인지과학, 제17권 제4호, 2006, 323~336.

정렬된 성경 코퍼스로부터 바꿔쓰기표현(paraphrase)의 자동 추출*

Automatic Extraction of Paraphrases from a Parallel Bible Corpus

이 공 주** (Kong Joo Lee)　　윤 보 현*** (Bo Hyun Yun)

요 약 바꿔쓰기(paraphrasing)는 동일한 내용을 다르게 표현하는 방식을 의미한다. 이러한 바꿔쓰기표현들(paraphrases)은 기계번역, 질의 응답 시스템, 문서 요약과 같은 다양한 분야에 매우 유용하게 사용될 수 있다. 그러나 이와 같은 바꿔쓰기표현의 유용성에도 불구하고 바꿔쓰기표현을 자동으로 추출할 수 있는 방법이 매우 어렵다. 우선 바꿔쓰기표현을 자동으로 추출할 수 있는 데이터를 구하는 것부터가 어려운 문제이다. 본 연구에서는 여러 버전의 한글 성경 코퍼스로부터 바꿔쓰기표현을 자동으로 추출해 보고자 한다. 성경은 각 문장이 절과 구로 나누어져 있어 문장과 문장을 정렬시키는 것이 매우 용이하다. 정렬된 여러 버전의 성경 코퍼스로부터 자율학습(unsupervised learning)을 통해서 자동으로 바꿔쓰기표현을 추출한다. 이와 같은 방법은 어휘수준의 바꿔쓰기표현뿐만 아니라 구문수준의 바꿔쓰기표현도 추출할 수 있음을 보여준다.

주제어 바꿔쓰기, 바꿔쓰기표현, 공동학습, 정렬코퍼스

Abstract In this paper, we present a pilot system that can extract paraphrases from a parallel corpus using co-training method. Paraphrases are useful for the applications that should create a varied and fluent text, such as machine translation, question-answering system, and multidocument summarization system. One of the difficulties in extracting paraphrases is to find a rich source from which we can extract paraphrases. The bible is one of the good sources for extracting paraphrases as it has several Korean versions in which every sentence can be easily aligned by the chapter and the verse. We can extract not only the lexical-level paraphrases but also the phrasal-level paraphrases from the parallel corpus which consists of the bibles using co-training method.

Keywords paraphrasing, paraphrases, co-training, parallel corpus.

---

* 이 논문은 2005년도 충남대학교 학술연구비의 지원에 의하여 연구되었음.

** 교신저자: 이공주, 충남대학교 전기정보통신공학부
305-764 대전역시 유성구 궁동 220, E-mail: kjoolee@cnu.ac.kr
*** 목원대학교 컴퓨터 교육과

서 론

바꿔쓰기 또는 바꿔 말하기(paraphrasing)는 동일한 정보나 내용을 다른 방식으로 표현하는 것을 의미한다[1]. 기계번역이나 문서 요약, 질의 응답시스템과 같은 문장을 생성할 필요가 있는 응용분야에서 이와 같은 바꿔쓰기표현(paraphrase)은 다양한 문장을 생성할 수 있게 해 줄 수 있는 유용한 수단이다. 그러나 바꿔쓰기표현들을 수동으로 구축하는 일은 많은 시간과 인력이 소모되는 작업이며 응용분야가 달라질 때마다 다시금 작성해야 한다는 단점을 갖고 있다. 우선 바꿔쓰기표현을 자동으로 구축하기 위해서는 바꿔쓰기표현을 추출할 수 있는 대량의 원본 데이터가 필요하다. 바꿔쓰기표현을 추출할 수 있는 원본 데이터로는 워드넷과 같은 대량의 어휘사전이나 잘 구축된 시소러스가 가능하다. 그러나 잘 구축된 한국어 어휘사전이나 시소러스를 구하기도 어려울 뿐더러 여기서 추출할 수 있는 바꿔쓰기표현은 주로 동의어(synonyms)에 국한된다[2]. 이에 대한 대안으로는 동일한 내용의 다양한 표현들이 자주 등장할 수 있는 대량의 코퍼스로부터 바꿔쓰기표현을 자동으로 추출하는 것이다. 이러한 데이터로는 다양한 한국어 번역본이 있는 외국 소설들이 가장 유용한 데이터로 사용될 수 있다.

성경은 오랜 세월을 거쳐 여러 차례 서로 다른 버전으로 한글 번역본이 만들어져 왔다. 즉 성경에는 동일한 내용에 대한 여러 다른 버전의 표현들이 존재하는 것이다. 또한 성경은 모든 문장들이 장과 절로 적절하게 나뉘어져 있어서 바꿔쓰기표현을 추출해 볼 수 있는 더할 나위없이 좋은 자료이다. 본 연구에서는 이와 같은 성경 데이터를 이용하여 바꿔쓰기표현을 자동으로 추출해 보고자 한다.

(그림 1)은 5개의 서로 다른 성경으로부터 추출한 레위기 15장 12절의 문장(들)이다. 동일한 내용을 담고는 있지만 표현이 서로 다르기 때문에 이와 같이 정렬된 문장으로부터 우리는 어렵지 않게 '목기'와 '나무그릇'이라는 동일표현을 추출할 수 있다. '질그릇'과 '오지그릇'이 유사하게 사용될 수 있다는 것도 쉽게 알 수 있다. 또한 '유출병'의 의미를 정확히는 모르지만 '유출병이 있다'와 '고름을 흘리다'가 유사한 의미로 사용된 표현임을 알 수 있다.

본 연구에서는 이미 잘 알려져 있는 공동학습법(co-training)을 이용하여 정렬된 성경 코퍼스로부터 자동으로 바꿔쓰기표현을 추출해보고자 한다. 본 연구에서 사용하게 될 공동학습법은 2001년 Regina Barzilay와 Kathleen R.

