한국어 비교 문장 유형 분류를 위한 변환 기반 학습 기법

(Transformation-based Learning for Korean Comparative Sentence Classification)

양 선 † 고 영 중 ††
(Seon Yang) (Youngjoong Ko)

요 약 본 논문은 비교마이닝(comparison mining)의 일환인 비교 문장 유형 자동 분류에 관하여 연구한다. 비교마이닝은 텍스트 마이닝의 한 분야로서 대용량의 텍스트를 대상으로 비교 관계를 분석하며, 크게 세 단계의 과정을 거치게 되는데 첫 번째 단계는 대용량의 문서에서 비교 문장만을 식별 후 추출해 내는 과정이고, 두 번째 단계는 추출된 비교 문장들을 비교 유형별로 분류하는 과정이며, 앞의 두 선행 과정이 끝나면 유형별로 비교 속성을 추출 및 비교 관계를 분석하는 세 번째 단계를 수행하게 된다. 본 연구에서는 변환 기반 학습(transformation-based learning) 기법을 이용하여 비교 문장들을 일곱 가지의 유형으로 자동 분류하는 두 번째 과제를 수행한다. 자연어 처리 분야 여러 부문에서 사용되고 있는 변환 기반 학습은 오류를 감소시키는 최적의 규칙을 자동으로 생성하여 정답을 찾아가는 규칙 기반 학습 방법이다. 웹상의 다양한 도메인에서 추출된 비교 문장들을 대상으로 유형 분류를 수행한 결과 정확도 80.01%의 성능으로 일곱 가지 유형을 분류할 수 있었다.

키워드 : 비교마이닝, 비교 문장, 변환 기반 학습

Abstract This paper proposes a method for Korean comparative sentence classification which is a part of comparison mining. Comparison mining, one area of text mining, analyzes comparative relations from the enormous amount of text documents. Three-step process is needed for comparison mining - 1) identifying comparative sentences in the text documents, 2) classifying those sentences into several classes, 3) analyzing comparative relations per each comparative class. This paper aims at the second task. In this paper, we use transformation-based learning (TBL) technique which is a well-known learning method in the natural language processing. In our experiment, we classify comparative sentences into seven classes using TBL and achieve an accuracy of 80.01%.

Key words : Comparison mining, Comparative sentence, Transformation-based learning

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2009-0065279)
· 이 논문은 2009 한글 및 한국어 정보처리 학술대회에서 '변환 기반 학습을 이용한 한국어 비교 문장 유형 분류'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 동아대학교 컴퓨터공학과
syang@donga.ac.kr
†† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr

논문접수 : 2009년 11월 10일
심사완료 : 2009년 12월 4일

정보과학회논문지 : 소프트웨어 및 응용 제37권 제2호(2010.2)

1. 서 론

비교 정보는 활용 범위 및 가치가 매우 큼에도 불구하고, 대용량의 텍스트에서 비교 정보를 추출 및 분석하는 비교마이닝 시스템 구축에 대한 연구가 아직 활발히 진행되고 있지 않다. 비교는 둘 이상의 대상을 직접 견주어 평가하게 되므로, 여러 평가 방법 중 가장 명확한 방법 중 하나이며, 다양한 분야에서 의사 결정의 주요 근거 정보로 활용되고 있다[1]. 특히 웹이 급속한 팽창과 발전을 거듭하면서 웹상의 많은 비교 정보 활용에 대한 중요성 인식도 증가하고 있다. 비교 문장은 하나의 대상에 대해 화자의 감정이나 의견을 표현하는 문장보다 더 극명하게 청자의 의사 결정을 좌지우지할 수 있다. 예를 들어 기존 고객들이 여러 제품들을 비교 평가

하여 작성한 리뷰들은, 현재 어느 제품을 구입할지 망설이는 고객의 구매 의사에 결정적인 역할을 할 수 있다. 또한 기업에서는 자사 제품과 경쟁사 제품을 비교한 자료에 의해 향후 마케팅 방향이 좌우될 수 있다. 이 외에도 선거 예측, 벤치 마킹 등 다양한 분야에서 비교 정보는 매우 직접적이고 중요한 인자로 활용될 수 있다.

