시맨틱 프레임을 이용한 한국어 패러프레이즈 자동 평가 방법

박한철*, 권가진*, 최호진**
*한국과학기술원 지식서비스공학과
**한국과학기술원 전산학과
e-mail:{parhan86, ggweon, hojinc}@kaist.ac.kr

An Automatic Evaluation Metric for Korean Paraphrase via Semantic Frame

Hancheol Park*, Gahgene Gweon*, Ho-jin Choi**
*Dept of Knowledge Service Engineering, KAIST
**Dept of Computer Science, KAIST

요 약

본 연구는 지능형 QA시스템과 관련한 연구에서, 자동 패러프레이즈 생성 시스템을 평가하는 새로운 방법을 제시한다. 기존의 패러프레이즈 생성 시스템의 자동 평가 방법은 참조할 수 있는 패러프레이즈 정보의 양이 크게 제한되어 있었으며, 원 문장의 콘텍스트(context)와 이에 의존하는 통사적 구조(syntactic structure) 및 의미적 구조의 유사성을 고려하지 않고, 단순 구/단어 수준의 의미 유사성을 기반으로 생성된 패러프레이즈를 평가하였다. 이러한 문제를 해결하기 위해 본 연구는 시맨틱 프레임(semantic frame)을 이용한 패러프레이즈 문장 평가 방법을 제시한다. 본 연구에서 제시하는 방법론은 문장의 콘텍스트를 표현하는 프레임과 이러한 프레임이 발생시키는 통사적, 의미적 구조의 유사성을 바탕으로 원 문장과 패러프레이즈 문장의 '의미 유사성', '어휘 형태 비 유사성'을 평가하는 방식이다.

1. 서론

본 연구는 지능형 QA시스템 관련 연구로써, 자연어 문장의 자동 패러프레이즈 생성 시스템을 평가하는 새로운 방법을 제시한다. 자동 패러프레이즈 생성은 기계 번역, 문서 요약(multi-document summarization)과 같은 자연어 처리 과정의 핵심이며, 지능형 QA 시스템과 정보 검색(IR) 분야에서 쿼리 확장(query expansion)의 용도로 활용된다. 자동 패러프레이즈 생성 시스템을 평가하기 위해 현재까지는 평가자들을 통한 '수동 평가'[1, 4]가 주로 이루어졌다. 그러나 이러한 수동 평가의 결과는 매우 주관적이며, 평가 결과의 재현이 어렵고, 타 시스템과의 성능 비교를 위한 지표로 사용되기에 부적합하다는 단점이 있었다.

이러한 단점을 해결하기 위해 최근에는 패러프레이즈 관계인 문장 쌍의 정보를 활용하는 '자동 평가' 방법이 제안되고 있다[2, 3, 6]. 패러프레이즈 문장 쌍을 묶은 말뭉치로써, 단일 언어로 작성 된 패러프레이즈 문장 쌍을 묶은 말뭉치(monolingual parallel corpora)(이하 M.P.C.) 혹은 외국어와 번역된 문장 쌍을 묶은 말뭉치(bilingual parallel corpora)(이하 B.P.C.)가 이용되어 왔다. 그러나 자동 평가에 있어서도 사용되는 말뭉치에 따라 각기 다른 문제들을 보여 왔다. M.P.C.는 획득의 어려움과 이에 수반되는 참조 가능한 패러프레이즈 정보량의 제한으로 평가 정확성의 문제를 야기해왔다[3, 6]. 이를 극복하고자 현실적으로 사용 가능 량이 많은 B.P.C.를 활용하는 방식이 제안되어왔다. 동일한 외국어 구/단어를 공유하는 번역된 언어의 구/단어들을 패러프레이즈로 간주함으로써 더 넓은 범위의 패러프레이즈 정보를 획득할 수 있었다. 그러나 두 말뭉치 모두 공통적으로 평가 방식에 있어서는 평가 대상이 되는 구/단어가 포함된 문장에서의 콘텍스트와 이에 수반되는 통사적/의미적 구조 유사성의 고려보다는 단순 독립적인 구/단어에서의 의미 유사성만을 평가한다는 단점이 있었다[2].