"유출병 있는 자의 만진 질그릇은 깨뜨리고 목기는 다 물로 씻을찌니라."
"유출병이 있는 자가 만진 질그릇은 깨뜨리고 나무 그릇은 다 물로 씻을지니라."
"고름을 흘리는 사람의 몸이 닿은 오지그릇은 깨뜨려야 하고, 나무그릇은 물로 씻어야 한다."
"고름을 흘리는 남자가 만진 오지그릇은 깨뜨려 버려야 한다. 그가 만진 것이 나무그릇일 때에는 모두 물로 씻어야 한다."
"고름을 흘리는 남자가 오지그릇을 만지면, 그 오지그릇은 깨뜨려라. 만약 그가 나무그릇을 만지면 그 그릇은 물에 씻어라."

(그림 1) 성경에서 추출한 정렬된 문장들

McKeown에서 제안한 방법[1]이며, 이 방법을 그대로 한국어 성경 코퍼스에 적용해 보고자 한다. 이 공동학습법에서는 서로 바꿔쓸 수 있는 표현들은 유사한 내지는 동일한 문맥정보를 갖고 있다는 가정하에 학습을 시작한다. 즉 바꿔쓰기표현을 추출할 수 있는 좌우 문맥정보를 먼저 학습하고, 학습된 문맥정보를 이용하여 바꿔쓰기표현들을 추출한다. 추출된 바꿔쓰기표현을 이용하여 문맥정보를 다시 학습하고 이를 이용하여 바꿔쓰기표현을 다시 추출하는 과정을 반복함으로써 최종의 바꿔쓰기표현들을 추출할 수 있다. 이와 같은 학습방법은 형태소 분석 결과만을 이용함에도 불구하고 단순한 어휘수준의 바꿔쓰기표현뿐만 아니라 구절단위의 바꿔쓰기표현도 추출할 수 있다는 장점을 갖고 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 관련연구를 간단히 살펴보고, 3장에서 데이터 추출을 위한 원본 코퍼스의 특성을 살펴본다. 4장에서 바꿔쓰기표현 추출 알고리즘을 설명하고 5장에서 실험결과로 추출된 바꿔쓰기표현들과 그 특징들, 그리고 이에 대한 간략한 평가를 수행한다. 마지막으로 6장에서 결론을 맺고자 한다.

관련 연구

일반적으로 바꿔쓰기표현의 자동 추출은 동일한 의미를 갖고 있는 여러 개의 문장으로 구성된 병렬 형태의 코퍼스로부터 단어 수준 혹은 구문 수준의 바꿔쓰기표현을 다양한 방법을 통해서 추출한다.

본 연구는 참고문헌 [1]의 연구를 기본으로 동일한 아이디어를 한국어 코퍼스에 적용해 보았다. 참고문헌 [1]의 연구에서는 외국소설의 다양한 영어 번역본을 기반으로 병렬 코퍼스를 구성하였으며 그렇기 때문에 여러 번역본들 사이의 문장단위의 정렬화 작업이 요구된다. 공동학습 방법을 이용하여 바꿔쓰기표현을 추출할 수 있는 문맥 정보와 바꿔쓰기표현 자체를 번갈아가며 학습하였다. 결과적으로 9,483개의 어휘수준의 바꿔쓰기표현과 25개의 형태-구문수준의 바꿔쓰기표현을 추출할 수 있었다.

참고문헌 [3]의 연구에서는 기계번역의 평가를 위해 작성된 중국어-영어 코퍼스를 이용하여 바꿔쓰기표현을 자동으로 추출하였다. 이 연구에서는 바꿔쓰기표현을 FSA(또는 word lattices)로 표현하였기 때문에 여러 개의 바꿔쓰기표현을 매우 효과적으로 표현할 수 있으며, 도출된 FSA로부터는 학습되지 않은 바꿔쓰기표현까지도 유도해 낼 수 있었다. 동일한 의미를 지니고 있는 문장 그룹을 모두 구문분석한 후, 결과로 나온 각각의 구문트리를 서로 병합하여 parse-forest를 만들어 낸다. 동일한 의미를 지닌 문장들을 모두 구문분석 하였기 때문에 기본적으로 주어의 NP들은 서로 바꿔 쓸 수 있으며, 서술어에 해당하는 VP들은 서로 바꿔 쓸 수 있다. 이와 같은 방법은 단순한 단어 수준의 바꿔쓰기표현뿐 아니라 좀더 일반화된 형태의 바꿔쓰기까지 추출할 수 있다는 장점을 갖고 있다. 반면에 바꿔쓰기표현 추출의 성능이 구문분석기의 분석능력과 정확도에 의존적일 수밖에 없다는 단점을 갖고 있다.

참고문헌 [4]의 연구에서는 바꿔쓰기표현이 완전 영역 독립적인 문제일 수 없기 때문에 바꿔쓰기표현의 추출을 단순 단어 단위나 구절 단위가 아닌 문장 단위로 추출하고자 하였

다. 이 연구에서는 동일한 사건(event)에 대해 서로 다른 신문사에서 제공한 기사로 구성된 비교 코퍼스(comparable corpus)를 사용하여 문장단위의 바꿔쓰기표현을 추출하였다. 우선 각각의 코퍼스에서 추출한 신문기사의 각 문장들을 유사한 표현이 사용된 문장들로 군집화(clustering)시킨다. Multiple Sequence Alignment 방법을 이용하여 군집화된 문장에서 단어 격자체 (word lattice)를 형성한다. 단어 격자체에서 지나치게 구체적인 표현들을 슬롯 (slot)으로 일반화시킨다. 각 코퍼스로부터 단어 격자체를 형성하고 난 후, 다른 비교 코퍼스의 단어 격자체들과 비교하면서 바꿔쓰기표현으로 사용할 수 있는 단어 격자체를 서로 연결시킨다. 이렇게 함으로써 문장단위의 바꿔쓰기표현을 단어 격자체의 연결(mapping)로 표현할 수 있게 되며 이를 이용하면 자연스러운 바꿔쓰기 문장을 생성해 낼 수 있다. 그러나 이 방법은 문장수준의 문형에 대한 바꿔쓰기만을 제공할 수 있으며 어휘수준의 바꿔쓰기표현은 다루고 있지 못하다.

