전자상거래를 위한 수사 추출 및 인식

김병주, 황도삼
영남대학교 컴퓨터공학과

Numeral Extraction and Recognition for Electronic Commerce

Byung-Joo Kim, Dosam Hwang
Dept. of Computer Engineering, Yeungnam University
and
Advanced Information Technology Research Center(AITrc)

요 약

최근 전자상거래 시스템을 이용이 많아짐에 따라, 상품의 정보나 거래를 위한 정보가 되는 수사 추출에 관한 연구가 필요하다. 수사는 표현 자체의 다양성과 다른 품사와는 구분되는 활용으로 인해 언어 분석에 있어 많은 문제점을 가지고 있지만, 일반문서에서는 발생빈도가 그다지 높지 않아 그에 관한 연구들은 적은 실정이다. 현재까지의 수사 추출에 관한 연구는 수사 어절이 다른 표현들과는 달리 어순이 뚜렷하다는 것을 이용하여 그 어순들의 결합정보의 조합을 이용하여 시도하였다. 본 논문에서는 이러한 수사 어절의 특징을 문법화함으로써 자연언어 질의에 의한 전자상거래 시스템에 관한 연구를 수행하였다.

1. 서 론[1)]

최근 인터넷 기술의 발달로 인해 원하는 정보를 웹에서 찾는 경우가 많아지고 있다. 웹에서의 정보 검색은 주로 단어 위주의 질의를 통한 검색이었다. 하지만, 자연언어처리 기술의 발달로 단어 위주의 검색에서 자연언어 질의문에 의한 검색도 가능하게 되어 가고 있다. 전자상거래 시스템에서는 자연언어 질의문과 대상 문서에서는 수사 어절이 많이 포함되어 있으며, 정보검색의 성능을 높이기 위해서는 수사 어절의 처리가 매우 중요하다.

전자상거래 시스템에서는 상품에 대한 가격 및 수요, 지불을 위한 신용카드 번호 등과 같이 상품 구입을 위한 대부분의 정보는 수사 어절로 이루어져 있다. 이러한 수사 어절에 관한 정보 처리는 수를 표현하는 단어들의 범위를 제대로 인식하여 구매자가 원하는 범위내의 제품을 검색할 수 있어야 한다. 이를 위해 우선적으로 문서에 포함된 수사 어절 및 수를 표현하는 단어들을 정확히 추출하여 수사 어절의 인식을 위한 시스템에 접목시키려고 한다.

자연언어 형태의 정보를 처리하기 위해서는 사전과 같은 언어 데이터를 바탕으로 형태소 해석 및 구문해석, 의미해석과 같은 방법을 통하여 단어 또는 구의 품사나 의미를 파악한다. 그러나, 기존의 방법에서는 전자 상거래를 위한 수사 어절 처리 및 수사의 범위를 인식하기 위한 단어나 구에 관한 연구가 미흡하였다. 수사 어절의 경우는 일반 문서에서는 그 출현 빈도가 낮으므로 수사 어절 자체를 형태소해석의 단위로 하는 것이 일반적이다[3]. 그러나 전자 상거래 시스템과 같은 특정 분야에서는 수사 어절이 가격 또는 거래를 위해 상품의 개수 등에 대한 색인어로 사용되기 때문에, 수사 어절의 표현에 대한 정규화 및 다양한 유형에 관한 정의가 필요하다. 또한 전자상거래 시스템에서 상품 정보에 대한 자연언어 질의 문장은 상품명과 가격의 범위를 인식하는 것이 가장 중요한 요소이다[4]. 가격의 범위를 인식하려면 가격 어휘와 가격 지정어로 이루어진 가격범위 구문에 대한 별도의 처리 방법이 요구된다.

2. 관련 연구

본 장에서는 한국어 수사 어절의 특징을 살펴보고, 수사 어절의 처리를 위한 현재까지의 방법을 분석한다.

1) 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음

2.1 한국어 수사 어절의 특징

사물의 수효나 차례를 나타내는 단어를 '수사'라 한다. 한국어 수사의 종류는 [표1]과 같다. 수사는 명사, 대명사와 함께 체언적 품사에 속해 체언의 특성을 지니고 있으면서도 명사, 대명사와는 구별되는 자체의 고유한 특성을 가진다[10, 11]. 수 개념을 표현하는 말에는 수사 외에도 수량 관형사가 있다. 한국어 수사가 가지는 특징을 살펴보면 다음과 같다.

