ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

모바일 환경에서 사용자 정보를 이용한 스토리 생성 방법

Story Generation Method using User Information in Mobile Environment

홍 진 표 차 정 원1*
Jeen-Pyo Hong Jeong-Won Cha

요 약

모바일 기기는 사용자가 늘 지니고 다니기 때문에 사용자의 주변 환경이나 행동 양상에 대한 매우 유용한 정보를 얻을 수 있다. 본 논문에서는 이들 정보를 하루 단위로 수집하여 하룻동안에 있었던 사용자의 행동에 대한 주제를 추출하고 이를 이용해 자동으로 일기를 생성하는 방법을 제안한다. 이를 위해 (1) 모바일 기기에서 사용자 행동 양상에 대한 정보를 모두 수집하고 (2) 수집한 정보로부터 개체명과 주제 연관 정보를 추출해 사용자가 그 날 있었던 일에 대한 주제를 추출한다. (3) (2)의 결과로부터 주제와 연관된 사건인 에피소드를 결정하고 (4) 문장 템플릿을 이용하여 문장을 생성한 후, 주제별 혹은 시간별로 스토리를 구성한다. 본 논문에서 제안한 방법은 기존의 방법보다 간단하기 때문에 모바일 기기 내에서도 수행이 가능하므로 개인 정보를 유출할 수 있는 문제를 최소화 할 수 있다. 또한, 본 논문에서는 문장의 형태로 정보를 제공하기 때문에 보다 많은 정보를 표현할 수 있다. 그리고 문장 생성 과정에 생성되는 주제 정보는 사용자의 행동 양상을 파악하는 자료로 이용할 수 있으므로 이를 바탕으로 한 사용자 맞춤형 서비스를 제공하는데 도움을 줄 수 있을 것으로 기대된다.

주제어 : 문장 생성, 정보 추출, 모바일 상황 정보, 자연어 처리

ABSTRACT

Mobile device can get useful user information, because users have always this device. In this paper, we propose automatically story generation method and user topic extraction using user information in mobile environment. Proposed method is follows: (1) We collect user action information in mobile device. Then, (2) we extract topics from collected information. (3) For the results of (2), we determine episodes for one day. Then, (4) we generate sentences using sentence templates and we compose stories which have theme-based or time-based. Because proposed method is simpler than previous method, proposed method can work only in mobile device. There's no room to leak user information. And proposed method is expressed more informative than previous method, because proposed method is provided sentence-based result. Extracted user-topic, a result of our method, can use to analyze user action and user preference.

☞ keyword : Sentence Generation, Information Extraction, Mobile Context, Natural Language Processing

1. 서 론

최근 젊은 층을 중심으로 모바일 기기에 오늘 하룻동안의 일을 일기로 작성하거나 사진을 찍어 기록을 남기는 욕구가 증가하고 있다. 사용자가 경험한 일에 대해 자동으로 이야기 형태로 작성하여 준다면 자신이 한 일을 되돌아 볼 수 있다. 또한, 이러한 정보를 바탕으로 일기를 작성하는데 도움을 줄 수 있을 것이다. 최근, 사용자가 늘 지니고 다니는 휴대폰이 스마트폰으로 대체되어 기존보다 다양한 정보 수집이 가능해졌으며, 휴대 인터넷 사용의 증가로 텍스트를 커뮤니케이션의 수단으로 삼는 경우가 늘어났다. 이러한 추세에 맞춰, 본 논문은 스마트폰과 같은 모바일 기기에서 수집 가능한 정보를 이용해 자동으로 일기를 생성해주는 방법을 제안한다.

기존에는 모바일 정보를 이용해 일기와 같은 하루 일과를 그림일기 형태로 생성하는 연구가 진행되었다[1, 2]. 이러한 그림일기는 문장에 비해 정보를 세부적으로 제공하지는 못하며, 하룻동안 사용자가 찍은 사진과 같은 멀티미디어 정보를 같이 표현하기에는 무리가 있다. 또한, 한정된 모바일 환경에서 동작하기에는 힘들기 때문에 서버와 같은 외부 하드웨어를 활용한다. 이 경우, 개인정보가 유출되는 위험성이 있다. 이를 개선하기 위해 본 논문에서는 일기를 문장 형태로 생성하였으며 하룻동안 사용자가 수집

1 Computer Engineering, Changwon National University, Uichang-Gu, Changwon-city, Kyungsangnam-do, 641-773, Korea
* Corresponding author (jcha@changwon.ac.kr)
[Received 26 April 2013, Reviewed 1 May 2013, Accepted 15 May 2013]

한 멀티미디어 정보도 함께 표시하여 기억을 회상하는 데 도움을 줄 수 있게 하였다. 또한, 개체명 인식과 주제를 기반으로 간단하면서도 모바일 환경에서도 사용 가능한 문장 생성 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 모바일 기기를 활용한 관련 연구와 이를 활용한 스토리 생성 방법에 대해서 살펴본다. 3장에서는 제안 방법에 대해 살펴보고 4장에서는 본 논문에 수집한 사용자 정보를 설명한다. 그리고 5장과 6장에서는 본 논문의 핵심에 해당하는 스토리 문장을 생성하는 방법에 대해 설명한다. 마지막으로 7장에서는 제안 방법의 효용성을 평가하고 8장에서 결론 및 앞으로의 연구 방향을 설명하고 본 논문을 마치도록 한다.

2. 관련 연구

2.1 모바일 기기 정보의 활용

모바일 기기는 이동성이 있고 사용자가 항상 지니고 다니기 때문에 사용자의 주변 환경과 행동 양상에 대한 매우 유용한 정보를 얻을 수 있다. 이러한 특성으로 모바일 기기의 로그 정보를 모을 수 있는 많은 방법이 연구되었다.

