인지과학, 제20권 제4호
Korean Journal of Cognitive Science
2009, Vol. 20, No. 4, 421~437

복합 자질 정보를 이용한 통계적 한국어 채팅 문장 생성*

김 종 환 장 두 성 김 학 수†
강원대학교 컴퓨터정보통신공학과 KT 중앙 연구소 강원대학교 컴퓨터정보통신공학과

채팅 시스템은 인간이 사용하는 언어를 이용하여 인간과 컴퓨터 간의 대화를 시뮬레이션하는 프로그램이다. 본 논문에서는 핵심어와 화행을 입력으로 받아 자연스러운 채팅 문장을 생성하는 통계 모델을 제안한다. 제안 모델은 먼저 핵심어를 포함한 어절을 말뭉치에서 선택하고, 해당 어절의 주위에 있는 어절의 출현 정보와 구문 정보를 이용하여 후보 문장들을 생성한다. 그리고 화행에 기초한 언어 모델, 어절간 공기 정보, 각 어절의 구문 정보를 이용하여 생성된 후보 문장 중 하나를 선택한다. 실험 결과에 따르면 제안 모델은 단순한 언어 모델에 기반한 기존의 모델보다 좋은 86.2%의 적합 문장 생성률을 보였다.

주제어 : 채팅 문장, 통계적 문장 생성, 채팅 시스템

---

* 이 연구(논문)는 지식경제부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다.
†교신저자: 김학수, 강원대학교 컴퓨터정보통신공학과, 연구세부분야: 자연어처리
E-mail: nlpdrkim@kangwon.ac.kr

Ⅰ. 서 론

채팅 시스템은 인간이 사용하는 언어를 이용하여 인간과 컴퓨터 간의 대화를 시뮬레이션(simulation)하는 프로그램을 말한다. 일반적으로 채팅은 특별한 목적 없이도 흐름을 유지한 채 대화를 진행해 나가는 특징이 있다. 따라서 채팅 시스템은 사용자로 하여금 마치 사람과 채팅하는 것 같은 느낌이 들 수 있도록 자연스러운 응답을 생성해야 한다. 채팅 시스템은 인간에게 익숙한 자연어를 사용한다는 점에서 사용자 친화적인 인터페이스이며, 목적지향 대화 시스템을 보완하여 시스템 환경을 인간 중심적으로 만든다는 점에서 중요한 의미를 가진다. 그러나 현재까지 연구된 채팅 시스템들은 특정 영역에 한정되거나 인간의 복잡하고 다양한 대화 현상을 설명하기에는 부족한 성능을 보여주었다. 이러한 문제를 해결하기 위해 본 논문에서는 입력 파라미터(parameter) 조절을 통해 자연스러운 한국어 채팅 문장을 생성하는 통계 기반의 문장 생성 방법을 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 채팅 시스템에 대한 기존의 연구를 살펴보고, 3장에서는 제안하는 한국어 문장 생성 시스템에 대해 자세히 설명한다. 4장에서는 제안된 시스템의 성능을 평가하기 위한 실험을 한 후, 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

기존의 채팅 시스템에 관한 연구로는 패턴 매칭(pattern matching) 방법, 마르코프 모델(Markov model), 유전 알고리즘(genetic algorithm) 등이 있다. 패턴 매칭을 이용한 채팅 시스템은 일반적으로 실제 사용자가 입력한 문장을 이해하고 응답하는 것이 아니라 입력 문장과 규칙의 비교를 통해 데이터베이스에 저장된 알맞은 응답을 찾아내는 방법으로 입력 문장과 응답 문장으로 이루어진 테이블을 사용한다. 채팅 시스템의 시초라 할 수 있는 ELIZA[1]는 사용자가 입력한 문장에 포함된 키워드로부터 미리 정의된 문장으로 응답하는 간단한 패턴 매칭 방법을 사용한다. 이를 계승한 시스템으로 Julia[2], Alice[3] 등이 있으며, 다양한 기능 추가와 학습을 통해 지

