단문 메시지 서비스의 준자동 응답을 위한 비지도학습 및 추론 방법

최 봉환 ○, 조 성배
연세대학교 컴퓨터과학과
bitbyte@sclab.yonsei.ac.kr○, sbcho@cs.yonsei.ac.kr

Unsupervised Learning and Inference Method for Semi-Autonomatic SMS Reply

BongWhan Choe○, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

모바일 상의 단문메시지 서비스는 등장한 이래 꾸준히 사용량이 증가하는 추세이며, 현재 세계적으로 가장 많이 사용되는 모바일 서비스이다. 모바일 기기에서 단문 메시지 작성의 불편함을 개선하기 위한 기술로 하드웨어적인 입력 방법 개선과 소프트웨어적인 입력보조 기능이 꾸준히 개발되었다. 소프트웨어적인 방법은 범용성이 넓고 적용이 쉽다는 장점이 있지만 제한된 자원에서 구현상의 어려움이 있어 연구가 미비한 분야이다. 본 논문은 소프트웨어적으로 단문 메시지의 작성을 보조하는 방법을 제시한다. 일상 생활의 반복성에 초점을 맞추어 반복 작성될 메시지에 대해 기존의 메시지를 제시해 자동적으로 응답하도록 하는 방법을 제안한다. 자동적으로 응답 메시지를 선택하기 위한 비교사 학습과 추론 기술로 "메시지 네트워크"를 제안하고, 실험을 통해 고안한 방법의 가능성을 보였다. 실험 결과로부터 반복적인 메시지의 작성에 제시한 방법이 유용함을 알 수 있었다.

1. 서 론

마케팅인사이드가 2005년도에 실시한 “무선인터넷 및 부가 서비스 이용실태에 대한 조사”에 의하면 2005년 3월 시점에서 단문 메시지 서비스(Short Message Service, SMS)가 가장 많이 이용되는 서비스였으며 월평균 31건에서 100건 사이가 전체 응답자의 약 30%에 해당할 정도로 이용 빈도가 높은 서비스로 조사 되었다[1]. 개인의 단문메시지 서비스의 대부분은 친밀한 관계에서 상호간 의사 교환이나, 신속한 상태 정보 전달을 목적으로 한다는 것을 알 수 있다[2].

이 논문에서는 SMS 작성의 불편함을 개선하기 위해 일상의 반복성에 초점을 맞춘 메시지 네트워크를 제시한다. 메시지 네트워크는 비지도 학습에 기반을 둔 방법으로 일상의 반복성 및 상호 작용의 정형성을 컴퓨터 프로그램으로 반영한 방법이다.

2. 단문 메시지 작성의 불편함

대부분의 모바일 장치에서 단문 메시지는 일반 키보드와 달리 12개의 숫자키와 몇가지의 기능키를 사용해 입력을 하도록 되어 있다. 따라서 메시지 작성 속도가 컴퓨터의 키보드에 비하면 매우 느리다. 따라서 단문 메시지의 작성에서는 적은 수의 글자로 많은 것을 표현하는 것이 중요한 문제가 된다.

이러한 불편함을 개선하기위한 기술은 지속적으로 제시되고 있으나, 키맵을 변경하여 입력 횟수를 감소시키려는 시도와 같이 하드웨어적으로 입력 횟수를 줄이는 방법에 치중되어 있다[3]. 소프트웨어적인 방식으로도 Amasasoft社의 quickta처럼 소프트웨어로 입력을 언어에 따라 최적화 하는 것으로 입력 시간을 단축하려는 노력도 있다[4].

소프트웨어적인 방법으로 LG전자의 ‘인텔리전트 SMS (싸이언 와인폰, LG-SV300/SV3000의 나만의 상용구)’ 기능은 조금 다른 관점에서 접근했다[5]. '인텔리전트 SMS'는 사용자가 유사한 문장을 다시 사용한다는 점에서 착안된 기술로, 기존의 사용자가 보낸 문장과 동일하거나 상당히 유사한 문장을 다시 작성하지 않도록 도와준다.

