

# 자질 선택 기법을 이용한 한국어 화행 결정 (Decision of the Korean Speech Act using Feature Selection Method)

김 경 선 <sup>†</sup> 서 정 언 <sup>\*\*</sup>

(Kyung-Sun Kim) (Jung-Yun Seo)

**요 약** 화행(speech act)이란 화자의 발화를 통해 나타나는 화자의 의도를 가르키며 자연어로 된 발화를 이해하고 이에 대한 응답을 생성하기 위해 중요한 요소이다. 본 논문에서는 한국어 화행 결정의 성능을 높이기 위해 두 단계 방법을 제안한다. 첫 번째 단계는 형태소 분석결과만을 이용하여 추출된 문장자질과 이전 화행을 이용하여 추출된 문맥자질 중 정보량이 높은 자질을 선택하는 단계이다. 이 단계에서는 형태소 분석 시스템을 사용하여 전체 자질을 구성하고 문서분류 분야의 자질 선택에서 높은 성능을 보인 카이제곱 통계량을 이용하여 효과적인 자질 선택한다. 두 번째 단계는 선택된 자질과 신경망을 이용하여 화행을 분석하는 단계이다. 본 논문에서 제시한 방법은 형태소 분석 결과만을 이용하여 자동적으로 화행을 결정할 수 있는 가능성을 제시하였으며 효과적인 자질 선택을 통해 자질의 수를 감소시키고 정보량이 높은 자질을 사용하여 속도와 성능을 향상 시켰다. 본 논문은 제안된 시스템을 실제 영역에서 수집되어 전사된 10,285개의 발화와 17개의 화행으로 이루어진 대화 코퍼스에 대해 실험하였다. 본 논문은 이 코퍼스에서 8,349개 발화를 학습 코퍼스로 사용하여, 실험 코퍼스의 1,936개 발화에 대해 1,709개에 대해 정확한 화행을 제시하여, 88.3%의 정확도를 보였다. 이는 자질 선택을 하지 않았을 때 보다 약 8%가 증가된 결과이다.  
키워드 : 화행 결정, 자질 선택, 신경망, 카이제곱통계량

**Abstract** Speech act is the speaker's intentions indicated through utterances. It is important for understanding natural language dialogues and generating responses. This paper proposes the method of two stage that increases the performance of the Korean speech act decision. The first stage is to select features from the part of speech results in sentence and from the context that uses previous speech acts. We use  $\chi^2$  statistics(CHI) for selecting features that have showed high performance in text categorization. The second stage is to determine speech act with selected features and Neural Network. The proposed method shows the possibility of automatic speech act decision using only POS results, makes good performance by using the higher informative features and speed up by decreasing the number of features. We tested the system using our proposed method in Korean dialogue corpus transcribed from recording in real fields, and this corpus consists of 10,285 utterances and 17 speech acts. We trained it with 8,349 utterances and have test it with 1,936 utterances, obtained the correct speech act for 1,709 utterances(88.3 %). This result is about 8% higher accuracy than without selecting features.

**Key words** : Speech Act Decision, Feature Selection, Neural Network, CHI Statistics

## 1. 서 론

화행이란 화자의 발화 속에 포함되어 있는 화자에 의

해 의도된 언어적 행위로, 자연어 대화를 처리하는 많은 시스템에서 화자의 의도를 파악하고 응답을 생성하는 중요한 역할을 한다[1, 2, 3]. 그러나 화행결정은 화행이 [그림 1]의 (1)에서 보듯이 같은 의도를 여러 가지 형태의 발화로 표현할 수 있고, 또 같은 발화라도 [그림 1]의 (2)에서 보듯이 발화가 되고 있는 상황이나 화자와 청자의 지식 상태, 화자의 어조 등에 따라 다르게 인식될 수 있기 때문에 어려운 작업이다. 그리고 시스템이

<sup>†</sup> 비 회 원 : 다이렉스트 연구소 선임연구원  
kksun@diquest.com

<sup>\*\*</sup> 종 신 회 원 : 서강대학교 컴퓨터학과 교수  
seojy@ccs.sogang.ac.kr

논문접수 : 2001년 11월 5일  
심사완료 : 2002년 11월 22일

현재까지는 화자의 어조나 화자의 지식상태 등을 화행 분석 시스템에게 정확하게 제공할 수 없다는 사실은 화행결정 문제를 더욱 어렵게 하고 있다.