속적인 대화 능력의 향상을 가져왔다. 패턴 매칭 방법은 구현이 매우 쉽다는 장점이 있지만 비슷하거나 동일한 대화가 계속되어 채팅의 유연성이 부족하고, 채팅의 수준을 향상시키기 위해 대량의 데이터베이스가 필요하다는 단점이 있다. 마르코프 모델을 이용한 방법은 인간의 대화 내용을 마르코프 모형으로 저장한 후 사용자의 입력에 대해 데이터를 조회하여 가장 적절한 문장을 생성한다. MegaHal[4]은 마르코프 모델을 적용한 대표적인 채팅 시스템으로써 영어와 같이 띄어쓰기 단위가 기본 단위인 언어에 적합하도록 설계 되었다. 따라서 한국어와 같이 띄어쓰기가 기본 단위가 아닌 언어에 대해서는 그대로 적용 할 수 없기 때문에 각 언어의 특성을 고려한 변형이 필요하다. 유전 알고리즘은 기존 DB에 저장되어 있는 데이터로부터 새로운 데이터를 생성하는 방법이다. Dana Vrajitoru[5,6]는 유전 알고리즘을 활용하여 기존 문장으로부터 더 나은 응답 문장을 생성하는 방법을 제안하였다. 그러나 유전 알고리즘을 이용한 방법은 항상 문법에 맞는 문장이 생성되는 것은 아니기 때문에 잘못된 결과를 필터링할 수 있는 방법에 대한 연구가 필요하다.

Ⅲ. 한국어 채팅 문장 생성 시스템

인간의 대화에서 문장은 대화의 상황과 사용자의 의도에 따라 결정되며, 채팅 시스템에서는 문맥 자질과 문장 자질을 통해 이러한 현상을 구현한다. 그러므로 본 논문에서는 화행과 핵심어가 문맥 자질과 문장 자질로 결정되어 있다고 가정한다[7]. 본 논문에서는 사용자의 의도에 적합한 응답 문장의 유형을 결정하는 문맥 자질로 화행을, 대화의 흐름을 유지하는 문장 자질로 핵심어를 사용한다. 본 논문에서 제안하는 문장 생성 과정을 살펴보면 그림 1과 같다. 먼저 주어진 핵심어를 포함한 어절을 말뭉치에서 찾고, 그것의 앞뒤에 나타나는 어절의 출현 정보와 문장의 구성 성분 정보에 따라 어절을 확장하여 후보 문장들을 생성한다. 그리고 각 후보 문장의 통계 정보와 화행에 따른 자질을 고려한 언어 모델(language model)을 적용하고 각 후보 문장에 포함된 어절 간의 공기 정보(co-occurrence)를 이용하여 연관도를 측정한다. 마지막으로 문장 구성 성분 정보에 따른 가중치를 통합하여 가장 수치가 높은 문장을 응답 문장으로 선택한다.

