 아니다.

채점과 연관된 연구는 ETS에서 많이 이루어졌다. ETS의 작문 평가는 기본적으로 훈련된 전문가 집단의 수동 채점으로 이루어지나, 비용이나 유지 등 여러 이유로 오랫동안 자동채점 시스템이 연구 및 고안되었다. E-rater라고 명명된 이 시스템은 인간 채점으로 축적된 빅 데이터를 기반으로 에세이 형식의 작문을 채점한다(Attali & Burstein, 2006). 이 시스템에서 활용하고 있는 자질들은 문법, 용례, 스타일, 구성, 논리구조, 단어 길이, 어휘 등등이다. 자체 평가로 볼 때 인간 채점 빅데이터와 상관성이 0.97에 가까울 정도로 정교한 채점 도구로 알려져 있다.

Madnani et al. (2013)은 긴 문장을 줄여 쓰는 짧은 작문의 채점과 연관된 ETS의 다른 문제 유형에 기초한 연구를 발표했다. 이 연구에서 본 연구와 유사하게 5점 척도 채점 기준을 제시했고, 짧은 문장을 작성하고 이를 채점하는 내용을 평가했다. 연구에서는 인간 채점과 기계적 채점 요소와 상관성도 측정했는데, 기계적 채점 요소는 n-gram 기반 기계번역 유사

도 측정 방식인 BLEU (Papineni et al., 2002), 재현율 기반의 자동 축약 평가 방식인 ROUGE (Lin & Hovy, 2003), 원래 문장에서 복사된 3 단어들의 연쇄 개수, 복사된 단락의 수치화된 점수, 첫 번째 문장, 문장 길이, 특정 담화 표지 단어들의 사용 빈도 등이 고려되었다.

Madani et al. (2013)의 연구에서 특이한 점은 특정 채점 평가 방식인 BLEU 가 다른 평가 요소들에 비해서 정보획득율이 로지스틱 회귀분석에서 높게 나타난다는 것이다. 따라서 로지스틱 분석의 특성에 따라 분류 기준으로 적합하다는 의미를 띤다. 연구에서는 n-gram 자질들과 같은 통계적 구성이 어떠한 연관성이 있는지를 살펴보았다. 또한 BLUE도 어떠한 연관성이 있는지도 살펴보았다.

Papineni et al. (2002)에서 제시한 BLEU는 원래 문장과 자동 번역된 문장간에 서로 얼마나 많은 단어가 일치하는가의 비율이다. 후보가 되는 전체 집단의 n-gram의 크기로 정규화하는 보상 처리를 한다. 이 연산 방식은 기본 문장에서 사용된 단어가 번역된 문장에서 얼마나 많이 사용되었는지 채점하는 방식이다. 본 연구에서는 채점의 기준이 되는 문장을 선정해서 이 문장과 후보 문장이 얼마나 유사한지를 보는 방식을 활용했는데, 채점 요소로 추상적 의미 표상이외에 BLEU를 도입해서 비교했다[1].

ETS에서 하나의 단락이나 몇 개의 문장을 채점하기 위해서 만든 것이 c-rater라는 시스템이다(Sukkarieh & Blackmore, 2009). 이 시스템은 본 연구에서 목표로 하는 문장간 유사도 측정과 가장 유사한 시스템이다. 본 연구가 추상적 의미 표상이라는 시스템을 활용한 의미적 논리 구조를 측정하는 도구인 smatch를 활용하는 반면에 이 시스템은 3단계로 구분된 처리 시스템이 자연어 처리 시스템과 언어 자원을 활용한 채점 방식을 활용한다.

C-rater의 채점 흐름을 자세히 설명하면 다음과 같다. 입력된 기준이 되는 문장이나 단락을 중심으로 기준이 되는 채점 문장을 구성한다. 채점 문장은 하나의 문장이 아니라 여러 개의 유사한 문장으로 구성하며, 이 문장들을 중심으로 채점 모델을 만들게 된다. 채점 모델과 채점 문장이 입력되면 채점 알고리즘이 채점 문장을 채점하게 되는데, 알고리즘의 기본 흐름은 다섯 가지 큰 영역의 자질을 원형이 되는 문장과 비교하게 된다. 다섯 가지 영역에 포함되는 것은 필수 단어 여부, 논항 개수, 감정영역[2], 동사와 형용사의 오류 관계에 대한 점검이다. 또한 문장들간이나 문제와 답안간에 논리 추론 관계도 채점한다.

기본적인 방법론은 본 연구와 유사성이 많다. 에세이 형식이 아닌 짧은 형식의 작문 채점

---

1) 작문 자동 채점의 경우에 중요한 요소는 글의 논리적 요소이다. E-rator의 경우에 상대적인 글의 구성이나 글의 논지의 발전은 글의 채점과 가장 높은 상관성을 보인다(Attali & Burstein 2006). 글의 구성은 서론, 본론, 결론과 같은 글의 논지를 구성을 말하고 글의 논지의 발전은 삼단논법의 전개에서 전제, 발전적 설명, 개별 결론화 같이 단락의 구성을 말한다.

2) 감정 영역은 찬성 또는 반대하는 지에 대한 채점이다. 즉 문제가 찬성을 선택하는 논조인데, 반대를 서술하면 감점한다던지 반대의 경우에는 가점한다던지 하는 방식이다.

에 있어서 채점에 기준이나 원형이 되는 문장들을 선택하고 이 문장들을 대상으로 채점 문장과 얼마나 유사성이 높은가를 평가한다. 평가의 기준은 의미적 논리 구조가 필수적인데, 논리 구조를 평가하기 위해서 논항 구조 및 기타 관련된 평가 정보를 포함한다. 그러나 c-rater는 기본 구조인 논항 구조인지 여부만을 판별할 뿐, 논항 구조가 언어적으로 어떻게 실현되었는지를 평가하지는 않는다. 반면에 smatch를 활용한 방식은 논항 구조의 연쇄를 수치화해서 연산하기 때문에 논항 구조가 어떻게 구성되었는지를 채점한다. 따라서 본 연구는 c-rater와 다르게 의미적 논리구조를 포함한 연산이 가능하다. 특히 논항간의 관계성을 채점에 포함하는 것은 c-rater구조로는 불가능하기 때문에, 추상적 의미 표상을 활용한 것이 더 향상된 방식이다. 연구에서는 c-rater가 고려한 여러 평가 요소를 반영해서 smatch를 활용한 방식들과 비교했다. 총 단어수, 총 논항숫자, 생략된 논항, 생략된 단어등은 이러한 정보를 반영하기 위해서 고려한 요소들이다.

2.2. 추상적 의미 표상

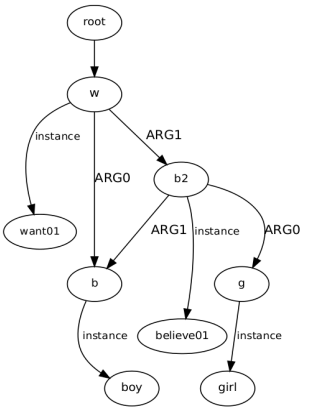
후기 데이비슨 방식은 개체, 사건, 속성, 상태에 대한 변수를 제공하므로, "b/boy"는 boy에 대한 개념을 제공한다. 이러한 방식으로 "(p/play02 :location (p/park))"는 "play in the park"의 개념을 나타내고, 다시 boy의 개념이 합쳐져서 "The boy plays in the park" 문장의 표상은 "(p/play02 :ARG0 (b/boy) :locate (p/park))"이 된다[3].

