정보 추출을 위한 이벤트 문장 추출

김태현° 임수종 윤보현 박상규

지식처리연구팀, 휴먼정보처리연구부, 한국전자통신연구원
{heemang, isj, ybh, parksk}@etri.re.kr

Event Sentence Extraction for Information Extraction

Tae-Hyun Kim°, Soojong Lim, and Bo-Hyun Yun

Knowledge Processing Research Team
Human Information Processing Dept.
Electronics and Telecommunications Research Institute

요 약

정보추출 시스템의 목적은 관심의 대상이 되는 특정 정보를 선택적으로 찾아내 제시하는데 있다. 따라서 도메인 정보에 의존적인 방법으로 정보추출이 이루어질 수 밖에 없고, 이에 따른 도메인 정보 구축의 부담이 컸다. 이러한 부담을 줄이기 위해 본 연구에서는 특정 주제영역과 관련한 문서로부터 자동으로 이벤트 문장을 추출하는 시스템을 제안한다. 이벤트 문장이란, 특정 도메인에서 다루어지는 이벤트의 구체적인 내용을 포함하고 있는 문장이다. 이러한 문장을 추출함으로써 기본적인 수준의 정보추출 요구를 만족시킬 수 있을 뿐만 아니라, 추출된 이벤트 문장을 도메인 정보 구축에 활용할 수 있을 것이다. 본 연구에서는 동사, 명사, 명사구, 및 3W 자질을 이용하여 문장추출의 성능을 최대화하기 위한 방안을 제안하고, 세 개의 평가 도메인을 대상으로 실험을 수행하였다. 실험 결과, when 및 where 자질과 동사, 명사, 명사구의 가중치를 이용하여 문장 가중치를 계산함으로써 최적의 이벤트 문장추출 성능을 얻을 수 있음을 알 수 있었다.

1. 서론

사용자의 정보에 대한 욕구는 단순한 정보검색의 결과를 넘어서 보다 정제되고 가공된 형태의 결과를 지향하고 있다. 이에 따라 단순히 사용자가 원하는 정보를 포함하고 있는 문서들을 찾아주는 정보검색(Information Retrieval)에서 벗어나 사용자가 원하는 정보 자체를 정확히 찾아주는 정보추출(Information Extraction)에 대한 요구가 부각되고 있다.

정보추출이란, 자연어를 분석하여 명시된 형식의 개체(entity)나 관계(relationship) 또는 이벤트(event)들에 대한 정보를 수집하는 프로세스이다[1]. 즉, 자연어 텍스트에서 개체들과 이들 간의 관계를 찾아내고, 이들을 이용해 텍스트 내에서 핵심 이벤트들에 대한 정보를 추출해내는 것이 정보추출의 궁극적인 목적이다. 정보추출의 대상이 되는 이벤트는 다음과 같이 정의된다.

> Something (non-trivial) happening in a certain place and a certain time[7].

이벤트는 하나의 정형적인 행위나 현상으로 규정지을 수 없는, 사용자의 관심을 끌만한 어떠한 일을 의미한다. 따라서, 모든 경우에 적용될 수 있는 일반화된 개념의 이벤트란 존재할 수 없다. 이러한 이유로 정보추출에서는 특정 도메인(주제 영역)과 관련이 있는 이벤트들을 추출하는 것을 목적으로 하고 있다. 뿐만 아니라, 이벤트가 발생한 장소 및 시간과 같은 일회적인 정보가 이벤트를 뒷받침하는 중요한 정보로써 함께 다루어지고 있다.

기존 연구들에서는 정보추출을 위해 대상도메인을 한정한 상태에서 패턴형태의 도메인 의존적인 정보를 구축하고 이를 이용해 텍스트의 특정 부분을 추출하는 방식을 주로 사용하고 있다[1, 2, 3]. 이는 다시 크게 두

가지 방식으로 나누어 볼 수 있다. 첫째는 텍스트에서 개체명들을 인식[1]하고, 템플릿 엘리먼트[2], 템플릿 릴레이션[3], 시나리오 템플릿[4]을 점차로 구성하면서 추출하고자 하는 정보를 획득하는 방식이고[1], 둘째는 우선 텍스트에서 중요 부분을 추출한 후에 이를 대상으로 수동으로 만들어진 패턴과의 비교를 수행해 원하는 정보를 찾아내는 방식이다[2, 3]. 첫 번째 방식의 경우 각 단계에서 이용하는 도메인 정보 구축을 위해 우선 해당 도메인에서 중요시 되는 정보들을 찾아내야 한다는 문제가 있다. 두 번째 방식의 경우는 중요부분 추출 문제를 단순히 어휘정보에만 의존해 해결하고 있어, 실질적인 정보추출 대상들을 효과적으로 추출해내지 못하고 있다는 단점이 있다.

