한국음향학회지 제28권 제1호 pp. 61-69 (2009)

다양한 언어적 자질을 고려한 발화간 유사도 측정 방법

A Method for Measuring Inter-Utterance Similarity Considering Various Linguistic Features

이 연 수*, 신 중 휘**, 홍 금 원*, 송 영 인*, 이 도 길***, 임 해 창*
(Yeon-Su Lee*, Joong-Hwi Shin**, Gumwon Hong*, Young-In Song*, Do-Gil Lee***, Hae-Chang Rim**)

*고려대학교 컴퓨터·전파통신공학과, **고려대학교 컴퓨터학과, ***고려대학교 민족문화연구원
(접수일자: 2008년 10월 10일; 채택일자: 2008년 1월 7일)

본 연구는 예제 기반 대화 시스템에서 응답을 결정하기 위한 핵심 요소 기술 중 하나인 발화간 유사도 측정 방법의 개선에 대해 논한다. 일반적인 문장간 유사도 측정과는 달리, 대화에서 발화간 유사도 측정은 단어 분포간 유사도 뿐만 아니라, 문형, 시제, 긍/부정, 양태등 대화의 자연스러움을 결정하는 문장의 다양한 언어적 요소 역시 중요하게 고려되어야 한다. 그러나 기존 연구에서는 이에 대한 고려가 부족 했던 것이 사실이며, 따라서 본 연구에서는 개선 방안으로서 발화의 형태적 유사성 뿐 아니라 다양한 언어적 자질들을 분석하고 이를 유사도 측정에 반영하여 정확도를 향상시키는 새로운 유사도 측정 방법을 제안한다. 또한, 발화의 자질별 유사도를 고려함으로써, 한정된 수의 예제들의 활용도를 높일 수 있는 방법을 제안하였다. 실험 결과 제안하는 방법이 기존 방식에 비해 10%p 이상의 정확도 성능 향상이 있었다.
핵심용어: 예제 기반 대화 시스템, 발화간 유사도 측정, 챗봇, 자연어 처리
투고분야: 음성처리 분야 (2.7)

This paper presents an improved method measuring inter-utterance similarity in an example-based dialogue system, which searches the most similar utterance in a dialogue database to generate a response to a given user utterance. Unlike general inter sentence similarity measures, the inter utterance similarity measure for example-based dialogue system should consider not only word distribution but also various linguistic features, such as affirmation/negation, tense, modality, sentence type, which affects the natural conversation. However, previous approaches do not sufficiently reflect these features. This paper proposes a new utterance similarity measure by analyzing and reflecting various linguistic features to improve performance in accuracy. Also, by considering substitutability of the features, the proposed method can utilize limited number of examples. Experimental results show that the proposed method achieves 10%p improvement in accuracy compared to the previous method.
Keywords: Example based, Dialogue system, Utterance similarity measure, Chatbot, Natural language processing
ASK subject classification: Speech Signal Processing (2.7)

I. 서 론

사람과 컴퓨터가 자연어 형태로 대화하며 정보를 주고 받는 대화 시스템 (dialogue system) 기술은 최근 음성인식, 음성 합성 기술과 접목되어 인터넷, 휴대 전화, 로봇 등에서 각광을 받고 있다. 컴퓨터와의 대화를 위해서는 컴퓨터가 사용자의 발화를 분석하여 응답을 생성할 수 있어야 한다. 최근에는 대량의 발화-응답 쌍으로 구성된 대화 예제를 사용자 발화에 대한 응답 생성에 활용하는 예제 기반 (example-based) 등의 방법이 많이 연구되고 있다 [1][2]. 예제 기반 방법에서는 입력된 사용자 발화에 대해서 가장 유사한 예제 발화를 탐색하여 해당 예제 발화의 응답을 시스템 발화로 활용한다. 이러한 방식은 자연어 발화에 대한 복잡한 언어 분석이나 대규모 언어 자원을 요구하지 않아 다양한 주제로 일상적인 대화를 처리해야하는 채팅 시스템 등의 분야에 강점이 있다.

이러한 예제 기반 방법에 기반하여 대화 시스템을 구축하기 위해서는 효과적 유사 발화 검색 방법, 특히 사용자 발화와 예제 발화간 유사도 측정 방법의 개발이 필수적이

책임저자: 임 해 창 (rim@korea.ac.kr)
136-713 서울시 성북구 안암동 5-1 고려대학교 컴퓨터·전파통신공학과
(전화: 02-924-2054; 팩스: 02-929-7914)

다. 예를 들어 사용자가 입력한 "*뭘 먹을까?*"와 정확히 일치하는 예제가 예제 데이터베이스에 존재다면 그에 대응하는 응답을 바로 출력하면 된다. 그러나 정확히 일치하는 것이 없다면 가장 유사한 예제 발화를 찾아야 한다. 만약 예제 데이터베이스에 사용자 입력 발화 "뭘 먹을까"와 의미적으로 동일한 예제 발화 쌍은 없지만 유사한 발화를 포함한 예제 쌍 〈발화:"*나 뭐 먹지?*, 응답:"*오늘은 칼국수 어때*"〉이 존재한다면, 해당 예제 쌍의 응답 발화를 시스템의 출력으로 사용하는 것이 바람직할 것이다. 이를 가능케 하기 위해서는 저장된 발화들과 사용자 발화와의 유사도를 측정할 수 있는 방법이 필요하다.

이를 위해서는 다음과 같은 사항들이 요구된다. 첫째, 발화는 일반적인 문장과는 다르게 상대방과의 대화를 전제로 한다. 따라서 발화를 통해 상대방에 대한 명령이나 요청, 질문, 칭찬, 비난 등 다양한 양태가 나타나고, 시제는 과거, 현재, 미래가 되기도 한다. 응답 역시 이에 맞게 이루어져야 자연스러운 대화가 된다. 그러므로 정확한 발화간 유사도 측정을 위해서는 일반 문장간의 유사도 측정과는 다른 접근법이 필요하다. 둘째, 예제 기반 시스템에서는 예제가 부족하거나 정확히 일치하지 않는 경우가 많다. 따라서 정확히 일치하지 않는 발화 예제라 할지라도 활용이 가능하다면 최대한 이용하는 전략이 필요하다. 즉 필요하다면 유사도의 기준을 완화하고 재현율을 높여 대화가 이어질 수 있어야 한다. 본 연구에서는 이러한 문제를 해결할 수 있는 방법을 제안한다.