VTT 리서치 센터에서는 이러한 유용한 정보를 얻어 처리할 수 있는 체계적인 방법을 프레임워크로 제공하였다[3]. 이것은 수집한 정보를 로그 종류에 따라 가공하여 클라이언트 프로그램을 통해 이벤트 방식으로 서버에 전송한다. J. Mantyjarvi는 프레임워크와 퍼지 컨텍스트를 이용하여 수집한 정보를 모바일 기기의 UI에 반영하기도 했다[4]. M. Raento와 D. Siewiorek에 의해서도 로그 수집 방법이 연구되기도 했다[5, 6].

2.3 모바일 정보를 이용한 스토리 생성

2.1절과 같은 모바일 정보로부터 스토리를 생성하는 연구는 S. Cho[1]에 의해 처음으로 시도되었다. S. Cho는 해당 정보로부터 스토리를 만화 형식으로 구성하여 시간 순으로 생성하는 것이 주 목적이다. 이를 위해, 모바일 정보들의 모든 사건들 중 중심이 되는 사건인 랜드마크를 찾아낸다. 랜드마크를 찾기 위해, 사건과 사건의 의존 관계를 표현하여 개인 정보에서 랜드마크를 찾아낸 N. Eagle과 비슷한 방법을 이용했다[7]. 그리고 이 랜드마크와 연관된 사건들로부터 시간 순으로 스토리를 구성한다. 여기서의 스토리는 이들 사건에 적합한 그림 정보를 가지고 있는 스토리 DB를 활용한다. Y. Lee는 기본적으로 S. Cho의 방법을 페트리 넷에 적용하였으며[2], 그림 정보 대신 텍스트 문장 샘플을 모아놓은 것을 스토리 DB로 이용해 블로그의 글을 자동 생성하는 방법을 제안하였다[8]. 이러한 방법은 대용량의 스토리 DB와 문장 샘플이 필요로 하며 이들 정보가 많더라도 사용자가 한 일을 압축하여 표현하지만 세밀하게 표현하기에는 부족하다. 이러한 점을 개선하기 위해 우리는 모바일 기기에서 수집할 수 있는 텍스트 정보를 활용해 사용자가 실제 경험한 일에 대한 내용을 수집한다. 그리고 문장 예제가 아닌 패턴 형식으로 구축한 문장 템플릿을 활용하여 문장을 생성하는 방법을 제안한다.

3. 제안 방법

기존 방법은 베이지안 네트워크 혹은 페트리 넷에 적용하기 때문에 자원이 한정된 모바일 기기에서 스토리를 생성하는 데에는 무리가 있다. 그래서 본 논문에서는 기존의 방법과 달리 모바일 기기 자체로만 가능한 문장 생성 방법을 제안한다.

과정을 간략화 하기 위해 기존 방법에서 사용하는 랜드마크 추출 방법은 활용하지 않았다. 대신, 사건을 시간 순으로 나열하여 해당 사건들을 모았을 때, 동일한 주제를 가지는 집합인 에피소드를 찾는다. 그 후, 에피소드의 결과로부터 문장을 생성하는 방법을 제안한다. 생성한 문장을 스토리로 구성하기 위해, 하루의 중심이 되는 주제인 주토픽을 찾아 이와 관련된 에피소드로 구성하는 주제별 구성과 에피소드를 시간 순으로 나열해 구성하는 시간별 구성 방법을 제안한다. 이 방법은 기존 방법과 달리 스토리 생성 결과에 인과 관계가 존재하지 않을 수 있다. 그러나 일기는 하루의 중심이 되는 주제로만 정리하거나 시간 순으로 적는 것이 일반적이기 때문에 일기를 생성하는 목적에 부합한다고 볼 수 있다.

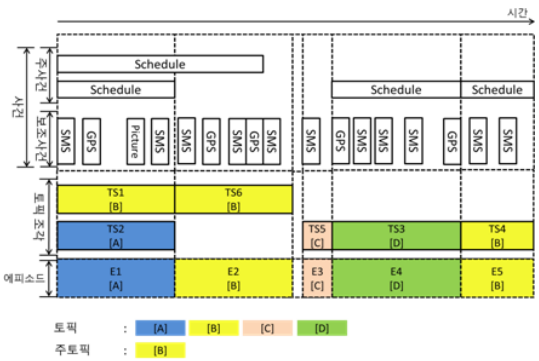
본 장에서는 우선 본 논문에서 사용하고 있는 각 용어에 대한 개념을 먼저 정의하고 제안 방법의 수행 과정을 설명한다.

3.1 용어 정리

용어는 S. Cho[1]의 것을 기초로 제안 방법에 맞게 용어를 재정의하였다. 각 용어의 정의는 (표 1)과 같다.

여기서 S. Cho와 가장 큰 차이점은 토픽 조각이 추가되었다는 것이다. 이는 특정 시간대에서 주제를 하나만 존재하는 것이 아니라 여러 주제가 존재할 수 있기 때문이다. 본 논문에서는 이 중 특정 시간대의 중심 주제인 최적 토픽을 선정해 이를 에피소드로 본다. 각 용어의 개념을 그림으로 정리한 결과는 (그림 1)과 같다. (그림 1)에서 대괄

호 "[]" 내에 있는 것이 토픽 정보이며 토픽 조각에서 "TS 숫자"로 구성되어 있는 것은 토픽 조각의 ID를 의미한다.