- (1) 화자가 시간을 알기 원할 때
    - a. 몇 시죠?
    - b. 시계 가지고 있으세요
    - c. 어디에 시계있는 곳이 있을까요?
  - (2) 화자발화 : "시계 가지고 있으세요"
    - a. 시계 소리 여부를 확인하려 할 때
    - b. 시간을 알기 원할 때

그림 1 여러 가지 발화와 화자의 의도

이러한 화행 결정 문제를 해결하기 위해 많은 연구들이 진행되고 있는데 최근에는 대화 코퍼스와 기계학습에서 사용되는 최대 엔트로피 모델(Maximum Entropy Model), 결정트리 모델(Decision Tree Model) 등의 방법을 화행 결정에 적용하려는 연구가 진행되어 왔다[4,5]. 이 연구들은 단어, 명사구 등의 정보와 문장 형태, 주동사, 시제, 부정형 여부, 보조동사로 이루어지는 구문 유형정보, 발화자 정보, 이전화행 등의 문맥 정보 등을 자질로 이용하고, 기계학습을 통해 화행을 결정하였다. 그러나 이와 같은 기존의 연구 방법은 기계학습의 방법적인 면에 초점을 두고 진행되었는데 화행을 결정하는 데는 학습 방법뿐만 아니라, 학습에서 사용되는 자질을 선택하는 문제도 중요하다. 또 사용되어진 자질이 형태소 분석 이상의 분석을 요구하기 때문에 자동으로 위의 화행분석에 필요한 정확한 자질을 생성하기 어려운 실제 응용프로그램에는 적용하기 어렵다.

문서 분류 분야에서는 대용량의 문서를 분류하기 위한 효과적인 자질 선택에 대한 많은 연구가 진행되어 왔다[6, 7, 8, 9]. 이 연구들은 대량의 문서로부터 다량의 단어 자질을 만들어 내고, 이를 상호정보(Mutual Information), 카이제곱 통계량( $\chi^2$  statistics), IG(Information Gain) 등을 이용해 정보량이 많은 효과적인 자질만을 선택하여, 문서 분류에 이용하였다. 이 방법은 대량의 문서로부터 뽑은 다량의 자질에 비해 선택된 적은 수의 자질만을 학습에 사용할 수 있게 하여 시스템의 속도를 증가시켰고, 또 정보량이 낮은 자질을 제거한 효과적인 자질을 학습에 이용할 수 있게 하여 시스템의 정확도도 증가시켰다.

화행 결정을 위해서도 문서분류에서 사용되는 방법들이 응용되었는데, Samuel은 화행결정을 위한 유용한 구나 단어를 추출하기 위해, 대량의 문서에 공기정보, 최

대 엔트로피 모델, 상호정보 등을 적용하여 좋은 결과를 얻었다[10, 11, 12].

본 논문은 기존의 연구를 바탕으로 한국어에 적합한 자동화된 화행분석 시스템을 제안한다. 본 논문에서 제안하는 시스템은 2단계로 이루어져 있다. 첫 번째 단계는 대량의 문서에서 추출된 자질로부터 효과적인 자질을 선택하는 단계로, 실제 상용프로그램에 응용될 수 있게 하기 위해 자연어 처리 분야에서 성능이 검증되어진 형태소 분석 시스템의 결과인 단어정보와 품사정보를 자질로 이용하였으며, 자질 선택을 위해서는 문서 분류 분야에서 가장 좋은 성능을 보인 카이제곱[9]을 이용하였다. 두 번째 단계는 선택된 자질과 신경망을 이용한 화행결정 단계로, 선택된 자질의 유용성을 보이기 위해 선택된 자질의 수를 변화 시켜가며 실험하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 전체 시스템 구조를 보이고, 3장에서는 시스템에서 사용된 방법들을 자세하게 설명한다. 4장에서는 제안된 방법의 유용성을 평가하기 위한 실험을 한 후 5장에서 결론을 맺는다.

