울산대학교 공학연구논문집 제33권 1호 pp. 245~251, 울산대학교.
Journal of Engineering Research Vol. 33, No. 1, pp. 245~251, 2002
University of Ulsan.

문장추상화를 활용한 설화요약*

배재학
울산대학교 컴퓨터·정보통신공학부
680-749 울산광역시 남구 무거동 산의 29번지
jhjbae@ulsan.ac.kr

<요 약>

본 논문에서 구상한 문단기준 문서요약 방법론은 다음과 같은 절차로 구성되어 있다: (1) 문단의 문장들을 추상화시킨다, (2) 잔류 문장구성성분들의 문장간 개연적 연결상황을 확인한다, (3) 연결집중도가 상대적으로 높은 문장을 문단의 화제를 담고 있는 것으로 인정한다. 이 중에서 본 논문에서는 문장추상화에 필요한 구문분석기와 온톨로지를 구체화하였고, 이 도구들을 설화문장 추상화에 활용하였다.

Narrative Summarization with Sentence Abstraction

Jae-Hak J. Bae
School of Computer Engineering and Information Technology
University of Ulsan, Ulsan, 680-749 Republic of Korea
jhjbae@ulsan.ac.kr

<Abstract>

The author proposes a paragraph-based methodology for text summarization, which has the following procedure: (1) applying sentence abstraction to a paragraph, (2) identifying abductive relationships between constituents of abstracted sentences, and (3) selecting a sentence whose connectivity degree is relatively high and taking it as a

* 이 논문은 2001년 울산대학교의 연구비에 의하여 연구되었음.

topic sentence of the paragraph. Among these, in this paper an ontology and a syntactic parser for the abstraction are set up and applied to abstracting sentences in narratives.

1. 서 론

인터넷과 멀티미디어 기술의 발전으로 인하여 개인이 수집할 수 있고 정리해야 하는 정보량이 막대해졌고 또한 그러한 정보는 시시각각으로 변하고 있다. 더욱이 이러한 정보량은, 한 사람의 정보처리 능력범위를 훨씬 상회하고 있다. 원하는 주제에 대한 최신정보를 신속·정확하게 그리고 요약된 형태로 얻고자 하는 바램은 예로부터 있어 왔지만, 정보·지식화 사회가 성숙되어 가고 있는 현재 더욱더 절실하다.

이상적인 요약은 원문(덩이글, Text)에 대한 깊은 이해를 바탕으로 이루어져야 한다. 그러나 현재까지 개발된 자연어 이해에 필요한 이론과 기술 그리고 축적된 지식베이스 등이 불충분하기 때문에 그것이 용이하지가 않다. 이러한 불충분성은 단기간에 해결될 문제가 아니다. 따라서 현시점에서는 원문에 나타나 있는 언어학적인 단서(Cue)를 충분히 활용하는 원문요약방법론을 개발하는 것이 보다 현실적이다[4, 5, 10]. 표층적 원문이해(Shallow Text Understanding)에 기반한 방법론의 개발이 바로 그것인데, 여기에 동원할 수 있는 언어학적인 자원으로는 견실한 구문분석기(Robust Parser), 시소러스(Thesaurus, 유의어 사전), 수사관계(Rhetorical Relation), 그리고 어휘사슬(Lexical Chain) 등과 같은 것이 있다.

원문요약은 글에 나타난 화제(Topic)들에 대한 논리적 연접의 추적과 그 결과의 요령있는 정리이다. 요약대상이 되는 원문은 그 표면적 구조로 보아 장(Chapter), 절(Section), 문단(Paragraph), 문장(Sentence), 절(Clause), 구(Phrase), 단어(Word), 글자(Letter) 등으로 세분할 수 있다. 여기에서 문단은 작문의 단위로서, 저자가 이야기하고자 하는 화제가 통상 한 개씩 들어간다[13]. 이러한 점에 착안하여 문단요약을 원문요약의 출발점으로 삼는다.

문단 요약과정은 일종의 추상화 과정이다[1]. 추상화에는 추출할 것에 대한 선별작업이 수반된다. 문단 추상화의 경우, 화제와 밀접한 관계가 있는 문장을 선택해야 하고 이렇게 발췌된 문장들에 대한 또 한번의 추상화 작업이 필요하다. 그러나 화제관련성이 높은 문장을 선정하는 작업이, 이해를 수반한다는 점에서, 그렇게 용이하지 않다. 그래서 전자와는 반대방향으로 문장 추상화에서 시작하여 문단 추상화로 접근해 가는 방식을 고려할 수 있다. 이와 같은 문단 추상화를 본 논문에서는 문장 추상화를 통해 실현하는 방법을 고안하고, 그것을 설화문단 추상화에 활용하였다.

