2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

# 기상해설 시스템 (Weather Commentator System)에서 자연언어 생성 기술의 적용

김정은, 백혜승, 최기선
한국과학기술원 전산학과, 전문용어언어공학연구센터
e-mail : {euni, hspaik, kschoi}@world.kaist.ac.kr

# Application of NLG technology to Weather Commentator System

Jungeun Kim, Haeseung Paik, Key-Sun Choi
Dept. of Computer Science, KAIST/KORTERM

요 약

본 논문은 기후자료로 제공되는 기상 데이터베이스로부터 사람이 이해할 수 있는 수준의 분석문을 생성하는 시스템에 자연언어 생성 기술을 적용한 연구에 관한 것이다. 기상청에서 제공되는 여러 가지 자료들을 이용하여 기상관련 지식을 획득하였으며, 제한된 영역에서 잘 구조화된 템플릿을 정의하고 담화관계를 설정함으로써 관련 기상자료에 대한 해설문을 생성할 수 있었다. 실험 결과, 본 시스템은 비교적 좋은 성능을 나타냄을 알 수 있었다.

## 1. 서론

자연언어 생성(Natural Language Generation)은 인공지능과 계산언어학의 한 분야로서 컴퓨터 시스템이 정보의 비언어적 표현형태로부터 사람의 언어로 된 텍스트를 생성하는 것을 연구한다[1].

본 논문에서는 수치로 표현된 기상자료로부터 한국어로 된 기상해설문을 생성하는 기상해설 시스템(Weather Commentator System)을 제안한다. 기상해설 시스템은 우리나라 각 기상대나 관측소에서 관측된 월별수치 자료표인 연월보 자료들을 입력으로 받아서 전반적인 분석과 중요하거나 특이한 기상 사건을 제시하는 텍스트를 생성한다.

기상해설 시스템은 다음과 같은 측면에서 도움이 될 수 있다.

- 사용자 : 기상자료는 강수량, 온도, 풍속, 일기일수 등의 여러 가지 수치로 복잡하게 표현되어 있는데, 이를 기상학과 관련된 전문 지식 없이도 쉽게 이해 할 수 있도록 도움을 줄 수 있다.
- 기상 전문가 : 컴퓨터가 기상 해설문을 생성하므로, 기상 전문가가 기상자료를 분석하여 기상 개황을 생성하는 시간과 노력을 줄일 수 있다.

본 논문에서는 기상해설 시스템의 입력으로 사용되는 기상수치표에 나타나는 중요한 정보를 추출하기 위하여 기상청에서 제공하는 기상관련 전문지식을 이용하였다. 또한 기상수치들이 담고 있는 내용을 적절하게 기술하기 위해서 자연언어 생성의 여러 가지 생성 기술들[4,5,6,7,8] 중 제한된 영역이라는 이점을 살려 템플릿에 기반한 생성 방법을 적용하였다.

## 2. 기상해설 시스템

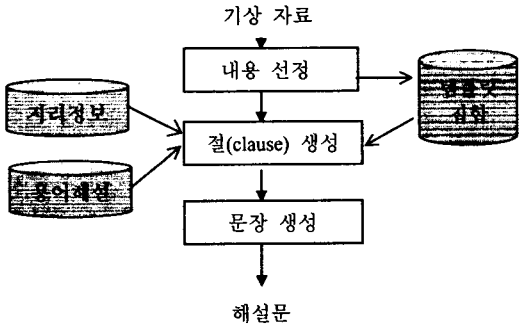
### 2.1. 시스템 구조

기상해설 시스템의 전체적인 구조는 [그림 1]과 같다. [그림 1]에서 기상 자료 입력이 들어오게 되면 기상 자료를 잘 표현할 수 있는 해설문을 생성하기 위하여 해설문에 포함될 내용을 선정(Content Selection)하는 작업이 필요하다. 이를 위해 기상청[2]에서 제공하는 기상분석문과 기상 관련 자료를 사용하였다.

