지지벡터기계(Support Vector Machines)를 이용한 한국어 화행분석

은 종 민[†] · 이 성 욱[††] · 서 정 연[†††]

요 약

본 연구에서는 지지 벡터 기계(Support Vector Machines)를 이용하여 한국어 대화의 화행을 분석하는 방법을 제안한다. 우리는 발화의 어휘 및 품사와 이진 품사 쌍을 문장 자질로 사용하고 이전 발화의 문맥을 문맥 발화로 사용한다. 카이 제곱 통계량을 이용해 적절한 자질을 선택하고 선택된 자질로 지지 벡터 기계를 학습하였다. 학습된 지지 벡터 기계 분류기를 이용하여 각 발화의 화행을 분석하였다. 호텔 예약 영역의 말뭉치에 대해 제안된 시스템을 이용하여 실험한 결과 약 90.54%의 정확률을 얻었다.

키워드 : 화행 분석, 지지벡터기계, 자질 선택기계 학습

An analysis of Speech Acts for Korean Using Support Vector Machines

Jongmin En[†] · Songwook Lee[††] · Jungyun Seo[†††]

ABSTRACT

We propose a speech act analysis method for Korean dialogue using Support Vector Machines (SVM). We use a lexical form of a word, its part of speech (POS) tags, and bigrams of POS tags as sentence features and the contexts of the previous utterance as context features. We select informative features by Chi square statistics. After training SVM with the selected features, SVM classifiers determine the speech act of each utterance. In experiment, we acquired overall 90.54% of accuracy with dialogue corpus for hotel reservation domain.

Key Words : Speech Act Analysis, Support Vector Machines, Feature Selection, Machine Learning

1. 서 론

자연어 대화에서 화자가 발화를 통해 상대방에게 나타내고자 하는 의도적인 행위를 화행이라 한다. 화행은 자연어 처리를 처리하는 많은 대화 시스템에서 화자의 의도를 파악하고 적절한 응답 발화를 생성하는데 중요한 역할을 한다[1, 2]. 이러한 화행을 대화 시스템이 분석하기 위해서는 주어진 발화의 의미 정보와 대화의 흐름을 고려해야 한다.

한국어 화행 분석과 관련해서 대화 말뭉치를 이용한 연구가 최근 많이 수행되어 왔다. [3]에서는 발화의 특성 정보와 담화의 구조를 고려한 통계적 화행 분석 모델을 제안하였고, 담화 구조를 반영한 통계 모델의 성능이 담화 구조를 반영하지 않은 통계 모델보다 성능이 나음을 보였다. [4]은 담화 구조 분석에 대한 통계적 모델을 제시하고, 최대 엔트로피 모델을 이용하여 화행 및 담화구조 분석을 통합하여 수행하는 통합 시스템을 제안하였다. [5]은 이전 문장의 화행으로부터 전이 확률을, 구문유형과 화자정보를 이용하여 관측 확률을 결정트리를 이용하여 추정하였고 이를 은닉 마르코프 모델(Hidden Makov Model)에 사용하여 화행을 분석하였다. [6]은 어휘-품사 자질과 이진 품사 쌍 자질을 구성하고 이전 발화의 화행을 사용하여 신경망을 학습하여 화행분석을 수행하였고 카이제곱 통계량을 이용해서 자질을 선택하는 방법을 제안하였다.

본 연구에서는 지지벡터 기계를 화행 분석에 이용하는 것을 제안한다. 지지벡터기계를 학습하기 위해 문장자질과 문맥자질을 사용하였다. 문장자질은 해당 발화로부터 의미 정보를 이용하기 위해서 추출한 자질이고, 문맥자질은 대화의 진행상에서 얻어지는 문맥정보를 고려하기 위해서 이전 발화로부터 추출한 자질이다. 이러한 자질들 중 유용한 자질을 [6, 9]에서 제안한 방법인 카이 제곱 통계량을 사용하여 선택하였고, 각 발화의 화행은 학습된 지지벡터 기계를 이용하여 결정하였다.