논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련한 기존 연구를 살펴보고 3장에서는 발화간 유사성을 측정하기 위해 분석해야 할 정보 및 유사도 측정 방법에 대해 설명한다. 4장은 제안하는 방법의 성능에 대한 실험 및 평가를 보여 준다. 5장에서는 본 연구의 결론과 향후 연구에 대해 논하도록 한다.

II. 기존 연구

발화간의 유사도 측정을 위한 기존 연구는 크게 두 가지로 분류할 수 있다. 발화의 형태 정보만을 이용하여 유사도를 측정하는 방법과 발화의 의미 정보까지 고려하여 유사도를 측정하는 방법이다.

발화의 형태 정보만을 이용하여 유사도를 측정하는 방법은 코사인 유사도 (cosine similarity) [3]와 같이 어휘의 통계 정보를 사용하는 방법, 음성 인식 분야에서의 WER (Word Error Rate) [4]을 이용하는 방법, 기계 번역의 자동 평가 척도인 BLEU [5]와 같이 공통 n-gram을 사용하는 방법 등이 있다. 코사인 유사도를 발화간 유사도 측정에 사용하면 입력 발화와 예제 발화는 각각 단어 공간의 벡터로서 표현되고, 두 벡터간의 코사인 값이 유사도로 정의된다. 따라서 발화를 구분하는 주요 단어의 개수가 얼마나 일치하는가가 유사도를 결정한다. 편집거리를 단어 수준에서 적용한 WER은 두 문장에서의 입력, 대체, 삭제된 단어의 비율로서 유사도를 측정한다. 이것을 단어의 위치를 고려하지 않고 계산하는 방식이 PER (Position-independent word Error Rate) [6]로서 기계 번역의 자동 평가 척도로 사용되었다. 자동 평가는 기계가 번역한 문장과 사람이 번역한 문장간의 유사성을 측정하여 시스템을 평가하기 때문에 발화간 유사도 측정 방법으로 볼 수 있다. 또 다른 척도인 BLEU는 두 문장간에 매칭되는 n-gram의 수를 세어 정확률 (precision)을 측정하며 기계 번역 평가 부분에서 사실상의 표준으로 인정된다.

그러나 앞의 방식들은 모두 단어 혹은 단어 n-gram에 기반하여 발화간의 형태적 유사성 만을 고려하는 방법이다. I장에서 언급했듯이 대화에서는 응답을 위해 양태나 문형, 시제 등이 중요하며, 이러한 정보는 단순히 단어 정보에 기반한 유사도 측정으로는 반영하기 어렵다. 예를 들어, 아래 그림 1의 발화들은 동일한 내용어 "비/NNG (일반명사) 오/VV (동사)"를 갖고 있지만 시제가 다르며 이로 인해 세 발화에 대한 응답 역시 각각 다르다. 이 경우, 형태적인 유사성만으로 유사 발화를 선택하는 기존 방법은 부적절한 결과를 낳을 수 있다. 예컨대 아래 그림 2에서 사용자 발화 "**많이 먹었다**"에 대해서 다른 유사한 발화들이 있음에도 불구하고 형태적 유사성만을 고려한 경우 "**많이** 못 **먹었어**"를 유사 발화로 선택하여 "다이어트 하니?" 라는 자연스럽지 못한 응답을 생성한 결과를 볼 수 있다.

이러한 점을 보완하여 보다 정확하게 유사 발화를 검색

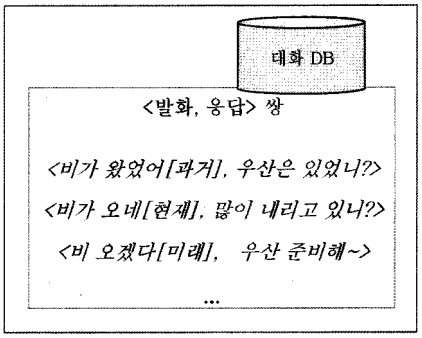

그림 1. 형태적으로는 유사하나 의미가 다른 문장의 예
Fig. 1. Example of utterances which is lexically similar, but semantically different.

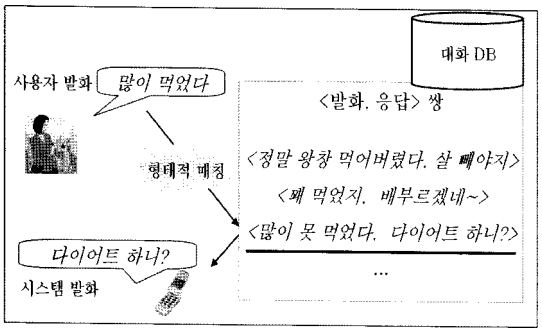

그림 2. 형태 정보만을 사용하는 유사도 측정 방법의 한계
Fig. 2. Limitation of measuring only lexical similarity.

하고자 최근의 예제 기반 대화 시스템에서는 형태 정보 외에 화행 (dialog act)과 같은 의미 정보를 유사도 측정의 주요 자질로 사용한다 [7][8]. [7]에서는 발화의 유사도를 측정하기 위해 키워드 (동사와 명사) 및 화행을 사용하거나 구문 구조의 거리 (structural distance)를 이용하는 방식을 제안하였다. 이와 유사하게 [8]에서는 화행을 이용하여 매칭한 후 그 결과를 편집거리에 의해 순서화하도록 하였다. 그러나 이들 방식은 첫째 사용된 자질 (키워드, 화행, 구문구조)만으로는 정확한 유사 발화 검색에 한계가 있다는 점, 둘째 단계적으로 자질이 다른 예제들을 제거하는 방식은 초기 단계에서 보다 유사한 발화가 제거 될 위험이 있다는 점, 셋째 자질값이 일치하지 않는 경우 얼마나 다른 지에 대한 정도가 고려되지 않은 점, 넷째 유사도에 하한선이 없어 유사하지 않는 발화도 최종적으로 선택될 수 있다는 점 등의 문제가 발생할 수 있다. 첫 번째 문제에 대해서는 화행 뿐 아니라 강/부정, 문형, 시제 등의 차이로 인해 의미가 크게 차이가 날 수 있다는 것을 앞의 그림 1, 그림 2를 통해 살펴보았다. 두 번째 문제는 아래 그림 3의 예로 설명이 가능하다. 사용자가