정렬된 코퍼스

본 연구에서 사용하는 코퍼스는 한국어 성경 코퍼스이다. 성경 코퍼스 중에는 문체가 다소 생소하고 자주 사용되지 않는 표현이 많다는 단점이 있으나, 반면에 각 문장이 장과 절로 나누어져 있어 정렬화시키는 것이 매우 용이하다는 장점이 있다. 한글로 번역된 성경은 번역된 시기와 번역 주체에 따라 여러 개의 성경 번역본이 존재하다. 본 연구에서는 그 중 가장 많이 사용되고 있는 5개의 성경을 정렬된 코퍼스로 사용하였다[5].

(A) 개역한글
(B) 개역개정
(C) 공동번역
(D) 표준새번역
(E) 쉬운성경(아가페)

5개의 성경번역본 중, (B)는 (A)을 개정한 것으로 5개의 성경 중 (A)와 (B)가 서로 가장 유사하다. 본 연구에서 조사해본 결과 (B)의 개역개정판은 (A)의 개역한글판과 69.73%의 어절이 완전히 동일했다. (개역개정에 모두 475,626 어절이 사용되었으며 이중 331,684의 어절이 개역한글판과 완전히 동일했다.) (A)와 (B)의 성경 번역본은 만연체 문장, 고어체 어투, 어려운 한자어의 사용으로 다소 이해하기 어려운 부분이 많다고 알려져 있다. 그에 비해 (C), (D), (E)는 (A)와 (B)에 비해서는 쉬운 현대말을 사용하여 번역되었다.

본 연구에서 사용한 정렬된 코퍼스에는 총 151,508 문장이 포함되어 있으며 정렬된 문장 집합으로는 모두 31,199 개의 문장집합이다. 전체 단어수는 2,472,238 단어이며, 총 고유단어 개수는 191,509 단어이다. (그림 2)는 정렬된 코퍼스의 일부를 보인다.

바꿔쓰기표현의 자동 추출

본 연구에서는 두 개의 구절(phrase)과 연관된 주변 문맥(context)이 충분히 유사하다면 이 두 구절은 바꿔쓰기표현일 확률이 높다는 가정하에 자동 추출을 시도하고자 한다. 여기서 사용될 수 있는 주변 문맥으로는 용언-목적어 (verb-object) 관계 또는 명사-수식어(noun-modifier) 관계 등이 사용될 수 있으나, 이 같은 관계를

고린도전서-13-5-A	무례히 행치 아니하며 자기의 유익을 구치 아니하며 성내지 아니하며 악한 것을 생각지 아니하며
고린도전서-13-5-B	무례히 행하지 아니하며 자기의 유익을 구하지 아니하며 성내지 아니하며 악한 것을 생각하지 아니하며
고린도전서-13-5-C	사랑은 무례하지 않습니다. 사랑은 사욕을 품지 않습니다. 사랑은 성을 내지 않습니다. 사랑은 앙심을 품지 않습니다.
고린도전서-13-5-D	사랑은 무례하지 않으며, 자기의 이익을 구하지 않으며, 성을 내지 않으며, 원한을 품지 않습니다.
고린도전서-13-5-E	사랑은 무례히 행동하지 않습니다. 사랑은 자기 유익을 구하지 않습니다. 사랑은 쉽게 성내지 않습니다. 사랑은 원한을 품지 않습니다.
시편-101-7-A	거짓 행하는 자가 내 집 안에 거하지 못하며 거짓말 하는 자가 내 목전에 서지 못하리로다
시편-101-7-B	거짓을 행하는 자는 내 집 안에 거주하지 못하며 거짓말하는 자는 내 목전에 서지 못하리로다
시편-101-7-C	사기 행각 일삼는 자, 내 집에 살지 못하고 거짓을 말하는 자, 내 앞에 서지 못하리니,
시편-101-7-D	속이는 자는 나의 집에서 살지 못하게 하며, 거짓말하는 자는 내 앞에 서지 못하게 하렵니다.
시편-101-7-E	속이는 사람들은 그 누구도 내 집에서 살지 못할 것입니다. 거짓말하는 사람들도 내 앞에 서지 못할 것입니다.
사무엘상-30-3-A	다윗과 그의 사람들이 성에 이르러 본즉 성이 불탔고 자기들의 아내와 자녀들이 사로잡혔는지라
사무엘상-30-3-B	다윗과 그의 사람들이 성읍에 이르러 본즉 성읍이 불탔고 자기들의 아내와 자녀들이 사로잡혔는지라
사무엘상-30-3-C	다윗의 일행이 성에 이르러서 보니 성은 불타고 있었고 아내와 아들딸들은 이미 사로잡혀 간 뒤였다.
사무엘상-30-3-D	다윗이 부하들을 거느리고 그 성읍으로 들어와 보니, 성은 불타 버렸고, 아내들과 아이들이 모두 사로잡혀 갔다.
사무엘상-30-3-E	다윗과 그의 부하들이 시글락에 와서 보니, 마을은 불타 버렸고 그들의 아내들과 아들딸들이 포로로 끌려갔습니다.
레위기-15-12-A	유출병 있는 자의 만진 질그릇은 깨뜨리고 목기는 다 물로 씻을찌니라
레위기-15-12-B	유출병이 있는 자가 만진 질그릇은 깨뜨리고 나무 그릇은 다 물로 씻을지니라
레위기-15-12-C	고름을 흘리는 사람의 몸이 닿은 오지그릇은 깨뜨려야 하고, 나무그릇은 물로 씻어야 한다.
레위기-15-12-D	고름을 흘리는 남자가 만진 오지그릇은 깨뜨려 버려야 한다. 그가 만진 것이 나무그릇일 때에는 모두 물로 씻어야 한다.
레위기-15-12-E	고름을 흘리는 남자가 오지그릇을 만지면, 그 오지그릇은 깨뜨려라. 만약 그가 나무그릇을 만지면 그 그릇은 물에 씻어라.

(그림 2) 정렬된 코퍼스 일부

추출하기 위해서는 입력 문장에 대한 구문관계를 파악해야 한다. 본 연구에서는 의미적으로 동일한 정렬된 문장들로부터 바꿔쓰기표현을 추출하기 때문에 구문관계의 분석없이 바꿔쓰기표현을 추출할 수 있다[1].

