제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

글의 응집성을 포착하기 위한 개연규칙

김곤, 양재군, 김민찬, 배재학
울산대학교 컴퓨터 · 정보통신공학부
e-mail:{gonkim, jgyang, tomatuli}@mail.ulsan.ac.kr,
jhjbae@ulsan.ac.kr

Abductive Rules for Text Cohesion

Gon Kim, Jae-Gun Yang, Min-Chan Kim, Jae-Hak J. Bae
School of Computer Engineering and Information Technology,
University of Ulsan

요 약

본 논문에서는 글의 응집성을 포착하기 위하여 개연규칙을 활용한다. 개연규칙은 문장 구성성분들의 문장간 개연적 연결상황을 나타내고, 글의 인과 성향이나 담화작용을 반영한다. 글을 이해하기 위한 대표적인 속성에는 글에 긴밀성을 부여하는 응집성이 있다. 글의 응집성을 파악하기 위한 대표적인 언어학적 도구나 지식으로는 어휘사슬을 들 수 있다. 이에 본 논문에서는 주어진 예문의 어휘사슬을 개연규칙으로 찾아낸 개연사슬과 비교해 보았다. 그 결과, 중요도가 높은 어휘사슬과 대응하는 개연사슬을 발견할 수 있었다. 개연사슬은 종래의 어휘사슬의 기능을 포함할 뿐만 아니라, 줄거리 단위, 단서구 용법, 문장사이의 개연성 등을 감지하여 문장간의 의미적 연관성을 포착할 수 있다. 이는 개연규칙을 활용하여 글의 화제문을 효과적으로 선별할 수 있음을 보인다.

1. 서론

글의 속성에는 조응과 응집이라는 것이 있다. 조응은 글을 조리있게 하고, 응집은 글에 긴밀성을 부여한다. 잘된 글은 글의 주제와 세부내용을 논리적으로 전개(조응성)하여 글의 뜻을 독자에게 분명하게 전달(응집성)한다. 글을 읽고 이해하기 위해서는 이러한 두 속성을 충분히 활용하여 문맥을 파악해야 한다. 이상적인 요약은 이러한 글의 응집성과 조응성이 상호보완적으로 작용하여야 한다.

글의 속성차원에서 보면, 어휘사슬은 글의 응집성을 표현하고 수사구조는 글의 조응성을 표현한다.

어휘사슬을 활용한 요약 방법론[1, 2]은 표층적 원문이해에 바탕을 둔 것으로 담화구조를 활용한다. 이 방법은 표층적 원문이해에 바탕을 두고 있어서 적용상 제약이 적으나, 형성된 어휘사슬의 정당성과 완전성을 판정할 논리적인 기준이 없다. 특히, 어휘결속을 확인할 때 여러 가지 뜻을 가진 단어의 중의성 해소단계의 문제점을 가지고 있다.

수사관계를 활용한 요약 방법론[1, 3]은 수사 구조이론(Rhetorical Structure Theroy)에 바탕을 둔 것이다. 요약문은 원문의 수사구조를 파악하여 주요문장을 발췌하여 얻는다. 이는 담화표지(Discourse Marker)라는 단서구(Cue Phrase)를 활용하는 표층적 원문이해 방식의 장점과 함께 원문표현 방식이 구조적으로 성형화 되었다는 점에서 심층적 원문이해 방식의 속성을 지니고 있다. 그러나, 담화표지의 중의성 해소, 원문분석용 수사관계 집합의 적정성 또는 완비성, 개별 수사관계의 의미적 독립성과 의존성을 보장해 줄 성형화 작업 등의 문제점 및 개선사항을 내포하고 있다.

본 논문에서는 문서요약의 과정에서 글의 응집성을 포착하기 위한 방법으로 개연규칙을 활용한다. 개연규칙은 글을 논리적으로 전개하고 그 뜻을 분명하게 하는 조응성과 응집성을 파악하기 위한 유용한 언어학적 도구이다. 개연규칙을 활용한 원문이해 방법은 계산가능한 범위에서 심층적인 원문이해를 지향한다. 그 결과 글의 화제문을 효과적으로 선별할 수 있다. 개연규칙을 활용한 원문이해는 종래의 표층적인 요약방법론이 가진 원문이해의 심도를 한 단계 깊게 하는데 상당한 기여를 할 것이다.

