일반적, 영역 의존적 특성을 반영한 감정 자질의 의미지향성 추정 방법

황재원○
동아대학교
컴퓨터공학과
sftcap@gmail.com

고영중
동아대학교
컴퓨터공학과
yjko@dau.ac.kr

A Semantic Orientation Prediction Method of Sentiment Features Based on the General and Domain-Dependent Characteristics

Jaewon Hwang○
Computer Engineering,
Dong-A University

Youngjoong Ko
Computer Engineering,
Dong-A University

요 약

본 논문은 한국어 문서 감정분류를 위한 중요한 어휘 자원인 감정자질(Sentiment Feature)의 의미지향성(Semantic Orientation) 추정을 위해 일반적인 특성과 영역(Domain) 의존적인 특성을 반영하여 한국어 문서 감정분류(Sentiment Classification)의 성능 향상을 얻을 수 있는 기법을 제안한다. 감정자질의 의미지향성은 검색 엔진을 통해 추출한 각 감정 자질의 스니핏(Snippet)과 실험 말뭉치를 이용하여 추정할 수 있다. 검색 엔진을 통해 추출된 스니핏은 감정자질의 일반적인 특성을 반영하며, 실험 말뭉치는 분류하고자 하는 영역 의존적인 특성을 반영한다. 이렇게 얻어진 감정자질의 의미지향성 수치는 각 문장의 감정 강도를 추정하기 위해 이용되며, 문장의 감정 강도의 값을 TF-IDF 가중치 기법에 접목하여 감정자질의 가중치를 책정한다. 최종적으로 학습 과정에서 긍정 문서에서는 긍정 감정자질, 부정 문서에서는 부정 감정자질을 대상으로 추가 가중치를 부여하여 학습하였다. 본 논문에서는 문서 분류에 뛰어난 성능을 보여주는 지지 벡터 기계(Support Vector Machine)를 사용하여 제안한 방법의 성능을 평가한다. 평가 결과, 일반적인 정보 검색에서 사용하는 내용어(Content Word) 기반의 자질을 사용한 경우보다 3.1%의 성능 향상을 보였다.

주제어: 감정분류, 의미지향성, 감정자질

1. 서 론

웹(web)의 출현과 인터넷, 데이터 베이스(database) 상의 디지털 컨텐트(digital content)의 급속한 증가로 인해 정보 검색(information retrieval)과 자연어 처리 영역에서 문서 분류(text classification)에 대한 관심이 증대되고 있다. 전통적으로 문서 분류 작업들은 문서의 주제(topic)에 따른 분류에 초점이 되어 진행되어 왔다[1]. 그러나, 최근 단지 주제에 따른 분류가 아닌 문서의 저자에 대한 감정, 의견, 의향을 분류하고자 하는 노력이 빠르게 증가되고 있다. 이 영역의 핵심은 목표 대상(책, 상품, 등)에 대한 긍정(positive) 또는 부정(negative)의 여부를 문서에 할당해 감정을 분류하는 것이다. 왜냐하면 이러한 작업은 감정 분석의 기본이 되며 넓은 적용 가능성을 가지고 있기 때문이다. 예를 들어, 우리가 특정 상품에 대한 다른 사람들의 의견을 통해 상품의 구매 여부를 결정하고자 한다면 위와 같은 분석은 상품에 대한 다른 사람들의 긍정/부정에 대한 의향을 쉽게 제공해 상품을 구매하는데 도움을 줄 수 있을 것이다.

현재까지 감정분류(sentiment classification)에 대한 많은 연구들이 수행되었지만 이러한 연구들은 크게 2가지 범주로 나눌 수 있다. 첫 번째는 기계 학습(machine learning) 기법이다[1]. 기계 학습 방법은 문서 내에 다양한 단어들의 출현 빈도에 기반하여 분류기를 학습시키는 기법이다. 다른 접근법은 의미지향성(semantic orientation)이다[2]. 이 기법은 단어들을 긍정 또는 부정의 2개의 범주로 분류하고 문서 내에 출현한 단어의 긍정/부정의 값들을 계산하는 기법이다. 지금까지 단어의 감정적인 극성을 식별하는 많은 연구가 수행되었다[2,3]. 예를 들어, “아름다운”은 긍정 지향적이며, “더러운”은 부정 지향적이다. 본 논문에서는 선행연구[4]를 통해 얻은 감정자질을 사용하며 감정자질(sentiment feature)의 의미지향성을 추정하고자 한다. 감정자질은