'인텔리전트 SMS'는 단순한 글자 비교를 통해 작성할 문장을 제시하므로 연속성이 없고, 문자열 비교라는 형태 비교를 통해서 응답 메시지를 제시하기 때문에 오류 발생의 가능성이 높아진다. 또한 우선 순위를 시간순이나 문장순으로 정렬하기 때문에 처음 부분이 동일한 메시지가 많아 질수록 적절한 메시지가 추천에 포함되어 있다 하여도 사용자에 의한 선택까지 많은 시간이 소요될 가능성이 높아진다.

3. 메시지 네트워크와 통계 기반 키워드 선택

언어 처리 기술은 데이타마이닝과 자연언어처리 분야에서 많이 연구 되었다. 데이타마이닝 분야에서 진행된 문서 처리 연구는 대량의 데이터로부터 소량의 의미있는 자료를 추출하기 위한 연구이다. 특히 TF-IDF는 다수의 문서로부터 각 문서의 차별성을 나타내는 단어를 통계적 기법으로 찾아내어 문서를 분류 하거나, 주제에 따른 내용요약 등을 위해 사용된다[6].

자연언어처리는 문법에 기반을 두고, 문법적인 분석을

통해 문장을 이해, 처리하는 방식으로 번역 서비스나, 정보 추출과 관련된 기술을 연구한다. 그러나 단문 메시지는 제한된 길이의 특성상 유행어, 축약어, 은어, 이모티콘 등, 비 문법적인 요소가 많고, 변화가 심하기 때문에 두가지 모두 적용이 힘들다[7][8].

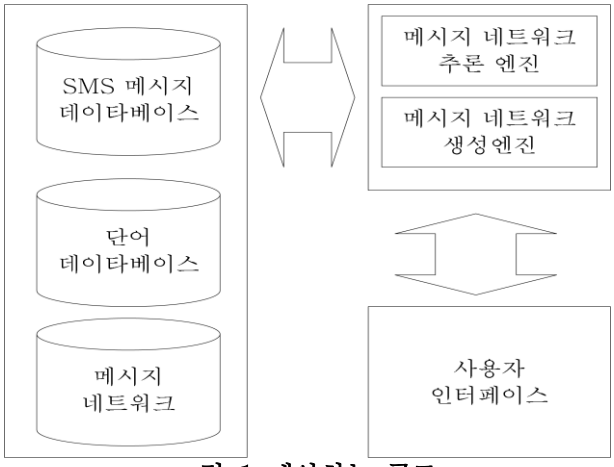

그림 1 제안하는 구조

기존의 자연언어 처리의 기법은 제한적인 모바일 환경 및 단문 메시지 서비스의 특수성으로 인해 적용이 힘들다. 그러나 TF-IDF나, k-means와같은 데이타 기반 기법을 적용할 수 있다.

본 논문에서는 기본적인 언어 형태분석 후, 통계기법을 통해 키워드 추출하고, 클러스터링 기법을 통해 네트워크를 구성하여 구성된 자료를 기반으로 다음 메시지를 추천하는 기술을 개발하였다. 본 논문에서는 통계기반 키워드 추출, 데이타 클러스터링, 그리고 형태소 분석 기술을 사용해 사용자의 송/수신 메시지를 비지도 학습하는 "메시지네트워크"를 생성하고, "메시지네트워크"에서 응답 메시지를 추천하기 위한 방법을 제안 한다.

