요약문 생성을 위한 중간 개념 표현*

서연경[U] 노태길 이상조
경 북 대 학 교 컴 퓨 터 공 학 과
ykseo@sejong.knu.ac.kr nayas@sejong.knu.ac.kr sjlee@bh.knu.ac.kr

Intermediate Concept Representation for Automatic Summary

Youn-Kyoung Seo[U] Tae-Gil Noh Sang-Jo Lee
Dept. of Computer Engineering, KyoungPook University

요 약

사건, 사고 관련 기사의 요약은 단순히 원문이 무엇을 말하는 가를 지시하는 것보다 가능한 요지를 판독하면서 필요한 정보를 누락시키지 않고 표현할 수 있는 것이 바람직하다. 이를 위하여 본 논문에서는 사건, 사고 관련 기사의 자동 요약문 생성을 위한 중간 개념 표현 방법을 제안한다. 단락 자동 구분을 통한 중요 문장 추출을 거쳐 각 단락의 중심문장을 파악하고, 단락내의 정보들을 의미 파악된 중심 문장에 추가, 병합하여 단락의 내용을 대표하는 Paragraph Representation Structure(PRS)를 생성한다. 이들은 통합과정을 거쳐 하나의 Unified Representation Structure(URS)로 만들어지며, 이것은 중간 개념 표현으로 다국어 자동 요약문 생성을 위한 기반이 될 수 있다. 본 연구에 이용한 코퍼스는 비행기, 선박, 차량, 열차 사고와 화재 폭발 및 사건 관련 신문 기사를 대상으로 한다.

1. 서론

빠른 속도의 네트워과 인터넷 보급으로 인해 대용량의 정보를 좀더 빠르고 정확하게 검색할 수 있는 여러 방법이 모색되고 있다. 지금까지의 정보검색 시스템은 그러 지능적이지 못해서 실제로 원하는 정보 이외의 다른 많은 정보를 제시하고 있으며, 적절한 키워드 사용에 미숙한 사람들은 좋은 검색 결과를 얻기가 어렵다. 하지만 문서가 검색되었을 때 그 문서의 요약문이 있다면, 그 문서가 소용에 닿는 것인지를 훨씬 정확하고 빠르게 파악할 수 있으며, 검색을 위한 DB구축에도 활용할 수 있다. 또한 요약문이 한 언어로 표현되어 제공되어지는 것보다 다른 언어의 요약문도 제공된다면 관련 문서의 파악에 많은 도움을 줄뿐만 아니라 그 활용 값어치가 증가할 것이다.

본 논문에서는 사건, 사고 관련 기사의 자동 요약문 생성을 위한 중간 개념 표현 방법을 제안한다. 우선 문서로부터 핵심적인 내용들을 인식하기 위해 단락을 나누어 중요 문장을 추출[7]하는 과정이 필요하다. 이것은 의미 해석을 위한 문장의 수를 줄이는 효과를 가져온다. 다음 단계로, 추출된 각 중요 문장이 포함된 단락으로부터 정보들을 추가하거나, 의미를 파악하여 Sowa의 개념 그래프[5]를 응용한 단락 재표현 구조(PRS)를 생성한다. 그리고 이들 각 PRS들을 통합하여 문서의 요약된 내용을 표현하는 Unified Representation Structure(URS)를 만든다. 이것은 향후 다언어 자동 요약문 생성을 위한 중간 개념 표현 구조로 응용될 수 있을 것이다.

2. 관련연구

자동 문서 요약 시스템은 문장 구성 요소 추출 시스템, 문장 이해 기반 시스템, 혼합된 형태의 시스템, 그리고 틀(template)기반 시스템 등 네 가지 형태로 구분할 수 있다. 문장 구성 요소 추출 시스템은 원문의 각 문장이 가지고 있는 언어와 구조 정보를 이용하여 문장을 단순히 추출, 나열하거나 재정렬한다.[1] 그래서 시스템은 간단하나 요약이 부자연스럽고 원문에 나오는 문장에 의존하는 단점이 있다. 문장 이해 기반 시스템[2]은 인간이 문서를 요약하는 과정을 고도의 자연 언어 처리 과정을 통해 재현하려는 시스템으로 주제를 표현하고 있는 정보를 식별한 후 문장 생성을 통해 요약하며 고품질의 자연스러운 요약을 생성할 수 있으나 복잡한 자연어처리 과정과 적용 분야마다 영역지식이 필요하여 시스템 구현이 어려운 단점이 있다. 틀 기반 시스템[4]은 요약문에 포함되어야 할 개념을 수작업을 통해 틀로 정의하고 텍스트 분석을 통해 미리 정의한 틀을 메운 후 요약문을 생성한다. 하지만 분야마다 틀을 재정의 해야하며 요약문의 형식도 획일화된다. 마지막으로 혼합 형태 시스템[3]은 개념 추출을 위해 통계적인 방법을 사용하고, 의미를 해석하기 위해 단어의 개념에 대한 지식을 사용하여 의미 해석을 위한 문장의 수를 줄임으로써 자연언어 처리의 단점을 해결할 수 있고, 일정 수준의 요약문을 생성할 수 있다.