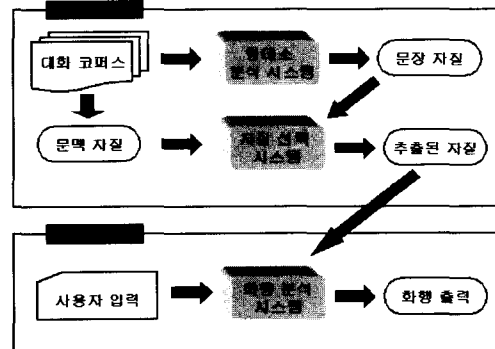


그림 2 전체 시스템 구조

## 2. 전체 시스템 구조

본 논문에서 제안한 시스템은 [그림 2]와 같이 2단계로 구성된다.

첫 번째 단계는 대화 코퍼스로부터 대량의 자질을 추출하고, 추출된 자질 중 효과적인 자질을 선택하는 단계이다. 먼저 본 논문에서는 형태소 분석 시스템[13]을 이용해 문장자질을 생성하고, 대화 코퍼스의 각 발화에 부착된 정답 화행을 이용해 문맥 자질을 생성한다. 그리고 자질 선택 시스템을 통해 정보량이 많은 자질들이 선택된다.

두 번째 단계는 선택된 자질을 바탕으로 화행을 결정

하는 단계이다. 화행 결정 시스템은 선택된 자질과 대화 코퍼스를 바탕으로 신경망을 학습하고, 사용자의 입력이 들어왔을 때 학습된 데이터를 바탕으로 화행을 결정하게 된다.

### 3. 자동 화행 분석 시스템

#### 3.1 대화 코퍼스

대화 코퍼스는 선행된 한국어 화행분석 시스템들에서 사용된 대화 코퍼스를 이용하였다[4, 5, 14, 15]. 이 코퍼스는 호텔, 비행기, 여행 예약에서 실제로 사용된 대화를 녹음후 전사한 것으로 528개 대화, 10,285개의 발화로 구성되어 있으며, 각 발화에 대해서는 수작업을 통해 정확한 화행이 부여되어 있다. [그림 3]은 사용되어진 대화 코퍼스의 일부분으로 SP는 화자, KS는 한국어 발화, SA는 부여된 화행이다.

SP : USER
KS : 숙소에 관해서 문의할 사항이 있어서요
SA : ask-ref
SP : AGENT
KS : 조지아대학의 어학연수 코스는 대학에 기숙사를 제공하고 있습니다.
SA : response

그림 3 대화 코퍼스

#### 3.2 자질 구성

자질은 문장 자질과 문맥 자질로 이루어진다. 문장자질은 각 발화에 나타난 단어와 그 품사의 순서로 문장에 나타난 정보만으로 화행을 결정하는 데 중요한 요소이며, 문맥 자질은 이전 발화의 화행에 대한 정보로 대화의 문맥을 통해 현재 화행 결정에 영향을 미치게 된다.

문장자질은 각 발화를 형태소 분석하여 그 결과로부터 생성되며, 단어/품사, 품사의 바이그램으로 구성된다. [그림 4]는 [그림 3]에서 사용된 USER의 발화를 형태소 분석하여 생성한 문장자질의 예로 KS는 USER의 발화, LP(Lexical/POS)는 형태소분석 결과, PB(POS tagged Bigram)는 품사의 바이그램을 나타낸다.