기상자료를 분석하여 내용이 선정되면 기상청[2]에

서 제공하는 예보용어 해설 정보 등을 이용하여 포함되어야 할 내용을 절(clause)로 생성하게 된다. 이 과정에서 템플릿을 구성하여 이용하게 되는데, 템플릿은 사용자나 응용 프로그램에 의해 실시간에 결정되는 매개변수를 가진 사전에 정의된 문자열[9]이다. 본 논문에서는 기상 해설문이 정형화된 패턴을 갖고 있다는 점을 이용해서 절 생성 단계에서 템플릿 간의 계층관계를 미리 설정하여 이용함으로써 템플릿의 구조화를 보다 쉽게 이루었다.

문장 생성 과정에서는 구조화된 템플릿들이 실제 문장으로 표현될 때 보다 자연스런 문장이 되도록 연결어미 선택 등의 과정을 거친다.




[그림 1] 기상 해설 시스템 구조도

## 2.2. 코퍼스 분석 및 전문 지식 획득

자연스러운 기상 해설문을 생성해 내기 위해서는 기존의 사람이 생성한 기상 해석문들[2, 3]을 분석하여 기상 해석문들에서 나타나는 정형화된 패턴을 파악하는 것이 필요하다. 본 논문에서는 기상해석문 분석 과정을 통하여 다음과 같은 세가지 유형의 정형화된 패턴을 파악하였다.

- 일반적 문장 : 가장 일반적인 정보를 담고 있는 문장이다.
  예) 이 달의 전국 월평균기온은 0.7~11℃의 분포를 나타내었다.
- 특징적 문장 : 입력 자료로부터 두드러진 특징에 관한 정보이다.
  예) 일조시간이 가장 많았던 곳은 거제로 251 시간이었다.
- 비교 문장 : 해당 수치를 다른 값과 비교하는 문장이다.
  예) 서울·경기지방은 평년대비 10% 미만의 강수분포를 나타냈다.

기상 관련 전문분야 지식은 기상청[2]에 제공하는 [표 1]과 같은 예보용어 해설에 관한 지식을 사용하였다. 또한, 기상 자료에 자주 나타나는 지방별 특징 분석을 효과적으로 처리하기 위하여 [표 2]와 같은 우리나라 지방별 분류표를 구축하였다.

| 용어 | 비교값 | | | 발생 확률 |
|---|---|---|---|---|
| | 반순 | 순 | 월 | |
| 높다 | +2.6 이상 | +2.1 이상 | +1.6 이상 | 10 |
| 조금 높다 | +1.1-+2.5 | +0.9-+2.0 | +0.6-+1.5 | 20 |
| 비슷 | -1.0-+1.0 | -0.8-+0.8 | -0.5-+0.5 | 40 |
| 조금 낮다 | -1.1--2.5 | -2.0--0.9 | -1.5--0.6 | 20 |
| 낮다 | -2.6 이하 | -2.1 이하 | -1.6 이하 | 10 |
| 주간, 월간 예보에 사용 | | | | |

[표 1] 예보 용어 해설 예: 기온 비교 표현

| | 서울경기 | 강원 | 충청 | 영남 | 호남 | 제주 |
|---|---|---|---|---|---|---|
| 기호 | SK/10 | KW/11 | CC/12 | KS/13 | JL/14 | CJ/15 |

[표 2] 지방 분류표

## 2.3. 템플릿 정의

기존의 기상해설문을 분석한 결과를 [그림 2]와 같은 템플릿으로 정형화한다. 각각의 템플릿은 하나의 절을 구성한다.

```
NAME
TEXT
SUBJECT
DISCOURSE
```

[그림 2] 템플릿 구조

NAME 은 템플릿의 이름을 나타내고 TEXT 는 템플릿을 구성하는 문자열을 나타낸다. TEXT 는 입력자료에 따라 다르게 채워지는 매개 변수인 슬롯을 포함하며, 슬롯은 슬롯 번호와 슬롯 종류로 구성된다. SUBJECT 는 템플릿의 주제어를 나타내고, DISCOURSE 는 앞 문장과의 담화관계를 나타낸다.