본 연구에서는 위와 같은 문제들을 해결하기 위해 텍스트로부터 정보추출에 유용한 부분만을 추출해내는 이벤트 문장 추출 시스템을 제안한다. 이 시스템은 주어진 특정 도메인과 관련된 문서집합을 대상으로 자동 학습을 수행하고, 학습된 정보를 이용해 도메인 특정적인 사건의 내용을 포함하고 있는 이벤트 문장을 추출한다. 추출된 문장은 이벤트의 주체, 객체, 발생 일시, 및 장소 등에 관한 정보를 포함하고 있다. 따라서, 추출된 문장만으로도 기본적인 정보추출 요구를 만족시켜줄 수 있을 뿐만 아니라, 이를 도메인 정보 자동구축을 위한 자료로 활용할 수 있을 것이다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 자동요약 분야에서 사용되는 문장추출 기법들에 대해 살펴본다. 3장에서는 이벤트의 특성을 고려해 선택한 다양한 자질들에 대해 소개하고, 4장에서는 선택한 자질들을 이용해 고안한 이벤트 문장 추출 시스템에 대해 설명한다. 5장에서는 이벤트 문장추출 결과를 평가하기 위한 실험 방안 및 실험 결과를 보이고, 6장에서 결론을 맺기로 한다.

2. 관련 연구

이벤트 문장추출은 기존의 자동요약 분야에서 사용되고 있는 요약을 위한 문장추출과 외형적으로 유사하지만 그 목적은 상이하다. 자동요약은 대상이 되는 문서의 핵심 내용을 찾아내는데 그 목적이 있는 반면, 추출은 특정 주제 영역과 관련이 있는 문서에서 사용자의 관심의 대상이 되는 이벤트성 정보들을 선택적으로 찾아내는데 목적이 있다. 본 연구에서는 자동요약에서 사용되는 문장추출 기법에서 기본 아이디어를 얻어와 이벤트 문장추출이라는 새로운 접근을 시도하였으므로 이와 관련된 기존 연구들에 대해 살펴보기로 한다.

2.1. 클러스터링을 이용한 문장추출

문장 단위의 클러스터링을 수행하고 그 결과로 만들어진 클러스터를 대상으로 문장을 추출함으로써 요약을 수행하는 방법이다. 클러스터링 결과로 만들어진 각 클러스터는 클러스터별 중요도와 클러스터 내의 문장별 중요도를 가지며, 이 중요도를 이용해 요약문장을 추출하게 된다[4].

클러스터링의 단위가 문장이기 때문에 발생하는 정보부족 현상을 줄이기 위해 문서에 있는 수사구조관계를 부가적으로 사용하여 서로 관계가 있는 문장을 결합해 하나의 입력단위가 되도록 한다. 그리고, 클러스터링된 결과에서 요약문장을 추출하기 위한 방법으로, 가장 중요도가 높은 클러스터 하나를 선택해 제시하거나, 각 클러스터내에서 문장 중요도가 가장 높은 문장들을 하나씩 선택하여 제시한다.

이러한 방법은 수사구조관계에 대한 정보를 제외하고는 어떠한 기본 정보도 사용하지 않고 있으므로, 도메인에 의존적이지 않은 방법으로 요약을 생성할 수 있다는 장점이 있지만, 클러스터링의 성능에 따라 요약의 성능도 좌우된다는 단점이 있다.

2.2. 확률에 기반한 문장추출

문서집합에 있는 여러 가지 자질에 대한 확률정보를 수집해, 이를 문장추출에 사용하는 방법이다. 이를 위해 우선적으로 고려되는 사항은 어떤 자질을 사용할 것인가 하는 점이다. 요약 문장추출을 위해 일반적으로 사용되는 자질로는 문장위치 자질, 문장길이 자질, 키워드 자질, 단서단어 자질, 제목단어 자질 등이 있다[5].

이 방법은 학습 문서 집합을 대상으로 위와 같은 자질들에 대한 통계정보를 수집하는 단계와 수집된 통계정보를 적용하여 실제로 요약문장을 추출하는 단계로 나뉘어 진다. 추출된 자질은 베이시안 분류(Bayesian classification) 기법에서 사용되는 벡터의 자질로 쓰여 분류기법을 이용해 문장을 추출하게 된다.

이 방법은 도메인에 독립적인 방법으로 적용할 수 있고 다양한 자질을 손쉽게 결합하여 사용할 수 있다는 장점이 있지만, 적용분야가 달라지는 경우에는 수집된 통계정보의 의미가 없어지게 되므로 다시 학습을 수행해야 한다는 단점이 있다.