(그림 1) 각 용어의 기본 개념
(Figure 1) The notion of terms

(표 1) 용어 정의
(Table1) Definition of terms

용어	정의
로그	모바일 기기에서 수집 가능한 사용자 행동 양상에 대한 모든 정보
스토리	에피소드들의 집합, 하루의 요약
에피소드	토픽과 관련되는 사건들의 집합, 반드시 하나의 토픽 조각만을 가짐
토픽	특정 시간대의 중심이 되는 주제
주토픽	하루의 중심이 되는 주제
사건	사용자의 행동과 경험
주사건	토픽 결정에 결정적인 영향을 주는 사건
보조사건	사건 중 주사건이 될 수 없는 사건
토픽 조각	동일한 토픽으로 볼 수 있는 시간대

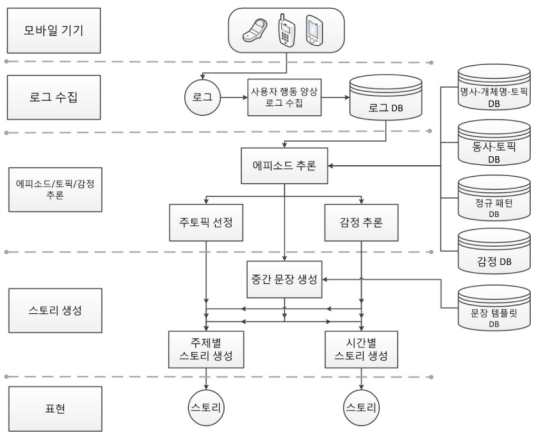
3.2 제안 방법의 수행 과정

본 논문에서 제안한 방법의 수행 과정은 (그림 2)와 같다. 이 과정은 크게 3 단계로 구성되어 있다. 첫 번째, 로깅 과정에서는 사용자의 기본 정보와 모바일 기기를 사용하면서 생성되는 사용자 행동 양상에 관한 모든 정보를 수집한다. 그 결과는 로그 DB에 저장된다. 그 후, 로그 DB의 결과를 이용하여 에피소드와 토픽, 감정을 추론한다. 먼저, 사건을 동일 시간대로 묶어 토픽 조각을 생성한 후 에피소드를 추론한다. 에피소드 추론 과정에서 문장 구성과 최적 토픽 계산에 영향을 줄 수 있는 자질, 감정과 관련된 자질을 각 에피소드 별로 추출한다. 추출한 자질로부터 오늘의 감정과 그날의 중심 주제인 주토픽을 선정한다. 세 번째 과정인 스토리 생성 과정에서는 각 에피소드들에 대해 문장 템플릿 DB를 이용하여 문장을 생성한다. 문장 생성 결과들을 주제별 혹은 시간에 따른 두 가지 형태로 스토리를 구성하여 그 결과를 사용자에게 보여준다.

4. 로그 수집

2장에서 언급했듯 로그는 모바일 기기에 종속적이다. 이러한 특성을 고려하여 본 논문에서는 모바일 기기로부터 수집할 수 있는 사용자 정보를 나열하여 토픽 생성 혹은 문장 생성에 필요한 정보를 분석했다. 그 결과, 모바일 기기에서 수집할 수 있는 사용자 행동 양상 정보는 (표 2)와 같다.

본 논문에서는 (표 2)의 로그 중 Schedule 정보를 주사건으로 이용했다. 그 외, SMS, Twitter, 사진은 보조 사건으로 활용했다. 주소록의 데이터는 사건으로 볼 수 없다. 하지만 보조 사건내의 인명 혹은 그룹 정보를 유추하는데 도움이 되므로 해당 자질을 추출하기 위한 보조 도구로 활용한다.



(그림 2) 제안 방법의 수행 과정
(Figure 2) A flowchart of proposed method

(표 2) 수집하는 로그의 종류
(Table 2) Collected Logs

분류	로그 종류	상세 정보
Personal	Personal	연령, 성별, 출생지
Social	SMS	받은 메시지
	Call Logs	모든 통화기록 (SMS 제외)
	Address Book	주소록(이름, 전화번호)
	Schedule	시간, 제목, 메모, 장소
Web	Twitter	자신이 작성한 글(시각, 내용, GPS, 주소)
	Weather	현재 위치의 날씨
Multimedia	Pictures	촬영한 사진(촬영시각, GPS, 주소, 촬영빈도)

5. 에피소드/토픽/감정 추론

로그 수집이 완료되었다면, 문장 생성에 필요한 에피소드와 토픽, 감정을 추출해야 한다. 본 장에서는 먼저 에피소드와 토픽, 감정을 추출하는 방법을 알아보고 이 과정과 함께 수행하는 자질 추출 과정에 대해 설명한다.

5.1 에피소드 및 토픽 추론

에피소드 추론 과정에서는 에피소드 후보가 될 수 있는 모든 토픽 조각들을 구성하고 이 결과로부터 에피소드들을 구한다. 이들 토픽 조각은 사건들로부터 생성된다. 이들 사건은 반드시 하나의 토픽 조각 내에 포함되어야 한다. 또한, 동일 시간대에 여러 토픽이 존재할 수 있으므로 하나의 사건이 여러 토픽 조각에 속할 수 있다고 본다. 이러한 조건들로부터 에피소드 추론은 아래와 같은 순서대로 진행된다.

1. 주사건들을 먼저 시간 순으로 나열한다. 그리고 이들 각각에 대해 토픽 조각을 만들고 여기에 속하는 주사건을 맵핑한다. (그림 1)에서 보면 TS1, TS2, TS3, TS4가 여기에 해당하는 예이다.
2. 보조사건들을 1에서 생성한 토픽 조각에 맵핑한다. 이 때, 맵핑될 수 있는 토픽 조각이 없다면 주사건의 시작 시간의 기점으로 1시간 크기 만큼의 모조(Dummy) 토픽 조각을 생성하고 보조 사건을 맵핑한다. 이는 (그림 1)에서 TS5에 해당한다.
3. 토픽 조각을 구성하는 과정에서 동일 시간대에 주사건이 여러 개가 있다면, 동일 시간대에 여러 개의 토픽이 있는 경우이므로, 길이가 짧은 것을 기준으로 자른다. (그림 1)에서 TS1은 TS6 과 하나의 토픽 조각이었으나 해당 시간대에 보다 시간대가 짧은 주사건(TS2 길이에 해당되는 주사건)이 있기 때문에 나뉘어진 것을 볼 수 있다.
4. 토픽 조각 중 토픽이 같으면서 한쪽의 끝시간과 시작시간이 같다면 이는 동일한 토픽이 이어지는 시간대이므로 합쳐준다. (그림 1)에서 TS6는 원래 주사건에서 생성된 토픽 조각과 모조 토픽 조각으로 구성되어 있었으나 본 과정에 의해 하나의 토픽 조각으로 합쳐진 예이다.
5. 1에서 4의 과정으로 모든 토픽 조각이 구성되었다면 각각의 토픽 조각 내에 속하는 사건들로부터 자질을 추출한다. 자질 추출 과정은 5.2절에서 상세히 설명한다.
6. 5에서 구한 자질들로부터 토픽 조각 내의 토픽을 결정한다. (그림 1)에서 토픽 조각 내에 [A], [B], [C], [D]가 이 때 결정된 토픽이다.
7. 각 시간대 별로 에피소드를 결정한다. 이는 동일한 시간대에서 토픽 조각이 여러 개가 있다면 4의 결과에서 결정된 토픽에 대한 토픽 스코어 값이 가장 높은 것을 해당 시간대의 에피소드로 본다. (그림 1)에서 TS2를 E1으로, TS6를 E2, TS5를 E3, TS3를 E4, TS4를 E5로 결정한 예를 보여주고 있다.