본 연구에서는 사고 관련 기사의 다국어 요약문을 생성하는 혼합 형태의 요약 시스템을 구현하기 위해 중간 개념 표현 구조를 제안한다. 품사 태거, 지역 의존 관계를 이용한 부분 파서, 사고, 사건 관련 기사에서 추출한 명사와 동사의 시소러스 및 동사의 하위범주화 사전 등을 이용하며, 의미 해석을 위한 문장의 수를 줄이기 위해 중요 문장을 추출하고, 단락을 재표현하는 구조(PRS)와 이들을 다시 통합한 중간 개념 표현 구조(URS)를 생성한다.

3. 중간 개념 표현

본 연구에서 제안한 중간 개념 표현 구조를 생성하기 위해 문서로부터 주제를 파악, 추출하고 각 단락으로부터 단락의 내용을 표현하는 단락 재표현 구조들을 생성 후, 이들을 다시 통합한다. 각 과정을 살펴보면 다음과 같다.

3.1 주제 파악을 위한 처리

하나의 문서는 여러 토픽을 가질 수 있으므로, 중요하고 핵심적인 토픽들로만 유지 되도록 원문의 문장들을 걸러야 한다. 이를 위해 단락을 나누어 중요 문장을 추출하는 과정이 필요하다. 이것은 의미 해석을 위한 문장의 수를 줄이는 효과를 가질 수 있다. 본 연구에서는 단락의 자동 구분을 통한 중요 문장 추출 시스템[7]을 이용하여 각 단락의 중요 문장을 추출한다. 그리고 이 추출된 문장들을 단락의 주된

* 본 연구는 정보통신연구진흥원의 대학기초연구지원사업 과제 "Web 상에서 적확한 검색을 위한 문서의 대표 개념어 생성 및 요약 시스템"의 일부로 수행되었음.

토픽으로 생각한다. 위 추출 시스템을 이용할 경우 추출된 문장의 부분 파싱된 구문적 정보로서 의존 그래프을 얻을 수 있으며, 의존 그래프의 각 노드는 격조사를 중심으로 지역적으로 단위화된 형태(chunk)를 갖추고 있다.

3.2. 의미 해석과 단락 재표현 구조

추출된 문장이 정보가 누락되어 있거나, 어휘적 결속이나 혼합적 결속 구조로 표현이 된 경우 정보를 보완하고, 결속 부분의 의미를 분석하여 단락의 내용을 재표현 한다. 단락 재표현 구조(PRS)는 Sowa의 개념 그래프[5]를 이용한 것으로 개념 노드와 관계 노드로 구성된다. 이들은 의존 그래프의 노드들이 변환, 생성된 것이거나, 동일 객체 파악을 통해 필요한 정보들이 노드로 생성되어 추가된 것이다.

PRS 노드 형성의 기본 규칙은 전치사, 접속사와 같은 기능어 들은 관계노드로, 명사, 동사, 형용사, 부사와 같은 내용어 들은 개념노드로 생성하는 것이다. 추출 문장의 의존 그래프를 입력으로 받아 중심용언 노드 아래에 연결된 모든 노드를 깊이우선탐색으로 탐색하면서 구문적 정보와 동사의 하위 범주화 사전에 나타나 있는 술어의 개념에 따른 문형 정보[6]를 이용하여 PRS의 노드를 생성한다. 동사의 하위 범주화 사전에는 술어에 따른 문형 정보가 나타나 있으며 여기에 의미 해석 규칙과 개념 패턴 규칙[8]을 추가하여 개념 노드 및 관계노드 생성에 이용한다.[그림1]