2.3. 언어이해에 기반한 문장추출

텍스트의 전체적인 구조를 분석하여 이용하는 담화론적 수준의 자동요약 방법으로, 주로 문서의 내용을 파싱하여 생성한 수사구조 트리(Rhetorical structure tree)를 이용해 문장의 논리적인 관계정보를 구조화한다. 전체 문서에 대해 트리가 생성되면 트리의 최상위 노드가 문서의 대표적인 문장이 되고, 이 노드에 가까울수록 문서 내용표현에 있어 중요한 역할을 하는 문장이 된다[6].

이 방법은 다른 방법보다 언어학적으로 보다 세련

---

[1] Named entity recognition: 인명, 조직명, 장소 등의 개체에 대한 이름을 인식하고 종류에 따라 분류하는 일. 날짜, 시간, 통화 등도 포함됨.
[2] Template element: 문서 내에서 동일 개체를 언급하기 위해 사용된 모든 형식의 이름 및 설명 등을 결합한 것.
[3] Template relation: 템플릿 엘리먼트(개체)들 간의 관계. 이러한 관계의 예로 "A is located in B"에서 *location-of*와 같은 것을 들 수 있음.
[4] Scenario Template: 템플릿 엘리먼트와 템플릿 관계들을 이용해 도메인과 정보추출 목적에 맞게 정해진 시나리오 템플릿을 채우는 일.

된 접근방법을 사용하고 있어 요약문장 추출에 있어 좋은 결과를 얻을 수 있으나, 대상 문서의 크기가 큰 경우 수사구조트리를 구성하기 어렵고, 이를 위해 필요한 분석 시간이 오래 걸린다는 단점이 있다.

자동요약을 위한 문장추출은 입력문서의 형식이나 문서 내 어휘빈도에 의존적인 방식으로 중요문장을 추출한다. 따라서, 문서의 형식이 다른 경우 문장추출의 조건을 수정해 주어야 하고 빈도가 낮은 중요정보를 포함하고 있는 문장을 추출하기 어렵다는 단점을 갖는다.

3. 자질 선택

본 논문에서는 정보추출의 효율을 높이기 위해 텍스트로부터 정보추출에 유용한 부분만을 추출해내는 시스템인 이벤트 문장추출 시스템을 제안한다. 제안된 시스템을 이용해 추출된 이벤트 문장은 도메인 정보 구축 및 정보추출의 기본 데이터로 활용될 수 있다. 본 연구에서 추출의 대상이 되는 이벤트 문장이란, 특정 도메인에서 다루어지는 사건의 구체적인 내용(주체, 객체, 일시, 장소 등)을 표현하고 있는 문장이다. 다음은 '비행기 사고'와 관련한 도메인에서의 이벤트 문장 예이다.

> 225명의 승객과 승무원을 태우고 대만을 떠나 홍콩으로 가던 대만의 중화항공 여객기가 25일 오후 대만해협에 추락했다.

자동요약을 위한 문장추출의 경우는 임의의 문서 또는 문서집합을 입력으로 받아 그것이 내포하고 있는 주제 및 핵심흐름을 파악하는데 목적이 있는 반면, 이벤트 문장추출은 특정 주제와 관련이 있는 문서 또는 문서집합을 입력으로 받아 해당 주제와 관련이 있는 이벤트의 주체, 객체 및 정황(시간, 장소 등)을 찾아내는데 목적이 있다. 따라서, 이러한 목적에 부합되는 자질을 선택하는 것이 중요하다.

본 연구에서는 형태소 분석 및 개체명 인식의 결과를 이용하여 도메인 의존적이고 일회적인 이벤트의 특성을 최대한 커버할 수 있는 자질들을 선택하였다. 선택된 자질은 동사 자질, 3W(Who, When, Where) 자질, 명사 및 명사구 자질의 세 종류로 나누어 볼 수 있다. 동사 자질은 도메인의 주제를 이끌어가는 핵심 행위 및 상황을 대표하는 자질이고, 3W 자질은 이벤트의 주체, 객체 및 정황을 찾아내기 위해 사용되는 자질이며, 명사 및 명사구 자질은 도메인 의존적인 정보를 반영하기 위해 사용되는 자질이다. 각 자질에 대해 자세히 살펴보기로 한다.

3.1. 동사 자질

형태소 분석 결과에서 'PV' 및 'NC+XSV'의 형태로 태깅된 어휘들을 추출하여 동사 자질로 삼는다. 다음은 동사 자질로 추출되는 형태의 예이다.

<PV:떠나>+<EF:아> → 떠나다
<NC:추락>+<XSV:하>+<EP:었>+<EC:다> → 추락하다

동사 중에서 '하다', '되다', '다하다' 등과 같이 보조적으로 사용되어 특별한 의미를 갖지 못하는 동사는 자질로 삼지 않도록 하였다.

