시간 및 다국어 공간에서 어휘 분포에 기반한 다국어 사건 링크 탐색

전북대학교 이경순

1. 서 론

사건 탐색 및 추적(TDT: Topic Detection and Tracking) 연구 [1]은 전세계 각 나라에서 매일 보도되고 있는 신문이나 방송 뉴스 기사에서 "어떤 중요한 사건이 발생했는가?" 또는 "새로운 사건이 일어났는가?"와 같이 그날 처음 발생한 사건을 탐색하거나, 같은 사건을 다루는 기사들을 탐색하거나, 예전에 발생한 사건과 관련된 사건인지를 추적해 나가는 것이다.

TDT에서의 'topic'은 구체적인 시간과 구체적인 장소에서 발생한 '사건(event)'을 나타내는 것으로, 정보검색에서의 '주제(subject)'와는 구분된다. 예를 들어, '비행기 사고'는 정보검색에서의 주제에 해당하고, '대한항공 007기 추락'은 사건에 해당한다. 사건 및 탐색에서의 사건은 정보검색에서의 주제에 대한 실례(instance)라고 볼 수 있다. 본 글에서는 정보검색에서 질의에 해당하는 topic과 용어 사용에서의 혼동을 피하기 위해, TDT에서의 topic을 '사건'으로 표현했다.

사건 탐색 및 추적은 그날 또는 오랫동안 신문이나 방송에서 보도된 대량의 뉴스 기사들을 모두 읽을 수 없는 상황에서 중요한 사건이나 관심 있는 사건에 대한 정보를 알기를 원할 때에 유용한 것으로, 기존의 정보검색에서와는 달리 사용자에게 '사건 중심적인 정보'를 제공한다.

사건 탐색 및 추적 연구와 정보검색, 질의응답 연구를 비교하면, 정보검색(Information Retrieval)은 사용자가 제시하는 질의에 대해서 정보검색 시스템이 질의와 관련된 내용의 문서를 검색해서 사용자에게 제시해 주는 것으로, 사용자의 질의에 의해 검색이 시작되고, 문서 내용에 중점을 둔 검색을 한다. 질의응답(Question Answering)에서도 사용자가 알고자 하는 내용을 구체적 질문으로 만들고, 질의응답 시스템은 사용자의 질문에 대해 문서에서 대답을 찾아서 제시한다. 예를 들면, 사용자가 "마틴 루터 킹을 죽인 사람은 누구인가?"라고 질문을 던지면, 질의응답 시스템은 질문에 나타난 실마리 어휘들을 기반으로 해서 문서에서 대답을 찾아서 사람 이름을 제시해 준다. 이와 같이 정보검색과 질의응답은 사용자가 찾고자 하는 내용에 대해 정확하게 질의를 만들 수 있을 때에는 좋지만, "어떤 일이 발생했는가?"와 같이 사용자가 알지 못하는 것에 대해 어떠한 정보를 얻고자 할 때는 질의를 만들지 못하기 때문에 정보를 획득하기가 어렵다.

사건 탐색과 추적 연구와 문서 클러스터링과 문서 범주화 연구를 비교해 보면, 비슷한 점과 다른 점은 다음과 같다. 문서 클러스터링(Document Clustering)과 사건 탐색은 학습을 위한 데이터 없이 문서들의 그룹을 만든다는 측면에서는 같지만, 생성되는 문서들의 그룹이 문서 클러스터링은 비슷한 내용의 문서들이고, 사건 탐색에서는 같은 사건을 다루는 문서들이라는 측면에서 구분된다. 또한 사건 탐색에서의 문서 그룹은 새로 들어오는 뉴스 기사에 대해서 점진적으로 클러스터를 형성해 나가는 반면, 문서 클러스터링에서는 전체 문서에 대해 일괄적으로 그룹을 형성하는 경우가 많다. 문서 범주화(Text Categorization) 또는 정보 필터링(Information Filtering)과 사건 추적은 주어진 학습데이타를 이용해서 학습을 하고, 주어진 문서의 범주를 선택한다는 측면에서는 같지만, 문서의 범주의 성격이 유사한 내용의 문서 범주이고, 같은 사건을 다루는 문서들의 범주라는 측면에서 구분된다. 또한, 학습할 수 있는 데이터의 양에 있어서 사건 추적에서는 적은 개수의 예제가 제공된다.