(2a)의 문장은 (2b)와 같이 두 개의 사건이 표상되는데, FrameNet과 PropBank에서 기술된 바와 같이 want의 사건은 ARG0을 나타내는 wanter와 wanted thing을 나타내는 ARG1, believe01의 사건은 believer가 되는 AGR0과 believed thing이 되는 ARG1로 구성된다. 여기에서 boy의 개념인 "b/boy"는 want의 사건에 포함되지만 believe의 사건으로 재진입된다. (2b)의 추상적 의미 표상은 시작점과 방향성이 있는 비대칭적 그래프 그림 2로 표현된다(Banarescu et al. 2013).

(2) a. The boy wants the girl to believe him.
b. (w/want01
:ARG0 (b/boy)
:ARG1 (b2/believe01
:ARG0 (g/girl)
:ARG1 b))

3) 2절의 기술과 예제는 Banarescu et al. (2013)을 참조하였다. 추상적 의미 표상의 기본적 구조는 Matthiessen와 Bateman (1991)의 PENMAN 기본 정신에서 시작되었다.

그림 2. 추상적 의미 표상 그래프



(2b)의 추상적 의미 표상은 (3)과 같은 3중항(triple)의 접속이다.

(3) instance(w, want01) ^ instance(b, boy) ^ instance(b2, believe01) ^ instance(g, girl) ^ ARG0(w, b) ^ ARG1(w, b2) ^ ARG0(b2, g) ^ ARG(b2, b)

또한, (2b)의 추상적 의미 표상은 (4)의 다양한 영어 표현을 동일하게 취급한다. 이러한 점에서 자막 영작문 채점 요소로 활용이 가능하다. 하나의 이미지에 대한 언어적 표현인 자막은 여러 개가 가능하므로, 각 문장들의 언어적 차이점이 정확한 채점이 된다. (4a-e)에 기술된 언어적 표현이 동일한 의미로 간주된다면, 동일한 채점이 되게 된다.

(4) a. The boy wants the girl to believe him.
b. The boy wants to be believed by the girl.
c. The boy has a desire to be believed by the girl.
d. The boy's desire is for the girl to believe him.
e. The boy is desirous of the girl believing him.

일반적으로 FrameNet (Baker et al. 1998), PropBank (Palmer et al. 2005)과 같은 언어 자원에 수록된 동사 위주의 프레임 집합을 활용하기도 하지만, 서술명사 위주의 논항구조에 기술한 NomBank (Meyers et al. 2004)를 활용한다. 프레임에 정의된 논항 구조를

일반적 통사 구조에서 추출한다. 정의되는 의미적 관계는 한정적으로 beneficiary, condition, direction, mode, name, part, path, time 등등과 같이 일정한 관계로 정의되는 경우로 제한하며, 수량적 관계, 시간이나 목록 등을 정의된 바에 따라 논항들을 나타낸다. 또한 동지시 관계도 나타내며 문장 요소들이 도치된 경우도 도치되지 않은 경우와 동일하게 취급한다. 'must, have to, should, can, shall'등과 같은 법조동사나 부정의 의미도 포착하며, 의문문도 arg-unknown 논항으로 기술한다. (5)는 부정어, (6)는 cannot, (7)은 의문사 where와 연관된 것이다.

(5) a. The boy did not go.
b. (g/go01
:arg0 (b/boy)
:polarity -))
(6) a. The boy cannot go.
b. (p/possible
:domain (g/go01
:arg0 (b/boy))
:polarity))
(7) a. Where did the girl find the boy?
b. (f/find01
:arg0 (g/girl)
:arg1 (b/boy)
:location (a/amr-unknown))

명사의 경우에는 PropBank의 동사 프레임 집합을 참조하거나, NomBank의 프레임 집합도 참조한다. (8a)의 동사 프레임 집합이 (8b-c)의 명사구의 프레임 집합과 동일하게 (9)와 같이 취급된다.

(8) a. What the girl opined
b. The girl's opinion
c. The opinion of the girl
(9) (t/thing
:arg1-of (o/opine01
:arg0 (a/girl)))

형용사는 PropBank에 정의된 프레임 집합을 그대로 적용하거나, tough 구문과 같이 특수한 구문 유형도 논리적으로 처리하게 한다.

(10) a. It is tough to please girls.
b. (t/tough
:domain (p/please01
arg1 (g/girl)))

전치사는 일정한 의미 관계로 포착이 된다. (11a)의 "in June"은 time이라는 의미적 관계로 기술된다.

(11) a. The nation defaulted in June.
b. (d/default01
:arg1 (n/nation)
:time (d2/date-entity
:month 6))

개체명은 특정한 이름으로 인식된다.

(12) a. Mollie Brown
b. (p/person
:name (n/name
:op1 "Mollie")
:op2 "Brown"))

이외에도 분사 구문이나 2어 동사, 경동사, 계사와 같은 특수한 동사 구조를 표현할 수 있는 논리구조를 포함한다. Banarescu et al. (2013)에 따르면 추상적 의미 표상은 굴절 형태를 표현하지 못하므로, 시제나 숫자, 관사 등의 문법적 표현이 필요가 없다. 이러한 방침은 일반 자연어 문장을 주석 처리하는 속도를 증가시킨다고 주장한다. 그러나 전반적으로 세밀하고 엄밀한 표현보다 자연어의 적절한 표현에 더 치중한다. 또한 특정한 언어 자원에 집중하기 때문에, FrameNet이나 PropBank에 수록된 프레임 집합을 그대로 반영하는 문제도 발생한다. 특정 실제 사건과 가상적 미래 사건에 대한 구분이 모호하다. 예를 들어서 "John wants to go."에서 want01사건과 go01사건은 동일하게 취급되기 때문에 go01사건이 미래 사건으로 발생하거나 발생하지 못하는 상황을 감지하기 못하게 된다. 또한 일반 수량사에

정확한 표현력이 부족해서, 수량적 표현을 엄밀하게 처리하지 못한다. 그러나 일반적인 자연어 문장의 논리를 이해하고, 전산적 처리 자원으로 변환하기에 적합한 수준으로 표현하는데 추상적 의미 표상이 적절하다.

추상적 의미 표상은 Treebank, FrameNet, PropBank, NomBank, WordNet 등의 여러 언어 자원을 활용한 여러 전산적 도구가 개발되어 공개되었는데, Flangian et al. (2014)은 자연언어처리 도구인 통사분석기, 품사부착기, 개체명인식기, 의미 중의성 해소 장치등을 활용해서 자동으로 문장 의미 논리 구조를 추상적 의미 표상으로 분석하는 JAMR을 제시하였고, Cai와 Knight (2013)은 서로 다른 문장의 추상적 의미 표상간의 유사도를 측정하는 smatch를 제시하였다. 연구에서는 smatch를 활용해서 채점 척도의 원형이 되는 문장과 채점 대상이 되는 문장간의 유사도를 측정하고 이를 채점에 활용한다. 관련된 내용을 간략히 설명하면 다음과 같다.

논항구조 언어자원 정보를 TreeBank의 파싱된 정보와 연동해서 파싱 결과를 언어자원 논항정보로 변환한다. 다시 말하면, 이전 시스템들은 파싱 결과와 논항구조를 별도로 취급해서, 서로 다른 차원의 정보로 활용했다면 추상 의미 표상과 연관된 시스템들은 파싱된 정보로부터 논항구조 정보를 포함시켜서 문장의 논리 의미 구조를 직접적으로 추출한다. Chiang et al. (2013)은 파싱 알고리즘인 CKY와 그래프 알고리즘을 직접적으로 활용한다. 또한 Flangian et al. (2014)는 파싱된 결과에서 논항구조를 적용한 그래프 유형의 데이터를 라그랑지안 완화(Lagrangian Relaxation) 기법을 적용한다.