2. 설화용 존재론: OfN

독자는 설화를 읽으면서 그럴듯한 추론을 하게 된다. 이 추론은 상식과 글의 내용에 관련된 주변지식에 기초한다. 그 결과 문장이나 저자가 다루는 화제들에 대한 논리적 연접을 추적할 수 있고, 종국에는 글 내용에 대한 심층적(In-Depth) 이해에 도달한다. 설화를 심층적으로 이해하는데 필요한 상식과 주변지식[3]은 (1) 등장인물의 심상과 동기, 목적, 그

리고 계획, (2) 인과적으로 연결된 사건과 상태, 그리고 (3) 등장인물이나 대상, 그리고 공간영역 등에 대한 정적인 속성 등에 관련된 것이다.

한편, 본 논문에서 취하는 원문요약방법론은 심층적(In-Depth)이 아닌 원문에 대한 중층적(Mid-Depth) 이해를 바탕으로 하고 있다[2]. 이것은 가용 언어학적 도구나 지식을 최대한 활용하여 표층적 원문이해의 방식의 제약점을 극복하고 심층적 이해에 접근하는 방식이다. 이러한 요약방법론을 구체화시키는 과정에서, 설화를 중층적으로 이해하는데 사용할 존재론(온톨로지, Ontology) - *OfN*(Ontology for Narratives)[2] - 을 설정하였다. 이 *OfN*에는 다음과 같은 7가지 유형(Type)이 포함되어 있다: (1) 등장인물(Character), (2) 심상(Affect State), (3, 4) 시공의 변화(Delta-{Space, Time}), (5) 사건(Event), (6) 상태(State), 그리고 (7) 담화표지(Discourse Marker) 또는 단서구(Cue Phrase).

이렇게 설정한 *OfN*을 구축하기 위해서 먼저 Roget 시소러스(Roget's Thesaurus)[12]의 범주를 심상, 시간과 공간, 사건, 그리고 상태 등으로 재편성하였다. 등장인물 유형에 속하는 어휘들은 고유명사 자원[11]을 이용하여 선정하였다. 담화표지의 경우는 수사구조의 연구결과[6]를 활용하였다. 이와는 달리 시공의 변화는, 구문분석 후 문장의 구성성분간의 상호작용에 의하여 확인되는 유형인 바, 그 기본유형은 시간과 공간이다.

*OfN*은 문장추상화 과정에서 추출할 문장 구성성분에 대한 선택기준을 제공한다. *OfN*과 함께 견실한 구문분석기도 문장추상화에 활용하는데, 그 과정은 다음과 같다: (1) 주어진 문장을 구문분석하여, 구성성분에 대한 구문상 중요도를 파악한다. (2) 중요 구성성분에 대한 *OfN* 유형을 확인한다. (3) 확인된 *OfN* 유형을 토대로, 필요하다면 구문상 중요도를 재평가한다. (4) *OfN* 유형으로 확인된 것만을 추상화된 문장의 구성성분으로 채택한다.

3. 구문분석기: LGPI+

구문분석기로는 LGPI(Link Grammer Parser Interface)[9]를 확장하여 사용하였다. 이것을 LGPI+라고 하자. LGPI+는 Link Grammar Parser[8]에 대한 SWI-Prolog[14] API(Application Program Interface)를 제공한다. 한편 LGP는 6만 어형을 수록한 사전을 내장하고, 다양한 구문구조를 처리할 수 있다. 이 사전은 필요에 따라 확장이 가능하다. 입력문장에 대한 LGP 구문분석 결과는, 표식고리(Labeled Link)의 집합으로 통사구조(Syntactic Structure)가 표현된다. 표식고리는 한 쌍의 단어를 연결함과 아울러 그것들의 문법적인 기능을 표시한다.

