韓國 컴퓨터情報學會 論文誌
第13卷 第7號, 2008. 12.

2008-13-7-1-7

인스턴트 메시징에서의 대화 주제 및 주제 전환 탐지

최윤정*, 신욱현*, 정윤재*, 맹성현**, 한경수***

Topic and Topic Change Detection in Instance Messaging

Yoonjung Choi*, Wookhyun Shin*, Yoonjae Jeong*, Sung-Hyon Myaeng**, Kyoung-Soo Han***

요 약

본 논문에서는 인스턴트 메시징(Instant Messaging), 채팅과 같은 텍스트 기반의 대화에서 현재 발화를 기준으로 대화의 주제를 파악하고, 대화 주제 전환 여부를 판단하는 기법에 대해 기술한다. 대화는 다른 종류의 글과 다르게 길이가 매우 짧아 적은 수의 단어를 사용하고, 두 사람 이상이 참여를 하며, 대화의 이력(History)이 현재의 발화에 영향을 미친다. 이러한 특성에 따라 본 논문에서는 사용자 발화 뿐 아니라 대화 상대자의 발화에서 추출한 키워드 기반으로 주제 탐지를 하며, 대화의 이력도 고려하여 대화 주제 탐지의 정확도를 높힌 연구 결과를 기술한다. 대화 주제 전환 탐지는 이전 발화와 현재 발화에서 탐지된 주제의 유사성을 계산하여, 유사성이 낮은 경우에 전환 탐지가 이루어졌다고 판단하였다. 본 논문의 실험에서 대화 주제 탐지는 88.20%, 대화 주제 전환 탐지는 87.36%의 정확도를 얻었다.

Abstract

This paper describes a novel method for identifying the main topic and detecting topic changes in a text-based dialogue as in Instant Messaging (IM). Compared to other forms of text, dialogues are uniquely characterized with the short length of text with small number of words, two or more participants, and existence of a history that affects the current utterance. Noting the characteristics, our method detects the main topic of a dialogue by considering the keywords not only the utterances of the user but also the dialogue system's responses. Dialogue histories are also considered in the detection process to increase accuracy. For topic change detection, the similarity between the former utterance's topic and the current utterance's topic is calculated. If the similarity is smaller than a certain threshold, our system judges that the topic has been changed from the current utterance. We obtained 88.2% and 87.4% accuracy in topic detection and topic change detection, respectively.

▸ Keyword : 주제 탐지(Topic Detection), 주제 전환 탐지(Topic Change Detection), 인스턴트 메시징(Instance Messaging)

• 제1저자 : 최윤정
• 접수일 : 2008. 10. 24, 심사일 : 2008. 12. 3, 심사완료일 : 2008. 12. 24.
* 한국정보통신대학교 공학부 ** 한국정보통신대학교 공학부 정교수 *** SK텔레콤
※ 본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음.
〔2008-F-047-01, Urban Computing Middleware 기술 개발〕

Ⅰ. 서 론

인터넷의 발전에 따라, 메일, 채팅, 인스턴트 메시징(Instant Messaging) 등과 같은 인터넷을 이용한 의사소통(Communication) 수단이 현대 사회에서 중요해지고 있다[2]. 실제 많은 회사나 학교 내에서 메일이나 채팅 등을 통하여 업무, 과제 등과 관련한 많은 정보 교환이 이루어진다. 특히, 인스턴트 메시징은 젊은 사람들로부터 빠르게 확산되었으며, 최근에는 메일을 능가하여 사람들 사이에서 의사소통을 하기 위한 주된 수단으로 이용되고 있다[1].

인스턴트 메시징은 인터넷 상에서 쌍방향 의사소통이 가능한 서비스로, 실시간으로 텍스트를 사용한 쪽지 교환, 1대1 대화 등을 통하여 사람들 간의 의사소통을 가능 하게 할 뿐 아니라, 상대방의 접속 유무 확인, 파일 공유 및 전송 등도 가능하다. 현재 대표적인 인스턴트 메시징 서비스로는 MSN 메신저, 구글 메신저 Google Talk, 네이트온(NateOn) 등이 있다.

최근에는 모바일 인스턴트 메시징(Mobile Instant Messaging)도 등장하여 사람들의 관심을 받고 있다. 이는 기존 인스턴트 메시징을 이동통신 환경에 도입하여 이동성이라는 모바일의 장점을 추가하여 휴대폰이나 PDA로 인스턴트 메시징을 즐길 수 있는 서비스이다[9]. 이로써 앞으로 인스턴트 메시징이 현대인의 중요한 의사소통 수단 중 하나로써 자리매김 될 것이다.