본 논문의 구성은 다음과 같다. 2장에서 지지벡터 기계의 학습에 사용된 자질의 종류를 설명하고 카이제곱 통계량을 이용한 자질 선택 방법에 대해 설명한다. 3장에서는 지지벡터기

※ 본연구는 과학기술부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다
† 정 회 원 : (주)하나로 드림 전임연구원
†† 정 회 원 : 동서대학교 컴퓨터공학과 전임강사
††† 종신회원 : 서강대학교 컴퓨터학과 교수
논문접수 : 2004년 12월 2일, 심사완료 : 2005년 4월 25일

계를 이용하기 위한 방법을 설명한다. 4장에서는 실험을 통해 제안된 방법의 성능을 보이고 5장에서 결론을 내린다.

2. 지지벡터기계의 학습에 사용된 자질

2.1 대화 말뭉치와 화행의 종류

본 연구에서는 전화 예약 영역(호텔 예약, 항공 예약, 여행 예약 등)에서 수집한 528 대화, 10,285 문장으로 구성된 말뭉치를 사용한다[3-6]. 대화 말뭉치에는 발화자 정보(SP), 발화 문장(KS), 구문 유형(ST), 화행(SA), 담화 구조(DS) 등의 담화 정보가 부착되어 있다. 구문 유형은 화행 분석에서 발화 대신에 사용하는 것으로 발화 문장의 구문 정보로 구성되어 있다. 구문 유형은 문장 유형, 본용언, 시제, 문장의 부정형 여부, 양상, 단서 단어로 구성되며, 본용언과 단서 단어에는 어휘 정보를 포함하고 있다. 발화자 정보는 발화자가 고객인지 직원인지를 나타낸다. (그림 1)은 담화 정보가 부착된 말뭉치의 일부분이다.

/SP/User /KS/미국 조지아대 어학연수에 참가 신청을 한 학생인데요. /ST/[decl,be,present,no,none,none] /SA/introducing-oneself /DS/[2]	/SP/Agnt /KS/조지아대학에 어학연수 코스는 대학에 기숙사를 제공하고 있습니다. /ST/[decl,pvg,present,no,none,none] /SA/response /DS/[2]
/SP/ User /KS/숙소에 관해서 문의할 사항이 있어서요. /ST/[decl,paa,present,no,none,none] /SA/ask-ref /DS/[2] ->계속	/SP/User /KS/그럼 식비는 연수비에 포함이 되어 있는 건가요? /ST/[yn_quest,pvg,present,no,none,그러면] /SA/ask-if /DS/[2.1]

(그림 1) 대화 말뭉치의 예

화행에는 대화 말뭉치에서 발견되는 17개의 화행을 사용하였다. 본 연구에 사용한 화행의 예는 <표 1>과 같다[3].

〈표 1〉 화행의 예

화행	예
Accept(호응)	예, 물론입니다.
Ackowledge(인정)	아 그렇군요.
Ask_confirm(확인 요구)	아 예약이요?
Ask_if(Y/N question)	예약하시겠습니까?
Ask_ref(WH question)	며칠간 머무르실 생각이십니까?
Closing(대화의 종료)	안녕히 계십시오.
Correct(대화의 수정)	잘못 말씀하신 거 같은데요.

2.2 문장자질

문장자질은 대화 말뭉치에 부착되어 있는 구문 유형 정보와 형태소 분석 결과를 이용하여 생성하며,어휘/품사, 이진(bigram) 품사 쌍, 구문유형 등으로 구성되어 있다.

어휘/품사 자질은 형태소 분석 결과에서 얻을 수 있는 어휘/품사의 쌍이며 이진 품사 쌍 자질은 품사의 이진 쌍인 품사-품사 쌍을 의미한다.

(그림 2)는 사용자의 발화를 각각의 문장자질로 나타낸 예이다.