본 논문에서는 사건 탐색 및 추적 연구의 핵심 기술에 해당하는 부분인 사건 링크 탐색 (Story Link Detection) 방법에 대해 소개하면서, 사건 탐색 및 추적 연구에서의 대상 문서인 뉴스 기사와 사건을 나타내는 어휘들의 특성을 살펴보고자 한다. 사건 링크 탐색은 임의의 두 문서가 주어졌을때 그것이 같은 사건을 다루고 있는지 아닌지를 판정하는 것으로, 사건 탐색 및 추적 연구에 적용될 수 있는 기법이다.

본 글의 2장에서는 사건 탐색 및 추적 연구의 동향을 살펴보고, 3장에서는 시간 및 다국어 공간에서 사건과 관련된 어휘의 분포 특성에 따른 사건 링크 탐색 방법을

설명하고, 4장에서는 3장에 소개된 방법에 대한 실험 결과를 보이고, 5장에서 결론을 맺는다.

2. 사건 탐색 및 추적 (TDT) 관련 연구

사건 탐색 및 추적 연구는 미국 국립표준기술연구소(NIST; National Institute of Standards and Technology)에서 1996년부터 미 국방성의 지원을 받아 시작되었다. 1999년에 TDT 평가대회를 시작하여 매년 개최하여, 영어, 중국어, 아랍어 등 다국어 신문 방송 기사에 대해 다루면서 다국어 테스트컬렉션을 구축해 오고 있고, 다음과 같은 세부 연구분야를 다루고 있다.

- 사건기사 분리(Story Segmentation)

라디오나 방송 등 뉴스 기사를 대해 사건과 사건보도의 경계를 탐색해서 분리하는 것. 오디오 기사에 대해 자동 음성변환기를 이용해서 변환된 텍스트에 대해서 한 사건에서 다음 사건으로 내용이 넘어가는 부분을 찾는 것이다.

- 사건 추적 (Topic Tracking)

사건에 대한 몇개의 예제 기사를 제시해 주고, 그 사건과 관련된 뉴스 기사를 추적하는 것. 사용자가 관심있는 사건을 지정해 주면 뉴스 기사에서 그와 관련된 사건이 보도되는지를 계속적으로 추적해서 사용자에게 알려주는 것이다.

- 사건 탐색 (Topic Detection)

같은 사건을 다루는 기사들의 클러스터를 생성하는 것. 새로 입력되는 뉴스 기사에 대해 점진적으로 클러스터를 형성해 나가는데, 기존에 탐색된 사건 클러스터와의 유사도가 임계치 이상이면 그 사건의 클러스터의 멤버로 포함시키고, 그렇지 않으면 새로운 사건으로 인식되어, 새로운 클러스터를 생성하게 된다.

- 새로운 사건 탐색 (New Event Detection)

기사가 예전에 발생하지 않은 새로운 사건을 다루는지를 탐색하는 것. 입력되는 각 뉴스 기사에 대해서 새로운 사건인지 새로운 사건이 아닌지를 결과로 준다. 사건 탐색에서 제일 처음 탐색된 사건을 나타내는 것으로, 사용자에게 새로운 사건에 대한 정보를 제공할 수 있다.

- 사건 링크 탐색 (Story Link Detection)

두 개의 뉴스 기사가 같은 사건을 다루는지 아닌지를 탐색하는 것. 이는 사건 탐색 및 추적 연구의 요소 기술에 해당한다.

사건 탐색 및 추적 연구에서 대부분의 접근 방법은 기존의 문서 내용 중심적 문제인 문서 클러스터링과 문서 범주화에 대한 접근 방법과 별로 다르지 않았다. 연구[2]에서는 TDT2002 평가대회에서 사건 링크 탐색을 위해서 두 뉴스 기사가 같은 사건을 다루는지의 유사도 측정을 하는데 있어서, 20여 가지의 유사도 측정 기법을 적용하여 계산을 하고, 그 유사도 값들을 조합하여 사건 링크 탐색을 수행하였다. 연구[3]에서는 뉴스 기사들을 표현하기 위해 명사, 동사, 형용사, 복합명사 등을 추출하여 표현하였고, 문서의 길이에 따른 유사도 측정값의 차이를 줄이기 위해서, 문서의 길이를 확장하는 방법을 이용하였다. 단일언어나 다국어에서 사건 링크 탐색을 위해서는 임계치에서 차이를 둔 정도이다.