본 연구팀의 최종 목표는 웹상의 비교 정보를 자동으로 추출 및 분석하는 한국어 비교마이닝 시스템 구축이다. 비교 문장 추출, 비교 문장 유형 분류, 유형별 비교 관계 분석이라는 비교마이닝 세 단계 중에서 첫 번째 단계인 한국어 비교 문장 자동 추출 시스템은 선행 연구[2]를 통해 이미 제안하였으며, 본 논문에서는 두 번째 단계인 비교 문장 유형 자동 분류 기법에 대해 제안한다. 본 논문에서의 유형 분류는 비교 구문 체계를 언어학적으로 확립하려는 시도가 아닌, 비교마이닝의 세 번째 단계인 유형별 비교 관계 분석에서 정보 처리를 용이하게 하기 위함이다. 예를 들어 아래 두 유형 예에서 보이듯이 비교 유형이 달라지면 분석해야 하는 비교 요소들도 달라진다.

- 동등비교 경우
  예문: *X폰과 Y폰은 디자인이 똑같다.*
  분석 요소: 비교 주체 {X폰, Y폰}
  속성 {디자인}

- 우열비교 경우
  예문: *내구성 면에서는 X폰보다 Y폰이 낫죠.*
  분석 요소: 비교 주체 {Y폰}
  비교 대상 {X폰}
  속성 {내구성}
  관계 {낫다}

비교 문장 유형 분류를 위해서 일차적으로는 비교키워드를 이용할 수 있다. 비교키워드는 선행 연구[2]에서 비교 문장 추출을 위해 177개의 키워드를 이미 정의한 바 있다. 그러나 비교키워드가 비교 문장 추출에서 핵심 역할을 하였음에도 불구하고 유형 분류에서는 한계점을 나타내었는데, 포함된 키워드 유형이 문장 유형과 일대일 대응한다는 보장이 없을 뿐만 아니라, 더 큰 문제점은 아래의 예문들에서 나타나듯 한 문장이 여러 유형의 키워드를 동시에 포함한 경우가 많다는 사실이다.

- *X폰이 튼튼하고 성능도 우수하기 때문에 가장 비싼 가격에도 불구하고 잘 팔리는 거라고 생각들 하시는데, 저는 내구성 면에서도 성능 면에서도 X폰보다 Y폰이 오히려 우수하다고 생각합니다.*

위 예문은 X폰보다 Y폰이 낫다는 화자의 의견을 표현한 비교 문장으로, 문장 유형으로 보면 우열(greater or lesser)비교에 속하지만, '가장'이라는 최상급 키워드와 '-보다'라는 우열비교 키워드를 동시에 포함하고 있다. 또한 아래 예문은 문장 유형은 상이(difference)비교에 속하지만, '-처럼'이라는 유사비교 키워드와 '다르-'라는 상이비교 키워드를 동시에 포함한 경우이다.

- *다른 회사들처럼 천연 재료를 사용한다고 광고했던 X사가 실은 다른 회사들과는 다르게 값싼 공업 재료를 사용하고 있었죠.*

이와 같은 이유로 비교 문장 유형 자동 분류는 비교 키워드 외에 또 다른 추가적인 프로세스를 필요로 하며, 본 연구에서는 규칙 기반 학습 모델로서 자연어 연구 분야에서 잘 알려진 변환 기반 학습(transformation-based learning) 기법을 이용한다. 그리고 5-fold cross validation을 수행하여 연구의 성과를 보여준다.

본 논문은 다음과 같이 구성되어 있다. 1장의 서론에 이어 2장에서는 관련 연구에 대해 기술하며, 3장에서는 비교 문장 유형 및 비교키워드에 대해 설명한다. 4장에서는 변환 기반 학습의 개념 및 적용에 대해 설명하고, 5장에서는 실험 결과에 대해 기술하며, 6장에서는 본 연구의 결론 및 향후 연구 계획을 기술한다.