USER : 숙소에 관해서 문의할 사항이 있어서요.
어휘/품사 : 숙소/nc 에/jca 관하/pv 어서/ecs 문의/pv 르/exm 사항/nc 이/jc 있/pa 어서요/ef ./s.
이진품사 : [nc,jca] [jca, pv] [pv, ecs] [ecs, pv] [pv, exm] [exm, nc] [nc, jc] [jc, pa] [pa, ef] [ef, s.]
구문유형 : decl, paa, present, no, none, none

(그림 2) 사용자 발화에 대한 문장자질의 예

화자가 전달하고자 하는 의도와 그러한 의도가 표면적인 발화로 나타날 때의 구문 유형은 상호 밀접한 관계가 있다. 즉 화자가 어떤 의도를 상대방에게 전달하고자 할 때 화자는 원하는 의도를 가장 잘 표현할 수 있는 구문 형태로 발화하게 되고 청자는 주어진 형태의 발화와 대화 상황으로부터 화자의 의도를 추론하게 된다. 그러므로 구문 유형은 화행을 분석하기 위한 가장 기본적인 정보로 사용될 수 있다. 표2는 구문 유형 정보의 예를 나타낸 것이다.

〈표 2〉 구문 유형 정보의 예

구문 유형 정보	예	종류
문장유형	yn_quest, decl, wh_quest, imperative	4
본용언	be(이다), pvg(일반동사), paa(상태 성상형용사), pad(지시형용사), pvd(지시동사), frag(동사없음), 알다, 감사하다, 좋다, ...	88
시제	present, future, past	3
문장의 부정형 여부	no, yes	2
양상	want, will, possible, serve, seem, intend, ...	29
단서 단어	예, 그리고, 그러면, 안녕, 대신, ...	26

<표 2>와 같이 화행과 관련된 구문 유형 정보로 문장 유형(sentence type), 본용언(main verb), 시제(tense), 문장의 부정형 여부(positive or negative), 양상(modal), 단서 단어(clue word)의 6가지 정보를 사용한다. 본용언은 동사와 형용사로 나뉘며, 화행을 지시적으로 나타내는 수행동사(performative verb)의 경우에는 어휘 그 자체를 사용한다. 양상은 한국어에서 보조용언에 나타나는 것으로 희망, 당위, 가능, 추측 등이 있으며, 이는 화자의 의도와 밀접한 관련을 갖는다.

동일한 발화라고 하더라도 문맥에 따라서 다른 화행을 갖는 것이 가능하다. 따라서 현재 발화의 화행을 분석하기 위해서 이전에 나타난 발화를 고려해야 한다. 이전에 나타난 발화를 자질로 사용하기 위해서 이전에 나타난 발화의 문장자질과 화행 그리고 화자정보를 현재 분석하려는 발화의 문맥자질로 사용한다.

2.3 자질 선택 및 가중치 부여

앞에서 살펴본 모든 자질을 사용하게 되면 매우 많은 수의 자질이 나타나게 된다. 이러한 자질들 중에서는 화행을 결정하는 데 기여를 하는 자질이 있기도 하지만 그렇지 않은 경우나 오히려 방해가 되는 자질들도 존재를 하게 된다. 따라서 우리는 카이 제곱 통계량을 이용해서 자질을 선택한다. 카이 제곱 통계량을 계산하는 식은 다음과 같다[9].

$$\chi^2(f,s)=\frac{(A+B+C+D)\times(AD-BC)^2}{(A+B)\times(A+C)\times(B+D)\times(C+D)} \quad (1)$$

A는 화행 s에 속해 있는 문서 중에 자질 f를 포함하고 있는 발화의 수이고, B는 화행 s외의 범주에 속해 있는 문서 중에 자질 f를 포함하고 있는 발화의 수이다. 또한, C는 화행 s에 속해 있는 발화 중에 자질 f를 포함하지 않는 발화의 수이며, D는 범주 s외의 화행에 속해 있는 발화 중에 자질 f를 가지고 있지 않는 발화의 수이다. 자질 f와 화행 s가 완전히 독립적이면 0의 값을 갖는다. 하나의 자질에 대해 카이 제곱 통계량의 값을 결정하는 방법은 전체 화행에 대한 평균값을 사용하는 방법과 전체 화행에 대해 최대값을 사용하는 방법이 있을 수 있다. 우리는 여러 실험 결과를 바탕으로 최대값을 이용하는 방법을 사용한다.