이러한 인스턴트 메시징의 대화 내용에는 사용자와 관련하여 많은 정보를 내포하고 있다. 대화 내용을 통해 사용자의 관심사(Interest) 등을 알 수 있기에 사용자 모델링(User Modeling)에도 도움이 되고, 응용프로그램 내에서 광고, 서비스, 뉴스, 음악 등의 추천서비스에도 쓰일 수 있다. 또한, 현재 대화 내용에 적합한 컨텐츠(Contents)를 제공할 수도 있다. 하지만, 한 세션(Session)내의 대화 내용에는 한 개 이상의 주제가 포함 될 수 있으며, 각 발화의 길이는 매우 짧고, 문법에 맞춰지지 않은 경우가 많기에 의미를 파악하기는 쉽지 않다. 이러한 어려움 때문에 채팅이나 인스턴트 메시징에서의 대화 주제를 파악하는 연구는 아직 드물다.

본 논문에서는 인스턴트 메시징의 대화에서 사용자의 발화, 대화 상대자의 발화, 그리고 대화의 이력을 고려하여 현재 발화의 대화 주제를 파악하고자 한다. 또한, 현재 발화와 이전 발화들과 비교하여, 현재 발화의 주제 전환 탐지 여부를 판단하려 한다.

본 논문의 구성은 다음과 같다. 2장에서는 대화 주제 탐지의 관련 연구를 살펴보고, 3장에서는 본 연구에서 제시하는 주제별 키워드 연관도에 대하여 소개한다. 4장에서 대화 주제 및 주제 전환 탐지 모듈에 대하여 설명하고, 5장에서 시스템의 성능을 알아보기 위한 실험 방법과 실험결과를 보여준다. 6장에서 실험 결과 분석하고, 7장에서는 본 논문의 결론 및 향후 연구를 제시한다.

Ⅱ. 관련 연구

주제 탐지에 관한 연구는 예전부터 많이 진행되어 오고 있다[1,2,5,6,7]. 하지만, 이들은 대부분이 문서나 메일과 같은 글의 길이가 짧지 않고 적지 않은 내용을 포함하며 주제가 자주 변하지 않는 글에 대하여 주제 탐지를 해왔다. 최근에 들어서 인스턴트 메시징과 채팅이 확산되면서, 이들의 대화 내용의 주제를 탐지하는 연구가 조금씩 진행되고 있으나 아직은 초기 단계이다.

[1]에서는 대화의 주제는 사용자의 관심사에 따라 다르기 때문에, 대화의 주제를 분류(Classification)하여 사용자 프로파일(User Profile)을 생성하는 시스템을 제안하였다. 대화 주제 분류를 위해 이들은 벡터 공간 모델에 근거한 분류기(classifier)를 이용하였다. 학습된 분류기는 각각의 주제 카테고리를 대표하는 단어들의 벡터로 나타내고, TF*IDF(Term Frequency * Inverse Document Frequency)를 기반으로 가중치를 계산하였다. 대화 내용이 입력되면 키워드를 추출하여 벡터로 표현하고, 이 벡터와 학습되어 있는 각 카테고리의 벡터간의 유사도를 계산하여 대화 주제를 분류하였다.

[2]에서도 TF*IDF 기반 벡터 공간 모델을 사용하여 대화의 주제 탐지를 하였다. 이들은 각 발화에 대해 주제를 탐지한 것이 아니라, 각 대화 세션에 대하여 대화 주제를 탐지하였다. 이에 따라 한 세션 내에서 시간차이에 따라 벌점(Penalty)을 부과하였고, 사용자의 별명(Nickname)과 WordNet[10]을 이용하여 탐지된 주어진 단어의 상위어(Hypernym)를 속성으로 추가하여 대화 주제 탐지의 성능을 향상시켰다.

대화의 경우, 일반 문서와 다르게 주제 전환이 빠르고 자주 일어난다. [3]은 텍스트 틸링(Text Tiling)을 이용하여 한 세션의 대화를 같은 주제를 갖는 대화 블록(Block)으로 나누어서 대화 주제 전환을 탐지하였다.

[4]에서는 새로운 단어의 출현여부와 미리 모아놓은 주제 전환을 의미하는 표현법 등을 이용하여 주제 전환 여부를 탐지하였다. 또한, 이들은 발화 내의 글만 고려한 것이 아니라, 입력시간도 고려하여 침묵(Silence)의 시간이 길어질수록 주제가 전환될 가능성이 높다고 판단하여 주제 전환 탐지를 위한 하나의 속성으로 사용하였다.

본 연구에서는 실시간으로 발화가 입력됨과 동시에 대화 주제 및 주제 전환 탐지가 가능하며, 소요시간이 길지 않으면서 높은 정확도를 갖는 기술 개발에 주안점을 주고 있다.