공동학습(co-training)

본 연구에서는 여러 분야에서 이미 많이 사용되고 있는 공동학습(co-training) 방법을 사용하여 바꿔쓰기표현을 자동으로 추출하고자 한다. 이 방법은 이미 단어의미 중의성 해소[6], 개체명 분류(named entity classification)[7] 등에 적용하여 좋은 결과를 보임을 입증하였다.

본 연구에서는 바꿔쓰기표현 추출에 [7]에 의해 제안된 DL-CoTrain 알고리즘을 기반으로 사용하고자 한다. 이 방법은 학습데이터에 대해 두 개의 서로 독립적인 속성 집합에 대한 분류기(classifier)를 반복적으로 학습해 나가는 방법이다. 본 연구의 경우 두 개의 속성 집합 중 하나는 바꿔쓰기표현 그 자체이며, 다른 하나는 바꿔쓰기표현 주변의 양쪽 문맥이다. 우선 바꿔쓰기표현을 추출할 수 있는 양쪽 문맥정보 추출기를 학습하고, 문맥정보를 이용하여 바꿔쓰기표현을 추출한다. 이 과정을 연속적으로 반복함으로써 바꿔쓰기표현을 점증적으로 추출할 수 있다.

본 연구의 바꿔쓰기표현자동 추출 알고리즘은 세 단계로 구성되어 있다. 첫째 초기화 단계, 둘째 문맥정보 추출기를 학습하는 단계, 셋째 바꿔쓰기표현 추출기를 학습하는 단계로 구성되며, 둘째 셋째 단계가 반복적으로 학습과정을 수행한다.

학습데이터는 5개의 성경으로부터 추출하여 정렬된 문장들로 구성되어 있다. 이러한 5개의 정렬된 문장에 대해 어떤 쌍에 대해 자동추출을 수행하느냐에 따라 추출되는 바꿔쓰기표현이 다르게 된다. 그렇기 때문에 5개의 집합에 대해 한꺼번에 수행하지 않고 5개 중에서 2개씩 쌍을 이루어 따로 자동추출을 수행하고자 한다. 즉, 자동추출 알고리즘을 10개의 쌍에 대해 각각 독립적으로 수행하였다.

초기화

정렬된 두 문장에서 서로 공통으로 발생되는 어절을 이용해서 초기 문맥 규칙을 만든다. (그림 3)의 문장1은 개역개정(B)에서 문장2는 표준새번역(D)에서 추출한 고린도전서 13장 5절의 문장이다. (그림 3)에서와 같이 정렬된 두 문장의 경우 공통으로 나타나는 어절 '자기의'를 이용해서 문맥정보 추출기 학습을 위한 초기 학습예제를 구할 수 있다. 문맥정보 추출기 학습을 위해서는 정례(positive) 학습예제뿐만 아니라 반례(negative) 학습예제도 필요하다. 반례 학습예제의 경우에는 정렬된 두 문장에서 공통으로 발생된 단어쌍의 각 단어에 대해 발생할 수 있는 모든 다른 단어쌍으로 구축한다. 그렇기 때문에 n개의 어절을 갖고 있는 문장과 m개의 어절을 갖고 있는 문

문장1: 무례히 행하지 아니하며 **자기의** 유익을 구하지 아니하며 성내지 아니하며 악한 것을 생각하지 아니하며

문장2: 사랑은 무례하지 않으며, **자기의** 이익을 구하지 않으며, 성을 내지 않으며, 원한을 품지 않습니다.

(그림 3) 정렬된 두 문장 (고린도전서 13장 5절)

정례 학습예제	word1 = '자기의' word2 = '자기의'	
반례 학습예제	word1 = '자기의' word2 = '사랑은'	word1 = '무례히' word2 = '자기의'
	word1 = '자기의' word2 = '무례하지'	word1 = '행하지' word2 = '자기의'
	word1 = '자기의' word2 = '않으며,'	word1 = '아니하며' word2 = '자기의'
	...	...
	word1 = '자기의' word2 = '않습니다'	word1 = '아니하며'word2 = '자기의'

(그림 4) 정렬된 두 문장(그림 3)에서 추출한 초기화 학습예제

장이 정렬되어 있을 때, 하나의 동일한 단어쌍에 대해 1개의 정례 학습예제와 (n-1)+(m-1)개의 반례 학습예제를 얻을 수 있다. (그림 4)에서는 (그림 3)의 정렬된 문장 중 "자기의" 어절로부터 추출할 수 있는 학습예제를 보이고 있다.

문맥정보 추출기 학습

문맥정보 추출기 학습을 위해 4.2절에서 얻어낸 초기 학습예제의 양쪽 문맥정보를 수집한다. 문맥정보를 추출하기 위해서 우선 정렬된 문장을 형태소 분석과 품사 태깅을 통해 품사 나열로 재구성하였다 사용된 품사집합은 21세기 세종계획 국어 기초자료구축 시 사용된 품사집합을 사용하였다. 부록에 사용한 전체 품사집합을 담았다. 문맥정보로는 초기화 학습 예제의 오른쪽과 왼쪽 각각의 한 어절을 추출하였다. 문맥정보로는 양쪽 어절의 어휘정보 자체를 추출하는 것이 가장 바람직할 것이나 데이터 부족현상(data sparseness)으로 인해 실질어는 어휘정보 대신 품사정보로 추상화시켜 추출하였다. 조사나 어미와 같은 기능어에 대해서는 간단한 이형태들('이'와 '가', '은'과 '는' 등)만을 표준화시킨 후 어휘정보 자체로 추출하였다.