7의 과정의 에피소드는 수식 (1)을 이용하여 선정한다.

$$T^* = argmax_{t \in T} S_t \tag{1}$$

여기서 T는 토픽의 집합을 의미하며 S_t는 토픽 t에 관한 토픽 스코어를 의미한다. 이 토픽 스코어 S_t는 수식 (2)로 계산하며, 이 과정은 위의 과정 중 6번째 단계에서 수행한다.

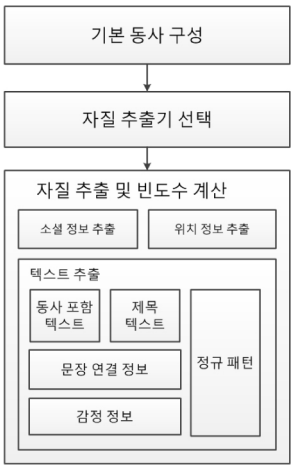
$$S_t = \Sigma_{x \in L} W_x \cdot \Phi(x, t)$$
$$\Phi(x, t) = \mathrm{P}(\mathrm{word}) + \mathrm{P}(\mathrm{exts}) \tag{2}$$

여기서 L은 토픽 조각 내 포함된 로그들을 의미한다. W_x는 해당 로그 종류에 대한 가중치이다. P(word)는 해당 토픽과 연관된 개체명 혹은 동사의 출현 빈도수를 전체 토픽 연관 단어의 출현 빈도수로 나눈 값이며, P(exts)는 문장 생성에 도움을 주지 않는 부가자질의 특성이 특정 토픽과 연관이 된다면 1, 아니면 0으로 부여한다. 부가 자질로는 GPS 값에 의한 거리 정보를 이용한다.

5.2 자질 추출

토픽과 감정, 문장 생성에 필요한 자질을 추출하는 단계이다. (그림 3)은 자질 추출의 세부 단계를 그림으로 표현한 것이다.

로그는 문장이 아닐 수가 있다. 그러나 이러한 로그는 종류에 따라 어떠한 행위를 했는가에 대한 동사를 유추할 수 있다. 예를 들어, GPS의 이동 거리 정보로부터는 이동 거리가 좁다면 걷는 것으로 간주할 수 있고 이동 거리가 매우 넓다면 여행을 한 것으로 간주할 수 있다. 또한, 문자 메시지와 같은 로그에서 텍스트 문장이 행위 정보가 생략된 불완전한 문장일 수 있다. 그러나 이 문장은 토픽이 결정된다면 대략적인 행위를 알 수 있다. 따라서, 이러한 종류의 로그는 기본동사를 정하지 않은 상태로 두고 토픽 선정 결과로부터 기본 동사를 결정한다.



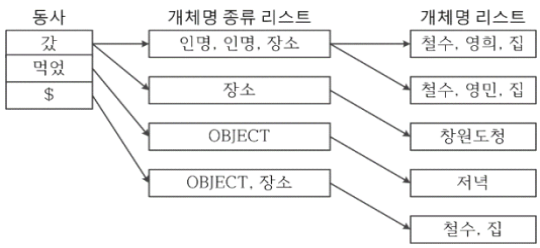
(그림 3) 자질 추출 과정
(Figure 3) A flowchart of feature extraction

자질 추출기 선택에서는 토픽 조각 내의 각 사건들에 대해 어떤 로그 종류인지를 판별하고 로그에 적절한 자질 추출기를 선택한다. 예를 들어, SMS는 동사를 포함한 텍스트 자질 추출기로 처리하라는 정보를 이 단계에서 알 수 있다. 자질 추출기가 선택되었으면 본격적으로 스토리 생성에 필요한 자질을 추출한다. 이 과정에서 추출하는 자질의 정보는 다음과 같다.

- 문장 구성과 관련된 자질
- 토픽 선정 시의 사용할 부가 자질
- 문장 연결 정보
- 로그 종류별 토픽 빈도 정보
- 감정 빈도 정보

문장 구성에 필요한 자질의 경우, (그림 4)와 같이 텍스트 내 문장에서 동사와 개체명을 쌍으로 추출한다. (그림 4)에서 "(갔)-(인명,인명,장소)-(철수,영희,집)"은 원 문장 "철수와 영희가 집에 갔다"로부터 추출된 것이다. 이러한 형식으로 저장하는 이유는 동사와 개체명 정보가 있다면 문장을 간단하게 요약할 수 있고 동일한 에피소드 내에 생길 수 있는 중복된 행위를 취하는 문장을 제거하기가 용이하기 때문이다.

동사 정보를 추출하기 위해, 본 논문에서는 동사 활용형 패턴 사전을 이용하였으며, 개체명을 추출하기 위해서는 한국어 개체명 인식기를 이용하였다. 개체명과 동사 정보에 토픽 연관 정보를 추가하여 결과를 반환할 때 해당



(그림 4) 문장 구성과 관련된 자질 구성의 예
(Figure4) A example of features construction

정보도 함께 전달해준다. 본 논문에서 사용하는 개체명 인식기는 인명, 단체명, 날짜, 시간, 위치, OBJECT(기타 개체명)을 개체명으로 인식한다.

