홈쇼핑 사이트를 위한 데이터베이스로부터의 한국어 텍스트 생성

노지은[0], 강신재, 이종혁

포항공과대학교 컴퓨터공학과

{jeroh,sjkang,jhlee}@kle.postech.ac.kr

Korean Text Generation from Relational Database for Homeshopping Sites

Ji-Eun Roh[0], Sin-Jae Kang, Jong-Hyeok Lee

Dept. of Computer Science & Engineering, POSTECH.

요약

국내에서는 한국어 생성에 있어서 기계 번역에 기반한 자연스러운 한국어 문장(sentence)의 생성에 관한 연구가 주로 이루어졌었다. 반면에 다양한 지식원으로부터 여러 문장이 긴밀히 결합되어 하나의 텍스트를 생성하는 텍스트 생성에 관한 연구는 거의 이루어지지 않았었다. 문장 단위의 기계 번역에서의 한국어 생성과는 또 다른 다양한 논점을 가지고 있는 텍스트 생성에 관해, 본 논문에서는 데이터베이스를 지식원으로 하여 하나의 일관된 정보를 전달하는 단락 단위의 자연스러운 한국어 텍스트를 생성하는 시스템을 제안한다.

1. 서론

자연어의 생성(Natural Language Generation)은 자연어로 이루어지지 않은 기저의 정보들을 자연어로 사상하는 언어처리의 한 분야로 일반적으로 두 가지의 범주로 나누어진다. 기계 번역에 기반한 단문장의 생성(sentence generation)과 텍스트의 생성(text generation)이 바로 그것이다. 특히 텍스트의 생성은 여러 가지 측면에서 단문장의 생성과는 서로 다른 논점을 갖는다. 여러 문장이 긴밀히 결합되어 하나의 정보를 전달하는 단위를 텍스트라 볼 때, 높은 질의 텍스트를 생성하기 위해서는 문장간의 순서, 문장간의 결합, 각 문장들에서의 지시어 생성 등을 적절히 처리해 주어야 한다. 본 논문에서는 홈쇼핑 사이트에서 제공되는 여러 가지 상품들의 정보들을 관계 데이터베이스로 구축하고 이를 도메인으로 하여 사용자의 요청이 들어올 때마다 해당하는 상품들의 카탈로그를 자동적으로 생성하는 시스템을 설계, 구현하고 이에 필요한 단계별 과정을 처리하는 방법을 보인다.

현재 제안하는 시스템은 크게 두 단계로 이루어진다. 첫번째 단계는 지식원으로 선정된 도메인의 모델링으로 오프 라인상에서 이루어지며 이 단계가 끝나면, 모델링된 도메인으로부터, 실시간으로 요청이 들어 올 때마다 텍스트를 생성하는 생성부분이 있다.

2. 도메인 모델링

관계 데이터베이스로 구축되어진 도메인을 생성 시스템의 지식원으로 이용하기 위해서는 각 테이블의 레코드들이 어떤 의미를 갖는지, 어떻게 표현되어야 하는지에 관한 정의가 이루어져야 한다

2.1 도메인 표현 모델링

구축되어진 데이터베이스의 테이블의 내용 일부를 간략하게 그림으로 나타내면 다음과 같다.

Id	Category	Name	Company	Price	Form.....
P1	Camera	Episode20s	E1(삼성항공)	300000	자동카메라...
P10	Camera	joycam	F1(폴라로이드)	395000	특석카메라...

<그림 1> 관계형 데이터베이스로 구축되어진 지식원

본 시스템에서는 각 칼럼의 이름(category, company, price...)을 '기초정보(BIU)'라 정의하고 기초정보는 두개의 인자를 가지며 이것이 가장 기본적인 정보의 단위가 된다. 모든 데이터 표현방법은 xml 형식을 이용하였다.

<BIU name="price" **primary**="p1" **secondary**="300000">

<그림 2> 기초정보 ' 가격' 의 정의 예

또한, 문장 순서 결정을 위해 각 기초정보에 역할을 부여하고 기초정보의 두개의 인자에 대한 의미 관계를 설정하기 위해 그림 3 과 같은 표현방식을 정의한다.

