변환 기반 학습을 이용한 한국어 비교 문장 유형 분류

양선○ 고영중
동아대학교 컴퓨터공학과
syang@donga.ac.kr, yjko@dau.ac.kr

Classifying Korean Comparative Sentences Using Transformation-based Learning

Seon Yang○ Youngjoong Ko
Computer Engineering, Dong-A University

요 약

본 연구의 목표는 비교 문장들을 일곱 가지 유형으로 자동 분류하는 것으로서, 비교 문장 추출, 비교 문장 유형 분류, 유형별 비교 관계 분석으로 이어지는 비교마이닝 세 단계 중 두 번째 과제이다. 본 연구에서는 변환 기반 학습(Transformation-based Learning) 기법을 이용한다. 자연어 처리 분야 여러 부문에서 사용되고 있는 변환 기반 학습은 오류를 감소시키는 최적의 규칙을 자동으로 생성하여 정답을 찾는 규칙 기반 학습 방법이다.

웹상의 다양한 도메인에서 추출한 비교 문장들을 대상으로 실험한 결과, 일곱 가지 비교 문장 유형을 분류하는데 있어서 정확도 80.01%의 우수한 성능을 산출하였다.

주제어: 비교 문장 유형, 비교마이닝, 변환 기반 학습

1. 서 론

사물이나 사람, 제도 등 어떤 대상을 평가함에 있어서 가장 직접적이고도 확실한 방법 중 하나로 비교(comparison)를 들 수 있다[1]. 비교 정보는 다양한 분야에서 중요한 근거 정보로 활용될 수 있는데, 예를 들어 고객 리뷰는 다른 고객의 구매 의사 결정에 영향을 주며, 신제품을 출시한 회사에서는 자회사의 신상품과 경쟁사 유사제품을 비교한 자료가 향후 마케팅 방향을 좌우할 수 있다.

비교마이닝(comparative mining)은 대용량의 문서에서 비교 문장들만을 자동으로 추출한 후 비교 유형별로 비교 관계를 정의 및 분석하는 학문으로서, 텍스트 마이닝의 한 분야로 볼 수 있다. 비교마이닝은 1)비교 문장 추출, 2)비교 문장 유형 분류, 3)유형별 비교 관계 분석이라는 세 단계 과제로 나눌 수 있는데, 선행 연구[2]에서 첫 번째 단계인 한국어 비교 문장 추출 시스템을 제안하였으며, 본 논문에서는 추출된 비교 문장들을 일곱 가지의 유형으로 자동 분류하는 두 번째 과제에 대해 연구한다.

이 과제를 수행하기 위해 먼저 비교키워드에 의한 분류를 시도할 수 있는데, 이 방법은 한계를 가진다.

Ex1.“고객님<u>처럼</u> 다른 분들도 소니가 전자제품 회사로는 <u>최고</u>라고 생각하시는 경우가 많은데, 제 의견을 말씀드리자면 카메라는 소니<u>보다</u> 캐논이 확실히 낫습니다.”

예를 들어 위 예문은 캐논 카메라가 소니 카메라보다 낫다는 화자(holder)의 의견을 표현한 비교 문장으로 문장 유형으로 보면 우열비교(greater or lesser)에 속하지만, ‘-처럼(유사)’, ‘최고(최상급)’, ‘-보다(우열)’라는 세 가지 유형의 키워드를 동시에 포함하고 있기 때문에 또 다른 추가적인 프로세스를 필요로 한다. 본 연구에서는 규칙 기반 학습 모델인 변환 기반 학습(Transformation-based Learning) 기법을 이용한다. 그리고 5-fold validation을 수행하여 연구의 성과를 보여준다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 관련 연구에 대해 기술하며, 3장에서는 비교 문장 유형 및 비교키워드에 대해 설명한다. 4장에서는 변환 기반 학습에 대해 설명하고, 5장에서는 실험 결과를, 6장에서는 본 연구의 결론 및 향후계획을 기술한다.

2. 관련 연구

본 연구는 자연어 처리 부문과 현대 한국어학 부문이

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2009-0065279).