USER : 숙소에 관해서 문의할 사항이 있어서요.
LP : 숙소/nc 예/jca 관하/pv 어서/ecs 문의/pv 르/exm 사항/nc 이/jc 있/pa 어서요/ef /s.
PB : [nc,jca] [jca, pv] [pv, ecs] [ecs, pv] [pv, exm] [exm, nc] [nc, jc] [jc, pa] [pa, ef] [ef, s.]

그림 4 문장자질

문맥 자질은 이전 화행을 고려하는 것으로 이전 발화의 화행과 현재 발화의 화행이 짝을 이루어 사용되며, 그림 3에서 생성된 문맥 자질은 [ask-ref, response]이다.

#### 3.3 자질 선택 시스템

##### 가. 카이제곱 통계량

카이제곱 통계량은 문서분류 분야에서 자질을 선택할 때 많이 사용되는 방법으로, 자질과 문서분류 사이의 정보량을 잘 나타낸다. 본 논문에서는 자질과 화행간의 정보량을 측정하기 위해 카이제곱 통계량을 사용하였다. 카이제곱 통계량은 자질 F와 화행 S간의 비독립성을 측정하는 것으로 자유도가 1인 카이제곱 분포와 비교될 수 있으며, 식1과 같이 계산되어 진다.

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

A : 자질 f를 가지면서, 화행이 s로 부여된 문서 수

B : 자질 f를 가지면서, 화행이 s로 부여되지 않은 문서 수

C : 자질 f를 가지지 않으면서, 화행이 s로 부여된 문서 수

D : 자질 f를 가지지 않고, 화행이 s로 부여되지 않은 문서 수

각 자질에 카이제곱 통계량을 부여하는 방법은 평균값을 이용하는 방법과 최대값을 이용하는 방법이 있다. 본 논문에서는 실험을 통해 좋은 성능을 보인 최대값을 이용하는 방법을 사용하여 각 자질에 카이제곱 통계량을 부여하였으며, 각 자질에 부여된 카이제곱 통계량은 식2와 같다.

$$\chi^2_{max}(f) = \max_{s=1}^m \{ \chi^2(f, s, ) \} \quad (2)$$

##### 나. 대화 코퍼스로부터 자질 추출

본 논문에서는 대화 코퍼스에서 추출된 문장자질과 문맥자질에 카이제곱 통계량을 부여하여 순위화 한 후 상위 순위의 자질을 선택하였다. [표 1]은 선택된 문장자질의 예로 대표화행은 자질이 나타난 화행 중 최고의 카이제곱통계량을 가진 것이다. “안녕하/pa”가 나타난 발화는 시작 인사말을 나타내는 화행인 opening에 대해 높은 정보량을 가져 1순위가 되었는데, 이것은 대부분의 사람들이 용무를 꺼내기 전에 인사로 시작한다는 사실에 기인한다. 그리고 보통 한국어 대화 습관 중의 하나가 거절을 위해서는 자신이 바쁘다는 말을 먼저 하는 것인데, 이런 사실은 9순위에 나타난 “바쁘/pa”라는 자질과 거절을 나타낸 reject라는 화행이 가지는 높은 정보량에서 찾아 볼 수 있다. 또 2 순위에 나타난 품사 바

이그램 [nn, nn]은 질문에 대한 응답을 나타내는 response와 밀접 관계를 가지고 있는데, 이 사실은 예약에 관한 대화의 질문이 주로 단답형 질문이며, 이 질문에 대해 많은 사람들이 긴 서술적 문장이 아니라, 명사구 위주의 짧은 대답을 하고 있는 영역적 특성을 보여 주고 있다.

