의견의 발안자를 찾기 위한 어휘점수의 부여와 확장

정헌영 ○, 김준기, 이예하, 이종혁

포항공과대학교 전자컴퓨터공학부 컴퓨터공학과

{blesshy○, yangpa, sion, jhlee}@postech.ac.kr

Expansion of Candidate Lexical Score for Opinion Holder Identification

Hun-young Jung○ , Jungi Kim, Yeha Lee, Jong-Hyeok Lee

Department of Computer Science and Engineering
Division of Electrical and Computer Engineering
Pohang University of Science and Technology

요 약

의견의 주체를 찾는 일은 의견 분석의 결과를 활용 하는데 있어 필수적인 분야이다. 본 논문은 발안자를 찾는 시스템의 성능을 높이기 위해 이전논문에 제안하였던 단어에 의견주체의 후보로서의 점수를 부여하는 방법을 개선하였고 미등록어 문제를 해결하기 위해 taxonomy에 의존하여 기존단어의 점수를 이용하는 방법을 제안하였다. 본 논문에서 제안한 방법은 Baseline과 비교하여 F1값이 18.9% 증가하였다.

1 서 론

의견분석은 문서에서 사실정보를 찾는 것이 아니라 의견이나 감정의 표현과 관련된 정보를 찾는 분야이다. 의견분석의 세부 분야로 의견을 포함하는 문장이나 문서를 찾는 주관성분리, 의견의 방향이 긍정인지 부정인지 혹은 중립인지를 판별하는 감정분류, 의견의 발안자와 대상 등의 의견과 관련된 정보를 추출하는 분야가 있다. 본 논문에서는 이러한 다양한 분야 중 의견의 발안자를 판별하는 것에 초점을 두었다.

의견의 발안자는 주어진 문서의 의견을 나타내는 주체를 말한다. 예를 들어 “인권위원회는 중국의 인권의 수준이 낮다고 평가하였다.”라는 문장의 경우 의견의 표현은 ‘낮다’이고 의견의 대상은 ‘중국의 인권의 수준’, 의견의 발안자는 ‘인권위원회’가 된다. 또한 의견의 발안자는 직접적인 발안뿐만 아니라 간접적인 발안도 해당된다. “그 보고서에 따르면 제품의 소비자 만족도가 높다.”의 경우 의견인 ‘만족도’의 직접적인 발안자는 ‘소비자’가 되지만 ‘보고서’도 의견의 간접적인 발안자로 볼 수 있다.

의견의 발안자를 찾아낸 결과는 여러 분야에 응용될 수 있다. 예를 들어 공통적인 발안자, 대상을 모아서 의견을 요약하거나[1], 의견과 관련된 질문에 대답하기 위한 질의응답 시스템[2]이 있다. 그리고 의견의 신뢰도를 판단하기 위한 연구도 진행되고 있다[3].

최근 들어 문서, 문장의 의견을 분석하는데 대한 관심이 커짐에 따라 의견분석에 대한 학회가 열리고 있다. NTCIR은 지난 2007년 6번째 학회에서부터 의견분석에 대한 Workshop을 추가하여 의견분석을 위한 말뭉치를 배포하고 이에 기반한 성능평가를 하고 있다.

발안자를 찾기 위한 방법은 정보분석(Information Extraction)이나 의미역 결정(Semantic Role Labeling)에 기반하여 시작되었다. [2]에서는 기존에 의미역 결정에서 사용되는 패턴을 자동으로 추출하는 방법을 사용하되 기존에 사용되는 의미역 태그가 아닌 의견과 발안자에만 관련된 태그를 정하여 사용하였다.

최근에 진행되는 발안자를 찾기 위한 방법은 크게 규칙기반 방법과 확률기반 방법으로 나눌 수 있다. 규칙기반 방법은 문장의 표층적인 형태를 기준으로 특정 단어나 품사에 기반한 규칙을 수동으로 만들어서 문장에 적용하는 방법이다. 이에 대한 대표적인 연구로는 NTCIR6에서 6개의 규칙을 기반으로 한 것으로 NTCIR6에서 의견의 발안자를 찾는 연구 중 가장 좋은 성능을 낸 방법이다[4].

확률기반 방법으로 각 단어나 구를 발안자와 그렇지 않은 것으로 분류하는 방법과 순차적인 태깅을 하는 방법이 있다. 분류를 사용한 방법으로 [5]가 있는데, 기계학습에 사용되는 Maximum Entropy (ME) 모델에 기반하여 문법적인 정보를 사용한 방법이다.