2. 관련 연구

본 연구는 자연어 처리 부문과 현대 한국어학 부문이라는 두 부문과 밀접히 관련되어 있다.

자연어 처리 부문에서 관련 연구를 살펴보면 다음과 같다. [1]에서는 키워드 및 Class Sequential Rules를 이용하여 영어 문서에서 비교 문장을 식별하는 방법을 제안하였고, [2]에서는 키워드 및 최대 엔트로피 모델을 이용하여 한국어 문서에서 비교 문장을 추출하는 시스템을 제안하였다. [3]에서는 영어 비교 문장 중에서 정도성(gradability)을 가진 우열비교 문장을 대상으로 비교 주체, 비교 대상, 비교 관계를 추출하는 방법에 대하여 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]에서 변환 기반 학습을 통한 문서 단위화(text chunking)에 대해 연구하였고, [6]에서는 언어 독립적 개체명 분류를 위하여 수정된 변환 기반 학습을 이용하였다. [7]에서는 변환 기반 학습을 이용한 정보 추출 방법을 제안하였다.

다음으로 현대 한국어학에서 비교 구문 관련 연구는 아래와 같다.

[8]에서 현대 한국어 비교 구문에 쓰이는 어휘를 검토하여 어휘의 특성과 비교구문의 체계를 정립하였다. [9]에서는 한국어 동등 비교 구문에 관해 연구하였고, [10]에서는 거의 동일한 기능으로 간주되던 '만큼'조사와 '처럼'조사의 이질성에 대하여 연구하였으며, [11]에서는 형용사 최상급 비교구문을 척도의 유형에 따라 분류하고, 정도부사 '가장'이 사용된 형용사 최상급 비교구문의 의미를 연구하였다.

3. 비교 문장의 일곱 유형

이 장에서는 비교 문장의 일곱 유형을 알아보고, 키워드만을 사용한 유형 분류에는 한계가 있음을 설명한다.

3.1 비교 유형

선행 연구[2]를 통해 비교 문장의 여덟 가지 유형을 정의하였는데, 본 연구에서는 8번째 유형(의문문처럼 비교의 결론이 나지 않은 경우)을 제외한 일곱 가지 유형에 대해 표 1과 같이 분류한다.

표 1 비교 문장 유형 및 유형별 비교키워드 예

	유형	비교키워드 예
1	동등	'같/pa'
2	유사	'비슷하/pa'
3	상이	'다르/pa'
4	우열	'보다/jca'
5	의사비교	'라기/ecx' *속성간 비교
6	최상급	'가장/ma'
7	함축비교	< '는/jxt', '지만/ecs', '는/jxt', '다/ef'>

주) pa, jca 등은 각각 특정 품사를 나타내는 기호이다.

위 유형 중 7번째 유형은 [2]에서 새롭게 확장된 유형으로 그 예는 아래와 같다.

- *A바나나우유는 바나나로 만들지만, B바나나우유는 바나나가 전혀 들어있지 않고 바나나 향으로만 맛을 낸다.*

위 문장은 언어학적 관점에서는 비(非)비교 문장으로 분류될 수 있다. 하지만 위 문장에서 화자는 'A가 B보다 낫다'는 의견을 나타내고 있다고 판단할 수 있으며, 설혹 화자가 A와 B를 비교하려는 의도 없이 순수하게 객관적 사실정보만 기술했다 하더라도 위 문장을 접한 청자들은 자연스럽게 'A가 B보다 낫겠구나'라고 받아들일 가능성이 크다. 따라서 위 문장은 비교마이닝 관점에서뿐만 아니라 감정(sentiment) 및 의견(opinion)으로서도 매우 중요한 의미를 가진다고 볼 수 있다. 본 연구에서는 이렇게 함축적으로 비교 의미를 내포한 문장들도 비교 문장으로 간주한다.