III. 주제별 키워드 연관도

주제별 키워드 연관도는 같은 주제 내에서 출현한 키워드의 빈도수에 기반한다. 한 키워드가 특정 주제를 가진 문장에서 많이 추출 된다면, 그 키워드는 이 주제와 연관도가 높다고 할 수 있다. K 는 전체 키워드의 집합이고, *freq(k, t)* 는 주제 t 에서 출현한 개별 키워드 k 의 빈도수라 한다면, 주제별 키워드 연관도는 다음 수식 (3.1)과 같다.

$$w_{t,k} = \min\left(1, \frac{freq(k,t)^2}{\sqrt{\sum_{i \in K} freq(i,t)}}\right) \quad \cdots\cdots (3.1)$$

하나의 주제에 대하여 존재 할 수 있는 키워드의 수가 많기 때문에, 간단한 빈도수의 확률로 계산하게 되면 값이 너무 작아지게 된다. 이러한 문제를 해결하기 위하여, 수 차례의 실험을 통하여 수식 (3.1)을 적용하여 값이 0과 1사이에 분포하도록 하였다. 주제별 키워드 연관도는 0과 1사이의 값을 나타내며, 특정 주제에 대하여 자주 발생하는 키워드 일수록 1에 가까운 값을 갖는다.

IV. 주제 및 주제 전환 탐지

대화 주제는 발화내의 키워드 기반으로 탐지한다. 먼저, 입력된 발화로부터 키워드를 추출하고, 3장에서 설명한 각 키워드에 대하여 주제별 키워드 연관도를 계산한다. 발화에 대한 대화 주제 관련도는 수식 (4.1)과 같이, 추출된 키워드와 각각의 대화 주제 사이의 주제별 키워드 연관도의 합으로 계산된다. 이때, 앞에서 설명했듯이 주제별 키워드 연관도는 빈도수에 의해 결정되며, 두드러진 몇몇의 단어 외에는 빈도수가 작아서, 관련도의 값이 매우 작다. 이를 보완하기 위하여 주제별 키워드 연관도에 제곱근을 한 뒤에 이들의 합을 계산한다.

$$TopicScore_{\ n}(t) = \sum_{k=1}^{N} \sqrt{w_{t,k}} \quad \cdots\cdots (4.1)$$

wt,k는 대화 주제 t와 각 발화에서 발견된 키워드 k의 연관도를 나타내고, N은 발화에서 추출된 키워드의 수이다.

본 논문에서는 사용자 발화, 대화 상대자 발화, 그리고 대화 이력을 고려하여 대화 주제 관련도를 계산하고, 가장 높은 관련도를 갖는 주제를 발화의 대화 주제로 선정한다. 또한, 현재 발화의 주제 관련도와 이전 발화의 주제 관련도를 비교하여 주제 전환을 탐지한다.

4.1 사용자 발화를 고려한 대화 주제 탐지

수식 (4.1)을 이용하여 사용자 발화로부터 추출된 키워드 기반으로 사용자 발화의 대화 주제 관련도는 수식 (4.2)와 같이 계산 할 수 있다. N_{user} 는 사용자 발화에서 추출된 키워드의 수를 의미한다.

$$TopicScore_{\ user,n}(t) = \sum_{k=1}^{N_{user}} \sqrt{w_{t,k}} \quad \cdots\cdots (4.2)$$

4.2 사용자와 대화 상대자 발화의 동시 고려

사용자 발화만 고려하는 경우, 올바른 대화주제를 탐지하는데 어려움이 있다. 대화는 두 사람 이상이 서로 이야기를 주고받는 것이기 때문이다. 사용자가 대화 상대자에게 많은 이야기를 하는 경우라면 문제가 되지 않지만, 사용자가 항상 많은 이야기를 하지 않는다. 경우에 따라서 대화 상대자가 주로 대화를 이끌어 나가고 사용자는 간단한 대답만 하는 경우도 종종 있다. 이 경우에는 사용자 발화만으로 주제를 탐지하기는 어렵다. 또한, 사용자 발화에서 추출된 키워드에서 모호성(Ambiguity)을 가지고 있으면 올바른 대화 주제를 탐지하기에 어려움이 있다. 이를 보완하여, 사용자 발화 뿐 아니라 대화 상대자 발화도 고려하여 대화 주제를 탐지한다.

대화 상대자 발화의 대화 주제 관련도는 수식 (4.3)으로 얻을 수 있다. 이때, N_{part} 는 대화 상대자(Partner) 발화에서 추출된 키워드의 수를 의미한다.

$$TopicScore_{\ part,n}(t) = \sum_{k=1}^{N_{user}} \sqrt{w_{t,k}} \quad \cdots\cdots (4.3)$$

대화 주제 탐지의 본래의 목적은 사용자 발화에서의 대화 주제 탐지이며, 대화 상대자 발화는 사용자 발화에서 키워드가 추출되지 않았을 경우나, 뚜렷한 대화 주제가 탐지되지 않았을 경우에 이를 보완하기 위해 사용되는 것이기에, 대화 상대자 발화로부터 계산된 주제 관련도보다 사용자 발화에서 탐지된 주제 관련도에 높은 가중치를 준다. w_{user} 는 사용자 발화의 가중치를 의미하며, 0.5보다 높은 값을 나타낸다. 사용자와 대화 상대자 발화 모두 고려한 키워드 기반 대화 주제 관련도는 다음 수식 (4.4)와 같다.

$$\begin{aligned} &TopicScore_{\ keyword,n}(t) \\ &\quad = w_{user} \times TopicScore_{\ user,n}(t) \\ &\quad + (1 - w_{user}) \times TopicScore_{\ part,n}(t) \end{aligned} \quad \cdots (4.4)$$

