근사 패턴매칭을 이용한 개선된 대화형 도우미 에이전트

김 수영, 조 성배
연세대학교 컴퓨터과학과

An Improved Conversational Help Agent Using Approximate Pattern Matching

Soo-Young Kim and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

인터넷의 성장에 따라, 많은 웹사이트가 생기고, 더 많은 정보들이 웹사이트에 등록되었다. 웹사이트에 등록되는 정보가 많을수록, 사용자가 원하는 정보를 얻기가 쉽지 않다. 따라서, 사용자가 찾는 정보를 쉽게 찾을 수 있도록, 웹사이트 내에 전문(full-text) 검색엔진을 도입하기도 한다. 본 논문에서는 사용자가 자연어를 이용한 대화를 통해 웹사이트 내의 정보를 습득할 수 있도록 하는 대화형 도우미 에이전트를 위하여 한글 전처리 방법과 근사패턴매칭기법을 제안한다. 사용자가 문장을 입력하면, 동의어처리와 형태소 분석을 통해 사용자의 문장을 분석하고, 이미 작성되어 있는 지식과의 매칭을 통해 사용자에게 알맞은 대답을 제시한다. 지식은 XML 형식으로 저장되며, 사용자가 입력한 문장과 아주 똑같지 않더라도, 어느 정도의 유사도를 가지고 대답을 이끌어낼 수 있다.

1. 서론

인터넷이 급속도로 성장함에 따라 인터넷의 웹사이트 숫자도 늘어나고, 많은 정보들이 등록되고 있다. 인터넷에 많은 정보가 등록되면서, 사용자의 정보획득을 위해 다양한 하이퍼링크를 제공하고 있으나, 대개 웹사이트의 구조는 정보제공자가 통일된 표준 없이 구성되기 때문에 인터넷 사용자들이 낯선 웹사이트에서 원하는 정보를 쉽고 빠르게 찾는 것이 쉽지 않은 경우가 많다. 이를 해결하기 위해, 웹사이트 내에 전문(full-text) 검색엔진을 도입하여, 사용자가 자신이 원하는 정보를 대표하는 키워드(keyword)를 입력함으로써, 검색을 할 수 있도록 하고 있다[1]. 그러나 정보검색 시스템의 대상문서의 양이 증가하면, 키워드 검색결과로 나오는 문서의 양 또한 늘어나게 된다. 그리고 결과로 나온 문서들을 일일이 읽어야 사용자가 원하는 답을 찾을 수 있는 번거로움이 있다.

여러 연구에서 사용자의 의도를 나타내는 데에는 키워드보다는 자연어(natural language)를 사용하는 것이 훨씬 효과적임이 밝혀져 있다[2,3]. 본 논문에서는 사용자가 원하는 정보를 키워드가 아니라, 자연어로 된 문장을 입력하게 함으로써, 사용자에게 좀더 친숙한 인터페이스를 제공한다. 입력받은 문장을 분석하여, 사용자가 원하는 답을 문서들의 집합이 아니라, 구체적인 문장으로써 답해주는 대화형 에이전트를 제안한다.

특정 도메인에서 사용할 수 있도록 지식을 정의하고, 사용자가 입력한 문장을 형태소분석과 매칭을 통해 구축된 지식 중 가장 근사한 답을 낼 수 있도록 한다.

2. 관련연구

최초의 대화형 로봇으로는 1966년 MIT대학의 Weizenbaum교수가 만든 ELIZA가 있다[5]. ELIZA는 사람과 기계사이의 자연어 의사소통을 연구하기 위해 만들어진 프로그램으로, 사용자가 던진 문장을 분석하여, 대답에 필요한 키워드를 추출한 후, 이미 프로그램 되어 있는 문장에 키워드를 치환하여 대답한다.