다음 문장을 생각해 보자: *Mike and Paul had been close friends ever since their high school days.* 이에 대한 구문분석 결과는 그림 1과 같다: (1) *Xp*는 문장의 마침표와 좌벽을 연결한다. (2) *MVs*는 동사를 접속사와 연결시킨다. (3) *Jp*는 전치사와 그것의 복수형 목적어를 연결한다. (4) *Opt*는 *be* 동사를 복수명사에 연결하고 *there* 구문 후처리에 참가한다. (5) *Dmc*는 정관사와 복수명사를 연결한다. (6) *Wd*는 평서문에서 주부를 좌벽에 연결한다. (7) *Ss*는 단수명사를 단수동사형과 연결한다. (8) *PPf*는 *have* 동사를 과거분사형과 연결하고 *it*나 *there* 구문 후처리에 참가한다. (9) *A*는 한정형용사를 명사와 연결한다. (10) *ID[X][Y]*는 숙어단어를 일렬로 연결한다. 여기에서 *X, Y*는 임의의 영문자이다. 마지막으로 (11) *AN*은 수식명사를 명사와 연결한다.

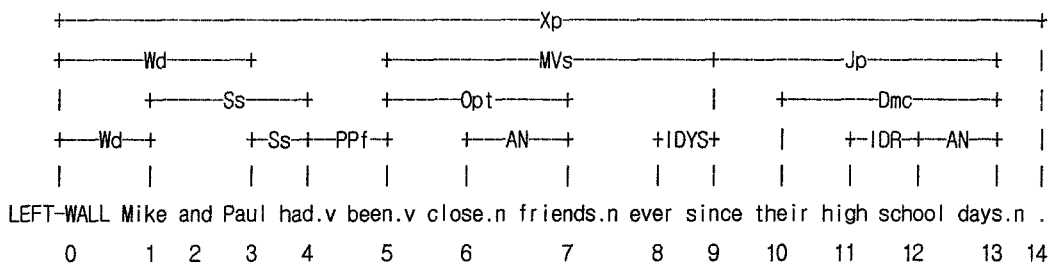


그림 1. 예문에 대한 LGP와 LGPI+의 출력결과

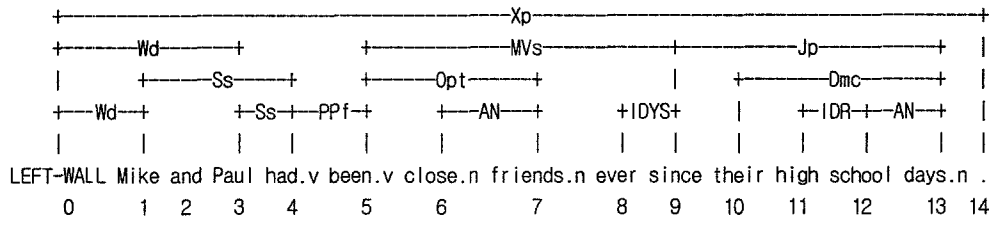
```
[link([m], connection(3-4, s-[s], paul(_G1443), had(v))),
 link([m], connection(0-3, w-[d], 'left-wall'(_G1413), paul(_G1415))),
 link([m], connection(11-12, idr-[], high(_G1383), school(_G1385))),
 link([m], connection(12-13, an-[], school(_G1356), days(n))),
 link([m], connection(10-13, d-[m, c], their(_G1329), days(n))),
 link([m], connection(9-13, j-[p], since(_G1296), days(n))),
 link([m], connection(8-9, idys-[], ever(_G1266), since(_G1268))),
 link([m], connection(6-7, an-[], close(n), friends(n))),
 link([m], connection(5-7, o-[p, t], been(v), friends(n))),
 link([m], connection(5-9, mv-[s], been(v), since(_G1181))),
 link([m], connection(4-5, pp-[f], had(v), been(v))),
 link([m], connection(1-4, s-[s], mike(_G1119), had(v))),
 link([m], connection(0-1, w-[d], 'left-wall'(_G1089), mike(_G1091))),
 link([], connection(0-14, rw-[], 'left-wall'(_G1059), 'right-wall'(_G1061)))]
```

그림 1. 예문에 대한 LGP와 LGPI+의 출력결과 (계속)