4.3 대화 이력(history)에 기반한 주제 탐지 보정

현재 발화의 대화 주제를 탐지 하는데 있어서, 이전 발화의 대화 주제를 고려하는 것도 중요하다. 대화의 특성상 대화 주제는 급격하게 변화하지 않으며, 동일한 주제가 유지되거나, 이전 주제와 관련 있는 주제로 변화하는 경향이 있다[3]. 따라서, 현재 발화(사용자 발화와 대화 상대자 발화)에서 키워드를 추출 할 수 없는 경우, 이전 발화에서 탐지된 대화 주제가 현재 발화의 대화 주제와 일치한다고 볼 수 있다. 또한, 발화에서 탐지된 주제가 있을 경우에도, 이전 발화에서 탐지된 대화 주제가 현재 발화의 올바른 대화 주제 탐지에 도움을 줄 수 있다. 현재 발화를 n 번째 발화라고 하면, 대화 이력을 기반으로하여 보정된 주제 관련도는 수식 (4.5)와 같다.

$$TopicScore_{history,n}(t) = w_{history} \times TopicScore_{n-1}(t) + (1-w_{history}) \times TopicScore_{keyworda,n}(t) \quad \cdots\cdots (4.5)$$

$TopicScore_{n-1}(t)$ 는 대화 주제 t 에 대하여 n-1 번째 발화(이전 발화)의 주제 관련도를 나타내며, $w_{history}$ 는 대화 이력의 가중치를 의미한다.

$TopicScore_{n-1}(t)$ 도 대화 이력을 고려하여 계산된 주제 관련도이기에, 이전 발화들에 대한 주제 관련도가 고려되어 있다. 따라서, n-1 번째 발화의 주제 관련도만 고려하더라도, 현재 세션에 있는 모든 이전 발화들의 주제 관련도를 포함한 결과를 얻을 수 있다. 이때, 매번 $w_{history}$ 이 곱해지기 때문에, 현재 발화로부터 거리가 먼 발화일수록 현 발화의 주제 관련도에 적은 영향을 미치게 된다.

4.4 주제 관련도에 기반한 대화 주제 탐지

앞에서 설명된 주제 관련도는 발화에서 탐지된 키워드의 주제별 연관도의 합에 기반하여 계산되었다. 따라서 탐지된 키워드의 수에 따라서 주제 관련도의 값이 달라진다. 올바른 대화 주제 탐지를 위해 수식 (4.6)과 같이, 마지막으로 계산된 대화 이력에 기반한 주제 관련도를 전체 키워드의 수로 나누어주어 정규화한다.

$$TopicScore_{n}(t) = \frac{TopicScore_{history,n}(t)}{N} \cdot (4.6)$$

현재 발화의 대화 주제는 수식 (4.7)과 같이, 각 주제에 대하여 주제 관련도가 가장 높은 주제로 결정한다.

$$Topic_{n}(t) = \arg max_{t \in T} TopicScore_{n}(t) \cdots (4.7)$$

4.5 대화 주제 전환 탐지

현재 발화에서 대화 주제 전환 여부 탐지는 코사인 유사도(Cosine Similarity) 방법을 사용한다. 이전 발화에서 탐지된 주제 목록의 주제 관련도와 현재 발화에서 탐지된 주제 목록의 주제 관련도를 비교하여, 유사도가 낮은 시점이 대화 주제가 전환되었다고 볼 수 있다. 이때, 탐지되지 않은 주제에 대한 주제 관련도는 0을 나타내는데, 0을 그대로 곱하게 되면 결과에 큰 영향을 미치게 된다. 이를 보완하기 위하여, 주제 관련도를 시그모이드(Sigmoid) 함수에 적용하여 재계산한 후에 코사인 유사성 방법을 이용하여 주제 전환 탐지를 한다.

수식 (4.8)은 본 논문에서 응용한 시그모이드 함수이다. 기존의 시그모이드 함수를 기반으로 본 연구에서 필요로 하는 형태로 변환하여, 그림 1과 같은 그래프를 얻었다. 주제 관련도가 0인 경우에는 최솟값(약 0.002473)을 갖고, 주제 관련도가 1인 경우에는 최댓값(약 0.997527)을 갖는다. 또한, 탐지된 주제 관련도 중 최댓값이 0.4 ~0.7 범위에 많이 존재하기 때문에, 시그모이드 함수를 이용하여 이 범위에 존재하는 주제 관련도에 편차를 주어 주제 전환 탐지의 정확도를 높였다.

$$TS_{n}(t) = \frac{1}{1+e^{-12 \times TopicScore_{n}(t)+6}} \cdots\cdots (4.8)$$