2. 글의 응집성

글의 속성 차원에서 보면, 어휘사슬은 글의 응집성(Cohesion)을 표현한다. 응집성은 잘된 글의 속성이다. 즉, 잘 쓴 글은 글의 전개가 분명하여 읽는 이가 그 뜻을 명확히 이해할 수 있다. 이렇게 되기 위

해서는 글의 주제가 분명해야 한다. 또한 이 주제를 독자에게 전달하기 위하여 주제를 보충 설명하는 세부내용이 있어야 한다.

글이 조응적이기 -- 논리적으로 전개되기 -- 위해서는 응집성을 가져야 한다. 또한 글이 응집적이기 -- 그 뜻이 명확해지기 -- 위해서는 조응성을 가져야 한다.

3. 개연규칙

개연규칙은 문장구성성분들의 인과성향이나 담화작용을 반영한다. 이는 문장간 구성성분들의 개연적 연결상황을 나타낸다. 문단의 화제와 관련이 깊은 문장을 선택하는 과정에서 줄거리 단위, 단서구 용법, 문장 사이의 개연성 등을 감지할 수 있는 개연규칙(Abductive Rules)[1, 4]을 통하여 문단내용에 대한 이해심도를 높인다.

3.1 개연규칙 고안

개연규칙은 문장내 구성성분들이 가지는 OfN(Ontology for Narratives)[1, 4, 7] 정보로서 표현된다. 2항 또는 3항인 개연규칙의 일반적인 모습은 다음과 같다.

Ante <= Post [{=+>, =->, =*>} Cons]

여기에서 (1) Ante, Post, Cons는 pred(args)의 형태를 가진다, (2) pred(args)는 OfN에 명시된 개념으로 7가지의 범주로 표현된다, (3) =+> and =->는 Post와 Ante에 제한사항이 있음을 나타내고, (4) =*>는 담화표지가 있음을 나타낸다. (그림 1)은 Mike와 Paul의 이야기[1]를 처리할 때 사용한 개연규칙의 예이다.

(1)	% 마음이 통하면 주고 싶어진다 affection(sympathetic) <= affection(offer)
(2)	% 풀이 죽으면 소극적이 된다 event(inactivity) <= event(descent).
(3)	% 장소를 바꾸고 싶을 때 여행을 한다 delta(space) <= event(journey) =-> *affection(prospective)*.
(4)	% 싫은 것을 권하면 흥미를 끌지 못한다 affection(advice) <= affection(cause(pleasure)) =*> cue_phrase(adversative)

(그림 1) 개연규칙

개연규칙은 크게 2항 규칙과 3항 규칙으로 나눌 수 있다. (그림 1)에서 규칙 (1)과 (2)는 2항 규칙이며, (3)과 (4)는 3항 규칙이다.

3.2 개연규칙 생성

이야기 문장에 일반적으로 적용할 수 있는 개연규칙을 만들기 위해 Dear Abby(http://www.DearAbby.com)에서 선택한 23편의 상담문을 분석하였다. 개연규칙의 추출과정을 요약하면 다음과 같다.

(1) 상담문의 주제 문장을 선택한다. 상담문의 제목과 가장 연관성이 높은 문장을 선택하고, 제목에 내포된 화자의 의도까지를 주제 문장의 범주에 포함시킨다.
(2) 선택한 주제 문장의 개념과 핵심 단어를 선별한다.
(3) 주제 문장의 개념, 핵심단어와 연결시킬 다른 문장의 단어를 추적하여 연결관계를 높여나간다.
(4) 연결시킬 단어는 문장의 주요 구성성분으로 한다.
(5) 정리된 상담문의 연결관계에 따라 개연규칙의 일반적인 형태로 정리한다.
(6) 개연규칙에 참여하는 단어가 OfN의 복수 범주에 해당 될 경우에는 상담문의 내용과 가장 연관성이 높은 범주를 선택한다.
(7) 선정된 OfN 범주에서 대표 개념을 정하고 개연규칙을 단순화한다.

3.4 개연사슬

개연사슬(Abductive Chain)은 문장간 개연성을 포착하기 위한 어휘사슬이다. 이는 추상화된 문장들의 구성성분을 대상으로 조응 및 응집성을 판정하여 적합한 두 문장을 개연고리(Abductive Link)로 각각 연결한다.

개연고리의 요소로 참가할 어휘는 품사정보나 빈출정도에 의거하여 결정되는 것이 아니라, 문장 내에서의 통사적인 그리고 의미적인 역할에 의해 결정된다. 이러한 개연고리의 요소가 될 후보단어는 문장추상기 SABOT[1, 4]의 출력에서 선택된다.