3.2 비교키워드의 한계

표 1에서 나타나듯 비교 유형별로 키워드가 따로 있으므로 비교키워드만을 이용한 비교 문장 유형 분류를 시도할 수 있는데 이러한 시도에는 한계가 있다. 그 이유는 서론에서 언급하였듯이 비교 문장 하나가 여러 유형의 비교키워드를 동시에 포함하는 경우가 많기 때문이다. 그림 1은 1,831개의 비교 문장 중에서 보유 키워드 개수별 비교 문장 분포를 보여준다.

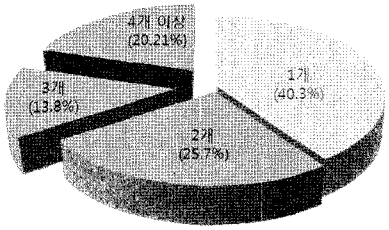

그림 1 보유 키워드 개수별 문장 분포

위 그림이 나타내듯 비교 문장들 중에서 키워드 1개만 보유한 문장의 비율은 약 40%에 불과하였으며, 60%에 가까운 문장들은 2개 이상의 키워드 보유하고 있었다. 특히 신문기사나 사설에는 한 문장의 길이가 매우 긴 경우도 많아서 비교 키워드를 5개 이상 보유한 경우도 다수 있었다. 따라서 키워드 유형과 문장 유형을 일대일로 대응시키는 분류 방법은 커다란 한계가 있으며, 다른 분류 기법을 필요로 한다고 판단하였다.

4. 제안하는 기법 : 변환 기반 학습

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 정의

변환 기반 학습은 1990년대 Brill[4]에 의해 처음 소개되었으며, 품사 태깅, 문서 단위화, 개체명 인식 등 자연어 처리 분야에서 널리 이용되어 왔다.

변환 기반 학습은 정답에 최대한 가까워질 때까지 오류를 줄인다는 아이디어를 기본으로 하는 규칙 기반 학습 과정이다. 변환 기반 학습 과정은 다음과 같다. 먼저 초기 정보 부착기(initial-state annotator)를 이용하여 학습 말뭉치에 초기 정보를 부착시킨다. 초기 정보 부착은 단순히 말뭉치에서 많이 나온 정보를 부착할 수도 있고 다른 분류기를 이용할 수도 있는 등 상황에 맞게 여러 가지 방법을 사용할 수 있다.

학습이 시작되면 이미 설정해 놓은 후보 규칙들을 학습 말뭉치에 모두 적용한 후 가장 오류를 많이 수정해 주는 규칙 하나를 선택한다. 선택된 규칙은 변환 규칙

리스트에 순서대로 저장된다. 이처럼 매 학습 회전마다 그리디 탐색(greedy search)을 적용하여 가장 정답 말뭉치에 가깝게 수정해주는 규칙을 계속 찾아나가며, 더 이상 정답 말뭉치에 가깝게 변환시켜주는 규칙을 찾을 수 없을 때 학습 회전은 중단된다.

본 연구의 목표인 비교 유형 분류는 이분 분류가 아닌 다중 분류로서, '비교'와 '비교 아님'으로 이분 분류하는 선행 연구[2]와는 다른 특성을 가지고 있다. 따라서 선행 연구에서 확률 기반 모델인 최대 엔트로피 모델을 이용하여 비교 문장 추출에서 높은 성능을 산출하였으나, 본 연구에서의 비교 유형 다중 분류를 위해서는 [2]와는 다른 접근이 필요하다고 판단되었다.

규칙 기반 학습은 확률 기반 학습에서 문제가 되었던 데이터 희소성 문제에 강한 특징을 가진다[4]. 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

후보 규칙들을 설정하기 위해서는 먼저 변환 규칙 틀(template)을 정의해야 하는데, 본 연구에서는 [4]를 참고로 하여 비교키워드 반경 2 이내의 품사 정보를 이용하여 표 2와 같이 여덟 가지 규칙 틀을 정의한다.

