인지과학, 제22권 제1호
Korean Journal of Cognitive Science
2011, Vol. 22, No. 1, 57～78.

이중 언어 기반 패러프레이즈 추출을 위한 피봇 차별화 방법*

박 에스더　　이 형 규　　김 민 정　　임 해 창†
고려대학교 컴퓨터・전파통신공학과

패러프레이즈는 같은 의미를 다른 단어를 사용하여 표현한 것을 말한다. 패러프레이즈는 일상적인 언어생활에서도 흔히 관측되며 자연어처리 분야에서 다양하게 활용할 수 있다. 특히 최근에는 통계적 기계 번역 분야에서 데이터 부족 문제를 보완하여 번역 성능을 향상시키기 위해 패러프레이즈를 활용한 연구가 많다. 이중 언어 병렬 말뭉치를 이용하는 패러프레이즈 추출 과정에서는 일반적으로 다른 언어를 피봇으로 사용하기 때문에 단어 정렬 및 구 정렬 과정을 두 번 거친다. 따라서 단어 정렬의 오류가 패러프레이즈로 전파될 수 있다. 특히 한국어와 영어와 같이 언어의 구조적인 차이가 큰 경우에는 단어 정렬 오류가 더 심각하기 때문에 피봇 프레이즈부터 잘못 추출되는 경우가 많아진다. 이러한 문제를 보완하기 위해 본 논문에서는 패러프레이즈 추출 과정에서 피봇 프레이즈를 차별화하는 방안으로서 어휘, 품사 정보를 이용해 올바른 피봇 프레이즈에 더 높은 가중치를 부여하는 방법을 제안한다. 실험 결과, 제안하는 피봇 가중치 부여 방법을 기존의 패러프레이즈 추출 방법에 추가했을 때 패러프레이즈 추출 정확률과 재현율이 모두 향상됨을 확인할 수 있었다. 또한, 제안하는 방법을 통해 추출한 패러프레이즈를 한영 기계 번역 시스템에서 활용하였을 때 번역률이 향상됨을 확인할 수 있었다.

주제어 : 패러프레이즈, 패러프레이즈 추출, 피봇 차별화 방법

---

* 이 논문은 교육과학기술부 한국연구재단(KRF-2007-361-AL0013) 및 2단계 BK21사업의 지원을 받아 수행되었음. 이 논문의 초고는 2010년 '한글 및 한국어 정보 처리 학술대회'에서 발표했던 내용을 보완한 것임.

† 교신저자: 임해창, 고려대학교 컴퓨터・전파통신공학과, 연구분야: 자연어처리
(E-mail: rim@nlp.korea.ac.kr)

서 론

패러프레이즈란 같은 의미를 나타내는 여러 가지 표현을 말한다[1]. 예를 들어, 그림 1에서 "_을 찾고 있_" 과 "_을 물색하고 있_"은 패러프레이즈로, 동일한 의미를 나타내기 때문에 서로 대체해서 사용할 수 있음을 확인할 수 있다.

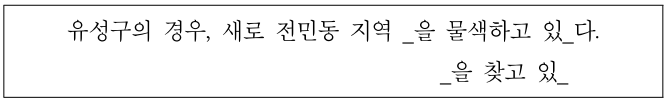

그림 1. 패러프레이즈의 예

자연어 처리 분야에서 패러프레이즈는 다양하게 활용할 수 있다. 특히 최근에는 통계적 기계 번역의 성능 향상을 위해 패러프레이즈를 활용한 연구가 많다. [2,3]은 학습 데이터의 확충을 위해 패러프레이즈를 이용하였으며, [4,5]는 패러프레이즈를 이용해 구 번역 테이블을 확장하여 데이터 부족 문제를 완화하였다.

패러프레이즈 추출을 위한 기존 연구는 크게 단일어 병렬 말뭉치를 이용하는 연구와 이중 언어 병렬 말뭉치를 사용하는 연구로 나눌 수 있다. [5,6]에서 사용한 단일어 병렬 말뭉치는 한 가지 소설의 여러 번역본을 수집한 말뭉치와 정해진 시간에 일어난 사건의 뉴스 기사를 수집한 말뭉치이다. 이러한 단일어 병렬 말뭉치는 구축이 매우 어려울 뿐만 아니라 도메인이 한정되어 있기 때문에 패러프레이즈 추출에도 제약이 따른다. 따라서 이중 언어 병렬 말뭉치를 사용하는 연구가 늘어나고 있으며, [7]은 이중 언어 병렬 말뭉치를 사용하여 추출된 패러프레이즈가 더 유용하다고 입증하였다.

[1]은 이중 언어 병렬 말뭉치를 이용한 방법 중 피봇 언어를 사용하는 방법으로, 언어에 독립적이고 매우 간단하다는 장점이 있으나 단어 정렬과 구[1] 정렬 과정을 두 번 거치는 과정에서 정렬 오류가 전파되는 단점이 있다. 특히 영어와 한국어 같이 구조적인 차이가 큰 언어쌍의 경우, 단어 정렬 성능에 한계가 있으므로 오류

1) 본 논문에서는 언어학적 구를 의미하는 것이 아니라, 구 단위 통계적 기계 번역에서의 구와 마찬가지로 단어열을 의미한다.

전파 문제가 더욱 심각하다. 본 논문에서는 이러한 문제점을 보완하기 위해 패러프레이즈 추출 중간 단계인 피봇 프레이즈에 중점을 두고, 올바른 피봇 프레이즈에 더 높은 가중치를 부여하는 패러프레이즈 추출 방법을 제안한다. 또한, 실험에서 피봇 프레이즈 가중치 방법을 기존 패러프레이즈 모델과 결합하여 패러프레이즈 추출 성능이 향상됨을 보이고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련한 기존 연구에 대해서 살펴본다. 3장에서는 피봇 프레이즈를 고려해야 하는 이유, 피봇 프레이즈를 고려한 자질 함수, 기존 패러프레이즈 추출 모델에 피봇 프레이즈 자질 함수를 결합한 방법과 전체 시스템에 대해 설명한다. 4장에서는 실험 및 평가에 대해서 분석한 후, 마지막으로 5장에서는 결론 및 앞으로의 연구 방향을 다룬다.