의존 그래프의 노드가 구체적일 필요 없는 객체들의 나열로 표현된 경우는 노드 생성 시 추상화하여 하나의 개념 노드로 만든다. 간단한 추상화 규칙의 예로 '여객선 선장 편징관 씨 등 승무원 8명과 승객 57명'과 같은 의존 그래프의 한 노드는 '탑승자 65명'으로 추상화되어 개념 노드가 된다. 그림 2와 같은 추상화 규칙은 동일 개념 파악 시 의미 해석을 줄여줄 뿐만 아니라 노드 생성시 드는 비용도 줄여준다. 그리고, "ㄴ(은/는) 것으로 추정된다.", "~라고 밝혔다./전했다."와 같은 보조 용언들은 PRS 노드 생성 시 필요 없는 내용으로 간주하여 노드생성에서 제외한다.

```
<Entry>
<Title> 타다 </Title>
<CaseFrame>
<frame> N0가이 N1을 타다.</frame>
<SIR> N0 <-AGNT,  N1 <- OBJ </SIR>
<synsem>
<SelRst1> N0=사람 N1={꽃물,커리,홍차...}</SelRst> //선택제약
<SIR1>  N1 <- OBJ </SIR1> //의미 해석 규칙
<ConceptP> 탑승하다</ConceptP> //개념 패턴 규칙
<SelRst2> N0=사람 N1={말, 버스..}</SelRst>
<SIR2>  N1 <- INS </SIR2>
<ConceptP> 승차하다</ConceptP>
<SelRst3> N0=사람 N1={거문고, 가야금...}</SelRst>
<SIR3>  N1 <- OBJ </SIR3>
<ConceptP> 연주하다 </ConceptP>
...
</synsem>
</CaseFrame>
</Entry>
```

[그림 1]하위범주화 사전에서 의미 해석 규칙과 개념 패턴 규칙

정보 추가의 기본 규칙은 생성된 개념노드들과 같은 개념이 있는지를 단락으로부터 찾아서 추가될 내용이 있을 경우 이를 추가하는 것이다. 개념 노드 중 중심 용언 노드를 확인 후 단락으로부터 같은 개념의 용언이 있는지를 찾는다. 이때 동사의 유의어 사전이 있다면 더욱 좋은 효과를 낸다. 본 연구에서는 코퍼스에 맞는 간단한 동사 유의어 사전 및 명사의 계층 사전을 만들어 이용하였다. 찾아진 경우 의존 관계를 지니고 있는 요소들을 살펴 필요한 정보를 노드로 생성하여 추가한다. 이 때, 의존 그래프의 요소들과 개념노드들 사이의 동일 의미 파악을 위해 그림 3의 휴리스틱한 규칙들을 바탕으로 용언에 딸려있는 필수

- humanN+{(,)|(와/과)+humanM}+(수량)-> humanK + (수량)
- humanN+수량$_{i-1}$+ {(,)|(와/과)+humanM+(수량i) }$_{i>=2}$
  -> nnp(humanK) + SUM(수량i)
- (규칙1의 앞 패턴) + 등 + [총] + (수량m)->humanK+(수량)
- {(고유명사)+,|와/과|} + 등 + (수량) ---> (수량)

...

[그림 2] 노드 생성 시 추상화 규칙

성분 개체들이 같은 개념인지를 확인한다. 이 규칙은 의존 그래프와 개념 그래프에서 개체의 표현 양상만을 보고 동일 개체 파악을 하고 있으므로 의미해석의 비용을 줄여준다. 더 구체화된 개체표현이 있을 시 개념 노드로 생성하여 관계노드(LINK)로 연결하거나, 딸려있는 수의적 성분을 개념노드와 관계노드로 생성하여 중심 용언에 단순 연결한다. 만약 관계노드가 명확히 보여지지 않는 경우는 LINK로 연결한다. 이 때 추가된 내용은 '( )'로 나타내어 정보 추가의 양상을 보이도록 한다.

그리고 중심용언이 단락 내 의존 그래프의 요소로 존재하지는 않지만 필수성분의 개념 노드들이 단락 내 문장에서 'A은 B이다.'와 같은 서술격 조사를 이용한 문 형태로 표현될 때, B를 개념 노드로 생성하여 관계노드(CHRC)로 연결한다. 즉, 기본적인 분류문과 A와 B가 서로 포함 관계에 있지 않지만 어느 것이 먼저인지 모르는 관계에 있는 기본 서술 구조에서 A와 B는 필요한 정보가 될 수 있다.