라는 두 부문과 밀접히 관련되어 있다.

자연어 처리 부문은 다음과 같다. [1]에서 영어 문서에서 비교 문장을 식별하는 방법을 연구하였고, [2]에서는 한국어 문서에서 비교 문장을 추출하는 시스템을 제안하였다. [3]에서는 영어 우열비교 문장에서의 비교 관계 추출에 관해 연구하였다. 그리고 본 연구에서 사용한 기법인 변환 기반 학습 관련 연구는 다음과 같다. [4]에서 오류 기반의 변환 기반 학습 기법에 대해 처음으로 소개하였으며, [5,6,7]에서 변환 기반 학습의 응용 분야를 확대하였다.

그리고 현대 한국어학 부문은 다음과 같다.[8]에서 현대 한국어 비교구문의 체계를 정립하였으며, [9]에서는 한국어 동등 비교 구문에 관해 연구하였고, [10]에서는 '만큼'조사와 '처럼'조사의 이질성에 대하여 연구하였다. [11]에서는 형용사 최상급 비교구문의 의미를 연구하였다.

3. 비교 문장 유형 및 비교키워드

이 장에서는 비교 문장의 일곱 가지 유형에 대해 설명하고, 키워드만을 사용한 비교 유형 분류에는 한계가 있음을 설명한다.

3.1 비교 문장 유형

선행 연구[2]를 통해 비교 문장의 여덟 가지 유형을 정의하였는데, 본 연구에서는 의문문처럼 비교 결론이 없는 경우를 제외한 일곱 가지 유형에 대해 분류한다.

[표 1] 비교 문장의 일곱 가지 유형

	유 형	비교키워드 예
1	동등	'같/pa', '동일/ncp+하/xsp'
2	유사	'비슷하/pa', '유사/ncp+하/xsp'
3	상이	'다르/pa', '차이/ncn+가/jcs+나/pv'
4	우열	'보다/jca', '훨씬/ma'
5	의사비교	'라기/ecx' *속성간 비교
6	최상급	'가장/ma', '최고/ncn'
7	함축적 비교	<'는/jxt','지만/ecs','는/jxt','다/ef'>

주) pa, ncp, xsp 등은 각각 특정 품사를 나타내는 기호이다.

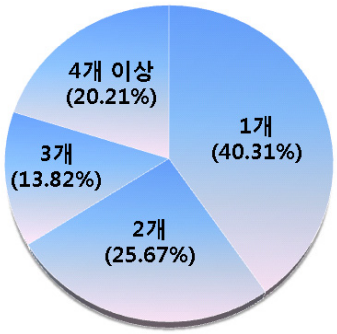
위 유형 중 7번째 유형은 한국어 비교마이닝 시스템 개발을 위하여 [2]에서 새롭게 확장된 유형으로 그 예는 아래와 같다.

Ex2."A-바나나우유는 바나나로 만들지만, B-바나나우유는 바나나가 전혀 들어있지 않고 바나나향으로만 맛을 낸다."

위 예문은 언어학적 관점으로 보면 비교 문장으로 분류되지 않을 수도 있지만, 'A-바나나우유가 B-바나나우유보다 좋다'라는 화자의 의견(opinion)을 함축하고 있다고 판단할 수 있다. 본 연구에서는 이처럼 비교 의미를 함축적으로 내포한 문장들도 비교 문장으로 간주한다.

3.2 비교키워드만을 이용한 유형 분류

[표 1]에서 보이듯이 키워드 자체의 유형을 분류할 수 있으므로 키워드만을 이용한 비교 문장 유형 분류를 시도할 수 있지만, 서론에서 언급하였듯이 비교 문장 하나가 여러 유형의 비교키워드를 동시에 포함하는 경우가 많기 때문에 이러한 시도에는 한계가 있다.



[그림 1] 비교키워드 보유 개수 별 문장 분포

[그림 1]은 1,831개의 비교 문장 중에서 보유 키워드 개수별 비교 문장 분포를 나타내고 있다. 이 그림에서 나타나듯 약 60%의 문장들이 2개 이상의 키워드 보유하고 있었다. 특히 일부 신문기사나 사설은 문장의 길이가 매우 긴 경우도 많아서 비교키워드를 5개 이상 보유한 경우도 다수 있었다. 따라서 비교 문장 유형 분류를 위해서는 추가적인 프로세스가 필요하다.