(그림 3)의 경우 정례 학습예제 '자기의'에서 추출할 수 있는 문맥정보는 문장1에서 "아니하며 자기의 유익을"에 해당하는 (왼쪽문맥1='VA+며' 오른쪽문맥1='NNG+을')와 문장2

<표 1> 추출된 문맥정보 일부

	왼쪽문맥1	오른쪽문맥1	왼쪽문맥2	오른쪽문맥2
(1)	VV+려	EOS	VV+려	EOS
(2)	BOS	VV+ㅏ	BOS	VV+서는
(3)	NNG+께서	VV+셨다+SP	NNG+XSN+께서	VV+셨다+SP
(4)	NNG+을	NNG+니	NNG+을	NNG+니
(5)	NNG+가	NP+에게	NNG+가	NP+에게
(6)	NNG+에	MM	NNG+에	MM
(7)	VA+며	VA+며	VA+며+SP	VA+며+SP

에서 "않으며, 자기의 이익을"에 해당하는 (왼쪽문맥2='VA+며+SP' 오른쪽문맥2='NNG+을') 이다.

수집한 모든 문맥에 대해 그 문맥이 바꿔쓰기표현을 추출할 수 있는 유용한 문맥정보인지 아닌지를 판단하기 위해서 문맥정보의 중요도를 다음과 같이 계산한다. 문맥 x의 전체 출현 빈도를 count(x)로 표시하고, 정례 학습예제의 문맥으로 발생한 빈도를 count(x+)로 나타낼 때, 문맥정보 x의 중요도를 count(x+)/count(x)로 계산하였다. 본 연구에서는 문맥정보 x의 절대 빈도와 문맥정보 x의 중요도를 함께 고려하여 바꿔쓰기표현을 추출할 수 있는 유용한 문맥정보를 결정하였다. 문맥정보 x의 절대빈도가 20회 이상 발생해야 하며 중요도 값이 85% 이상인 경우에 대해서만 바꿔쓰기표현을 추출할 수 있는 문맥정보로 추출하였다. <표 1>은 추출된 문맥정보의 일부이다. ('BOS'는 문장시작(Beginning Of Sentence)이며, 'EOS'는 문장끝(End Of Sentence)이다.)

바꿔쓰기표현 추출기 학습

4.3절의 문맥정보 추출기로부터 추출한 유용한 문맥정보를 정렬된 코퍼스에 적용하여 바꿔쓰기표현을 추출할 수 있다. (그림 3)의 예제로부터 4.3 절의 <표 1>에서 추출된 문맥정보 (7)을 이용하여 ('성내지', '성을 내지')와 같은 바꿔쓰기표현을 추출할 수 있다. 문맥정보를 적용하는 방법에 따라 한 어절뿐만 아니라 여러 어절로 구성된 바꿔쓰기표현도 추출할 수 있다. 본 연구에서는 정렬된 문장에 대해 왼쪽 문맥과 오른쪽 문맥을 한 어절에서 네 어절까지의 거리를 두고 적용함으로

<표 2> 추출된 바꿔쓰기표현의 단순화된 형태

추출된 바꿔쓰기표현 원형	단순화된 바꿔쓰기표현
사망/NNG+의/JKG ↔ 죽음/NNG+의/JKG	사망 ↔ 죽음
여호와/NNG+께서/JKS ↔ 주/NNG+님/XSN+께서/JKS	여호와 ↔ 주님
내/NP+가/JKS 진실/NNG+로/JKB ↔ 내/NP+가/JKS 진정/NNG+로/JKB	진실 ↔ 진정
숫양/NNG 한/MMN마리/NNB+이/VCP+며/EC ↔ 수양/NNG하나/NNG+이/VCP+며/EC	숫양 한 마리 ↔ 수양 하나
유출병이/unknown 있/VV+는/ETM ↔ 고름/NNG+ㄹ/JKO 흘리/VV+는/ETM	유출병이 있다 ↔ 고름을 흘리다
부활/NNG+이/JKS 없/VA+다/EC 하/VV+ㄴ/ETM ↔ 부활/NNG+이/JKS 없/VA+다고/EC 주장하/VV+ㄴ/ETM	VA+다 하다 ↔ VA+다고 주장하다
살/VV+ㅆ/EP+던/ETM 것/NNB+이/VCP+다/EF+./SF ↔ 살/VV+ㅆ/EP+다/EF+./SF	VV+EP+던 것이다 ↔ VV+EP+다
살리/VV+ㅓ/EC 주/VV+소서/EF+./SF ↔ 살리/VV+ㅓ/EC 주/VV+십시오/EF+./SF	VV+소서 VV+십시오 ↔
왕/NNG+이/JKS 시바/NNP+에게/JKB ↔ 왕/NNG+이/JKS 시바/NNP+더러/JKB	NNP+에게 ↔ NNP+더러

써 최대 네 어절까지의 바꿔쓰기표현을 추출할 수 있다. 한국어의 경우 한 어절은 보통 한 개 이상의 형태소로 구성되기 마련이다. 예를 들어, 추출된 바꿔쓰기표현이 '사망/NNG+의/JKG'와 '죽음/NNG+의/JKG'일 경우, 서로 공통되는 최대한의 형태소를 제거하고 '사망'과 '죽음'만을 단순화된 바꿔쓰기표현으로 추출하였다. 여러 어절로 구성된 바꿔쓰기표현의 경우에도 공통되는 최대한의 형태소를 제거한다. <표 2>는 단순화된 바꿔쓰기표현의 일부이다.

이와 같이 추출된 바꿔쓰기표현은 다시 4.3절의 문맥정보 추출기에 사용되어 새로운 문맥정보를 만들어 내고, 새로이 만들어진 문맥정보를 이용하여 바꿔쓰기표현을 추출하는 과정을 반복하게 된다. 새로운 바꿔쓰기표현이 더 이상 추가되지 않을 때, 학습과정은 종료된다.

실험 결과

추출된 바꿔쓰기표현들

추출된 바꿔쓰기표현들을 한국어의 특성에 맞게 다음과 같이 세 종류로 나누어 보았다. 첫번째가 단순 어휘 수준에서 바꾸어 쓸 수 있는 바꿔쓰기표현들, 두번째가 기능어 수준에서 바꾸어 쓸 수 있는 바꿔쓰기표현들, 마지막으로 구절 단위에서 바꾸어 쓸 수 있는 바꿔쓰기표현들이다. 어휘 수준의 바꿔쓰기표현은 주로 단순 어절을 단순 어절로 바꾸어 쓰는 경우이고, 기능어 수준의 바꿔쓰기표현은 기능어만을 대치하는 경우이며, 구절 단위의 바꿔쓰기표현은 구절을 구절로 바꾸어 쓰는 경우이다.