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

감정분류의 기본이 되는 자원이며 응용(application)에 대한 큰 잠재력을 가지고 있을 것이라 믿어진다. 그러나, 감정자질을 어떻게 효과적으로 이용하여 감정분류의 성능을 높일 것인지에 대한 문제는 여전히 남아있다.

비록 감정을 지닌 단어라 할지라도 일반적으로 사용될 때와 특정 상품이나 정책 등에 대한 의견 및 감정을 표현할 때는 차이가 있다. 저자는 감정을 지닌 단어를 일반적으로 사용할 경우에 비해 특정 도메인에 대해 자신의 의견이나 감정을 표현할 경우 비슷한 감정을 지닌 단어들을 함께 사용하게 될 가능성이 높다. 본 논문에서는 이와 같은 특징을 감정자질의 의미지향성에 반영하고자 한다.

먼저, 검색 엔진[2]을 통해 얻은 감정자질의 스니핏을 통해 감정자질의 일반적인 특성을 반영하고, 실험 말뭉치를 통해 영역(domain) 의존적인 특성을 함께 반영한다. 문서 내에 나타나는 문장들 중에는 해당 문서의 감정을 잘 나타내는 문장과 그렇지 못한 문장들이 있으며, 이러한 문장 감정 강도의 차이는 각 문장에 나타나는 감정자질의 중요도에도 영향을 미친다. 그러므로, 본 논문에서는 감정자질의 의미지향성을 이용하여 문장이 지닌 감정의 강도를 추정하여 선행 연구[4]의 자질 가중치 기법에 적용하여 성능을 향상시키고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 앞서 연구된 관련 연구와 배경에 대해 살펴보고, 3장에서는 감정자질의 의미지향성 추정 및 문장의 감정 강도를 반영한 감정자질의 가중치 책정과 기계 학습 방법에 대해 자세히 설명한다. 그리고 4장에서는 실험 및 평가를 하며, 5장에서는 결론 및 향후 연구를 기술한다.

2. 배경 및 관련 연구

감정분류는 영화 리뷰, 상품 리뷰, 고객 피드백(feedback) 리뷰와 같이 다양한 영역에서 시도되고 있다[1,2,5]. 또한 분류의 대상이 문서뿐만 아니라, 문장[5,6], 그리고 문장의 감정 패턴 분석을 통해 문장의 여러 감정적 표현을 인식하고 분류하는 연구도 수행되었다[7]. 현재까지 여전히 많은 연구들이 지지 벡터 기계(support vector machines)와 같은 기계 학습 기법들에 초점이 맞추어져 수행되었다. 그리고 긍정/부정 term-counting 기법과 단어의 긍정 또는 부정을 자동적으로 결정하는 연구[8]도 수행되었다.

감정분류 역시 문서분류의 한 영역이기 때문에 분류를 위한 자질의 추출도 중요한 문제이다. 영어권 선행 연구에선 감정 분류에 적합한 자질을 추출하는 연구[7]와 어휘 자원을 이용하여 감정 자질의 가중치를 결정하는 연구[9]도 수행되었고, 한국어 권 연구에서는 한국어 감정자질을 추출하는 연구[10]가 수행되었다.

일반적으로 사용되는 벡터 공간 모델(vector space model)의 단점을 보완하기 위해 제목과 문장 간의 유사도를 이용하여 중요한 문장을 결정하여 자질의 가중치에 적용하는 연구가 수행되었다[11].