4. 문장추상기: SABOT

문장추상화 과정에서는 구문분석된 결과에서 주어, 동사, 목적어, 그리고 동사 수식어 등과 같은 어구들을 고려하되, 최상구(Top-Level Phrase)의 주요어(Head Word)가 우리의 주된 관심대상이 된다. 전치사구의 경우, 전치사와 그 목적어만 고려한다. 숙어(Multiword Unit)는 한 단어로 취급한다. 동사가 의미상 변화를 내포하고 있거나 검토중인 단어가 심상에 연관되어 있을 때에는, 검토심도를 한 단계 깊게 한다. 이 경우, 문제단어의 목적어구나 수식어구의 주요어를 검토대상으로 삼는다. 만일 어떤 단어의 *OfN* 범주값이 2개 이상이면 최소값을 선택한다. 본동사가 의미상 변화를 포함하고 있고 이 동사구의 구성성분 유형이 시간 (또는 공간)이라면, 구성성분의 *OfN* 범주는 *delta-time* (또는 *delta-space*)이 된다. 이러한 방식으로 우리는 한 문장 안에서 추상화에 참가할 요점어(Pivot Word)를 선택할 수 있다. 요점어란 *OfN* 범주가 확인된 문장추상화의 후보단어이다.

이러한 문장추상화 과정을 앞서 본 예문에 적용한 결과가 그림 2에 나타나 있다. Prolog로 구현한 문장추상기 ***SABOT***가 요점어 *mike, paul, been, friends*, 그리고 *ever since* 등을 선별해내었다. 문장표식 *sent(X,Y)*에서 *X*는 문단 내에서 위치를 그리고 *Y*는 절의 위치를 각각 나타낸다. 술어 *affect_state*와 *cue_phrase*는 각각 *OfN*의 심상과 담화표지에 대응한다.



[sent(1, 1/2): [affect_state([friend, social, sympathetic]):friends/ (mike<->paul),
state([identity, absolute, relation]):been/ (mike<->paul)],
sent(1, 2/2): [cue_phrase([temporal, durative]):[ever, since]/ (mike<->paul)]]

그림 2. 추상화된 예문

문장추상기 SABOT을 문단을 구성하고 있는 각 문장에 적용하여 추상화된 문단을 얻는다. Mike와 Paul에 대한 이야기[7]의 경우를 보자. 그림 3에 추상화시킬 한 문단이 있다. SABOT이 문단을 처리한 결과가 그림 4에 보인다.

Mike and Paul had been close friends ever since their high school days. But now Mike wanted Paul out of town for a few days so that *he could build a patio in Paul's backyard as a surprise birthday present.* He suggested to Paul that he get away for a weekend, but *Paul said he wasn't interested.* On another occasion Mike casually spoke about the joys of fishing or camping trips. But Paul told him he enjoyed puttering around the house much more. Paul was getting very settled in his old age.

그림 3. 추상화시킬 문단

[sent(1, 1/2): [affect_state([friend, social, sympathetic]):friends/ (mike<->paul),
state([identity, absolute, relation]):been/ (mike<->paul)],
sent(1, 2/2): [cue_phrase([temporal, durative]):[ever, since]/ (mike<->paul)]]

[sent(2, 1/2): [affect_state([requirement, conceptional, prospective]):wanted/ (paul<-mike),
delta(space):[out, of, town]/ (paul<-mike),
delta(time):days/ (paul<-mike),
cue_phrase([temporal, repetitive]):[but, now]/ (paul<-mike)],

그림 4. 추상화된 문단

sent(2, 2/2): [affect_state([wonder, contemplative]):surprise/ (paul<-mike),
affect_state([giving, intersocial]):present/ (paul<-mike),
delta(space):patio/ (paul<-mike),
delta(space):backyard/ (paul<-mike),
event([production, power, causation]):build/ (paul<-mike),
cue_phrase([causal, specific, purpose]):[so, that]/ (paul<-mike)]]

[sent(3, 1/1): [affect_state([advice, voluntary]):suggested/ (paul<-mike),
delta(space):away/ (paul<-mike),
delta(time):weekend/ (paul<-mike),
state([conversion, change]):get/ (paul<-mike)]]

[sent(4, 1/2): [cue_phrase([adversative, proper]):but/ (paul<-paul)],
sent(4, 2/2): [affect_state([pleasurableness, passive]):interested/ (paul<-paul),
state([identity, absolute, relation]):'wasn\'t'/ (paul<-paul)]]

[sent(5, 1/1): [affect_state([pleasure, passive, personal]):joys/ (paul<-mike),
event([journey, land, motion]):trips/ (paul<-mike)]]

[sent(6, 1/2): [cue_phrase([adversative, proper]):but/ (paul<-paul)],
sent(6, 2/2): [affect_state([pleasure, passive, personal]):enjoyed/ (paul<-paul),
event([inactivity, voluntary, individual]):puttering/ (paul<-paul)]]

[sent(7, 1/1): [state([identity, absolute, relation]):was/ (paul<-paul),
state([stability, change]):settled/ (paul<-paul)]]

그림 4. 추상화된 문단 (계속)

5. 결 론

현재의 자연어이해 이론과 기술의 수준 그리고 가용 지식베이스의 완전성 수준에서 볼 때, 심층적 원문이해의 성취는 단기적으로 불가능하다. 따라서 문제의 본질상 원문이해가 필요한 문서요약의 경우, 표층적 원문이해에 기반을 둔 방법론에 의존할 수밖에 없다.