순차적인 태깅에 기반한 방법은 각 단어에 의견의 발안자가 표시된 구절의 시작단어인지 중간에 있는 단어인지, 발안자를 나타낸 구절에 포함되지 않는 단어인가를 표시하여 연속적인 단어의 묶음으로 의견의

발안자를 찾는 방법으로 주로 Conditional Random Field (CRF)가 사용된다[6]

최근에는 이러한 방법을 혼합하여 확률적인 방법으로 의견의 발안자가 문장에 명시적으로 드러나 있는지를 판단하고 명시적으로 표현된 문장에 대해서 규칙에 기반한 방법을 적용하는 연구가 있었다[7]. 또한 기계학습적인 방법을 사용하는데 있어서 규칙기반에 활용되는 패턴에 사용되는지를 자질로서 사용하는 연구가 있었다[8].

이전 연구에서는 각 단어에 의견의 발안자로서 가능성에 대한 점수를 부여하고 이를 ME 모델에서 사용하기 위한 방법에 대해 연구하였다[9].

본 논문에서는 이전연구에서 제안하였던 방법을 개선하기 위해 말뭉치에서 점수를 부여하는 방법을 개선하고 미등록어 문제를 해결하기 위해 단어와 단어 사이에 유사도를 사용한 방법을 제안하였다.

2 접근방법

본 논문에서는 문장에서 의견의 발안자를 찾기 위해 기계학습 방법을 사용하여 각 단어가 의견의 발안자로 나타나는 확률을 계산하는 방법을 사용하였다. 이러한 확률을 계산하기 위하여 ME 모델을 사용하였다. ME 모델은 주어진 조건에서 Entropy를 최대화 하는 방향으로, 즉 제약조건 이외의 부분에 대해서는 동일한 확률을 부여하는 모델로서 자연어처리 같이 본래의 확률분포를 특정하기 힘든 상황에서 많이 사용된다[5]. 본 논문에서는 주어진 문장에서 각 단어가 의견의 발안자를 나타낼 확률을 계산한 후 그 확률이 일정 값 이상을 넘어서는 경우 해당 단어가 의견의 발안자를 나타낸다고 분류하였다.

2.1 점수 부여 방법의 개선

의견의 발안자를 나타내는 명사구는 '연구원', '경제학자', '회장' 같이 의미적으로 '생각'이나 '발언' 등의 동작이 가능한 단어나 '보고서'처럼 '분석'의 내용을 포함하는 단어가 나타난다. 반대로 이러한 동작이나 의미를 내포하지 않는 단어만으로 이루어진 명사구는 의견의 발안자로서 사용되지 않는다. 각 단어가 이러한 특성을 가지고 있는지 판단하기 위해서 말뭉치로부터 점수를 계산하였다.

이전 연구에서는 주어진 단어 'w'에 대해서 단어가 발안자로서의 특성을 얼마나 높게 가지고 있는지를 판별하기 위하여 학습 말뭉치를 사용하여 다음과 같은 점수를 부여하였다[9].

$$Score(w) = \frac{count(w \mid holder, corpus)}{count(w \mid corpus)}$$

단어가 한 문서에서는 의견의 발안자로서 자주 사용되는 반면 다른 문서에서는 발안자로 사용되지 않는 경우가 있다. 그러나 한 문서에서라도 발안자로서 자주 사용된다면 그 단어는 발안자로서의 특성을 많이 가지고 있는 단어이다. 따라서 위와 같은 방법은 여러 문서에서 등장하는 단어를 혼합하여 사용하기 때문에 단어의 특성을 정확하게 판별할 수 없다. 이러한 문제를 해결하고 더 정확하게 단어의 특성을 추측하기 위하여 단어에 점수를 부여하는 방법을 다음과 같이 변경하였다.

$$Score(w) = \max\{score_{doc \in corpus}(w, doc)\}$$

$$Score(w, doc) = P(holder \mid w, doc)$$

$$= \frac{P(w \mid holder, doc)P(holder \mid doc)}{P(w \mid doc)}$$

$$= \frac{\frac{count(w \mid holder, doc)}{size(holder \mid doc)} \times \frac{size(holder \mid doc)}{size(doc)}}{count(w \mid doc)/size(doc)}$$

$$= \frac{count(w \mid holder, doc)}{count(w \mid doc)}$$

2.2 미등록어의 점수부여

위와 같은 점수 부여방법은 전적으로 학습 데이터에 의존하는 지도학습방법이기 때문에 학습 데이터에 등장하지 않는 단어에 대해서는 점수를 부여할 수 없다. 이를 해결하기 위해서 학습 말뭉치에 등장하는 단어와 그에 부여된 점수를 이용하여 새로운 단어의 점수를 추측하는 일이 필요하다. 이를 위해서 단어 사이의 유사성을 측정하여 사용하였다.