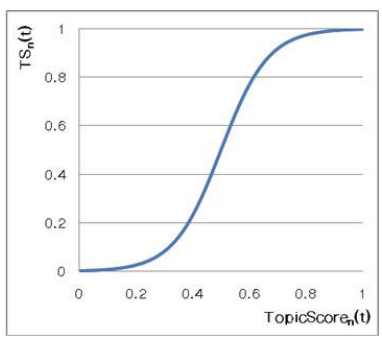

그림 1. 시그모이드 함수 그래프
Fig 1. The Graph of Sigmoid Function

시그모이드 함수에 의해 수정된 주제 관련도를 수식 (4.9)와 같이 코사인 유사성을 이용하여 이전 발화와의 유사도를 계산하여, 임계값(threshold)보다 작은 경우 주제가 전환되었다고 탐지한다. Un 은 n 번째 발화를 의미하며, T 는 전체 주제 분류의 개수를 의미한다.

$$Sim(U_n, U_{n-1}) = \frac{\sum_{t=1}^{T}(TS_n(t) \times TS_{n-1}(t))}{\sqrt{\sum_{t=1}^{T} TS_n(t)^2} \times \sqrt{\sum_{t=1}^{T} TS_{n-1}(t)^2}} \quad \cdots\cdots (4.9)$$

V. 실험 및 실험결과

본 논문에서 제안한 시스템을 평가하기 위하여 사용자식별 정보가 삭제된 이동통신사 대화 서비스 로그 중에서 일부의 발화를 선택하여 사용하였다. 사용자식별정보가 삭제된 이동통신사 대화 서비스 로그는 하루 동안 모바일 대화 서비스를 이용한 여러 사람들의 대화 로그를 저장한 것이다. 총 276세션(4,231개의 발화)을 선택하여 실험하였다. 각 발화에 대한 주제는 3명의 사람이 판단하여 결정하였다.

본 연구를 위하여, 모바일 인스턴트 메시징에 적합한 대화 주제 분류를 구축하였으며, 총 13개의 주제로 분류 된다 (표 1).

표 1. 대화 주제 분류 및 정의
Table 1. The Topic Category and its Definition

주제명	정의 및 예시
생활/건강	우리의 일상생활과 건강에 관련된 발화 (ex. 날씨, 다이어트, 음식 등)
엔터테인먼트	즐거움과 여유를 나타내는 발화 (ex. 만화, 영화, 유머 등)
로맨스/성	이성관계 및 결혼 관련된 발화 (ex. 연애)
모바일서비스	핸드폰에서 제공하는 서비스에 관련된 발화 (ex. 문자, 벨소리 등)
음악	음악에 관련된 발화 (ex. 노래 제목, 장르 등)
스포츠/레포츠	경기를 하는 스포츠와 레져, 모험을 겸비한 레포츠 관련된 발화 (ex. 스포츠 종목, 대회, 선수 등)
게임	게임에 관련된 발화 (ex. 게임명, 게임관련용어 등)
여행/지리	지역 정보에 관련된 발화 (ex. 음식점, 맛집, 교통정보 등)
쇼핑	상품 구매 활동 및 관련 정보에 대한 발화 (ex. 상품명, 쇼핑몰 등)
컴퓨터/인터넷	컴퓨터 관련 기술 및 서비스 관련 발화 (ex. 웹사이트, 바이러스 등)
교육/학문	지식습득 및 시험/진학 관련 발화 (ex. 대학, 시험, 입시정보 등)
사회/정치	사회적 이슈와 관련 기관/인물에 관련 발화 (ex. 정부, 대기업 등)
경영/경제/금융	현물이 오가는 경제활동 및 금융상품 (ex. 은행, 부동산 등)

대화 주제 및 주제 전환 탐지를 하기 위해 3장에서 언급한 주제별 키워드 연관도를 학습 데이터를 이용하여 먼저 계산하고, 이를 기반으로 주제 및 주제 전환 탐지에 대하여 실험한다.

5.1 주제별 키워드 연관도

본 논문에서 제안한 방법은 실시간으로 대화 주제 및 주제 전환 탐지에 사용 가능하도록 하기 위해 소요 시간을 최대한 단축시키도록 시스템 설계를 하였다. 이를 위해 주제별 키워드 연관도는 학습 데이터를 통해 계산해놓고, 이를 저장하는 주제 사전(Topic Dictionary)을 구축하였다.

주제 사전 구축을 위해 사용된 학습 데이터는 총 118,308개의 문장으로, 모바일 검색어 분류 결과, 웹 디렉토리 구조, 사용자 대화 샘플 등으로 구성되어 있으며, 각각의 문장은 하나의 주제가 입력되어 있다. 각 문장에서 형태소 분석기를 이용하여 단어를 추출하고, 추출된 단어는 입력된 주제로 분류된다. 주제별 단어 집합을 모은 뒤에 3장에서 언급한 수식을 이용하여 주제별 키워드 연관도를 계산하여, 주제 사전에 저장한다.

5.2 대화 주제 탐지 실험

총 3가지 실험으로, 사용자 발화만 고려한 경우(실험1), 사용자와 대화 상대자자 발화를 모두 고려한 경우(실험2), 그리고 대화 이력을 고려한 경우(실험3)에 대해 각각 실험하였다. 실험결과는 표 2와 같다.

