채팅 텍스트로부터의 화자 감정상태 학습

문현구 장병탁
서울대학교 컴퓨터공학부
{hkmoon, btzhang}@scai.snu.ac.kr

Learning Emotional States of Chatting Partners from Text Data

Hyun-Ku Moon Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

현재 인터넷 환경에서 텍스트는 다루기 쉽고 부하가 적어 가장 많이 사용되는 통신 수단이다. 그러나 화상 채팅과는 달리 자신의 표정이나 제스춰를 전달할 수 있는 방법이 없기 때문에 표현상의 한계가 있다. 이 글은 일상 대화를 텍스트로 입력받아, naive Bayes 알고리즘을 사용해 미리 정의된 감정 범주, 즉 울기, 웃기, 화내기 등으로 분류해 주는 방법에 관해 다루고 있다. 채팅사이트에서 수집된 학습데이터는 사람에 의해 해당 감정 범주로 태깅되고, 이렇게 태깅된 데이터가 학습엔진에 의해 통계 정보로 구축되면, 실제 채팅사이트에서 감정인식 엔진은 입력된 데이터를 분석해 해당 감정으로 분류한다. 연령별로 5개의 그룹으로 나눈 대화방에서 각각 1000문장씩 테스트해 본 결과 평균 91.6%의 정확도를 얻을 수 있었다.

1. 서론

일반적으로 전자메일, 채팅, 메신저 등에서는 텍스트 위주의 의사소통을 하게되므로 일상어 전달시 필요한 화자의 표정 등을 표현하는데는 어려운 면이 있기 마련이다. 이러한 한계를 극복하기 위해서 젊은 층을 중심으로 이모티콘(그림문자)의 사용이 활성화되어왔다. 이모티콘은 UNIX 시스템의 네트워크 사용자들이 사용했던 스마일리(Smiley)에서 그 유래를 찾을 수 있는데, 국내에서는 채팅 사용자를 중심으로 다양한 표정이 개발되었다. 표 1은 현재 채팅에서 많이 쓰이는 그림문자와 인터넷에서 자주 사용되는 이모티콘의 예이다.

이모티콘(인터넷)		그림문자(국내통신)	
:-)	가장 기본적인 웃음	^_^	가장 평범한 웃음
:-D	박장대소하는 모습	^o^	박장대소하는 모습
:-(	찡그린 모습	^^;	엇퍽은 웃음
:-<>	깜짝 놀란 표정	T.T	눈물을 흘리는 표정

표 1. 그림문자의 예

본 논문에서는 입력된 문장을 해당 감정으로 분류하는 시스템인 TexMo의 원리와 실험 결과에 대해 다루고 있다. TexMo는 일상어를 입력받아 미리 정의된 범주 코드를 출력한다. TexMo가 수신자측에 있으면서 이러한 범주 코드에 따라 캐릭터나 각종 이미지를 보여주게 되면, 통신 부하에 전혀 영향을 미치지 않으면서 사용자의 표정이나 제스춰를 전달할 수 있으므로 비주얼하고 재미있는 통신이 가능하다. 이 글에서는 먼저 전체 시스템의 구조에 대해 설명한 후, 분석 단계에서 naive Bayes 알고리즘을 어떻게 적용했는지를 보이며, 마지막으로 테스트 데이터를 가지고 실험한 결과를 제시한다.

2. 감정분석 처리 시스템 구조

그림 1에서와 같이 TexMo는 크게 두 부분으로 나뉜다. 하나는 학습데이터를 가지고 통계 정보를 구축하는 학습 시스템이고, 다른 하나는 위에서 구축된 통계정보를 사용해 실제 입력 데이터를 분석하는 분석 시스템이다.

2.1 학습 시스템

통계 데이터가 구축되기 위해서는 실제 채팅에서 모은 데이터를 사람이 각 범주별로 분류해 주는 작업이 필요하다. 이러한 작업을 '데이터 태깅'이라 하는데, 각 범주의 선험적인 확률인 $P(v_j)$를 구하기 위해 전체 문장의 범주를 태깅하고, 이 문장을 다시 어절 단위로 태깅하여 $P(a_i|v_j)$를 구한다. 여기서 v_j는 감정을 분류한 범주이고 a_i는 입력 문장의 i번째 어절이다.