2.1 의미지향성

단어의 의미지향성 추정에 관한 연구는 Hatzivassiloglou와 McKeown에 의해 시작되었다[2]. 의미지향성 추정을 위해서는 비지도 학습 알고리즘이 사용된다. 이 알고리즘은 대표적인 7개의 긍정/부정 단어로부터 시작되며 알타비스타 검색 엔진의 NEAR 연산을 이용해 7개의 긍정/부정 단어들 근처에서 검색 단어가 등장하는 문서가 얼마나 많은지로 단어의 의미지향성을 추정하게 된다.

2.2 감정분류를 위한 기계 학습

문서를 긍정/부정으로 분류하는 가장 일반적인 방법 중의 하나가 기계 학습 알고리즘을 문서를 분류하기 위해 학습하는 것이다. 몇 가지의 ML 알고리즘이 비교[1,5]되었고 SVM[12]이 다른 분류기에 비해 보다 나은 성능을 보였다. 유니그램(unigram), 바이그램(bigram), POS(part of speech) 정보, 단어의 위치 정보들이 자질로 사용되었지만 단지 유니그램만을 사용한 경우에 가장 좋은 성능을 보였다.

3. 감정자질의 의미지향성 추정 방법 및 적용

3.1 일반적 의미지향성 추정

일반적 의미지향성을 추정하기 위해서는 일반적으로 사용되는 표현들의 현상을 반영해야 한다. 이러한 표현들은 검색 엔진을 통해 얻을 수 있으며 PMI(pointwise mutual information) 수치를 통해 의미지향성을 추정할 수 있다[2]. 그러나, Google API에서 NEAR 연산이 제공되지 않기 때문에 의미지향성을 찾고자 하는 단어의 스니핏을 대상으로 PMI 수치를 계산한다.

두 단어 w_1과 w_2의 PMI 수치는 식 (2)와 같이 두 단어가 동시에 출현한 확률에 각 단어가 출현한 확률을 나누어 얻을 수 있다.

$$PMI(w_1,w_2)=\log\frac{P(w_1,w_2)}{P(w_1)P(w_2)}=\log\frac{hits(w_1,w_2)N}{hits(w_1)hits(w_2)} \quad (2)$$

단어의 의미지향성에 관한 식은 다음과 같이 표현할 수 있다.

$$SO\text{-}PMI(word)=PMI(word,p\_query)-PMI(word,n\_query) \quad (3)$$

긍정/부정 참조 단어는 영어권에서 각 7개씩 사용하였지만 한국어는 긍정/부정을 나타내는 단어들이 더 다양하기 때문에 영어권 대표 단어 7개를 한영사전을 통해 각 63개씩으로 확장을 하였다. 그 예는 [표 1]과 같다.

2) Google Open API

[표 1] 확장된 참조 단어의 예

p_query	좋은, 만족한, 여유로운, 아름다운, 행복한, 훌륭한, 친절한, 맛있는, 잘하는, …
n_query	나쁜, 불량한, 해로운, 불쾌한, 불길한, 더러운, 지저분한, 불행한, 괴로운, …

NEAR 연산을 사용할 수 없기 때문에 모든 문서를 대상으로 Window Size를 잡아 계산하는 것을 비효율 적이다. 그렇기 때문에 검색 단어가 반드시 포함되는 스니핏을 NEAR 연산의 범위로 고려하여 PMI 수치를 계산할 수 있다. 각 단어마다 검색 엔진이 반환하는 상위 1,000개의 스니핏을 사용하였으며, 검색 대상이 되는 전체 문서 수(N)는 무시할 수 있다. 변경된 최종 식은 다음과 같다.

$$SO\text{-}PMI(word) = \log \frac{\sum_i Snippet_i(word, p\_query)}{\sum_i Snippet_i(word, n\_query)} \quad (4)$$

*Snippet(word,p_query)*은 단어의 스니핏에서 *p_query*를 포함하는 횟수를 반환하는 함수이다. 식 (4)를 통해 감정자질의 일반적인 의미지향성을 추정할 수 있다.