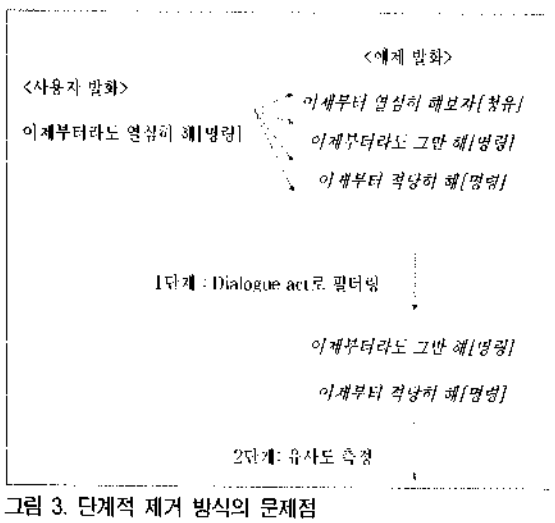

그림 3. 단계적 제거 방식의 문제점
Fig. 3. Problem of multi-step elimination of examples.

"이제부터라도 해 [명령]"라는 말을 한 경우 이 발화의 화행은 [명령]이다. 따라서 화행이 [청유]인 "이제부터 열심히 해보자"라는 문장은 화행이 달라서 제거된다. 결국 나머지 문장들 중에서 편집거리가 가장 가까운 "이제부터 그만 해 [명령]"가 가장 유사한 문장으로 선택된다.

세 번째 문제는 그림 4의 예로 설명할 수 있다. 이 예에서 발화들의 화행은 모두 다르지만 "*나 나가려고* [의지]"는 "*나 나가야겠다* [의무]", "*나 나간다* [서술]"와는 유사한 용도로 사용될 수 있는 반면에, 또 다른 발화 문장 "*나 나갈까?* [의문]"와는 의미적으로 상당히 큰 차이를 보인다. 따라서 사용자 발화로 "나 나가려고"가 입력되었을 때 동일한 화행의 발화가 없다면 "*나 나간다*"보다는 "*나 나간다*"를 유사 발화로서 선택하는 것이 보다 적절하나, 기존 방법으로는 이러한 발화간 의미 차이의 정도를 구분할 수 없다.

네 번째 문제는 사용자가 "나 학교 못 간다"를 입력하였는데 "나 학교 가고 싶어", "나 학교 간다", "나 학교 갔다" 등 전혀 유사하지 않은 발화만 대화 DB에 존재하는 경우이다. 이 때 어느 것을 선택해도 응답은 적절치 못할 것이다. 따라서 시스템은 이 상황을 인지하여 '재질의' 등 다른 대응을 할 수 있어야 한다. 즉, 유사도의 한계를 정할 수 있어야 한다.

예제 기반 대화 시스템은 대량의 예제를 가정하지만 이를 확보하는 것은 쉽지 않다. 일부 주제에 대해서는 예제가 많지만 다른 주제에 대해서는 예제가 부족할 수도 있다. 따라서 가장 유사한 것을 정확하게 선택하되 예제가 부족할 때에도 대화를 자연스럽게 이어갈 수 있어야 한다. 그러나 기존 연구에서는 앞에서 지적한 문제점들로 인하여 이러한 요구 사항을 충족시키기에는 부족함이 있었던 것이 사실이다. 그러나 본 연구에서는 형태 정보 외에 다양한 의미 정보를 사용하여 정확률을 높이는 한편 모든 자질을 동시에 고려하고 자질간에도 유사도를 고려하여 예제 부족 문제에도 견고할 수 있는 효과적인 방법을 제안한다.

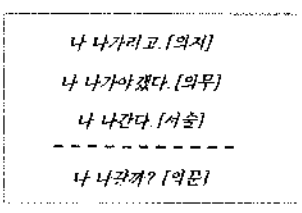

그림 4. 자질값이 일치하지 않더라도 유사한 발화들의 예
Fig. 4. Example of similar utterances expressing different feature values.

III. 유사도 측정 방법

본 장에서는 발화의 형태적·의미적 자질을 종합하여 발화간 유사도를 측정하는 방법과 말뭉치를 통해 자질간 유사도값을 어떻게 결정하는가에 관하여 논한다. 제안하는 방법이 기존의 연구와 다른 점은 다음과 같다. 첫째 발화의 형태적 자질과 의미적 자질을 모두 고려한다. 둘째 모든 자질들을 동시에 고려한다. 셋째 유사도에 영향을 미치는 각 자질들이 두 발화에서 일치하지 않더라도 각 자질간에 유사한 정도를 측정하여 전체 유사도에 반영한다.

3.1. 발화의 분석

본 연구에서는 발화간 유사도 측정을 위해 하나의 발화를 형태적, 통사-의미적 차원에서 분석하여 표현한 후 이를 비교한다. 여기에는 다음과 같은 두 가지 가정이 전제되어 있다. 첫 번째는 발화의 의미를 여러 가지 자질로서 분석하여 표현할 수 있다는 것이다. 다른 하나의 가정은 이러한 자질이 두 발화에서 일치하는 비율이 높을수록 두 발화가 유사하다는 것이다. 다음 표 1는 본 논문에서 발화의 의미를 분석하기 위해 고려한 자질들이다.