이 외, 문자 메시지는 신용카드 내역과 같이 정형화 된 구조로 구성된 텍스트 정보도 있다. 이러한 정보는 사용자의 소비 내역을 추적하는데 유용한 정보로 활용할 수 있으므로 정규 패턴을 이용하여 추출한다.

추출하는 자질에는 빈도 정보도 있다. 토픽 빈도는 동사 및 개체명 추출시의 결과 중 토픽 연관 정보를 이용하여 토픽별 연관 단어의 출현 빈도를 센다. 감정 정보의 경우, 감정과 관련된 단어를 모아놓은 감정 DB를 이용하여 해당 단어의 출현 빈도를 센다.

본 논문에서 사용하는 문장 생성에 필요한 자질은 인과관계에 대한 정보가 없기 때문에 단문 형태가 반복되는 문장을 생성한다. 이러한 구조는 스토리 구성이 단조로울 수 있다. 따라서 이를 개선하기 위해 복문의 문장도 구성하도록 하였다. 이를 위해 텍스트 정보를 수집할 때 문장 연결 정보를 어미 패턴을 이용하여 추출한다. 본 논문에서 문장의 연결 종류와 어미 패턴은 (그림 5)와 같다.

연결 어미	연결 종류
고,며,고서, 다, 다가, …	합동형
거나, 든지, 든가, 나, …	분리형
나, 건만, 나마, 되, …	대립
므로, 더니, 나니, …	원인종속
면, 더라도, 거든, 았자, …	조건종속
려, 고저, …	목적/의도종속
고저, 면서, …	방식/정도종속

(그림 5) 문장 연결 패턴의 예
(Figure5) Examples of sentence connection patterns

5.3 주토픽 선정과 감정 추론

주토픽 선정에서는 에피소드에서 결정한 토픽으로부터

ID	문장 템플릿	필수 개체명
1	<날짜> <시간>에 <장소>에서 <OBJECT>을/전해줬	OBJECT
2	<날짜> <시간>에 <장소>에서 <OBJECT>을/건내줬	OBJECT
3	<단체명> <장소>로/집합했	장소
4	<날짜> <시간>에 <인명><이> <OBJECT><을>/통과했	OBJECT
5	<날짜> <시간>에 <장소>에서 <인명><와> <OBJECT><으로>/운동했	
6	휴일에 여자친구와 데이트 약속을 잡았다	
7	오늘은 화가 나는 일이 있었다	

(그림 6) 문장 템플릿 DB의 예
(Figure6) A example of DB of sentence template

하루의 중심이 되는 주토픽을 선정한다. 주토픽은 수식 (3)으로 계산된다.

$$T^{**} = argmax_{t \in T} \sum_{x \in EP} d \cdot S_t \quad (3)$$

여기서 d는 에피소드의 길이이다. 이 값은 에피소드가 끝나는 시간과 시작하는 시간의 차로 계산할 수 있다. EP는 에피소드 추론에서 추론된 전체 에피소드의 집합을 의미하며, S_t는 수식 (2)로 계산한 토픽 스코어 값이다.

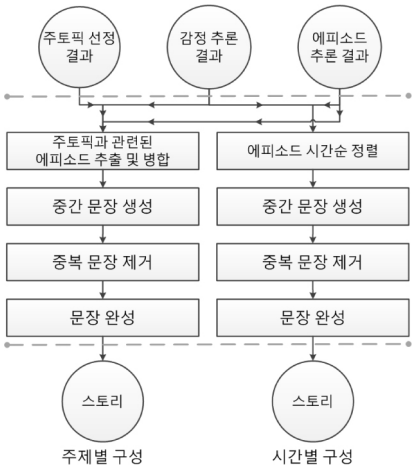
감정 추론은 에피소드의 자질로부터 그날의 감정을 추론한다. 우리는 감정을 무감정, 기쁨, 분노, 슬픔의 5가지로 분류하였으며, 수식 (4)를 이용하여 그날의 최적 감정을 추론한다.

$$E^{*} = argmax_{e \in E} \sum_{x \in EP} M_e \quad (4)$$

E는 감정의 분류이며, M_e는 에피소드에서 해당 감정과 연관된 단어의 빈도 정보이다. 만약, 여기서 슬픔 3번, 기쁨 3번, 분노 1번과 같이 감정 정보의 빈도수의 최대값이 동일한 것이 등장하는 경우, 최대값이 동일한 후보, 슬픔, 기쁨 모두를 그날의 감정으로 추론한다. 감정 정보의 빈도수가 모두 0이라면, 이 때의 감정은 무감정으로 추론한다.

6. 스토리 생성

이 장에서는 5장의 결과를 바탕으로 문장을 생성하고 생성한 문장을 스토리로 구성한다. 스토리를 구성하는 방법은 주토픽을 중심으로 한 주제별로 구성하는 방법과 시간대별로 일어난 일을 문장으로 요약하는 시간별 구성 방법이 있다. 세부 과정은 (그림 7)과 같다. 주제별 구성과 시간별 구성에서의 가장 큰 차이점은 주제별 구성에는 주토픽과 관련된 에피소드만 추출하여 이들을 병합하는 과정이 있다는 점이며, 시간별 구성에는 이 과정 대신 에피소드를 시간 순서대로 정렬하는 과정이 있다는 점이다.



(그림 7) 스토리 생성의 세부 과정
(Figure7) A flowchart of story generation

문장을 생성할 때 종결 어미를 어떤 형태로 구성할 것인지 혹은 복문으로 구성할 것인지 여러 조건을 판단해야 한다. 또한, 생성된 문장이 다른 생성된 문장에 의미적으로 포함되어 있는 문장인지도 판단해야 한다. 처음부터 하나의 완성된 문장으로 생성하면 이러한 처리가 힘들기 때문에 본 논문에서는 먼저 중간 문장을 생성한다.