표 3 선택된 문장자질의 예

순위	문장자질(대표화행)	카이제곱 통계량
1	안녕하/pa (opening)	4,984.74
2	[nn, nn] (response)	4,236.83
3	[nqc, jcp] (introducing-oneself)	3,372.94
4	알/pv (acknowledge)	3,069.66
5	그렇/pad (ask-confirm)	2,821.44
6	[pad, ef] (opening)	2,170.02
7	[pa, efp] (opening)	1,921.91
8	?/s. (ask-ref)	1,641.81
9	바쁘/pa (reject)	1,588.95
10	잘/a (acknowledge)	1,534.23

[표 2]는 선택되어진 문맥자질의 예로, 첫 번째 발화는 이전 화행이 존재하지 않으므로, "start"라는 가상의 화행을 이용하여 나타냈다. 1순위인 [opening, introducing-oneself]는 "안녕하세요, XXX 여행사입니다"와 같은 시작인사의 응답으로 "예 안녕하세요. 홍길동입니다"와 같이 자신을 소개하는 발화가 이어짐을 나타내 주고 있으며, 대화에는 [질문, 응답], [제안, 수용], [제안, 거절]과 같은 발화의 쌍이 많이 나타나는데, [표 2]는 이런 사실을 잘 보여주고 있다.

표 4 선택된 문맥자질의 예

순위	문맥자질	카이제곱 통계량
1	[opening, introducing-oneself]	2,830.68
2	[start, opening]	2,261.26
3	[ask-ref, response]	1,922.22
4	[introducing-oneself, opening]	1,146.57
5	[reject, suggest]	1,122.55
6	[closing, closing]	898.85
7	[start, introducing-oneself]	708.87
8	[ask-if, response]	679.90
9	[suggest, accept]	424.25
10	[request, accept]	360.37

다. 화행 결정

본 논문에서는 선택된 자질과 신경망을 이용하여 화행을 결정하였다. 본 논문에서는 선택된 자질 수 만큼으로 이루어진 입력층(input layer)과 17개의 화행으로 이루어진 출력층(output layer)을 가지며, 은닉층(hidden layer)를 가지지 않는 선형 분류기 형태의 신경망을 이용하였으며, 입력층의 각 노드 값은 자질 출현 여부에 따라 "0" 또는 "1"의 값을 주었다. 이 신경망은 식 3과 같은 vanilla backpropagation 알고리즘[16]을 이용하여 학습되어, 사용자 입력에 대해 화행을 결정한다.

$$\Delta w_{ij} = \eta \delta_j o_i$$

$$\delta_j = f'_j(\text{net}_j)(t_j - o_j) \quad j \text{ is a output-unit} \quad (3)$$

$\eta$  is learning factor eta(a constant)  
 $\delta_j$  is error of unit  $j$   
 $t_j$  is teaching input of unit  $j$   
 $o_i$  is output of the preceding unit  $i$   
 $i$  is index of a predecessor to the current unit  $j$  with link  $w_{ij}$ , from  $i$  to  $j$   
 $j$  is index of the current unit

4. 실험 및 평가

4.1 실험

본 논문은 기존의 연구에서 사용된 실험데이터와 동일한 2.1절에서 언급된 대량의 코퍼스를 428개 대화, 8,349 발화로 구성된 학습데이터와 100개의 대화, 1,936개의 발화로 구성된 테스트 데이터로 구성하여 사용하였다.

실험은 두 가지 방법으로 진행된다. 실험 1은 우선 문장자질만을 이용하여 화행을 결정하여 보았다. 대량의 문서로부터 3,483개의 문장 자질을 추출하였으며, 자질 선택에 따른 시스템의 성능을 알아보기 위해 선택될 자질의 수를 일정하게 감소시키며 실험하였다. 실험 2는 문장 자질과 문맥 자질을 모두 사용하여 실험하였다. 이 실험에는 문맥자질과 문장자질을 위한 각각의 신경망이 사용되었으며, 문장자질에 의해 결정된 화행이 현재 화행으로 문맥자질을 위한 신경망에 입력되어 최종 화행이 결정된다. 본 논문에서는 대량의 문서로부터 204개의 문맥 자질을 추출하였으며, 역시 자질 선택에 따른 시스템 성능 평가를 위해 선택될 자질의 수를 감소시키며 실험하였다.