각각의 자질에 가중치를 부여하는 방법은 여러 방법이 있으나 본 연구에서는 이진가중치, 용어 및 역문헌 빈도 (Term Frequency-Inverse Document Frequency) 가중치, 용어 및 역범주 빈도 (Term Frequency- Inverse Category Frequency) 가중치를 각각 사용하며 그 중 가장 좋은 성능을 보이는 가중치 방법을 사용한다. 화행 분석에 적용하기 위해 TF-IDF값이나 TF-ICF값을 계산하는 경우, 일반적으로 정보검색분야에서 사용하는 의미와는 달리 용어(term)는 자질로, 문서(document)는 발화로 범주(category)는 각 화행으로 간주하여 계산한다.

3. 지지 벡터 기계의 학습

지지 벡터 기계(Support Vector Machine)는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik[5]에 의해 소개된 학습기법으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 결정면(decision surface)을 찾는 모델이다. 선형 분리가 가능한 공간에서의 결정면은 초월면 $H : y = \boldsymbol{w} \cdot \boldsymbol{x} - b = 0$ 이며 이 초월면에 평행하고 동일 거리에 있는 두 개의 초월면은 아래 식의 H_1, H_2와 같으며, H_1와 H_2 사이에 어떠한 데이터 포인트도 존재하지 않는 조건을 만족시키며 H_1와 H_2 사이의 거리는 최대가 된다.

$$H_1 : y = \boldsymbol{w} \cdot \boldsymbol{x} - b = +1,$$
$$H_2 : y = \boldsymbol{w} \cdot \boldsymbol{x} - b = -1.$$

H_1와 H_2 사이의 거리를 최대로 만드는 것이 지지벡터 기계의 학습 목적이 된다. 따라서 H_1에는 양의 값을 갖는 데이터가 존재하게 되고 H_2에는 음의 값을 갖는 데이터가 존재하게 되는데, 이러한 데이터들을 지지벡터(support vectors)라 부르며 이들이 분리 경계면을 결정하는 역할을 한다. 다른 데이터들은 H_1와 H_2를 교차하지 않도록 분리 경계면 주위로 이동되거나 제거된다. H_1와 H_2 사이의 거리를 최대로 하기 위해서 H_1와 H_2 사이에 어떠한 데이터 포인트도 존재하지 않도록 하면서 ||w|| 을 최소화시키면 된다.

$$\boldsymbol{w} \cdot \boldsymbol{x} - b \geq +1 \text{ for } y_i = +1,$$
$$\boldsymbol{w} \cdot \boldsymbol{x} - b \leq -1 \text{ for } y_i = -1.$$

지지 벡터 기계의 문제는 이러한 w와 b를 찾아내는 문제이며, 이것은 이차 프로그래밍(quadratic programming) 기술에 의해 풀 수 있다[7].

문서 분류 분야에서 좋은 성능을 보여 주고 있는 분류기인 지지 벡터 기계를 우리는 화행을 분석하는데 사용하였다. 지지벡터 기계는 이진 분류기이므로 우리는 각각의 화행을 위한 지지벡터기계를 각각 따로 학습한다. 지지벡터 기계의 학습을 위한 자질은 2장에서 설명한 각각의 자질들의 가중치로 벡터를 구성하였다. 지지벡터 기계를 이용한 화행 결정 단계에서는 주어진 발화에 대해 각각의 화행을 위한 지지 벡터 기계의 출력값들을 서로 비교하여 가장 큰 값을 출력하는 지지 벡터 기계에 해당하는 화행을 주어진 발화의 화행으로 최종적으로 결정한다. 본 연구에서는 SVM^{light}[8]를 이용하였고 선형 커널을 이용하여 학습하였다.