이러한 자질들을 분석하기 위해서는 먼저 사용자가 입력한 발화에 대해 띄어쓰기, 철자 교정 등의 오류 교정을 수행한 후 형태소 분석 및 품사 부착을 수행한다. 그 후 술어부를 추출하여 시제 긍/부정을 결정하는 규칙이 나타나는지를 살펴본다. 예를 들어 "같이 영화 봤었지"의 경우는 "같이/MAG (부사) 영화/NNG (일반명사) 보/VV (동사)+았었/EP (선어말어미)+지/EF (종결어미)"와 같이 형태소 분석 및 품사 부착을 수행한 후 "았었/EP (선어말어미)"이 나타남에 따라 〈과거〉시제로 분류한다.

표 1. 발화의 의미를 분석하기 위해 고려한 자질
Table 1. Features for the analysis of an utterance.

분석 수준	자질	설명
형태적 요소	형태소/품사열	발화의 형태소/품사열
통사-의미적 요소	문형	평서형, 의문형, 명령형, 청유형 등 문장 종결형
	시제	과거, 현재, 미래
	긍/부정	긍정문, 부정문
	문형-시제-긍/부정의 조합	문형-시제-긍/부정의 조합
	양태	서술, 지각, 추측 등 화자의 의도

3.2. 유사도 척도

본 연구에서 제안하는 유사도 측정식은 식 (1)과 같다. 사용자 입력 발화 S_1와 후보 발화 S_2간의 유사도 Sim는 각 자질의 유사도 (자질 별 유사도 함수) f_i의 선형합으로 측정된다. 또한 각 자질의 유사도 함수의 값은 그 중요도에 따라 w_i로 가중된다. 여기서 각 자질의 유사도 f_i는 각 발화의 i 번째 자질 (1번 째:문형 등)간 유사도를 나타내는 함수이다. 사용된 자질은 문형, 시제, 긍부정, 문형-시제-긍/부정 조합, 양태 그리고 품사를 고려한 어휘 유사도이다.

$$Sim(S_1,S_2)=\sum_i w_i f_i(e_{1,i},e_{2,i}) \quad (1)$$

- S_1 : 입력 발화
- S_2 : 후보 발화
- $e_{1,i}$: 입력 발화의 i번째 의미 자질 값
- $e_{2,i}$: 후보 발화의 i번째 의미 자질 값
- f_i : i번째 자질에 대한 유사도 함수
- w_i : f_i의가중치

표 2에서처럼 본 유사도 측정 방법은 형태적 자질과 의미적 자질을 모두 동시에 고려한다. 예를 들어 f_1 (문형 유사도 자질 함수)은 입력된 사용자 발화의 문형과 후보 발화의 문형간 유사도를 출력하는 함수로서 모든 문형 (평서형, 의문형, 명령형, 청유형)의 조합 ${}_4C_2=6$개에 대해 각각이 발화간 유사도에 미치는 정도를 통계적으로 측정하여 $0\sim1$ 사이의 값으로 대응시킨 것이다. 이것은 동일한 의문형이라도 완전히 같다고 할 수 없으며, 의문형과 평서형이라도 유사성이 있을 수 있다는 점, 그리고 그 정도는 연속성을 갖는다는 점을 반영한 함수이다. 그림 5는 자질 함수의 의미를 보여준다. 여기서 함숫값은

표 2. 자질 별 유사도 함수
Table 2. Similarity functions for each feature.

유사도 함수	설명
f_1: 문형 유사도 자질	입력된 사용자 발화의 문형과 후보 발화의 문형간 유사도 $f_1: f_1$ (*S_1의 문형, S_2의 문형*)→ 유사도 (0~1)
$f_2 \sim f_5$:시제, 긍부정, 문형-시제-긍부정 조합, 양태의 유사도 함수	f_1과 동일한 방식으로 정의됨. $f_2: f_2$ (*S_1의 시제, S_2의 시제*)→ 유사도 (0~1) ...
f_6: 품사를 고려한 어휘 유사도 함수	입력된 사용자 발화의 형태소/품사열과 예제 발화의 형태소/품사열의 유사도 $f_6: f_6$ (*S_1의 형태소/품사집합, S_2의 형태소/품사집합*)→ 유사도 (0~1)

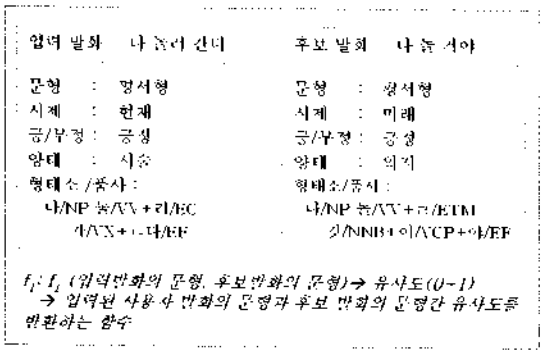

그림 5. 자질 함수의 의미
Fig. 5. Example of the feature function.

유사 발화 말뭉치를 통해 결정되며 방법은 3.3절에서 설명한다.

f_6 (품사를 고려한 어휘 유사도)는 두 발화의 어휘가 어느 정도 유사한가를 나타내는 함수로서 두 발화의 형태소/품사 집합에서 차이가 나는 어휘들에 대해 품사의 중요도를 반영하여 편집 비용을 설정한 후 편집 거리를 측정한 것으로서 식 (2)와 같다. 분모는 두 발화가 완전히 다르다고 가정할 때 상대 발화로 만드는데 필요한 최대 비용을 의미한다. 분자는 실제 두 발화간에 차이가나는 형태소들만을 동일하게 만드는 데 필요한 비용이다. 이것은 정규화된 편집 거리 (normalized edit distance)의 응용이다. 그러나 이때 각 형태소를 편집함에 있어 해당하는 품사를 고려하여 편집 비용을 측정한다.

$$f_6 = 1 - \frac{\sum_{k_{1,i} \in sSet} c(k_{1,i}, I) + \sum_{k_{2,j} \in sSet} c(k_{2,j}, S)}{\sum_{i=1}^{|U_1|} \max[c(k_{1,i}, I), c(k_{1,i}, S)] + \sum_{j=1}^{|U_2|} \max[c(k_{2,j}, I), c(k_{2,j}, S)]} \quad (2)$$