3.2. 명사 및 명사구 자질

형태소 분석 및 개체명 인식된 결과에서 명사형으로 사용되는 단어를 추출해 명사 자질로 삼는다. 이때 특성상 변형이 많은 어휘의 경우는 품사정보를 자질로 삼고, 그렇지 않은 경우는 어휘 자체를 자질로 삼는다. 즉, 'NC', 'PERSON', 'LOCATION', 'ORGANIZATION' 등의 품사를 갖는 단어들은 어휘 자체를 자질로 사용하고[5], 'NN', 'PERCENT', 'DATE', 'TIME', 'MONEY', 'QUANTITY' 등의 품사를 갖는 단어들은 품사정보를 자질로 사용한다. 이는 이벤트에서 중요한 정보인 이벤트 발생 일이나 시간, 수량 등에 대한 정보가 단순히 어휘 빈도가 낮다는 이유로 학습 데이터에서 누락되지 않도록 하기 위함이다. 다음은 각 경우에 대한 예이다.

<NC:승객> → 승객
<LOCATION:대만> → 대만
<DATE:25일>, <DATE:84년12월> → DATE

명사구 자질은 동일 문장 내에 인접해 나타나는 명사 자질들을 결합한 것을 사용한다. 본 연구에서는 두 개의 명사 자질을 결합하는 간단한 방법을 이용하였다. 다음은 명사구 자질로 선택된 예이다.

'국제 공항', '비행기 추락', 'QUANTITY 승무원'

3.3. 3W 자질

영어권에서 사용되는 Who, When, Where에 해당되는 개념을 각각 이벤트의 주체 및 객체, 일시, 장소에 해당되는 정보를 식별하기 위해 사용한다. 개체명 인식 결과로 얻어진 태그정보를 이용하여 다음과 같이 그 정보를 매칭시킨다.

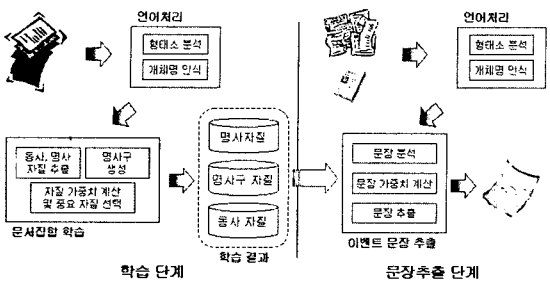
'PERSON', 'ORGANIZATION' → WHO
'DATE', 'TIME' → WHEN
'LOCATION' → WHERE

4. 이벤트 문장 추출 시스템

본 연구에서 제안한 이벤트 문장추출 시스템은 크게 학습 단계와 추출 단계로 나누어 볼 수 있다. 다음의 [그림 1]은 학습 및 추출 단계에 대한 간략한 시스템 구성도이다.

---

[5] 'NC+XSN'의 형태로 태깅되는 어휘의 경우 이를 결합한 어휘를 자질로 삼는다. ex) <NC:비행>+<XSN:기>→ 비행기



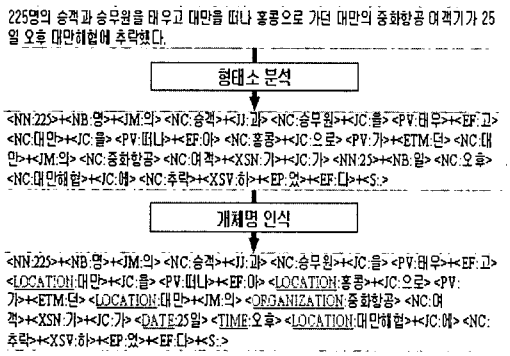
[그림 1] 시스템 구성도

학습 단계에서는 특정 주제와 관련이 있는 문서집합을 대상으로 언어처리를 수행한 결과로부터 동사, 명사 및 명사구 자질에 대한 도메인 정보를 자동 학습하고, 추출 단계에서는 입력으로 주어진 문서 또는 문서집합을 언어처리한 결과와 학습된 정보를 이용하여 실제로 이벤트 문장을 추출하게 된다.

본 장에서는 시스템의 구성요소를 언어처리, 문서집합 학습, 이벤트 문장 추출의 세 부분으로 나누어 설명하고자 한다.

4.1. 언어처리

이 단계에서는 주어진 입력 문서에 대해 형태소 분석과 개체명 인식 과정을 차례로 수행하여 최종적으로 다음의 [그림 2]와 같은 형태의 결과를 만들어낸다.