사건과 관련된 어휘들의 행태를 고려한 연구들이 있다. 사건 추적에서 연구[4]는 사건과 관련된 어휘는 여러 문단에 걸쳐서 두루 나타나지만 주제와 관련된 어휘는 그렇지 않다고 가정하고, 사건과 관련된 어휘들의 영역 의존도를 고려하여 어휘 가중치를 부여하였다. 연구[5]는 시간상에서 어휘의 중요도를 계산하여, 뉴스기사들에서 주요하게 다루고 있는 정보를 파악하기 위해 어휘들의 클러스터를 생성하여 제시하였다. 많은 연구들에서 사건을 나타내는 요소에 해당하는 개체 인식 (Named Entity)을 포함하고 있다[6,7].

다국어 사건 탐색 및 추적 연구에서는 아랍어와 영어 뉴스 기사에 대한 사건 링크 탐색을 위해서 아랍어-영어 사전과 번역 확률을 이용한 연구가 있다[8]. 아랍어, 중국어, 영어에 대한 사건 추적에서 연구[9,10]에서는 통계사전에 기반한 두개의 가장 좋은 대역어 선택과 번역후 문서확장 방법이 기계번역 시스템에 의한 하나의 대역어를 선택하는 것보다 더 좋은 성능을 보였다. 다국어 사건 탐색 기법은 사전기반 번역 또는 기계번역 시스템 기반 번역 등과 같이 언어번역 과정을 거친 후에는 대부분 언어 중심적인 정규화 과정에 중점을 두었다[8,3,11].

3. 시간 및 다국어 공간에서 어휘 분포에 기반한 사건 링크 탐색

본 논문에서는 사건 탐색 및 추적 연구의 핵심 기술에 해당하는 사건 링크 탐색 기법에 대해 자세히 설명하면서, 사건 탐색 및 추적 연구에서의 대상 문서인 뉴스 기사와 사건을 나타내는 어휘들의 특성을 살펴보고자 한다.

다국어 사건 링크 탐색을 위해서 시간 및 다국어 공간에서 어휘 분포에 따라 사건을 나타내는 어휘들의 가중치에 변별력을 줌으로써, 두 문서가 같은 사건을 다루는지 관련도 측정시 영향을 미칠 수 있도록 한다. 다음과 같은 관찰과 가정에 기반을 두어서 다국어 사건 링크 탐색에 접근하는 방법을 설명한다.

- 사건 어휘 그 자체의 특성: 사건은 "누가, 언제, 어

디서, 무엇을, 왜, 어떻게 했다"와 같은 요소들로 기술된다. 이들의 개체에 해당하는 〈사람〉, 〈지역〉, 〈시간〉 등 사건 요소에 해당하는 개체 인식은 사건의 주요 객체에 대한 인식에 도움이 될 것이다.

- 사건 어휘가 문서에서 나타나는 행태: 사건과 관련된 어휘들은 그 사건을 설명하기 위해 문서의 전체에 걸쳐서 두루 나타난다.
- 사건 어휘의 시간의 흐름에서의 분포 특성: 시간상의 한 시점에서 새로운 사건을 보도하는 기사에서는 사건과 관련된 중요한 어휘들이 새로 등장하고, 어휘 빈도수에 있어서 빠르게 변화한다. 한 시점에서의 어휘의 분포와 지속적인 시간 동안의 어휘의 분포를 상대적으로 비교함으로써 한 시점에서의 중요한 어휘를 파악할 수 있다.
- 사건 어휘의 다국어 공간에서의 분포 비교: 어떠한 사건에 대해 신문이나 방송에서 보도되는 양의 정도는 사건의 중요도로 볼 수 있는데, 이는 각 나라마다 그 나라에 중요하거나 관심있는 사건인가에 따라 다를 것이다. 따라서 다른 언어 공간에서의 어휘의 분포를 참조함으로써 다국어에 대해서 같은 사건을 다루는지 탐색에 도움이 될 수 있다.

3.1 다국어 사건 탐색을 위한 언어 번역

다국어 문서에 대해서 같은 사건을 다루는지를 탐색하기 위해서는 같은 언어 공간으로 변환을 해주어야 한다. 본 논문에서는 한국어와 일본어 뉴스 기사에 대해서 다룬다. 한국어와 일본어 뉴스 기사의 언어 공간을 하나로 하기 위해, 한국어-일어 문서 번역기를 이용하여 한국어를 일본어로 변환하였다.

