결정트리를 이용한 한국어 화행 분석

이성욱, 서정연
서강대학교 컴퓨터학과 자연어처리 연구실

Korean Speech Act Analysis Using Decision Tree

Songwook Lee, Jungyun Seo
Natural Language Processing Lab., Dept. of Computer Science, Sogang Univ.

요약

담화 분석에서 화자의 의도와 대화의 흐름을 이해하기 위해서 화행 분석이 중요하다. 최근에 대화 말뭉치를 이용하여 화행을 결정하는 방법들이 많이 연구되어 왔다. 발화 특성 정보를 이용한 통계적 화행 분석과 담화 구조를 최대 엔트로피 모델에 적용한 연구가 있었다. 그러나 이러한 연구에서 발화의 어떤 특성 정보가 실제 화행 결정에 중요한 역할을 하는지 알기가 어렵다. 그러나 결정 트리를 이용한 본 연구는 결정트리의 분리자를 통해 어떤 정보들이 화행결정에 영향을 끼치는지 알 수 있다는 장점이 있다. 본 연구는 결정트리를 이용하여 화행을 결정하였으며, 현재 발화의 이전 발화 정보만을 고려한 bigram, 이전 두 발화의 화행을 고려한 trigram, 또한 담화 구조를 고려한 trigram모델을 비교 분석하였다.

1. 소개

자연어 대화에서 화자가 발화를 통해 상대방에게 나타내고자 하는 의도적인 행위를 화행이라 한다. 담화 분석에서 화자의 의도와 대화의 흐름을 이해하기 위해서 화행 분석이 중요하다. 발화의 화행을 시스템이 분석하기 위해서 해당 발화의 의미 정보와 대화의 흐름을 고려해야한다.

최근에는 대화 말뭉치를 이용하여 화행을 결정하는 방법들이 많이 연구되어 왔다. [1]에서는 발화의 특성 정보와 담화의 구조를 고려한 통계적 화행 분석 연구 결과, 담화 구조를 반영한 통계 모델의 성능이 담화 구조를 반영하지 않은 통계 모델보다 성능이 나음을 보였다. [3]은 최대 엔트로피 모델을 이용하여 화행을 결정할 때 가능한 담화 구조를 모두 살펴본다. 따라서 화행 결정 모델에 담화 구조의 결정이 포함되었다. 이러한 통계 기반 모델은 말뭉치를 이용하여 학습된 확률값을 이용하기 때문에 영역 지식이 필요없고 영역 확장이 쉬운 장점이 있다. 그러나 이러한 통계 기반 연구에서는 어떤 지식이 화행 결정에 중요한 역할을 하는지 알기 어렵다. 결정 트리를 이용하면 말뭉치 속에 포함된 통계적 정보를 자동적으로 학습하여 규칙으로 보여주는 장점이 있다. 결정트리의 분리자(splitter)를 살펴보면 문제 해결을 위해 사용된 지식의 분석도 가능하다.

결정 트리는 자연어처리의 품사태깅, 통계적 구문 분석, 의미 중의성 해소 등 여러 분야에 이용되어왔다[4,5,6,7]. 본 연구는 결정트리를 이용하여 화행을 결정하고자 한다. 본 연구는 결정트리를 이용하여 화행을 결정한다. 그리고 학습된 결정트리로부터 화행결정에 필요한 언어적 지식을 얻는다. 현재 발화의 이전 발화의 화행과 발화자 정보를 문맥으로 사용한 bigram 모델, 이전의 두 발화의 화행과 발화자 정보를 문맥으로 사용한 trigram 모델, 또한 담화 구조를 반영한 trigram 모델에 대해 비교 분석하였다. 결정트리를 만드는데는 CART를 이용하였다[5,9].

본 연구는 과학재단 특정기초연구 97-01-02-03-01-3의 지원으로 이루어진 것임.

2. 대화 말뭉치

본 연구에서는 [1]과[3]에서 사용된 전화 예약 영역(호텔 예약, 항공 예약, 여행 예약 등)에서 수집한 528 대화, 10,285 문장으로 구성된 말뭉치를 사용한다. 대화 말뭉치에는 발화자 정보(SP), 발화 문장(KS), 구문 유형(ST), 화행(SA), 담화 구조(DS) 등의 담화 정보가 태깅되어 있다. 그림1은 담화 정보가 태깅된 말뭉치의 일부분이다.

구문 유형은 화행 분석에서 발화 대신에 사용하는 것으로 발화 문장의 구문 정보로 구성되어 있다. 구문 정보는 문장 유형, 본용언, 시제, 문장의 부정형 여부, 양상, 단서 단어로 구성되며, 본용언과 단서 단어에는 어휘 정보를 포함하고 있다.