(그림 5)에 보이는 어휘수준의 바꿔쓰기표현은 한 어절 또는 두 어절의 단어 대 단어로 바꿔쓸 수 있는 대체표현들을 의미한다.

어휘 수준의 바꿔쓰기표현에는 바꾸어 사용할 수 있는 유용한 대체표현뿐만 아니라 코퍼스의 특성에서 기인한 다음과 같은 결과도 나왔다.

(1) 한글맞춤법의 변화로 인한 결과 (개역한글을 개역개정으로 수정하면서주로 나타남)

예제: (진찰할지니, 진찰할찌니)

(2) 번역개정에서 수정된 단어들로 인한 결과 (개역한글을 개역개정으로 수정하면서 '저 → 그', '저희 → 그들'로 수정함)

예제: (저희, 그들)

(3) 띄어쓰기의 비일관성:

예제: (소리지르다, 소리 지르다)

(4) 고유명사 표기방법의 변화

예제: (바리사이파, 바리새파) (갈릴래아, 갈릴리)

(5) 성경의 특성 (성경 내용에 비추어 볼 때에만 허용될 수 있는 바꿔쓰기표현들)

예제: (딸, 여자) (백성, 자손) (족속, 사람) (식물, 음식) (호숫가, 바닷가)

(6) 고어체 표현

예제: (보수하다, 보복하다) (여종, 비자) (궤휼, 거짓) (신낭, 고환) (유벽하다, 구석지다)

(7) 복수형, 높임말, 단수/복수 사용의 비일관성:

예제: (여쭈다, 묻다) (너, 너희) (원수들, 대적)

(하늘에 올라가다, 승천하다) (어머니께서 죽다, 어머님이 숨을 거두다) (고름을 흘리다, 유출병이 있다)
(행복하다, 유복하다, 복이 있다) (결어서, 도보로) (흡족히 마시다, 양껏 마시다) (더러운 악령, 악한 귀신)
(넉넉하게 살다, 풍족하다) (효과를 나타내다, 효력을 내다) (함께하다, 함께 있다) (폐회하다, 모임이 끝나다)
(하찮게 생각하다, 멸시하다) (한밤중에, 밤중쯤 되어) (구조되다, 구원을 얻다) (안전하게 거주하다, 안전히 거하다)
(영접하는 자, 맞아들이는 사람) (말씀하여 이르시되, 이르어 가라사대) (살아있는, 산) (일곱째 날에, 제 칠일)
(~의 말씀이니라, ~께서 가라사대) (명령을 따라, 명을 좇아) (애통하는 자, 슬퍼하는 사람) (욕을 받다, 치욕을 당하다)
(지시, 시키는 것) (거룩하신 이름, 성호) (시험삼아, 시험적으로) (학대받다, 억눌리다) (화를 내다, 노하다)
(그리고나서, 그런 다음에) (범사에, 모든 일에서) (그리고, 그런 다음에) (할렐루야, 여호와를 찬양하십시오)
(명령하신 대로, 명하신 대로) (복병, 매복한 군인들) (이와 같이 말씀하시되, 가라사대) (~에 있는, ~에 사는)
(너희, 여러분 모두) (귀를 기울이다, 듣다) (내게 말하는, 나와 더불어 말하던) (유업으로 받다, 상속받다)
(자세히 묻다, 캐어묻다) (거짓을 행하다, 속이다) (병이 낫다, 병 고침을 받다) (그리고 나서, 그런 다음에)
(염병이 그치매, 재앙이 그치자) (범죄한 자들, 죄를 지은 사람) (그 때가 되다, 그 때가 이르다) (소리를 지르다, 외치다)
(환난에서, 고통 가운데서) (죄를 짓다, 범죄하다) (굴복하다, 무릎을 꿇다) (~가 마음이 괴로와서, ~는 괴로운 마음으로)
(아버지, 아비, 부친) (송사하다, 고발하다) (문둥병, 나병) (맹인, 소경) (어떤, 어느) (그런즉, 그러면) (어찌하여, 왜)
(연고, 까닭) (온, 모든) (여자, 여인, 처녀) (갑자기, 홀연히) (군인, 군병) (도둑, 도적) (화목, 친교) (한편, 그런데)
(청년, 소년) (혹시, 만일) (매우, 심히) (맏아들, 장자, 첫아들) (찬미하다, 찬송하다) (허다하다, 수많다) (불쌍히, 긍휼히)
(유대인, 유대 사람) (중에, 가운데) (북쪽, 북방, 북) (결혼하다, 혼인하다) (정혼하다, 약혼하다) (여러, 많은) (즉시, 곧)
(화염, 불길) (나병환자, 문둥이) (무렵, 즈음) (이제, 지금) (정탐, 탐지) (찾다, 구하다) (송사, 고발) (끝내, 마침내)
(방백, 우두머리) (서른, 삼십) (오직, 다만) (올리브, 감람) (네거리, 사거리) (며칠, 수일) (모든, 온갖) (오히려, 차라리)
(우선, 먼저) (가령, 이를테면) (경배하다, 숭배하다) (악마, 마귀, 귀신) (임신, 잉태, 수태) (친히, 스스로) (악인, 악한 자)
(쳐죽이다, 도륙하다) (원하건대, 청컨대) (제각기, 각기) (질투하다, 질시하다) (경외하다, 두려워하다) (수일 후, 며칠 뒤)
(태곳적부터, 옛부터) (종일토록, 온종일) (저마다, 각각) (매일, 날마다) (무질서, 어지러움) (가까운, 부근의)
(가두어 두다, 금고하다) (말미암아, 인하여) (영원히, 영원토록) (아니하다, 않다) (간계, 간사한 꾀) (온 무리, 모든 사람)