ALICE (Artificial Linguistic Internet Computer Entity)라는 대화형 로봇은 자연스러운 대화를 이끌어 내기 위하여, 패턴매칭 기법을 사용하였다[6]. 대화에 사용되는 지식, 즉 패턴은 AIML (Artificial Intelligence Markup Language)라는 XML형태의 파일로 저장되어 있다. AIML파일에 사용자가 물어볼 만한 질문과 그에 해당하는 대답을 기술한다. ALICE는 사용자가 물어보는 문장과 AIML파일에 패턴으로 기술된 문장을 비교하여, 하나의 와일드카드(*)를 허용하는 범위 내에서 정확하게 매칭 되는 것만을 선택하여 답을 해주는 시스템이다.

인터넷에서 사용되고 있는 검색엔진에 많은 진보가 이루어지고 있다. 특히, 사용자 쿼리를 자연어로 입력받고, FAQ에

기반하여 검색결과를 내는 검색엔진도 있다. 대표적인 검색엔진으로 Askjeeves.com이 있는데, FAQ스타일 검색엔진의 장점은 키워드기반 검색엔진이 수천 개의 결과를 내는 반면, 상호작용과 카탈로그 방식을 통해 더 정확한 결과를 내 준다는 것이다. 이러한 스타일의 검색엔진은 웹 기반 기술 지원과 같은 제한된 도메인에서 괄목할만한 성공을 보여주고 있다[4].

이와 같은 맥락에서, 사용자가 입력한 자연어 문장을 순차적 패턴 매칭 기법을 이용하여, "패턴-답변" 형태로 저장되어 있는 스크립트 파일로부터 사용자에게 알맞은 대답을 해주고 대화를 이끌어 가는 대화형 로봇을 개발한 바 있다[7]. 본 논문에서는 순차적 패턴 매칭 기법을 좀 더 발전시켜, 사용자의 문장이 약간 다르게 들어오더라도, 스크립트에 기술된 패턴과 근사하게 매칭 시켜 답을 해줄 수 있는 대화형 로봇을 제안한다.

3. 시스템 구현

대화형 도우미 에이전트는 다음과 같은 구조를 가진다.

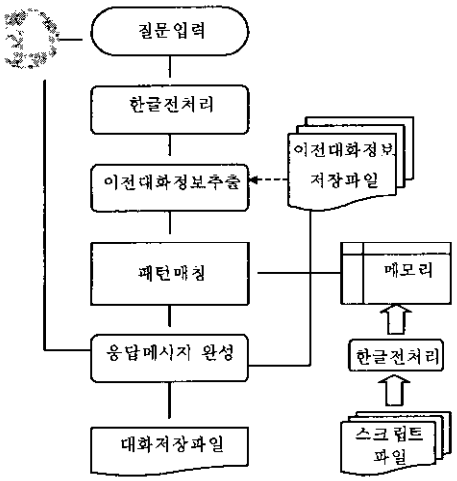

그림 1. 대화형 도우미 에이전트의 구조

3.1 한글 전처리

대화형 도우미 에이전트에서 이루어지는 문장은 대화체문장으로서, 문어체에 비해 단어의 축약, 조사의 생략, 수정 또는 반복발화, 간투어등의 특성으로 인해 분석하는데 많은 문제점을 포함하고 있다[8,9]. 이러한 대화체 문장을 분석하기 위한 가장 대표적인 방법이 개념기반 분석기법이다. 개념기반 분석기법은 강건성을 가장 큰 장점으로 가지며, 비문법적인 요소를 많이 포함하고 있는 자연발화 처리에 유리한 기법 중 하나로 평가되고 있다[10,11].

본 논문에서는 각 도메인에 대해 발화될만한 문장을, 그리고 대화를 통해 남겨진 문장을 분석하여, 문장내의 의미 있는 단어들을 개념으로 정의하고, 개념을 중심으로 지식을 만들어 매칭을 수행하게 한다.