표 2. 대화 주제 탐지 실험 결과
Table 2. The Result of Topic Detection Experiment

	발화개수	정확도
실험 1	1322	31.24%
실험 2	3078	72.74%
실험 3	3732	88.20%

표 2는 각 실험에서 가장 높은 정확도를 갖는 경우의 결과이다. 사용자 발화만 고려한 경우에는 매우 낮은 정확도를 갖지만, 대화 상대자 발화를 고려하고, 대화 이력을 고려하면서 정확도가 높아져 감을 알 수 있다.

실험1은 사용자 발화만 고려하여 대화 주제를 탐지하였고, 정확도는 31.24%로 측정되었다. 실험2에서 사용자 발화에 대

한 가중치가 0.9일 때 위와 같은 정확도(72.74%)로 가장 좋은 성능을 보였다. 즉, 사용자 발화의 주제 관련도가 많은 영향을 미치며, 대화 상대자 발화의 주제 관련도는 많은 영향을 미치지는 않지만 사용자의 발화에서 추출된 키워드가 없을 경우에는 낮은 가중치(0.1)를 주더라도 큰 영향을 미친다는 것을 의미한다. 그러나 표 3에서도 알 수 있듯이, 가중치에 따른 정확도의 차이가 크지 않다. 가중치보다도 대화 상대자의 발화를 고려한다는 자체가 큰 영향력을 지닌다는 것을 알 수 있다.

두 사람 이상의 대화에서 주로 대화를 이끌어 가는 사람이 많은 이야기를 하는 반면 상대방은 대답 위주로 대화를 하기 때문에, 대화를 이끌어 가는 사람의 발화로부터 주제 탐지를 위한 키워드가 주로 추출된다. 실험1과 실험2의 결과를 비교해보면, 대화 상대자의 발화를 고려하였을 경우에 사용자의 발화만 고려한 경우보다 약 2배 이상의 상승을 보였는데, 이로써 대화의 반은 사용자가, 나머지 대화의 반은 대화 상대자가 대화를 이끌어 간다는 것을 알 수 있다.

표 3. 대화 주제 탐지 실험2 결과
Table 3. The Result of Topic Detection Experiment2

가중치	발화개수	정확도
0.5	3061	72.34%
0.6	3065	72.44%
0.7	3065	72.44%
0.8	3077	72.72%
0.9	3078	72.74%

실험3에서 사용자 발화의 가중치는 실험2에서 가장 높은 정확도를 갖은 0.9로 설정하였다. 이 실험에서는 대화 이력의 가중치가 높아질수록 높은 정확도를 보였으며, 가중치가 0.5일 때 높은 정확도를 보였다. 하지만, 너무 높은 가중치를 부여하였을 경우, 대화 이력에 비해 현재 탐지된 주제의 비중이 적어지면서 주제 탐지의 정확도가 낮아졌다 (표 4). 이 실험 결과를 통해, 대화 이력이 대화 주제를 판단하는데 있어서 중요한 요소가 되지만, 너무 큰 비중을 차지하게 되는 경우 오히려 오류를 발생시킨다는 것을 보였다.

또한, 가중치가 0.1인 경우에는 실험2의 결과와 비교하였을 때 약 15%의 상승을 보였다. 대화에 있어서 항상 사용자와 대화 상대자 모두 많은 이야기를 하는 것이 아니기에 주제 탐지를 위한 키워드가 추출되지 않는 경우가 있다. 이러한 경우에 낮은 가중치라도 대화 이력을 고려하게 되면 대화 주제를 탐지 할 수 있음을 보여주는 실험 결과였다.

표 4. 대화 주제 탐지 실험3 결과
Table 4. The Result of Topic Detection Experiment3

가중치	발화개수	정확도
0.1	3692	87.26%
0.2	3698	87.40%
0.3	3695	87.33%
0.4	3726	88.06%
0.5	3732	88.20%

5.3 대화 주제 전환 탐지 실험

이전 발화와 비교했을 때, 현재 발화의 주제가 바뀐 시점을 대화 주제가 전환되었다고 한다. 주제가 전환된 시점에서 바로 전환 탐지가 되는 경우도 있지만, 대화 이력을 고려하고 있기 때문에 이전 발화의 영향에 의해서 바로 전환 탐지가 되지 않는 경우가 발생한다. 이러한 오류는 사람도 마찬가지이다. 현재 발화만으로 상대방이 이전 발화와 다른 주제를 이야기하고 있다는 것을 바로 알 수 있는 경우도 있지만, 한두번 지난 뒤에 상대방이 다른 주제를 이야기하고 있다는 것을 인식하는 경우도 있다. 이러한 경우를 고려하여, 주제가 전환된 시점에서 바로 전환 탐지가 되는 경우와, 전환 시점의 다음 발화에서 전환 탐지가 되는 경우에 대해서 올바르게 대화 주제 전환을 탐지하였다고 하였다.