- U_1 : *입력 발화의 형태소/품사 집합*
- U_2 : *후보 발화의 형태소/품사 집합*
- $k_{1,i}=m_{1,i}/t_{1,i}$: U_1의 *i* 번째 형태소/품사

- $sSet=\{k_{1,i}$ or $k_{2,j} \mid m_{1,i} \neq m_{2,j} \wedge t_{1,i}=t_{2,j}\}$
- $iSet=\{k_{1,i}$ or $k_{2,j} \mid m_{1,i} \neq m_{2,j} \wedge t_{1,i} \neq t_{2,j}\}$
- $c(k_i, X)$: k_i를 편집하는데 필요한 편집 비용, $X \in \{I, S\}$
- I : *입력/삭제*
- S : *대체*

3.3. 자질별 유사도 함숫값의 결정

각 자질의 유사도는 함수로서 결정된다. 이 함수는 유사도가 평가된 정답 말뭉치에서 통계적으로 결정하였다. 말뭉치는 700개의 그룹 (5274개의 발화쌍)으로 구성되는데 한 그룹은 하나의 사용자 입력 발화와 여러 개의 유사도가 다양한 예제 발화들로 구성된다. 정답 말뭉치를 구축하기 위해 유사도 *0*에서 *4* (완전 다름, 약간 유사, 의미 동일, 거의 동일, 완전동일)까지의 수준을 정의하여 평가하도록 하였다. 유사도 0은 두 발화가 완전히 다른 경우로서 유사 발화로 검색되어서는 안 되는 임계값 (threshold)을 의미한다.

자질 함수는 f_1~f_5까지 동일한 방식으로 결정하였는데 문형의 경우를 예를 들면 식 (3)과 같은 통계적 분석을 통해 결정한다. 특정 문형 조합의 분포와 특정 유사도 분포간의 관련성을 상호 정보 (Point-wise Mutual Information) [9]를 통해 측정한다. 식 (3)은 i번째 자질의 유사도 함수로서 두 발화의 *i*번째 자질 한 쌍을 입력으로 하여 그들간의 유사도를 출력한다. 예를 들어 말뭉치에서 사용자 발화의 문형이 '의문형'이고 예제 발화의 문형이 '평서형'인 모든 경우에 대해 유사도가 얼마인지 빈도를 조사하여 관련성을 계산하는 것이다.

$$f_i(e_{1,i}, e_{2,i}) = \sum_{j=0}^{5} a_j \times \log \left| \frac{p(<e_{1,i}, e_{2,i}>, A_j)}{p(<e_{1,i}, e_{2,i}>) p(A_j)} \right| \quad (3)$$

- $e_{1,i}$: *입력 발화의 i 번째 자질 값*
- $e_{2,i}$: *후보 발화의 i 번째 자질 값*
- A_j : *j 유사도를 가진 발화 쌍*
- $p(<e_{1,i}, e_{2,i}>)$: *i 번째 자질 값의 쌍이 나타나는 비율*
- $p(A_j)$: *j 유사도를 가진 발화 쌍이 나타나는 비율*
- $p(<e_{1,i}, e_{2,i}>, A_j)$: *i 번째 자질 값의 쌍과 j 유사도를 가진 발화쌍이 동시에 나타나는 비율*
- a_j : *유사도 j에 대한 가중 값*

여기서 a_j는 유사도 *j*에 대한 가중 값으로 유사도*0*에서 이러한 조합이 분포되어 있는 경우와 유사도1에서 분포되어 있는 경우, 유사도3인 발화 쌍에서의 경우를 차별화하기 위함이다. 유사도*3*에서 많이 분포되어 있다면 유사한 발화의 특징이라고 할 수 있지만 그렇지 않고 유사도*0*인 발화 쌍에서 많이 분포되어 있다면 유사하지 않는 발화의 특징이라고 할 수 있기 때문이다. 또한 유사도2 보다는 유사도3에서 나타나는 것이 더 큰 영향을 준다고 할 수 있다. 이와 같은 계산 결과 자질 별 유사도 함수는 표 3의 예에서와 같이 결정되었다.

품사를 고려한 어휘 유사도 함수 f_6는 식 (2)와 같다. (2)는 품사별 편집 비용인 $c(k_i, X)$를 포함하는데 이것은 식 (4)로서 형태소의 편집 비용을 품사를 고려하여 책정한 것이다. 이 식은 유사도 평가 말뭉치에서 특정 품사가

표 3. 자질 별 유사도 함숫값의 예
Table 3. Example of similarity function values for each feature.

f_1	값	f_2	값	f_3	값
청유, 청유	3.25	미래, 미래	0.71	부정, 부정	0.63
의문, 의문	0.82	과거, 과거	0.67	긍정, 긍정	0.56
명령, 명령	0.67	현재, 현재	0.64	긍정, 부정	0.05
명령, 평서	0.21	현재, 미래	0.18		

유사도와 어떤 관련이 있는가의 경향성만을 분석하기 위한 식으로 품사를 수준별로 군집화하는데 사용하였다. 예를 들어 부사 (MAG)가 유사도 평가가 0인 발화쌍에서 나타나고 두 발화에서 공통으로 나타났다면 발화를 차별화하는데 중요한 역할을 못한 것으로 간주한다. 그렇지 않고 두 발화 중 한 곳에서만 나타났다면 발화를 차별화하는데 중요한 역할을 한 것으로 간주하여 비용이 증가한다. 마찬가지로 부사가 유사도 평가 3인 발화 쌍에서 둘 중 한 곳에서만 나타났다면 발화를 유사하게 만드는 데 중요한 역할을 한 것으로 간주하여 비용이 증가한다.

$$c(k,X)=a_0\times\frac{D_{0,k}}{I_{0,k}}+a_1\times\frac{D_{1,k}}{I_{1,k}}+a_2\times\frac{D_{2,k}}{I_{2,k}}+a_3\times\frac{D_{3,k}}{I_{3,k}} \quad (4)$$

