문장추상화: 문서요약을 위한 문장교열 방법론

°김곤†, 배재학†
울산대학교 컴퓨터·정보통신공학부

Sentence Abstraction: A Sentence Revision Methodology for Text Summarization

°Gon Kim†, Jae-Hak J. Bae†

School of Computer Engineering and Information Technology
University of Ulsan

요 약

본 논문에서는 문서요약을 위한 문장교열 방법론으로 문장추상화를 생각하였다. 이에 문장추상화의 판단기준이 되는 요소들을 구문분석기를 통해 얻은 정보와, 문장의 구성성분들이 가지는 온톨로지 정보를 바탕으로 선정하였다. 문장추상화에는 Roget 시소러스에 기반한 온톨로지 OfN, 구문분석기 LGPI+, 그리고 이를 활용하는 문장추상기 SABOT를 이용하였다. 본 논문을 통하여 문장추상화가 문서요약을 위한 문장교열 방법의 하나로 가능함을 보였다.

1. 서론

문서요약 방법[1, 7]은 본문문법(Text Grammar), 수사관계(Rhetorical Relation), 어휘사슬(Lexical Chain) 등을 핵심적으로 활용하는 방법과 원문에 나타나는 용어나 문장의 빈도수(Frequency), 단어 가중치(Word Weight) 등에 기반한 방법론들이 있다. 문서요약을 위한 문장교열(Sentence Revision) 방법[2]에는 (1) Text Compaction, (2) Sentence Reduction, (3) Sentence Compression 등이 있다. 본 논문에서는 문서요약을 위한 또 다른 문장교열 방법론으로 문장추상화를 생각하였다. 이러한 문장추상화에는 그 판단기준과 자동화 도구가 필요하다.

2. 문장교열 방법론

문장교열은 주어진 원문의 초기 요약을 구성하는 과정을 말한다. 이러한 문장교열의 방법 중에서 (1) Text Compaction[2]은 주어진 문장에서 나타나는 용어에 대해서 개념적으로 모호한 부분을 해결하고 문장내 구성성분들의 문법적인 관계에 따라 문장을 축소하는 방법이다. 이러한 문장의 축소를 위한 기본적인 규칙에는 명사와 형용사를 다른 구성성분들보다 더 중요하다고 보고, 명확한 의미를 가진 명사를 중요시하며, 문장에서의 주절과 반의어들을 중심으로 하는 것 등이 있다. (2) Sentence Reduction[2]은 원문을 축소하는 방법으로 문장에서 나타나는 어휘들의 연결관계를 파악하고, 그 응집도에 따라 문장에서 불필요한 부분을 제거하는 것이다. (3) Sentence Compression[2]은 문법적으로 문장을 분석한 자료를 토대로 문장에서의 주요어를 선별하여 요약 과정을 수행하는 방법이다.

이러한 문장교열 방법론들은 문장이나 문장 내 구성성분들이 가지는 의미론적인 정보보다는 문장

의 통사구조를 바탕으로 얻을 수 있는 문법적인 정보를 토대로 문서요약을 하고자 하는 방법이다.

본 논문에서 실험한 문장추상화는 기존의 문법적인 문서요약 방법과는 달리 문장의 구성성분들 간의 관계와 단어들이 가지는 존재론 정보를 바탕으로 하여 문장이 내포하고 있는 의미를 찾고자 하는 방법이다.

3. 문장추상화의 판단기준

요약과정은 추상화과정의 일종으로, 원문에서 상대적으로 중요한 부분을 발췌하여 일반화시키는 과정을 말한다. 이를 위해서는 원문의 각 문장 구성성분들 중에서 추상화할 대상을 찾아야 한다.

추상화할 대상선정의 판단기준을 구하기 위하여 통사론과 의미론적인 관점에서 접근하였다. 통사론적 입장에서는, 구문분석기를 통하여 얻을 수 있는 문장내 구성성분들이 가지는 통사론적인 중요도를 판단기준으로 하였다. 의미론적 입장에서는, OfN(Ontology for Narratives)[1, 4] 범주정보, 주요어(Head Word)와 그 부속어들의 OfN 범주정보를 판단기준으로 하였다.

4 문장추상화 도구

본 논문에서 이용한 문장추상화 도구들은 (1) 문장의 구성성분들 사이의 관계를 파악하는 구문분석기, (2) 주어진 문장에서 중요정보를 분별하기 위한 온톨로지(Ontology, 존재론), (3) 문장의 구성성분들이 가지는 온톨로지 정보와 추상화를 위한 선호규칙(Preference Rules)을 적용하여 문장추상화 작업을 수행하는 문장추상기이다.

구문분석기로는 LGPI(Link Grammar Parser Interface)[6]를 확장시킨 LGPI+를 이용하였다. LGPI+는 Link Grammar Parser[5]에 대한 SWI-Prolog API를 제공한다.