관련 연구

[1]은 처음으로 이중 언어 병렬 말뭉치를 사용하여 영어 패러프레이즈를 추출한 연구이다. [1]에서는 영어 구 e_1과 e_2가 다른 언어의 동일한 구인 c(피봇 프레이즈)와 정렬되었다면, 두 영어 구 e_1, e_2는 서로 패러프레이즈라는 가정 하에 패러프레이즈를 추출하였다. 예를 들어 한영 병렬 말뭉치에서 영어의 패러프레이즈를 찾는다고 할 때, 그림 2와 같이 “를 조사하”에 동일하게 정렬된 “look into”와 “investigate”는 패러프레이즈가 된다.

[1]에서는 패러프레이즈 확률을 다음 식(1)과 같이 구하였다. e_1은 하나 이상의 피봇 프레이즈와 정렬될 수 있기 때문에 가능한 피봇 프레이즈와 동일하게 정렬된 모든 구가 후보 패러프레이즈가 된다. 그러므로 패러프레이즈 확률 모델은 가능한 모든 피봇 프레이즈로의 번역 확률의 합으로 구성된다.

$$p(e_2|e_1) = \sum_c p_{MLE}(c|e_1) p_{MLE}(e_2|c) \quad (1)$$

$p_{MLE}(c|e_1)$과 $p_{MLE}(e_2|c)$는 각각 e_1이 c로, c가 e_2로 번역될 확률을 말한다.

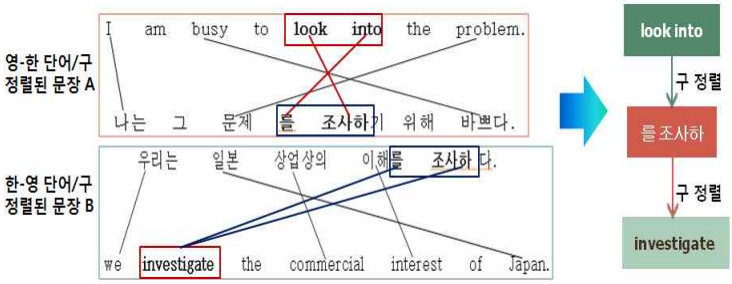

그림 2. 병렬 말뭉치를 이용한 패러프레이즈 추출 예

이 값은 병렬 말뭉치를 이용하여 최대 우도 측정(MLE)방법으로 다음 식과 같이 구하였다. 식(2)에서 $count(c, e_1)$는 말뭉치에서 c와 e_1이 정렬된 빈도수를 말한다. $p_{MLE}(e_2|c)$ 또한 식(2)와 같은 방법으로 계산하였다.

$$P_{MLE}(c|e_1) = \frac{count(c,e_1)}{\sum_c count(c,e_1)} \quad (2)$$

이 방법은 언어에 독립적이고 비교적 간단한 방법이라 할 수 있다. 하지만 패러프레이즈 후보를 추출하고 패러프레이즈 확률을 구하는 과정에서 e_1에서 c로, c에서 e_2로 두 번의 단어 정렬, 구 정렬 과정을 사용하게 되면서 단어 정렬과 같은 이전 단계의 오류가 패러프레이즈 추출에 많은 영향을 미친다.

[8]은 이중 언어 병렬 말뭉치를 이용하여 패러프레이즈 패턴을 추출한 연구이다. [8]에서 제안한 패러프레이즈 추출 모델 중 모델 1은 [1]의 모델에 단어 단위 가중치 자질[2] 을 추가하여 구 대역 확률의 오류를 완화하고자 하였다. [8]의 모델 1에서 패러프레이즈 확률을 구한 수식은 다음과 같다.

2) 원문에는 lexical weighting(LW)이라고 되어 있으나, 단어 단위로 단어 정렬 정보를 반영하는 것이므로 본 논문에서는 단어 단위 가중치 자질이라고 해석한다.

$$score(e_1|e_2) = \sum_c \exp\left[\sum_{i=1}^{4} \lambda_i h_i(e_1, e_2, c)\right] \quad (3)$$

식(3)에서 사용한 4개의 자질 함수 중 h_1과 h_2는 대역 확률 값이고, h_3과 h_4는 단어 단위 가중치 자질 값이다. 이를 풀어 쓰면 다음 수식과 같으며 λ_i는 각 자질 함수의 가중치를 나타낸다.

$$\begin{aligned} h_1(e_1, e_2, c) &= score_{MLE}(c|e_1) \\ h_2(e_1, e_2, c) &= score_{MLE}(e_2|c) \\ h_3(e_1, e_2, c) &= score_{LW}(c|e_1) \\ h_4(e_1, e_2, c) &= score_{LW}(e_2|c) \end{aligned} \quad (4)$$

자질함수 h_1과 h_2는 식(2)와 같이 대역 확률로 구한다. 자질함수 h_3과 h_4를 구하는 수식은 식(5)와 같다.

$$score_{LW}(c|e) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{|\{j|(i,j)\in a\}|} \sum_{\forall (i,j)\in a} w(c_i|e_j)\right) \quad (5)$$

a는 c와 e의 단어정렬이며, n은 구 c를 구성하는 단어 개수이고, c_i와 e_j는 c와 e의 각 단어를 나타낸다. $w(c_i|e_j)$을 구하는 방식은 식(6)과 같은데, 이 때 $count(c_i, e_j)$는 말뭉치에서 단어 c_i와 e_j가 정렬된 빈도수이다. $score_{LW}(e|c)$ 또한 식(6)과 같은 방법으로 계산된다.

$$w(c_i|e_j) = \frac{count(c_i, e_j)}{\sum_{c_i} count(c_i, e_j)} \quad (6)$$

지금까지 소개한 패러프레이즈 추출 연구는 다른 언어를 피봇 언어로 사용하며

공통적으로 병렬 문장에서의 단어 정렬 및 구 추출 기법을 사용한다. 이 방법은 언어에 독립적이고 비교적 간단하다. 하지만 패러프레이즈 후보를 추출하고 패러프레이즈 확률을 구하는 과정에서 e_1에서 c로, c에서 e_2로 두 번의 단어 정렬, 구 정렬 과정을 거치면서 이전 단계의 오류가 패러프레이즈 추출에 많은 영향을 미친다. 또한, 기존 모델인 [1]과 [8]의 패러프레이즈 추출 모델은 모든 피봇 프레이즈를 동일하게 사용함으로서, 패러프레이즈 추출 성능에 좋은 결과를 주지 못하였다.