- 규칙 1: 한 개 이상의 명사로 이뤄진 개체는 그 개체를 이루는 각 명사들을 순서대로 조합한 원소 중 하나로 다시 표현된다.
- 규칙 2: 한 개체를 표현하고 있는 어휘에서 (항공기, 공항, 상가, 버스, 유조선 ...등) 실마리 단어가 존재하면 개체는 그 일반명사의 개념을 가지게 되므로, "#+(실마리 단어|유사어)" ('#' 기호는 '이', '그', '사고', ' ')로 다시 표현된다.
- 규칙 3: '이날', '이 시간'과 같은 시간의 대한 참조 표현이 있을 경우 앞선 문장에 참조대상이 나타난다. 참조 대상이 없는 경우는 그 기사가 실린 날짜를 대부분 따르고 있다.

...

[그림 3] 동일 개체 파악을 위한 휴리스틱 규칙

3.3. 요약문을 대표하는 통합된 재표현 구조

단락의 내용을 대표하는 PRS들을 응집성 있도록 재형성 하는 과정을 거쳐 요약을 대표하는 개념구조로 만든다. 요약된 내용을 표현하는 통합된 재표현 구조(URS) 생성은 PRS생성 시와 같이 개념 노드들이 같은 의미를 지닌 개체(피해자, 장소, 시간, 사물...)인지를 파악할 수 있어야 한다. PRS 생성 과정에서 적용한 휴리스틱한 규칙들을 이용하여 동일 개념 파악 후에 내용을 추가시키거나 병합한다. 병합은 Sowa의 결합(Join)규칙[5]을 이용하였다. 즉, 한 PRS내에 있는 개념 노드 c와 다른 PRS에 있는 개념 노드 d가 같을 때 d를 제거하고 d에 붙어있던 모든 노드를 c에 붙여서 통합한다. 그리고 중심 용언에 연결된 객체 노드가 동일 한 경우 뒤에 오는 PRS의 객체 노드를 '# 객체'화 하여 표현한다. 각 단락을 대표하는 단락 재표현 구조들은 우선 인접한 두 개씩 병합되어 차례대로 올라가며 URS를 형성해나간다.

4. 실험 및 고찰

단락의 자동 구분을 통한 중요 문장 추출 시스템[7]을 이용하여 중요 문장을 입력문[그림 4]에 대해서 40%로 추출하였다. 첫 번째 단락에서

는 1이 추출, 문장 2,3,4로 이뤄진 단락에서는 3 추출, 문장 5,6,7,8로 이뤄진 단락에서는 5 추출, 마지막 단락에서는 9가 추출되었다.

1. 승객과 승무원 1백 79명을 태운 싱가포르 에어라인 소속 SQ006 항공기가 31일 밤 타이베이국제공항에서 이륙 직후 폭발했다. 2. 이 사고로 탑승자 상당수가 사망한 것으로 추정된다고 공항측은 밝혔다. 3. 목격자들은 사고 항공기가 이날 오후 11시 20분쯤 치앙카이섹 공항에서 이륙한 직후 거센 바람으로 중심을 잃으면서 곧바로 추락, 폭발했다고 전했다. 4. 사고비행기는 로스앤젤레스 행이었다. 5. 이 비행기는 추락하면서 공항에 대기 중이던 중국항공 비행기와 충돌했다. 6. 그러나 중국항공 소속 비행기에는 탑승객이 없어 인명피해는 없었다. 7. 사고 직후 공항구조대가 출동, 곧바로 비행기 화재진압에 나섰으며 불길이 잡히자마자 생존자 구출작업을 벌여 자정 무렵까지 부상자 18명을 병원으로 후송했다. 8. 이날 타이베이에는 태풍이 상륙 비바람이 심하게 몰아쳤으나 항공사측이 무리하게 이륙을 고집한 것으로 알려졌다. 9. 사고비행기는 보잉747기종이다.

[그림 4] 추락사고 관련 기사 원문

추출된 각 문장과 각 단락으로부터 PRS 생성 과정을 거친 결과는 아래와 같다.[그림 5]

```
[폭발하다] <-(past)
  ->(loc)->[타이베이국제공항:공항]
  ->(ptim)->[31일 밤]
  ->(agent)->[싱가포르 에어라인 소속 SQ006항공기]
              <-(obj)<-[타다]
                       ->(agent)->[탑승객 1백 79명]
  <-(succ)-[이륙하다]
-------------------------------------------------------
[추락,폭발하다] <-(past)
  <-(succ)<-[이륙하다]
            ->(agent)->[# 항공기]
                      ->(link)->[(로스앤젤레스 행)]
            ->(ptim)->[이날 오후 11시 20분쯤]
            ->(loc)->[치앙카이섹 공항]
  ->(manr)->[곧바로]
  ->(caus)->[잃다]->(obj)->[중심]
             ->(caus)->[바람]->(attr)->[거세다]
-------------------------------------------------------
[충돌하다]<-(past)
 ->(agent)->[# 비행기]
            ->(accm)->[중국항공 (소속) 비행기]
                     <-(agent)<-[대기하다]<-(ing)
                                ->(loc)->[공항]
  ->(link)->[추락하다]
• 단락 4의 PRS
-------------------------------------------------------
[# 비행기]->(chrc)->[보잉747기종]
```