학습을 위한 데이터는 인터넷의 채팅사이트 4곳을 골라 골고루 수집하였다. 분석 결과 전체 모아진 30만 문장 중 표 2에서 언급된 24개의 감정으로 분류되는 문장은 약 30% 정도였다. 즉 입력 데이터를 모두 무동작 상태로 처리하면 정확도가 70% 정도 나온다는 뜻이다. 이는 TexMo로 하여금 '안전한 전략'을 취하도록 하는데, 다시 말해 범주가 모호해서 어떠한 범주를 취할지 결정할지 모를 때는 무조건 25번 범주인 '무동작'으로 결정하도록 한다.

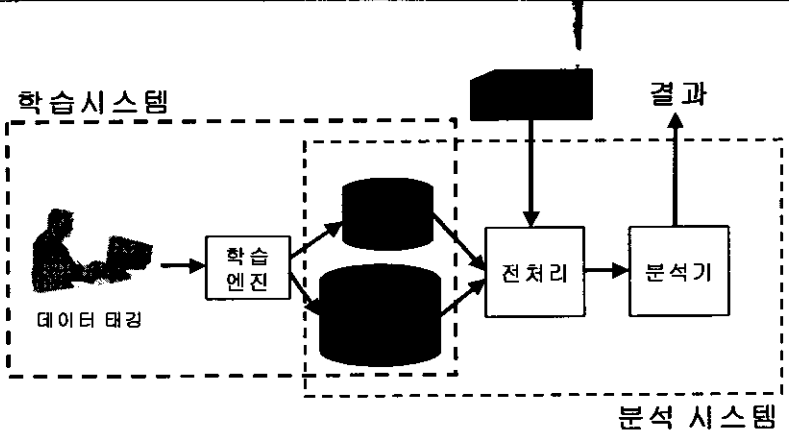

그림 1 감정 인식 엔진 TexMo의 전체 흐름도

2.2 분석 시스템

분석 시스템은 두 개의 모듈로 구성된다. 첫째로 전처리 단계에서는 채팅 데이터에서 분석에 불필요한 문자(@#$^&◀♤※☆ 등)를 제거하는 일과, 감정 상태에 결정적인 역할을 하는 그림문자(^^, -.-; 등)를 찾아내는 일을 한다. 그림문자는 글을 쓰는 사람의 표현 의도를 명시적으로 나타내는 것이므로, 만약 이러한 그림문자가 검출될 경우 대화내용을 분석하지 않아도 된다.

분석기에서는 위에서 언급한 naive Bayes 알고리즘을 기초로 하여 가장 높은 확률의 감정 범주를 선택한다. 이때 문제가 되는 부분은 실제 입력 데이터의 길이가 너무 짧은 즉, 학습데이터의 희소성으로 인해 계산된 확률의 결과값이 0인 경우가 많다는 것이다. 그러나 실험 결과, 이러한 현상이 나타나는 범주는 주로 18번 '질문'이었다. 다행이도 이러한 문제는 문장 내에 물음표 '?'의 존재를 파악함으로써 상당 부분 해결이 가능하였다. 또 한가지 문제는 감정을 표현하는 의성어들, 예를 들어 "하하하"나 "흑흑흑" 등의 길이기 일정하지 않게 쓰인다는 것이다. 즉 "하하하"와 "하하하하하"는 같은 뜻으로 사용됨에도 불구하고 다른 통계 정보가 구축된다. 이의 해결을 위해 분석기에 별로로 이러한 의성어의 연속을 찾아내서 같은 내용의 감정으로 인식하도록 하였다. 이러한 처리들은 naive Bayes 알고리즘을 적용하는데 문제시 되는 데이터의 희소성에 대한 좋은 해결책이 될 수 있다.

그밖에 필요한 처리로는 하나의 문장에 상반되는 범주의 두 가지 그림문자가 검출되었거나, 문장이 너무 길어 여러 가지 감정범주로 분류되는 경우의 범주 선택 원칙이 있다. 첫 번째 문제의 경우 TexMo는 가장 나중에 나온 그림문자를 선택하는 것을 원칙으로 했으며, 두 번째 문제의 경우 문장의 길이가 어느 정도 이상이 되면 25번 범주인 '무상태'를 선택하도록 하였다. 문장의 길이가 길어지는 것은 감정 표현이라기 보다는 그냥 일반적인 '서술'이라고 보기 때문이다. 이러한 특성으로 인해, 채팅이 아닌 메일 등의 데이터에 TexMo를 적용하기 위해선 지금과는 다른 처리가 필요하다.