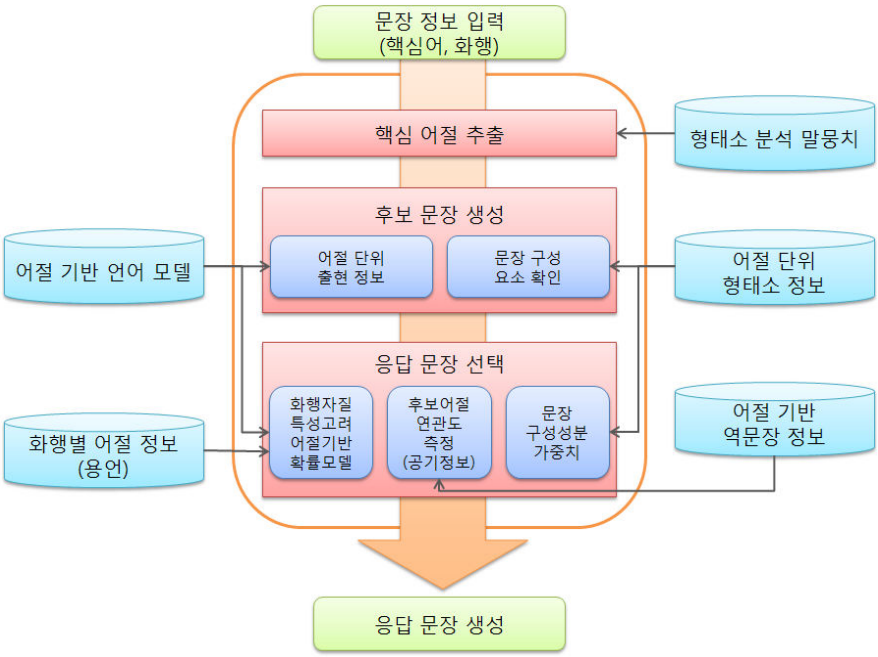

그림 1. 제안한 채팅 문장 생성 시스템의 구성도

1. 후보 문장 생성

입력된 핵심어를 포함하는 후보 문장을 생성하기 위해 형태소 분석 말뭉치를 이용한다. 형태소 분석 말뭉치에서 핵심어가 포함된 어절들을 추출하고, 그 어절을 기준으로 앞뒤에 나타나는 어절을 확장하며 후보 문장을 생성한다. 그러나 어절의 연속 출현 정보만을 이용하여 후보 문장을 생성할 경우 문장의 시작과 끝이 나타나지 않아 어절이 무한히 확장되는 문제가 발생할 수 있다. 또한 후보 문장이 너무 많이 생성되어 계산량이 폭발적으로 증가하는 문제가 발생할 수 있다. 이러한 문제들을 해결하기 위해 본 논문에서는 그림 2와 같이 어절의 위치 정보와, 문장 길이, 구성 성분을 이용한다.

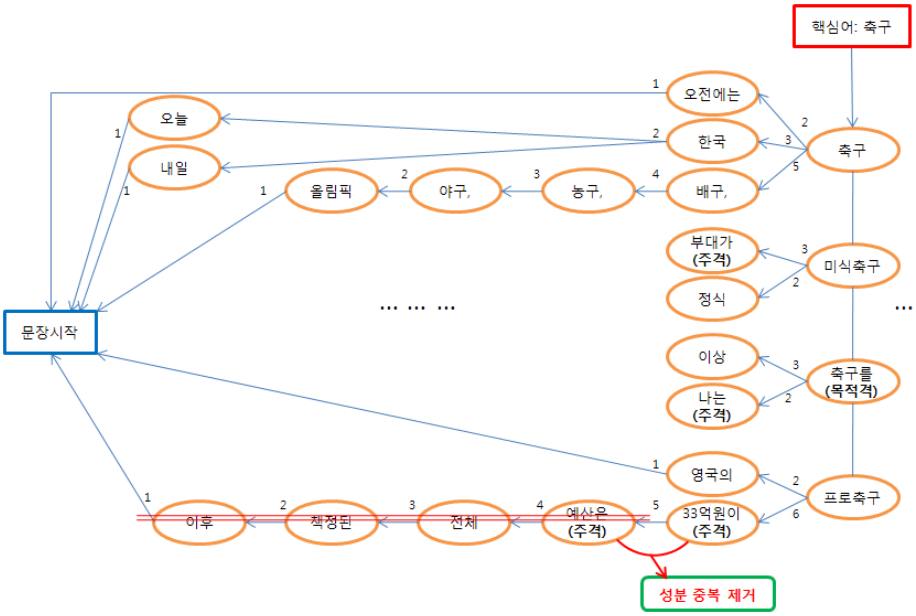

그림 2. 후보 문장 생성 예

그림 2에서 타원은 후보 문장을 구성하는 어절이며, 숫자는 문장에서 나타나는 순서의 위치 정보를 의미한다. 이와 같이 문장에서의 위치 정보에 적합한 어절만을 선택하기 때문에 문장 구성 후보 어절의 무한 확장을 막을 수 있다. 또 후보 문장의 수를 조절하기 위해 후보 문장 구성 성분, 문장의 길이, 동일 어절 비교 정보에 기초한 휴리스틱을 이용한다. 먼저 문장의 구성 성분을 확인하여 후보 문장을 제한하는 방법을 적용한다. 한국어 문장의 기본 형식에서 가장 많이 나타나는 주성분은 주어, 목적어, 서술어로 이 세 가지 요소는 문장의 가장 기본이 되는 구성 성분이다[8]. 표 1과 같이 각 성분의 판단은 조사를 이용하여 수행했으며, 학습 코퍼스의 분석 결과 세 가지 성분이 1회 이하로 조합되어 나타난 문장의 비율이 92.42%를 차지했다. 이에 따라 본 논문에서는 어절이 갖는 형태소 정보를 이용하여 후보 문장의 문장 구성 요소를 확인하고, 그림 2의 예와 같이 각 성분이 2회 이상으로 조합되어 나타나는 후보 문장을 제거한다. 그리고 학습 말뭉치에서 나타나는 문장 길이의 평균(6.00)과 표준 편차(1.97)를 식 (1)을 이용하여 신뢰구간 95%

로 책정하고 후보 문장의 최대 길이를 10으로 제한한다. 문장의 길이는 문장이 갖는 어절의 수를 의미하며 후보 문장 생성 과정에서 후보 어절의 수가 10개를 초과하면 해당 후보 문장을 제거한다. 마지막으로 어절이 중복되어 나타나는 경우 문장의 후보 문장을 제거하여 후보 문장의 수를 줄인다.