3.2 영역 의존적 의미지향성 추정

분류하고자 하는 영역에서의 의미지향성은 해당 영역의 학습 말뭉치를 통해 반영할 수 있다. 학습 말뭉치 내에서 의미지향성은 추정하고자 하는 단어가 포함된 문장만을 대상으로 식 (4)를 적용하여 추정한다. 하지만, 이 방법은 학습 말뭉치가 필요하지만 해당 문서가 긍정/부정인지에 대한 레이블(label)은 필요가 없다는 장점이 있다.

3.3 일반적 + 영역 의존적 의미지향성

본 논문에서 사용한 의미지향성 결합 방법은 두 값 중에 큰 값을 이용하는 방법이다. 단어의 의미지향성은 그 단어가 비슷한 뜻을 가진 단어들과 함께 많이 쓰였을 경우에 의미지향성이 증가하기 때문에 높은 값을 사용함으로써 성능향상을 얻을 수 있었다.

3.4 감정자질 및 문장의 감정 강도 계산

본 연구에서 사용된 감정자질과 가중치 책정기법은 선행연구[4]에서 제한한 확장된 감정자질과 개선된 가중치 책정기법을 사용하였다. 이렇게 추출된 감정자질의 의미지향성을 추정하여 개선된 가중치 책정기법에 적용하여 성능을 향상시키는 것이 본 논문의 목표이다.

4. 실험 및 결과

본 논문에서는 SVM Light[13]를 사용하였다.

4.1 실험 말뭉치

실험에 사용된 문서 말뭉치는 총 2,440개의 문서이며, 3개의 분야를 나누어 수집하여 신문기사 721개, 영화리뷰 1,323개, 상품리뷰 396개의 문서로 실험하였다. 모든 문서를 사람이 직접 읽고 감정 여부를 판단하여 실험 말뭉치를 구축하였다.

[표 2] 실험에 사용한 실험 말뭉치

분야	긍정	부정	총합
신문기사	402	319	721
영화리뷰	715	608	1323
상품리뷰	216	180	396
총합	1333	1107	2440

4.2 성능평가 방법

본 논문에서는 5-fold cross validation 방법으로 실험을 하였으며, 인터넷 사이트상에서 수집된 문서 집합의 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확률(precision)과 재현율(recall)을 사용하였다.

4.3 실험 결과

4.3.1 전처리 과정

본 연구에서는 내용어(content word)로서 형태소 분석의 결과 중 명사, 형용사, 동사, 부사만을 고려하였다. 전처리 과정을 통해 입력 문서는 문장 단위로 내용어를 추출하게 되고, 추출된 내용어를 사용하여 문장 벡터들을 구성한다.

4.3.2 실험 환경

실험은 아래의 3가지의 의미지향성 값을 적용하여 실험하며, 카테고리별 선택적 가중치 강화 기법 적용[4]의 유/무로 나누어서 실험한다.

실험1) 일반적 의미지향성
실험2) 영역 의존적 의미지향성
실험3) 일반적 + 영역 의존적 의미지향성

4.3.3 기본 시스템

실험은 실험 말뭉치의 카테고리의 구분 없이 실험하였으며, 내용어를 사용한 실험(Base1)에서는 77.08%의 결과를 얻었으며 선행연구[4]를 통한 실험(Base2)에서는 78.09%의 성능을 얻었다. 본 논문에서는 위의 두 가지를 기본 시스템으로 한다.

4.3.4 일반적 의미지향성 반영 실험

[표 3] 실험1) 결과(F_1-Measure)

구분	강화기법미적용	강화기법적용	비고
실험1	***78.83***	***79.51***	+0.68
Base1	77.08		+2.43
Base2	78.09		+1.42

감정자질에 일반적 의미지향성을 반영한 실험 결과는 [표 3]과 같다. 가중치 강화 기법을 적용한 경우가 적용하지 않은 경우보다 0.68%의 성능 향상을 보였으며, 두 가지의 기본 시스템들보다 각 2.43%, 1.42%의 나은 성능을 보였다.

4.3.5 영역 의존적 의미지향성 반영 실험

실험 결과는 [표 4]와 같다.