4.2 평가

[표 3]은 선택된 문장 자질의 수에 따른 실험 1의 결

파이다. 본 논문에서는 학습데이터 실험에서 자질의 수가 많아질수록 높은 결과를 얻었으나, 비학습 데이터 실험에서는 전체 자질을 사용하는 것 보다 선택된 자질(2,300개)을 사용할 때 더 좋은 성능을 얻을 수 있었다. 그러나 최소자질(200개)보다 적은 수의 자질을 사용하면 시스템의 성능이 급격히 감소함을 볼 수 있었다. 이는 모든 자질을 사용하면 시스템이 그 자질에 편향되어 학습되기 때문에 오히려 비학습 데이터 실험 시 시스템의 성능을 저하시킬 수 있어 정보량이 높은 자질만을 선택하여 사용하는 것이 시스템 성능을 위해 필수적이며, 시스템의 성능을 위해서는 일정 수의 자질이 꼭 필요함을 보여 주고 있다.

표 5 선택된 문장자질 수에 따른 실험결과

선택된 문장 자질수	비학습 데이터 실험	학습 데이터 실험
모두(3,483)	78.86%	93.89%
3,200	79.38%	93.74%
2,900	79.07%	93.59%
2,600	79.12%	93.40%
2,300	79.64%	92.84%
2,000	79.53%	92.24%
1,700	78.55%	91.91%
1,300	79.02%	90.84%
1,000	78.50%	89.75%
700	77.31%	88.40%
500	77.73%	86.90%
300	78.09%	83.95%
200	75.56%	79.89%
100	66.51%	69.51%

[표 4]는 전체 자질에서 각 품사와 품사 바이그램이 전체에 대한 참여율과 최소 자질에서 각 품사와 품사 바이그램이 전체에 대한 참여율을 비교한 것으로 검게 표시된 부분이 전체 자질에 비해 최소 자질에서 참여율이 높아진 품사와 품사 바이그램이다. 기존의 연구에서 중요한 요소로 사용되었던 용언과 보조용언, 단서단어 등과 관련된 품사의 참여 비율이 높아진 것을 볼 수 있는데, 이는 자동으로 추출된 문장 자질들이 기존 연구에서 사용한 자질들을 반영하고 있음을 보여주며, 보통 명사 중 자질로 선택되었던 것들은 대상영역의 특징을 반영한 보통명사로 “안녕하세요 xxx 여행사입니다”에서 “여행사”와 같이 도메인에 따라서 화행을 결정에 중요

한 역할을 할 수 있는 보통 명사들이다. 품사 바이그램은 각 화행을 이루는 문장의 구문구조가 반영된 것으로 “미국 보스턴이요”와 같이 질문에 대한 짧은 응답 시 나타나는 “명사+명사”의 형태나, “저는 흥길동입니다”와 같이 자기 소개에 나타나는 명사+서술격 조사 등이다.

표 6 전체 자질과 최소 자질에서 각 품사와 품사 바이그램의 참여율

품사/ 품사바이그램	전체에 나타난 빈도	참여율	선택된 200개 문장자질에 나타난 빈도	참여율
부사	183	0.0525	16	0.0800
품사 바이그램	183	0.0525	32	0.1600
형용사	114	0.0327	16	0.0800
동사	360	0.1033	29	0.1450
보조동사	25	0.0071	6	0.0300
어미	186	0.0534	27	0.1350
감탄사	38	0.0109	4	0.0200
조사	94	0.0269	9	0.0450
관형사	37	0.0106	3	0.0150
수사	33	0.0094	11	0.0550
대명사	42	0.0120	3	0.0150
기호	4	0.0011	2	0.0100
서술형 동사	182	0.0522	7	0.0350
외국어	2	0.0005	0	0.0000
의존명사	145	0.0416	3	0.0150
보통명사	927	0.2661	30	0.1500
고유명사	894	0.2566	2	0.0100
접사	34	0.0097	0	0.0000
계	3,483	1.0	200	1.0