$$\text{신뢰구간}(95\%) : \mu \pm (1.96 \times \sigma) \tag{1}$$

표 1. 문장 구성 성분 판단

	주어	목적어	서술어
형태소	주격조사	목적격조사	종결어미

2. 응답 문장 선택

후보 문장의 생성이 끝나면 적합 문장의 선택을 위해 3가지의 통계 정보를 이용한다. 각 후보 문장 별로 계산된 통계 정보는 식 (2)의 가중치로 적용되어 최적합 응답 문장의 결정에 사용된다.

$$Response\ Sentence = \arg\max_{S_i}(\alpha \times P(S_i) + \beta \times R(S_i) + \gamma \times CW(S_i)) \tag{2}$$

식 (2)에서 S_i는 i번째 후보 문장을 의미하며, $P(S_i)$, $R(S_i)$, $CW(S_i)$의 값은 각각 언어 모델 통계 정보, 공기 정보를 이용한 연관도, 구성 성분 통계 정보를 나타낸다. α, β, γ 값은 각 요소의 중요도로써 실험적으로 $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$의 값을 가진다.

응답 문장의 선택을 위해 사용된 가중치의 의미는 다음과 같다. 먼저 어절 기반 언어 모델을 적용하여 통계 정보의 가중치를 계산한다. n개의 어절 E로 구성된 i번째 후보 문장 $S_i=E_1E_2...E_n$을 생성했을 때 문장 S_i가 나타날 확률 $P(S_i)$는 식 (3)과 같다.

$$P(S_i) = P(E_1, E_2, \ldots, E_{n-1}, E_n) \tag{3}$$

식 (3)에서 후보 문장 S_i가 학습 말뭉치에서 그대로 나타나는 것이 사실상 불가능하기 때문에 식 (3)을 연쇄규칙(chain rule)과 1차 마르코프 가정을 적용하여 식 (4)의 형태로 근사한다.

$$\begin{aligned} P(S_i) &= P(E_1, E_2, \cdots, E_{n-1}, E_n) \\ &= P(E_1)P(E_2|E_1)P(E_3|E_1, E_2)P(E_4|E_1, E_2, E_3)\cdots P(E_n|E_1, \cdots, E_{n-1}) \\ &= \prod_{j=1}^{n} P(E_j|E_1, \cdots, E_{j-1}) \\ &= \prod_{j=1}^{n} P(E_j|E_{j-1}) \end{aligned} \tag{4}$$

식 (4)에서 $P(E_j|E_{j-1})$는 어절 E_{j-1}가 나오고 어절 E_j가 나타날 확률로써 본 논문에서는 학습 말뭉치 추출된 출현 빈도를 이용하여 계산한다.

$$P(E_j|E_{j-1}) = \frac{freq(E_{j-1}E_j)}{freq(E_{j-1})} \tag{5}$$

그리고 응답 문장의 적합도를 향상시키기 위해 화행별 대표 용언 정보를 이용하여 식 (5)의 확률 값을 변경한다. 본 논문은 특정 영역에 국한되지 않는 말뭉치 기반이므로, 영역에 종속적이지 않은 용언을 화행의 자질로 사용한다[9]. 화행별 대표 용언은 식 (6)과 같은 카이 제곱 통계량으로 결정한다.

$$\chi^2(e, f) = \frac{N \times (AD - BC)^2}{(A+B) \times (A+C) \times (B+D) \times (C+D)} \tag{6}$$

식 (6)에서 A는 화행이 f인 문장 중 용언 e를 포함하고 있는 문장의 수, B는 화행이 f가 아닌 문장 중 용언 e를 포함하고 있는 문장의 수, C는 화행이 f인 문장 중 용언 e를 포함하지 않는 문장의 수, 그리고 D는 화행이 f가 아닌 문장 중 용언 e를 포함하지 않은 문장의 수이다. 그리고 N은 전체 문장의 수이다.