중간 문장은 "동사를 제외한 문장/동사 활용형/동사" 형태로 구성된 문장을 말한다. 이 문장은 (그림 6)과 같은 문장 템플릿 DB와 문장 생성 관련 자질을 이용하여 생성한다. 만약, 이 두 정보를 이용해 그대로 문장을 생성할 경우, 일부는 텍스트에서 추출한 자질이므로 로그와 동일한 형태의 나오기 때문에 약간의 변화를 줄 필요가 있다. 이를 위해 동사 정보를 그대로 활용하지 않고 유의 동사를 활용

(표 3) 실험에 사용한 토픽 분류
(Table 3) Topic list

대분류	토픽
모든토픽에 포함	신용카드, 통화, 편지(이메일, 문자메시지 포함)
생활	종교, 금융, 계획/약속, 집안일, 다이어트, 아픔, 기념일, 특별한 휴일(개인적인 기념일)
대인관계	대인관계(회식, 외식 등 가족이나 친구, 직장 동료들과의 활동이면서 다른 토픽에 속하지 않는 것)
취미활동	운동, 독서, 낚시, 음악, 게임(컴퓨터 게임, 보드 게임만), 그림, 사진, 파티, 요리, 애완동물, TV시청
야외활동	여행(해외,국내여행,동물원,식물원), 관람(연극,영화, 뮤지컬, 콘서트), 쇼핑
학교생활	수업, 공부(혼자서 뭔가를 습득하는 활동), 체육대회, 축제
발표회	발표회(세미나, 강연, 학술회)
시험/고시	시험/자격증 (중간/기말고사, 고시, 자격증시험), 입시/입사(대학면접, 회사입사등)
직업	회의(주간회의, 월간회의, 업무관련회의), 출장(업무로 인해 외출하는 것), 업무(타 직업과 관련된 토픽과 연관없는 일상적으로 직장에서 일어나는 일)

한다. 또한, “운동장에서 건네주다”와 같이 자질 추출 시 필수격 정보가 없어 의미적으로 부정확한 문장이 될 수 있다. 이러한 현상을 방지하기 위해 문장 템플릿에 반드시 필요한 필수 개체명 정보를 둬 필수 개체명 정보가 없으면 해당 문장은 생성되지 않도록 하였다.

본 과정을 수행하면 에피소드 내의 모든 자질 리스트에 대해 문장이 생성된다. 그러나 위의 문장 생성 과정은 “오늘/시험쳤/시험치”라는 문장과 “오늘 학교에서/시험쳤/시험치”라는 문장과 같이 에피소드 내에서 동일한 의미를 내포하고 있는 중간 문장이 발생할 수 있다. 따라서, 이렇게 의미를 내포한 문장을 중복 문장 제거 작업을 통해 전자의 문장을 삭제한다.

그 후, 최종적으로 중간 문장을 완전한 문장으로 복원한다. 이 작업은 먼저 (1) 단문을 복문으로 구성할지를 판단하고 (2) 문장의 종결 어미를 어떻게 구성할 것인지를 선정하는 과정을 거친다. 만약, “오늘 학교에서/시험쳤/시험치”와 “철수와/농구했/농구하” 라는 문장이 있고 두 문장 간 연결 종류의 관계가 합동형으로 추출되었다고 하자. 그렇다면 이들 문장을 연결해 “오늘 학교에서 시험쳤고 철수와 농구했다”로 문장을 만든다. 그러나 이러한 문장 결합은 너무 많은 문장을 연결하면 사용자가 읽는데 부담을 주므로 최대 2문장만 연결하도록 하였다. 한국어는 종결 어미의 변화를 통해 좀 더 사용자의 연령, 지역, 성별에 따라 그 사람의 성향에 맞는 문장을 만들 수가 있다. 이를 위해 연령, 지역, 성별에 따라 종결 어미를 분류한 정보로부터 사용자의 성향에 맞는 문장을 생성하도록 했다.

스토리 생성 결과, 문장수가 너무 작다면 대체로 한 곳에 머물러 행동하거나 모바일 기기로 정보를 수집할 수 없는 활동을 한 경우이다. 이 경우, 본 제안 방법으로는 문장을 생성할 수 없다. 그래서 오늘이 휴일인지 계절은 어떠한 지 혹은 감정 추론 결과를 이용하여 기본 문장을 출력하도록 하였다. 이 기본 문장은 사용자의 행동과 토픽과는 무관한 문장으로 구성하였다.

7. 실험 및 평가

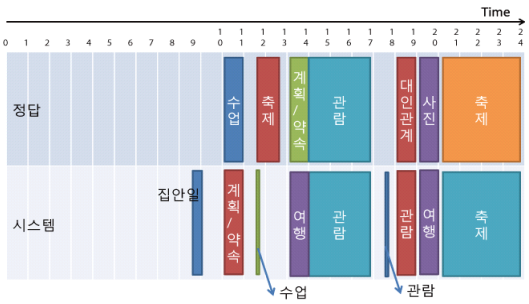
7.1 실험 환경

본 논문에서는 실험에 필요한 로그 정보를 모으기 위해 하룻동안 총 3명의 실험자가 직접 스마트폰을 사용하면서 얻은 정보를 바탕으로 총 5개의 시나리오를 작성했다. 그리고 토픽 분류를 일반화시키기 위해 실험자와 함께 오랜 기간에 걸쳐 일상에 있을 수 있는 토픽을 총 37개로 정하였다. 그 결과는 (표 3)과 같다. 이 토픽 정보를 바탕으로 명사-개체명 DB와 동사 DB의 단어에 대해 연관 토픽 정보를 붙였다. 토픽 연관 단어는 주제별로 잘 분류되어 있는 DC 인사이드 갤러리 게시판*의 게시물을 대상으로 수집한 문서 총 2,524,483개를 이용하였다. 수집한 결과를 총 9명이 수작업으로 정제하여 토픽 연관 단어 정보를 얻었다. 이렇게 반자동으로 토픽 연관 단어를 학습한 토픽은 16개이다. 나머지 21개의 토픽은 수작업으로 총 50여개의 토픽 연관 단어 목록을 만들어 해결하였다. 실험에 사용한 토픽 연관 단어의 개수는 반자동으로 학습한 토픽과 수작업으로 학습한 개수를 동일하게 맞추었다. 분류한 토픽 중 신용카드, 통화, 편지와 같은 토픽은 모든 토픽과 동시에 일어날 수 있다. 이러한 토픽은 다른 토픽에 속하더라도 상관 없는 것으로 간주하였다.