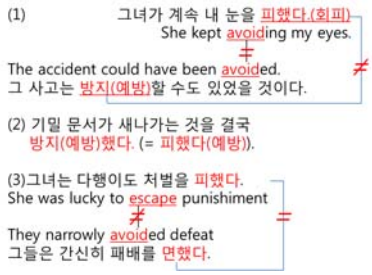
이를 해소하는 방법으로써 본 연구는 시맨틱 프레임[7]을 활용한 자동 평가 방식을 제안한다. 시맨틱 프레임은 문장의 요소들이(구/단어) 나타내는 사건/행동들에 대한 스키마적인 표현인 프레임과 이에 통사적, 의미론적으로 수반 되는 문장 내의 참여자와 이들의 의미적 역할(semantic role)을 표현하는 프레임 워크이다. 따라서 프레임은 문장의 콘텍스트를 표현하는 것으로 볼 수 있다. 이를 활용하여, 본 연구는 문장의 콘텍스트와 이러한 콘텍스트에 의존하는 통사적, 의미적 구조 유사성을 동시에 고려한 패러프레이즈 평가 방법을 제안한다.

2. 관련 연구

기존의 자동 평가 방식은 M.P,C를 참조하여 원 문장과 패러프레이즈 문장과의 N-gram 일치도 (의미, 표면적 형태 유사 정도)를 측정하는 방식으로 시스템 성능을 평가하였다[3, 6]. 그러나 생성된 패러프레이즈 문장의 유효한 평가를 위해서는 참조할 수 있는 가능한 많은 패러프레이즈 정보가 담긴 M.P.C를 사전에 만들어야 한다. 모든 패

러프레이즈 문장을 다룰 수 있을 양으로 단일 언어 패러프레이즈 쌍의 말뭉치를 수동으로 작성하는 것은 현실적으로 상당한 시간적 노력이 요구된다. 이러한 문제는 결과적으로 참조할 수 있는 정보의 양을 제약하여, 평가 결과를 신뢰하기 어렵게 만든다.

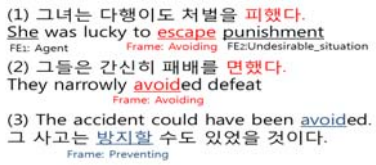
이 문제를 해결하는 방법으로 상대적으로 양이 많은 번역 말뭉치를(B.P,C.) 이용하는 방법이 제안되었다[2]. 기존 연구에서 현실적으로 더 많은 양이 존재하는 번역문들을 이용함으로써 M.P.C.보다 더 많은 패러프레이즈 정보를 생성할 수 있었다. 그러나 번역 말뭉치를 활용하는 방법 또한 몇 가지 문제가 발생할 수 있다. 첫 번째는 공유되는 외국어의 구/단어가 동일한 의미를 지니는지 확신할 수가 없다. 그림 1 의 (1) 예제에서 동일한 외국어 단어를 공유하더라도 다른 의미로 사용 될 수 있다는 점을 확인할 수 있다. 반면 문장의 콘텍스트에 따라서 '피하다'와 '방지하다'는 그림 1의 (2) 예제에서와 같이 동일한 의미를 나타낼 수 있다. 또 하나의 문제점은 그림 1의 (3) 예제와 같이 공유되는 외국어 구/단어가 다르나 한국어 의미가 같은 경우 이들이 패러프레이즈임을 인식하기 어렵다는 점에 있다.



(그림 1) 번역 말뭉치 활용 상 문제점

평가 방법에 있어서는, 기존의 연구들은 패러프레이즈 말뭉치를 참조하여 단순 구/단어 수준에서의 의미 유사성만을 평가해왔다. 그러나 해당 구/단어가 쓰인 문장의 콘텍스트에 따라 그림 1의 (1)과 같이 의미가 다를 수 있다. 기존 연구는 이러한 콘텍스트적인 측면을 크게 고려하지 않았다. 만약 원 문장과 생성된 패러프레이즈 문장의 콘텍스트 내에서의 구/단어의 의미와 그 콘텍스트가 발생시키는 통사적, 의미적 의존 구조의 유사성까지 높다고 판단되면 구/단어 수준에서의 의미 유사성이 결과적으로 문장 전체 수준에서 동일하다 볼 수 있다[5]. 본 연구는 이러한 가정에서 출발한다.