4. 변환 기반 학습 방법을 이용한 비교 유형 분류

이 장에서는 변환 기반 학습 방법의 개념에 대해 설명하고, 이 학습 방법을 비교 유형 분류에 적용하는 과정에 대해 기술 한다.

4.1 변환 기반 학습의 개념

변환 기반 학습은 1990년대 Brill[4]에 의해 처음 소개되었으며 자연어 처리 분야에서 널리 이용되어 왔다. 변환 기반 학습은 정답에 최대한 가까워질 때까지 오류를 줄인다는 아이디어를 기본으로 하는 규칙 기반 학습 과정이다. 먼저 초기 정보 부착기를 이용하여 학습 말뭉치에 초기 정보를 부착시키고, 미리 설정해 놓은 후보 규칙들을 학습 말뭉치에 모두 적용한 후 가장 오류를 많이 수정해주는 규칙 하나를 선택한다. 선택된 규칙은 변환 규칙 리스트에 순서대로 저장된다. 이처럼 매 학습 회전마다 그리디 탐색을 적용하여 가장 정답 말뭉치에 가깝게 수정해주는 규칙을 계속 찾아나가며, 더 이상 정답 말뭉치에 가깝게 변환시켜주는 규칙을 찾을 수 없을 때 학습 회전은 중단된다.

규칙 기반 학습은 확률 기반 학습에서 문제가 되었던

데이터 희소성 문제에 강한 특징[4]을 가진다. 본 연구에서 사용하는 비교 말뭉치는 웹에서 추출하였으며, 이러한 비교 문장들 중에는 뉴스 기사 같은 정형화된 문장들도 많지만, 구어체의 고객 리뷰 혹은 개인 블로그도 상당 부분 포함되어 있다. 이처럼 비정형적인 문장들이 다수 포함된 다중 분류에서 데이터 희소성 문제에 상대적으로 강하고 오류를 감소시키는 방향으로 학습하는 변환 기반 학습은 다른 기법들보다 상대적으로 높은 성능을 보일 수도 있다고 기대할 수 있다.

4.2 변환 규칙 틀 및 변환 규칙 선택 함수

후보 규칙들을 설정하기 위해서는 먼저 변환 규칙 틀을 정의해야 하는데, 본 연구에서는 [4]를 참고로 하여 규칙 틀을 [표 2]와 같이 여덟 가지로 정의한다.

[표 2] 변환 규칙 틀

아래와 같은 경우 비교 유형 a를 b로 변환시킨다.

1. 키워드가 k일 때
2. 1을 만족하면서, 앞(preceding) 품사가 z일 때
3. 1을 만족하면서, 뒤(following) 품사가 z일 때
4. 1을 만족하면서, 앞의 앞(two before) 품사가 z일 때
5. 1을 만족하면서, 뒤의 뒤(two after) 품사가 z일 때
6. 1을 만족하면서, 앞 품사가 z이고 뒤 품사가 w일 때
7. 1을 만족하면서, 앞 품사가 z이고 앞의 앞 품사가 w일 때
8. 1을 만족하면서, 뒤 품사가 z이고 뒤의 뒤 품사가 w일 때

다음 예는 2번 규칙 틀에 의해 생성된 변환 규칙의 예이다.

if (키워드 = '-보다') and (앞 품사 = '고유명사')
then 유형 = '우열비교'

다음으로는 후보 규칙들 중 어떤 규칙을 선택할 것이냐에 기준이 되는 변환 규칙 선택 함수를 결정한다. 본 연구에서 변환 규칙 선택 함수는 각 후보 규칙을 적용했을 때 적용 전 틀린 상태였다가 적용 후 맞는 상태로 변환된 문장의 수를 나타내는 C(Correction)와 그 반대 경우의 수 E(Error)를 계산하여 C-E가 가장 큰 후보 규칙 하나를 선택한다. 이 과정을 반복 수행하면서, 각 수행마다 하나의 규칙이 선택되어 규칙리스트에 차례대로 저장되며, 변환 규칙 선택 함수가 더 이상 정답에 가까워지는 규칙을 찾을 수 없을 때 학습은 중단된다.