본 논문에서는 패러프레이즈를 추출하는 중간 단계 산출물인 피봇 프레이즈에 차별화를 줌으로서, 이러한 문제를 완화하고자 한다. 또한, 기존 모델인 [1]과 [8]의 패러프레이즈 추출 모델에 피봇 프레이즈 가중치를 추가하는 방법을 제안한다.

패러프레이즈 추출을 위한 피봇 차별화 방법

피봇 프레이즈를 고려해야 하는 이유

[1]의 가정에 근거해 추출한 패러프레이즈를 분석한 결과, 잘못된 피봇 프레이즈가 많았으며, 피봇 프레이즈의 오류가 패러프레이즈 결과에도 영향을 미치는 것을 확인할 수 있었다. 그림 3은 “를 받”에 대한 패러프레이즈를 패러프레이즈 확률 순위, 피봇 프레이즈와 함께 나열한 것으로, “의 지시”의 확률이 제일 높다. “의 지시”와 “를 그리다”와 같은 잘못된 패러프레이즈가 추출된 경우를 보면, 구 정렬 오류로 인한 잘못된 피봇 프레이즈와 정렬되어 있음을 확인할 수 있다. 내용어가 하나도 없는 피봇 프레이즈인 “by”와 정렬되어 “의 지시”와 같은 잘못된 패러프레이즈가 추출되는 결과는 [1]에서 많이 나타나는 결과이며, “를 그리다”와 같이 한쪽의 구 정렬 오류로 인해 잘못된 패러프레이즈가 추출되는 결과는 [1,8]에서 많이 나타나는 결과이다. 올바른 패러프레이즈 “를 얻”, “를 수령하”가 추출된 경우에는 양방향 구 정렬이 모두 올바르게 되어서 올바른 피봇 프레이즈 “obtain”, “receive”를 피봇프레이즈로 사용하였다.

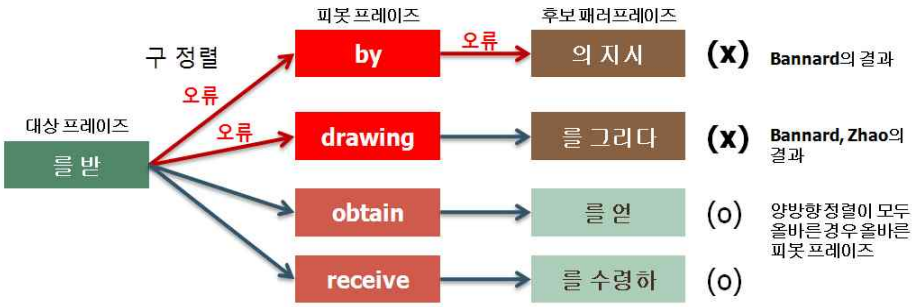

그림 3. “를 받” 패러프레이즈 추출 순위

이러한 분석 결과를 통해 구 정렬 오류로 인한 피봇 프레이즈 추출 오류가 심각하며, 피봇 프레이즈의 오류가 패러프레이즈 추출에도 전파된다는 사실을 알 수 있다. 그러나 단어 정렬의 오류 전파 때문에 피봇 프레이즈를 분별하는 데는 어려움이 있다. 따라서 단어, 구 정렬 정보 외에 다른 정보를 이용하여 올바른 피봇 프레이즈를 선택할 수 있는 방법이 필요하다.

본 논문에서는 올바른 피봇 프레이즈에 대한 정의를 하고, 그와 같은 올바른 피봇 프레이즈를 선별하는 방법을 제안한다. 또한 이를 기존 모델과 결합하여 기존 모델을 개선하고자 한다.

제안하는 방법을 설명하기 전에 우선 본 논문에서 사용하는 용어를 정리하면 다음과 같다.

- 대상 프레이즈: 패러프레이즈를 찾고자 하는 프레이즈. 그림 3에서 “를 받”에 해당
- 피봇 프레이즈: 대상 프레이즈와 구 정렬이 된 프레이즈. 그림 3에서 “by”, “drawing”, “obtain”, “receive”에 해당
- 후보 패러프레이즈: 구 정렬을 거쳐 추출한 패러프레이즈의 후보. 그림 3에서 “의 지시”, “를 그리다”, “를 얻”, “를 수령하”에 해당

올바른 피봇 프레이즈 가중치 부여

본 논문에서는 올바른 피봇 프레이즈(c)를 다음과 같이 대상프레이즈(e_1)와 패러프레이즈(e_2)의 모두에 대해 올바르게 번역된 프레이즈라고 정의한다. 예를 들어 그림 4와 같이 (a)와 (b)의 c는 e_1과 e_2에 올바르게 번역된 프레이즈라고 볼 수 있으므로, “obtain”과 “receive”는 올바른 피봇 프레이즈가 된다.

만약 구 정렬외에 다른 정보를 이용해 e_1과 e_2가 패러프레이즈일 가능성이 높다는 사실을 알 수 있다면, 그 때의 e_1과 e_2를 연결해주고 있는 피봇은 올바른 피봇 프레이즈일 가능성이 높다. 관찰 결과, e_1과 e_2가 같은 내용어를 포함할 때의 피봇 프레이즈와 e_1과 피봇 프레이즈가 같은 품사를 포함할 때의 피봇 프레이즈가 대체로 올바른 피봇 프레이즈임을 관찰할 수 있었다. 따라서 다음과 같은 가정을 한다.