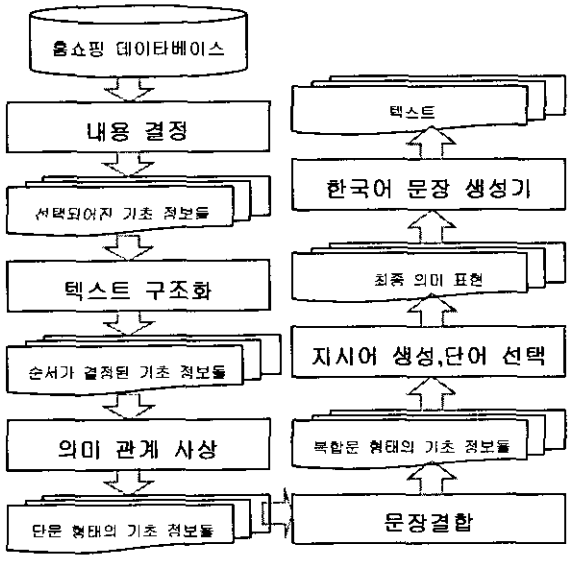
```
<SemRep-Expression BIU="price">
    <root verb="be">
        <feature>
            <mood>declarative</mood>
            <voice>active</voice>
            <tense>present</tense>
        </feature>
        <case name="EXP">
            <head type="NP">lexicalise(" price" )</head>
            <case name="DET">
                <head type="NP">lexicalise(primary)</head>
            </case>
    </case>
    <case name="RLT">
```

```
        <head type="NP">lexicalise(secondary)</head>
    </case>
  </root>
</SemRep-Expression>
```

<그림 3> 기초정보 ' 가격' 의 두 인자에 대한 의미관계 정의 예

3. 텍스트 생성

제안하는 생성 시스템의 두번째 단계인 텍스트 생성 부분의 전체 시스템 구조도는 아래 그림과 같다.



<그림 4> 생성 부분의 전체 시스템 구조도

파이프 라인 형식의 순차적인 구조를 따랐으며, 각 단계별 피드백을 허용하지 않는다. 이와 같은 순차적인 구조는 피드백이 허용되는 것에 비해 유연성은 떨어지지만, 모듈화가 가능하고 구현상 편리하다는 장점이 있다[1].

3.1 내용 결정(Content Determination)

이 단계는 어떠한 상품의 정보를 얻고자 요청이 들어 왔을 때, 생성될 텍스트에 최종적으로 나타날 모든 정보들을 지식원(데이터베이스)에서 뽑아내는 과정으로, 알고리즘을 제안하기에 앞서 이상적인 텍스트(target text)를 먼저 살펴보면 다음과 같다

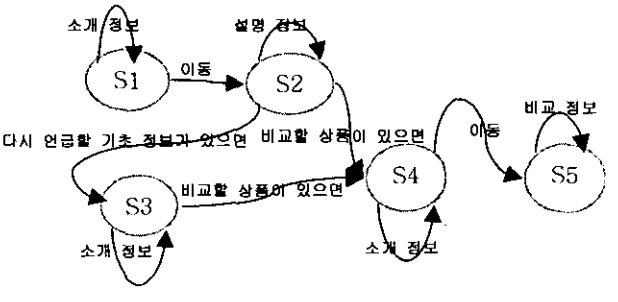
삼성항공에서 생산된 삼성 캐녹스 카메라 **KENOX-140IP** 는 줌렌즈 내장 전자동 35mm 렌즈셔터 카메라로 **가격**은 304000 이다. 폭이 115mm, 깊이 44mm,높이 65mm 이고 무게는 230g 으로 앞서 본 캐논 카메라보다 *작고 가볍다*. 줌과 파노라마가 가능하고 리모콘 촬영이 가능하다. 촬영 날짜가 기록되는 이 카메라는 셀프 타이머가 있으며 촬영매수가 표시되고 플래시가 있다. --------①
이 카메라와 같은 **가격대**의 제품은 다음과 같은 것들이 있다.
SLR 카메라
펜탁스 카메라 ----------②
SLR 카메라는 역시 삼성항공에서 생산된 전문 카메라로 가격은 KENOX-140IP 보다 *조금 비싸지만, 크기는 작고 더 많은 기능*을 가지고 있다. ---------- ③