[그림 2] 언어처리 예

4.2. 문서집합 학습

특정 주제와 관련이 있는 문서집합을 언어처리한 결과로부터 동사, 명사 및 명사구 자질을 추출하고 중요 자질들에 대한 통계정보를 수집하는 단계이다. 동사 및 명사 자질은 3.1.과 3.2.에서 설명한 것과 같은 기준에 따라 추출한다. 그리고 각 자질에 대한 출현 빈도(tf)와 문헌 빈도(df), 자질이 출현한 위치정보를 수집한다. 자질이 출현한 위치정보는 자질이 출현한 문장 번호와 동일 문장 내에서 현재 자질의 바로 앞에 위치하는 자질의 ID로 구성된다. 이 위치정보를 이용하여 동일 문장에 인접해 나타나는 명사들을 결합함으로써 명사구 자질을 생성한다.

동사, 명사 및 명사구 자질은 각기 다음의 [표 1]과 같은 식에 의해 가중치가 부여되고, 그 중에서 상위에 위치한 중요 자질들이 학습 결과로서 최종적으로 저장된다.

[표 1] 자질 가중치 부여

동사 및 명사	명사구
$w_i = \frac{tf_i \times \left(\log \frac{D}{df_i} + 1\right)}{w_{\max}}$	$w_{ij} = \frac{w_i + w_j}{2}$

4.3. 이벤트 문장 추출

특정 주제와 관련이 있는 문서 또는 문서집합을 언어처리한 결과를 입력으로 받아 문장분석, 문장 가중치 계산의 과정을 거쳐 최종적으로 이벤트 문장을 추출해 내는 단계이다.

문장분석 단계에서는 문장에 포함되어 있는 각 자질요소에 대한 정보를 수집하고, 학습된 도메인 정보에서 각 자질에 대한 가중치 및 출현문장목록을 얻어오게 된다. 이 과정에서 태그정보를 참조해 문장 별로 Who, When, Where 정보를 얼마나 포함하고 있는지(1)와 문장 내에 포함되어 있는 각 자질의 가중치가 얼마나 되는지(2)에 대한 정보를 파악하게 된다. 아래의 수식에서 i는 문장번호를 나타낸다.

$$C_{i,who},\ C_{i,when},\ C_{i,where} \quad (1)$$

$$W_{i,verb},\ W_{i,noun},\ W_{i,np} \quad (2)$$

문장 가중치 계산에서는 각 문장 별로 다음과 같은 수식을 이용해 문장 가중치를 계산한다. 문장에 포함되어 있는 명사 자질과 명사구 자질을 문장 가중치 계산에 반영하기 위해, i번째 문장에서 동사-j와 공기하는 명사 자질들에 대한 가중치 합의 평균을 나타내는 $Co\_vn_{i,j}$ 값과 i번째 문장에서 동사-j와 공기하는 명사구 자질들에 대한 가중치 합의 평균을 나타내는 $Co\_vp_{i,j}$ 값을 아래와 같이 각각 구하고, 여기에 동사 자질의 가중치를 수식 (3)과 같은 방법으로 반영함으로써 각 문장의 가중치를 계산한다.

$$Co\_vn_{i,j} = \frac{\sum_{k=1}^{C_{i,noun}} \left(W_{n_k} \times Co_{v_j,n_k}\right)}{C_{i,noun}}$$

$$Co\_vp_{i,j} = \frac{\sum_{l=1}^{C_{i,np}} \left(W_{np_l} \times Co_{v_j,np_l}\right)}{C_{i,np}}$$

$$W_i = \sum_{j=1}^{C_{i,verb}} \left(W_{v_j} \times \left(\alpha \cdot Co\_vn_{i,j} + \beta \cdot Co\_vp_{i,j} \right) \right) \Big/ C_{i,verb} \quad (3)$$

위의 수식들에서 $C_{i,noun}$, $C_{i,np}$, $C_{i,verb}$는 각각 문장-i 내에 출현한 명사, 명사구, 동사 자질의 수를 나타내고, W_{n_k}, W_{np_l}, W_{v_j}는 학습의 결과로 얻은 각 자질의 가중치를 나타낸다. 또한 Co_{v_j,n_k}, Co_{v_j,np_l}는 각각 동사-j와 명사-k의 공기빈도, 동사-j와 명사구-l의 공기빈도를 나타내고, α와 β는 명사와 명사구 자질이 문장추출에 기여하는 정도에 따라 조정되는 상수 값이다. 각 문장의 가중치가 계산되면 이를 이용해 단일 문서 내에서 문장들을 정렬해 문장추출 시에 조건항목으로 사용한다.