[그림 3]은 기상 해설문의 분석을 통해서 작성된 템플릿의 예이다.

```
NAME : AveMaxTemp
TEXT : 전국 월평균최고기온은
[SLOT[NUMBER:1][TYPE:NUM]]~[SLOT[N
UMBER:2][TYPE:NUM]]℃의 분포였다
SUBJECT : 기온
DISCOURSE : APPOSITION
```

[그림 3] 템플릿의 예

## 2.4. 해설문 생성

구조화된 템플릿 집합들은 다음의 규칙에 따라 문장으로 생성하게 된다 [9].

- 할당 규칙 : 템플릿에 포함된 슬롯의 값을 결정하여 적절하게 채워 넣는다.
- 병합 규칙 : 같은 개념을 가지고 있는 여러 개의 문장을 하나의 문장으로 묶는다.
- 선택 규칙 : 해당 템플릿 이전에 어떤 템플릿이 선택되었느냐에 따라 다른 템플릿을 선택한다.

템플릿의 집합을 이용하여 생성된 문장들은 절의 형태이기 때문에 부자연스러운 표현이 될 수 있다. 이러한 문장들을 자연스런 문장으로 만들기 위하여 다음과 같은 처리를 한다.

- 담화관계에 따라 연결어미를 결정한다.
- 템플릿의 주제에 따라 문단을 구분한다.

본 논문에서는 기상 해설문에서 가장 자주 사용되는 [표 3]과 같은 세가지 담화관계를 고려하여 연결어미를 선택하고 생성된 템플릿의 결합에 이용하고자 한다[1].

| 담화관계 | 정 의 | 선택될 연결어미 |
|---|---|---|
| 대립 | 선행문과 후행문의 내용이 반대이거나 비교, 대조된다. | (으)나<br>(이)고 |
| 상술 | 선행문의 주제를 후행문에서 부연, 보충한다 | (으)로 |
| 대등 | 선행문과 후행문이 대등한 관계이다. | (으)며<br>(이)고 |

[표 3] 가장 대표적인 담화 관계

이해하기 쉽고 자연스런 텍스트를 만들기 위해서는 문장들을 연결하고 끊는 것이 적절해야 하므로 담화관계가 설정되지 않은 문장들의 경우 적절한 연결어미를 선택해서 문장의 길이를 조절하는 것이 필요하다. 기상해설문 생성시 사용되는 제한된 비교 용언들에 대한 분석을 통하여 연결어미의 선택에 사용될 수 있는 [표 4]와 같은 대조 어휘들을 찾을 수 있다.

| | 긍정값 (-1) | 부정값(-1) |
|---|---|---|
| 대조어휘 | 높다<br>최고<br>많다<br>맑다<br>덥다, 무덥다, 따뜻하다<br>있다, 나타나다 | 낮다<br>최저<br>적다<br>흐리다<br>춥다, 쌀쌀하다<br>없다 |

[표 4] 대조어휘표

담화 관계에 따라 연결어미를 결정한 후 담화 관계가 설정되지 않은 문장들에 대해서는 [표 4]의 대조어휘표를 이용하여 [표 5]과 같이 연결어미를 선택할 수 있다.

| | 선행문 -> 후행문 | 연결어미 | 예 문 |
|---|---|---|---|
| 비교관계 | 0 -> +1 | -고 | 이 달의 평균 기온은 13~24℃이고, 더운 날씨가 계속되었다. |
| | +1 -> +1 | -이며 | 최고기온은 34℃이며 중부지방에서 더운 날씨가 계속되었다. |
| | -1 -> 0 | -고 | 최저기온은 강릉에서 □ 10℃이고, 수원에서 초하루와 말일의 기온차이가 12℃였다. |
| 대조관계 | -1 -> +1<br>+1 -> -1 | -(으)나 | 전국이 대체로 평년보다 낮은 기온 분포를 보였으나, 영남 내륙 일부지방은 1~2℃ 높았다. |

[표 5] 대조어휘에 의한 연결어미 생성 예

## 3. 실험 및 평가

### 3.1. 실험

실험에 사용한 입력 자료는 기상청[2]에서 제공하는 기압, 기온, 일조, 강수량, 바람, 일기일수, 극값 등의 월별 수치 자료를 사용하였으며 33 개의 템플릿을 이용하여 문장을 생성하였다. 2000 년 1 월~2000 년 12 월의 자료를 이용하였으며, 결과로 [그림 4]와 같은 기상해설문을 생성하였다.