한글 전처리에서는 사용자가 입력한 문장을 시스템이 다룰 수 있는 문장으로 규격화하는 작업을 수행한다. 한글에서는 영어와 달리 조사와 용언의 어미등 변화가 복잡하므로, 주된 의미에 영향을 미치지 않는 조사나 보조용언, 동사의 어미들을 제거하여 주된 의미만 가지고 매칭을 한다. 한글 전처리 과정은 부호(따옴표, 쉼표, 마침표, 느낌표, 물음표 등) 제거, 동의어처리, 형태소분석, 개념화로 구성된다. 등 부호를 제거한다. 동의어 처리부분에서는 동의어목록을 가지고 같은 의미를 가진 단어를 지식에 쓰인 단어로 치환한다. 동의어 처리 후 형태소 분석을 하여 개념을 뽑아내게 된다. 문장의 의미에 가장 많은 영향을 주는 명사, 용언의 어근, 의문사만을 추출한다. 의문사는 "무엇", "어디", "어떻게", "왜" 등으로 형태소 분석시 대명사나 부사등으로 분석되지만, 질문문장에서 중요한 역할을 하므로, 따로 정의하여 개념으로 추출될 수 있도록 한다. 즉 "주소록은 어떻게 만드나요?"라는 문장은 무호제거 단계를 거쳐 "주소록은 어떻게 만드나요"로 변하고, 동의어 처리 후 형태소분석을 하게 된다. 다음은 형태소 분석을 한 결과이다[12].

주소록은: 주소록/NNIN2+은/PPAU
어떻게 : 어떻게/ADCO
만드나요 : 만들/VBMA+나요/ENTE

표1. 형태소분석

이것을 바탕으로 개념화하면 "주소록 어떻게 만들" 이라는 3가지의 단어가 나오게 된다.(주소록'은 보통명사, '어떻게'는 부사지만, 의문사로 취급, '만들'은 동사)

3.2 지식처리

도우미에이전트가 대화를 나눌 수 있는 지식은 XML형식을 가진 스크립트파일에 저장된다.

지식의 기본단위는 FAQ에서 쓰이는 각각의 "질문-답변"이라고 보면 된다. 지식은 질문패턴, 답변, 관련 URL으로 구성된다. 답변은 질문패턴에 해당하는 답변이며, 관련 URL은 답변과 관련된 URL로 사용자 인터페이스에서 대화프레임 아래 프레임에 보여지게 된다. 질문패턴은 사용자가 입력할 만한 질문으로 같은 의미를 가진 질문을 여러 개 기술할 수 있다. 이렇게 함으로써, 같은 의미지만 다양하게 나타나는 질문들을 하나로 처리할 수 있다. 스크립트파일에 저장되는 질문패턴은 자연어로 된 완전한 문장이며, 메모리에 로딩될 때에는 한글 전처리 과정을 거쳐 개념화한 형태로 바뀌게 된다. 다음은 스크립트 파일의 예이다.

```
<qna>
<pattern>
<li>보낸 메일을 취소할 수 있나요?
<li>보낸메일을 취소하는 법은?
</pattern>
<answer>
인터넷메일은 전송된 메일을 PC통신에서처럼 전송취소를 할 수가 없습니다. 따라서 보내실 때는 신중하게 보내셔야 합니다.
</answer>
<link>http://...</link>
</qna>
```

표2. 스크립트 파일의 예

3.3 패턴매칭

사용자가 인터페이스를 통해 입력한 문장과 메모리에 로딩되어 있는 질문 패턴들과 비교하는 작업을 한다. 패턴매칭 전에 사용자의 입력문장은 한글 전처리 과정을 통하여 개념화되어 있고, 마찬가지로 지식도 스크립트 파일에서 한글 전처리 과정을 거쳐 개념화 되어 있다. 사용자가 입력한 문장의 개념단어집합 $C = \{ c_1, c_2, .. , c_n \}$, 지식의 질문패턴집합 $P=\{P_1,P_2,..,P_n\}$, $P_j=\{pc_1,pc_2,...,pc_n\}$ 라고 하면, 입력 문장에 대한 질문패턴 j 와의 스코어는 다음과 같이 구한다.

$$S_j = \sum_{i=0}^{n} exst(c_i, P_j)$$

$exst(c_i, P_j)$은 P_j의 집합내에 c_i와 같은 단어가 있으면 1을 반환, 그렇지 않으면 0을 반환하는 함수이다.