이러한 콘텍스트 문제를 다루기 위해서 본 연구에서는 시맨틱 프레임을[7] 활용한 패러프레이즈 생성 시스템의 자동 평가 방법을 제안한다. 프레임이란 어떤 행동/사건에 대한 스키마적인 표현을 의미하며, 즉 구/단어의 콘텍스트를 의미한다(e.g., 프레임 'avoiding'의 콘텍스트: 회피, 프레임 'preventing의 콘텍스트: 예방). 또한 프레임에 따라 의존하는 참여자와 그들의 역할이 결정된다. 다음 그림 2 에서 'avoid', 'escape'는 'Avoiding'과 'Preventing'이라는 프레임을 발생시키는 Lexical unit(LU)이며, 그림 2 (1)에서 'she(Agent)', 'punishment(Undesirable_situation)'는 LU에 통사적, 의미적으로 수반되는 참여자와 그들의 역할을 의미하며 이를 Frame Element(FE)라 한다[7]. FE는 '주어', '목적어'와 같은 통사적 의존 구조를 내포할 뿐만 아니라, 의미적인 역할, 지위 등을 포함한다. 이러한 정보는 문장의 콘텍스트와 이에 수반되는 통사 및 의미구조 평가의 부재를 해결해 줄 뿐 수 있다.



(그림 2) 시맨틱 프레임의 예제

이러한 접근법은 또한 그림1의 (3)의 문제를 해결해 줄 수 있다. 그림 2의 (2), (3)에서 같은 단어('avoid')를 공유하더라도 이는 서로 같은 의미가 아니다. 오히려, 같은 콘텍스트(Avoiding)를 공유하는 (1)의 'escape'와 (2)의 'avoid'만이 서로 패러프레이즈라는 것을 알 수 있다. 따라서 구/단어가 다른 형태의 외국어 구/단어를 공유하더라도 프레임이 동일하므로 패러프레이즈임을 인식할 수 있어 더 많은 패러프레이즈 정보를 획득 할 수 있게 해준다.

본 연구는 기존의 패러프레이즈 자동 평가 기준인 원문장과의 '의미 유사성'과 '어휘 형태의 비 유사성'을 측정하는데 있어 시맨틱 프레임을 활용하는 방법을 제안한다.

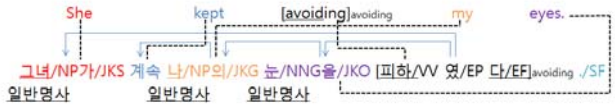
3 제안 방법

본 장에서는 한국어 프레임 학습을 위한 말뭉치 획득 및 자질 추출 방법(3.1)과 패러프레이즈 문장과 원 문장 사이의 '의미 유사성'과 '어휘 형태 비 유사성'을 측정하는 방법(3.2)을 제시한다.

3.1 한국어 프레임 학습을 위한 말뭉치 획득 및 자질 추출

본 연구를 위한 번역 말뭉치는 웹에서 제공되는 한/영 번역 뉴스를 이용한다. 먼저 말뭉치 태깅을 위해, 영어/한국어 번역 쌍에 대해서 구/단어 수준에서 의미가 같은 것끼리 얼라인먼트한다 (e.g., she=그녀가, eyes=눈을). 이는 구/단어 단위에서 번역 사전을 참조하거나 정규 표현식을 활용하여 수행된다. 그 후 Automatic SRL (Semantic Role Labeling) 도구인 SAMAFOR[8]를 활용하여 영어 문장에서 프레임을 나타내는 LU를 추출하고 이에 대한 프레임을 태깅한다. 마지막으로 한국어에 대해서는 ETRI 언어 분석기를 통해 Bag-of-words, NE(Named Entity), 의존 문법 파싱, 형태소 분석 결과를 얻은 후 이들을 각 각 태깅한다. 의존 문법 파싱과 형태소 분석을 하는 이유는

LU에 의해 의존적으로 수반되는 통사(격조사(주격조사, 목적격 조사 등), 구의 품사 등)구조에 대한 정보를 얻기 위함이며, Bag-of-words, NE(Named Entity)는 의미적 구조(의미상의 지위, 역할 등)의 정보를 얻기 위함이다. 상기 정보들은 한국어 프레임을 분류하기 위한 학습 자질로 사용된다. 다음 그림 3은 태깅이 완료된 문장의 예다.