3.1 메시지네트워크

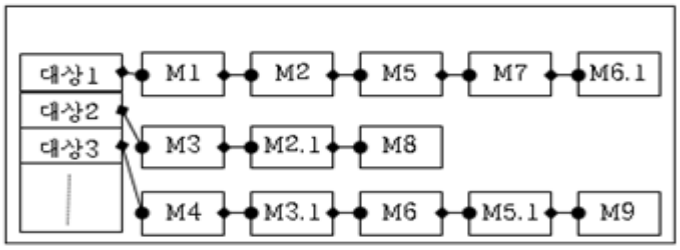

그림 2 메시지 전후 관계 예시

"메시지네트워크(Message Network)"는 메시지와 메시지 사이의 인과 관계를 네트워크의 형태로 표현한 자료구조이다. 반복되는 일상에 있어서 반복되는 유사한 내용의 메시지로부터 응답패턴을 학습하고, 학습된 응답 패턴을 사용해 사용자의 응답메시지를 제시한다. 따라서 2회 이상의 동일하거나 유사한 내용의 메시지를 송/수신한 경우 응답 메시지를 추론하여 제시한다.

일반적으로 동일한 대상과 주고받은 일련의 메시지를 묶으면 하나의 주제를 가진 대화가 된다. 따라서 메시지의 순차적 나열로부터 대상 및 전/후 관계에 따른 구조를 표현하면 그림 2와 같다.

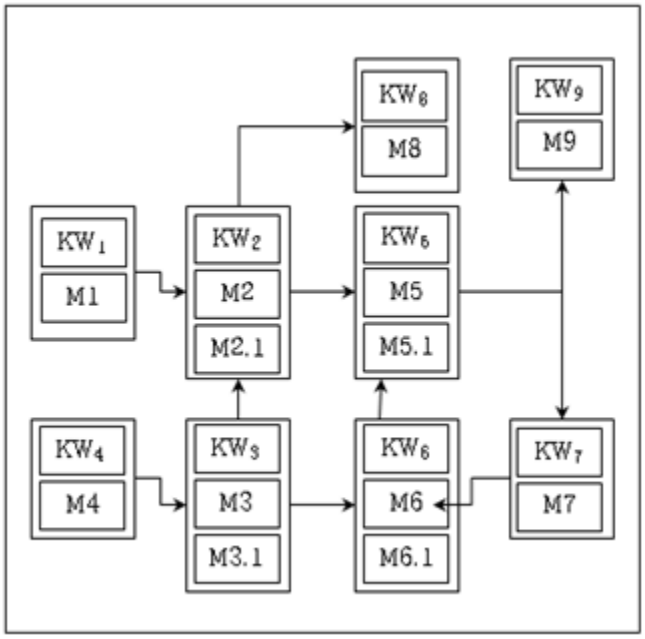

그림 3 그림 2에 대응하는 메시지 네트워크 예시

이렇게 구성된 연관된 메시지의 나열을 다시 유사한 메시지끼리 묶고, 기존의 순서 관계를 지향성 연결선으로 구성하면 그림 3과 같은 네트워크의 형태를 띄게된다. 이렇게 구성된 네트워크 구조를 "메시지네트워크"로 정의 한다.

3.2 통계기반 키워드 선택

일상적인 문장은 “너", "나", "~은” 와 같이 모든 내용에 반복되는 요소가 있으며, “신촌"과 같이 특정한 경우에만 사용되는 특수한 단어도 있다. 통계적으로 모든 문장에 포함되면 문장과 문장의 변별력이 떨어지며, 반대로 너무 작은 빈도로 사용되는 단어는 동일한 목적으로 다시 사용될 가능성이 매우 낮다. 따라서 모든 단어를 동일한 가중치로 사용한 문장 비교는 오류를 내포하게 된다. 따라서 메시지 네트워크를 구성하기 위해 각 메시지의 유사성을 판단하는 기준으로 모든 단어를 사용하지 않고, 핵심이 되는 키워드 만을 골라낼 필요가 있다.