표 5는 임계값에 따른 대화 주제 전환 탐지 실험의 결과이다. 결과에서 알 수 있듯이, 임계값이 너무 높을 경우에는 주제 전환이 자주 탐지되어, 주제 전환의 시점이 아닌 곳까지 탐지되는 경우가 발생하여 정확도가 낮아지게 된다. 반면, 임계값이 너무 작을 경우에는 주제 전환이 드물게 탐지되어, 주제 전환 탐지가 되어야 하는 시점임에도 불구하고 탐지되지 않는 경우에 발생하여 낮은 정확도를 갖는다. 따라서, 적당한 임계값을 주어 주제 전환 탐지 여부를 판단해야 한다. 본 논문의 실험에서는 임계값이 0.2일 때 가장 높은 정확도를 보였다.

표 5. 대화 주제 전환 탐지 실험 결과
Table 5. The Result of Topic Change Detection Experiment

임계값	발화개수	정확도
0.1	3644	86.13%
0.2	3696	87.36%
0.3	3660	86.50%
0.4	3575	84.50%
0.5	3525	83.31%

Ⅵ. 실험분석

표 2에서 볼 수 있었듯이, 주제 탐지 실험에서 실험의 고려 요소를 증가할수록 정확도가 증가하였으며, 소요 시간도 조금씩 길어졌다(그림 2). 그림에서 막대그래프는 각 실험의 평균 소요 시간을 의미하며, 실험에서 최대 소요 시간은 154.55ms이며, 최소 소요 시간은 9.80ms으로 측정되었다.

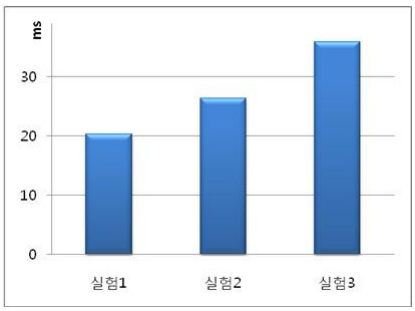

그림 2. 각 실험의 평균 소요 시간
Fig 2. The Average Elapsed Time of Each Experiment

사용자 발화만 고려한 경우, 사용자들이 발화를 입력하지 않거나, 키워드가 없는 "응", "아니" 등과 같이 짧은 내용만을 입력하게 되는 상황이 종종 발생하여 낮은 정확도를 가질 수 밖에 없게 된다.

반면, 대화 상대자 발화도 같이 고려하면 이러한 문제를 해결할 수 있다. 두 사람 모두 키워드가 존재하지 않는 짧은 대화를 주고받는 경우가 아니라면, 최소한 한 사람은 키워드가 있는 의미 있는 발화를 입력하기 때문이다. 이 경우 소요 시간은 사용자 발화만 고려한 경우보다 약간 길어지지만 높은 정확도를 나타내었다.

대화 이력을 고려함으로써 소요 시간이 길어지기는 하지만, 그만큼 정확도가 많이 증가하였다. 대화 이력의 고려는 글과는 다른 대화의 특성을 활용한 것으로써, 현재 발화의 주제를 결정할 중요한 키워드가 없어도 이전 발화를 고려하여 대화 주제를 결정 할 수 있다.

대화 탐지에 있어서 대부분의 오류는 키워드 때문에 많이 발생하였다. 주제 사전으로부터 주제별 키워드 연관도를 가져와 주제 관련도를 계산하는데, 주제 사전에 키워드가 존재 하지 않는 경우에는 주제별 키워드 연관도가 0이 된다. 본 논문에서 주제 사전을 제작하기 위해 많은 데이터를 이용하여 많은 키워드를 포함하고 있지만, 모든 키워드를 포함하지는 못하기에 문제가 발생하였다. 특히, 사람이름, 작품명과 같은 고유명사나 신조어 같이 시간이 변함에 따라 새로 생기는 단어들을 모두 포함하기는 쉽지 않았다. 또한 축약어의 경우, 본래의 단어는 주제 사전이 포함하고 있으나, 축약어와 일치시키지를 못하여 0의 값을 갖는 경우도 발생하였다.

대화 주제 전환 탐지는 표 5에서 볼 수 있듯이, 임계값이 0.2인 경우에 가장 높은 정확도를 보였다. 전환 탐지의 오류는 대부분이 주제 탐지의 오류에 의해 많이 발생하였다. 주제 전환 여부가 현재 발화와 이전 발화의 주제 관련도 유사성으로 판단되기 때문에, 앞서 말한 오류에 의해 주제 관련도가 잘못 계산되면 주제 전환 탐지에도 오류가 발생하게 된다. 또한, 주제 관련도에 대화 이력이 고려되어 있기 때문에, 현재 발화와 이전 발화의 주제 관련도 차이가 크게 나지 않아 오류가 발생하기도 하였다.

Ⅶ. 결론 및 향후 연구

인스턴트 메시징을 이용한 대화 내용에는 문서나 메일과 같이 많은 내용을 포함하고 있지 않다. 또한, 한번에 쓰는 글이 매우 짧으며, 문법에 잘 맞춰진 문장을 보기 힘들다. 따라서 문서와 메일에서 주제를 탐지하는 것보다 어렵고 해결해야 할 문제들이 많이 있다.