- $D_{i,k}$: 품사 k가 유사도 평가 i인 발화 쌍 중 한 쪽에서만 나타난 횟수
- $I_{i,k}$: 품사 k가 유사도 평가 i인 발화 쌍 양 쪽에서만 나타난 횟수
- a_i : 유사도 평가 i (0~3)에 대한 가중치

이러한 분석을 통해 품사의 중요도에 대한 경향성을 분석하였으며, 품사는 7개의 그룹으로 분류하여 단순히 1에서 7까지의 정수를 비용으로 책정하였다. 식 (1)의 가중치 w_i는 학습 말뭉치에서 10%의 말뭉치를 대상으로 모든 가중치를 조합한 경우에 대해 일정 범위 안에서 (0~10) 작은 단위 (0.2)로 변화시키면서 성능을 측정하여 결정하였다. 식 (3), (4)에서의 가중치 a_i, a_i는 유사도평가 값이 높을수록 유사도에 미치는 영향이 커지도록 임의의 상수값을 주었다. 따라서 추후에는 이러한 가중치들의 값을 결정하기 위한 적절한 방법이 연구되어야 할 것이다.

IV. 실험 및 평가

제안하는 발화간 유사도 측정 방법의 효과성을 평가하기 위하여 3장에서 정의한 유사도 기준에 따라 유사도 평가 말뭉치를 구축한 후 여기에 본 연구에서 제안하는 측정 방법 및 기존 연구 인 PEBL [6]과 [8]의 방법을 적용하여 정확률 및 유사도 재현율을 비교하였다. 본 장에서는 먼저 말뭉치 구축 등 실험 환경에 대해 살펴보고 실험 결과를 소개한 후 이에 대한 평가를 하도록 한다.

4.1. 실험 환경

본 연구에서 유사 발화 말뭉치를 구축한 목적은 다음 네 가지에 이용하기 위해서이다. 첫째 유사도 측정 방법의 자질값 함수 결정, 둘째 각 자질의 가중치 학습, 셋째 유사도 임계값 결정, 넷째 제안하는 모델의 성능 평가 등을 위해서이다. 유사 발화 말뭉치 구축을 하려면 먼저 유사 발화를 수집해야 하며 그 다음 각 발화들 간에 유사도를 앞에서 정의한 유사도 기준에 의해 평가해야 한다.

말뭉치는 3.3절의 유사도 기준에 따라 표 4의 형태로 구축되었다. 실험의 편의상 키워드로 군집화한 후 가장 빈도가 높은 발화를 사용자 입력 발화로 선택하고 다양한 유사도 수준의 예제 발화와 유사도를 앞서 언급한 다섯 가지 기준에 의해 평가하도록 하였다. 그룹 내에 유사발화가 없는 경우도 있으며 가장 유사한 발화가 2개 이상인 경우도 있다.

이렇게 하여 구축된 발화 그룹은 700개이며, 5274개의 유사 발화 쌍으로 이루어져 있다. 즉 그룹내의 1개의 기준 발화에 대해 평균 7.53개의 후보 발화가 존재한다. 이 중 유사도가 0인 발화 쌍은 2365쌍 (44.84%)이며, 유사도 1인 발화 쌍은 1002쌍 (19%), 유사도가 2인 것은 1027 (19.47%), 유사도가 3인 것은 880쌍 (16.69%)이다. 따라서 전체 유사한 발화는 쌍 (55.16%)이다. 실험은 이 말뭉치를 임의로 10개로 나누어 번갈아 가며 9개는 학습을 위해 1개는 평가를 위해 사용하는 10-묶음 교차 검증 (10-fold cross-validation) [10] 방법으로 이루어졌다.

실험을 통해 평가하고자 하는 바는 다음과 같다. 첫째 유사 발화 검색은 보다 정확한 응답을 위해서 사용자 발화와 가장 유사한 예제를 찾을 수 있어야 한다. 둘째 유사

표 4. 유사도 말뭉치 구축의 예
Table 4. Example of manual labeling of similarities.

사용자 발화	예제 발화	유사도
뭐 해?	뭐 하니	3
	아까 뭐 하고 있었어	1
	요즘 뭐 하고 있니	1
	요즘 뭐 해?	1
	뭐 하라고	0

한 발화 예제가 없는 경우에는 부적절한 응답을 하지 않고 시스템이 다른 대응을 할 수 있도록 '정답 없음'을 출력해 주어야한다. 이 두 가지를 충족하는지의 여부가 '정확률'로서 식 (5)와 같다. 단, 앞서 언급했듯이 가장 유사한 발화가 2개 이상인 경우 그 중 하나를 맞히면 맞는 것으로 하였다. 마지막으로 가장 유사한 발화를 찾지 못한 경우라도 활용 가능할 정도의 유사한 발화가 존재하면 이를 유사하다고 판단하여 출력할 수 있어야 한다. 이것은 시스템의 견고성 (robustness)에 해당하며 식 (6)과 같은 '유사 발화 재현율'로서 평가할 수 있다.

$$\text{정확률} = \frac{\text{정답존재여부맞힌1위발화를맞힌경우의수}}{\text{전체평가횟수}} \quad (5)$$

$$\text{유사 발화 재현률} = \frac{\text{시스템평가임계값이상인발화의수}}{\text{전체유사한 발화의 수}} \quad (6)$$

정확률은 각축한 말뭉치에서 발화 그룹을 대상으로 사용자 발화를 입력으로 하여 평가한 것이며 유사 발화 재현율은 그룹과 관계없이 유사도를 평가한 발화쌍을 대상으로 평가한 것이다. 앞서 3.3절에서 언급한 임계값을 학습 말뭉치에서 유사도 기준 1과 같도록 학습한 후에 평가 말뭉치 상에서 실제 유사도 1 이상인 발화가 임계값을 적용했을 때 얼마나 나타나는가를 본 것이다.