[표 5]는 문장자질과 문맥자질을 함께 고려한 실험 2의 결과이다. 전체 문맥 자질을 사용했을 때 보다, 적절한 문맥자질(40 개)을 사용했을 때 최고의 성능을 얻었으며, 실험 1에서 가장 높은 성능을 보인 2,300개의 문장자질을 사용할 때 보다 실험 1에서 사용된 300개의 문장자질을 사용하였을 때 높은 성능을 얻었다. 이것은 많은 수의 자질을 사용한 학습이 학습데이터에 편향된

결과(overfitting problem)를 가져와 그것에서 발생할 수 있는 오류를 문맥자질을 사용하여 수정하기 어렵다는 사실에 기인한다.

표 7 선택된 문맥자질의 수에 따른 실험결과

		문장자질 300개 결과		문장자질 2300개 결과	
		O	C	O	C
선택된 문맥 자질수	20	85.12	90.01	83.98	96.35
	40	88.32	92.91	86.05	96.95
	60	87.75	92.82	85.94	96.71
	80	87.24	92.48	85.68	96.47
	모두 (204)	83.72	89.95	82.27	95.68

\* 단위 : 백분율

O : 비학습 데이터 실험

C : 학습 데이터 실험

[표6]은 화행 결정을 위한 기존 최원석의 연구결과(System1)[4], 이성욱의 연구결과(System2)[5]를 본 논문의 연구결과(System3)와 비교한 것으로 본 논문에서 제시한 자질 선택방법과 학습 알고리즘을 사용한 시스템이 좋은 성능을 나타내고 있다.

표 6 화행 결정을 위한 기존 연구와의 비교

System	실험 유형	성능 (%)
System1	비학습데이터실험	83.3
	학습데이터실험	90.6
System2	비학습데이터실험	83.7
	학습데이터실험	88.2
System3	비학습데이터실험	88.3
	학습데이터실험	97.0

### 5. 결론 및 향후 과제

화행은 자연어로 된 발화를 이해하고, 이에 대한 응답을 생성하기 위해 중요한 요소이다. 그러나 기존의 화행 결정에 대한 연구는 자질의 선택보다는 학습 알고리즘 선택에 초점이 맞추어져 왔다. 본 논문에서는 추출되어지는 모든 자질을 이용하는 것 보다 정보량이 많은 자질만을 선택하는 것이 화행 결정 시스템의 정확도를 높일 수 있음을 보였다. 본 논문은 자질 선택의 방법으로 문서분류에서 높은 성능을 보인 카이제곱 통계량을 이

용하였고, 자질 선택방법이 효과적임을 증명하기 위해 신경망을 이용한 화행분석시스템을 구성하여, 자질 선택에 따른 시스템의 정확도를 측정하여 좋은 결과를 얻었다. 또 본 논문에서는 자질 선택을 이용해 화행결정에 참여하는 자질의 수를 감소시킴으로서, 시스템의 속도 향상과 리소스 사용감소라는 부가적인 장점을 얻었으며, 이런 장점은 상용 시스템에서 필요로 하는 부분이다. 그리고 기존의 한국어 화행분석 연구[4, 5]에서 사용되던 구문정보를 사용하지 않고, 형태소 분석 결과만을 이용하여, 학습데이터 작성 시 정답화행을 부여하는 것 이외의 모든 작업을 자동화시킴으로서 기존의 다른 시스템 보다 타영역으로 이식성이 좋고 일반적인 화행분석 시스템을 구성하였다.

그러나 본 논문에서 제시한 시스템은 여행과 관련된 예약이라는 제한된 영역에서 수행되어, 일반성을 증명하기 위해서는 다른 영역들을 대상으로 한 실험이 필요하다. 또 사용되어진 화행도 기존의 연구에서 정의한 화행을 사용하였는데, 이 화행 중에는 영역에 특화된 화행도 존재하여, 보다 일반적인 화행 분석 시스템을 구성하기 위해서는 화행에 대한 재정의가 필요하며, 이를 위해서는 다양한 영역에 대한 대화 코퍼스를 구성 및 분석이 필요하다.