[그림 5] 각 단락으로부터 생성된 PRS

단락 대표 개념구조를 생성한 후, 인접한 두 개씩 병합되어 차례대로 올라가며 URS를 생성한다. 그림 6에서 ①은 PRS1, PRS2의 통합된 결과이며, 이것은 ②의 동일 개념노드 '추락하다'에 의해 결합규칙으로 통합되어진다. 이와 같이 요약을 대표하는 구조(URS)를 생성하면 일정한 수준의 요약문을 생성하기가 쉬워 질 것이며, 또한 다국어 요약 생성을 위한 중간적 개념 표현 구조로서 역할을 수행하게 될 것이다.

5. 결론 및 향후 연구

본 논문에서는 사건, 사고 관련 기사의 자동 요약문 생성을 위한 중간 언어적 재표현 방법을 제안한다. 각 단락으로부터 중요 문장을 선정하여 의미 해석을 위한 문장의 수를 줄이고 필요한 정보를 추가하거나 간략화 하여 단락을 재표현 하는 구조(PRS)를 만들고 이들을 다시 통합하여 요약 내용을 재표현한 구조(URS)로 만든다. 이것은 향후 다언어 자동 요약문 생성을 위한 중간 언어적 재표현 구조로 응용될 수 있을 것이다. 앞으로 이 구조를 이용하여 문서의 다국어 요약문을 생성하는 혼합 형태의 요약 시스템을 구현하고자 한다. 이를 위해서 URS로부터 문장을 생성하는 연구가 추가 진행되어야 한다.

```
①
[추락,폭발하다]<-(past)
  ->(agent)->[싱가포르 에어라인 소속 SQ006항공기]
               ->(link)->[(로스앤젤레스 행)]
              <-(obj)<-[타다]->(agent)->[탑승객 1백 79명]
    ->(ptim)->[31일 (오후 11시 20분쯤)]
    ->(loc)->[타이베이국제공항:공항]
             ->(link)->[(치앙카이섹 공항)]
    <-(succ)-[이륙하다]
    ->(caus)->[잃다]->(obj)->[중심]
              ->(caus)->[바람]->(attr)->[거세다]
    ->(manr)->[곧바로]
-------------------------------------------------------
②
[충돌하다]<-(past)
   ->(agent)->[# 비행기]
              ->(chrc)->[(보잉747기종)]
              ->(accm)->[중국항공 (소속) 비행기]
                       <-(agent)<-[대기하다]<-(ing)
                                  ->(loc)->[# 공항]
   ->(link)->[추락하다]-> (link) -> ①
```

[그림 6] 통합 과정을 거친 요약 대표 개념구조

참고 문헌

[1] J.Kupiec, J.Pedersen, and F.Chen. "A Trainable Document Summarizer" In Proceedings of the 18th ACM-SIGR Conference, pp. 68-73, 1995.

[2] R.Barzilary and M.Elhadad. "Using Lexical Chains for Text Summarization" In Inderjeet Mani and Mark T. Maybury, editors, Advances in automatic text summarization. MIT Press, Cambridge, MA, pp. 111-121.

[3] Eduard Hovy and Chin-Yew Lin. Automated Text Summarization in SUMMARIST. IN ACL/EACL97 Workshop on Intelligent Scalable Text Summarization, pp18-24, 1997.

[4] Chris D. Paice and Paul A. Jones. "The Identification of Important Concepts in Highly Structured Technical Papers". In Proceedings of the 16th Annual International ACM-SIGIR Conference, pp 69-78, 1993.

[5] John F. Sowa, Conceptual Structure : Information Processing in Mind, Machine, Addison Weseley Pub. Com, 1984.

[6] 강은국, 조선어 문형에 관한 연구, 박이정 출판사, 1995

[7] 김계성 외, "단락 자동 구분을 통한 중요 문장 추출", 제12회 한글 및 한국어 정보처리 학술발표 논문집. pp233~238, 2000

[8] 박인철 외, "A Study on Generation of Conceptual Graphs using Sentence Patterns in Korean", In Proceedings of Natural Language Processing Pacific Rim Symposium, pp. 685-690, 1995.