Cai와 Knight (2013)은 추상적 의미 표상들로 분석된 논리 구조 간의 유사성을 측정하는 도구인 smatch를 제안했다. 유사도 측정 원리는 다음과 같다. 추상적 의미 표상은 문장의 논리 의미 구조를 (13)과 같이 3중항의 접속으로 처리한다(Cai & Knigt, 2013; p. 749).

(13) instance(x, like01) ^ instance(y, boy) ^ instance(z, girl) ^
ARG0(x, y) ^ ARG1(x, z)

또한 “The boy likes soccer”는 (14)에 제시된 3중항들의 접속이다.

(14) instance(a, like01) ^ instance(b, boy) ^ instance(c, soccer) ^
ARG0(a, b) ^ ARG1(a, c)

각각이 개체집합 {x, y, z}는 {a, b, c}에 대해서 6가지의 서로 다른 동일한 관계를 만들어 내고, (15)와 같이 각각의 일치(M), 정확률(P), 재현률(R)과 F-점수(F)로 측정된다.

(15)			M	P	R	F
x=a,	y=b,	z=c	4	4/5	4/6	0.73
x=a,	y=c,	z=b	1	1/5	1/6	0.18
x=b,	y=a,	z=c	4	0/5	0/6	0.00
x=b,	y=c,	z=a	4	0/5	0/6	0.00
x=c,	y=a,	z=b	4	0/5	0/6	0.00
x=c,	y=b,	z=a	2	2/5	2/6	0.36

여기서 M은 정확히 일치한 3중항 명제의 값으로 총 6개의 3중항 중 몇 개가 일치하는가의 값이다. P는 두 번째 문장 3중항 구조 중 첫 번째 문장 구조에 비해서 어느 정도가 일치하는 것의 비율이다. R은 전체 중 어느 정도가 일치하는지에 대한 비율인 재현율이 된다. 그리고 정확률과 재현율의 조화평균이 F가 된다. smatch는 F-점수를 토대로 두 개나 그 이상의 문장들의 추상 의미 표상들 간의 유사도를 측정한다. 예에서 총 4개의 3중항이 일치하므로 smatch 점수는 0.73이 된다.

추상적 의미 표상의 정신은 일반적으로 주석 및 사용이 용이하고, 계산적으로 그래프상에서 이동 연산이 간편한 점에 초점을 맞추었다. 이러한 연유로 매우 상세한 의미 분석보다는 논리를 구조적으로 분석, 표현, 이해, 처리가 비교적 용이한 점을 목표로 했다. 또한 전산적으로 여러 언어 자원들을 활용하고, 자연어 처리 장치들을 적극적으로 활용하는데 그 목표가 있다. 그러므로 양화 현상이나 2차 술어 논리와 같은 복잡한 논리 표현에는 적합하지 않지만, 파서(paser), 태거(tagger), 청거(chunker) 등 여러 언어 처리 장치와 적절히 결합이 가능하도록 구성하였다.

3. 실험

실험은 피실험자 집단에게 사진을 제시하고 설명하는 영어 자막을 작성하게 해서, 순위를 매기고 측정치와 비교해 보았다. 우선 실험 순서를 설명하면 다음과 같다. 먼저, 피실험자 집단을 50명으로 선정하고[4], 그림 1, 3, 4의 간단한 사진 5개를 제시했다.[5]

4) 피실험자 집단은 대학교에 재학중인 대학생1~4학년으로 영어를 12년 이상 학습한 경험을 갖고 있다. 학습자의 프로필이 영어 능력과 연관된 실험에서 중요한 요소로 작동하기 때문에 실험자들의 영어 능력과 연관된 적절한 선출 조건을 제시하여야 했다. 이런 연유로 적당한 조건을 제시해야 했으며, 초등학교 6년, 중고등학교 6년 도합 12년 동안 정규교육과 사교육을 포함해서 영어 학습을 시작한 경우가 있는지 질문을 했고, 모든 실험 참가자의 조건이 12년 이상 학습했다는 답변을 받았다. 이 점은 한 심사자의 지적으로 부연된 설명이다.

그림 3. 그림1(좌), 그림2(우)

그림 4. 그림3(좌), 그림4(우)

사진의 이미지는 추상적 의미 표상이 가진 제한이 포함될 수 있는 경우가 가능한 경우를 제외했다. 2절에서 논의한 바와 같이 양화사를 표현하지 못하기 때문에, 양화사가 포함되지 않고 기술될 경우만을 고려했다. 또한 복합적인 전치사가 필요 없을 경우로 간단한 단문으로 자막이 기술될 것 같은 경우만을 고려했다. 동사는 2어 동사나 분사가 가능하지 않고, 의문사, 상적 표현이 필요 없는 경우만을 고려했다. 형용사는 크게 서술적 경우와 수식하는 경우로 구분되는데, 서술하는 경우는 논항이 필요한 경우이므로 수식의 경우만을 고려했다. 그 외에 사람 이름, 장소명, 단체명 등과 같은 개체명은 포함되지 않는 일반적 경우만을 고려했다. 또한 be동사로 간단히 서술되는 것도 피하려고 했지만,6) 피실험자의 자막 작문에 포함된 경우도 있었다.7)

---

5) 사진의 경우에는 저작권법의 문제가 있으므로, 저작권이 해결된 경우만을 고려했다. 연구에서는 인터넷 상에서 CCL (Creative Common License)나 PD (Public Domain) 의 태그가 가능한 사진들 경우만을 고려했다.

6) 예를 들어서 4번 그림의 경우에 "The umbrella is yellow."와 같은 문장도 포함되어 있다.

7) 실험의 특성을 고려해서, 단문으로 구성될 수 있는 경우로 and, but과 같은 복문이나 관계절이 불필요한 경우를 고려했다. 이러한 고려들은 본 실험전에 10명 규모의 간단한 규모의 선실험을 미리 실시했는데, 복

피실험자 집단에게 채점 기준을 설명하고, 각 채점이 1, 2, 3점 척도에 맞는 자막을 5개씩 작성하게 했다. 총 3,500개의 문장이 수집되었다. 수집된 문장 중 중복이 되는 경우를 제외하고 1,000 여개의 문장을 고려했다. 사진이 이해하기 쉽기 때문에 기술된 내용들을 비교해보면 거의 유사했다. 이를 다시 분류해서 1, 2, 3점 척도에 적합한 문장들로 분류하고[8], 이 문장들 중 채점 척도에 원형이 될 문형을 10개 이하로 추출했다.

원형이 될 문형의 경우에 2.2절에 기술한 추상적 의미 표상의 특성에 따라 각각 서로 다른 추상적 의미 표상으로 도출될 경우에만 원형으로 선택된다. 구체적으로 다음과 같은 형식들이 동일한 것으로 취급된다.

(16) a. 명사구에 정관사나 포함되어서 서로 다른 문장인 경우는 같은 구조가 된다(The book, A book, The books, Books).
b. 명사구로 구현된 논항 구조가 동사구로 구현된 논항 구조와 동일한 구조가 된다(The destruction of Rome by Huns = Huns destructed Rome).
c. 동일 논항 구조를 갖지만, 문형이 다른 경우는 같은 것으로 취급했다(도치, 수동태 등과 같은 경우).
d. 여격 교체가 있는 경우도 논항 설정이 동일하므로 같은 경우로 간주했다(John gave Mary a book = John gave a book to Mary).
e. There 구문은 아닌 구문과 추상적 의미 표상 구조가 다르며, 서로 다른 구조로 취급한다(There is a book on the table. ≠ A books is on the table.).