- 아래 두 가지 조건을 만족하는 피봇 프레이즈는 올바른 피봇 프레이즈이다.
  ① e_1과 e_2의 내용어 어휘가 일치하는 피봇 프레이즈(c)
  ② e_1과 품사 정보가 일치하는 피봇 프레이즈(c)

위 가정에 근거하여 올바른 피봇 프레이즈 집합을 추출하고, 추출한 피봇 프레이즈 집합 내 빈도 정보에 기반하여 가중치를 부여한다. 예를 들어 그림 5에서 (a)와 (b)는 모두 e_1과 e_2가 동일한 동사인 ‘받’를 포함하고 있으므로 위 가정의 ①번 조건을 만족한다. 하지만 (a)의 피봇 프레이즈는 동사를 포함하고 있으므로, ②번 조건을 만족하는 반면 (b)의 피봇 프레이즈는 동사를 포함하고 있지 않다. 따라서

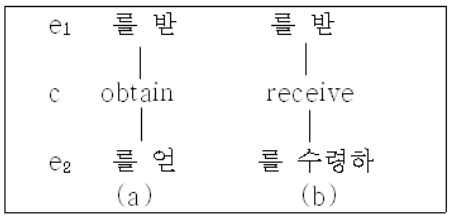

그림 4. 올바른 피봇 프레이즈의 예

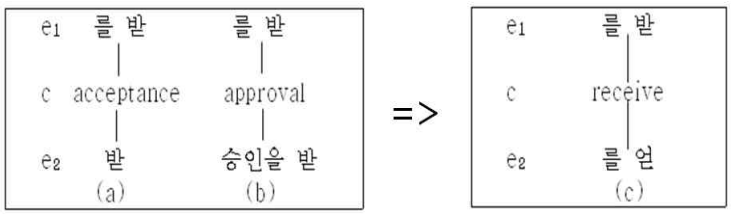

그림 5. A 피봇 집합 예

“를 받”의 패러프레이즈 추출을 위한 좋은 피봇 프레이즈 집합에 (a)의 “receive”는 포함하며, (b)의 “approval”는 포함하지 않는다. (a)의 e_2는 e_1의 패러프레이즈라고 볼 수 없지만 가정의 두 조건을 만족하기 때문에 (c)와 같이 “receive”는 e_1의 올바른 패러프레이즈인 e_2를 추출할 수 있게 해주는 올바른 피봇 프레이즈라 할 수 있다. 가중치를 계산하는 수식은 다음과 같다.

$$W_A(c|e_1) = \frac{count(c|e_1)}{\sum_{c \in A|e_1} count(c|e_1)} \quad (7)$$

이렇게 추출한 피봇 프레이즈의 양이 너무 적어서 자료 부족 문제가 발생할 수 있다. 따라서 평탄화(smoothing)를 하기 위해 위 가정의 조건을 완화한 두 집합을 추가적으로 추출하였다. 이렇게 추출한 두 집합을 B 피봇 집합, C 피봇 집합이라고 한다.

B 피봇 집합은 e_1과 e_2의 내용어 어휘가 일치한 피봇 프레이즈만을 말한다. 예

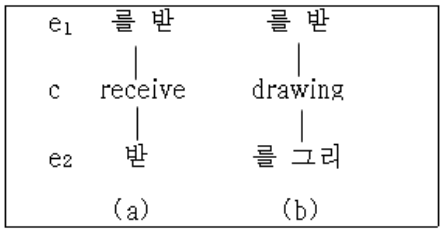

그림 6. B 피봇 집합의 예

를 들어, 그림 6에서 (a)는 e_1과 e_2가 동일한 내용어 어휘인 “받”을 포함하고 있으므로 e_1(“를 받”)의 패러프레이즈 추출 과정에서 피봇 프레이즈 “receive”는 B 피봇 집합에 포함한다. 하지만 그림 6의 (b)는 e_1과 e_2가 동일한 내용어 어휘를 포함하지 않으므로 e_1(“를 받”)의 패러프레이즈 추출 과정에서 피봇 프레이즈 “drawing”은 B 피봇 집합에 포함하지 않는다. 마지막으로 C 피봇 집합은 e_1과 e_2의 다른 정보를 고려하지 않은 모든 피봇 프레이즈를 말한다.

B 피봇 집합, C 피봇 집합을 이용하여, 식(8)과 같이 빈도에 기반하여 가중치 부여를 하였다. 각 집합의 구는 내용어 하나 이상을 가진 것을 대상으로 하고, 단어 원형이 같다면 동일하게 취급하였다. 각 e_1마다 세 개의 피봇 집합은 $A \subset B \subset C$와 같은 포함 관계를 가진다.

$$W_B(c|e_1) = \frac{count(c|e_1)}{\sum_{c \in B|e_1} count(c|e_1)}$$
$$W_C(c|e_1) = \frac{count(c|e_1)}{\sum_{c \in C|e_1} count(c|e_1)} \quad (8)$$

피봇 가중치 계산은 총 3개의 피봇 프레이즈 가중치를 식(9)와 같이 결합하여 구한다.

$$W_{pivot}(c|e_1) = \alpha_1 W_A(c|e_1) + \alpha_2 W_B(c|e_1) + \alpha_3 W_C(c|e_1) \quad (\alpha_1 + \alpha_2 + \alpha_3 = 1) \quad (9)$$

위의 식은 올바른 피봇 프레이즈를 잘 찾는다면 올바른 패러프레이즈를 추출할 수 있다는 가정에 기반한다. 그러나 올바른 피봇 프레이즈를 잘 찾았지만 추가적으로 그 피봇 프레이즈의 대역어 오류의 문제가 있음을 확인할 수 있었다. 예를 들어 그림 7에서 (a)와 (b)는 올바른 피봇 프레이즈를 포함하고 있지만, “receive”의 대역어로 우리는 (a)와 같은 올바른 패러프레이즈 “를 얻”의 결과뿐만 아니라 대역어 오류로 인한 (b)와 같이 잘못된 패러프레이즈 “배정”의 결과를 얻을 수 있다.