개연고리는 개연규칙의 용례이다. 이 규칙의 개연성은 두 문장 사이에서 문장 구성성분들의 인과성향이나 담화작용을 반영한다. 개연사슬을 형성하는 과정에서 기존의 어휘사슬을 활용한 연구[5]에서 간과되었던 기능어와 빈출어도 적절한 전처리를 통해 사슬의 고리역할을 할 수 있게 개념확장을 하여 개연사슬에 참가하도록 하였다.

3.4 개연사슬 형성

문장추상화와 개연규칙을 적용하여 문단의 주제와 가장 연접한 화제문을 선정하기 위하여 Prolog로 구현한 SICHA(A Situation Chainer)[1, 4]를 활용하였다. SICHA는 문단을 구성하는 문장간 연결집중도를 보여준다. 문장들의 연결집중도는 개연규칙에 부합하는 연결유형의 급수(Degree)로써 나타낼 수 있다. 연결집중도의 일반적인 모습은 다음과 같다.

$$SDs = \{D_1\text{-sent}(N_1), \ldots D_n\text{-sent}(N_n)\}$$

여기에서 SDs는 문장의 연결집중도(Sentence Degrees)를 나타내며 D는 그 문장의 급수(Degree)를, sent(Nn)는 문단내에서 각 문장의 순서(Sentence Number)를 나타낸다. 개연규칙에 의해 한 문장이 다른 문장과 연결되는 경우에 Degree를 1로 한다. (그림 2)는 SICHA를 통해 얻은 문장간 연결 집중도의 한 예이다.

```
SDs = [1-sent(8), 3-sent(10), 3-sent(12), 4-sent(4),
       4-sent(6), 4-sent(15), 5-sent(1), 5-sent(3),
       5-sent(11), 6-sent(14)];
```

(그림 2) 문단내 문장들의 연결 집중도

SICHA는 (그림 2)와 같이 문단내 문장들의 연결

집중도를 오름차순으로 정렬하여 보여준다. (그림 2)의 경우, 문단의 14번째 문장의 연결집중도가 6으로 가장 높게 나타나 있다. 즉, 문단의 주제를 내포하고 있는 문장들의 연결집중도가 높게 나타난다.

4. 글의 응집성 포착

다음은 아인슈타인의 글에서 어휘사슬을 파악한 예[6]이다. 예문에 대한 3개의 어휘사슬이 보인다. 사슬의 고리는 이탤릭체로 표기한 단어들이고 사슬 번호가 부여되어 있다. <표 1>은 생성된 어휘사슬이다. 표의 각 어휘사슬에서 고리들의 의미적 연관성을 파악할 수 있다.

(1) We suppose a very long ***train*$_1$** ***travelling*$_2$** along the ***rails*$_1$** with the constant ***velocity*$_2$** and in the ***direction*$_2$** indicated in Figure 1. (2) People ***travelling*$_2$** in this ***train*$_1$** will with advantage use the ***train*$_1$** as a rigid ***reference-body*$_3$**. (3) They regard all events in ***reference*$_3$** to the ***train*$_1$**. (4) Then every event which takes place along the ***line*$_1$** also takes place at a particular ***point*$_1$** of the ***train*$_1$**. (5) Also, the definition of simultaneity can be given relative to the ***train*$_1$** in exactly the same way as with respect to the ***embankment*$_1$**.

<표 1> 예문에 대한 어휘사슬

번호	어휘사슬
(1)	{train, rails, train, train, train, line, point, train, train, embankment}
(2)	{traveling, velocity, direction, traveling}
(3)	{reference-body, reference}

(그림 3)은 <표 1> 문단의 추상화된 모습이다. 이는 구문분석기 LGPI+[1, 4], 문장추상기 SABOT을 거쳐 얻은 결과이다. 이러한 문단추상화 작업 후에는 개연사슬을 계산한다.

일반적으로, 문단 안에서 각 문장은 다른 문장과 의미적으로 조응하거나 응집한다. 조응과 응집은 문장간 연결도(Connectivity Degree)로 추정할 수 있다. 따라서 문장이나 절의 연결도는 문단 내에서 그것의 중요도를 반영한다고 할 수 있다. 이러한 문장이나 절의 연결성을 개연사슬로 포착한다.