본 논문에서는 발생빈도를 계산해 중간발생 빈도를 가지는 단어만을 사용하는 방법으로 이러한 문제를 풀었다. 즉, 발생 빈도가 매우 높은 단어는 유사성을 비교하는 기준이 되기 힘들고, 반대로 매우 적게 발생되는 단어는 특수성을 가졌다고 보고, 전체 입력된 메시지의 단어의 발생 빈도를 계산하여, 발생 빈도가 상위 20%에서 하위 20%까지를 키워드로 사용했다.

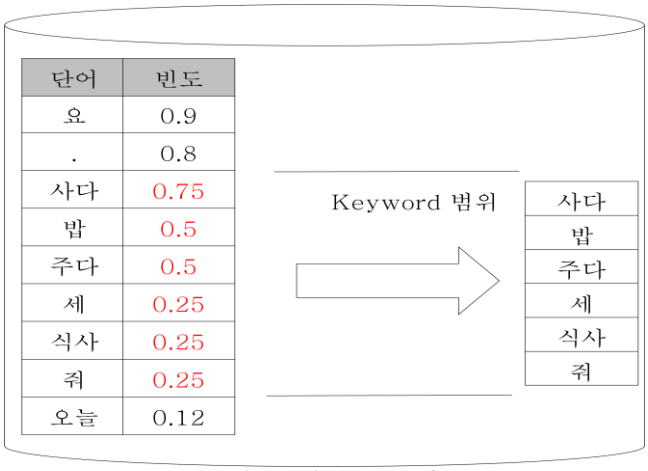

그림 4 키워드 결정

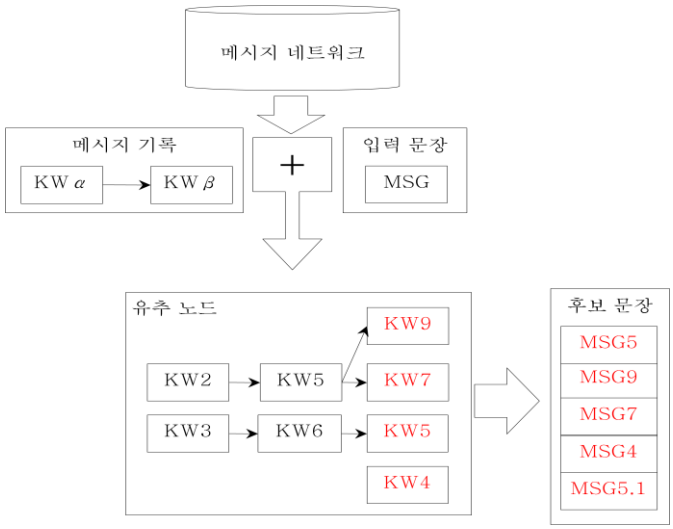

그림 5 대화패턴 네트워크의 동작

4. 메시지 네트워크의 생성 및 추론

4.1 메시지 유사성

메시지네트워크를 생성하거나 사용해 추천을 하기 위해서는 문장의 유사성을 수치화하여 표현할 필요가 있다. 본 논문에서 메시지 유사성 $S(msg_i, msg_j)$은 두 메시지의 키워드의 유사성을 통해 계산하였다. 두 메시지의 전체 키워드 중에 동일하거나 유사한 키워드의 개수를 비교하였다. 즉, 키워드의 키워드 가중치를 $fq(k)$라고 정의하고, 메시지의 유사도 함수 $S(msg_i, msg_j)$를 다음 수식 (1)과 같이 정의한다.

$$S(msg_i, msg_j) = \frac{\{\sum fq(k_c)\} \times 2}{count(keyword(msg_i)) + count(keyword(msg_j)) \{k_c \mid k_c \in \{keyword(msg_i) \cap keyword(msg_j)\}\}}, \quad (1)$$

note : $fq(k)$는 전체 메시지 데이타중 해당 키워드 k의 발생 빈도를 의미하다.

이때, 문장의 길이가 키워드 비교시 주는 영향을 줄이기 위해 문장에 포함된 단어의 총 개수로 정확도를 나누는 것으로 수치를 정규화 하였다.