문장추출 단계에서는 문장 단위로 얻어진 (1), (2), (3)의 정보를 조합하여 문장을 추출한다. 실험을 통해 얻은 이벤트 문장 추출에 가장 적합한 문장추출 조건은 다음 [그림 3]의 알고리즘과 같다.

```
FOR all Sententces IN this Document
    IF (C_{i,when} ∧ C_{i,where} ∧ W_i) THEN SELECT_SENT;
FOR all Remained_Sentences IN this Document
    SELECTmax_weight SENT;
    IF ((W_i > θ_1) ∨ ((selected < θ_2) ∧ W_i)) THEN
        SELECT_SENT;
ENDFOR
```

[그림 3] 이벤트 문장추출 알고리즘

조건식에서 θ_1은 문장 가중치의 임계값, θ_2는 문장 선택 개수의 임계값을 나타내고, *selected*는 문서 내에서 선택된 이벤트 문장의 개수를 나타낸다.

5. 실험

5.1 실험 자료

이벤트 문장추출 시스템의 성능을 평가하기 위해 '비행기 사고', '교통 사고', '재해'의 세 개 도메인에 대해 각각 40개의 문서를 온라인 뉴스 기사에서 수집하였다. 각 문서집합에 있는 모든 문장들을 대상으로 도메인과의 관련성과 이벤트 관련 정보의 포함 정도에 따라 0~4의 점수를 부여하도록 하였다. 다음은 각 도메인 별 문장 점수의 분포를 나타낸 표이다.

[표 2] 도메인 별 문장점수 분포

도메인	크기 (Kb)	0	1	2	3	4	총 문장 수
비행기 사고	160.0	77	146	147	64	37	471
교통 사고	33.3	18	40	100	43	45	246
재해	51.2	25	65	155	77	62	384

5.2 평가 척도

시스템 평가의 척도로는 정확도와 양호도를 사용한다. 정확도(P: precision)는 정보검색 분야에서 일반적으로 사용되는 정확도의 개념에 실험 문서집합의 단계별 문장점수 부여 특성을 반영하여 수식 (4)와 같이 정의하였다. 수식에서 n_i는 시스템에서 추출한 문장 중에서 문장점수 i를 갖는 문장의 개수를 나타낸다.

$$P = \frac{\sum_{i=0}^{4} 0.25i \times n_i}{\sum_{j=0}^{4} n_j} \quad (4)$$

양호도(G: goodness)는 시스템의 재현율을 이용하여 얻어지는 수치로써 '3 또는 4점의 문장점수를 갖는 문장들이 많이 추출되고 0 또는 1점의 문장점수를 갖는 문장들이 적게 추출되는 시스템일수록 좋은 시스템이다'라는 평가관점을 반영하기 위한 평가 척도이다. 이를 위해 수식 (5)와 같이 긍정적 재현율(PR: positive recall)과 부정적 재현율(NR: negative recall)을 이용하여 양호도를 계산한다. 수식에서 S_j는 실험 문서집합 내에서 문장점수 j를 갖는 문장의 개수를 나타낸다.

$$G = PR - NR \qquad \text{where } -1 \le G \le 1 \quad (5)$$

$$PR = \sum_{i=3}^{4} n_i \Big/ \sum_{j=3}^{4} S_j$$

$$NR = \sum_{i=0}^{1} n_i \Big/ \sum_{j=0}^{1} S_j$$

또한 정확도와 양호도로 나타나는 시스템의 성능을 하나의 평가척도로 표현하기 위해 다음의 수식 (6)과 같이 간단하게 이 두 평가척도의 평균값(M)을 이용한다.

$$M = \frac{P+G}{2} \quad (6)$$

5.3 실험결과

실험을 위해 각 도메인 별로 20개의 문서를 학습시키고, 40개의 문서에 대해 이벤트 문장을 추출하도록 하였다. 선택한 각 자질의 기여도를 평가하고 최적 문장 추출 조건을 찾아내기 위해 자질들을 19가지 방법으로 조합하여 이벤트 문장 추출 실험을 수행하였다.

다음의 [표 3]과 [표 4]는 각각 3W 자질과 명사, 명사구, 동사 자질들만을 이용하여 문장을 추출하였을 경우에 대해 시스템 성능을 평가한 결과를 나타낸 표이다. 명사, 명사구, 동사 자질을 사용하는 경우는 문장에서 각 자질 가중치의 평균값이 학습 데이터에서의 평균 가중치 값보다 큰 값을 갖는 경우들만 문장을 추출하도록 하였다. 추출 결과 '비행기 사고' 도메인에서는 When >

Who > Where와 Verb > NP > Noun의 순으로, '교통 사고'와 '재해' 도메인에서는 When > Where > Who와 NP > Verb > Noun의 순으로 문장추출 성능이 높게 나타남을 알 수 있다.