다국어 언어 공간을 하나의 공간으로 매핑하기 위해서는 사전 등을 이용한 어휘 번역 방법이나 기계 번역기를 이용해서 문서를 번역하는 방법을 선택할 수 있는데, 한국어와 일본어는 기계 번역기가 비교적 좋은 성능을 보이기 때문에 문서 번역기를 이용할 수 있다.

3.2 시간상에서 어휘의 분포 변화에 따른 중요도 계산

각 뉴스 기사 문서를 표현하기 위해 문서에 나타나는 어휘들에 대해서 품사 태깅을 거쳐서 명사, 고유명사, 형용사와 동사를 선택하였다. 또한, 사건을 구성하는 주요 개체의 인식을 위해 〈사람〉, 〈조직〉, 〈나라〉, 〈지역〉, 〈시간〉을 나타내는 개체를 인식하여 표현한다. 품사가 동사로 태깅된 것이나 동사적 명사로 태깅된 것은 〈동작/상태〉를 나타내는 것으로 한다. 어휘의 표현 단위는 명사의 나열이나 구 단위로 표현되어 자주 나타나는 것을 사건 표현의 한 단위로 다루기 위해 모든 가능한 어휘들의 조합으로 생성하여 표현한다.

추출된 어휘에 대해서, 시간상에서 '어느 한 시점'에서의 어휘 분포와 '어느 연속적인 시간'의 어휘 분포를 상대적으로 비교함으로써 한 시점에서의 중요하게 다뤄지는 사건의 어휘를 파악한다. 이때, '어느 한 시점'을 '그날 하루'의 뉴스 기사들로 하고, '어느 연속적인 시간'을 예전부터 '그날까지'의 모든 뉴스 기사들로 하여, 어휘의 중요도는 카이제곱(x^2)으로 계산한다.

카이제곱 측정은 어휘 t와 시간범주 t0의 독립 정도를 측정하는 것으로, 보통 문서 범주화에서 그 범주에서 중요한 어휘 (자질)를 추출하는데 많이 이용하고 있는 방법이다. 본 논문에서는 범주 t0를 '시간'으로 두었다.

표 1 시간상에서 중요 어휘를 계산하기 위한 분할표

	어휘 t를 포함하는 문서	어휘 t를 포함하지 않는 문서
t0에 속하는 문서	a	b
t0에 속하지 않는 문서	c	d

표 1은 2 x 2로 된 분할표인데, t0는 시간상에서 하루에 해당하는 '그날'이고, 해당하는 그날에 보도된 뉴스 기사들로 이뤄진 범주이다. 즉, '그날의 모든 뉴스 기사들'이 하나의 범주에 속하는 문서가 된다. 따라서 하루하루의 뉴스 기사들은 각각 하나의 범주가 된다. 전체 문서는 그날 이전의 모든 문서들이다. 분할표를 이용하여 각 어휘의 중요도는 다음과 같이 계산한다.

$$\chi^2 = \frac{(a+b+c+d)\cdot(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \qquad (1)$$

여기서, a는 시간 범주인 t0의 문서들에서 어휘 t를 포함하는 문서의 개수를 나타내고, b는 시간 범주 t0의 문서들에서 어휘 t를 포함하지 않는 문서의 개수를 나타낸다. 그러므로 a와 b를 더한 값은 t0 하루동안 보도된 문서의 개수가 된다. c는 현재 보이는 모든 시간상에서의 문서집합에서 시간 범주인 t0를 제외한 모든 시간동안의 문서들에서 어휘 t를 포함하는 문서의 개수를 나타낸다. a는 시간범주 t0에서의 문서 빈도수(document frequency)를 나타내고, a와 c를 더한 것은 전체 문서집합에서의 어휘 t의 문서 빈도수를 나타낸다.

어떤 어휘의 카이제곱 값이 높다는 것은 그 어휘가 '전체' 뉴스 기사들에서 나타난 빈도에 비해서 '그날' 뉴스 기사들에서 나타난 빈도가 비교적 높은 것이다. 즉, 그날의 중요 사건과 관련된 어휘일 가능성이 높다는 것이다.