7.2 평가 자질

실험을 평가하기 위해 우리는 시스템이 유추한 토픽이

* http://gall.dcinside.com



(그림 8) 시나니오 S5의 에피소드 및 토픽 생성 결과
(Figure 8) A result of generation of episodes and topics for scenario S5

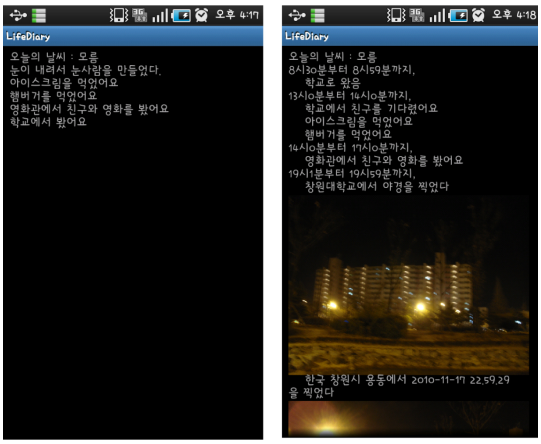
얼마나 정확한지 알아보는 척도가 되는 토픽 정확도와 해당 에피소드에서 생성된 문장이 실제 사건과 얼마나 부합되는지를 평가하기 위해 문장 정확도를 이용하여 평가했다. 토픽 정확도의 수식은 수식 (5)와 같다.

$$\text{토픽정확도} = \frac{\textit{시스템이 맞은 토픽의 개수}}{\textit{전체 토픽의 개수}} \quad (5)$$

여기서 전체 토픽의 개수는 시스템과 정답에서 등장한 에피소드 두 결과를 종합하여 구한다. 이 때, 정답과 시스템에서 동시에 에피소드가 있을 때, 동일한 시간대의 에피소드가 아니라면 해당 에피소드를 작은 쪽을 기준으로 자른다. 예를 들어, (그림 8)의 경우 이 값은 10이 된다. 그리고 시스템이 맞은 토픽의 개수를 셀 때 정답에서 에피소드가 없으나 제안 시스템에서 에피소드가 있다고 하면 이 때의 시스템의 결과는 무조건 맞는 것으로 평가한다. 그 이유는 우리가 정의한 에피소드의 토픽이 모든 토픽에 해당될 수 있는 토픽인 신용카드, 통화, 편지로만 구성된 에피소드를 나머지 34개의 토픽에서도 처리하기 때문이다. 문장 정확도의 수식은 수식 (6)과 같다.

$$\text{문장정확도} = \frac{\textit{해당 시간대에 실제 실현된 문장의 개수}}{\textit{시스템이 생성한 전체 문장의 개수}} \quad (6)$$

여기서 시스템이 생성한 전체 문장의 개수는 멀티미디어 정보와 시간 정보를 제외하고 시스템이 생성해낸 전체 문장의 개수를 말한다. 그리고 에피소드의 문장이 실제로 해당 시간대에서 있었던 일인지를 고려해 해당 시간대에 실제 실현된 문장의 개수를 센다.



(그림 9) 시나리오 S5의 출력 결과 (왼쪽 : 주제별 구성, 오른쪽 : 시간별 구성)
(Figure 9) A result of scenario S5 (Left : story by main topic, Right : story by time)

(표 4) 실험 결과
(Table 4) A result of the experiment

시나리오	토픽 정확도	시간별 문장 정확도	주제별 문장 정확도
S1	72.73%	57.14%	100.00%
S2	50.00%	50.00%	75.00%
S3	61.54%	75.00%	0.00%
S4	100.00%	100.00%	100.00%
S5	76.47%	100.00%	100.00%
평균	72.15%	76.43%	75.00%

7.3 결과

본 논문의 실험 결과는 (표 4)과 같다. 해당 정보는 시나리오에 대한 실제 로그를 수집한 실험자가 1차적으로 평가한 후, 그 결과를 제 3자가 재평가 해 결과를 얻었다. 이 때, 문장 정확도는 시간별, 토픽별로 각각 그 정확도를 평가했다.

(표 4)에서 S3의 문장 생성 결과가 0.00%가 나온 것은 그 날 주토픽은 제대로 유추했으나 주토픽 내의 문장 생성 결과가 없어 생긴 문제이다. 이 이유는 S3의 시나리오의 하루 일과는 대부분 시험치는 일인데 시험 시간에는 모바일 기기를 사용하지 않기 때문에 수집할 수 있는 정보가 없기 때문이다. 또한, 자질 추출에서 띄어쓰기 문제로 인해 개체명 인식 결과가 제대로 나오지 않아 문장 생성에 필요한 중요한 개체명을 잡아 내지 못하였기 때문이다. 하지만 본 실험 결과는 토픽과 이와 관련된 문장이 만족할 수준은

아니지만 예상 외로 괜찮은 결과를 얻을 수 있었다.

그러나 에피소드 결정에 결정적인 단서가 되는 주사건은 에피소드의 시간대를 결정하는 결정적인 역할을 하지만 악영향을 줄 수도 있다. 예를 들어, 사용자가 하루 전체의 Schedule을 쇼핑으로 잡았을 때의 경우를 보면 이 중 대다수가 그에 관한 문장을 만든다. 이는 주제별 관점에서는 해당 주제에 맞는 문장을 생성해내므로 좋은 역할을 하나 특정 시간대에 주제와 연관 없지만 중요하다고 보여지는 일에 대한 행위의 문장이 사라질 수도 있다.