5. 실험 및 평가

본 연구에서는 [2]에서 수집하고 레이블링한 비교 문장들 중 1번부터 7번까지의 유형에 속하는 1,831개의 문장을 대상으로 실험한다. [표 3]은 레이블링 후 비교 문장의 유형별 분포를 나타낸다.

[표 3] 레이블링 후 유형별 비교 문장 수

	유 형	문장 수
1	동등	65 (3.6%)
2	유사	131 (7.2%)
3	상이	88 (4.8%)
4	우열	998 (54.5%)
5	의사비교	23 (1.3%)
6	최상급	206 (11.3%)
7	함축적 비교	320 (17.5%)

변환 기반 학습을 적용하기 위한 초기 정보 부착은 키워드만을 이용하는 분류방법을 사용하였다. 키워드를 1개만 보유한 문장은 해당 키워드 유형을 그대로 부착하기로 하고, 2개 이상 보유한 경우는 한국어의 특성상 종속절보다 주절이 문장 후반부에 위치하는 경우가 많기 때문에 그 문장 안에서 가장 나중에 위치한 키워드의 유형을 부착하였다. 초기 정보 부착 결과 그 정확도는 66.24%였는데, 키워드만 사용한다는 한계에도 불구하고 비교적 높은 성능을 나타내었다. 변환 기반 학습 환경을 요약하면 아래와 같다.

- 초기 상태(키워드만 이용) 정확도: 66.24%
- 변환 규칙 틀 수: 8개
- 생성된 변환 규칙 후보 수: 5,116개
- 선택된(C-E>0) 변환 규칙 리스트 사이즈 : 227

그런데 위의 227개의 규칙 리스트 중에는 학습 말뭉치에서 단지 한 문장만 개선시키는 효과(C-E=1)를 가진 규칙들도 많았는데 이러한 규칙들이 실제 실험 말뭉치 적용 시에는 오히려 과적용(Overfitting) 문제를 발생시킬 수도 있다고 판단되었다. 따라서 본 연구에서는 학습 말뭉치에서 두 문장 이상에 대한 개선 효과를 가지는 변환 규칙들만 실험 말뭉치에 적용하도록 제한 조건을 추가하였다.

- 제한 후(C-E>1) 변환 규칙 리스트 사이즈 : 51

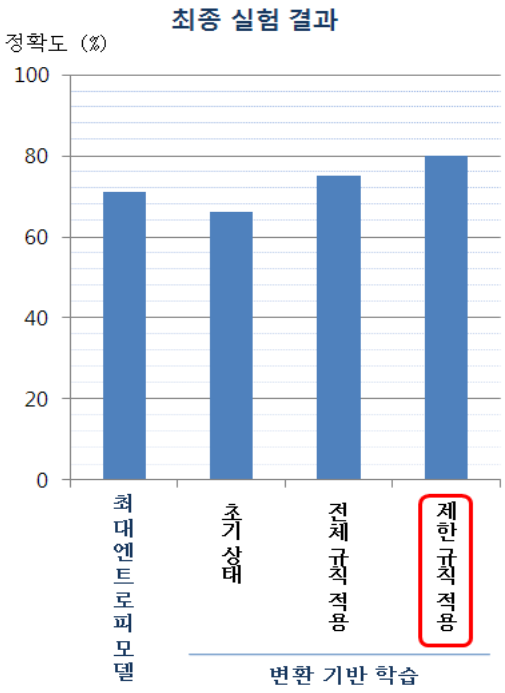
이렇게 하여 최종적으로 변환 기반 학습은 비교 문장 유형 분류에서 정확도 80.01%의 성능을 산출하였다. 본 연구에서는 변환 기반 학습의 성능과 비교하기 위하여 확률 기반 모델인 최대 엔트로피 모델을 이용한 실험을 병행하였다. 자질로는 [2]에서 비교 문장 추출을 위해 사용했던 키워드 중심 반경 3 이내에 있는 모든 연속된 품사시퀀스 자질을 사용하였다.