시간에서의 그 어휘의 중요도는 다음과 같이 각 언어에서 그 어휘의 카이제곱 값으로 한다.

$$wTime(t,t0,l) = \chi^2(t,t0) \qquad (2)$$

여기서, t는 어휘를 나타내고, t0는 '그날'에 해당하는 시간의 범주를 나타내고, l은 언어 공간을 나타낸다. 그림 1은 한국어 뉴스 기사에 대해서 시간의 흐름에 따라 어휘의 중요도인 카이제곱 값의 변화를 보여주고 있다.

어휘 '김일성'은 '김일성 사망' 사건이 발생한 시점인 1994년 7월 10일에서 어느 기간동안 급작스럽게 많이 분포해서 높은 중요도를 갖고, 어휘 '지진'은 '고베 지진' 사건이 발생한 시점인 1995년 1월 18일에서 어느 기간동안 많은 분포를 나타내서 높은 중요도를 갖게 되었다. 이와같이 사건과 관련된 어휘는 사건이 발생한 시점에서 어느정도 시간이 흐르면서 어휘의 분포에 따른 중요도에서도 낮은 값을 갖고 있음을 알 수 있다.

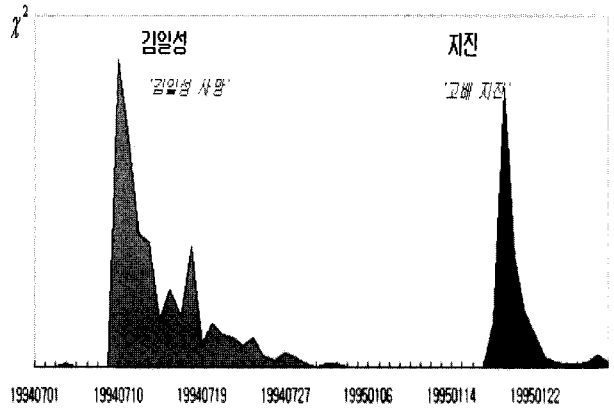

그림 1 시간상에서 어휘의 중요도 변화

3.3 다국어 공간에서 어휘의 분포 비교

각 나라에서 다루는 뉴스는 그 나라에서 발생한 사건 또는 그 나라와 관련이 있는 사건을 주요하게 다루고, 다른 나라에서 발생한 사건들은 그 나라와의 관련 정도에 따라 다르게 다룬다.

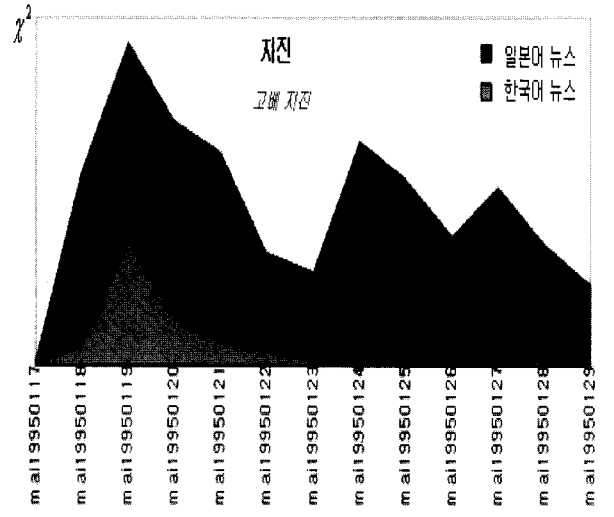

그림 2 다국어 공간에서 어휘 분포 비교

표 2 다국어 공간에서 어휘의 날짜 빈도수

어휘	한국어	일본어	영어
인천	261	31	3
히로시마	61	307	25
NHK	32	292	4
플로리다	17	48	249

각 나라/민족에는 사람이름, 지역이름, 회사, 조직 등 그들 고유의 어휘들이 많이 있으므로, 표 2에서 보는 바와 같이 각 나라마다 나타나는 어휘의 분포에는 차이가 있다. 각 나라에서의 어휘 공간을 언어 공간이라고 하자.