표 2 키워드와 품사 정보를 이용한 변환 규칙 틀

아래와 같은 경우 비교 유형 a를 b로 변환시킨다.
1. 키워드가 k일 때
2. 1을 만족하면서, 앞(preceding) 품사가 z
3. 1을 만족하면서, 뒤(following) 품사가 z
4. 1을 만족하면서, 앞의 앞(two before) 품사가 z
5. 1을 만족하면서, 뒤의 뒤(two after) 품사가 z
6. 1을 만족하면서, 앞 품사는 z, 뒤 품사는 w
7. 1을 만족하면서, 앞 품사는 z, 앞의 앞 품사는 w
8. 1을 만족하면서, 뒤 품사는 z, 뒤의 뒤 품사는 w

위의 8가지 규칙 틀을 기반으로 학습 말뭉치에서 틀에 맞는 모든 경우를 추출하여 후보 규칙들을 생성하게 되는데, 아래는 규칙 틀에 의해 생성된 변환 규칙 예이다.

2번 규칙 틀에 의해 생성된 규칙 예 :
if ((키워드 = '보다/jca') and
(앞 품사 = '고유명사'))
then 비교 문장 유형 = '우열'

6번 규칙 틀에 의해 생성된 규칙 예 :
if ((키워드 = '가장/ma') and
(앞 품사 = '부사격조사') and
(뒤 품사 = '형용사어간'))
then 비교 문장 유형 = '최상급'

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 전에 틀린 상태였다가 적용 후 맞는 상태로 변환된 문장의 수를 나타내는 C(correction)와 그 반대 경우의 수 E(error)를 계산하여 그 차이인 C-E가 가장 큰 후보 규칙 하나를 선택한다. 모든 후보 규칙들을 전체 학습 말뭉치에 적용하여 가장 적합한 후보 규칙 하나를 선택하는 과정을 계속 반복하면서, 각 수행마다 하나의 규칙이 선택되어 규칙리스트에 차례대로 저장되며, 변환 규칙 선택 함수가 더 이상 정답에 가까워지는 규칙을 찾을 수 없을 때 학습은 중단된다.

5. 실 험

5.1 실험 대상 및 과정

본 연구에서는 [2]에서 수집하고 레이블링한 비교 문장들 중 1번부터 7번까지의 유형에 속하는 1,831개의 비교 문장을 대상으로 실험한다. 이 문장들은 웹상의 뉴스 기사, 제품 비교 사이트, 개인 블로그 등 다양한 도메인을 통해 수집되었다. 그림 2는 레이블링 후 비교 문장의 유형별 분포를 나타낸다.

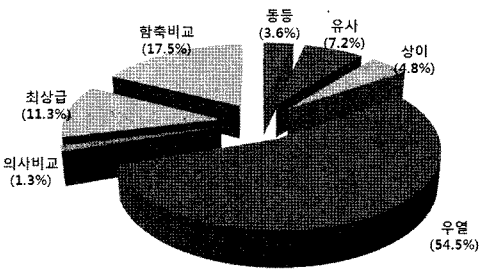

그림 2 유형별 비교 문장 비율

변환 기반 학습을 적용하기 위한 초기 정보 부착은 키워드만을 이용하는 분류방법을 사용하였다. 키워드를 1개만 보유한 문장 경우는 해당 키워드의 유형을 그대로 부착하고, 키워드를 2개 이상 보유한 문장 경우는 그 문장 안에서 가장 나중에 나온 키워드의 유형을 부착하였다. 가장 나중에 나온 키워드를 선택한 이유는 한국어의 특성상 종속절보다 주절이 문장 후반부에 위치하는 경우가 많기 때문이다. 이렇게 초기 정보를 부착하였을

때 그 정확도는 66.24%였는데, 키워드만 사용한다는 한계에도 불구하고 비교적 높은 성능을 나타내었다.