비문법적인 경우는 고려의 대상이 되지 않도록 했지만, 짧은 단문으로 구성된 경우이어서 적은 범위의 문법 오류가 발견됐다. 영어의 여러 문법적 오류 중 대부분의 오류가 관사가 생략된 오류가 많았다. 그러나 추상적 의미 표상 구조는 관사를 고려하지 않으므로 실질적 고려의 대상이 되지는 않았다. 수일치 오류가 there 구문에서 많이 발견되었는데(*There are a book on the table.), 이러한 경우는 대상에서 제외시켰다. 또한, 개체명이 필요하지 않은 사진을 제시했는데 피실험자가 개체명을 포함한 경우는 대상에서 제외시켰다(The cat

---

잡한 사진의 경우에는 기술되는 문장의 편차가 너무 심하게 나타났다. 이것은 실험 참가자들이 그림에 대한 이해가 각각 다르기 때문이기 때문에, 직관적으로 이해가 되는 사진을 중심으로 실행하게 되었다.

8) 작성된 자막의 채점 기준은 ETS Toeic writing test 채점 기준을 참조했다. 3점: 문법적 오류가 없고, 하나의 문장으로 구성된다. 사진과 연관성이 있다. 2점: 약간의 문법적 오류가 있고, 하나의 문장이 아닌 두 개 이상의 문장으로 구성된다. 사진과 연관성이 있다. 1점: 문법적 오류가 있어서 의미 해석이 어렵고, 사진과 연관성이 없다. (민선식 2008; p. 26)

named Leo is sitting on the couch.).

대부분의 문장이 단문으로 구성되기 때문에 비문법성은 거의 발견되지 않았다. 1, 2, 3 척도로 구분한 경우에, 3점은 구조적으로 완전하며 상황을 가장 잘 표현한 경우로 설정했다. 2점은 사진 속 상황을 적절히 기술하기는 했어도 3점보다 덜 적절히 기술한 것들로 설정했다. 1점은 성의 없이 써진 것들로 일부 상황만 묘사한 경우이다. 예를 들어서 그림 5의 경우에 (17a-c)는 3점으로 가장 상황을 적절히 표현한 것이고, (17d-e)는 2점으로 상황이 덜 적절히 기술된 것이다. (17f-k)는 상황을 기술하기 보다는 일부만 묘사된 경우로 1점 척도에 해당한다.

(17) a. A polar bear is standing on an iceberg.
b. There is a polar bear standing on an iceberg.
c. The polar bear is standing on an ice cap.
d. A white bear is staring at me.
e. The polar bear looks at me.
f. There is a polar bear.
g. It's a polar bear.
k. The bear is very big.

각 척도의 원형이 되는 문장을 여러 개를 설정한 이유는 하나의 사진에 대한 기술이 하나씩 일대일로 대응하는 것이 아니라 여러 개의 묘사가 가능하기 때문이다.

실제 피실험자의 작문들은 원형 문장과 유사하거나 약간 다른 많은 예들이 발견된다. (17a) 원형 문장을 기준으로 했을 때 (18a)와 같이 시제가 다르거나, (18b, d)와 같이 수식 형용사가 사용되거나, (18c)와 같이 주어 명사가 다르게 사용되는 등 여러 가지로 변이된 형도 발견이 된다.

(18) a. A polar bear stands on an iceberg.
b. A white polar bear is standing on an iceberg.
c. A bear is standing on an iceberg.
d. The big polar bear is standing on an iceberg.

4. 결과 및 평가

연구에서 고려한 평가 요소들은 다음의 세 가지 원칙에 의해서 선출되었다. 첫 번째는 어떠한 채점 요소들이 작문이 사진과 얼마나 연관성이 높은지를 평가할 수 있는가 이다. 두 번째는 어떠한 채점 요소들이 작성된 문장들에서 가장 유사성이 높은 문장들을 구분할 수 있는가 이다. 다시 말하면 인간이 작성한 답안이 표본이 되는 문장과 비교했을 때 정확하게 동일한 단어나 문구, 문법이 사용되지 않았더라도, 얼마나 유사성이 높은지를 발견할 수 있는 요소들인가 이다. 마지막으로는 서술이 얼마나 세밀하게 사진을 설명하는지에 대한 요소들인지 이다.

연관성은 표본 문장과 자막 작문이 연관성이 있는가를 채점하기 위한 요소로 기본 통계와 더불어 어휘적, 구조적 요인을 포함시켰다. 총 단어 수, 총 논항숫자, 생략된 논항, 생략 단어들, 문장에서 활용된 문맥 자유 규칙 등이 이 항목을 위해서 포함되었다. 유사성은 통계적 유사성과 더불어 구조적 유사성을 살펴보았다. 통계적 유사성으로 google 1T n-gram, BLEU 평가 수치가 이 항목에 포함된다. 세밀한 묘사를 위해서 의미적 문제와 문법을 포함한 논리적 구조를 포함 시켰다. 의존 관계와 smatch가 세밀한 묘사를 평가하기 위해서 포함시켰다.

원형이 되는 문장들 각각과 채점할 문장들의 smatch의 점수로 일대일로 측정한다. 다시 말하면, 채점할 문장 하나당 모든 원형 문장과의 유사도를 smatch로 측정하고, 가장 점수가 높은 경우에 원형 문장 척도로 채점을 매긴다. 예를 들어서 자세히 설명하면 다음과 같다. 3번 그림의 경우에 (19a), (19b), (19c)가 각각 3, 2, 1점 척도이고, "Someone is pouring a glass of milk."라는 문장이 채점이 대상이고, 각각의 smatch 점수가 0.57, 0.33, 0.29라면 가장 높은 smatch 점수가 매겨진 원형 문장의 척도 점수인 3점을 채택한다. "Someone is pouring a glass of milk."의 수동 채점은 3점 척도로 smatch로 판정한 값과 일치한다.

(19) a. Milk is being poured into a glass.
b. There's a cup of milk.
c. This is fresh milk.

Smatch에 의한 자동 채점결과와 수동 채점의 결과를 상관성 분석으로 비교했다. 그림 1에서 그림 5까지 평균값은 0.637로 뚜렷한 상관관계로 측정이 된다. 각 그림별 상관성은 아래와 같다[9].

9) 상관성분석은 두 변수의 선형적 관계를 측정하는 통계적 방식이다. 계수는 +1.0~0~-1.0사이에서 관찰이 되는데, 양수의 경우에 +1.0~+0.7은 강한 양적 선형관계, +0.6~0.3은 뚜렷한 양적 선형관계, +0.2~+0.1은 약한 선형관계를 이룬다. ETS의 e-rater의 경우에도 자동 채점과 수동 채점을 상관성분석으로 측정해

표 1. 그림별 상관관계

그림	그림 1	그림 2	그림 3	그림 4	그림 5
상관관계 계수	0.691	0.844	0.351	0.688	0.546

또한 일치의 신뢰도를 통계적으로 연산으로 카파 통계도 0.614를 나타냈다.[10] 그림 3과 연관되어서 특징적으로 점수가 낮은 까닭은 추상적 의미 표상의 잘못된 분석에 있다. 그림 3은 우유가 채워지는 그림으로 fill과 pour동사가 작문에 사용되었다. Pour의 경우는 (20a)과 같은 문장은 AGENT에 해당하는 ARG0, THEME에 ARG1, DESTINATION에 ARG2를 (20b)와 같이 부여해야 한다.