3 월의 전국 월평균기온은 0.7℃에서 11.0℃의 분포로, 서울경기지방 4.2~6.3℃, 강원지방 0.7~8.1℃, 충청지방 3.2~6.1℃, 영남지방 3.2~9.1℃, 호남지방 3.4~8.3℃, 제주지방 9.0~11.0℃의 분포를 나타내었다. 이 달의 전국 월평균최고기온은 6~16℃의 분포였고 전국 월평균최저기온은 -5~7℃의 분포였다
3 월의 월강수량은 전국적으로 2~83 ㎜의 분포로, 서울・경기지방 2~6 ㎜, 강원지방 9~22 ㎜, 충청지방 4~26 ㎜, 영남지방 16~59 ㎜, 호남지방 13~38 ㎜, 제주지방 37~83 ㎜의 분포를 나타내었다.
3 월의 월평균풍속은 전국이 1~9 *m/s* 로, 서울・경기지방 2~4 *m/s*, 강원지방 1~6 *m/s*, 충청지방 1~5 *m/s*, 영남지방 1~4 *m/s*, 호남지방 1~5 *m/s*, 제주지방 3~9 *m/s* 였다. 홍천에서 최대풍속 11.3 *m/s*, 산청에서 최대풍속 15.8 *m/s* 의 바람이 불어 최대풍속 극값을 갱신하였다. 이 달에 비교적 바람이 강했던 날은 24 일, 27 일이었다. 폭풍일수는 제주고층이 14 일로 가장 많았으며 완도・여수・군산 6 일, 흑산도 4 일이었다.
3 월의 일조시간은 전국적으로 153~250 시간으로 일조율은 41~67%였다. 각 지역별 일조시간은 서울・경기지방 153~232 시간, 강원지방 194~233 시간, 충청지방 204~238 시간, 영남지방 201~250 시간, 호남지방 187~239 시간, 제주지방 198~220 시간이었다. 일조시간이 가장 많았던 곳은 진주로 250.1 시간(일

조율은 67%)이었으며, 일조시간이 가장 적었던 곳은 울릉도로 152.8 시간(일조율은 41%)이었다. 이 달의 부조일수는 전국적으로 1~4 일의 분포를 나타내었으나, 울진 · 대구는 없었다

[그림 4] 생성된 기상 해설문의 예

### 3.2. 평가

자연언어생성 시스템의 평가에는 다음과 같은 방법이 있다 [1].

- 과업(task) 평가 : 사용자가 자연언어생성 시스템이 목표로 하는 해당 분야에서의 과업을 수행할 수 있는지의 성공 여부를 본다. 평가 비용이 큰 단점이 있다.
- 품질(quality) 평가 : 전문가가 생성된 텍스트의 품질을 평가한다.

두번째 방법인 품질 평가는 비교적 비용이 적게 들고 빠른 시간에 수행할 수 있다는 장점이 있다. 따라서 본 실험에서는 이 방법을 이용하여 평가를 수행하였다. 평가 기준으로는 다음의 세 가지를 사용하였다.