또한 변환 기반 학습 수행 시 227개(C>E)의 최종 규칙 리스트가 결정되는데, 이 227개의 규칙 리스트 중에는 학습 말뭉치에서 단지 한 문장만 개선시키는 효과(C-E=1)를 가진 규칙들이 많은 비중을 차지하였다. 비록 앞서 기술하였듯이 변환 기반 학습은 학습 말뭉치를 통해 결정된 모든 변환 규칙을 실험 말뭉치에도 적용하는 것을 기본으로 삼지만, 실험 말뭉치보다 네 배 사이즈의 학습 말뭉치에서 단지 한 문장 개선 효과만 가지는 변환 규칙이라면, 실제 실험 말뭉치 적용 시에는 오히려 과적용 문제를 발생시킬 위험이 있다고 판단되었다. 따라서 본 연구에서는 학습 말뭉치에서 두 문장 이상에 대한 개선 효과를 가지는 변환 규칙들만 실험 말뭉치에 적용하도록 제한 조건을 추가하였다. 이와 같이 변환 규칙 선택 조건을 강화하자 실험 말뭉치에 적용되는 변환 규칙 리스트 사이즈는 51로 감소하였다. 변환 기반 학습 환경을 요약하면 표 3과 같다

표 3 변환 기반 학습 환경

키워드만 이용한 초기 상태 정확도	66.24%
품사 정보를 활용하는 변환 규칙 틀 수	8개
틀에 의해 생성된 변환 규칙 후보 수	5,116개
학습 후 선택된 모든(C>E) 변환 규칙 리스트 사이즈	227개
제한 조건(C>E+1)을 만족하는 변환 규칙 리스트 사이즈	**51개**

5.2 최종 성능

이렇게 하여 최종적으로 변환 기반 학습은 그림 3과 같이 비교 문장 유형 분류에서 정확도 80.01%의 성능을 산출하였다.

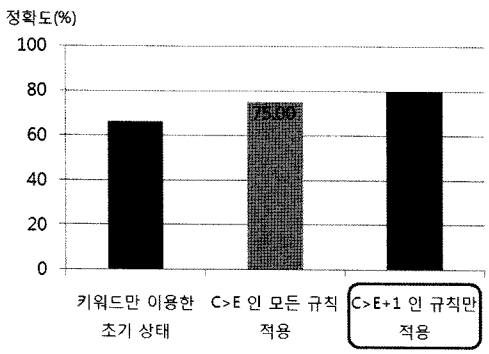

그림 3 변환 기반 학습 성능

그리고 표 2에서 기술한 규칙 틀은 키워드 외에는 품사 정보만 이용하는데, 품사 대신 아래 표 4처럼 실제 단어를 이용한 규칙 틀로도 변환 기반 학습을 적용해 보았다.

표 4 키워드와 실제 단어 정보를 이용한 변환 규칙 틀

아래와 같은 경우 비교 유형 a를 b로 변환시킨다.
1. 키워드가 k일 때
2. 1을 만족하면서, 앞(preceding) 단어가 z
3. 1을 만족하면서, 뒤(following) 단어가 z
4. 1을 만족하면서, 앞의 앞(two before) 단어가 z
5. 1을 만족하면서, 뒤의 뒤(two after) 단어가 z
6. 1을 만족하면서, 앞 단어는 z, 뒤 단어는 w
7. 1을 만족하면서, 앞 단어는 z, 앞의 앞 단어는 w
8. 1을 만족하면서, 뒤 단어는 z, 뒤의 뒤 단어는 w

그리고 표 2와 표 4에 있는 총 15개의 규칙 틀을 이용한 경우의 실험도 수행하였는데, 표 5에서 나타나듯 품사만 사용한 규칙 틀 경우가 가장 적합하다는 결론을 내릴 수 있었다.

표 5 변환 규칙 틀 변경 실험 결과

구 분	틀 수	후보 규칙 수	정확도(%)
품사 사용	8개	5,116	80.01
실제 단어 사용	8개	8,062	77.69
모두 사용	15개	13,001	79.13

또한 변환 기반 학습의 성능과 비교하기 위하여 확률 기반 모델을 이용한 실험을 병행하였다. 확률 기반 모델로는 선행 연구인 비교 문장 추출[2]에서 가장 높은 성능을 보였던 최대 엔트로피 모델을 선택하였으며, 자질도 마찬가지로 비교 문장 추출을 위해 사용했던 키워드 중심 반경 3 이내에 있는 모든 연속된 품사시퀀스 자질을 그대로 적용하였다. 예를 들어, "3G 요금이 WiMAX보다 비싸다."라는 문장의 경우 아래와 같이 16개의 자질이 생성된다.