단어 사이의 유사성을 측정하는데 여러 문서에서 각 단어가 등장하는 빈도를 행렬로 표현하여 이용하는 방법이나 사전에서 유의어를 사용하는 방법이 있다. 본 논문에서는 단어의 의미에 따라 계층관계를 구성하는 taxonomy를 사용하여 단어 사이의 유사성을 측정하였다. 이들 중 많이 사용되고 있는 taxonomy에서 두 단어 사이의 거리를 이용한 방법을 사용하였다. 유사성은 다음과 같은 방법으로 측정된다.

먼저 상의어/하의어 관계만을 사용하여 한 단어로부터 다른 단어로의 Path를 찾고 유사도를 다음과 같이 계산한다.

$$Similarity(w1, w2) = \frac{1}{dis(path) + 1}$$

여기서 *'dis(path)'*는 두 단어 사이에 설정된 Path의 길이이다.

이를 이용하여 새로운 단어 *'us'*의 점수는 다음과 같이 계산된다.

$$w = \arg\max_{w}\{dis(path)\}$$

$$Score(us) = similarity(us, w) \times Score(w)$$

2.3 점수의 사용방법

이전 연구에서는 각 단어에 부여한 점수를 ME 모델의 자질로 사용하였다[9]. 본 논문에서는 이전 연구에서와 같이 자질로서 사용하는 방법뿐 아니라 이 점수를 각 단어의 선행점수로 간주하여 ME 모델로 계산한 확률에 일정 비율을 곱하여 더하는 방법(interpolation)을 적용하여 자질로서 사용하는 방법과 비교하여 보았다.

3 실험 방법

3.1 Baseline

성능향상의 기준을 정하기 위해서 [6]에서 사용된 자질들 중 일부를 사용하여 ME 모델을 학습하였다. 각 자질들은 다음과 같다

표 1 시스템에 사용된 자질들

자질	의미
F_1	단어의 첫 글자가 대문자인가
F_2	단어에 대문자가 포함되어 있는가
F_3	단어의 형태소 태그 정보(Part of Speech)
F_4	단어가 의견을 표현하는 단어인지의 여부
F_5	단어를 포함하는 구의 문법적 역할
F_6	선행단어를 포함하는 구의 문법적 역할

단어의 형태소 정보는 해당 단어뿐 아니라 전후단어의 정보도 사용하였다. 본 논문에서는 [-2.+2] 범위의 단어의 형태소 태그 정보를 F_3으로 사용하였다. F_4에서 단어가 의견을 표현하는 단어인지를 판별하기 위해서 Senti-WordNet[1]의 점수를 사용하였다. Senti-WordNet은 각 단어에 대해서 긍정점수와 부정점수를 부여하는데 이 두 점수 중에서 높은 점수를 단어의 의미표현 점수로 사용하여 10단계로 표현하였다.
또한 구의 문법적 역할에 대한 정보를 알기 위해서는 문장을 파싱(parsing)해야 하는데 이를 위해서 Stanford parser[2]를 사용하였다. ME 모델의 구현에 있어서는 기존에 개발되어 사용되고 있는 "Maximum Entropy Modeling Toolkit for Python and C++"[3] 를 사용하였다.

3.2 실험 데이터

본 논문에서는 학습데이터로서 NTCIR7에서 학습 데이터로 제공된 말뭉치를 사용하였다. 이 말뭉치는 241개의 문서를 포함하며 이중 3836개의 문장이 의견을 나타내는 문장이다. 평가 데이터로서는 NTCIR7에서 평가데이터로 제공된 말뭉치를 사용하였다. 이 말뭉치는 142개 문서를 포함하며 이중 2721개의 문장이 의견을 나타내는 문장이다.

3.3 평가방법

말뭉치에 있는 문장의 각 단어에 대해서 의견의 발안자를 표현하는 단어와 그렇지 않은 단어를 분류하여 표시하고 각 단어를 ME 모델로 분류한 결과를 구 단위가 아닌 단어 단위에서 측정하여 비교하였다. 성능의 평가를 위해서 본 논문에서는 정확률(precision)과 재현률(recall)및 F1(F-measure)값을 계산하였다.

이전 연구[9]에서는 단어 점수를 재조정하고 미등록어 문제를 해결하기 위해 Self-training을 제안하였다. 본 논문에서 미등록어 문제를 해결하기 위해 제시한 방법과 비교하기 위하여 추가적인 실험을 진행하였다.