- 적합성 : 일기 분석문이 수치자료로부터 어느 정도로 적합하게 해당월의 특성을 잘 나타내고 있는가의 정도이다. "1:아주 부적합, 2:약간 부적합, 3:그저 그렇다, 4: 적합, 5: 아주 적합"의 1 에서 5 의 수치로 나타낸다. 따라서 수치가 클수록 더 적합하다.
- 가독성(readability) : 연결어미가 적절하여 문맥이 자연스러운가의 정도

$$\text{가독성} = 1 - \frac{\text{틀린 연결 어미수}}{\text{전체 연결 어미수}}$$

- 수용성(acceptance) : 기상 해설문이 종합적으로 어느 정도 수용 가능한가의 정도. "1:절대 수용불가, 2: 수용 불가, 3:그저 그렇다, 4:수용 가능, 5:수용 매우 적합"의 1 에서 5 의 수치로 나타낸다. 따라서 수치가 클수록 더 적합하다.

기상분야 전문지식이 필요한 적합성과 수용성 평가는 전문가 3 명*이 수행하였으며, 기상분야의 전문지식이 필요없는 가독성 평가는 일반인 2 명이 수행하였다.

| 평가자 | 가독성 | 평가자 | 적합성 | 수용성 |
|---|---|---|---|---|
| A | 0.97 | C | 3.92 | 4.67 |
| B | 0.87 | D | 4.50 | 4.33 |
| | | E | 5.00 | 4.25 |
| 평균 | 0.92 | | 4.47 | 4.42 |

[표 6] 평가 결과

[표 6]은 평가 결과로, 높은 가독성을 나타낸다. 수용성은 종합적인 품질을 나타내는데, 수용할만하다는 전문가의 평가를 얻었다. 적합성과 수용성은 가독성에 비해 낮은 수치를 보이고 있는데 이는 기상학 분야의 전문 지식의 부족으로 인한 용어의 부적절한 사용에 기인한 것으로 보인다.

## 4. 결론 및 향후연구

본 논문에서는 템플릿을 기반으로, 지나간 달의 기상에 관한 수치 자료들로부터 해당월의 기상에 대한 분석문을 제공하는 시스템을 구현하여 92%의 가독성과 비교적 높은 적합성과 수용성을 얻었다.

보다 신뢰성 있는 정보를 제공하기 위해서는 기상분야의 전문가와의 심도 있는 협력이 필요하다. 또한, 본 논문에서는 일반적인 사용자를 가정하였으나, 사용자에 따라 중점을 두는 정보가 다르므로 각각의 사용자에게 보다 유용한 정보를 제공하기 위해서는 먼저 사용자의 요구를 분석하여 사용자에 따라 다른 분석문을 제시하는 연구가 필요하다. 마지막으로, 학습을 통한 템플릿의 자동생성이 이루어진다면, 적응성 있는 시스템이 만들어질 수 있을 것이다.

* 평가에 참여해 주신 대전 기상청 관계자 분들께 감사를 드립니다.

## 참고문헌

[1] Ehud Reiter, Robert Dale. "Building Natural Language Generation Systems", Cambridge University Press, 2000.
[2] 기상청 홈페이지 http://www.kma.go.kr/
[3] w365 기상정보 http://www.w365.com/
[4] Gerd Herzog (DFKI), FLUIDS Technology Watch Report –Part4 "Natural Language Generation"
[5] Goldberg, Eli, Norvert Driedger, and Richard Kittredge, Using natural language processing to produce weather forecasts, IEEE Expert, 9(2):45-53, 1994.
[6] R. Paterson et al., The Forecast Production Assistant, Proc. Fourth AES/CMOS Workshop on Opearational Meteorology, Canadian Meteorological and Oceanographic Soc., Toronto, 1002, pp 262-269
[7] Polguere, Grammatical and Lexical Formalisms in FOG-89, internal report, Atomospheric Envireonment Service, Environment Canada, Downsview, Ontario, Canada, 1989.
[8] CLINT - A Hybrid Template/Word-based Text Generator http://www.cs.bgu.ac.il/~elhadad/clint.html
[9] Susan W. McRoy, Songsak Channarukul, Syed S. Ali, YAG:A Template-Based Generator for Real-Time Systems, Proceedings of The First International Natural Language Generation Conference (INLG 2000), pp. 264-267, 2000.