온톨로지 OfN은 다음의 7가지 범주로 구성되어 있다: 등장인물(Character), 심상(Affect State), 사건(Event), 상태(State), 시간과 공간의 변화(Delta-{Time, Space}), 담화표지(Discourse Marker). 이렇게 설정한 OfN을 구축하기 위해서 먼저 Roget 시소러스[3]의 범주를 심상, 시간과 공간, 사건, 그리고 상태 등으로 재편성하였다. 등장인물 유형에 속하는 어휘들은 고유명사 자원[8]을 이용하여 선정하였다. 담화표지의 경우는 수사구조의 연구결과[7]를 활용하였다. 이와는 달리 시공의 변화는, 구문분석 후 문장의 구성성분간의 상호작용에 의하여 확인되는 유형인 바, 그 기본유형은 시간과 공간이다.

문장추상화 도구로는 Prolog로 구현한 문장추상기를 활용하였다. 이는 OfN, Roget 시소러스, LGPI+등을 이용하여 추상화 선호규칙[1]들을 토대로 추상화 작업을 수행한다.

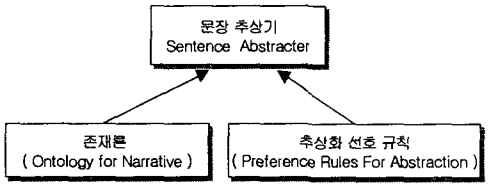

그림 1. OfN과 추상화 선호규칙을 활용한 문장추상기

5 추상화 선호규칙

추상화 선호규칙은 문장에서 그것의 의미범주가 OfN에 있고, 문장추상화의 적절한 대상이 될 수 있는 요점어(Pivot Word)들의 위치를 파악하기 위한 것이다. 대표적인 추상화 선호규칙은 다음과 같다: (1) 최상위구의 주요어를 우선적으로 고려한다. (2) 전치사구의 경우에는 전치사와 그 목적어를 고려한다. (3) 숙어는 하나의 단어로 간주한다. (4) 단어가 의미상 변화를 내포하거나, 심상을 나타낼 때, 고려범위를 확장하여 그 목적어나 수식어들의 주요어를 고려한다. (5) 단어가 OfN의 범주에 중복되어 있으면, 범주 우선순위를 따른다. (6) 주동사가 의미상 변화를 내포하고 있고, 부속 구성성분의 범주가 시간(또는 공간)일 때는 이 성분의 범주를 시간(또는 공간)의 변화로 정한다. 이러한 방식을 토대로 각 문장마다 요점어들의 위치를 파악한다.

6 문장추상화 적용예

다음의 예문을 생각해보자: *She is a girl that I really like to ask out.* 이에 대한 구문분석 결과

는 [그림 2]와 같다.

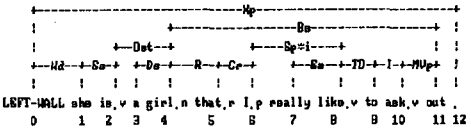

그림 2. 예문의 구문분석 결과

[그림 2]는 주어진 예문에 대한 구문분석 결과로서 문장 구성성분들 간의 다양한 연결유형을 보여줌으로써 문장의 통사구조를 표현한다. 이러한 연결유형들은 한 쌍의 단어를 연결하며 그것들의 문법적인 기능을 표시한다. 예문에서 나타나는 각 연결유형은 다음과 같다: (1) Xp는 좌벽(Left Wall)과 문장의 끝을 연결한다. (2) Bs는 명사와 관계절을 연결한다. (3) Ost는 동사와 단수명사인 보어를 연결한다. (4) Sp*i는 주어와 동사를 연결하며 주어가 I인 경우이다. (5) MVp는 동사와 그것의 수식어구를 연결한다. (6) Wd는 주절을 좌벽에 연결하기 위해 사용한다. (7) Ss는 주어와 동사를 연결하며 주어가 단수명사인 경우이다. (8) Ds는 명사에 대한 한정사가 단수임을 나타낸다. (9) R은 명사와 관계사를 연결한다. (10) Cr은 접속사와 종속절의 주어를 연결한다. (11) Em은 동사를 수식하는 부사를 나타낸다. (12) TO는 부정사의 보어를 가지고 있는 동사와 형용사를 to에 연결한다. (13) I는 부정사의 동사와 to를 연결한다.