각 화행별로 얻어진 카이 제곱 통계량 값은 식 (7)에 의해 계산된 가장 큰 값이 해당 용언의 자질 값이 되며, 이 값을 순위화하여 자질 값이 높은 순으로 정렬한

다. 그리고 카이 제곱 통계량에 의해 얻어진 순위를 바탕으로 정해진 순위 안에 포함된 자질들만을 화행의 특징을 나타내는 어휘로 판단하여 대표 용언으로 사용한다.

$$\chi^2_{\max}(e) = \max_{i=1}^{m} \chi^2(e, f_i) \tag{7}$$

대표 용언의 선택이 끝나면 해당 용언을 포함하는 어절의 가중치를 식 (8)과 같이 변경한다.

$$P(E_j|E_{j-1}) = \frac{freq(E_{j-1}E_j)}{freq(E_{j-1})} \approx \frac{k \times freq(E_{j-1}E_j)}{freq(E_{j-1})} \quad \begin{cases} k=2, E_j \in F \\ k=1, E_j \notin F \end{cases} \tag{8}$$

식 (8)에서 F는 화행 f에 특징적으로 나타나는 대표 용언의 집합을 의미한다. $k=2$는 실험적으로 결정된 대표 용언에 대한 가중치이다. k가 3이상의 값을 가질 경우 후보 문장 내의 다른 어절과의 연관성보다 대표 용언의 값에 집중되는 현상으로 어색한 문장이 만들어 지는 빈도가 높아진다.

그러나 식 (4)에서 후보 문장의 길이에 따라 연산 횟수에 차이를 보이므로 문장의 길이가 다른 후보 문장 간에는 공정한 계산 수치의 비교가 이루어 질 수 없다는 문제점이 있다. 이러한 문제를 해결하기 위해서 식 (9)과 같이 후보 문장의 길이에 대한 정규화를 수행한다.

$$P(S_i) = \left\{ \prod_{j=1}^{n} P(E_j|E_{j-1}) \right\}^{\frac{1}{n}} = \sqrt[n]{\prod_{j=1}^{n} P(E_j|E_{j-1})} \tag{9}$$

식 (9)는 후보 문장 길이의 차이에 따른 연산 횟수 차이 문제를 해결하고, 상대적으로 문장의 길이가 짧은 문장에 높은 가중치를 반영하는 특성을 포함하고 있기

때문에 효과적인 계산이 가능하다.

문장별 언어 모델 통계 정보의 가중치 계산이 끝나면, 후보 문장의 적합성 판단을 위해 문장 내 어절간 연관도를 측정한다. 학습 말뭉치 내 동일 문장으로부터 생성된 어절이 많을수록 그렇지 않은 것보다 높은 가중치를 부여하기 위함이다. 본 논문에서는 후보 문장 어절들이 학습 말뭉치에서 같이 나타난 문장의 공기 정보를 이용한다. 후보 문장 S_i의 각 어절들의 대한 연관도는 식 (10)을 이용하여 측정한다.

$$R(S_i) = \prod_{j=1}^{n} \frac{\sum_{k=1}^{n} |E_j \cap E_k|}{|E_j|}, j \neq k \tag{10}$$

식 (10)에서 S_i는 n개의 어절로 구성된 i번째의 후보 문장이며, E는 후보 문장에 포함된 어절을 의미한다. $|E_j|$와 $|E_k|$는 각각 어절 E_j와 E_k가 나타난 문장의 수, $|E_j \cap E_k|$는 어절 E_j와 E_k가 동시에 나타난 문장의 수를 의미한다. 이들 후보 문장 내 어절 간의 공기 정보를 이용하여 후보 문장별 연관도에 대한 가중치를 계산한다.

연관도 측정이 끝나면 문장 구성 성분 조합에 따른 가중치를 적용한다. 후보 문

표 2. 문장 구성 성분 조합에 따른 가중치

(O: 포함, X: 미포함)

주어	목적어	서술어	CW(Si)
X	X	X	0.16
O	X	X	0.07
X	O	X	0.06
X	X	O	0.34
O	O	X	0.02
O	X	O	0.20
X	O	O	0.10
O	O	O	0.05

장의 어절이 갖는 형태소 정보를 이용하여 후보 문장의 문장 구성 성분을 분석하고, 후보 문장 생성 판단에 사용된 세 가지 문장 구성 성분, 주어, 목적어, 서술어가 후보 문장에서 어떤 조합으로 나타났는지 확인한다. 그리고 각 후보 문장의 주요 구성 성분의 조합에 따라 학습 말뭉치에서 통계적으로 얻은 정보를 이용한 표 2를 기준으로 가중치를 부여한다.

후보 문장에 수식을 적용한 과정을 예를 들어 설명하면 그림 3과 같다.