(20) a. Someone is pouring milk into a cup.
b. ARG0:Someone, ARG1:milk, ARG2: a cup

그런데, (21a)의 문장에서 THEME에 해당하는 요소를 AGENT인 ARG0로 분석하는 오류를 범하고 있다. 또한, (22a)의 문장에서 DESTINATION에 해당하는 요소와 THEME에 해당하는 요소를 동일하게 ARG1으로 분석하고 있다.

(21) a. Milk is pouring into the glass.
b. ARG0:Milk, ARG2:the glass
(22) a. Someone is filling the glass with milk.
b. ARG0:Someone, ARG1:the glass, ARG1: milk

(13-4)과 같이 smatch를 활용한 유사도 측정이 각각의 3중항을 분해해서 얼마나 같은 3중항이 있는지를 연산하기 때문에, (21-22)와 같이 서로 다른 논항들의 구조는 서로 다르게 판별한다. 따라서 구조적 문제를 잘못 연산한 추상적 의미 표상에서 문제가 비롯된다.

Habash와 Dorr (2001), Langkilde와 Knight (1998), Rebecca et al. (2001)은 어휘 개념 구조를 포함된 논항 구조에 기초한 추상적 의미 표상이 의미 구조적으로 유사한 통사적으로 변형된 문장들을 다양하게 생성할 수 있는 방안이라고 주장한다. 바꾸어서 말하면, 통사적으로 다양한 문장들이 서로 같은 구조인지 아닌지 판별하는 방법은 논항 구조의 유사성이

서, 자동 채점의 정확도를 산출하고 있다(Attali and Burstein, 2006).

10) 카파 통계는 0에서 1사이에 존재하는데, 0은 완전한 불일치, 1은 완전한 일치를 나타낸다. 0.1~0.20 아주 적은 일치, 0.21~0.40 조금 일치, 0.41~0.60 적절한 일치, 0.61~0.80실질적인 일치, 0.81~1.00 완전한 일치를 나타낸다.

요소 기술이 될 수 있는데, 이 방법은 어휘 개념 구조(Lexical Conceptual Structure)를 포함하고 있기 때문이다. (21-22)와 같은 문제는 동사 논항 구조뿐만이 아니라 논항들의 의미적 요소를 판별해야 한다. ARG0가 되는 문장 구성 요소가 AGENT의 속성 부여가 가능한지 ARG1이 되는 문장 요소가 THEME이 가능한지 여부를 판단 가능해야 한다.

이와 같은 판별 작업은 대상에 대한 지식 체계와 이 지식 체계에서 중의적인 요소들을 구분하는 방식이 필요하다. 다시 말하면 (22a)의 "Milk"는 행동할 수 있는 대상이 아니고 무생물 물질이라는 지식이 필요하며, "Milk"가 무엇을 지시하는지에 대해서 구분이 가능해야 한다.[11]

이러한 문제를 종합적으로 처리하기 위해서는 표상 자체도 다양한 어휘 사전과 지식 체계를 수용 가능하거나 포괄적이어야 한다. 추상적 의미 표상이 사건 구조를 표상하며, 이 구조를 WordNet과 같은 어휘 사전과 이를 온톨로지와 같은 지식기반 시스템과도 연결가능하다고 한다(Knight & Luk, 1994). 그러나 Flanigan et al. (2014)의 자연어 처리 시스템을 활용한 파싱 시스템은 WordNet의 의미 정보가 포함된 어휘에 기초할 뿐, 더 복잡한 어휘 개념 구조나 온톨로지와 같은 지식 기반 시스템을 활용하지는 않는다. 현재 제안된 자연어 처리 기반 시스템은 입력된 어휘 정보에 의해 연산된 3중항의 연쇄들이 계산적으로 얼마나 유사한지에만 초점이 맞춰져 있는 한계가 있다. 향후 개발되어져야 할 시스템들은 이러한 한계점을 극복하기 위해서 동의적 관계와 같은 어휘 정보, 논항구조와 같은 어휘 개념 정보, 그리고 전체 지식 체계의 개념 구조에 대한 정보 등을 통합적으로 고려해야 할 것이다.[12]

다음으로는 n-gram 방식으로 문장 간 비교방식인 BLEU를 적용해서 smatch를 적용한 방식처럼 가장 높은 점수를 받은 원형 문장의 점수를 채점으로 선택해서 수동 채점 테스트 집단과 비교해 보았다. BLEU는 대상이 되는 문장들간에 모든 n-gram의 합에 대한 일치하는 n-gram의 비율들의 합의 비율이다. 이 방법은 unigram, bigram, trigram, quadrigram등의 다양한 n-gram을 적용할 수 있다. 또한 적용되는 코퍼스 길이에 대해서 유연한 적용도 가능하고 너무 짧게 줄인 경우 감점이 되기도 한다. 이 방식은 재현율 기반으로 smatch 방식과 유사성이 있다. 2.1절에서 논의한 바와 같이 짧은 작문의 경우에 채점의 하나의 요소로도 활동된다(Madani et al. 2013). 이러한 기존 연구에 근거해서 본 연구에서는 smatch와 비교할 방법으로 BLEU를 선택했다. 연구와 smatch 측정법과 동일하게 원형이 되는 문장들과 채점 대상 문장들의 BELU값을 측정하고 가장 높은 측정값을 산출하는 원형의 되는 문장의 척도를 채택했다. BLEU의 경우에는 정규화 수치가 항상 부가되기 때문에

---

11) "Milk"가 사람을 가리킬 수도 있는데, 드레곤볼 만화에서는 등장인물을 가리킨다 (https://en.wikipedia.org/wiki/Milk_(disambiguation) 참조). 이 경우에 의미를 판별하기 위한 작업이 필요하며, 의미 중의성 해소 작업이 필수적이다.

12) 추상적 의미 표상 방식의 개념을 확장해서 의미적 논리 구조를 언어 중립적으로도 활용하고 있다. Vanderwende et al. (2014), Xue et al. (2014)는 다국어간 자동 번역 시스템에서 중간 표현의 의미적 논리 구조로도 활용하고 있다.

같은 수치가 하나 이상 나오는 경우가 있었는데, 이 경우에 일괄적으로 높은 척도를 채택했다.[13] 상관성 분석에서 BLEU는 음의 상관성을 보이는데, -0.01정도의 매우 낮은 수치를 보였다. 특히 n-gram은 단어의 순서에만 기반을 두기 때문에 의미 구조적 내용은 고려되지 않는다. BLEU도 이러한 문제점을 포함하고 있기 때문에 적절하지 못한 채점 방법이 된다.

Google 1T n-gram으로 추출된 trigram, BLEU적용수치, smatch적용수치를 로지스틱 다중 회귀를 통해서 어떠한 요인들이 채점과 직접적인 연관성이 있는지 비교해 보았다. 이러한 접근법은 어떠한 자질이 채점에 결정적 요인으로 작용하는지를 살펴보기 위함이다.[14]

Attali와 Burstein (2006)은 ETS e-rator의 자동채점 요소로 n-gram 모델을 활용했다. 자막을 활용한 사진 정보 검색에서 n-gram은 자막의 언어적 표현의 자연스러움을 평가하기 위해서 활용된다(Kuznetsova et al., 2013; Ordonez et al., 2015). 또한 사건들의 의미적 외연들을 묘사하는 표현들의 후보에서 자연스러운 것부터 순서화하기 위해서도 활용된다(Young et al., 2013). Dodge et al. (2012)는 자막에서 술어와 논항관계의 자연스러움의 순서화를 위해서 google 1T n-gram을 적용했다. 이와 같이 여러 기존 연구에서 활용된 n-gram을 적용해서 채점에 얼마나 결정적 요인으로 작동하는지를 살펴보았다.