3. 감정상태 학습

TexMo가 입력으로 받는 데이터는 주로 채팅이나 메신저에서 받아오는 것들이다. 이러한 데이터들의 특징은 비어나 속어 등이 많다는 것과 맞춤법과 띄어쓰기를 무시하는 경우가 종종 있다는 것이다. 이러한 데이터를 형태소 분석을 하는 등의 자연언어 처리 방식으로 접근하게 되면 비어나 속어에 대한 특별한 사전을 구성하지 않는 한 정확도가 떨어질 수밖에 없다. 따라서 우리는 입력된 문장을 자연언어로서의 처리가 아닌, 어절 단위 각각을 통계 정보로 분석하는 것을 원칙으로 했다.

TexMo에서 텍스트를 감정에 따라 분류하기 위해 사용된 알고리즘은 기본적으로 naive Bayes 방법이며 다음과 같은 식으로 표현된다.

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

여기서 v_{NB}는 결정된 범주를 나타내고, V는 분류해야 할 범주의 집합, a_i는 입력 데이터에 있는 각각의 단어를 나타낸다. 위와 같은 식이 가능한 이유는 Bayes 정리에서 각각의 a_i값이 독립이라고 가정했기 때문이다.

한국어는 영어와 달리 문장 내의 어절 하나만 가지고도 그 문장의 내용 분석이 가능하다. 예를 들어 영어에서 "I don't know"를 어절 단위로 분석하면 "don't"와 "know"가 분리되어 분석 대상이 둘로 나뉘지만 한국어는 , "나는 잘 몰라"에서 보듯이 하나의 단어 "몰라"가 전체 문장을 대표하기 때문에 어절 단위의 손쉬운 분석이 가능하다. 이는 영어보다는 한국어가 naive Bayes의 "a_i값이 독립이어야 한다"는 조건을 더 만족시킴을 의미한다.

결과값을 얻기 위해서는 각 범주가 나올 확률인 $P(v_j)$

와 각 범주에서 해당 단어가 나올 확률인 $P(a_i|v_j)$가 미리 주어져 있어야 한다. 따라서 미리 학습 데이터를 사용해 문장별, 단어별로 태깅 작업을 하는 단계가 필요하다. 태깅은 아래에 있는 표 2에 따라 학습데이터의 문장을 어절단위로 분리해 사람이 직접 각각의 어절에 해당 번호를 부여하는 작업이다.

분류 범주	입 력 예
1. 첫인사	하이루, 안녕하세요, 하2, 안녕
2. 끝인사	빠이, 안녕, 담에봐
3. 미소	^_^ ^^ ^.^ ^.^ 헤헤, 히
4. 큰웃음	^o^ :-D 하하, 푸하하, 캬캬
5. 킥킥웃음	킥킥, 큭큭, 힉힉힉, 쿄쿄
⋮	⋮
18. 질문	뭔데?, 먹었어?, 어디?, 뭐하세여
19. 이해/동의/긍정	글쿤, 아항, 알았어, 그래, 예
20. 거절/부정	시로, 안돼, 싫다, 아니, 아뇨
21. 황당함	황당하군, 허걱
22. 썰렁함	추워, 썰렁하다~
23. 어이없음	--; -.-; -_-; -.-a --a
24. 꾸벅	꾸벅
25. 무상태	아무런 행동을 안하는 상태

표 2 현재까지의 분류 가능 범주 (일부)

학습엔진은 문장별로 태깅된 결과를 가지고 선험적 확률인 $P(v_j)$를 만들고, 단어별로 태깅된 결과는 해당 범주에서 각 단어 나올 확률인 $P(a_i|v_j)$를 만듦으로서 학습데이터가 구축된다.

4. 실험 및 결과

실험은 2종류로 진행되었다. 하나는 그림문자와 의성어, 물음표 인식 모듈을 제외한 순수한 통계적 방법에 의한 실험이고, 또 하나는 정확도를 높이기 위하여 이러한 부가적인 처리를 모두 해준 경우이다.