<그림 5> KENOX-140IP 카메라에 대한 이상적인 텍스트

그림 5 에서 ①은 기술대상(KENOX-140IP)에 관련된 모든 내용이 포함된 핵심 단락이고 ②는 기술대상을 설명하면서 부가적으로 추가되는 부가 단락이며, ③은 현재의 기술대상과 비교할 만한 상품을 소개하는 비교 단락이다. 이를 근거로 해서 구체화된 알고리즘을 소개하면, 먼저 기술대상을 첫번째 인자로 갖는 모든 기초정보를 선택하고, 그 중 중요도가 높은 기초정보(그림 5 에서 " 가격대")와 관련된 인자를 갖는 또 다른 기초정보를 선택한다. 다음 생성기록(Discourse History)를 고려하여 비교할만한 상품이 있는가를 판단하고, 있으면 이것을 첫번째 인자로 갖는 기초정보를 선택한다. 그리고 최종적으로 요약이 필요한 부분에 대해서 처리해 준다.(그림 5 에서 밑줄 친 부분)

3.2 텍스트 구조화(Text Structuring)

앞 단계에서 선택되어진 모든 기초정보들을 대상으로 그것들의 순서를 정하는 단계이다. 문장간의 순서를 결정하는 방법은 크게 RST 방식[3]과 Schema 방식[4]으로 나눌 수 있다. RST 방식은 메시지간의 수식관계(rhetorical relation)를 이용하여 최적의 문장관계를 결정할 수 있게 하는 것으로 유연성이 높지만 탐색 문제이므로, 시간 복잡도가 높고 구현하기 힘들다는 단점이 있다. 본 시스템은 적용하는 도메인의 성격상 각 기초정보간의 정교한 수식관계가 존재하지 않으므로 RST 방식을 적용하기에는 무리가 있어 Schema 방식을 기반으로 한 Explanation-Schema 를 만들고 이를 이용해 문장 순서 및 단락의 순서를 결정한다.



<그림 6> Explanation-Schema

위의 그림에서 상태(state)간의 이동은 앞서 도메인 모델링에 의해 정의되었던 기초정보의 역할에 의해 결정된다. S2 에서 S3 으로의 이동은 그림 5 에서 보았던 " 가격대" 처럼 다른 단락을 생성 가능하게 하는 기초정보가

존재하여야 가능하고 S2 에서 S4, S3 에서 S4 로의 이동은 그림 5 에서 " SLR 카메라" 처럼 비교 가능한 상품이 있으면 가능하다. 그리고 S5 는 기술대상이 되는 상품과 비교대상이 되는 상품의 기초정보의 비교가 이루어 진다. 최종적으로 S1 과 S2 가 결합되어 하나의 핵심 단락이 되고, S3 이 부가 단락, S4 와 S5 가 결합되어 비교 단락이 된다.

3.3 의미 관계 사상(Lexicalisation)

순서가 결정되어진 각 기초정보에 대해, 그림 3 과 같은 의미관계를 사상하는 과정으로 단문 수준의 초기 의미 표현(그림 4, 단문 형태의 기초 정보들)이 생성된다.

3.4 문장 결합(Aggregation)

의미 관계 사상이 끝나고 생성된 초기 의미 표현은 문장 구성상 주어와 서술어가 하나인 단문의 수준으로 볼 수 있다. 이러한 단문을 계속 열거하는 것보다 단문들을 둘 이상 결합하여 복합문을 이루면 가독성이 훨씬 향상될 수 있다. 본 시스템에서는 [2]에서 제안한 4 가지 방식 중 접속(conjunction)과 내포(embedding)만을 선택하고 한국어에 맞게 적용하였다. 접속문을 구성할 때는 공통 성분이 삭제(conjunction reduction)되는데 한국어에서는 동사가 동일할 때는 앞 문장의 동사가 삭제되고 명사가 동일할 때는 영어와는 반대로 뒤쪽이 삭제된다.[5]