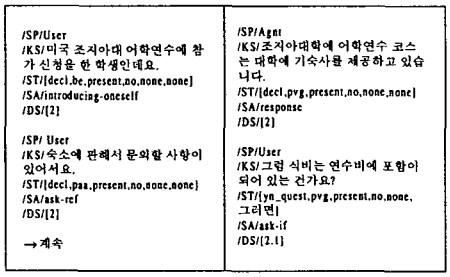

그림 1. 대화 말뭉치의 예

3. 결정트리 학습에 사용된 담화 지식

결정트리를 학습하기 위해, 대화 말뭉치에서 화행과 구문 유형 정보, 발화자 정보를 추출하여 사용했다. 발화자 정보는 발화자가 고객인지 직원인지를 나타낸다. 화행에는 대화 말뭉치에서 발견되는 17개의 화행을 사용하였다. 본 연구에 사용한 화행과 구문 유형 정보의 예는 표1, 표2와 같다[1,3].

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.

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

화행을 결정하기 위한 i번째 발화를 U_i이라 하면 결정트리의 학습에 사용하는 변수들은 다음과 같다.

SA_{i-2} : U_{i-2}의 화행
SA_{i-1} : U_{i-1}의 화행
SA_i : U_i의 화행, 목표 변수
SP_{i-2} : U_{i-2}의 발화자 정보
SP_{i-1} : U_{i-1}의 발화자 정보
SP_i : U_i의 발화자 정보
SenType : U_i의 문장 유형
MainV : U_i의 본용언
Tense : U_i의 시제
Neg : U_i의 부정형여부
Modal : U_i의 양상
Clue : U_i의 단서 단어

bigram 모델에서 화행을 결정하기 위해 사용하는 변수는 다음과 같다.

{SA_{i-1}, SP_{i-1}, SP_i, *SenType, MainV, Tense, Neg, Modal, Clue*}

trigram모델에서 사용하는 변수는 다음과 같다.

{SA_{i-2}, SP_{i-2}, SA_{i-1}, SP_{i-1}, SP_i, *SenType, MainV, Tense, Neg, Modal, Clue*}

담화구조를 고려한 trigram 모델의 이전 발화는 현재 발화와 선형적인 인접 발화가 아니고 담화 구조를 고려한 계층적인 인접 발화이다.

4. 실험

대화 말뭉치 528대화 중 428대화 8349문장을 학습 말뭉치로 사용하였고 100대화 1936문장을 평가 말뭉치로 사용하였다. 본 연구의 평가 말뭉치와 학습 말뭉치는 [1]에서 사용한 말뭉치를 다시 태깅한 것으로 [1]의 말뭉치보다 화행과 담화 구조 정보의 일관성이 높다. 본 연구에서 사용한 학습 말뭉치와 평가 말

뭉치는 [3]에서 사용한 말뭉치와 동일하다.

CART를 이용하였고 결정트리 구성을 위한 파라미터는 다음과 같다[5,9].

Best Tree : minimum cost tree regardless of size
splitting method : gini
priors: learn
categorical variables: SA_{i-1}, *MainV*, *Modal*, *Clue*

표3은 학습된 최적의 결정 트리에 대한 평가 말뭉치와 학습 말뭉치에 대한 정확률과 다른 관련 연구들의 정확률을 나타낸다.

표 3. 결정트리 모델과 기존 연구들의 정확률 비교

모델		평가 말뭉치	학습 말뭉치	사용 문맥
선형 인접 기반 모델	(a)bigram	83.7	88.2	SA_{i-1}, SP_{i-1}
	(b)trigram	77.8	80.8	SP_{i-1}, SP_{i-2}
	(c)trigram	83.6	89.2	SA_{i-1}, SP_{i-1} SA_{i-2}, SP_{i-2}
계층 인접 기반 모델	(d)trigram	84.1	90.2	SA_{i-1}, SP_{i-1} SA_{i-2}, SP_{i-2}
	(e)trigram	80.8	85.6	SP_{i-1}, SP_{i-2}
	(f)trigram	85.1	89.7	SA_{i-1}, SP_{i-1} SA_{i-2}, SP_{i-2}
*L-Model		71.2	86	SA_{i-1}, SA_{i-2}
**H-Model		74.5	88.6	SA_{i-1}, SA_{i-2}
***MEM		81.9	90.6	SA_{i-1}, SA_{i-2}

*L-Model은 [1]의 담화 구조를 고려 안한 화행 분석 모델.
**H-Model은 [1]의 담화 구조를 고려한 화행 분석 모델.
***MEM은 [3]의 최대 엔트로피를 이용한 화행 분석 모델.