4.2. 메시지 네트워크를 사용한 추론

메시지 네트워크를 사용한 추론은 그림 5와 같은 과정을 거치게 된다.

가) 추론하기 위한 현재 대상의 기록과 현재 입력중인 문장을 단어 단위로 분할한다.
나) 분할된 단어에서 키워드를 구분해 낸다.
다) 현재 메시지를 수신할 대상의 마지막 기록을 기준으로 $S(msgi, msgj)$를 사용해 기록 가중치 w_c를 계산한다.
라) 입력중인 메시지와 $S(msgi, msgj)$를 사용해 메시지 가중 치 wm을 계산한다.
마) $w(msg, context) = w_c * \omega * w_e + w_m$, $1 > \omega > 0$, w_e = Weight of edge 으로 계산된 가중치 순으로 상위 n개를 출력한다.

가)의 과정에서는 강승식[9]의 한글 형태소 분석기 2.1의 데모버전을 사용하였으며, 그 밖의 과정은 자체 제작된 모듈을 통해 수행하였다.

4.3 메시지 네트워크의 생성

메시지 네트워크는 메시지의 송/수신 시점에서 다음과 같은 과정을 거쳐 구성한다.

가) 메시지를 단어 단위로 분할한다.
나) 분할된 단어를 포함하여 단어 개수를 갱신한다.
다) 기존 네트워크의 키워드를 갱신한다.
라) 새로 입력된 키워드를 사용해 가장 유사한 노드를 찾는다.
1) 가장 유사한 노드의 유사성이 임계치 ε을 넘으면, 해당 노드로 묶고 마)단계를 수행한다.
2) 가장 유사한 노드의 유사성이 임계치 ε을 넘지 못하면 새로운 노드를 생성한다.
마) 기존의 노드에 포함된 경우 노드의 연관성을 갱신한다. 만약 기존에 이미 동일한 노드를 참조하고 있었다면 가중치를 상향 조정한다.

5. 실험 및 결과 분석

5.1 실험용 데이터의 구성

모든 메시지는 띄어쓰기 맞춤법이 정확하다는 가정하에 실험용 메시지 데이터를 구성하였다. 이는 한글 형태소 분석기가 부 정확한 메시지에 대한 처리가 정상 적으로 이루어지지 않았기 때문이다.

실험을 위한 데이터는 시나리오 기반으로 작성되었다. 대상은 평범한 남자 대학생의 학기중 생활 패턴을 기준으로 잡았다. 실험 데이타의 예는 표1과 같다.

5.2 결과 분석

표 1과 같이 구성한 실험용 데이터를 미리 입력한 후 특정 단어로부터 의도한 주제의 메시지를 선택하게 되는지 실험하였다. 표 2와 같이 초기에 입력한 단어에 따라서 두가지 상반된 실험 결과가 나왔다.

좋은 결과를 도출한 경우로 "밥"이라는 키워드에 대해서는 식사관련 메시지를 정상적으로 추천하는 것을 볼 수 있다. 따라서 메시지 네트워크를 사용할 경우 사전 지식 없이 학습된 소량의 데이터로부터 다양한 응답이 추천 가능함을 알 수 있다.