이러한 문장분석 결과를 기계 가독형으로 바꾸어서 문장추상기에 적용하여야 한다. [그림 3]은 [그림 2]가 나타내는 예문의 구문분석 정보를 LGPI+를 통하여 출력한 결과이다. 연결유형은 문장의 단어에 대한 색인번호로 대상을 지정한다.

```
/* link information */
linkage(1,
[link([r, m], connection(10-11, mv-[p], ask(v),
      out(_G1366))),
 link([r, m], connection(9-10, i-[], to(_G1331),
      ask(v))),
 link([r, m], connection(8-9, to-[], like(v),
      to(_G1303))),
 link([r, m], connection(7-8, e-[m], really(_G1271),
      like(v))),
 link([r, m], connection(6-8, s-[p, _G1232, i], i(p),
      like(v))),
 link([r, m], connection(5-6, c-[r], that(r), i(p))),
 link([m], connection(4-5, r-[], girl(n), that(r))),
 link([r, m], connection(4-11, b-[s], girl(n),
      out(_G1141))),
 link([m], connection(3-4, d-[s], a(_G1106),
      girl(n))),
 link([m], connection(2-4, o-[s, t], is(v), girl(n))),
 link([m], connection(1-2, s-[s], she(_G1043),
      is(v))),
 link([m], connection(0-1, w-[d], left-wall(_G1013),
      she(_G1015))),
 link([], connection(0-12, rw-[], left-wall(_G983),
      right-wall(_G985)))],
s(np(she/1), vp((is)/2, np(np(a/3, girl/4),
  sbar(whnp(that/5), s(np(i/6), vp(advp(really/7),
  like/8, s(vp(to/9, vp(ask/10, pp(out/11))))))))))).
```

그림 3. 예문의 구문분석 정보

문장에서의 요점어를 선별하기 위해 앞에서 정의한 추상화 선호규칙에 따라 구문분석기에서 얻은 정보와 각 단어에 대한 OfN 범주정보를 이용한다. 단어가 OfN의 복수 범주에 해당될 경우에는 다음과 같은 범주 우선순위를 따라 해당범주를 지정한다: ① Character, ② Affect State, ③ Cue Phrase, ④ Event, ⑤ State, ⑥ Space, ⑦ Time.

[표 1]은 위 예문에서 요점어들을 얻은 결과이다. 문장추상기는 요점어에 대한 정보를 (1) 요점어, (2) OfN 범주, (3) 요점어에 대한 부가정보의 형태로 받아들인다. 요점어가 가지는 부가정보는 주로 Roget 시소러스의 표제정보이다. 만약에, 요점어가 등장인물인 경우에는 별도의 부가정보를 제공한다.

표 1. 예문의 요점어

단어	OfN 범주	요점어에 대한 부가정보
i	character	[man, vitality(special), vitality]
ask	affect state	[request, volition(special, intersocial), volition(intersocial)]
girl	character	[woman, vitality(special), vitality]
like	affect state	[desire, affection(prospective), affection(personal)]
really	affect state	[wonder, affection(contemplative), affection(personal)]
out	state	[exteriority, general, dimension(centrical)]

이상의 과정을 거쳐 문장추상기를 통해 [그림 4]의 결과를 얻을 수 있다.

```
sent(1, 1/2):[
  state([existence, being]): (is)/ (girl<-i)
],
sent(1, 2/2):[
  affect_state([wonder, contemplative])
          :really/ ( girl<-i),
  affect_state([desire, prospective, personal])
          :like/ ( girl<-i),
  affect_state([request, special, intersocial])
          :ask/ ( girl<-i),
  state([exteriority, general, centrical])
          :out/ (girl<-i)
]
```

그림 4. 문장추상화 결과

[그림 4]의 문장추상화 결과는 문장의 요점어와 그에 대한 OfN 범주, 등장인물, 요점어의 부가정보를 보여주고 있다. 부가정보를 토대로 한 요점어의 추상화된 개념은 다음과 같다: (1) 실체, 존재, (2) 놀라움, 관조적, 묵상적, (3) 갈망, 개인적인 기대, (4) 요청, 상호사회적이고 자발적인 의지작용, (5) 외형적, 중심 영역, 공간 등이다. 이러한 추상화된 개념들이 문장에서 is, really, like, ask, out과 같은 어휘로 표상되어 있다. 즉, 예문은 girl을 향한 i의 심상이 간절히 원하고 있으며, 자발적인 의지작용에 의한 요청의 내용이 내재되어 있음을 알 수 있다.

위에서 보인 문장추상화 과정을 한 문단에 적용해 보았다. 문단 추상화의 경우에는, 화제와 밀접한 관계가 있는 문장을 선택해야한다. 이를 위해서는 문단을 구성하고 있는 문장들에 대한 추상화 작업이 선행되어야 한다. [그림 5]는 한 여인에게 관심을 가지고 있는 어떤 사람의 이야기이며, [그림 6]은 문장추상기가 이 문단을 처리한 결과이다.