위의 각 모델에서 문맥으로 사용되는 이전 발화들의 화행 정보는 항상 올바르다고 가정했다. 즉 현재 발화에 대한 화행 분석 오류가 누적되지 않는다고 가정했다. 현재 발화의 화행을 잘못 결정하여도 그 오류가 다음 발화의 화행 결정에 영향을 미치지 않는 것이다. 또 계층 인접 기반 모델에서는 담화 구조를 이미 알고 있다고 가정했다. 담화 구조를 결정하는 것은 화행과 관련된 또다른 연구 주제이다. L-Model과 H-Model의 정확률도 이와 동일한 가정을 적용했을 때의 정확률이다[1].

모델(b)와 모델(e)는 결정트리 학습에서 이전 발화의 화행들을 문맥으로 사용하지 않고 이전 발화들의 발화자 정보를 문맥으로 사용한 모델이다. 모델(b)와 모델(e)는 오류 누적 문제가 근본적으로 발생하지 않는 모델이다.

선형 인접 모델에서 현재 발화의 이전 발화는 선형적인 이전 발화이고 계층 인접 기반 모델에서 이전발화는 담화 구조를 고려한 계층적인 이전발화이다[1].

모델(d)과 모델(f)의 차이점은 담화 구조에서 이전 발화의 범위가 다른데 있다. 모델(d)의 경우 U_i가 U_{i-1}의 부대화일 때 U_{i-1}을 U_i의 이전 발화로 간주한다. 그러나 모델(f)의 경우 U_i가 U_{i-1}의 부대화인 경우에 U_i의 이전 발화를 U_{i-1}로 간주하지 않는다. 만약 담화 구조상 부대화가 이전 발화의 영향을 많이 받는다면 모델(d)가 신빙성이 있을 것이고 부대화의 화행이 이전 발화의 영향을 적게 받는다면 모델(f)의 성능이 더 좋아야 할 것이다. 실험 결과를 보면 모델(d)와 모델(f)의 차이가 크지 않음을 알 수 있다. 이는 부대화의 화행이 항상 이전발화의 영향을 받을수도 있고 안받을 수도 있다는 것으로 해석된다. 실제로 어떤 대화가 이전 대화의 부대화인지 새로운 대화의 시작인지를 구별하는 것은 상당히 어려운 문제이다. 대화 말뭉치에 태깅된 담화 구조 정보 자체에 이러한 문제가 내재되어 있다고 여겨진다.

본 연구에서 담화 구조를 결정하는 것은 고려하지 않으므로 [1]과 [3]의 통계적 모델이 담화 구조를 결정하는 점에서 장점을 가진다. 결정트리 모델과 [3]의 MEM 모델을 객관적으로 비교하기는 어렵다. 오류의 누적을 고려하는 이상적인 결정 트리 모델이 연구된다면 그 모델의 정확률은 선형 인접 기반 모델에서는 모델(b)와 모델(c)의 정확률 사이, 계층 인접 기반 모델에서는 그 모델은 모델(e)와 모델(f)사이의 정확률을 보일 것이라 추정할 수 있다. 모델(e)에서 얻어지는 각 화행별 확률을 이용하고 모델(f)에서 얻어지는 확률을 결합하면 담화 구조를 결정하는 기존의 통계적 방법을 적용하여 담화 구조를 결정할 수 있을 것이다[1].

결정트리의 분리자의 비율을 보면 화행을 결정하는 변수의 상대적 중요도를 알 수 있다. 화행 결정 변수의 상대적 중요도는 표4와 같다. 상대적 중요도는 가장 중요도가 높은 변수를 100으로 보았을 때 다른 변수의 중요도를 수치로 나타낸 것이다.