[표 3] 3W 자질 평가결과

조건	평가척도	비행기 사고	교통 사고	재해
Who	P	0.5196	0.6792	0.5923
	G	0.3954	0.2884	0.0348
	M	0.4575	0.4838	0.3135
When	P	0.6308	0.8509	0.7516
	G	0.4414	0.4252	0.692
	M	0.5361	0.638	0.7218
Where	P	0.4651	0.6572	0.6104
	G	0.2349	0.4334	0.3417
	M	0.35	0.5453	0.4761

[표 4] 명사, 명사구, 동사 자질 평가결과

조건	평가척도	비행기 사고	교통 사고	재해
Noun	P	0.4695	0.4695	0.5593
	G	0.2517	0.2517	0.0274
	M	0.3606	0.3606	0.2934
NP	P	0.5473	0.5473	0.5907
	G	0.5626	0.5626	0.2292
	M	0.5549	0.5549	0.4099
Verb	P	0.5553	0.5553	0.577
	G	0.6482	0.6482	0.1077
	M	0.6018	0.6018	0.3423

[표 5]와 [표 6]은 3W 자질과 명사, 명사구, 동사 자질들을 조합하여 문장을 추출한 실험결과를 나타낸다. '비행기 사고' 도메인의 경우 ④>①>③>②와 ⑦,⑧>⑤>⑥, '교통 사고' 도메인의 경우 ③>①>④>②와 ⑧>⑦>⑤>⑥, '재해' 도메인의 경우 ③>④>①>②와 ⑦,⑧>⑤>⑥의 순으로 문장추출 성능이 높게 나타남을 알 수 있다.

[표 5] 3W 자질 조합 평가결과

조건	평가척도	비행기 사고	교통 사고	재해
① Who & When	P	0.7465	0.9737	0.7429
	G	0.4439	0.2159	0.1321
	M	0.5952	0.5948	0.4375
② Who & Where	P	0.5775	0.7444	0.648
	G	0.4246	0.306	0.0687
	M	0.5011	0.5252	0.3584
③ When & Where	P	0.6977	0.8883	0.7756
	G	0.461	0.4028	0.6063
	M	0.5794	0.6456	0.691
④ Who & When & Where	P	0.7778	0.9722	0.7768
	G	0.433	0.2045	0.1106
	M	0.6054	0.5884	0.4437

[표 6] 명사, 명사구, 동사 자질 조합 평가결과

조건	평가척도	비행기 사고	교통 사고	재해
⑤ Noun & NP	P	0.5951	0.6478	0.5907
	G	0.6344	0.5121	0.2292
	M	0.6147	0.58	0.4099
⑥ Noun & Verb	P	0.5591	0.6507	0.5775
	G	0.586	0.4957	0.1044
	M	0.5725	0.5732	0.3409
⑦ NP & Verb	P	0.6638	0.6921	0.6174
	G	0.7069	0.5823	0.2376
	M	0.6853	0.6372	0.4275
⑧ Noun & NP & Verb	P	0.6638	0.6938	0.6174
	G	0.7069	0.5823	0.2376
	M	0.6853	0.638	0.4275

마지막으로 [표 7]은 위의 실험들에서 좋은 성능을 보인 문장추출 조건들과 문장 가중치(W_i)를 조합하여 문장 추출실험을 수행한 결과이다. 위의 실험과 실험 ⑨, ⑩의 결과 3W 자질 중에서 Who 자질을 제외하고 When과 Where 자질을 함께 사용하는 편이 시스템 양호도 수치가 높아 M 값이 좋게 나타남을 알 수 있다. 따라서, 실험 ⑪에서는 When 및 Where 자질과 함께 문장 가중치 값을 이용해 문장을 추출하도록 하였다. 실험 ⑫에서는 문서 내에서 문장 가중치가 높은 문장들을 추출하도록 하였고(θ_1과 θ_2는 3.2.3.의 마지막 부분에서 설명한 임계값들임), 실험 ⑬에서는 ⑪ 및 ⑫의 문장추출 방법을 모두 적용하여 문장을 추출하도록 하였다. 실험결과 실험 ⑬에서와 같이 When 및 Where 자질을 독립적으로 사용하고, 여기에 명사, 명사구 및 동사 자질을 이용해 계산된 문장 가중치를 사용함으로써 이벤트 문장 추출에 있어 좋은 성능을 얻을 수 있음을 알 수 있었다.