이렇게 하면, 패턴의 개념단어들과 입력문장의 단어들의 순서가 스코어에 영향을 미치지 않는다. 입력문장에 대한 모든 질문패턴의 스코어를 구한 후 가장 높은 점수를 획득한 질문패턴을 선택하여, 그에 해당하는 응답을 내보내게 된다.

4. 결론 및 향후 연구방향

본 논문에서는 특정 도메인에 대해 사용자의 질문에 대답하는 대화형 도우미 에이전트를 제안하였다. 질문문장이 한국어와 대화체라는 것을 감안하여 형태소분석과 동의어 사전을 통해 질문을 개념화하였다. 지식 생성시에는 동일한 뜻을 가진 질문을 여러 패턴으로 입력할 수 있게 함으로써, 융통성을 보였고, 또한 시스템 운영중에 생성된 로그를 바탕으로 지속적으로 쉽게 질문패턴을 추가할 수 있도록 하였다. 매칭스코어 계산시 입력문장의 개념단어들과 패턴의 개념단어들과의 순서를 고려하지 않음으로써, 한국어 어순의 자유로움을 고려하였다.

본 논문에서 제안된 시스템은 질문의 정확한 의미분석을 하기보다는, 질문패턴을 여러 가지로 입력함으로써, 계산 부하를 줄이면서도 사용자가 원하는 답을 뽑아낼 수 있게 하였다. 또한 많은 양의 질문패턴 데이터가 있을 때 효과적으로 성능을 발휘할 수 있을 것이다.

그러나 스크립트 파일을 일일이 사람이 작성하기에는 한계가 있으므로 자동으로 도우미 에이전트의 지식을 축적할 수 있는 방향이 연구되어야 하겠다.

참고문헌

[1] J.Cho, H.Garcia-Molina, L. Page, "Efficient crawling through URL ordering", *The 7th Int. WWW Conf.* (WWW 98), Brisbane, Australia, 14-18, April 1998.

[2] A.Pollok and A.Hockley, "What's wrong with Internet Searching". D-Lib Magazine. http://www.dlib.org/dlib/march97/bt/03pollock.html

[3] Y.J. Yang, L.F. Chien, L.S. Lee. "Speaker intention modeling for large vocabulary mandarin spoken dialogues", *Proc. of Fourth Int. Conf. on Spoken Language*, 1996.

[4] Q.Yang, H.F.Wang, J.R.Wen, G.Zhang, Y.Lu, K.F. Lee and H.J.Zhang "Towards a Next Generation Search Engine, " *Proc. of the Sixth Pacifiic Rim Artificial Intelligence Conference*, Melborne, Australia, August, 2000.

[5] Weizenbaun, J. "ELIZA-a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, 9(1):36-45, 1965.

[6] http://www.alicebot.org

[7] 김수영,조성배, "순차적 패턴 매칭 기법을 이용한 대화형 도우미 에이전트", *한국정보과학회* 2000 추계학술발표회, Vol 27, No.2, pp 24-26. 2000.

[8] 노서영, 정천영, 서영훈 "핵심개념 기반의 강건한 한국어 대화체 파싱", *한국정보처리학회 논문지*, 제6권, 제8호, pp. 2113-2123, 1999.

[9] 정천영, 임희동, 서영훈 "대화체 기계번역을 위한 중심어 기반 한국어 분석", *충북대학교 산업과학기술연구소 논문지*, 제 13권 제1호, pp.47-56, 1999

[10] Levin, E. and R. Pieraccini "Concept-based Spontaneous Speech Understanding System", *Eurospeech'95*, pp. 555-558, 1995

[11] L.J. Mayfield, M. Gavald`a, Y-H.Seo, B.Suhm. W. Ward, A. Waibel "Parsing real input in JANUS : a concept based approach to spoken language translation" *Proceeding of TMI95*, pp196, 1995

[12] 윤준태, 이충희, 김선호, 송만석 "연세대 형태소 분석기 morany: 말뭉치로부터 추출한 대량의 어휘 데이터베이스에 기반한 형태소 분석" *한글 및 한국어 정보처리 학술대회* 1999.