표 4. 모델(f)에서 변수들의 상대적 중요도

변수	상대적 중요도	변수	상대적 중요도
MainV	100	SP_{i-2}	21
SA_{i-1}	68	*Clue*	18
Modal	49	SP_i	8
SenType	37	*Tense*	3
SP_{i-1}	35	*Neg*	3
SA_{i-2}	22		

표4에서 화행 결정에 가장 중요한 비중을 차지하는 것은 *MainV*(본용언)과 SA_{i-1}(이전 발화의 화행)임을 알 수 있다. 상대적 중요도가 높다는 것은 화행 결정에 기여하는 바가 크다는 것을 의미한다. 이런 변수는 변수를 이루는 특성 정보들이 잘 구축되었다고 볼 수 있다. 그러나 상대적 중요도가 낮은 정보들은 화행 결정에 기여하는 바가 적다고 할 수 있고, 특성 정보가 불필요하다고 볼 수 있다. 그렇지만 상대적 중요도가 낮은 정보도 특정 화행의 결정에서 중요하게 사용될 수도 있으므로 이러한 정보는 유지하고 불필요한 정보는 없애는 것이 필요하다. 화행 결정 트리를 살펴보면, 화행 결정에 관여하지 않는 불필요한 정보를 찾아낼 수 있다. 즉, 대화 말뭉치 구축에서 상대적 중요도가 낮은 정보들을 통합하거나 화행 결정에 기여할 수 있는 정보를 더 추가하는 것이 요구된다.

그림2와 그림3은 모델(f)에서 평가 말뭉치에서 각 화행별 정확률과 각 학습 말뭉치의 각 화행의 비율을 나타낸 그림이다.

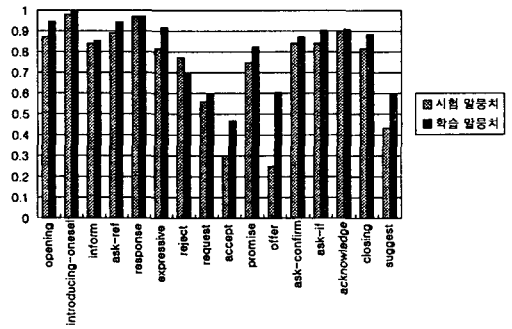

그림 2. 화행의 정확률

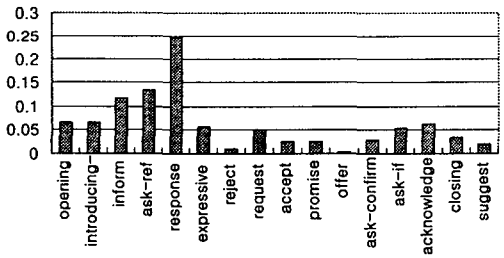

그림 3. 학습 말뭉치에서 화행의 비율

그림 2를 보면 화행 결정에서 낮은 정확률을 보이는 화행은 'offer', 'accept', 'suggest', 'request' 등이다. 이 중 'offer'와 'suggest', 'request'는 대화의 주도권을 가지는 발화이므로 문맥으로 예측이 상대적으로 더 어렵다[10]. 게다가 그림3을 보면, 이러한 화행의 학습 말뭉치에서 출현 비율도 다른 화행에 비해 비교적 낮은 것을 확인할 수 있다. 이런 화행들을 결정할 수 있는 규칙을 학습하기 위한 자료가 부족해서 더욱 다른 화행에 비해 정확률이 낮게 나타났다.

한편, 'promise', 'reject' 등과 같은 화행은 학습 말뭉치에서 낮은 빈도수에도 불구하고 다른 저빈도 화행에 비해 비교적 높은 정확률을 가진다. 이런 화행은 본용언과 양상, 문장 유형, 부정형 유무, 그리고 이전 발화의 화행으로 화행 결정 규칙을 만들어냈다. 이런 발화는 대화의 주도권을 가지는 발화가 아니고 이전 발화에 의해 유도되는 발화이기 때문에 상대적으로 적은 정보만으로도 다른 화행과 구별되는 특성을 가진다고 볼 수 있다[10].

표5는 모델(f)에서 화행 결정의 대표적인 오류 유형들을 보인 것이다. 표5를 보면 어떤 화행이 어떤 화행과 혼동해서 사용되는지 보여준다.

표 5. 대표적 오류 유형

올바른 화행	잘못 결정된 화행
opening	closing
inform	reponse, introducing-oneself
ask-ref	inform
request	inform,expressive, reject
reject	inform
request	inform, ask-if
accept	inform
promise	inform
offer	inform
ask-confirm	ask-if
suggest	ask-ref, request
acknowledge	inform
closing	request
suggest	inform, ask-ref, ask-if

모델(f)의 결정트리를 적용한 대표적인 오류 유형들은 대부분 동일한 구문 유형 정보를 갖고 있었다. 이런 오류는 혼동하는 화행들을 구분할 수 있는 정보가 부족하기 때문에 발생하는 오류라고 할 수 있다. 예를 들어, 표5의 'request'와 'inform'같은 화행을 가지는 서로 다른 발화에 대해 [decl, pvg, present, *no*, *want*, 예] 등으로 동일한 구문 유형 정보를 가지는 경우가 있었다. 이와 같이 서로 다른 화행을 가지는 발화에 대한 구문 유형 정보가 서로 동일한 경우가 많아서 각 발화에 대한 변별

력이 떨어졌다.