[표 4] 실험2) 결과(F_1-Measure)

구분	강화기법미적용	강화기법적용	비고
실험2	***78.88***	***80.09***	+1.12
Base1	77.08		+3.01
Base2	78.09		+2.00

영역 의존적 의미지향성을 반영한 실험 역시 가중치 강화 기법을 적용한 경우가 그렇지 않은 경우보다 나은 성능을 보였으며, 기본 시스템보다 나은 성능을 보였다.

4.3.6 일반적 + 영역 의존적 의미지향성 반영 실험

실험 결과는 [표 5]와 같다.

[표 5] 실험3) 결과(F_1-Measure)

구분	강화기법미적용	강화기법적용	비고
실험3	***78.98***	***80.18***	+1.20
Base1	77.08		+3.10
Base2	78.09		+2.09

두 가지의 특성을 모두 반영한 결과 실험1과 실험2에서 단독으로 사용한 경우보다 모두 나은 성능을 얻을 수 있다.

모든 경우에 가중치 강화 기법을 적용하는 것이 성능 향상에 도움이 되며, 실험 말뭉치가 존재한다면 영역 의존적인 특성을 이용하는 것이 성능 향상에 도움이 된다는 결과를 얻었다. 그리고 일반적 특성과 영역 의존적 특성을 모두 반영한 실험에서 가장 높은 성능을 얻을 수 있었다. 이는 내용어를 사용한 실험(Base1)보다 3.1% 향상된 성능이다.

5. 결론 및 향후 연구

본 논문에서는 감정자질의 일반적 특성과 영역 의존적인 특성을 고려하여 감정자질의 가중치 책정에 반영하여 한국어 문서 감정분류의 성능을 향상시키는 방법을 제안하였다. 검색 엔진을 통해 반환되는 문서로 단어의 일반적인 의미지향성을 추정할 수 있으며, 분류하고자 하는 영역의 실험 말뭉치를 통해 영역 의존적인 의미지향성을 추정할 수 있다. 이 두 가지 특성을 적절하게 결합한 결과 기본 시스템에 비해 3.1%의 성능향상을 얻을 수 있었다.

향후 연구로는 일반적 의미지향성과 영역 의존적 의미지향성을 효과적으로 결합할 수 있는 방법을 연구할 것이다.

참고 문헌

[1] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," *In Proceedings of the EMNLP*, pp.79-86, 2002.

[2] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," *In Proceedings of the 35th ACL/8th EACL*, pp.174-181, 1997.

[3] P.D. Turney and M.L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *In Proceedings of the ACM Transactions on Information Systems*, pp.315-346, 2003.

[4] 황재원, 고영중, "문장 감정 강도를 반영한 개선된 자질 가중치 기법 기반의 문서 감정 분류 시스템", *한국정보과학회논문지*, 소프트웨어 및 응용 제36권 제6호, pp.491-497, 2009.

[5] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *In Proceedings of the ACL*, pp.271-278, 2004.

[6] Y. Mao and G. Lebanon, "Isotonic Conditional Random Fields and Local Sentiment Flow," *In Proceedings of the NIPS*, 2007.

[7] A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," *In Proceedings of the CIKM*, pp.617-624, 2005.

[8] P. Turney and M. Littman, "Unsupervised learning of semantic orientation from a hundred-billon-word corpus," Technical Report ERB-1094, National Research Council, Institute for Information Technology, 2002.

[9] A. Esuli and F. Sebastiani, "PageRanking WordNet Synsets: An Application to Opinoin Mining", *In Proceedings of the ACL*, pp.424-431, 2007.

[10] 황재원, 고영중, "감정 분류를 위한 한국어 감정 자질 추출 기법과 감정 자질의 유용성 평가", *한국정보과학회논문지*, 컴퓨팅의 실제 및 레터 제14권

제3호, pp.336-340, 2008.
[11] Y. Ko, J. Park, and J, Seo, "Automatic Text Categorization using the Importance of Sentences", *In Proceedings of the 19th International Conference on COLING*, pp.474-480, 2002.
[12] C. Cortes and V. Vapnik "Support-Vector Networks," Machine Learning, Vol.20, pp. 273-297, 1995.
[13] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many relevant Features," *In Proceedings of the ECML*, pp.137-142, 1998.