문장이 (23a)와 같은 단어의 연쇄모델로 표현된다면 단어출현 확률인 (23b) 연쇄확률로 문장 확률을 측정가능하다. 연구에서는 이러한 방식에 기반을 두고 (23c)와 같은 trigram으로 다루었다.

(23) a. $s = w_1, \ldots, w_n$

b. $P(s) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2) = \prod P(w_k|w_1 \ldots w_{k-1})$

c. $P(w_n|w_{n-2}w_{n-1}) = \prod P(w_k|w_{k-N-2}w_{k-N-1})$

문장 내의 총 단어 수는 작문과 연관되어서 중요한 요소이다. 척도가 높은 원형 문장일수록 더 많은 단어로 작성되었고, 반대로 척도가 낮으면 더 적은 수의 단어로 작성되는 것이 관찰되었다. 척도 1의 경우에는 5~10 (평균 7.5), 척도는 2는 7~20 (평균 9.1), 척도 3은 8~22 (평균 13.3)으로 확연하게 척도가 높을수록 문장 내 단어 수가 많았다. Brew와 Leacock (2013), Sukkarieh와 Blackmore (2009)은 ETS의 단문 평가 시스템인 c-rater의 경우에 문장내 총 단어 수가 평가 요소로 포함되어 있다.

기존 연구의 문장의 통사 규칙과 연관되어서 크게 문맥 자유 규칙을 활용하는 경우와 의

13) 높은 수치를 선택한 이유는 동일한 n-gram이 높은 수치로 나오게 되므로, 동일 n-gram으로 판단한 경우에는 길이가 길거나 내용이 더 많은 쪽이 더 높은 점수로 나오는 것이 적절하기 때문이다.

14) Google 1T n-gram은 구글에서 만든 코퍼스로 주로 서적을 자동 스캐닝 작업을 해서 만든 코퍼스이다. 이 코퍼스는 1800년대부터 2000년대 초반까지 서적들에서 추출된 자료를 토대로 1조 이상의 유형과 9백만 이상의 문장이 포함된 현재까지 만들어진 코퍼스 중 가장 양적으로 방대하다(Brants & Franz, 2006).

존관계를 활용한 경우로 나뉜다. Kuznetsova et al. (2013)는 특정 어휘들의 의존적 관계 사진 자막과 연관되어 있다는 가정을 기초로, Stanford dependency parser를 활용한 de Mernefe와 Manning (2009)의 여러 의존 관계 중 (24)와 같은 의존관계를 분석에 활용했다.

(24) advcl, acomp, advmod, agent, amod, complm, ccomp, prep, npadvmod, nn, num, parataxis, partmod, purpcl, preconj, predet, quantmod, rcmod, ref, tmod, xcomp, xsubj, attr, auxpass, cc, cop, aux, csubj, csubjpass, expl, mark, infmod, mwe, nsubj, nsubjpass, conj, conj_and, dobj, iobj, neg, pobj, number, pcomp, possessive, prt, rel

연구에서는 이러한 의존 관계들이 채점에 미치는 영향을 통계적으로 살펴보기 위해서, Stanford dependency parser를 활용해서 모든 의존 관계를 살펴보고 (24)의 의존관계가 있는지 조사해서 통계적 수치를 추출했다.

Dodge et al. (2012), Ordnonez et al. (2015), Young et al. (2013)은 특정 단어가 포함된 문맥 자유 규칙의 확률적 분포나 규칙 자체를 자막이 이미지에 적합한지를 고려했다. 연구에서는 문장 내에서 추출되는 모든 문맥 규칙들을 통계 수치화했다. 이러한 규칙의 수치는 중복되지 않는 규칙들로 더 많은 규칙들은 더 복잡한 수식관계를 나타내고, 더 많은 논항 관계를 나타내기 때문에 규칙수 자체와 자막의 자세한 표현은 서로 상관성이 있다. 이와 같은 주장은 Young et al. (2013)의 더 많은 문맥 자유 규칙이 포함되면 더 세밀한 묘사가 가능하다는 관찰과도 일치한다. 따라서 문맥 자유 규칙의 경우에 통계적으로 높은 수치를 갖으면 사진을 더 세밀히 더 자세히 묘사할 수 있다는 것을 의미한다.

ETS의 c-rater의 경우에 논항은 단문 채점의 경우에 중요한 채점 요소로 간주된다. 특히 Sukkarieh와 Blackmore (2009)는 채점 요소로 정답이 되는 원형 문장과 비교가 되는 문장을 비교해서 생략된 단어의 숫자, 생략된 논항의 숫자를 고려했다. 이러한 고려는 ETS의 다른 연구에서 인간 수동 채점과 기계 자동 채점을 비교했을 경우에, 생략된 단어의 숫자가 가장 중요한 요소로 간주된다는 것과 일맥상통한다(Brew & Leacock 2013). 연구에서도 두 가지 요소를 고려의 대상에 포함시켰다. 그 외에 추상적 의미 표상이 비교 대상들 간에 논항 숫자를 고려하는 점에 착안해서 비교하는 문장의 논항숫자도 고려했다.

로지스틱 회귀분석 모델은 가설 모델의 적합성을 검정하거나 회귀계수 유의성 검정에 활용된다. 일반적 선형 회귀 모델이 각 계차항이 계수값과 연산되는 수식으로 연산되는 $f(x)=\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i+\epsilon$이라면, 로지스틱 회귀분석은 로짓(logit) 함수 $f(x)=\frac{1}{1+e^x}$에 기초한 (25)의 모델로 정의된다.

$$(25)\ f(x)=\frac{1}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i}}+\epsilon_i$$

이러한 로지스틱 회귀분석은 다음과 같이 연구에 적용했다. 연구에서 적용한 채점은 1, 2, 3 척도로 다항이 되며, 개별 범주로 고려해서 다항 로지스틱 회귀분석으로 처리한다. 각각의 채점 요인들을 독립 변인으로 고려해서 종속 변인인 채점 1, 2, 3 척도에 어떻게 분류되는가를 통계적으로 처리한다. 채점 요인들이 일정한 비율로 채점 척도에서 발생하게 되는데, 발생할 확률을 로짓 변환을 수행하고 이를 일정 선형 함수로 결합해서 이를 발생 확률 로짓 함수로 연산한 결과가 분석의 결과가 된다.