(그림 5) 추출된 어휘수준의 바꿔쓰기표현

(~라, ~십시오) (~더라, ~쓰다) (~며, ~고) (~리라, ~ㄹ 것입니다, ~ㄹ 것이다, ~겠다, ~리로다) (~느냐, ~ㅂ니까, ~느뇨)
(~어본즉, ~어보니) (~느니라, ~습니다) (~기에, ~관대) (~시옵고, ~옵시고) (~에 관한, ~에 대한) (~도록, ~게) (~매, ~니)
(~ㄹ지어다, ~기를 빕니다) (더냐, 느냐) (~려 하심이라, ~시려는 것입니다) (~냐, ~뇨) (및, ~와/과) (~처럼, ~같이)
(~같이, ~처럼) (~겠나이다, ~겠습니다) (~도다, ~구나) (~께서, ~가/이) (~ㄴ데, ~지만) (~고 나서, ~ㄴ 다음에)
(~나이까, ~습니까) (~옵소서, ~십시오) (~리로다, ~련다) (~들아, ~여러분) (~에게, ~더러) (~처럼, ~만큼) (~니, ~ㄴ즉)
(~라, ~시오) (~도다, ~소, ~습니다, ~다) (~거든, ~면) (~이/가, ~은/는) (~소서, ~어/아 주십시오) (~이/가 다, ~은/는 모두)
(~리라, ~ㄹ 것입니다) (~께서, ~님께서, ~님이) (~ㄹ 것입니다, ~겠습니다) (~시오, ~십시오) (~에서부터, ~에서)
(~ㄴ 까닭, ~기 때문) (~때문에, ~의 연고로)

(그림 6) 추출된 기능어 수준의 바꿔쓰기표현

기능어 수준의 바꿔쓰기표현은 대치할 수 있는 기능어들의 모임이다. 성경 코퍼스가 만연체 표현을 많이 사용했기 때문에, 종결형 어미와 연결형 어미의 대치가능한 표현들도 추출되었다. 본 연구에서는 종결형 어미와 연결형 어미의 바꿔쓰기표현은 추출하지 않도록 하였다.

구절단위의 바꿔쓰기표현은 어순이 바뀌면서 사용될 수 있는 대체표현들이 해당된다. (그림 7)이 구절단위의 바꿔쓰기표현에 해당한다.

PRON+JKB 진정으로 ↔ 진실로 PRON+JKB (예제) 너희에게 진정으로 ↔ 진실로 너희에게	NNG+JKO 모두 ↔ 모든 NNG+JKO (예제) 기름을 모두 ↔ 모든 기름을
한 NNG ↔ NNG 하나 (예제) 한 바위 ↔ 바위 하나	NNG+JX 얼마든지 ↔ 모든 NNG+JX (예제) 약속은 얼마든지 ↔ 모든 약속은
또 NNP+JKB ↔ NNP+JKB 다시 (예제) 또 기드온에게 ↔ 기드온에게 다시	MAG NNG+JKS ↔ NNG+JKS MAG (예제) 이미 도끼가 ↔ 도끼가 이미

(그림 7) 추출된 구절단위 수준의 바꿔쓰기표현

추출된 바꿔쓰기표현의 평가

본 연구에서 추출한 바꿔쓰기표현은 모두 7,170개이다. 그 중 어휘수준의 바꿔쓰기표현이 6,857개이며 기능어 수준이 83개, 구절단위의 바꿔쓰기표현이 32개 등으로 각각 추출되었다.

대명사는 사람이나 사물의 이름을 대신하여 쓰는 단어이다. 그렇기 때문에 추출된 바꿔쓰기표현 중 일부는 대명사로 바꿔쓴 경우들이 포함되어 있었다. 즉 '(주님, 그)', '(것, 말씀)'와 같은 바꿔쓰기표현들이 많이 추출되었다. 본 연구에서는 이와 같은 대명사 또는 대동사('하다')와 같이 추출된 바꿔쓰기표현들을 자동으로 제거하였다. 제거 결과 어휘수준의 바꿔쓰기표현 중 159개가 제거되었다.

추출된 바꿔쓰기표현의 정확도에 대한 평가는 그리 용이하지 않다. 본 연구에서는 159개가 제거된 6,698 개의 어휘수준 바꿔쓰기표현 중에서 임의로 600개를 추출하여 수동으로 평가를 수행하였다. 두 명의 평가자에게 문맥정보 없이 추출된 바꿔쓰기표현을 제시하였다. 충분히 바꿔쓸 수 있다고 평가되면 '○'를 바꿔쓸 수 없다고 평가되면 'X'를 표시하도록 하였다. 평가기준 중, 높임말 표현이나 단/복수형 표현, 시제, 띄어쓰기 등의 불일치에 대한 평가는 수행하지 않도록 하였다. 첫번째 평가자는 486개(81%)의 바꿔쓰기표현이 유용하다고 판단하였으며 두번째 평가자의 경우 474개(79%)가 유용하다고 평가하였다. 기능어 수준의 바꿔쓰기표현과 구절단위의 바꿔쓰기표현은 그 개수가 많지 않기 때문에 추출된 결과 모두에 대해서 평가를 수행하였다.

틀렸다고 평가된 바꿔쓰기표현들을 살펴보면 <표 4>와 같다. 오류 유형 (1)은 문맥정보가 주어진다면 바꿔쓰기표현으로 간주될 수 있는 오류들이었다. 오류 유형 (2)는 일반적인 바꿔쓰기표현 추출 오류이다. 오류 유형 (3)의 경우는 의미는 동일하나 품사나 상태/동작, 능동/피동 등과 같이 문장에서 쓰임의 차이로 인해 발생한 오류이다. 실제 성경 코퍼스에서는 다음과 같은 문장들이 함께 정렬되어 있었다.