4. 실 험

본 연구에서 사용한 말뭉치는 예약영역(호텔예약, 항공예약 등)에서 수집한 528개의 대화, 10285개의 발화로 구성된 대화 말뭉치를 428개의 대화, 8349개의 발화를 학습 말뭉치로, 100개의 대화, 1936개의 발화를 평가 말뭉치로 사용한다. 화행 분석 시스템은 주어진 문장에 가장 적절한 화행을 분석하는 시스템이다. 따라서 제안하는 시스템은 하나의 문장에 하나의 화행을 제시하여 얼마나 정확히 화행을 분석해 냈는지를 나타내는 척도로 정확률(accuracy)을 사용한다.

<표 3>은 각 자질의 종류에 따른 시스템의 성능을 나타낸다(이중 가중치를 사용하였고 자질의 수는 3000개로 제한했다).

〈표 3〉 자질에 따른 성능 비교

사용된 자질		정확률(%)
①	문장자질만 사용	79.96
②	① + 구문정보	83.83
③	② + 이전 문장의 자질	85.12
④	② + 결정된 화행	88.12
⑤	④ + 화자정보	90.54

①은 3.2에 기술된 어휘-품사와 이진 품사쌍 자질로 구성된 문장자질만을 사용한 결과이고, ②는 문맥정보로 이전 발화의 문장자질까지 포함하여 화행을 결정한 경우, ③은 이전 문장에서 결정된 화행을 문맥정보로 사용한 경우이다. 이전 문장의 자질을 문맥정보로 반영하는데 있어서 결정된 화행을 자질로 사용하는 것이 좀 더 나은 결과를 보여주고 있으며, 여기에 ⑤와 같이 화자정보를 더해 수행한 결과가 가장 좋은 성능을 보여준다.

<표 4>는 자질의 개수에 따른 시스템 성능을 나타낸 것이다.

카이 제곱 통계량을 이용하여 자질을 제한하는 것이 더 좋은 결과를 보였으며 자질이 너무 적거나 너무 많은 것보다 적절히 사용되었을 때 좋은 성능을 보이는 것을 알 수 있다.

〈표 4〉 자질의 개수에 따른 성능비교

사용된 자질 수	정확률(%)
1000	87.71
2000	89.77
2500	90.27
3000	90.54
3500	87.51
4000	87.51

<표 5>는 카이 제곱 통계량을 이용하여 선택한 3000개의 자질의 가중치에 따른 성능을 나타낸다.

〈표 5〉 자질의 가중치에 따른 성능비교

가중치 방법	정확률(%)
이진 가중치	90.54
TF-IDF가중치	88.52
TF-ICF가중치	85.53

TF-IDF나 TF-ICF를 사용하였을 때 성능 향상이 없었는데 그 이유는 다음과 같다. 한 문장의 화행 분석을 위한 자질 추출 과정에서 대부분의 자질이 중복되지 않고 한번만 출현한 것이 대부분이었다. 그래서 용어 빈도수(TF)가 1이 되어 용어 빈도수(TF)로 인한 변별력이 거의 없었다. 그래서 단순한 이진가중치를 사용한 경우가 가장 좋은 결과를 보인 것이라 할 수 있다.

<표 6>은 다른 연구의 결과와 본 시스템의 성능을 비교하고 있다.

〈표 6〉 다른 시스템과 비교

화행 분석 시스템	정확률
최대엔트로피 모델[4]	80.5%
결정트리 모델[5]	81.5%
신경망 모델[6]	88.32%
제안 시스템	90.54%

동일한 실험 데이터를 이용하여 다른 연구의 결과와 비교하였는데 다른 기계학습 방법을 사용한 방법보다 지지벡터 기계를 사용한 본 연구가 가장 좋은 결과를 보였다.