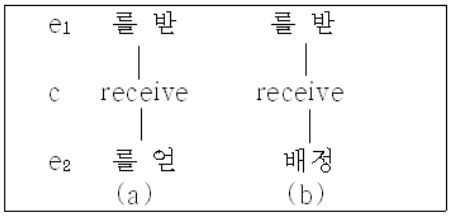

그림 7. “receive”의 대역어 예

그래서 올바른 피봇 프레이즈를 구성하는 A 피봇 집합을 이용하여 추가로 후보 패러프레이즈 가중치를 식(10)과 같이 구한다.

$$W_{cand}(e_2|c) = \beta_1 W_A(e_2|c) + \beta_2 W_{A*}(e_2|c) \quad (\beta_1 + \beta_2 = 1) \quad (10)$$

$W_A(e_2|c)$는 A 피봇 집합에 대한 후보 패러프레이즈의 최대 우도 측정 방법으로 구한다. $W_{A*}(e_2|c)$는 A 피봇 집합과 동일한 품사를 가진 후보 패러프레이즈에 대해 최대 우도 측정으로 구하며, $W_A(e_2|c)$와 $W_{A*}(e_2|c)$를 계산할 때 후보 패러프레이즈는 내용어 한 개 이상을 갖는 패러프레이즈를 대상으로 한다.

$W_A(e_2|c)$와 $W_{A*}(e_2|c)$는 최대 우도 측정에 근거하여 식(11)과 같이 구한다.

$$W_A(e_2|c) = \frac{count(e_2, c)}{\sum_{c \in A} count(e_2, c)}$$
$$W_{A*}(e_2|c) = \frac{count(e_2, c)}{\sum_{c \in A} count(e_2, c)} \quad (11)$$

기존 모델과의 결합

본 논문에서는 이전 연구의 패러프레이즈 확률 모델 (1)과 (3)에 제안하는 두 개의 가중치를 다음과 같이 결합하였다. 가중치를 결합할 수 있는 방법은 다양하며, 사전 실험에서 다음 두 식(12), (13)와 같이 결합할 때 패러프레이즈 추출에 좋은

성능을 주었다.

▪ [1]모델 + 제안하는 방법

$$p(e_2|e_1) = \sum_c (p_{MLE}(c|e_1) p_{MLE}(e_2|c)\, W_{pivot}(c|e_1)\, W_{cand}(e_2|c)) \quad (12)$$

▪ [8]모델 + 제안하는 방법

$$score(e_1|e_2) = \sum_c \left(\exp\left[\sum_{i=1}^{6} \lambda_i h_i(e_1, e_2, c) \right] \right) \quad (13)$$

식(13)에서 $h_5 = W_{pivot}(c|e_1), h_6 = W_{cand}(e_2|c)$이다.

시스템 구성도

본 논문에서 제안하는 시스템은 그림 8과 같이 구성된다. 시스템의 입력으로 영한 병렬 말뭉치가 들어오면 단어/구 정렬 모듈에 의해 구 번역 테이블(phrase-table)

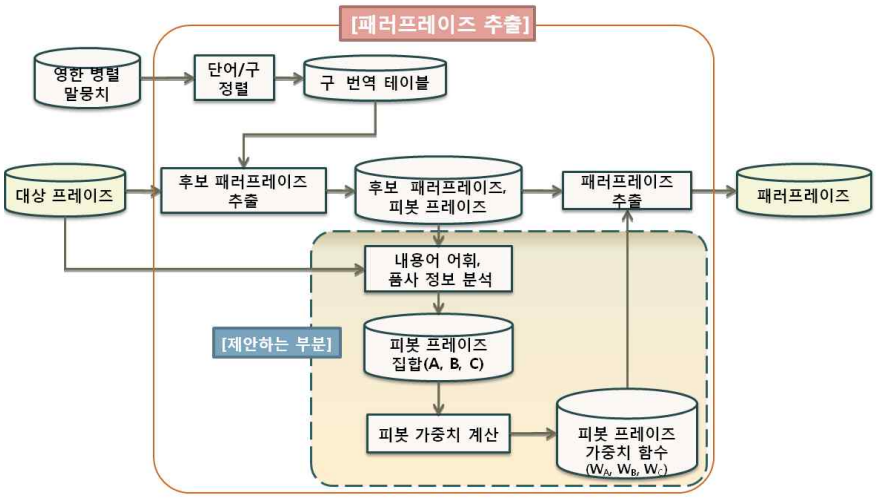

그림 8. 시스템 구성도

이 구축된다. 구축된 구 번역 테이블을 이용하여 입력으로 들어온 찾으려는 대상 프레이즈에 대한 패러프레이즈를 [1]의 가정에 의해 후보 패러프레이즈, 피봇 프레이즈를 추출한다. 추출된 피봇 프레이즈에 대해 대상 프레이즈의 내용어 어휘, 품사 정보를 분석하여 피봇 프레이즈에 대한 가중치를 구한다. 이때 구한 피봇 프레이즈의 가중치를 이용해 패러프레이즈를 재추출하게 된다. 그 방법은 이전과 같이 구 번역 테이블을 이용하되 대상 프레이즈에 대한 패러프레이즈를 추출할 때 피봇 프레이즈 가중치를 반영하여 패러프레이즈를 추출 한다.

실험 및 평가

본 논문에서 제안한 피봇 프레이즈 가중치 방법을 검증하기 위해, 기존 연구인 [1], [8]모델과 제안하는 패러프레이즈 추출 모델의 비교 평가 실험을 하였다. 두 가지 측면에서 비교 평가 하였는데, 첫 번째는 패러프레이즈 추출 결과의 정확률과 재현율 평가이고, 두 번째는 추출된 패러프레이즈의 응용 측면으로서 한영 통계 기계 번역 시스템에 적용했을 때의 번역 성능 평가이다.