S	정장에	스타일리쉬한	넥타이지.
출현빈도(E_j)	6	3	1
출현빈도($E_{j-1}E_j$)	4	1	1
형태소	-	-	-
공기정보	1	1	1
입력: 핵심어 – 정장, 화행 – Expressive			

$$P(S) = \left(\frac{4}{6} \times \frac{1}{6} \times \frac{1}{3}\right)^{\frac{1}{3}} = 0.3333$$

$$R(S) = \left(\frac{1}{6} \times \frac{1}{3} \times \frac{1}{1}\right) = 0.0556$$

$$CW\ (S) = 0.16$$

$$Weight\ (S) = 0.4 \times 0.3333\ + 0.4 \times 0.0556\ + 0.2 \times 0.16\ = 0.1876$$

그림 3. 후보 문장 수식 적용 예

그림 3에서는 핵심어 '정장'과 화행 'Expressive'의 입력으로 생성된 후보 문장 *S*의 예이다. 먼저 통계 정보 *P(S)*는 '정장에 스타일리쉬한'과 '스타일리쉬한 넥타이지.'의 출현빈도와 바이그램(bigram) 출현빈도를 이용하여 구한다. 이때 화행 'Expressive'의 대표 용언이 있는지 확인하여 가중치를 적용하지만 후보 문장 *S*에는 포함되어 있지 않기 때문에 적용되지 않는다. 통계 정보 계산이 끝나면 공기 정보를 이용하여 연관도 *R(S)*를 구한다. 그림 3의 공기정보는 각 어절과 나머지 어절이 함께 나타난 문장의 수를 의미한다. 연관도 계산이 끝나면 마지막으로 형태소 정보를 이용하여 구성 성분을 파악하고 구성 성분 조합에 따라 가중치를 부여받는다. 후보 문장 *S*는 본 논문에서 판단하는 주어, 목적어, 서술어 성분이 포함되어

있지 않기 때문에 그에 따른 가중치 0.16을 부여 받는다. 세 가지 가중치 계산이 끝나면 각 정보별 가중치를 반영하여 문장의 적합도를 나타내는 수치를 얻는다. 후보 문장 S의 최종 수치는 0.1876이며 최적합 문장을 결정하기 위한 다른 모든 문장과의 비교를 위해 사용된다.

Ⅳ. 실험 및 평가

어절의 통계 정보와 공기 정보, 문장 구성 성분 정보를 얻기 위해서 대화형 언어 말뭉치 43,351문장(260,362어절)을 자체 수집하고 표 3과 같이 9개의 화행을 부착하여 학습 말뭉치[1]로 사용했다. 또 문장 생성 실험을 위한 입력으로 학습 말뭉치에서 나타난 화행 분포에 따라 임의로 추출한 명사 100개를 핵심어로 사용하였다. 표 3은 본 논문에서 사용한 화행의 분류와 말뭉치에서의 분포를 보여준다.

본 논문에서 제안한 한국어 채팅 문장 생성 시스템은 보다 완성도 높은 응답 문장을 생성하는 것을 목표로 한다. 이러한 문장 생성 시스템은 특정 수치로 나타낼 수 있는 측정 방법이나 다른 시스템과의 비교가 매우 어렵다. 따라서 본 논문

표 3. 화행 분류

화행	구분	비율	화행	구분	비율
Describe	일반적 사실이나 경험 표현	48.12%	Ask_WH	질문, 질의(WH)	9.81%
Expressive	감정, 선호	8.95%	Ask_Confirm	확인	19.26%
Greet-Bye	인사	0.04%	Hope	희망	3.39%
Guess	추측	6.76%	Other	그 외, 복합적 의미나 구분 모호	0.65%
Request	명령, 행동	3.04%			

1) 본 논문에 사용된 학습말뭉치는 2007년도 KT “대화형 언어코퍼스 구축” 과제의 지원으로 구축된 것입니다.

표 4. 생성 문장 완성도 구분

문장 완성도	의미	예
3	Reasonable	댄스스포츠와 라틴댄스의 차이점을 알고 있니?
2	Grammatically	예산은 어느 그룹 사장이지?
1	Incorrect	운동회를 활동하기 좋은 날씨여서 가을에 헤어지는 것 같아.

은 시스템의 성능을 평가하기 위해 표 4와 같이 생성 문장의 완성도 수준을 3단계로 구분하여 평가했다[6].

표 4에서 3단계 Reasonable은 발화 문장으로서 적합한 문장, 2단계 Grammatically는 문법적으로는 맞지만 주어와 서술어와의 의미 관계 등이 올바르지 않은 문장, 1단계 Incorrect는 부적합한 문장을 의미한다.