<표 2>는 추상화된 문장들이 개연사슬로 연결된 모습이다. <표 2>의 (2), (3), (4)는 각각 심상, 사건, 상태에 관련된 개연사슬의 형성결과를 보이고 있다. 각 개연사슬과 <표 1>에서 보인 어휘사슬의 고리에 대응하는 어휘들은 볼드체로 표기해 놓았다. (5)는 문장간 연결도를 계산한 결과이다. 다른 문장과의 연결정도가 그 문장의 중요도를 반영한다고 할 때, 문단 안에서 최대의 연결도를 가진 (1), (2), (3)번 문장이 <표 1> 예문의 화제문(Topic Sentence)라고 할 수 있다.

```
sent(1, 1/2):[affect_state([supposition, thought(creative),
                formation(idea), word(faculty(intellectual))])
                :suppose/ (somebody<-we)].
sent(1, 2/2):[event([vehicle, motion(general), motion,
                word(space)]):train/ (somebody<-we),
              event([journey,locomotion(land), motion(general),
                motion,word(space)]):travelling/
                (somebody<-we),
              state([indication, means(natural),
                means(communication(idea)),
                communication(idea),
                word(faculty(intellectual))]):indicated/
                (somebody<-we),
              state([representation, means(natural),
                means(communication(idea)),
                communication(idea),
                word(faculty(intellectual))]):figure/
                (somebody<-we)].
              sent(2, 1/1):[affect_state([use,
                measures(precursory), volition(prospective),
                volition(individual),
                word(power(voluntary))]):use/
                (people<->travelling),
              affect_state([mankind, vitality(special), vitality,
                matter(organic), word(matter)]):people/
                (people<->travelling),
              event([journey, locomotion(land),
                motion(general), motion,
                word(space)]):travelling/
                (people<->travelling),
              event([vehicle, motion(general), motion,
                word(space)]):train/ (people<->travelling),
              state([hardness, matter(solid), matter(inorganic),
                word(matter)]):rigid/ (people<->travelling)].

              ... 중략 ...

sent(5, 1/1):[event([receiving, transfer(property),
                relation(possessive), volition(intersocial),
                word(power(voluntary))]):given/
                (definition<->simultaneity),
              event([vehicle, motion(general), motion,
                word(space)]):train/
                (definition<->simultaneity),
              state([addition, quantity(conjunctive), quantity,
                word(relation(abstract))]):also/
                (definition<->simultaneity),
              state([interpretation, nature(idea(communicated)),
                communication(idea),
                word(faculty(intellectual))]):definition/
                (definition<->simultaneity),
              state([belief, result(reasoning), formation(idea),
                word(faculty(intellectual))]):be/
                (definition<->simultaneity),
              state([relation, relation(absolute), relation,
                word(relation(abstract))]):relative/
                (definition<->simultaneity),
              state([limitation, relation(partial), relation,
                word(relation(abstract))]):exactly/
                (definition<->simultaneity)].
```

(그림 3) 추상화 된 문단

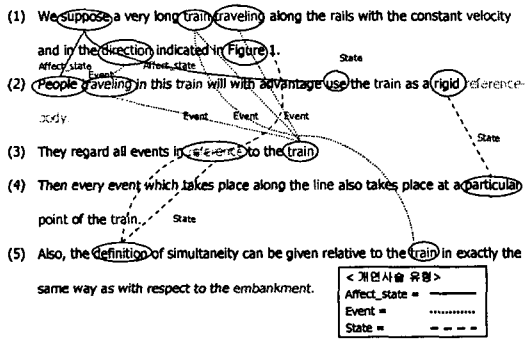
이러한 연결도의 크기 순으로 문장을 추출하면 문단에 대한 발췌요약을 얻을 수 있다.

<표 2>의 결과에서 어휘사슬과 대응하는 것은 (3) 사건에 관련된 개연사슬이다. (2), (3)은 어휘사슬과는 대응하지는 않으나, 주어진 글의 화제문을 선정하는 과정에서 요점어로 선별된 어휘들로 연결된 사슬이다. 이는 문장간 연결도에 반영된다.

<표 2> 개연사슬 형성

(1)	?- process_cue_phrases(CPs).
(2)	?- process_affect_states(CPs). CPs=[(sent(2,1/1)->[sent(1, 1/_G448)])] ; people/use suppose
(3)	?- process_events(CPs). CPs = [(sent(2,1/1)->[sent(1,2/_G512)]), traveling direction (sent(3,1/1)->[sent(1, 2/_G547), sent(2, 1/_G535)]), train train traveling (sent(5, 1/1)->[sent(1, 1/_G720)])] ; train train
(4)	?- process_states(CPs). CPs=[(sent(4, 1/2)->[sent(2, 1/_G581)]), particular rigid (sent(5, 1/1)->[sent(1, 2/_G711), sent(3, 1/_G759)])] ; definition figure reference
(5)	?- test(SDs). SDs = [1-sent(4), 2-sent(5), 3-sent(1), 3-sent(2), 3-sent(3)] ;