패러프레이즈 추출 성능 평가

실험 환경 및 평가 방법

본 논문에서 제안하는 방법의 효과를 실험하기 위해 수동으로 패러프레이즈 여부를 판단하여 정답을 부착한 평가 집합을 구축하였다. 우선 약 50만 문장 쌍 영한 병렬 말뭉치에 통계 기계 번역의 구 정렬 기법을 적용하여 구 번역 테이블을 구축하였다. 그리고 구축한 구 번역 테이블 내에서 빈도 2 이상이고 내용어를 최소 한 개 이상 포함하는 한국어 구 중 75개를 임의로 추출하여 실험 집합으로 선정하였다. 전체 말뭉치에서 대상 프레이즈로 선정한 한국어 구 75개에 대해 [1]의 방법으로 패러프레이즈 후보를 추출한 후 사람이 직접 평가하였다. 총 66,600개의 후보 패러프레이즈 중 올바른 패러프레이즈는 3,290개, 잘못된 패러프레이즈는 60,000개로 평가하였다. 사람이 평가하였을 때 옳고 그름이 불분명한 후보 패러프

레이즈 3,310개는 평가집합에서 제외하였다. 평가 집합 구축 시, 한 사람이 평가를 하였으며 두 사람이 평가 집합에 대해 검증을 하였다.

실험 평가 방법으로 정확률, 재현율, F-measure를 이용하였다.

$$정확률(P) = \frac{시스템이 정확하게 추출한 패러프레이즈 개수}{시스템이 추출한 전체 패러프레이즈 개수}$$

$$재현율(R) = \frac{시스템이 정확하게 추출한 패러프레이즈 개수}{정답 패러프레이즈 개수}$$

$$F-measure = \frac{2PR}{P+R}$$

실험 결과

다음 4개의 패러프레이즈 확률 모델에 대해 실험을 하였다.

- **Bannard :** {1} 모델, 식(1)
- **Bannard+Pivot Weight** : {1} 모델에 피봇 가중치 추가, 식(12)
- **Zhao :** {8} 모델, 식(3)
- **Zhao+Pivot Weight :** {8} 모델에 피봇 가중치 추가, 식(13)

각 모델의 결과로 추출된 후보 패러프레이즈 중 임계치(T) 이상인 값을 갖는 패러프레이즈에 대해 평가를 하였다. 각 모델의 자질 함수의 파라미터 α_i, β_i와 최적 임계치 T는 실험에 의해 결정하였다.

실험 결과는 표 1과 같다. 결과에서 볼 수 있듯이 제안한 방법은 패러프레이즈 추출 성능에 긍정적인 효과를 주었다. Bannard 모델과 Zhao 모델은 피봇 프레이즈의 차별화를 주지 않으므로 인해 잘못된 패러프레이즈가 추출됨을 확인 할 수 있었다.

Bannard 모델에서 추출된 패러프레이즈는 내용어가 하나도 없는 전치사, 조동사 등으로만 구성된 피봇 프레이즈와 많이 정렬됨으로 인해 올바른 패러프레이즈가

표 1. 각 모델에 따른 패러프레이즈 추출 결과

모 델	정확률	재현율	F-measure
Bannard	21.34%	44.71%	28.89%
Bannard+**Pivot Weight**	**33.65%**	**64.07%**	**44.13% (+15.24%p)**
Zhao	44.56%	50.49%	47.34%
Zhao+**Pivot Weight**	**46.81%**	**56.02%**	**51% (+3.66%p)**

추출되지 않았다.

Zhao 모델에서는 단어 단위 가중치 자질을 추가함으로 인해 패러프레이즈의 내용어가 하나도 없는 피봇 프레이즈는 많이 제거 되었다. 그러나 내용어가 포함된 피봇 프레이즈의 오류로 인해 잘못된 패러프레이즈가 추출되는 문제까지는 해결하지는 못하였다. 예를 들어, “를 받”과 “치료”는 동일한 피봇 프레이즈 “treated”와 정렬되어서 “치료”가 “를 받”의 패러프레이즈로 추출되었다. 이는 “를 받”과 “treated”의 대역 확률이 낮아도 “치료”와 “treated”의 대역 확률이 높아서 생긴 오류로 볼 수 있다.

제안하는 피봇 가중치를 두 모델에 결합한 결과, 피봇 프레이즈의 오류를 줄임으로써 보다 양질의 패러프레이즈를 추출할 수 있었다. 또한 제안하는 피봇 가중치 방법이 올바른 패러프레이즈에 높은 점수를 줌으로써 재현율이 기존 모델에 비해 많이 향상 되었다.

패러프레이즈를 이용한 번역 성능 평가

제안하는 방법을 통해 추출한 한국어 패러프레이즈를 응용 측면에서 평가하기 위해 한영 기계 번역 시스템에 활용하는 실험을 수행하였다. [4]에서는 패러프레이즈를 이용해 구 기반 통계 기계 번역 시스템의 필수 요소인 구 번역 테이블의 소스 언어 부분을 확장함으로써 번역 성능을 향상시킬 수 있음을 보였다. 본 절에서의 실험은 [4]에서 제안된 구 번역 테이블 확장 방법을 그대로 사용하여 제안하는 패러프레이즈 추출 방법이 한영 기계 번역 성능 향상에 기여할 수 있는지를 확인

하는 데에 초점을 맞추었다.

실험 환경 및 평가 방법

Baseline 번역 모델로는 구 기반 통계 기계 번역 모델(Phrase-based Statistical Machine Translation, PBSMT)[9]을 채택하였으며, 번역 시스템은 현재 가장 널리 이용되고 있는 Moses 툴킷[10]을 이용하여 구현하였다.

구 번역 테이블을 확장하기 위해 우선 패러프레이즈를 추출할 대상 프레이즈를 선정해야 하는데, 튜닝 데이터와 테스트 데이터의 Baseline 번역 결과를 이용하여 선정하였다. Baseline 번역 결과에서 번역되지 않고 남은 한국어 단어를 모두 추출한 후, 확장 전 테이블에 존재하는 프레이즈 중에 이 단어들을 하나 이상 포함하고 있는 프레이즈를 추출하여 대상 프레이즈로 삼았다. 이렇게 선정된 대상 프레이즈는 3,923개였으며 이 프레이즈들의 패러프레이즈를 추출하여 구 번역 테이블을 확장하였다.