(그림 3) 태깅이 완료 된 한/영 번역 문장

위 그림에서 'avoiding'은 프레임을 의미하며, 'avoiding', '피하였다'는 LU이다. 점선은 얼라인먼트를 의미하며, 화살표는 의존 문법 파싱의 결과이다. 그리고 한국어 단어 '/' 옆에는 형태소 분석결과가 있으며, '일반명사'라고 적힌 부분은 해당하는 명사의 NE를 의미 한다. 태깅 정보를 바탕으로 표 1과 같이 LU의 학습 자질들을 나열할 수 있다.

<표 1> 한국어 프레임 학습 자질

Class (Frame)	Lexical Unit(LU)	Bag-of-words	NE	Proposition Pattern
Avoiding	피하다	눈 나	일반명사 일반명사	-의/JKG -을/JKO
Avoiding	회피하다	사람 시선	PS_OTHER 일반명사	-의/JKG –을/JKO

표 1의 두 번째 행은 표 1의 예제 문장에 태깅된 정보를 바탕으로 자질의 인스턴스를 나열한 것이다. Bag-of-words, NE, 그리고 조사 패턴은 프레임에서 FE에 해당하는 단서로써 학습된다. 단, 이러한 자질 정보들은 프레임을 나타내는 LU가 포함 된 의존 트리 경로에 (의존 문법 파싱결과 LU에 의존하는 것들의 경로) 포함되는 것만을 취한다. 또한 어떤 LU가 한 문장 내 두 개의 트리에서 중복 될 경우에는 문장 내에서 가장 가까운 트리로만 LU를 할당시킨다. 그림 3의 문장의 의존 트리 경로는 [그녀가 - 피하였다] [계속-나의-눈을-피하였다]의 두 가지로 나타난다. LU인 '피하였다'와 '계속 나의 눈을'은 '그녀가' 보다 문장에서 위치가 더 가깝기 때문에 [그녀가 -피하였다]에는 LU가 제거되며, 따라서 이 트리는 학습자질에서 제외된다. 이는 한국어가 head-final 언어라는 점에서 기인한다.

3.2 생성된 패러프레이즈 문장의 평가

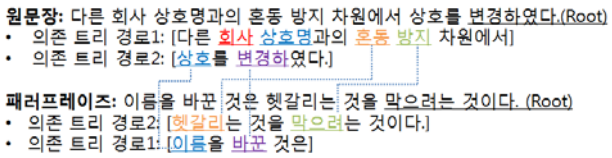
본 절에서는 학습이 완료 된 후, 새로운 입력 문장에 대한 패러프레이즈 생성 결과를 평가하는 방법을 소개한다. 자동 생성된 패러프레이즈 문장은 원 문장과의 콘텍스트 유사도를 가중치로 하여, 두 가지 평가 기준인 "의미 유사성", "어휘 형태의 비 유사성"이 측정된다. 이는 적절한 패러프레이즈 문장은 원 문장의 의미를 보존하면서, 최대한 다른 형태로 표현되어야 한다는 점에서 기인하며, 실제 이 기준에 의한 자동 평가 점수가 사람에 의한 수동 평가 점수와 높은 상관관계를 보여 왔다는 점에 착안한다[2, 6].

이를 위해 먼저 생성된 문장과 원 문장의 프레임과 프레임에 의존적인 통사 구조의 유사성을 측정한다. 세부적으로는 우선 원 문장과 생성 된 문장이 지니는 프레임들을 분류하는데서 출발한다. 이를 위해 3.1에서 언급된 자질 추출 방법과 동일한 방식으로 원 문장과 패러프레이즈 문장에서 자질을 추출한 후, 다음의 결정 식을 바탕으로 각 문장의 프레임들을 분류한다.