(그림 4)는 연결된 개연사슬 유형을 보인다. <표 1>의 어휘사슬에서 (1), (2)번 사슬만이 개연사슬의 유형에서 나타나고 있다. 그러나, 어휘사슬에 속한 모든 어휘들을 대상으로 개연사슬이 형성되지는 않았다. 이는 어휘사슬의 경우, 같은 뜻을 지닌 동의어나 어휘의 반복을 기준으로 사슬을 형성하는 반면, 개연사슬의 경우에는 개연규칙에서 설명하는 주요 문장구성성분간의 인과성을 기반으로 하고 있기 때문이다. 개연사슬은 종래의 어휘사슬의 기능을 포함하면서, 줄거리 단위, 단서구 용법, 문장사이의 개연성 등을 감지하고 있다.



(그림 4) 개연사슬 연결 유형

5. 결론

현재의 자연어 처리 이론과 기술수준 그리고 축적된 기계가용 지식자원의 완전성 수준에서 볼 때, 원문에 대한 심층적 이해의 성취는 응용영역을 제한하지 않는 한, 단기적으로 불가능하다. 그러나, 현재까지의 언어처리 자원과 기술을 충분히 활용하여 원문에 대한 표층적 이해(Shallow Understanding) 수준에만 머무르지 않고, 글의 화제연접 상황을 밝힘으로써 원문에 대한 이해심도를 높이는 방법은 가능하다. 이는 글의 논리전개와 뜻의 명확성에 관여하는 응집성과 조응성을 파악함으로써 가능하다.

본 논문에서는 개연규칙(Abductive Rule)을 활용하여 글의 응집성을 파악하고자 하였다. 개연규칙은 문장의 구성성분들 간의 개연적 연결상황을 나타내며, 글의 인과성향이나 담화작용을 반영한다. 개연규칙의 용례인 개연고리로 이루어진 개연사슬은 추상화된 문장들의 구성성분들을 대상으로 조응이나 응집성을 판정하여 적합한 두 문장을 연결하게 된다.

글의 응집성을 파악하기 위한 대표적인 언어학적 도구나 지식으로는 어휘사슬을 들 수 있다. 이에 본 논문에서는 주어진 예문의 어휘사슬과 개연규칙으로 찾아낸 개연사슬을 비교해 보았다. 그 결과, 중요도가 높은 어휘사슬과 대응하는 개연사슬을 발견할 수 있었다. 또한, 개연사슬은 어휘사슬의 기능을 포함하면서, 문장사이의 개연성을 감지하는 추가적인 사슬도 포착하였다.

개연규칙을 활용한 원문이해는 낙관적이다. 그 이유는 (1) 온톨러지 OfN이 개방형이어서 새로운 개연규칙을 쉽게 추가할 수 있다. (2) 개연규칙에 담화기능을 추가하여 체계적으로 정교하게 만들 수 있다. (3) 개연규칙 자동생성 방법에 관한 연구[8]가 활발히 이루어지고 있다.

<Acknowledgements>

본 연구는 한국과학재단 목적기초연구 R05-2004-000-12362-0 지원으로 수행되었음. 또한 울산대학교 디지털 제조 정보기술 연구센터 (DMITRC)의 부분적인 지원을 받았음.

[참고문헌]

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

[2] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization", in *Proc. ISTS'97 (The Intelligent Scalable Text Summarization Workshop, ACL)*, Madrid, Spain (Jul. 1997), pp. 10-17.

[3] D. Marcu, "From Discourse Structures to Text Summaries", in *Proc. ACL'97 and EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, pp. 82-88, 1997.

[4] 김곤, 양재군, 배재학, 이종혁, "문장추상화: 개념추상화를 도입한 문장교열", 정보처리학회논문지B, 제11-Brnjs 제5호, pp.563-572, 2004.

[5] J. Morris, and G. Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", *Computational Linguistics* 17:"1, pp. 21-48, 1991.

[6] D. St-Onge, "Detection and Correction Malapropi는 with Lexical Chains", M.Sc. Thesis, University of Toronto. Department of Computer Science Technical Report CSRI-319, March 1995.

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

[8] 양재군, 강인수, 배재학, 이종혁. "Factotum SemNet을 활용한 개념간 개연사슬 발견", 한국정보과학회 추계학술발표대회논문집, 제30권 제2-1호, pp.139-141, 2003.