평가는 번역 성능 평가 척도인 BLEU[11]을 사용하였다. BLEU는 시스템 번역 문장과 참조 번역 문장 간에 매칭되는 n-gram의 비율을 통해 시스템의 번역 성능을 측정한다. 또한 번역 결과에서 번역되지 않은 단어의 개수의 비교를 통해 번역 테이블의 확장이 번역률에 미치는 영향을 살펴본다.

실험 결과

다음 패러프레이즈 모델을 이용해 구 번역 테이블을 확장하여 번역 성능 실험을 하였다.

- Zhao: [8] 모델, 식(3)
- Zhao+**Pivot Weight**: [8] 모델에 피봇 가중치 적용, 식(12)

각 패러프레이즈 추출 모델의 임계치(T) 이상을 갖는 패러프레이즈를 구 번역 테이블에 확장하였다. 그리고 [4]에서와 마찬가지로 패러프레이즈 확률값을 번역 모델의 새로운 자질로 추가를 하여 실험하였다.

추출한 한국어 패러프레이즈를 이용해 구 번역 테이블을 확장한 후 한국어에서

표 2. 패러프레이즈를 이용한 번역 성능 결과

번역 모델	패러프레이즈 추출 모델	BLEU	구 번역테이블 크기(번역 쌍 개수)	번역에 실패한 단어(개/문장)
Baseline (PBSMT)	N/A	7.36	777,544개	0.726개
PBSMT+구 번역 테이블 확장	Zhao	7.22	1,110,279개	0.617개
PBSMT+구 번역 테이블 확장	Zhao+**Pivot Weight**	**7.52**	**1,353,207개**	**0.594개**

영어로 번역 실험을 하였다. 실험 결과는 표 2와 같다. 결과에서 볼 수 있듯이, 제안하는 가중치 함수를 결합한 패러프레이즈 추출 모델에 의해 번역에 실패한 단어의 개수는 감소하였으며 번역 성능은 미미하게 향상되었다.

번역에 성공한 단어 개수는 증가하였으나 번역 성능 향상이 미미한 점에 대해 오류 분석을 한 결과, 다음과 같이 오류 유형을 나누어 볼 수 있었다.

번역 성능 평가 방식의 문제. 시스템의 번역 결과와 번역 성능을 평가할 때 사용하는 정답 데이터간의 단어가 일치하지 않아서 번역 성능 향상이 반영되지 않은 경우가 다수 존재하였다.

예를 들어, 그림 9의 “십대”라는 단어를 패러프레이즈를 이용해 “teen”이라고 번역을 하였음에도 불구하고 정답 문장에서는 “teenager”로 번역되어 있어 단순히

실험 데이터	몇 년 전엔 십대들이 주로 인터넷에서 썼지만 요즘엔 오프라인에도 보인다.
정답 데이터	a few year ago, the expression be frequently use online among **teenager,** but these day, we can often hear people say it regarding daily life.
Baseline 번역 모델의 번역결과	a few year ago, **십대** be mostly write on the internet, but nowadays it seem as offline.
패러프레이즈 이용한 확장 모델의 번역결과	a few year ago, **teen** be mostly write on the internet, but these day, offline also be see.

그림 9. 패러프레이즈 효과가 번역 성능에 미반영 된 예

n-gram 매칭으로 계산되는 BLEU 스코어에는 미반영 되었다.

패러프레이즈는 맞게 추출하였으나 번역이 잘못 되는 문제. 구 번역 테이블에 나타나지 않은 단어나 구에 대한 올바른 패러프레이즈를 찾았음에도 불구하고 기존에 있던 구 번역 테이블 번역 쌍의 오류로 인해 번역이 잘못되는 문제도 다수 관찰되었다.

예를 들어, 그림 10을 보면, 패러프레이즈를 이용해 “역전되”라는 단어에 대해 “too”라고 잘못 번역되어 있다. “역전되”의 패러프레이즈로 “뒤집히”를 옳게 찾았으나, 그림 11과 같이 기존 구 번역 테이블에서 “뒤집히”가 “too”로 잘못 번역되어 있다. 따라서 그림 12와 같이 잘못된 번역 쌍이 추가된다.

따라서 올바른 패러프레이즈가 추출되었음에도 불구하고 기존의 구 번역 테이블 번역 쌍의 오류로 인해 잘못된 번역 쌍의 정보가 그대로 확장되어 번역 성능에 도움을 줄 수 없게 됨을 확인할 수 있었다. 이 오류 유형은 [2]에서 제안된 구 번역 테이블 확장 방법에 개선의 여지가 있음을 보여준다고 할 수 있다.

실험 데이터	전통적으로 일본이 더 비쌌던 경유 가격도 최근 역전 되었다.
정답 데이터	all the more surprisingly, the price of light oil in korea recently surpass the price in japan, although the price have be traditionally higher in japan.
Baseline 번역 모델의 번역결과	japan have traditionally be more expensive 역전되 also recently diesel price.
패러프레이즈 이용한 확장 모델의 번역결과	traditionally, japan, the more expensive the price of diesel, which be recently, too.

그림 10. 기존 구 번역 쌍의 오류로 인한 번역 오류

```
뒤집히 ||| too ||| 0.00460829 0.0018282 0.25 0.0857143 2.718
```

그림 11. “뒤집히”의 구 번역 테이블 정보

```
뒤집히 ||| too ||| 0.00460829 0.0018282 0.25 0.0857143 2.718 1
역전되 ||| too ||| 0.00460829 0.0018282 0.25 0.0857143 2.718 0.56
```

그림 12. “역전되” 구 번역 테이블에 추가

결론 및 향후 연구

본 논문에서는 이중 언어 병렬 말뭉치에서 패러프레이즈를 추출하는 연구를 수행하였으며, 피봇 프레이즈에 중점을 두어 피봇 프레이즈 자체에 차별화를 통해 패러프레이즈 추출 성능을 높이고자 하였다. 기존의 방법들은 피봇 프레이즈에도 옳고 그름에 대한 차별을 주지 않았으며, 피봇 프레이즈의 오류가 패러프레이즈 추출의 오류를 가져다 준다는 것을 고려하지 못하였다.