3.5 지시어 생성 및 단어 선택(Referring Expression Generation and Lexical Choice)

자연스런 지시어 생성을 위한 과정으로, 같은 지시어를 반복해서 사용함으로 인한 어색함을 피하고 가독성을 높이기 위해 다양한 지시어 생성 방법이 필요하다. 그림 5 에서 4 번째 문장의 주어의 실제 지시어는 "KENOX-140IP"이지만 "이 카메라"로 대체됨으로써 타입에 의해(referring-by-type) 지시어가 생성된 경우라 볼 수 있다. 본 시스템에서는 고유명사를 이용한 생성(referring-by-name), 분류 체계를 이용한 지시어 생성(referring-by-type), 지시어의 생략(referring-by-ellipsis)에 의한 처리의 3 가지 방식을 만들어 적용하였다. 마지막으로 적합한 한국어 단어를 선택하면 그림 4 에서 보이는 최종 의미 표현들이 생성되고, 이것이 포항공대 KLE 연구실에 있는 한국어 문장 생성기를 거치면 최종적인 텍스트가 생성된다.

4. 실험 및 결론

홈쇼핑의 상품 도메인 중에서" 카메라" 의 품목의 70 종에 제한시켜 도메인을 만들고 실험이 이루어졌다. 평가 범주는 앞서 설명한 4 가지(내용선택,텍스트 구조화,문장결합,지시어 생성)과정과 전체적인 가독성 평가를 포함하였다. 평가 방식은 각 과정별로 4 단계(9,6,3,0)의 점수를 제시하고 사용자에게 점수를 매기게 한 후 평균을 취하였다. 실험 대상으로 사용한 전체 텍스트의 수는 20 개로 실험 결과는 다음과 같다.

내용	단계별 평가				가독성
단계	내용선택	텍스트구조화	문장결합	지시어 생성	
평균	9	7.1	5.7	8.2	7.4

<표 1> 실험결과

실험결과를 분석해 보면 내용 선택은 높은 점수를 얻어, 내용의 충실성 측면에서는 잘 처리가 되었다고 본다. 문장의 순서 및 단락의 순서 역시 비교적 자연스럽게 이루어 진 것으로 인정된다. 하지만 복합문의 구성에 있어 그 연결이 매끄럽지 못하거나, 복합문으로 구성되는 단문의 수가 획일화 되어 있어 가독성이 떨어지는 결과를 초래하였다.이런 결과를 얻게 된 이유 중 하나가 문장 결합 시 결합 가능한 최대 문장 개수를 제한하였고 결합 범위가 인접한 문장으로 제한되기 때문이다. 이는 문장의 순서가 결정된 이후에 문장 결합을 처리하는 순차적인 구조를 따르는데 기인하며 문장 결합과 텍스트 구조화간의 피드백이 필요한 동기가 된다.

추후에 보완해야 할 점으로는 한국어 특성에 맞는 복합문 구성을 위한 여러가지 방안과, 더욱더 자연스러운 표층문의 생성을 위해 현재 사용하고 있는 문장 생성기를 개선하는 것이다. 그리고 좀더 유연성이 있는 문장 순서, 단락 순서를 결정하기 위해 텍스트 구조화에 적용하고 있는 스키마 방식에 RST 방식을 결합하여 적용해 볼 수 있겠다.

참고문헌

[1] Hui-Feng Li, Sin-Jae Kang, and Jong-Hyeok Lee, "Korean Sentence Generation in Mtran-C/K System Based-on a Multi-level Processing Strategy", 2000 International Conference on Chinese Language Computing, Chicago, pp. 12-18, July 8-9, 2000

[2] H. Dalianis & E. Hovy. "On Lexical Aggregation and Ordering." In Demonstrations and posters of the 8th International Workshop on Natural Language Generation, INLG'96, pp. 29-32, Herstmonceux, Sussex, UK, June 13-15, 1996

[3] Marcu, Daniel. ``Building up rhetorical structure trees." *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96),* pp.1069—1074, August 1996,

[4] Kathleen R.Mekwown "Text Generation" Cambridge university press,1985

[5] 이익섭,이상억,채완 "한국의 언어",신구 문화사,1999