연구에서는 로지스틱 회귀분석의 계수에 초점을 맞추어 분석을 했다. 계수는 회계 계수로 하나의 요인이 변화하는데 대한 예측된 로짓 변환을 상수화한 것이다. 로짓 계수는 발생 확률과 관련해서 해당되는 변인의 발생 확률이 유의한 수준인지 아닌지에 대한 z-검정을 할 수 있으며, z-분포의 발생 확률을 예측할 수도 있다. 따라서 가설 검정이 가능하며, 영가설과 연구가설을 세울 수도 있다. 즉, 채점 변인인 총 단어수, 의존관계, 문장에서 활용된 문맥 자유 규칙, 총 논항숫자, 생략된 논항, 생략 단어, smatch, BLEU 평가 수치, google 1T n-gram 변인들이 채점이 미치는 영향이 유의미한지를 살펴보게 된다. 표 2에서 회귀식은 smatch + BLEU + google 1T n-gram + total (총 단어수) + dep (의존관계) + cfg (문맥자유 규칙의 수) + TotalArg (총 논항숫자) + MissingArg (생략된 논항) + MissingWord (생략된 단어) 이다.[15]

표 2. 로지스틱 회귀 분석 결과1

변인	예측	표준오차	z-값	z의 p-값
smatch	3.456e+00	4.718e-01	7.325	2.40e-13***[16]
BLEU	-1.999e-12	1.051e-12	-1.903	0.05713
google 1T n-gram	-1.513e+00	9.173e-01	-1.649	0.09912
total	0.80404	0.33182	2.42	0.01539*
dep	0.06622	0.33064	0.20	0.84126
cfg	0.0226	0.00109	2.07	0.03847*
TotalArg	-0.67544	0.31649	-2.13	0.03283*
MissingArg	1.11266	0.37398	2.975	0.00293**
MissingWord	1.36167	0.48422	2.812	0.00492**

15) 연구에서 활용한 통계 및 기계학습 패키지는 R 3.2.4이다.

16) ***는 유의성이 높은 수준의 변인을 나타내고, **은 더 낮은 *은 낮은 수준을 가리킨다. *이 없으면 유의성이 없는 경우를 가리킨다.

통계적 결과에서 관찰되는 것은 smatch, 총 단어수, 문맥자유 규칙의 수, 총 논항숫자, 생략된 논항, 생략된 단어들이 유의미한 변인들로 관찰된다. 또한 smatch의 통계적 결과가 가장 중요한 요인으로 나타난다. 또한 논항 구조와 연관성이 높은 자질들인 생략된 논항, 생략된 단어등도 주요한 자질이다. Brew와 Leacock (2013), Sukkarieh와 Blackmore (2009)에서 논의한 ETS의 c-rater 특정 요소와 같이 논항이나 논항과 연관성이 높은 자질들이 가장 주요한 채점 요소가 된다는 것과 일치한 결과가 나타났다. 그 외에 주요한 자질로는 문장 내 총 단어수, 문맥 자유 규칙이 유의미한 자질로 나타났다. 앞서 언급한 바와 같이 총 단어 수는 전체 논항 대표하는 자질로 활용될 수 있다. 즉 논항의 수는 일정한 단어 수와 연관성이 있는데, 이러한 자질들이 통계적 유의미하게 측정된 것이다. 그러나 언어 자질 중 의존 관계 자질은 중요하지 않은 자질로 나타나지만, 문맥 자유 규칙은 중요하게 나타났다. Young et al. (2013)은 사진 자막 문장에서 의미적 문제는 어떠한 술어가 어떠한 논항을 취하는가도 결정된다는 것에 기초해서 의미 자질들을 특징하는 방법으로 문맥 자유 규칙을 활용했다. 이러한 논의와 유사하게도 논항과의 유사성을 나타내는 방식으로 문맥 자유 규칙의 통계 정보가 활용될 가능성이 있다.

여러 자질들 중 로지스틱 회귀분석의 경우에 의미 있는 경우는 smatch와 google 1T n-gram을 적용한 결과이다. 특이하게도 google 1T n-gram이 의미있는 자질로 나타나는데, 다른 여러 결합에서는 의미 없는 자질로 판명된다.

표 3. 로지스틱 회귀분석 결과2

	예측	표준오차	z-값	z의 p값
smatch	3.342e+00	4.634e-01	7.213	5.49e-13***
google 1T n-gram	-2.061e-12	1.046e-12	-1.971	0.0487*

여러 가지 해석이 가능하지만, google 1T n-gram이 논항과 연관된 정보를 포함한 것으로도 해석할 수 있다. Google 1T n-gram은 단어의 연쇄에 의한 언어 자질이므로, 이 자질이 특정 논항이나 논항-술어의 연쇄를 포착할 수 있는 방법이 될 수도 있다.

연구에서는 또 모든 자질들 토대로 SVM을 적용한 분류 결과를 비교했다. SVM은 $\{i=1,2,...,n\}$ 데이터 D가 분류 공간 D=$\{(x_i,y_i)\}$에서 존재하고, y_i가 +1, -1로만 분류된다면 데이터 집합인 공간을 구분하는 초평면(hyperplane)을 구성하는 선형적 함수 $h(x)$를 (26)과 같이 설정할 수 있다.[17] (26)을 통해서 분류 공간 D에서 최대 마진(margin)을 찾아내는 분류기이다.

17) SVM은 여러 종류의 커널(kernel)이 있는데, 연구에서는 C-classification을 활용하고, 10, 20차 교차검증(10/20-fold cross-validation)을 활용했다.

(26) $h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + ... + w_d x_d + d$

여러 채점 요인들을 결합시켜서 어떠한 요인들이 가장 정확하고 의미 있는 결과를 나타내는지를 측정했다. 결과는 크게 정확도와 재현률로 측정되었다. 정확도는 측정한 것 중에서 얼마나 많은 정확한 분류를 했는가를 측정한 것이다. 재현률은 모든 정답 중에서 얼마나 많은 정확한 분류가 이루어졌는지를 측정한 것이다. 정확률은 분류한 것 중에서의 정확한 정도라면, 재현률은 정확한 분류가 이루어진 것의 얼마만큼 많은지 이다. 다시 말하면, 아무리 정확한 분류 방식이라도 매우 적은 정답만 도출한다면 의미가 없을 것이고, 반대로 아무리 높은 수준의 적용 범위가 있더라도 낮은 수준의 정확성을 갖는다면 의미가 없을 것이다.

표 4. 정확도 연산

요인	정확도 (%)	재현률(%)
smatch + BLEU + google 1T n-gram + total + cfg + dep + TotalArg + MissingArg + MissingWord	85.041	66.811
smatch + total + cfg + + TotalArg + MissingArg + MissingWord	83.231	65.217
smatch	65.225	53.478
BLEU + google 1T n-gram + dep	36.341	23.478
total + cfg + TotalArg + MissingArg + MissingWord	45.329	33.478

결과적으로 smatch를 적용한 결과가 가장 좋은 정확도와 재현률을 갖는 것으로 나타났다. Smatch를 제외한 요인들의 결합인 경우에는 정확도가 낮고 재현률이 낮게 나타났다. 따라서 smatch는 가장 중요한 요인으로 나타났다. 특히 정확도뿐만 아니라 재현률에도 주요한 요인으로 나타났기 때문에, 적용 범위가 넓은 요인으로도 판단된다.

5. 결론

스토리텔링은 언어 능력을 평가하는 일환으로 활용되며, ETS의 TOEIC Writing Test에서 사진 속 상황을 설명하는 평가에서 이용된다. 자막은 단문 형식으로 긴 에세이 형식과 다르며 채점 방식도 다르다. 이 논문에서 어떠한 평가 요소가 단문인 사진 자막 영작문 평가에 적절한지를 살펴보았다.

단문은 특정한 상황을 설명하기 위한 문장이 정해져 있다. 자막 단문 채점은 표본 문장과

작성된 문장이 얼마나 유사한지, 얼마나 관련성이 높은지, 사진속 내용을 얼마나 세밀히 묘사하는지를 평가해야 한다. 연구에서는 이러한 점을 고려해서 9개의 평가 요소를 설정했다. 여러 평가 요소 중 의미-논리 구조를 평가하기 위한 방식인 추상적 의미 표상 방식에서 활용되는 smatch가 가장 적절한 요소로 나타났다. Smatch는 문장 사이에 유사성을 논리-의미 구조를 비교하는 것으로 Brew와 Leacock (2013)의 단문 평가 요소와 일치한다.