본 연구에서는 패러프레이즈 추출에 사용되는 피봇 프레이즈들 중에 옳은 것과 잘못된 것이 있다는 것을 고려하여 올바른 피봇 프레이즈에 높은 가중치를 부여하는 방법을 제안하였다. 먼저 올바른 피봇 프레이즈의 요건에 대한 가설을 세우고 이를 바탕으로 기존의 구 정렬 정보뿐만 아니라 피봇 프레이즈, 후보 패러프레이즈의 어휘와 품사 정보를 추가적으로 이용하여 각 피봇 프레이즈의 가중치를 계산하는 함수를 고안하였다. 추가적인 정보 사용으로 인한 자료 부족 문제를 완화하기 위해 어휘, 품사 정보 이용 방법을 세 가지로 구성함으로서 다양한 피봇 프레이즈들의 가중치를 좀 더 정확하게 계산할 수 있도록 하였다. 한편 피봇 프레이즈가 올바른데도 불구하고 구 정렬 오류로 인해 최종 패러프레이즈가 잘못 추출되는 문제를 완화하기 위해 패러프레이즈에도 동일한 방법을 적용하여 가중치를 부여하였다.

제안한 두 가지 가중치 방법을 기존의 패러프레이즈 추출 모델에 결합한 결과 기존 모델에 비해 성능이 향상됨을 확인할 수 있었다. 또한 제안한 방법의 실제적인 유용성을 평가하기 위해 본 방법을 번역 시스템에 적용한 실험에서도 번역률 및 번역 품질이 향상됨을 보였다.

향후에는 더 견고한 피봇 프레이즈를 구성하기 위해 다른 외부 자원인 사전이나 웹을 이용해 볼 것이다. 또한, 본 논문에서는 패러프레이즈 구 추출 방법에서만 피

봇 가중치 방법을 적용하였지만 패러프레이즈 패턴 추출에도 적용하여 평가해 볼 것이다.

참고문헌

Bannard, C. and Callison-Burch, C. (2005). “Paraphrasing with Bilingual Parallel Corpora”, In Proceedings of the Annual Meeting of the ACL, Ann Arbor, MI, pp. 597-604.

Nakov, P. (2008). “Improved Statistical Machine Translation Using Monolingual Paraphrases”, In Proceedings of the European Conference on Artificial Intelligence (ECAI), pp. 338-342.

Bond. F., Nichols, E,, Scorr Appling, D. and Paul, M. (2008). “Improving Statistical Machine Translation by Paraphrasing the Training Data”, In Proceedings of the International Workshop on Spoken Language Translations (IWSLT), pp. 150-157.

Callison-Burch, C., Koehn, P. and Os-borne, M. (2006) “Improved Statistical Machine Translation Using Paraphrases”, In Proceedings of the Human Language Technology conference North American chapter of the Association for Computational Linguistics (HLT-NAACL), pp. 17-24.

Marton, Y., Callison-Burch, C. and Resnik, P. (2009). “Improved Statistical Machine Translation Using Monolingually Derived Paraphrases”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 381-390.

Barzilay, R. and McKeown, K. R. (2001). “Extracting Paraphrases from a Parallel Corpus”, In Proceedings of the Annual Meeting of the ACL, Toulouse, France, pp. 50-57.

Zhao, S., Lan, X., Liu, T. and Li, S. (2009). “Application Driven Statistical Paraphrase Generation”, In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 2-7, pp. 834 – 842.

Josef Och, F. and Ney. H. (2001). “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation”, In proceedings of the Annual Meeting of the

ACL.

Zhao, S., Wang, H,, Liu, T. and Li, S. (2009). “Extracting Paraphrase Patterns from Bilingual Parallel Corpora”, NLE. Vol. 15. No. 4. pp. 503-526.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran Richard Zens, C., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007). “Moses: Open Source Toolkit for Statistical Machine Translation”, In Proceedings of the 45th Annual Meeting of the ACL, Poster and Demonstration Sessions, pp. 177-180.

Papineni, Roukos, S., Ward, T. and Zhu, W. J. (2002). “BLEU: A Method for Automatic Evaluation of Machine Translation”, In Proceedings of the Annual Meeting of the ACL, pp. 311-318.

1 차원고접수 : 2011. 1. 24
2 차원고접수 : 2011. 3. 5
최종게재승인 : 2011. 3. 7

(*Abstract*)

Pivot Discrimination Approach for Paraphrase Extraction from Bilingual Corpus

Esther Park Hyoung-Gyu Lee Min-Jeong Kim Hae-Chang Rim

Dept. of Computer and Radio Communications Engineering, Korea University

Paraphrasing is the act of writing a text using other words without altering the meaning. Paraphrases can be used in many fields of natural language processing. In particular, paraphrases can be incorporated in machine translation in order to improve the coverage and the quality of translation. Recently, the approaches on paraphrase extraction utilize bilingual parallel corpora, which consist of aligned sentence pairs. In these approaches, paraphrases are identified, from the word alignment result, by pivot phrases which are the phrases in one language to which two or more phrases are connected in the other language. However, the word alignment is itself a very difficult task, so there can be many alignment errors. Moreover, the alignment errors can lead to the problem of selecting incorrect pivot phrases. In this study, we propose a method in paraphrase extraction that discriminates good pivot phrases from bad pivot phrases. Each pivot phrase is weighted according to its reliability, which is scored by considering the lexical and part-of-speech information. The experimental result shows that the proposed method achieves higher precision and recall of the paraphrase extraction than the baseline. Also, we show that the extracted paraphrases can increase the coverage of the Korean-English machine translation.

Key words : Paraphrase, Paraphrase extraction, Pivot discrimination