5. 결 론

본 연구에서는 지지벡터기계를 이용하여 한국어 대화의 화행을 분석하였다. 지지벡터기계는 문장 자질과 문맥자질을 이용하여 학습하였고, 각 자질들은 카이 제곱 통계량을 이용하여 선택하였다. 학습된 지지벡터기계를 이용하여 한국어 화행 분석을 수행하였으며 약 90.54%의 정확률을 얻었다. 본 연구에서 제안한 시스템은 여행과 예약이라는 제한된 영역에서만 실험을 수행한 것으로 이것이 일반적임을 보이기 위해서는 다른 영역에 대한 적용이 필요하다. 본 연구에 사용된 화행은 호텔 예약 영역에 특화된 화행도 존재하므로 일반적인 대화시스템을 위한 화행 분석 시스템을 구성하기 위해서는 화행의 종류에 대한 재정의가 필요하며, 이를 위해서는 다양한 영역에 대한 대화 말뭉치의 수집 및 분석이 필요하다.

참 고 문 헌

[1] Lambert, L. and S. Caberry. A Tripatite Plan-Based Model of Dialogue. In Proceedings of ACL, 1991. pp.47-54.

[2] Chu-Carroll, J. and S. Carberry. Response Generation in Collaborative Negotiation. ACL-95, 1995.

[3] 이재원, 통계적 화행처리를 이용한 대화체 기계번역에서의 효율적인 대화분석, 박사학위논문, 한국과학기술원, 1999.

[4] Choi, Won Seug, Jeong-Mi Cho, and Jungyun Seo. Analysis System of Speech Acts and Discourse Structures Using Maximum Entropy Model. In Proceedings of the 37th Annual Meeting of the Association for computational Linguistics, 1999, pp.230-237.

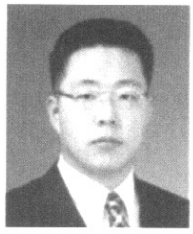
[5] Songwook Lee, Jungyun Seo, "An analysis of Korean speech act using Hidden Markov Model with decision trees", In Proceedings of the 19 ICCPOL, pp.397-400. 2001.

[6] Kyungsun Kim, Jungyun Seo, "Feature selection in automatic speech act tagging", In Proceedings of the 19 ICCPOL, pp.379-383. 2001.

[7] V. Vapnik. The nature of statistical learning theory, Springer, New York, 1995.

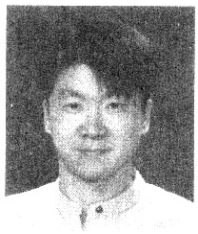
[8] Joachims, T. http://svmlight.joachims.org.

[9] Yang, Yiming and Jan O. Pedersen. A comparative study on Feature selection in text categorization. In proceedings of the 14th International conference on Machine Learning, 1997.



은 종 민

e-mail : uook@hanafos.com
2001년 서강대학교 컴퓨터학과 학사
2003년 서강대학교 컴퓨터학과 석사
2003년~현재 (주)하나로드림 전임연구원
관심분야 : 문서 자동 분류, 및 검색, 텍스트 필터링



이 성 욱

e-mail : leesw@dongseo.ac.kr
1996년 서강대학교 컴퓨터학과 학사
1998년 서강대학교 컴퓨터학과 석사
2003년 서강대학교 컴퓨터학과 박사
2003~2004년 서강대학교 산업기술연구소 연구원
2003년~2005년 서강대학교 정보통신대학원 대우교수
2004년~2005년 LG전자 기술원 선임연구원
2005년~현재 동서대학교 컴퓨터공학과 전임강사
관심분야 : 형태소 및 구문 분석, 단어의미분별, 대화 언어처리

서 정 연

e-mail : seojy@sogang.ac.kr
1981년 서강대학교 수학과 학사
1985년 University of Texas at Austin, Computer Science, M.S.
1990년 University of Texas at Austin, Computer Science, Ph.D
1990년~1991년 UniSQL Inc. 선임연구원
1991년~1995년 KAIST 전산학과 조교수
1995년~현재 서강대학교 컴퓨터학과 교수
관심분야 : 자연어처리, 대화형 에이전트, 한국어 정보처리, 텍스트마이닝