아래의 표 3은 채팅 사이트에서 연령대에 따라 대화방을 선택한 후 1000문장씩 데이터를 모아 감정인식 결과를 요약한 것이다. 표에서 알 수 있듯이 혼합적 방법에 사용되는 부가 처리가 정확도 향상에 많은 영향을 끼치고 있다.

	직장인	대학생	고등학생	중학생	초등학생	평균
통계적 방법	84	87	85	84	83	84.6
혼합 방법	91	92	96	92	87	91.6

표 3 실험 결과 (accuracy)

위의 실험 결과를 분석해 보면 초등학생 데이터를 가지고 한 실험에서의 결과가 좋지 않게 나타남을 볼 수 있는데, 이는 이 데이터에 참여한 채팅 집단의 타이핑 습관 때문이다. 즉 대화에 주도적으로 참여한 사람은 그림문자를 전혀 쓰지 않았으며, 질문하는 문장의 마지막에 물음표를 사용하는 경우도 적었다. 예를 들어 "밥 먹었어?" 라고 써야하는데도 "밥 먹었어"로 질문하는 등, 사람도 그 문장만 보고는 판단할 수 없는 내용이 많았다. 대학생 같은 경우 반대로 그림문자의 사용이 많은 채팅 데이터였다. 주로 고등학생이 사용하는 채팅 언어는 그림문자나 은어, 속어들이 많다. TexMo의 경우 정규 언어보다는 이러한 은어, 속어들이 많은 데이터에 강한 특징을 갖는다. 이는 그러한 단어의 경우 의미가 한가지로 굳어버리기 때문인 것으로 해석된다.

실험에서 나타나는 에러들을 분석해 본 결과, 에러의 50% 정도가 말뭉치의 부족 때문으로 나타났으며, 30%는 태깅을 잘못해서 노이즈가 생긴 경우, 그리고 나머지는 단어가 여러 가지 뜻을 내포해서 분석이 불가능 한 경우이다. 실제로 채팅사이트에 적용했을 땐 캐릭터 에니메이션의 중의적인 표현으로 채감 정확도가 더 늘어나는 것을 감안한다면 만족할 만한 성과라 할 수 있다.

5. 결론

지금까지 감정인식 엔진이 채팅사이트에서 감정 상태를 학습하는 방법과, 실제 실험 결과에 대해 알아보았다. naive Bayes 방법에 덧붙여 전처리 과정에서 몇 가지 처리를 추가한 결과, 입력 데이터의 희소성으로 인해 결과가 좋지 않을 것이라는 우려와 달리, 높은 정확도를 얻을 수 있었다. 이는 영어보다는 한국어가 하나의 단어로 분석하기 편하다는 특징 때문이기도 하다.

현재 TexMo에서 해결되어야할 과제는 하나의 시스템을 만들기 위해 각 범주에 따른 태깅을 하는데 너무 많은 인력과 시간이 소요된다는 것이다. 이러한 문제는 EM 알고리즘의 사용으로 해결이 가능할 것으로 예상한다. EM 알고리즘은 사람이 수동으로 하는 태깅 작업을 자동화해 줄 수 있는 한가지 대안으로써, 노이즈에 대한 해결책이 마련된다면 도전해 볼 가치가 있는 방법이다.

또 하나의 문제는 문형 특징상 18번 범주인 '질문'에 1회만 사용되는 단어들이 너무 많다는 것이다. 실제 전체 통계 데이터의 50% 이상이 '질문'범주에 있는 내용들인 반면, 여기 있는 데이터가 분석에 사용될 비율은 극히 적고, 오히려 물음표를 인식하여 '질문'범주를 인식하는 경우가 더 많다. 이는 "***했어?"에서 알 수 있듯이 질문에서 어간의 변화가 무한하다는 특징 때문이다. 이의 해결을 위해 통계 정보에 영향을 끼치지 않으면서 불필요한 데이터를 삭제할 방법을 찾아내야 할 것이다.

감사의 글: 본 연구는 교육부 BK 21의 지원과 첨단정보기술 연구센터 (AITrc)를 통한 과학재단의 지원을 받았음.

<참고문헌>

[1] Tom M. Mitchell. Baysian Learning, In *Machine Learning*, The McGraw-Hill Companies, Inc., pp. 154-200.

[2] McCallum, A., and Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Categorization*. 1998.

[3] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Learning to Classify Text from Labeled and Unlabeled Documents. In *AAAI-98*, 1998.