단어의 문맥적 위치와 문장 유사도를 이용한 상품 특성 추출 및 계층화

김세종○ 이용훈 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과
{sejong○, yhlee95, jhlee}@postech.ac.kr

Extraction and Hierarchicalization of Product Features Using Word Contexts and Sentence Similarities

Se-Jong Kim○ Yong-Hun Lee Jong-Hyeok Lee
Dept. of Computer Science and Engineering, Division of Electrical and Computer Engineering, POSTECH

1. 서 론

특성 기반 의견 요약(feature-based opinion summarization)이란 각 의견 대상에 대한 긍정적 의견 및 부정적 의견을 선별하여 요약하는 것을 목적으로 한다. 본 분야는 주로 상품평을 대상으로 연구되어 왔으며 상품의 구성요소 및 속성과 같은 상품 특성을 의견 대상으로 삼아 이들에 대한 사용자의 의견을 종합한다. 상품 특성 추출에 관하여, Hu와 Liu[1]는 고빈도 상품 특성을 추출하고 해당 특성과 함께 나타나는 의견 단어들을 수집하여 본 단어가 포함된 문장 내에서 해당 단어와 가장 가까운 곳에 위치한 명사구를 저빈도 상품 특성으로서 추출하였다. 그들은 이후에 순차적 패턴을 활용하여 보다 효과적으로 상품 특성을 획득하였다[2]. Popescu와 Etzioni[3]는 상위 상품 특성과 패턴으로 이루어진 하위 상품 특성 식별자와 후보 하위 상품 특성 간의 상호 연관성(*PMI* : pointwise mutual information)을 측정하여 본 값이 매우 작을 경우에는 해당 후보 하위 상품 특성을 배제하였다. 이러한 기존의 상품 특성 추출 방법론은 한정된 규칙과 패턴을 사용함으로 인해서 다양한 상품 특성을 추출할 수 없고, 상품 특성들 간의 계층화를 소홀히 함으로써 보다 명료하고 구조화된 요약 결과를 제공해줄 수 없다. 본 논문은 기존 방법론의 한계를 극복하기위해 한국어 상품평 도메인에서 상품 특성을 추출하고 이를 계층화하는 새로운 방법론을 제안한다. 단어의 문맥적 위치를 이용하여 후보 상품 특성을 추출하고, 상위 상품 특성을 포함하는 문장과 후보 하위 상품 특성을 포함하는 문장 간의 유사도를 측정하여 상품 특성 선별 및 계층화를 수행한다.

2. 본 론

본 논문은 두 가지 가설을 기술한다. 가설 1은 “상품평 저자는 상위 상품 특성을 언급한 후에 하위 상품 특성을 언급한다”는 것이고, 가설 2는 “상품평 저자는 하위 상품 특성에 대한 내용을 상위 상품 특성에 대한 내용으로서 사용할 때가 많다”는 것이다.

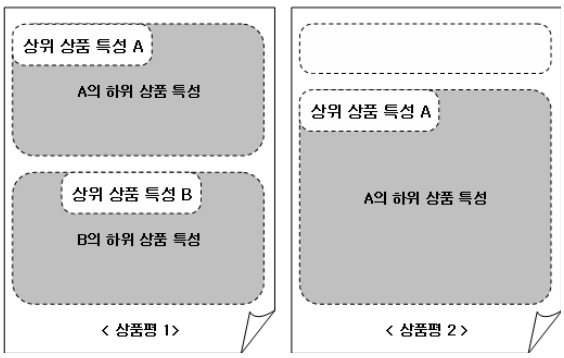

그림 1 상품 특성 단어들의 문맥적 위치(가설 1)

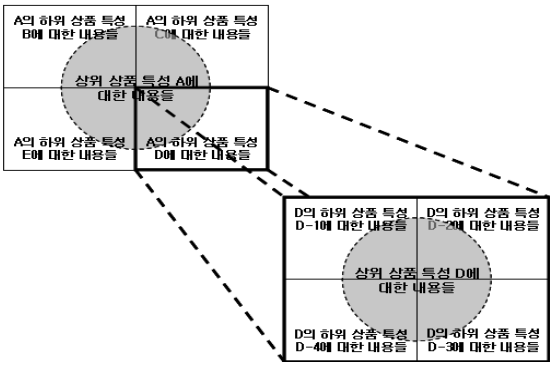

그림 2 상품 특성 내용 사이의 포함 관계(가설 2)

가설 1은 일반적인 문서 작성 방법을 고려하여 유추한 내용으로서, 상위 상품 특성이 포함된 문장에 해당 특성의 하위 상품 특성이 함께 포함될 수 있고 이후 문장에서도 하위 상품 특성이 나타날 수 있음

을 뜻한다. 초기의 상위 상품 특성은 상품 자체의 명칭을 사용하는데, 만일 상품평 도메인이 '모니터'인 경우 {모니터, B2430L, TGL-2410AT 무결점,…} 중 최소한 하나가 포함된 문장에 함께 나타나는 단일명사 및 복합명사를 후보 하위 상품 특성으로서 추출하고 이후 문장에서 나타나는 단어들도 후보 하위 상품 특성으로서 추출한다. 이렇게 추출된 후보 하위 상품 특성들은 이후 제안하는 방법론을 사용하여 선별하고 선별된 상품 특성들은 새로운 후보 하위 상품 특성들을 추출하기위한 상위 상품 특성으로서 사용한다. 이때 초기의 상위 상품 특성 등과 같은 이전 특성이 후보 하위 상품 특성을 추출하는 과정에서 나타나면 현재의 상위 상품 특성이 나타날 때까지 후보 하위 상품 특성을 추출하지 않는다.