표 1 실험용 메시지 데이타 샘플

수신	후배	형 밥 먹었어요?
송신	후배	아직 안 먹었어. 같이 먹을래?
수신	후배	예. 사주실 거죠?^^
송신	후배	어,대신 학관으로 와라.
수신	후배	ㅎㅎ 지금 가요.
수신	후배	엇, 형 어디세요?
송신	후배	중도, 곧 갈께
수신	친구1	어~이 밥 먹자.
송신	친구1	이미 먹고 있는데 ㅎㅎ
수신	친구1	어딘데? 같이 먹자.
송신	친구1	학관 올려면 빨리와.
수신	친구1	어 지금 간다. 기다려.
수신	친구1	도착.. 어디냐?
송신	친구1	입구 근처 빨리와.
수신	친구2	오후 7시에 숙제 모임 할껀데 시간돼?
송신	친구2	어, 약속 없다.
송신	친구2	그럼 오후 7시에 연락할께.
송신	친구1	어디냐?
수신	친구1	음. 밥먹고 있다. 어디서 모일껀데?
송신	친구1	중도로 와. 기다릴께.
수신	친구1	조금 늦게 갈께.
수신	후배	같이 밥먹어요.
송신	후배	지금 숙제중이야
수신	후배	밥사달라고 하려고 했더니.
수신	후배	형 밥먹어요.
송신	후배	수업 중 이었다. 이제 먹자.
수신	후배	이미 먹고 있어요.

6. 결론 및 향후 과제

본 논문에서 진행한 실험은 시나리오를 기반으로 실시된 실험으로, 수치적인 효용성을 결론 짓기는 힘들다. 다만, 결과로부터 통계적인 기법을 사용해 SMS로부터 메시지를 추천하는 방식이 가능성이 있음을 확인 할 수 있었다. 정확한 증명을 실험데이터의 추가 후 재실험이 요구된다.

이 방법이 실제적으로 사용하기 위해서는 (1) 단문 메시지 단어 분석 기술, (2) 단문 메시지를 위한 키워드 추출 기술, (3) 메시지 네트워크의 갱신 알고리즘 개선 등 많은 과제가 남아 있다. 또한 외적 정보를 수집, 처리함으로써 추천 메시지의 정확도를 향상시킬 필요가 있다.

표 2 실험 결과

입력문장	대상	추천 메시지
밥	친구1	학관 올려면 빨리와
		이미 먹고 있는데 ㅎㅎ
		어 지금 간다. 기다려.
	친구2	어 지금 간다. 기다려.
		이미 먹고 있는데 ㅎㅎ
		학관 올려면 빨리와
		어~이 밥 먹자.
숙제	친구1	오후 7시에 숙제 모임 할껀데 시간돼?
		학관 올려면 빨리와.
		수업 중 이었다. 먹자.
		이미 먹고 있는데 ㅎㅎ
	형	오후 7시에 숙제 모임 할껀데 시간돼?
		수업 중 이었다. 먹자.

감사의 글

한국과학기술원의 위탁과제에 의해서 일부 지원받았음

참고 문헌

[1] (주)마케팅 인사이드, "[텔레콤 리포트 24호] 무선인터넷/부가서비스 중 가장 많이 쓰는 것은 문자 메시지," *http://www.mktinsight.co.kr/*, 2008.

[2] (주)마케팅 인사이드, "이동통신 TELECOM / SMS 이용 현황," *http://www.mktinsight.co.kr/*, 2007.

[3] K. Sörensen, "Multi-objective optimization of mobile phone keymaps for typing messages using a word list," *European Journal of Operational Research*, vol.179, no.3, pp.838-846, 2007.

[4] amasasoft社,"Quickta(퀵타)제품 설명," *http://amasasoft.kr.ecplaza.net/1.asp*.

[5] LG전자, "LG-SV300제품 특징, Intelligent SMS," *http://www.cyon.co.kr/*.

[6] G.Salton, C.Buckley, "Term-weighting approaches in automatic text retrieval.," *Information Processing & Management*, vol.24, no.5, pp.513-523, 1988.

[7] *http://www.world-english.org/SMS.htm*, 2002.

[8] 강승식,장두성, "SMS 변형된 문자열의 자동 오류 교정 시스템," *한국정보과학회 2007 한국컴퓨터종합학술대회 논문집,* vol.34, no.1, pp.59-60, 2007.

[9] 강승식, "한국어 형태소 분석기 (KLT 2.10b)", *http://nlp.kookmin.ac.kr/*, 2008.