$$\hat{f} = argmax_{f_k} P(f_k|l,b,n,p) \quad (1)$$

$f = \{f_1, f_2, ... f_n\}$: possible Frame in the lexcal unit
$l = \{l_1, l_2, ... l_i\}$: lexcial units for f
$b = \{b_1, b_2, ... b_j\}$: bag of words for f
$n = \{n_1, n_2, ... n_k\}$: NER for f
$p = \{p_1, p_2, ... p_t\}$: postpositon pattern for f

원 문장과 생성된 패러프레이즈 문장내의 각 LU들의 프레임은 상기 4개의 자질(lexical unit, bag-of-words, NE, proposition pattern)을 바탕으로 가장 큰 확률을 도출하는 것으로 결정 된다. 그 후 의존 문법 파서를 통해 원 문장과 패러프레이즈 문장 각각의 의존 트리 경로를 추출한다. 이는 서로 의존 관계에 있는 요소들은 관련 있는 하나의 행동/사건을 표현하기 위해 서로 의존할 가능성이 높기 때문이다. 의존 트리 경로에서 서술어에 해당하는 Root는 복수의 의존 트리에서 중복될 수 있다. 이런 경우 Root는 3.2절에서 언급한 바와 같이 문장 내에서 거리가 가까운 요소들이 위치한 경로로 할당 된다. 그림 4는 각 문장에 대한 의존 트리 경로를 보여준다. 해당 예제에서 '변경하였다' 와 '막으려는 것이다'는 2개의 경로에 대해서 중복되는 서술어(root)이기 때문에 문장 내에서 가까운 요소가 있는 트리 경로로 할당된다. 따라서 두 경로는 결과적으로 상호 배타적인 경로가 된다.



(그림 4) 의존 트리 경로의 예

각 의존 트리 경로에 대해서 원 문장과 패러프레이즈 문장 사이에 프레임과 이에 따른 통사적 구조가 얼마나 유지되는가를 다음 식 (2)을 통해 구한다.

$$Score_{context} = \begin{cases} \frac{N(f_s \cap f_c)}{N(f_s)}, & if\ there\ is\ at\ least\ one\ LU\ in\ tree\ path \\ 1, otherwise \end{cases} \quad (2)$$

여기서 $Score_{context}$는 패러프레이즈 문장이 나타내는 콘텍스트와 그 콘텍스트에 따른 의존적인 통사 구조(의존 트리 내에서 주변 프레임)들이 하나의 의존 트리 경로 내에서 얼마나 원 문장과 일치하는가를 의미한다. 여기서 f

는 프레임, s는 원 문장 c는 생성된 패러프레이즈, 그리고 N은 개수를 의미한다. 그림 4에서 의존 트리 경로에서 점선으로 연결 된 부분은 서로 같은 프레임을 의미한다. 따라서 그림 4에서 의존 트리 경로 1의 $Score_{context}$는 0.5, 의존 트리 경로 2는 1.0 이다.

본 단락부터 본론 적으로 '의미 유사성'과 '어휘 형태 비 유사성'을 측정 방법을 소개한다. 패러프레이즈 문장이 원 문장의 콘텍스트와 콘텍스트에 따른 통사적 구조를 따른다면 구/단어 단위에서의 유사성은 전체 문장의 의미 유사성과 크게 관련된다는 가정 하에, 위에서 구해진 $Score_{context}$는 문장의 구/단어 의미 유사성에서 가중치로 작용한다.

의미 유사성은 각 문장의 구성요소(구/단어)가 의미적으로 얼마나 일치하는가를 각 의존 경로 별로 구한 후 각 경로 별 콘텍스트 일치도인 $Score_{context}$를 가중치로 곱한 뒤 이들 모두의 평균을 구한다. 각 구/단어의 의미 일치는 동일한 영어 구/단어를 지닌 것 혹은 같은 프레임을 지닌 것을 같은 의미를 지닌 것으로 간주 하여 계산한다. 수식 (3)은 이를 표현하는 것이다. 수식에서 C는 구/단어에 해당하고 i는 의존 트리 경로를 의미한다.

$$\frac{\sum_i Score_{context_i} * \frac{N(c_s \cap c_c)}{N(c_s)}}{N(i)} \quad (3)$$

어휘 형태 비 유사성은 의미 유사성과 구하는 방식이 유사하다. 단, 식 (3)에서 분자는 하나의 의존 트리 경로 내에서 같은 의미를 지녔지만 표면적 형태가 다른(e.g., 상호-이름) 구/단어의 개수가 몇 개 인가가 된다.