There is this girl I see all the time. She works in customer service departments at two places I often patronize. She has a nice smile and seems very friendly, and I'd love to ask her out for lunch or dinner sometime. However, the only time I ever see her is when she is at work, and I worry that asking her while she is busy with other customers would be inappropriate. I thought about handing her a note but also thought that would be inappropriate. I'd really like to ask her out, but don't know how.

그림 5. 추상화할 문단

```
sent(1, 1/2):[
 state([existence, being]): (is)/ (girl<-i)  ]
sent(1, 2/2):[
 event([vision,  perception, sensation]):see/ (girl<-i)]
sent(2, 1/2):[
 affect_state([ornament, discriminative, personal])
          :works/ (girl<-i),
 affect_state([good, volition, general])
          :service/ (girl<-i),
 event([purchase, interchange(property),
          relation(possessive)]):customer/ (girl<-i),
 state([duality, determinate, number,]):two/ (girl<-i),
 state([circumstance, relative, existence])
          :places/ (girl<-i) ]
```

그림 6. 문장추상화를 문단에 적용한 결과

```
sent(2, 1/2):[
 affect_state([ornament, discriminative, personal])
          :works/ (girl<-i),
 affect_state([good, object, volition])
          :service/ (girl<-i),
 event([purchase, property, possessive])
          :customer/ (girl<-i),
 state([duality, number, determinate])
          :two/ (girl<-i),
 state([circumstance, relative, existence,])
          :places/ (girl<-i)
]
sent(2, 2/2):[
 affect_state([aid, active, antagonism])
          :patronize/ (girl<-i) ]

          - 중간 생략 -

sent(7, 1/2):[
 affect_state([wonder, contemplative])
          :really/ ( girl<-i),
 affect_state([desire, prospective, personal])
          :like/ ( girl<-i),
 affect_state([friendship, social, sympathetic])
          :know/ (girl<-i),
 state([exteriority, centrical])
          :out/ (girl<-i)
 state([absolute, existence, modal])
          :do/ (girl<-i),
 state([negation, communication])
          :not/ (girl<-i),
 cue_phrase([unconformity, order, category])
          :but/ (girl<-i)
]
sent(7, 2/2):[
 state(method, path, prospective])
          :how/ (girl<-i),
 state([absolute, existence, modal])
          :do/ (girl<-i)
]
```

그림 6. 문장추상화를 문단에 적용한 결과 (계속)

이러한 문단내의 문장들에 대한 추상화 결과에서 상대적으로 다른 문장과의 연결집중도가 높은 문장을 문단의 주제문으로 할 수 있다. 문단은 작문의 단위로서, 이야기하고자 하는 화제가 통상 한 개씩 들어간다[9]. 따라서, 원문요약의 출발점을 문단요약으로 볼 수 있다.

7. 결론 및 향후 과제

본 논문에서는 구문분석기의 기능을 확장시킨 LGPI+, OfN 범주정보, 추상화 선호규칙과 OfN을 바탕으로 한 문장추상기를 이용하여 주어진 문장에 대한 추상화를 시도하였다. 이러한 문장추상화 방법은 다음과 같은 의의를 가진다. 첫째, 구문분석기에서 얻은 정보와 문장 구성성분들의 OfN 정보를 이용하여 주어진 원문에서 요점어들을 찾아낼 수 있었다. 둘째, 문장에 내포된 주요개념들이 추상화된 문장에 나타날 수 있음을 확인하였다. 셋째, 문장의 구성성분들이 가지는 의미론적인 문장교열 방법으로 문장추상화가 가능함을 확인하였다.

문장추상화 결과의 정확도를 높이고 그 타당성을 뒷받침하기 위하여 다수의 원문에 문장추상화 방법을 적용하고 보완해야 할 필요가 있다. 이를 위해서 설화를 대상으로 다수의 원문들에 대해 문장추상화를 수행하고 그 결과를 정리하여 지식기반으로 활용할 수 있는 규칙을 정리하고 있다. 향후에는, 이러한 문장추상화 지식기반을 토대로 주어진 문장에 대한 추상화와 문장교열, 나아가 문서요약의 자동화를 위한 보다 나은 방법론을 찾고자 한다.

참고 문헌

[1] Bae, J.-H. J. and Lee, J.-H. "Topic Sentence Selection with Mid-Depth Understanding." Proc. of ICCPOL, pp. 199-204, 2001.

[2] Inderjeet Mani, Automatic Summarization, John Benjamins Publishing Company, 2001.

[3] Roget's Thesaurus. http://promo.net/cgi-promo/pg/t9.cgi?entry=22&full=yes&ftpsite=ftp://ibiblio.org/pub/docs/books/gutenberg/.

[4] 양재군, 배재학. "온톨로지 정보를 이용한 범주 재편성: Roget 시소러스의 경우." 한국정보처리학회, 제 9권, 제 1호, pp.515-518, 2002.