(예제1) 세겜 사람들이 산들의 꼭대기에 사람을 매복시켜

<표 3> 추출된 바꿔쓰기표현에 대한 평가

	1st 평가자	2nd 평가자
어휘수준의 바꿔쓰기표현 (600개 평가)	81%	79%
기능어수준의 바꿔쓰기표현 (83개 평가)	97%	95%
구절단위의 바꿔쓰기표현 (32개 평가)	91%	91%

<표 4> 오류 유형별 예제

오류유형	예 제
(1)	(빵, 떡) (힘 없는 자, 가난한 사람) (많은 재물을, 전리품을 많이) (수금, 거문고) (악한, 더러운)
(2)	(한, 어떤) (자손, 사람) (족속, 사람) (예수께서, 대답하여) (세례자, 세례) (먼젓번, 처음) (끊어, 모두 끊어서) (많다, 크다) (사울은 주님, 여호와) (그러나, 또한)
(3)	(기록하다, 기록되다) (담겨있다, 담다) (매복하다, 매복시키다) (미혹하다, 미혹되다) (기쁘다, 기뻐하다) (놀라다, 놀래다) (둘, 두) (사형을 받다, 죽이다)

세겜 사람들이 산들 꼭대기에 사람을 매복하여

(예제2) 누가 어떻게 하여도 너희가 미혹되지 말라

누가 아무렇게 하여도 너희가 미혹하지 말라

추출된 바꿔쓰기표현은 그 정확도뿐만 아니라 코퍼스로부터 바꿔쓰기표현을 놓치지 않고 얼마나 많이 추출해 낼 수 있는가도 중요한 문제이다. 즉 바꿔쓰기표현 추출의 재현율(recall)도 중요한 성능 평가 중의 하나라고 할 수 있다. 본 연구에서는 이에 대한 평가를 위해 임의대로 추출한 정렬된 문장 집합으로부터 수동으로 바꿔쓰기 표현을 추출해 보도록 시켰다. 50개의 정렬된 집합으로부터 바꿔쓰기표현을 추출했으며 사람이 추출한 결과 모두 86개의 바꿔쓰기표현이 추출되었다. 이 중 알고리즘에 의해 자동으로 추출된 것과 공통되는 것이 모두 53개로 약 62%의 재현율을 보임을 알 수 있었다.

결 론

본 논문에서는 공동학습법이라는 자율학습을 이용하여 정렬된 한국어 성경코퍼스로부터 바꿔쓰기표현을 자동으로 추출해 보았다.

바꿔쓰기표현을 이용하면 다양한 문장을 생성해 낼 수 있기 때문에 바꿔쓰기표현의 추출은 문장 생성 모듈에서는 매우 중요한 요소 중의 하나이다. 그러나 바꿔쓰기표현들을 자동으로 추출할 수 있는 원본 데이터를 구하는 것이 쉽지 않다. 본 연구에서는 다양한 번역본과 문장 단위로 정렬이 되어 있는 성경을 갖고 바꿔쓰기표현을 자동으로 추출하였다.

추출된 바꿔쓰기표현은 어휘 단위, 기능어 단위, 구절 단위로 나눠볼 수 있었다. 본 연구에서 추출한 바꿔쓰기표현들은 동일한 의미의 다양한 표현이라는 것 외에도 성경에서 볼 수 있는 고어체 표현이나 사어(死語)들을 쉬운 풀이말로 바꿔쓰기 할 수 있도록 해준다. 그렇게 함으로써 어려운 고어체 표현이나 사어들에 대한 학습효과도 기대할 수 있었다.

본 연구에서 추출한 바꿔쓰기표현들은 주로 한 두 단어의 어휘 수준의 바꿔쓰기표현들이 대부분이었고, 길어봤자 3~4 단어 수준의 구절단위의 바꿔쓰기표현들이 추출되었다. 앞으로의 연구는 좀더 넓은 범위의 바꿔쓰기표현 - 예를 들면, 문장 수준 - 을 추출하는 것이라 하겠다.

참고문헌

[1] Regina Barzilay and Kathleen R. McKeown. Extracting Paraphrases from a Parallel Corpus. In *Proc. of the Meeting of the Association for Computational Linguistics*, 2001.

[2] I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *Proc. of the COLING-ACL*, 1998.

[3] Bo Pang, Kevin Knight and Daniel Marcu. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proc. of the HLT-NAACL 2003*, 2003.

[4] Regina Barzilay and Lillian Lee. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proc. of HLT-NAACL 2003*. 2003.

[5] http://www.holybible.or.kr/.

[6] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*. 1995.

[7] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. 1999.

[8] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proc. of the 11th Annual Conference on Computational Learning Theory*, 92-100, 1998.

[9] Regina Barzilay and Lillian Lee. Bootstrapping Lexical Choice via Multiple-Sequence Alignment. In *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing*. 2002.

1 차원고접수: 2006. 8. 23
2 차원고접수: 2006. 10. 28
최종게재승인: 2006. 12. 2

부 록

대분류	소분류	세분류
(1) 체언	명사 NN	일반명사 NNG
		고유명사 NNP
		의존명사 NNB
	대명사 NP	
	수사 NR	
(2) 용언	동사 VV	
	형용사 VA	
	보조용언 VX	
	지정사 VC	긍정지정사 VCP
		부정지정사 VCN
(3) 수식언	관형사 MM	
	부사 MA	일반부사 MAG
		접속부사 MAJ
(4) 독립언	감탄사 IC	
(5) 관계언	격조사 JK	주격조사 JKS
		보격조사 JKC
		관형격조사 JKG
		목적격조사 JKO
		부사격조사 JKB
		호격조사 JKV
		인용격조사 JKQ
	보조사 JX	
	접속조사 JC	

대분류	소분류	세분류
(6) 의존형태	어미 E	선어말어미 EP
		종결어미 EF
		연결어미 EC
		명사형전성어미 ETN
		관형형전성어미 ETM
	접두사 XP	체언접두사 XPN
	접미사 XS	명사파생접미사 XSN
		동사파생접미사 XSV
		형용사파생접미사 XSA
		(부사파생접미사 XSB)
	어근 XR	

(7) 기호		
	마침표,물음표, 느낌표	SF
	쉼표, 가운뎃점, 콜론, 빗금	SP
	따옴표, 괄호표, 줄표	SS
	줄임표	SE
	붙임표(물결, 숨김, 빠짐)	SO
	외국어	SL
	한자	SH
	기타 기호	SW
	명사추정범주	NF
	용언추정범주	NV
	숫자	SN
	분석불능범주	NA