실험은 본 논문에서 제안한 방법으로 생성한 문장과 동일 입력으로 일반 언어 모델을 적용하여 생성한 문장을 이용했다. 일반 언어 모델은 비교 평가를 위한 기저모델(baseline model)로 바이그램 통계 수치를 적용한 언어 모델이며, 후보 문장 생성에 사용된 휴리스틱은 동일하게 적용했다. 그리고 생성된 문장이 말뭉치에서 그대로 나타난 문장인 경우 차순위의 적합 문장을 실험에 사용했다. 자연어 처리를 전공한 석사과정 학생 5명이 명사와 화행의 쌍으로 이루어진 입력으로부터 생성된 문장에 대한 완성도 평가를 수행했으며 평가 결과는 표 5와 같으며, 표 6은 모델별로 생성된 문장의 예를 보여준다.

평가를 위해 식 (11)과 같은 적합 문장 생성률을 사용하였다.

$$\text{적합 문장 생성률} = \frac{Reasonable\text{로 평가된 문장 수}}{\text{전체 평가 문장 수}} \tag{11}$$

평가 결과 기저모델인 바이그램 언어 모델은 평균 61%의 적합 문장 생성률을 보였고, 화행 자질의 특성을 고려한 언어 모델은 평균 71.6%, 본 논문에서 제안한 방법은 이보다 높은 평균 86.2%의 적합 문장 생성률을 보였다. 이는 문맥 자질로 사용된 화행 자질과 어절간 연관도 측정을 위한 공기 정보, 그리고 구성 성분 정보가 적합 문장 생성에 긍정적으로 작용했음을 알 수 있다.

표 5. 생성 문장 평가

모델	문장완성도	A	B	C	D	E	평균
바이그램 언어 모델	3	63	59	61	61	61	61
	2	13	20	12	10	21	15.2
	1	24	21	27	29	18	23.8
바이그램 언어 모델 (화행고려)	3	72	76	65	64	81	71.6
	2	8	16	14	12	18	13.6
	1	20	8	21	24	1	14.8
제안 시스템	3	93	89	80	78	91	86.2
	2	2	11	15	11	5	8.8
	1	5	0	5	11	4	5

A, B, C, D, E는 평가자이며, 표 안에 있는 수치는 평가된 문장의 수를 의미함

표 6. 생성 문장 예

핵심어	생성 문장		
	바이그램 언어 모델	바이그램 언어 모델(화행고려)	제안 시스템
소파	소파는 정말 멋진 가슴 근육을 키워준다며?	소파는 정말 대단한 것 같아.	소파는 정말 대단한 것 같아.
무좀	무좀이 생기면 각막염을 의심해야 해.	무좀이 생기면 각막염을 의심해야 해.	무좀이 생기면 어떻게 해야 해.
루즈벨트	루즈벨트는 민주당에서 4선까지 한 거야?	루즈벨트는 민주당에서 4선까지 한 거래.	루즈벨트는 민주당에서 4선까지 한 거야?
연장전	연장전까지 이어지는 인간애가 느껴지는 글이다.	연장전까지 이어지는 인간애가 느껴지는 글이다.	연장전까지 이어지는 축구경기를 뛸 생각이야.
정장	난방용품을 전부 정장만 입나봐.	정장은 5벌 있어.	정장에 스타일리쉬한 넥타이지.
미인	용기없으면 미인을 얻나?	용기없으면 미인을 얻나?	용기있는 자가 미인을 쟁취한다.

표 5의 생성 문장 평가 결과를 보면 평가자마다 실험 결과에 차이를 보이고 있다. 이러한 차이의 이유는 같은 문장이라 하더라도 개인의 평가가 다르기 때문에 나타나는 현상이다. 평가 결과에 차이를 보이는 실험 문장의 예는 표 7과 같다.

표 7. 평가 차이 예

문장	평가자 D	평가자 E
애완용으로 기르는 남성을 어떻게 생각해?	1	3
싸이에 배경음악으로 뭐가 있을까?	2	3

표 7과 같이 문장 완성도에 대한 개인의 수긍 정도에 차이가 존재하며, 그 결과가 수치상에 나타나는 것으로 보인다. 이는 문장 생성 시스템의 기준이 될 만한 평가 방법이 없기 때문에 발생하는 문제로 향후 연구에서는 더 평가를 위한 명확한 방법이 필요할 것으로 생각된다.