표 2는 1년 동안의 뉴스 기사에서 각 어휘가 나타난 날짜의 빈도수 (date frequency)를 나타낸 것이다. 1년 동안의 전체 뉴스 기사에서 한국에 위치한 지역을 나타내는 어휘인 '인천'은 한국어 공간에서 261일 나타났고, 일본어 공간에서는 31일 나타났다. 일본에 있는 조직을 나타내는 'NHK'는 일본어 공간에서는 292일 나타났고, 한국어 공간에서는 32일, 영어 공간에서는 4일 나타났다. 이와같이 그 나라와 관련된 어휘들은 그 나라의 언어 공간의 뉴스에서 자주 보도되는 경향을 보인다. 같은 사건을 보도하는 문서의 개수에 있어서도 사건에 대한 그 나라의 관심의 정도에 따라 의존한다.

그림 2는 '고베 지진' 사건이 발생한 시간대에서 한국어 뉴스와 일본어 뉴스에서 사건과 관련된 어휘인 '지진'의 분포를 비교한 것이다. 그 사건과 직접 관련이 있는 일본어 공간에서 어휘의 중요도가 훨씬 높게 나타나는 것을 볼 수 있다. 이를 통해서, 사건과 관련된 어휘는 시간의 흐름뿐만 아니라 다국어 언어 공간에 따라 영향을 받는다는 것을 알 수 있다.

$$wTimeSpace(t,t0) = \max \arg_l wTime(t,t0,l) \quad (3)$$

두개 이상의 언어 공간은 하나의 언어 공간에서 보다 더 많은 정보를 제공할 수 있다고 보기 때문에, 본 논문에서는 다국어 언어 공간을 합쳐서 어휘의 분포를 계산하지 않고, 서로 다른 언어 공간에서 어휘의 분포를 각각 측정하여 중요도로 계산하고, 각 시간에서의 다른 언어 공간에서의 어휘의 중요도를 비교 참조한다.

하나의 언어 공간에서 높은 중요도를 갖는 사건과 관련된 어휘는 다른 언어 공간에서 낮은 중요도로 나타났다 할지라도 그 사건을 나타낼 가능성이 있다. 그러므로 다국어 공간에서 어휘의 분포를 비교하여 높은 값을 갖는 것을 취함으로써, 다국어 사건 탐색에서의 다리 역할을 하도록 한다.

3.4 사건 관련도 측정

시간 및 다국어 공간에서 어휘의 분포를 이용하여 어

휘의 가중치는 다음과 같이 계산한다.

$$wgt_t = tf_t \cdot wNE_t \cdot wTimeSpace_t \quad (1)$$

여기서 wgt는 어휘 t의 가중치를 나타내는 것으로, tf는 어휘 빈도수, wNE는 사건의 요소에 해당하는 개체인가에 따라, 〈사람〉, 〈지역〉, 〈국가〉 등에 개체 인식 단계에서 인식한 개체인 것에 대해 2의 값을 부여해서, 일반 어휘 (디폴트=1) 보다 높은 가중치를 갖도록 했다. 그리고 시공간상에서의 어휘 중요도 wTimeSpace를 곱한 값이다.

각 문서는 각 어휘에 대한 가중치의 벡터로 표현을 한다. 사건 링크 탐색에서 두 문서가 같은 사건을 다루는지를 관련도를 측정하기 위해서 두 문서 벡터에 대한 코사인 계수를 계산한다. 유사도에 대한 임계치에 따라 같은 사건 또는 다른 사건을 다룬다고 판단을 한다.

4. 실험 및 평가

시간 및 다국어 공간에서 어휘의 분포를 이용하여 가중치를 계산하여 사건 링크를 탐색하는 방법이 유효한지를 보기 위해, 한국어 뉴스 기사와 일본어 뉴스 기사로 구성된 다국어 테스트 컬렉션을 이용하여 평가를 하였다.

4.1 실험 환경 설정

문서 집합은 한국어와 일본어 신문 기사로 구성되어 있는데, 한국어는 인터넷에 보도된 뉴스 기사를 수집한 것이고, 일본어는 마이니치 신문 기사이다. 문서의 날짜는 1998년 1월에서 1998년 6월까지 보도된 것으로, 문서의 개수는 한국어는 40,000개, 일본어는 61,637개이다.

각 문서의 어휘들은 품사 태거 ChaSen 시스템 [12]을 이용하여 추출하였다. 한국어 문서 공간에서 나타난 어휘는 193,730개이고, 일본어 문서 공간에서 나타난 어휘는 353,210개였다.