자질 예: <jxt ncn jcs **같/pa** ep ef sf> -> 동등비교

이와 같은 자질로 비교 유형 분류에 대한 최대 엔트로피 모델을 적용한 결과는 71.25%였다. [표 4]와 [그림 2]는 최종 실험 결과를 보여주고 있다.

[표 4] 성능 비교 (단위: %)

실험		정확도
최대 엔트로피 모델		71.25
변환 기반 학습	키워드만 이용한 초기 상태	66.24
	선택된(C-E>0) 변환 규칙 모두 적용	75.00
	선택된 변환 규칙들 중 C-E>1로 제한한 경우	**80.01**



[그림 2] 최종 성능 비교 (단위: %)

6. 결 론

본 연구에서는 한국어 문장 특징을 참고하여 키워드만을 이용하여 초기 정보를 부착시키고, 키워드 좌 우 반경 2까지의 품사 정보를 이용하는 변환 기반 학습 기법을 수행함으로써, 한국어 비교 문장 일곱 가지 유형 분류에서 정확도 80.01%의 우수한 결과를 산출하였다.

본 연구의 성과는 다음과 같이 요약할 수 있다.

1) 한국어 비교마이닝 시스템 개발 관련 연구가 국내에서는 처음이고 선진 외국에서도 초기 단계임을 고려할 때, 본 연구의 국내 텍스트 정보처리 연구 영역을 넓히고 기술 수준을 높이는 계기가 될 것이다.

2) 본 연구에서 진행하는 비교마이닝은 텍스트 마이닝 기술의 발전에 기여할 것이며, 이들 기술은 영역 이식성이 매우 높기 때문에 감정/의견마이닝 등 다양한 응용영역에 손쉽게 적용될 수 있을 것이다.

3) 본 논문에서의 비교 문장 유형 분류는 향후 비교 문장에서 비교주체, 비교대상, 비교 관계 등 다양한 세부 정보를 추출할 수 있는 기반이 될 수 있다.

향후 유형 분류에서의 정확도 향상을 위해 실험을 계속 진행할 것이며, 비교 주체, 비교 대상, 비교 관계 추출을 통한 분석 등 다양한 연구를 지속할 것이다.

참고 문헌

[1] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents", In *Proceeding of SIGIR*, pp.244-251, 2006

[2] 양선, 고영중, "기계학습 기법을 이용한 한국어 비교 문장 추출", *제20회 한글 및 한국어 정보처리 학술대회*, pp.182-287, 2008

[3] N. Jindal and B. Liu, "Mining Comparative Sentences and Relations", In *Proceeding of AAAI*, pp.1331-1336, 2006

[4] E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging", In *Poceeding of Computational Linguistics*, pp.543- 565, 1995

[5] L. Ramshaw and M. Marcus, "Text Chunking using Transformation-Based learning", In *Proceeding of the Third ACL workshop on Very Large Corpora*, pp.82-94, 1995

[6] W. J. Black and A. Vasilakopoulos, "Language -Independent Named Entity Classification by Modified Transformation-Based learning and by Decision Tree Induction", In *Proceeding of CoNLL*, 2002

[7] 장유식, *변환 기반 학습을 이용한 정보 추출 방법*, 석사학위논문. 서강대학교 대학원: 컴퓨터학과, 2006

[8] 하길종, *현대 한국어 비교 구문 연구*, 도서출판 박이정, 1999

[9] 하길종, "현대 한국어 동등비교 구문의 의미 연구", *한국어학 5권 봄호*, pp.229-265, 1999

[10] 오경숙, "'만큼' 비교구문과 '처럼' 비교구문의 이질성", *한국어 의미학회 제 14차 전국학술대회*, pp.197-221, 2004

[11] 정인수, "국어 형용사 최상급 비교구문의 의미", *한민족어문학 제36집*, pp.61-86, 2000