4.2. 실험 결과 및 분석

정확률에 대한 기준선 (base line)은 [6]을 형태소 단위에 적용한 것과 [8]의 유사도 측정 기준을 선정하였다. [6]은 간단하면서도 많이 사용되는 방법이고, [8]의 방법은 동일하게 한글에 대하여 대화 시스템을 위해 제안된 방법이기 때문이다. 기존 연구와의 비교 실험 결과는 표 5와 같다. 기존의 방법은 가장 유사한 것만 선택할 수 있을 뿐 임계값 학습을 통해 모든 유사한 것을 찾는 것이 불가능 하므로 재현율은 비교 실험을 하지 않았다. 제안하는 방법의 정확률은 73.97%, 재현율은 78.15%로 나타났다. 이는 기존의 형태 자질만을 이용한 방식에 비해 본 연구에서 제안하는 방법이 약 10%p 내외의 성능 향상이 있음을 보여준다.

이러한 성능 향상은 II장에서 제기한 문제점들을 본 논문에서 제안한 방법을 통해 해결할 수 있음을 보여준다. 그림 6은 그림 3, 그림 4의 예들이 어떻게 처리되는 지, 실험 결과를 보여 준다. 왼쪽은 그림 3의 예로서 문형 및

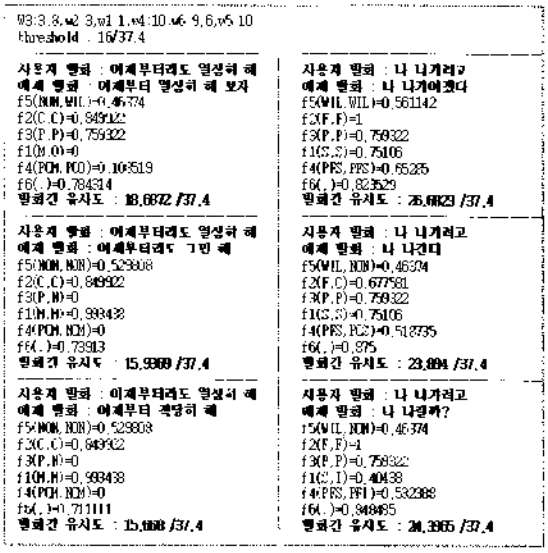

그림 6. 그림 3, 그림 4의 예에 대한 실험 결과
Fig. 6. Experimental results of figure 3 and 4.

표 6. 자질 별 제거 실험 결과
Table 6. Performance results when excluding features one by one.

제거 자질	정확률 (%)
모든 자질 사용	73.97
-문형	68.59
-시제	69.18
-긍/부정	68.76
-조합	68.96
-양태	66.52
-어휘 유사도	68.36

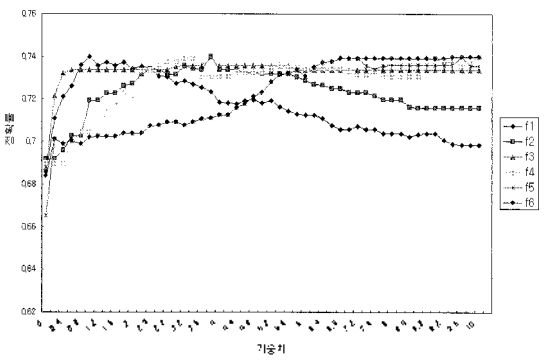

그림 7. 가중치 변화에 따른 정확률 측정
Fig. 7. Precision based on weight changes.

표 5. 1-best 정확률 비교 실험결과
Table 5. Experimental results of accuracy

방법	정확률 (%)	재현율 (%)
Base-Line 1 [6]	61.43	-
Base-Line 2 [8]	64.29	-
제안하는 방법	73.97	78.15

양태가 다르더라도 제거 되지 않고 "이제부터 열심히 해 보자"가 가장 유사한 것으로 선택되는 것을 볼 수 있다. 오른쪽은 그림 4의 예로서 서로 다른 문형과 양태를 갖는 문장들의 유사한 정도가 측정된 것을 알 수 있다.

한편 표 6은 각 자질 별 제거 실험 결과이다. 자질 별 제거 실험에서는 양태를 제거했을 때 성능 하락이 가장 컸고 그 다음은 어휘 유사도로 나타났다. 그림 7은 실험적으로 추정한 가중치를 적용하고 각 자질 별로 나머지 자질의 가중치를 고정 시킨 채 해당 가중치만을 변화시키면서 성능을 측정한 결과이다. 문형, 시제의 경우 반드시 사용해야하는 자질이나 가중치를 높이면 성능이 떨어짐을 알 수 있다. 긍/부정 자질, 문형/시제/긍부정 조합 자질의 경우 반드시 필요한 자질이나 가중치 영향을 크게 받지 않는 것으로 나타났다. 어휘 유사도나 양태의 경우는 자질의 비중을 높일수록 성능이 높아졌으나 결국 수렴하는 모습을 보인다. 실험 결과 제안하는 모든 의미 자질이 발화간 유사도 측정에 유효하다는 것과 각 의미 자질의 존재 여부뿐 아니라 유사한 정도를 측정하는 것이 유효하다는 것을 알 수 있었다.

V. 결론 및 향후 연구

본 논문에서는 예제 기반 대화 시스템에서 사용자 발화와 예제 발화들 간에 유사도를 효과적으로 측정하는 새로운 방법을 제안하였다. 대화의 자연스러움을 위해서는 정보 전달을 위한 내용어 뿐만 아니라 문형, 시제, 긍/부정, 양태 등 발화의 다양한 요소를 고려하여 사용자 발화와 가장 유사한 것을 선택해야 한다. 또한, 예제 기반 시스템에서는 제한적인 예제를 최대한 활용하여 대화를 이어가되 유사한 정도에 대한 한계를 설정하는 전략이 필요하다. 그러나 기존의 연구들은 이러한 요구를 충족하기에는 부족함이 있었다.