V. 결 론

본 논문에서는 어절 기반의 한국어 채팅 문장 생성 방법을 제안하였다. 입력된 핵심어가 포함된 어절을 말뭉치에서 찾아 해당 어절의 앞뒤에 나타나는 어절의 출현 정보를 이용하여 후보 문장들을 생성한다. 그리고 각 후보 문장의 통계 정보와 화행에 따른 자질을 고려한 언어 모델, 후보 문장 내 어절 간의 공기 정보를 이용한 연관도, 문장 구성 성분 정보에 따른 가중치를 통합하여 응답 문장을 선택한다. 어순이 자유로운 한국어의 특성으로 인해 구문 분석과 같이 높은 수준의 자연어 처리 정보 없이도 적합 문장을 얻을 수 있었다. 실험 결과에 따르면 제안한 방법은 평균 86.2%의 적합 문장 생성률을 보였다. 향후 연구로는 문장의 완성도를 보다 높일 수 있는 추가 자질에 관한 연구가 필요하며, 말뭉치에 나타나지 않은 단어로부터 문장을 생성하는 모델 개발이 필요하다.

참고문헌

[1] Joseph Weizenbaum, “ELIZA-A Computer Program For the Study of Natural Language Communication Between Man can Machine”, Communications of the ACM, Vol.9, No.1, pp. 36-45, 1996.

[2] Michael L. Mauldin, “ChatterBots, TinyMuds, and the Turing test: entering the Loebner Prize competition”, Proceedings of the twelfth national conference on Artificial intelligence, Vol.1, pp. 16-21, 1994.

[3] Robert P. Schumaker, Ying Liu, Mark Ginsburg, Hsinchun Chen, "Evaluating mass knowledge acquisition using the ALICE chatterbot: the AZ-ALICE dialog system", International Journal of Human-Computer Studies, Vol. 64, No 11, pp. 1132-1140, 2006.

[4] Jason L. Hutchens, Michael D. Alder, “Introducing MegaHAL”, NeMLaP3 / CoNLL98 Workshop on Human-Computer Conversation, ACL, pp. 271-274, 1998.

[5] Dana Vrajitoru, “Evolutionary Sentence Building for Chatterbots”, GECCO, 2003.

[6] Dana Vrajitoru, Jacob Ratkiewicz, “Evolutionary Sentence Combination for Chatterbots”, AIA, 2004.

[7] 박훈민, 대화 시스템을 위한 CRF와 Active Learning 기반의 효율적 의미 구조 분석 (석사학위논문, 서강대학교, 한국, 2006)

[8] 김혜숙, “한국어 기본 문형 설정에 대하여”, 국어국문학회, 국어국문학 제122권, pp. 13-47, 1998.

[9] 김민정, 한경수, 박재현, 송영인, 임해창, “도메인에 비종속적인 대화에서의 화행 분류”, 한국정보과학회 언어공학연구회 학술발표 논문집, pp. 246-253. 2006.

[10] 이일주, 김민구, “단어의 공기정보를 이용한 클러스터 기반 다중문서 요약”, 정보과학회논문지, 제33권 제2호, 2006.

[11] 손강민, 손승범, 강태근, 문애경, 정인철, 김현, 함호상, “디지털 생명체 연구: 채팅 로봇(Chatterbot) 기술”, 정보통신연구진흥원 학술정보, 주간기술동향 1115호, 2003.

[12] 이강천, 서정연, “의미 중심어에 기반한 한국어 문장 생성 시스템”, 정보과학회논문지(C), 제4권 제5호, pp. 718-727. 1998.

1 차원고접수 : 2009. 9. 13
2 차원고접수 : 2009. 10. 27
최종게재승인 : 2009. 12. 17

(*Abstract*)

Statistical Generation of Korean Chatting Sentences Using Multiple Feature Information

JongHwan Kim
Kangwon National University

Du-Seong Chang
KT Central R&D Group

Harksoo Kim
Kangwon National University

A chatting system is a computer program that simulates conversations between a human and a computer using natural language. In this paper, we propose a statistical model to generate natural chatting sentences when keywords and speech acts are input. The proposed model first finds Eojeols (Korean spacing units) including input keywords from a corpus, and generate sentence candidates by using appearance information and syntactic information of Eojeols surrounding the found Eojeols. Then, the proposed model selects one among the sentence candidates by using a language model based on speech act information, co-occurrence information between Eojeols, and syntactic information of each Eojeol. In the experiment, the proposed model showed the better correct sentence generation rate of 86.2% than a previous conventional model based on a simple language model.

Keywords : Chatting sentence, statistical sentence generation, chatting system