### 참고 문헌

- [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] Kim, jin Ah, et al. A response generation in dialogue system based on dialogue flow diagrams. *In Proceedings of NLPRS*, 1995.
- [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] Lee, Songwook, and Jungyun Seo. Korean Speech Act Analysis Using Decision Tree. *In Proceedings of the Conference on Hangeul and Korean Language Information Processing*, 1999. pp. 377-381
- [6] Lweis, D.D. and M. Ringuette. 1994. Comparison of two learning algorithms for text categorization. *In Proceedings of the Third Annual Symposium*

on Document Analysis and Information Retrieval (SDAIR'94), 1994.

- [7] Schütze, H., D.A. Hull, and J.O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *18<sup>th</sup> Annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '95)*, 1995.
- [8] Wiener, E., J.O. Pedersen and A.S. Weigend. A neural network approach to topic spotting. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, 1995.
- [9] Yang, Yiming and Jan O. Pedersen. A comparative study on Feature selection in text categorization. In *proceedings of the 14<sup>th</sup> International conference on Machine Learning*, 1997.
- [10] Samuel, K., S. Carberry, and K. Vijay-Shanker. Dialogue Act Tagging with Transformation-Based Learning. In *Proceedings of the 17<sup>th</sup> International Conference on computational Linguistics and the 36th Annual Meeting of the Association for computational Linguistics*, 1998. pp 1150-1156.
- [11] Samuel, K., S. Carberry, and K. Vijay-Shanker. An Investigation of Transformation-Based Learning in Discourse. *Machine Learning: Proceedings of the 15<sup>th</sup> International Conference*. 1998.
- [12] Samuel, K. and S. Carberry, and K. Vijay-Shanker. Automatically Selecting Useful Phrases for Dialogue Act Tagging. In *Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics*, 1999.
- [13] Jae-hoon Kim, Jungyun Seo, Gilchang Kim, Estimating Membership Functions in a Fuzzy Network Model for Part-Of-Speech Tagging, *Journal of Intelligent and Fuzzy Systems*, Vol. 4, pp.309-320, 1996.
- [14] Lee, Jae-won, Jungyun Seo, Gilchang Kim. "A dialogue analysis Model with statistical speech act processing for Dialogue Machine Translation," *Proceedings of Spoken Language Translation (Workshop in conjunction with (E)ACL'97*, page 10-15, 1997.
- [15] Lee, Hyunjung, Jae-Won Lee, Jungyun Seo. "Speech Act Analysis Model of Korean Utterances for Automatic Dialog Translation," *Journal of Korea Information Science Society (B): Software and Applications*, 25(10) : 1433-1552, 1998.
- [16] Rumelhart, D.E. and J.L. McClelland. *Parallel Distributed Processing, volume 1*. MIT Press. 1986.

김 경 선

1997년 서강대학교 전자계산학과 학사. 1999년 서강대학교 컴퓨터학과 석사. 1999년~현재 서강대학교 컴퓨터학과 박사과정. 2001년~현재 다이렉트 연구소 선임연구원. 관심 분야는 자연어 처리, 대화 시스템, 화행 분석, 정보 검색, 정보 분류



서 정 연

1981년 서강대학교 수학과 학사. 1985년 미국 Univ. of Texas, Austin 전산학과 석사. 1990년 미국 Univ. of Texas, Austin 전산학과 박사. 1990년~1991년 미국 Texas Austin, UniSQL Inc. Senior Researcher. 1991년 한국과학기술원 인공지능 연구센터 선임연구원. 1991년~1995년 한국과학기술원 전산학과 조교수. 1995년~1996년 서강대학교 전산학과 조교수. 1996년~현재 서강대학교 컴퓨터학과 부교수. 관심분야는 한국어정보처리, 자연언어처리, 기계번역, 대화처리