이런 화행들을 잘 구분하기 위해서는 구문 유형 정보의 세분화가 필요하다. 화행 결정에 중요한 비중을 차지하는 변수들을 중심으로 구문 유형 정보의 세분화가 필요하며 반대로 지나친 세분화는 자료부족 문제를 일으키므로 유사한 정보에 대해서는 범주화가 요구된다. 특히 본용언과 양상, 단서 단어 등의 세부 쓰임에 따른 범주화가 필요하다.

결정트리를 살펴보면 어떤 변수가 어떤 화행을 결정하는지 알 수 있다. 예를 들어 화행 'response'를 결정하는 결정트리의 일부분을 규칙으로 보이면 그림4와 같다. 그림4를 보면 'response'가 이전 발화의 화행과 현 발화의 본용언과 양상에 따라 결정된다는 것을 알 수 있다. 이와 같이 결정 트리를 구성하는 각 노드를 살펴보면 각각의 화행을 결정하는 언어적 규칙을 얻을 수 있는 장점이 있다.

```
if
(
  (
      MAINV == "알리다" ||
      MAINV == "가르치다"
  ) &&
  (
      SA_{i-1} == "ask-confirm" ||
      SA_{i-1} == "ask-if" ||
      SA_{i-1} == "correct"
  ) &&
  (
      MODAL == "seem" ||
      MODAL == "try, will" ||
  )
)
{
   class = "response";
}
```

그림 4. 'response'에 대한 결정트리 규칙의 일부분

5. 결론 및 향후 과제

본 연구는 대화의 화행 결정과 화행 결정 규칙을 얻기 위해 결정트리를 이용하였다. 결정트리를 이용하면 각 화행을 결정하는 언어학적 규칙을 얻을 수 있으며, 그 규칙은 쉽게 다른 응용 시스템에 적용될 수 있는 장점이 있다. 결정트리를 구성하기 위한 정보는 이전 발화의 화행과 현재 발화의 구문 유형 정보를 이용하였다. 담화 구조를 고려하여 이전 두 발화의 발화자 정보와 화행을 이용하는 화행 분석 모델이 담화 구조를 고려하지 않은 모델보다 나은 성능을 보였다.

화행 분석 모델의 성능을 향상시키기 위해서는 많은 양질의 대화 말뭉치 구축이 요구된다. 또 대화 말뭉치 구축에 있어 발화에 대한 구문 유형 정보의 세분화가 필요하다. 특히 담화 구조를 결정할 때 일관성있는 상세한 태깅 지침이 요구된다.

앞으로, 얻어진 결정트리의 각 화행별 확률을 담화 구조 결정에 이용하여 화행과 담화 구조를 동시에 결정할 수 있는 방법의 연구가 필요하다. 결정트리에서 담화 구조를 결정할 수 있고, 이전 발화의 화행 결정에 대한 오류에도 일정한 성능을 보장한다면 다른 시스템에 쉽게 적용할 수 있는 화행 분석 모듈이 될 것이다.

참고문헌

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

[2] 이현정, 한국어 대화체 문장의 화행 분석, 석사학위논문, 서강대학교, 1997

[3] 최원석, 최대 엔트로피 모델을 이용한 화행 및 담화구조 분석 시스템, 서강대학교, 석사학위 논문, 1998

[4] Black, E. etc, "Decision Tree Models applied to the labeling of text with part-of-speech," in Proceeding of workshop on speech and natural language, pp. 117-121, 1992.

[5] Breiman, L. etc, Classification and Regression Trees, Wadsworth international, 1984.

[6] Márquez, L. etc, "POS Tagging Using Decision Tree," ECML-98, 1998.

[7] Samuel, Ken, Carberry, Sandra, and Vijay-Shanker, K.,"Dialogue Act Tagging with Transfomation-Based Learnging, " In Proceeding of COLING-ACL, 1998.

[8] V. Siegel, E. etc, "Emergent Linguistic Rules from Inducing Decision Trees: Disambiguating Discourse Clue Words," Proceeding of AAAI, 1994.

[9] Steinberg, D. etc, CART$_{TM}$, Salford Systems, 1997.

[10] Nagata M. etc, "Rhetorical structure theory: Description and construction of text structures," Natural Language Generation, Martinus Nijhoff, pp.83-203, 1994.