① 수사는 관형사와 형용사의 수식을 받을 수 없다.[10]
예) 새 하나는 헌 둘보다 적다
② 수 개념의 말이 조사를 취하면 수사이고, 조사를 취하지 않고 다음에 오는 체언을 수식하면 관형사이다.[11]
③ 수 개념의 말에 조사가 연결되지 않아도 문장의 주기능을 가지면 수사이다.[11]
④ 고유어 수사는 고유어로 된 단위명사나 한자말로 된 단위 명사 앞에도 제한 없이 쓰이나, 한자말 수사는 고유어로 된 단위명사 앞에서는 원칙적으로 쓰이지 않는다.[1]
⑤ 수사는 복수 표시를 할 수 없다.

<표1> 한국어 수사의 종류

명칭	계열	구분	보 기
양수사(量數詞)	고유어계	정수	하나, 둘, 셋, 넷, 다섯, 여섯, 열, 스물, 서른, …
		부정수	한둘, 두셋, 서넛, 너덧, 댓, 대여섯, 예닐곱, 일여덟, 열아홉, 여남은, …
	한자어계	정수	일, 이, 삼, 사, 오, 육, 십, 이십, 삼십, 백, 천, 만, …
		부정수	일이, 이삼, 삼사, 사오, 사오륙, 오륙, 육칠, 칠팔, …
서수사(序數詞)	고유어계	정수	첫째, 둘째, 셋째, 넷째, 다섯째, 여섯째, 일곱째, …
		부정수	한두째, 두세째, 서너째, 너댓째, 너더댓째, 댓째, 대여섯째, 예닐곱째, 일여덟째, 열아홉째, 여남은째, …
	한자어계	정수	제일, 제이, 제삼, 제사, 제오, 제육, 제칠, 제팔, …

2.2 기존 연구 분석

한국어 수사 어절에 관한 관련 연구는 텍스트에서 수사 어절을 추출하는 것[3]과 수사 어절의 표기를 통일화하기 위한 알고리즘 제시[2]가 있다. [2]의 경우는 한자어계 양수사 중에서 정수[표1]들을 대상으로 한 연구였는데, 수사 어절 표기시 한 어절에서 띄어쓰기나 한글 영어의 혼합된 사용의 오류까지 수정하지는 못하고 있다. [3]은 수사 어절의 발생 패턴을 일반화하여 이를 문법화함으로써 이를 이용한 수사 표현의 처리를 시도하였다. 이는 언어 분석에서 수사 어절이 자주 분석 오류를 발생시킴에 따라 제시된 방법이다. 이 연구에서는 언어 분석 시 오분석의 수정을 목적으로 하였기 때문에 수사 어절에 나타날 수 있는 모든 품사 패턴을 문법에 포함시킴으로서 문제를 해결였다. 따라서, 수표현이 아닌 단어열이 수표현 문법을 만족하는 경우와 미등록어 단위에 대한 단위 인식 오류나, 단위후치사 인식에 있어서의 범용성 판단 기준의 미비 등이 발생하였다.

그 외 수사의 수치적 범위를 나타내는 단어들을 정의하고 이들의 범위를 정의한 연구[4]가 있다.

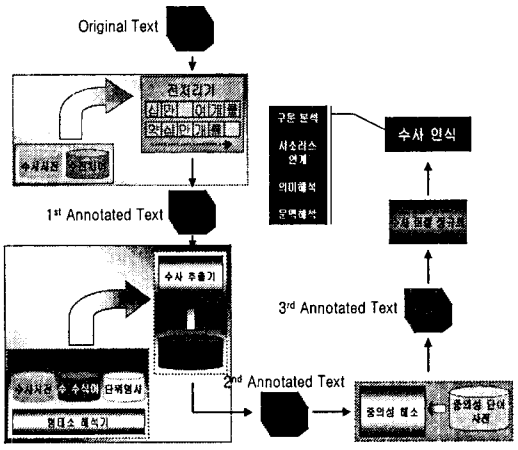
3. 수사 추출 방법 및 시스템 구성

수사 추출은 텍스트에서 수사 어절을 추출하고 이들의 문법적 요소에 따라 수표현을 위한 어절들을 인식하기 위한 연구이다. 그리고, 나아가 전자상거래 시스템에서 자연언어 질의를 통한 상품 구매를 위해서 수표현들이 가지는 수의 개념이나 의미적 요소를 판단하는 시스템에 도입하기 위한 것이다.

이를 위한 시스템 구성도는 [그림 1]과 같다.