[표 7] 조건 및 문장 가중치 조합 평가결과

조건	평가척도	비행기 사고	교통 사고	재해
⑨ (③ & ⑧)	P	0.8095	0.9516	0.7813
	G	0.4663	0.3237	0.2976
	M	0.6379	0.6376	0.5394
⑩ (④ & ⑧)	P	0.8802	0.9643	0.8167
	G	0.4158	0.1591	0.0569
	M	0.648	0.5617	0.4368
⑪ (③ & W_i)	P	0.8148	0.9188	0.85
	G	0.3979	0.3805	0.4205
	M	0.6064	0.6496	0.6353
⑫ (($W_i > \theta_1$) || ((selected < θ_2) & W_i)	P	0.7558	0.7736	0.7791
	G	0.3123	0.4882	0.2446
	M	0.5341	0.6309	0.5118
⑬ (⑪ || ⑫)	P	0.7941	0.8312	0.8182
	G	0.5068	0.7269	0.4997
	M	0.6505	0.779	0.6589

6. 결론 및 향후 연구

본 논문에서는 특정 주제 분야에서 다루어지는 이벤트들에 대한 구체적인 정보를 포함하고 있는 문장들을 자동으로 추출하도록 하는 이벤트 문장추출 시스템을 제안하였다. 이는 정보추출 분야에서 문제가 되는 도메인 별 정보구축의 어려움을 극복하기 위해 자동요약 분야에서 사용되는 문장추출의 개념을 도입한 새로운 시스템이다.

본 시스템에서는 이벤트의 주체, 객체, 발생 일시 및 장소에 대한 정보를 포함하는 이벤트 문장을 추출하기 위해서 동사, 3W, 명사 및 명사구의 세 가지 자질을 사용하였다. 이벤트 문장 추출 시스템은 학습 단계와 추출 단계로 나누어 진다. 학습 단계에서는 특정 도메인을 대상으로 명사, 명사구 및 동사 자질을 추출한 후 이에 대한 통계정보를 수집한다. 추출 단계에서는 문장분석 결과에 학습 단계에서 수집한 도메인 정보를 반영함으로써 이벤트 문장을 추출한다.

각 자질의 유용성과 시스템의 성능을 평가하기 위해 ' 비행기 사고' , ' 교통 사고' , ' 재해' 도메인을 대상으로 각 자질 및 자질의 조합을 이용해 실험을 수행하였다. 실험 결과, 선택한 자질들을 적절히 조화시킴으로써 도메인에 의존적이면서 일회적인 정보를 포함하고 있는 이벤트 문장을 효과적으로 추출할 수 있었다. 3W 자질 중에서는 When과 Where 자질이 Who 자질에 비해 이벤트 문장을 추출하는데 더 유용함을 알 수 있었다. 또한 문장 가중치를 계산하는 단계에서 명사, 명사구, 동사 자질을 독립적인 조건으로 사용하는 것보다 복합적으로 반영하는 것이 문장추출에 보다 효과적임을 알 수 있었다.

향후 연구에서는 동사 자질 선택 방법을 동사의 의미적인 측면까지 고려하여 선택하도록 할 것이다. 또한 공기정보를 반영할 때 동사 자질과 다른 자질 간의 격관계를 이용하도록 할 것이다. 그리고, 이벤트 문장추출의 결과를 재귀적으로 문장추출의 새로운 부가 정보로 활용하거나 정보추출을 위한 도메인 패턴 자동구축에 이용하는 방안을 모색하고자 한다.

7. 참고 문헌

[1] Ralph Grishman, *"Information Extraction: Techniques and Challenges"*, In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD, April 1998.

[2] 임수종, 정의석, 황이규, 윤보현, *"이벤트 템플릿을 이용한 정보 추출에 관한 연구"*, 2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호, pp.585-588, 2002.

[3] 오효정, 임정묵, 이만호, 맹성현, *"유한 오토마타를 이용한 정보 추출 시스템의 구현 및 분석"*, 제10회 한글 및 한국어정보처리 학술대회, pp.97-104, 1998.

[4] 장동현, *"문장 클러스터링을 통한 텍스트 자동요약에 관한 연구"*, 충남대학교 컴퓨터과학과 박사 학위 논문. 2002.02.

[5] J. Kupiec, J. Pedersen, and F. Chen, *"A Trainable Document Summarizer"*, Proceedigns of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.68-73, 1995.

[6] Eduard Hovy and Daniel Marcu, *"Automated Text Summarization"*, Pre-conference Tutorial of the COLING/ACL'98, Université de Montréal Montréal/Canada, August 1998.

[7] Yiming Yang, Tom Pierce, and Jaime Carbonell, *"A Study on Retrospective and On-Line Event Detection"*, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.28-36, 1998.

[8] 김계성, 이현주, 정영규, 서연경, 손기준, 이상조, *"단락 자동 구분을 통한 중요 문장 추출"*, 제 12 회 한글 및 한국어 정보처리 학술대회, pp.233-237, 2000.