4. 제안하는 방법의 성능 평가 계획

본 연구에서 제안하는 평가 방식과 기존에 제안 되었던 평가 방식을 비교하기 위해서, 기존의 자동 패러프레이즈 생성 시스템[4]과 사람에 의해서 생성된 다수의 패러프레이즈 문장들이 활용된다. 주어진 패러프레이즈 문장들에 대해서, PEM[2], ParaMetric[3], BLEU/PINC[6]와 본 연구에서 제안하는 방식으로 각각 자동 평가한다.

또한 3명의 평가자들을 고용하여 동일한 데이터에 대해서 다음 3가지 기준으로 평가하도록 한다[2].

의미 유사성: 원 문장의 의미가 적절히 보존 되어있는가?
어휘 형태 비 유사성: 얼마나 패러프레이즈 문장이 원 문장의 형태를 변화시켰는가?
전체: 전체적으로 패러프레이즈가 잘 되었는가?

평가자 들은 각 문장에 대해서 상기 기준에 따라 1(매우 미비함)에서 5(매우 완벽함)까지의 점수를 부여하게 된다. 결과적으로, 수동 평가에 의한 점수와 자동평가에 의한 점수 사이의 피어슨 상관계수를 측정함으로써 기존 평가 방식과의 성능 비교를 수행할 수 있다[2, 6].

5. 결론

본 연구에서는 자동으로 생성된 패러프레이즈 문장의 품질을 측정하기 위해 시맨틱 프레임을 활용하였다. 이는 단어/구 수준에서의 의미 유사성만을 측정하는 기존 평가 방법을 확장하여, 원 문장과 패러프레이즈 문장 간의 콘텍스트 및 콘텍스트에 따른 통사적, 의미적 구조의 유사성까지 평가하는 방식이다. 차후 논문에서는 본 연구에서 제안한 방식과 기존 평가 방식의 비교연구가 수행 될 것이다.

본 연구에서 제안하는 평가 기준을 적용한다면, 쿼리의 정확성이 요구되는 지능성 QA 시스템에서의 쿼리 확장 과정에서 적합하지 않은 쿼리들을 생성하는 패러프레이즈 문장을 필터링함으로써 좀 더 정확성 높은 정답을 이끌어 낼 수 있을 것으로 예상된다.

감사의 글

본 연구는 미래창조과학부 산업융합원천기술개발사업의 "휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발"과제의 지원으로 수행되었음 (과제번호 10044577)

참고문헌

[1] Ali Ibrahim, Boris Katz, and Jimmy Lin. "Extracting Structural Paraphrases from Aligned Monolingual Corpora.", Association for Computational Linguistics, pp. 57-64 , 2003

[2] Chang Liu, Daniel Dahlmeier, Hwee Tou Ng, PEM: "A Paraphrase Evaluation Metric Exploiting Parallel Texts.", Conference on Empirical Methods in Natural Language Processing, pp. 923-932, 2010

[3] Chris Callison-Burch, Trevor Cohn, Mirella Lapata, ParaMetric: "An Automatic Evaluation Metric for Paraphrasing.", International Conference on Computational Linguistics, Vol.1, pp. 97-104, 2008

[4] Colin Bannard and Chris Callison-Burch, "Paraphrasing with Bilingual Parallel Corpora.", Association for Computational Linguistics, pp. 597-604, 2005

[5] Daniel Andrade et al, "Detecting Contradiction in Text by Using Lexical Mismatch and Structural Similarity", NTCIR Conference, pp. 512- 517, 2013

[6] David L. Chen, William B. Dolan, "Collecting Highly Parallel Data for Paraphrase Evaluation", Association for Computational Linguistics, pp. 190-200, 2011

[7] Charles Fillmore, Collin Baker, "A Frames Approach to Semantic Analysis. In B. Heine and H. Narrog (eds.), The Oxford Handbook of Linguistic Analysis, pp. 313-340, 2010

[8] SEMAFOR, http://www.ark.cs.cmu.edu/SEMAFOR/