우선적으로 수사 어절을 포함할 가능성이 있는 후보어절들을 추출하는 전처리 과정을 거쳐, 결과로 얻어진 구나 단어등을 문법과의 비교를 통해 자세하게 수사 어절들을 추정한다. 마지막으로 추출된 수사 어절이 다른 품사로 사용된 경우[예1]를 위해 중의성 해소를 위한 품사 및 문법 검증을 하여 수사 어절을 분리해낸 후, 이들을 정규화 작업을 거침으로써 통일된 표현 형식의 수사 어절을 나타낸다.

[예1] 백사, 영원, 삼삼, 영만, 영영, 영오, …



(그림 1) 시스템 구성도

3.1 전처리기

한국어 수사 어절들의 발생 패턴을 보면 수전치어 + 수표현 + 단위표현[표2]의 정규화된 형태로 정의할 수 있다. 이러한 수사 어절의 추출을 위해 우선적으로 수사 후보어절의 추출이 필요하다. 한국어 문장 전체를 대상으로 두고 수사 어절을 추출하기에는 많은 시스템의 소요가 필요하기 때문이다. 따라서, 본 시스템에서는 전처리 단계에서 한국어 문장 중 수사 및 수사를 수식하는 어절을 우선적으로 후보로 하여 추출한다.

이를 위해 필요한 방법은 다음과 같다.

1) 문장 입력 중 어절 단위 분리를 통해 어절의 첫 음절이 수사인지를 검사한다. 이는 한국어 수사의 특성상 체언임에도 불구하고 관형사, 형용사등의 수식을 받을 수 없기 때문이다.

2) 수사가 포함된 어절이 검색되면, 이들의 앞 뒤 어절을 검사하여, 수사를 수식하기 위한 단어들이 포함된 어절인지를 확인한다. [표2]에서 정의한 것과 같이 한국어 수사 어절에는 순수하게 수사로만 표현된 어절외에도 이를 수식하기 위한 어절들이 동반되는 경우가 존재하기 때문이다. 수사를 수식하기 위한 어절을 찾기 위해서는 수사 어절 앞에 위치하는 두 어절과 뒤의 네 어절을 검사한다. 이는 수전치어와 수사표현 중 띄어쓰기 오류가 많은 것을 감안하여 발생할 수 있는 최대 어절이다.

3) 위의 1), 2)의 과정을 거친 어절들을 후보로 하여 수사 표현은 우선 nnn(숫자) 또는 nn(수사)사전에 등록된 수사, nbu(단위성 의존명사)로 태깅된 단어와 그 단어를 포함하는 단어열을 수사 표현의 대상으로 인식하였다. 그러나 어형의 변화가 없는 한자어 수사와는 달리 한국어 수사 같은 경우는 "한반, 두반"등과 같이 수사 어절을 포함하고 있지만 오분석 되어 나오는 경우가 존재하기 때문에 2)의 방법을 고려하였다.

이를 위해 사용한 시스템은 한국과학기술원의 KTS(Korean Tagging System)를 사용하였다.

3.2 수사 추출기

전처리기에서 생성된 후보 어절들은 수사 표현이 아닌 어절들을 포함하고 있다. 따라서 순수한 수사어절만을 추출하기 위한 작업이 요구된다. 관련연구[3]에서 제시한 한국어 수사 표현의 발생 패턴[표2]을 바탕으로 새로운 문법을 제시한다. 수사 추출의 대상은 전처리 과정을 거친 수사 후보 어절들로 한다. 또한 이러한 표현들은 다시 세부 항목으로 나눌 수 있는데, 이들 항목의 발생패턴을 [표2]와 같이 문법으로 정리하였다. 문법 구조를 보면 수사표현은 수전치어 + 수표현 + 단위표현이라는 기본 골격을 가지고 있지만, 수전치어와 단위 표현은 생략이 가능하다. "약 십만개"라는 표현이 가능한 반면 "십만"이란 표현도 가능하기 때문이다. 그리고 수표현이란 비종단 기호는 한글 표현과 아라비아 수표현, 한글과 아라비아 수의 혼합된 표현 형태로 나타난다. 단위 표현은 다시 단위전치어, 단위표현, 단위 후치어로 세분할 수 있다[표2].