일반적으로 요약대상이 되는 원문은 표면상 계층구조를 가진다. 그 중에서 문단은 작문의 한 단위인데, 한 가지 화제를 다룬다는 독립성을 지닌다. 이 점에 착안하여 원문요약 문제를 문단요약 문제로 환원하여 생각할 수 있다. 한편, 문단 요약과정은 일종의 추상화 과정이다. 문단 추상화과정에서 화제관련성이 높은 문장을 선정해야 하지만, 이 작업은 이해를 수반한다는 점에서 그렇게 용이하지 않다. 그래서 본 논문에서는 문장추상화에서 시작하여 문단추상화로 접근해 가는 방식을 취했고 그것을 설화문단 추상화에 활용하였다.

본 논문에서 소개한 설화요약 방법론을 개관하면 다음과 같다: (1) 문단 안의 문장들을 추상화시킨다, (2) 이들의 논리적 연결상황을 확인한다, (3) 연결집중도가 상대적으로 높은

것을 그 문단의 화제를 담고 있는 문장으로 인정한다, (4) 추상화된 문단들을 대상으로 (2, 3)의 과정을 적용하여 원문의 주제전개 상황을 파악한다. 이러한 방법론의 구현에서, 비교적 풍부한 어휘를 내장하고 있는 존재론 *OfN*과 실용적인 구문분석기 *LGPI+*, 그리고 문장추상기 *SABOT* 등의 유용성을 본 논문에서 확인해 보았다. 구체적으로, *OfN*은 (1, 2) 단계에서, *LGPI+*는 (1) 단계에서, 그리고 SABOT는 (1, 4) 단계에서 그 유용성을 발휘한다.

참고문헌

1. Bae, J.-H. J. and Lee, J.-H. Another Investigation of Automatic Text Summarization: A Reader-Oriented Approach. In Proceedings of ANZIIS '94 (Australian and New Zealand Conference on Intelligent Information Systems), pp. 472-476, 1994.
2. Bae, J.-H. J. and Lee, J.-H. Topic Sentence Selection with Mid-Depth Understanding. In Proceedings of ICCPOL2001 (The 19th International Conference on Computer Processing of Oriental Languages), pp. 199-204, 2001.
3. Dahlgren, K. Naive Semantics for Natural Language Understanding. Boston: Kluwer Academic Publishers, 1988.
4. Jones K. S., What might be in a summary? In G. Knorz, J. Krause, and C. Womser-Hacker, editors, Information retrieval '93: von der modellierung zur anwendung, pp. 9-26. Konstanz, Universitatsverlag Konstanz, 1993.
5. KAML Research Group. Research on Text Summarization (fragments of a grant proposal, HTML). http://www.csi.uottawa.ca/~szpak/proposals/text-summ-1996_ToC.html, 1996.
6. Knott, A. A Data Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh, 1996.
7. Lehnert, W. G. Plot units: A narrative summarization strategy. In W. G. Lehnert and M. H. Ringle (Eds.), Strategies for natural language processing, Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.
8. Link Grammar. http://www.link.cs.cmu.edu/link/.
9. LPG - Link Grammer Parser Interface. http://gollem.swi.psy.uva.nl/twiki/pl/bin/view/Library/LinkGrammer.
10. Paice, C. D. Constructing literature abstracts by computer. Information Processing and Management, Volume 26, Issue 1, pp. 171-186, 1990.
11. Proper Names Wordlist. http://clr.nmsu.edu/cgi-bin/Tools/CLR/clrcat#I4.
12. Roget's Thesaurus. http://promo.net/cgi-promo/pg/t9.cgi?entry=22&full=yes&ftpsite=ftp://ibiblio.org/pub/docs/books/gutenberg/.
13. Strunk, W. Jr. The Elements of Style. http://www.bartleby.com/141/.
14. SWI-Prolog. http://www.swi-prolog.org/.