가설 2는 상품평이 저자의 편의성 및 해당 상품에 대한 지식에 따라 다양한 표현으로 기술될 수 있음을 고려한 것으로, 상위 상품 특성이 포함된 각 문장과 후보 하위 상품 특성이 포함된 각 문장 간의 유사도를 측정하여, 상위 상품 특성이 포함된 각 문장에 대해 가장 높은 유사도를 가진 후보 하위 상품 특성을 최종적인 하위 상품 특성으로서 선별하고 이를 상위 상품 특성과 연결함으로써 상품 특성 간의 계층화를 수행한다. 이때 문장 간의 유사도는 코사인 유사도를 사용하고 문장을 이루는 모든 형태소를 비교 대상으로 삼는 것이 아니라 의존 구문 분석을 통하여 각 상품 특성과 1차 및 2차 의존 관계를 맺고 있는 보통명사, 고유명사, 의존명사, 동사, 형용사, 일반부사, 외국어만을 대상으로 한다.

본 실험은 국내 가격비교 사이트 중 하나인 에누리닷컴에서 모니터와 휴대폰에 대한 전문가 상품평과 덧글을 추출하여 본 연구실에서 개발한 언어 분석기를 통해 형태소 분석 및 의존 구문 분석을 수행한 말뭉치를 사용한다. 또한 본 논문은 자유롭게 작성된 상품평을 대상으로 상품 특성을 추출하고 계층화하는 것을 목적으로 하기 때문에 Popescu와 Etzioni[3]가 제안한 *PMI* 방법론을 비교 실험 대상으로 삼는다. 표 1은 상품 특성 추출 결과만을 비교 실험 평가한 것으로, 모니터 도메인에 대해서는 기존 방법론보다 4.37%, 휴대폰 도메인에 대해서는 5.05%의 정확률 향상을 보였다. 상대적 재현율은 모두 2배 이상의 성능 향상을 보였는데 이러한 사실은 기존의 패턴 기반 방법론이 재현율 향상 면에서 취약하다는 것을 나타낸다. 표 2는 상품 특성 계층화 결과만을 비교 실험 평가한 것으로, 모니터 도메인에 대해서는 3.79%, 휴대폰 도메인에 대해서는 21.94%의 정확률 향상을 보였다. 특히 휴대폰 도메인에서 기존의 *PMI* 방법론의 성능이 낮게 나온 이유는 초기에 잘못된 하위 상품 특성들을 많이 추출함으로써 이후 처리과정에 잘못된 상위 상품 특성과 계층 관계를 가진 하위 상품 특성을 추출하게 되었기 때문이다.

표 1 상품 특성 추출 비교 실험 결과 (단위:%)

도메인	적용 방법론	정확률	상대적 재현율
모니터	*PMI*	82.76	1
	제안 방법	87.13	3.67
휴대폰	*PMI*	74.92	1
	제안 방법	79.97	2.35

표 2 상품 특성 계층화 비교 실험 결과 (단위:%)

도메인	적용 방법론	정확률	상대적 재현율
모니터	*PMI*	65.52	1
	제안 방법	69.31	3.68
휴대폰	*PMI*	33.22	1
	제안 방법	55.16	3.65

3. 결 론

본 논문은 빈도수 및 패턴 등을 기반으로한 기존 방법론들의 한계점을 극복하고자, 단어의 문맥적 위치를 이용하여 후보 상품 특성을 추출하고 문장 유사도를 이용한 상품 특성 선별 및 계층화를 수행하였다. 하지만 본 방법론은 중복되어 나타나는 상품 특성 및 유의어나 동의어의 처리를 고려하지 않아 최적화된 계층 정보를 구축할 수 없었고 잘못된 상품 특성 추출로 인해 이후 처리 단계에서 발생하는 누적된 오류를 해결할 수 없었다. 앞으로의 연구에서는 해당 방법론을 보다 적절하게 구현할 수 있는 알고리즘을 개발하고 상품 특성들의 재계층화 및 최적화 단계를 추가할 계획이다.

감사의 글

본 논문은 2010년도 두뇌한국21사업, 포항공과대학교 정보통신연구소 자체 학술연구과제(선도과제), 그리고 한국과학재단 기초연구사업(No. 2010-0012662)의 지원으로 수행되었습니다.

참고문헌

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168-177, 2004.

[2] M. Hu and B. Liu, "Opinion feature extraction using class sequential rules," AAAI, 2006.

[3] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, pp. 339-346, 2005.