본 논문에서는 키워드 기반 주제 탐지를 기반으로 사용자와 대화 상대자의 발화에서 키워드를 추출하고, 대화 이력을 고려하여 대화 주제 및 주제 전환을 탐지하였다. 또한, 대화에 적합한 주제 탐지를 하기 위하여, 대화 주제 분류체계를 구축하였으며, 주제별 키워드 연관도를 담고 있는 주제 사전을 구축하였다. 이로 인해 비교적 높은 정확도를 얻을 수 있었다.

향후 연구에는 크게 두 가지 작업이 필요하다. 첫째는 주제 사전 구축 확장이다. 본 논문에서 사전 구축을 위해 많은 데이터를 사용하였지만, 모든 단어를 포함하기에는 부족함이 있었다. 따라서, 더 많은 데이터를 이용하여 충분한 단어를 포함하는 사전 구축이 필요하다. 또한, 사전에 없는 단어라도 추론(Reasoning) 등을 통하여 주제별 키워드 연관도를 계산하는 새로운 접근 방법이 요구된다. 둘째는 사용자 모델링을 이용한 대화 주제 탐지가 이루어져야 한다. 사용자들은 주로 자신이 관심이 있고 흥미가 있는 주제에 대하여 대화를 한다. 이에 따라, 사용자 모델링을 통하여 사용자의 관심사를 대화 주제 탐지에 적용하여, 주제 탐지의 정확도를 높이고 단어의 모호성도 감소시키는 것이 필요하다.

참고문헌

〔1〕 Jason Bengel, Susan Gauch, Eera Mittur, and Rajan Vijayarahgavan, "ChatTrack: Chat Room Topic Detection using Classification", 2nd Symposium on Intelligence and Security Informatics, 2004.

〔2〕 Paige H. Adams and Craig H. Martell, "Topic Detection and Extraction in Chat", IEEE International Conference on Semantic Computing, pp.581-588, 2008.

〔3〕 Alan P. Schmidt and Trevor K. M. Stone, "Detection of Topic Change in IRC Chat Logs", http://www.trevorstone.org/school/ircsegmentation.pdf.

〔4〕 Ichikawa Hiroshi and Tokunaga Takenobu, "An Empirical Study on Detection and Prediction of Topic Shifts in Information Seeking Chats", 11th Workshop on the Semantics and Pragmatics of Dialogue, pp.173-174, 2007.

〔5〕 Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi, "Simple Semantics in Topic Detection and Tracking", Information Retrieval, Vol. 7, pp.347-368, 2004.

〔6〕 Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi, "Topic Detection and Tracking with Spatio-Temporal Evidence", 25th European Conference on Information Retrieval Research (ECIR), pp.251-265, 2003.

〔7〕 Ramesh Nallapati, "Semantic Language Models for Topic Detection and Tracking", HLT-NAACL 2003 Student Research Workshop, Vol. 3, pp.1-6, 2003.

〔8〕 한승현, 임영환, "키워드 분석을 이용한 개인화 모바일 웹 뉴스 컨텐츠 생성에 관한 연구", 한국컴퓨터정보학회 논문지, Vol. 12, No. 3, pp.277-285, 2007.

〔9〕 Mobile Instant Messaging, www.mobilein.com/MIM.htm.

〔10〕 WordNet, wordnet.princeton.edu

저 자 소 개

최 윤 정

2007년: 한국정보통신대학교 전산학 학사

2007년~현재: 한국정보통신대학교 전산학과 석사과정

관심분야: Trend Analysis, User Interest, Topic Detection, Mood Detection, Ads Placement

신 욱 현

2007년: 한국정보통신대학교 전산학 학사

2007년~현재: 한국정보통신대학교 전산학과 석사과정

관심분야: Blog Search, Trend Analysis, Social Intelligence, Ads Placement.

정 윤 재

1998년: 포항공과대학교 전산학 학사

2007년: 한국정보통신대학교 전산학 석사

2007년~현재: 한국정보통신대학교 전산학과 박사과정

관심분야: Knowledge Discovery, Social Intelligence, Complex System.

맹 성 현

1983년: California State University, Hayward, 전산학 학사.

1985년: Southern Methodist University, Dallas, Texas, 전산학 석사.

1987년: Southern Methodist University, Dallas, Texas, 전산학 박사.

2003년~현재: 한국정보통신대학교 공학부 정교수.

관심분야: Information Retrieval, Text Mining, Natural Language Processing.

한 경 수

1998년: 고려대학교 컴퓨터학과 학사

2000년: 고려대학교 컴퓨터학과 석사

2006년: 고려대학교 컴퓨터학과 박사

2006년: 고려대학교 컴퓨터정보통신연구소 연구조교수.

2006년~현재: SK 텔레콤 재직.

관심분야: Information Retrieval, Text Mining, Natural Language Processing.