기존 연구들이 논항 구조, 단어 길이 등등의 형식적 통계 요인들에 초점을 맞춘 반면에 이 연구는 의미-논리의 복합 구조를 통계화 했다. 이를 통제된 실험을 통해서 수집된 코퍼스를 활용해서 측정했다. 최종적으로 통계 요인들을 결합해서 수동 채점을 예측하기 위한 기계학습 방식을 적용해서 smatch가 가장 적절하다는 것을 입증했다.

이러한 연구결과는 컴퓨터를 활용한 언어 학습(Computer-Assisted Language Learning)에서 직접적으로 활용된다. Sukkarieh와 Blackmore (2009), Sukkarieh와 Stoyanchev (2009)에서 ETS 단문 채점 시스템인 c-rater를 소개하며 c-rater는 원형 문장과 기타 관련된 채점 기준을 자동으로 생성하며 여러 복잡한 비교 대상을 생성한다. 그러나 본 연구는 원형 문장과 작성된 문장을 추상적 의미 표상을 이용해서 자동으로 비교해서 여러 복잡한 기재 생성이라는 절차가 불필요하게 했다. 따라서 ETS의 c-rater보다 더 효율적인 방식이 활용됐다.

향후 연구에서는 여러 지식 기반을 적용할 필요성이 있다. 유사 논항 구조 판별을 위한 논항 구조 정의와 연관된 지식 기반이 필요하다. 이와 같은 측면은 Sukkarieh와 Stoyanchev (2009)의 연구에서도 지적하고 있다. 따라서 이를 적용한 확장된 연구와 작업도 향후 요구된다.

참고문헌

민선식. (2008). Toeic Writing Test 공식문제집. 서울: 시사영어사.

Achananuparp, P., Hu, X., & Shen, X. (2004). The evaluation of sentence similarity measures. In Song, I.-Y., Eder, J., and Nguyen, T. M. (Eds.), *Lecture Notes in Computer Science* (pp. 305-316). Berlin: Springer-Verlag.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rator V.2. Journal of Technology, *Journal of Learning, and Assessment*, 4(3), 3-30.

Baker, C., Fillmore, C., & Lowe, J. (1998). The Berkeley FrameNet project. In *proceedings of ACL*. 86-90.

Banarescu, L., Bonail, C., Cai, C., Georgescy, M., Griffitt, K., Hermajakob, U., Knight, K., Kohen, P., Palmer, M., & Schneider., N. (2013). Abstract

meaning representation for semanbanking. In *proceedings of Linguistics Annotation Workshop*. 178-186.

Botvin, G. & Sutton-Smith, B. (1977). The development of structural complexity in children's fantasy narratives. *Developmental Psychology*, 13(4), 377 – 388.

Brants, T., & Franz, A. (2006). *The Google web 1T 5-gram corpus version 1.1*. Linguistic Data Consortium 2006T13, Philadelphia, PA. Retrieved from Novermber 11, 2016, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13.

Brew, C., & Leacock, C. (2013). Automated short answer scoring. In Shermis, M. & Burnstein, J. (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 136-153). New York: Routledge.

Cai, S., & Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In *proceedings of the ACL*. 748-752.

Chiang, D., Andreas, J., Bauer, D., Hermann, K., Jones, B., & Knight, K. (2013). Parsing graphs with hyperedge replacement grammars. In *proceedings of ACL*. 924-932.

Dodge, J., Goyal, A., Han, A., Mensch, A., Mitchell, M., Stratos, K., Yamaguchi, K., Choi, Y., Daume, H., Berg, A., & Berg, T. (2012). Detecting visual text. In *proceedings of Conference of the NACACL*. 762-772.

Flangian, J., Thomson, S., Carbonell, J., Dyer, C., & Smith, N. (2014). A discriminative graph-based parser for the abstract meaning representation. In *proceedings of ACL*. 1426-1436.

Habash, N., & Dorr, B. (2001). Large scale language independent generation: using thematic hierarchies. In *proceedings of the MT-Summit*. 139-144.

Knight, K., & Luk, S. (1994). Building a large-scale knowledge base for machine translation. In *proceedings of AAAI*. 773-778.

Kuznetsova P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. (2013). Generalizing image captions for image-text parallel corpus. In *proceedings of ACL*. 790-796.

Langklide, I., & Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. In *proceedings of COLING*. 704-710

Li, Y., Bandar, A., McLean, D., & O'Shea, J. (2004). A method for measuring sentence similarity and its application to conversational agents. In *proceedings of the International FLAIRS Conference*. 820 – 825

Lin, C., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *proceedings of NACAC*. 71-98.

Liu, H., & Wang, P. (2008). Assessing sentence similarity using WordNet based word similarity. *Journal of Software*, 8(6), 1451-1458.

Madnani, N., Burstein, J., Sabatini, J., & O'Reilly, T. (2013). Automated scoring of a summary writing task designed to measure reading comprehension. In *proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. 163-168

McKeough, A. & Malcolm, J. (2011). Stories of family, stories of self: Developmental pathways to interpretive thought during adolescence. *New Directions for Child & Adolescent Development*, 2011(131), 59-71.

Matthiessen, C., & Bateman, J. (1991). *Text Generation and Systemic-functional Linguistics: Experiences from English and Japanese*. London: Pinter Publishers.

de Marneffe, M.-C., & Manning, D. (2008). The Stanford typed dependencies representation. In proceedings of *COLING*. 1-8.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. (2004). The NomBank project: An interim report. In *proceedings of NACACL*. 24-31.

Ordonez, V., Han, X., Kuznetsova, P., Kulkarni, G., Mitchell, M., Yamaguchi, K., Stratos, K., Goyal, A., Dodge, J., Mensch, A., Daume, H., Berg, A., Choi, Y., & Berg, T. (2015). Large scale retrieval and generation of image description. *International Journal of Computer Vision*, 119(1), 46-59.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71-106.

Papineni, K., Roukos, S., Ward, T., & Zhu. W. (2002). BLEU: A method for automatic evaluation of machine translation. In *proceedings of ACL*. 311-318.

Rebecca, G., Pearl, L., Dorr, B., & Resnik, P. (2001). Mapping WordNet senses to a lexical database of verbs. In *proceedings of ACL*. 244-251.

Sukkarieh, J., & Blackmore, J. (2009). C-rater: automatic content scoring for short constructed responses. In *proceedings of the International FLAIRS Conference*. 290-295.

Sukkarieh, J., & Stoyanchev, S. (2009). Automating model building in c-rater. In *proceedings of the 2009 Workshop on Applied Textual Inference*. 61-69.

Somasundaran, S., Lee, C., Chodorow, M., & Wang, X. (2015). Automated

scoring of picture-based story narration. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 42 – 48.

Sun, L., & Nippold, M. (2012). Narrative writing in children and adolescents: Examining the literate lexicon. *Language, Speech, and Hearing Services in Schools*, 43(1), 2 – 13.

Vanderwende, L., Menezes A., & Quirk, C. (2014). An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus. In *proceedings of NAACL-HLT*. 26-30.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ACL*, 2(10), 67 – 78.

<웹사이트>

Milk_(disambiguation). (n.d.) In Wikipedia. Retrieved July 25, 2016, from https://en.wikipedia.org/wiki/Milk_(disambiguation)

김동성
03760 서울시 서대문구 이화여대길 52
이화여대 인문과학대학
전화: (02)3277-4339
이메일: dsk202@ewha.ac.kr

Received on June 29, 2016
Revised version received on November 11, 2016
Accepted on December 30, 2016