그리고 스팸 정보에 대한 문제이다. 이것은 우리가 모바일 기기로부터 얻을 수 있는 토픽과 문장 생성에 연관되는 자질 정보 중 사용자가 직접 보내는 문자 메시지를 활용할 수 없어 받는 문자 메시지를 자질 추출에 활용했기 때문이다. 만약, 특정 시간대에 의미 있는 로그 정보가 많이 나타난다면 이들 받는 문자메시지도 도움을 준다. 그러나 의미 있는 로그 정보보다 많은 양의 스팸이 나타나게 되면 해당 토픽을 결정하는데 악영향을 준다. 또한, 실제 사용자의 행동이 아닌 받는 사람의 행동 양상이 사용자가 했던 행동으로 보고 문장을 생성하는 문제가 있다.

8. 결 론

본 논문에서는 모바일 환경에서 수집할 수 있는 로그 정보로부터 그 날의 중심이 되는 토픽을 선정하고 그날의 사용자 행동 양상을 일기 형태의 문장을 생성하는 방법을 시도했다. 이를 위해 주사건으로부터 에피소드를 결정하고 이들의 토픽을 결정하였다. 그리고 그 결과로부터 그 날의 중심이 되는 주토픽을 선정하였다. 이 정보를 바탕으로 그날의 주토픽을 기준으로 스토리를 생성하는 주제별 구성과 에피소드를 시간 순으로 나열해 문장을 구성하는 시간별 구성 방법을 제안했다. 그 결과, 기존 시스템보다 간단하면서 각 에피소드에 대해 만족할 만한 토픽 선정 결과를 얻을 수 있었다.

그러나 결과 분석에서와 같이 에피소드 시간대를 결정하는데 주사건을 맹목적으로 고려하는 문제, 모바일 데이터에서 관련 토픽과 문장 생성과 관련된 자질이 매우 한정적인 문제를 볼 수 있다. 본 연구에서 사용한 모바일 기기에서는 하드웨어 정책 상 정보 수집과 관련된 많은 부분이 제한되어 있어 데이터가 부족해 위와 같은 문제를 개선하지 못했다. 향후에는 모바일 기기의 수집 데이터를 확충하여 본 방법에서 제시한 문제점을 개선할 수 있는 방향으로 연구를 진행하고자 한다.

그리고 실험을 보다 더 확장하여 하루 단위가 아닌 장기적인 관점에서 해당 정보를 수집했을 때 얼마만큼 정확한 사용자 행동 양상 정보를 얻어낼 수 있을지에 대한 연구가 필요하다. 또한, 최근에 많은 스팸 정보가 모바일에 축적되는데 여기에 대한 효과적인 필터링 작업을 수행할 때, 본 연구 방법이 얼마만큼 효과가 있는지에 대한 연구도 진행할 계획이다.

참 고 문 헌(Reference)

[1] S. Cho, K. Kim, K. Hwang, I. Song, “AniDiary: Daily Cartoon-Style Diary Exploits Bayesian Networks,” *Journal of IEEE Pervasive Computing*, vol.6, No.3, pp.66-75, 2007.

[2] Y. Lee, S. Cho, “Petri Net-Based Episode Detection and Story Generation from Ubiquitous Life Log,” *Proceedings of the 5th international conference on Ubiquitous Intelligence and Computing(UIC’08)*, pp.158-168, 2008.

[3] P. Korpipää, Jani Mänityjärvi, J. Kela, H. Keränen, E.J. Malm, “Managing context information in mobile devices,” *Journal of IEEE Pervasive Computing*, vol.2, No.3, pp.42-51, 2003.

[4] J. Mäntyjärvi, T. Seppänen, “Adapting applications in handheld devices using fezzy context information,” *Interacting with Computers*, vol.15, No.4, pp.521-538, 2003.

[5] M. Raento, A. Oulasvirata, R. Petit, and H. Toivonen, “ContextPhone: A Prototyping Platform for Context-Aware Mobile Applications,” Journal of IEEE Pervasive Computing, vol.4, No.2, pp.51-59, 2005.

[6] D. Siewiorek, A. Smailagic, J. Furukawa, N. Moraveji, K. Reiger, and J. Shaffer, “SenSay: A Context-Aware Mobile Phone,” *Proceeding of the 7th IEEE International Symposiumon Wearable Computers*, pp.248-249, 2003.

[7] N. Eagle and A. Pentland, “Social Serendipity: Mobilizing Social Software,” *Journal of IEEE Pervasive Computing*, vol.4, No.2, pp.28-34, 2005.

[8] Y. Lee, S. Cho, “Automatic weblog generation form mobile context using Bayesian network and Petri net,” *Journal of KIISE: Computing Practices and Letters*, vol.16, No.4, pp. 467-471, 2010. (in Korean)

◐ 저 자 소 개 ◑

홍 진 표

2003년 창원대학교 컴퓨터공학과 졸업(학사)
2007년 창원대학교 대학원 컴퓨터공학과 졸업(석사)
2009년~현재 창원대학교 대학원 컴퓨터공학과 재학(박사)
관심분야 : 자연어 처리, 기계학습
E-mail : virnmer@changwon.ac.kr

차 정 원

1996년 숭실대학교 컴퓨터공학과 졸업(학사)
1999년 포항공과대학교 대학원 컴퓨터공학과 졸업(석사)
2002년 포항공과대학교 대학원 컴퓨터공학과 졸업(박사)
2002년~2003년 USC/ISI (박사후연수)
2004년~현재 창원대학교 컴퓨터공학과 교수
관심분야 : 자연어 처리, 기계학습, 정보검색
E-mail : jcha@changwon.ac.kr