표 2 시스템 성능 평가 결과

시스템	정확률	재현률	F1
Baseline	0.146	0.272	0.190
Baseline+Score(w)	**0.224**	0.226	0.225
Baseline+Score(w) +미등록어 점수부여	**0.224**	0.228	**0.226**
Basline+Score(w) +self-training	0.186	**0.279**	0.223

표 3 단어의 점수 사용법에 대한 비교

시스템		정확률	재현률	F1
Score사용	미등록어 점수			
자질	0	0.224	0.226	0.225
	부여	0.224	**0.228**	**0.226**
Interpolation	0	**0.280**	0.127	0.175
	부여	0.232	0.209	0.220

4 결과 및 분석

단어의 점수를 사용하지 않은 기준시스템에 비교해서 단어의 점수를 사용한 경우 재현률이 약 17.9%가량 감소하지만 정확률의 경우 53.4%로 크게 향상되는 것으로 나타났다. 이 수치는 이전 연구와 비교해서도 큰 폭으로 증가한 것으로 개선된 점수 계산 방법이 성능을 향상시키는데 효과가 있다는 것을 나타낸다. 재현률의 하락을 막기위해 Self-training을 사용하여 학습 데이터를 증가시킨 경우 정확률의 상승은 27.4%로 낮아지지만 재현률이 하락하지 않고 약 9.9% 상승하는 것으로 나타났다.

또한 미등록어 문제의 해결을 위해 사용한 단어의 확장 방법은 재현률의 향상이 0.9% 정도로 높지 않지만

[1] http://sentiwordnet.isti.cnr.it/
[2] http://nlp.stanford.edu/software/lex-parser.shtml
[3] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

정확률의 하락이 없는 것을 알 수 있다. 이는 상의어/하의어 관계에 기반한 단어의 유사도 측정 방법이 두 단어가 얼마나 공통적인 개념을 가지고 있는지를 잘 반영하기 때문으로 생각된다.

본 논문의 단원 2.3에서 제시한 바와 같이 단어의 점수를 자질로서 사용한 경우와 선행점수로서 Interpolation을 사용한 것을 비교하였다(표 3). 단어의 점수를 선행점수로 사용한 경우 자질로서 사용한 경우와 비교하여 정확률은 상승했으나 재현률이 하락하는 결과가 나왔다. F1을 기준으로 비교하였을 때 자질로서 사용하는 것이 더 높은 성능을 보였다.

5 요약 및 결론

본 논문에서는 단어의 어휘점수를 부여하기 위한 더 정교한 방법을 제안하였고 미등록어 문제를 해결하기 위해 taxonomy를 기반으로한 단어 사이의 유사도를 사용하는 방법을 제안하였다. 또한 단어에 부여된 점수를 자질로 활용하는 방법과 선행점수로 활용하는 방법을 비교하여 보았다. 본 논문에서 단어별 점수부여 방법과 미등록어 해결을 위해 제시한 방법이 이전 논문에서 제시한 self-training을 이용한 방법에 비교해서 F1을 기준으로 1.3% 더 높은 성능을 보였다. 특히 정확률을 기준으로 20.4% 더 높은 성능을 보였다.

감사의 글

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

참고문헌

[1] V. Stoyanov and C. Cardie. Toward Opinion Summarization: Linking the Sources. In Proceedings of the Workshop on Sentiment and Subjectivity in Text. pp. 9-14. 2006.

[2] S. Bethard, H. Yu, A. Thornton, V. Hativassiloglou, and D. Jurafsky. Automatic extraction of opinion propositions and their holders. In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text. pp. 22-24. 2004.

[3] V. Rubin. Trust Incident Account Model: Preliminary Indicators for Trust Rhetoric and Trust or Distrust in Blogs. 3rd International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media, May 17 - 20, 2009, San Jose, California.(ICWSM-2009) pp. 300-303.

[4] Y. Kim and S. Myaeng . Opinion Analysis based on Lexical Clues and their Expansion. In Proceedings of the NTCIR-6 Workshop. pp. 308-315. 2007.

[5] S. Kim and E. Hovy. Identifying and Analyzing Judment Opinios. In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. pp. 200-207. 2006.

[6] Y. Choi, C. Cardie, E. Riloff and S. Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language. pp. 355-362. 2005.

[7] Y. Seki, N. Kando and M. Aono. Multilingual opinion holder identification using author and authority viewpoints. Information Processing & Management, Volume 45, Issue 2, Pages 189-199. 2009.

[8] Y. Kim, Y. Jung and S. Myaeng. Identifying Opinion Holders in Opinion Text from Online Newspapers. In Proceedings of IEEE International Conference on Granular Computing. pp. 699-702. 2007.

[9] 정헌영, 김준기, 이예하, 이종혁. 의견의 주체를 찾기 위한 의견주체 후보점수 부여방법과 Self-training. 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집 제36권 제1호(C), pp. 341-345. 2009.