<표2> 수사 표현의 발생 패턴

한국어 수사 어절의 형태	
수사표현	수전치어 + 수표현 + 단위표현
수전치어	수 | 제 | 약
단위표현	단위전치어 + 단위명사 + 단위후치어
단위전치어	여
단위후치어	대 | 째 | 씩 | 전
수표현	아라비아수 | 한글표현 | 혼용
아라비아수	0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
한글표현	한글 표현
n	일 | 이 | 삼 | … | 하나 | 둘 | … | 한 | 첫 | …

<표3> 수사 어절 문법

NME	→ npre NE UE	NME	→ NE UE
NME	→ NE	NE	→ arn
NE	→ kon	NE	→ mixn
UE	→ upre nbu upos	UE	→ nbu
UE	→ nbu upos		

3.3 중의성 제거

수사 어절과 형태는 같지만 다른 품사로 쓰이는 단어와의 구별을 위해서 중의성을 가진 단어들의 사전을 수표현 어절과 양방향 검색을 통해 중의성을 해소한다. 미리 작성해둔 600개의 중의적 수사 사전과 수사 추출에서 얻어진 수 표현 결과들을 비교하여 문자열과 품사성보를 기본값으로 하는 검색 방법을 시도

한다.

수사와 다른 체언이 문장내에서 결합하는 품사 패턴 정보를 이용하여 서로 다른 패턴 정보를 이용한다.

4. 실험 및 평가 고찰

수사 어절을 포함시킨 테스트용 문장 200개를 대상으로 추출에 관한 실험을 하였다. 샘플 문장은 텍스트에서 발췌하였거나 연구실 학생들이 직접 생성한 문장을 이용하였다. 실험 결과 추출에서는 189문장이 성공함으로써94.5%의 재현율을 보였다.

$$\text{재현율} = \frac{\text{시스템에의한추출}}{\text{수작업에의한추출}}$$

5. 결론 및 향후 계획

본 논문은 전자 상거래 시스템에서 상품 및 가격정보 인식을 위해 우선적으로 이루어져야 할 처리 중 수사 추출에 관한 연구를 수행하였다.

이를 위해 자동색인을 목표로 수사 처리를 수행했던 기존 논문에서 제시한, 한국어 수사 어절 발생 패턴을 바탕으로 본 시스템에 맞게 문법을 구성하였다. 그리고, 수사 중에서 다른 의미로도 쓰이는 단어와의 구별을 위해 중의성 제거 시스템을 제안하였다. 중의성 제거 시스템의 경우는 수사가 다른 품사로 사용되었을 때 결합하는 품사들의 접속정보를 이용하였다. 그러나 현재까지의 시스템은 그리 많은 실험을 거치지 않아 좋은 결과를 기대하기는 어려울 것이고 앞으로 더 많은 평가와 실험을 통해 부족한 부분들을 계속 추가하여야 할 것이다. 또한, 본 논문의 추출 시스템을 바탕으로 수사 어절의 범위 인식 및 수 개념을 표현하는 단어들의 의미적 분석을 위한 연구를 수행할 계획이다.

[참고문헌]

[1]. 김민정, 권혁철, "한국어 형태소 분석에서의 수사 처리", 제3회 한글 및 한국어 정보처리 논문집, pp. 178-187, 1991.

[2] 강승식, "한국어 수사 어절의 유형 분류 및 정규화", 1999년도 한국정보과학회 가을 학술발표논문집 Vol. 26 No. 2, pp. 187~189, 1999.

[3] 한영석, "한국어 문서의 자동색인을 위한 수사, 고유명사 표현의 처리(Preprocessing numerical and title expressions for indexing Korea documents)", 제2회 과학기술정보 워크샵, 연구개발정보센터 pp.61-66, 1997. 11.

[4] 강승식, "전자거래 시스템에서 가격지정 연산자의 인식", 제11회 한글 및 한국어 정보처리 학술발표 논문집, pp. 85~88, 1999.

[5] Pierre Boullier, "Chinese Numbers, MIX, Scrambling, and Range Concatenation Grammars", In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL '99), pp. 53-60, 1999.

[6] Daniel Radzinski, "Chinese Number-Names, Tree Adjoining Languages, and Mild Context-Senstivity." In Computational Linguistics, 17(3), pp 277-299, 1991.

[7] 황도삼, 최기선, 김태석, "자연언어처리", 홍릉과학출판사.

[8] 황도삼, 최기선, 김태석, "자연언어이해", 홍릉과학출판사.

[9] 이상호, "미등록어를 고려한 한국어 품사태깅 시스템 구현", 한국과학기술원 석사학위 논문, 1993.

[10] 조규빈, "고교문법 자습서", 지학사.

[11] 곽우종, "고교국문법", 동아출판사.