본 연구에서는 발화간의 형태적 유사도 뿐 아니라 의미적인 유사도를 측정하기 위하여 형태 자질과 함께 문형, 시제, 긍/부정, 양태 등 통사-의미적으로 주요한 자질들을 사용하는 방법을 제안하였다. 또한 자질들을 사용함에 있어 두 발화의 자질이 다르더라도 각 자질간에 유사할 확률을 계산하며, 마지막에 모든 자질의 유사도 값을 고려하여 전체 유사도를 계산하도록 하였다. 각 자질 의 유사도 함수는 해당 자질 쌍에 대한 유사도를 측정하는 것으로서 유사도 평가 말뭉치에서 특정 자질 쌍이 나타난 분포와 특정 유사도 평가 분포간의 상관관계로부터 통계적으로 결정된다. 이를 위해 발화간의 유사도 기준을 5개의 수준으로 정의하고 이 기준에 따라 발화쌍의 유사도를 평가하여 말뭉치를 구축하였다. 또한 이 유사도 평가 말뭉치로 부터 형태적 자질에서 사용하는 품사별 중요도 및 발화의 유사성에 대한 하한선을 결정하였다.

제안하는 방법은 기존의 방식에 비하여 무엇보다 검색의 정확률를 향상시킬 수 있다는 장점이 있다. 즉, 대량의 예제가 존재할 때 다양한 의미 자질 및 형태 자질을 이용하여 의미적으로 가장 유사한 발화가 선택되어 대화를 자연스럽게 이루어지도록 한다. 그리고 자질간 유사도를 측정하고 또 이 값들을 동시에 고려함으로써 정확률과 함께 재현율을 높여 예제의 개수가 제한적일 때에도 최대한 대화를 지속할 수 있도록 한다. 또한 본 연구에서 제안한 방법은 각 자질값의 가중치 및 유사도 임계값 조정을 통해 응용 시스템의 요구에 따라 유사도 방향 및 한계를 조정할 수 있는 유연성을 제공한다. 마지막으로 의미 수준에서의 유사성을 고려하고 품사별 편집 가중치를 고려함으로써 생략이나 문체 변화 등 구어체 발화가 가지는 특성에 견고하다는 특징이 있다.

향후에는 각 자질의 가중치를 기계학습을 통해 최적화하고 아직 고려하지 않은 구문 구조나 서술어-논항 구조, 화용론적 분석을 결합하여 성능을 향상시킬 여지가 있는지에 대한 연구를 수행할 수 있을 것이다.

감사의 글

이 논문은 2단계 BK21사업과 2008년도 한국과학재단의 지원을 받아 수행된 연구임 (No. R01-2006-000-11162-0).

참 고 문 헌

1. E Levin, R Pieraccini, and W Eckert., "Using markov decision processes for learning dialogue strategies", In Proceedings of ICASSP98, 1, 201-204, 1998.
2. S Young, "Talking to machines (statistically speaking)", In Proceedings of ICSLP-2002, 9-16, 2002.
3. G Salton, "The SMART Retrieval System - Experiments in Automatic Document Processing", Prentice Hall Inc., EngleWoodCliffs, NJ, 1971.
4. I McCowan, D Moore, J Dines, D Gatica-Perez, M Flynn, P Wellner, and H Bourlard., "On the Use of Information Retrieval Measures for Speech Recognition Evaluation", IDIAP-RR 04-73, 2004.
5. K Papineni, S Roukos, T Ward, and WJ Zhu, "BLEU:A

method for automatic evaluation of machine translation", In Proceedings of ACL02, pp. 311-318, 2002.
6. C Tillmann, S Vogel, H Ney, A Zubiaga, and H Sawaf, "Accelerated DP based search for statistical translation", InEUROSPEECH-1997, 2667-2670, 1997.
7. N Inui, T Koiso, J Nakamura andY Kotani, "Fully Corpus-Based Natural Language Dialogue System", AAAI Spring Symposium, 2003.
8. C Lee, S Jung, M Jeong, and GG Lee, "Chat and Goal-Oriented Dialog Together: A Unified Example-based Architecture for Multi-Domain Dialog Management", Proceedings of the IEEE/ACL 2006 workshop on spoken language technology (SLT), 2006.
9. Y Yang and JO Pedersen, ., "A comparative study on feature selection in text categorization", In Proceedings 14th International Conference on Machine Learning (ICML-97), 412-420, 1997.
10. R Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2(12): 1137-1143, (Morgan Kaufmann, San Mateo), 1995.

저자 약력

•이 연 수 (Yeon-Su Lee)

1975년 10월 18일생
2000년 : 고려대학교 컴퓨터학과 학사
2008년 : 고려대학교 컴퓨터학과 석사 과정
2008년 ~ 현재 : 고려대학교 컴퓨터·전파통신공학과 박사과정

•신 중 휘 (Joong-Hwi Shin)

1982년 4월 4일생
2006년 : 고려대학교 컴퓨터학과 학사
2006년 ~ 현재 : 고려대학교 컴퓨터학과 석사과정

•홍 금 원 (Gumwon Hon)

1974년 6월 26일생
2000년 : 고려대학교 컴퓨터학과 학사
2002년 : 고려대학교 컴퓨터학과 석사
2007년 ~ 현재 : 고려대학교 컴퓨터·전파통신공학과 박사과정

•송 영 인 (Young-In Song)

1975년 12월 4일생
2001년 : 고려대학교 컴퓨터학과 학사
2003년 : 고려대학교 컴퓨터학과 석사
2008년 : 고려대학교 컴퓨터학과 박사
2008년 ~ 현재 : 고려대학교 컴퓨터학과 연구교수

•이 도 길 (Do-Gil Lee)

1974년 11월 23일생
1999년 고려대학교 컴퓨터학과 학사
2001년 : 고려대학교 컴퓨터학과 석사
2005년 : 고려대학교 컴퓨터학과 박사
2005년 ~ 2006년 : 고려대학교 컴퓨터정보통신연구소 연구교수
2006년 ~ 2008년 : NHN (주) 과장
2008년 ~ 현재 : 고려대학교 민족문화연구원 HK연구교수

•임 해 창 (Hae-Chang Rim)

1953년 2월 26일생
1981년 : Missouri 주립대학 학사
1983년 : Missouri 주립대학 석사
1990년 : Texas 주립대학 박사
1991년 ~ 1994년 : 고려대학교 전산과학과 조교수
1994년 ~ 1999년 : 고려대학교 컴퓨터학과 부교수
1999년 ~ 현재 : 고려대학교 컴퓨터·전파통신공학과 교수