<**보다/jca**> -> 우열
<nq **보다/jca**> -> 우열
<**보다/jca** pa> -> 우열
<jcs nq **보다/jca**> -> 우열
...
<ncn jcs nq **보다/jca** pa ef sf> -> 우열

이와 같은 자질로 비교 유형 분류에 대한 최대 엔트로피 모델을 적용한 결과는 71.25%였으며, 표 6에서 나타나듯 변환 기반 학습을 이용한 경우보다 약 9% 낮았다.

표 6 확률 기반 모델과의 성능 비교

구 분	정확도(%)
최대 엔트로피 모델	71.25
변환 기반 학습	80.01

6. 결 론

본 연구에서는 한국어 문장 특징을 참고하여 키워드만을 이용하여 초기 정보를 부착시키고, 키워드 앞 뒤 반경 2까지의 품사 정보를 참고하는 변환 기반 학습 기법을 이용하여 학습을 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 성공적인 수행은 국내 텍스트 정보처리 연구 분야의 연구 영역을 넓히고 기술 수준을 높이는 계기가 될 것이다.
2) 또한 본 연구에서 진행하는 비교마이닝 관련한 내용은 텍스트 마이닝 기술의 발전에 기여할 것이며, 이들 기술은 영역 이식성이 매우 높기 때문에 감정/의견마이닝 등 다양한 응용영역에 손쉽게 적용되어 웹 마이닝 산업 발전에 이바지할 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 계속해서 진행할 것이며, 비교 주체, 비교 대상 등 비교 관계 추출을 통한 분석 등 다양한 연구를 지속할 것이다.

참 고 문 헌

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

[2] S. Yang, Y. Jo, "Extracting Korean Comparative Sentences by Machine Learning Techniques," *Proc. of the HCLT-2008*, pp.182-287, 2008. (in Korean)

[3] N. Jindal, B. Liu, "Mining Comparative Sentences and Relations," *Proc. of the 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," *Proc. of the Computational Linguistics*, pp.543-565, 1995.

[5] L. Ramshaw and M. Marcus, "Text Chunking using Transformation-Based learning," *Proc. of the 3th ACL workshop on Very Large Corpora*, pp.82-94, 1995.

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

[7] Y. Jang, *An Information Extraction Method Using Transformation Based Learning*, Sogang University Press, Seoul, 2006. (in Korean)

[8] G. Ha, *Korean Modern Comparative Syntax*, Pijbook Press, Seoul, 1999. (in Korean)

[9] G. Ha, "Research on Korean Equality Comparative Syntax," *Proc. of the Association for Korean Linguistics*, vol.5, pp.229-265, 1999. (in Korean)

[10] K. Oh, "The Difference between 'Man-kum' Comparative and 'Cheo-rum' Comparative," *Proc. of the Society of Korean Semantics*, vol.14, pp.197-221, 2004. (in Korean)

[11] I. Jeong. "Research on Korean Adjective Superlative Comparative Syntax," *Proc. of the Korean Han-min-jok Eo-mun-hak*, vol.36, pp.61-86, 2000. (in Korean)

양 선

1995년 서강대학교 전자계산학과 학사 1995년~1999년 동양시스템즈. 2001년~2003년 삼성카드. 2007년 동아대학교 정보컴퓨터교육학과 석사. 2008년~현재 동아대학교 컴퓨터공학과 박사과정. 관심분야는 자연어처리, 텍스트 마이닝(감정, 의견, 비교 마이닝) 등

고 영 중

1996년 서강대학교 수학과 학사. 1996년~1997년 LG-EDS 근무. 2000년 서강대학교 컴퓨터학과 석사. 2003년 서강대학교 컴퓨터학과 박사. 2004년~현재 동아대학교 컴퓨터공학과 조교수. 관심분야는 자연어처리, 텍스트마이닝, 의견마이닝, 정보검색, 대화시스템, 소프트웨어공학 등