매일 보도되는 뉴스 기사에 대한 사건 탐색이기 때문에, 그날 뉴스 기사가 추가될 때마다, 가중치 계산에서 사용되는 문서 빈도수는 점진적으로 계산하였다. 개체 인식을 위해서는 NExT 개체인식 시스템 [13]을 이용하여 사건의 요소에 해당하는 개체들을 인식하였다.

본 실험의 사건 탐색에서 다룬 사건은 13개로 구성되어 있는데, 이는 TDT2 테스트 컬렉션에 포함된 사건의 일부이다. 같은 시기에 한국어, 일본어, 영어로 보도된 뉴스 기사에 대한 다국어 사건 탐색을 위해 이를 이용한 것인데, 현재 본 논문에서는 한국어와 일본어에 대해서만 실험을 한 것이다. 사건은 표 3에 나타난 것으로 국제적인 사건에 해당되는 것들이다.

사건 'Upcoming Philippine Elections'에 대한 구체적인 설명은 다음과 같이 기술되어 있다.

- **WHAT : National elections in the Phil-ippines**
- **WHERE : Manila, Philippines**
- **WHEN : January 1998 (cabinet resignations) through May 1998 (new president elected)**

각 사건을 다루고 있는 뉴스 기사에 대한 정답 평가는 한국어와 일본어 각 언어에 대해 각 두명의 평가자가 평가를 하였다. 13개의 사건에 대해 5,902개의 문서를 평가하였는데, 이는 사람이 다양한 키워드를 넣어 정보검색을 여러번 수행하여 사건과 관련이 높은 기사들을 추출한 것이다. 그 중에서 3,875개가 사건을 다루는 기사로 평가되었다. 평가를 위한 기준은 LDC (Linguistic Data Consortium)에서 TDT2 테스트 컬렉션을 구축하기 위해 정의한 것을 따랐다. 다국어 사건 링크 탐색을 위해서 관련이 있는 사건의 쌍 1,731,419개와 관련이 없는 사건의 쌍 5,224,891개에 대해서 평가를 하였다.

표 3 탐색할 사건 리스트

Upcoming Philippine Elections
1998 Winter Olympics
Current Conflict with Iraq
China Airlines Crash
Tornado in Florida
Asteroid Coming
Viagra Approval
India, A Nuclear Power
Israeli-Palestinian Talks (London)
Anti-Suharto Violence
Anti-Chinese Violence in Indonesia
Afghan Earthquake
Clinton-Jiang Debate

시스템의 성능 평가는 정확률(precision), 재현률(recall), 누락률(miss), 오류률(false alarm), 마이크로 평균 F1(micro-average F1)으로 측정하였다. 사건 탐색 연구에서 누락률과 오류률을 이용해서 성능 평가를 하고는 있으나, 이것이 정확률과 재현률 평가에 대해서 다른 의미를 보여주고 있지는 못하다.

4.2 실험 결과

시간 및 다국어 공간에서 어휘의 분포 특성을 이용하여 어휘의 가중치를 부여한 것의 사건 탐색 성능을 비교 평가하기 위해 일반적으로 어휘의 가중치 계산에 많이 이용되고 있는 어휘 빈도수(tf)와 역문서 빈도수(idf)에 의한 tfidf 가중치 계산의 성능과 비교 평가를 하였다.

표 4 사건 링크 탐색 성능 비교

	같은 언어 쌍				같은 언어 및 다른 언어쌍	
	한국어 뉴스 기사		일본어 뉴스 기사			
	tfidf	proposed	tfidf	proposed	tfidf	proposed
정확률	0.3865	0.4240	0.2899	0.3313	0.3025	0.3559
재현률	0.8506	0.9042	0.9808	0.9131	0.9657	0.8970
누락률	0.1494	0.0958	0.0192	0.0869	0.0343	0.1030
오류률	0.2983	0.3298	0.6929	0.5765	0.5870	0.4601
마이크로평균 F1	**0.6593**	**0.7735**	**0.7349**	**0.8040**	**0.6896**	**0.7880**

표 5 다국어 공간의 적용에 따른 성능 비교

	한국어-일본어 다국어 뉴스 기사 쌍		
	tfidf	다국어 공간 적용 않음	다국어 공간 적용
정확률	0.3468	0.3678	0.3769
재현률	0.8799	0.7560	0.8324
누락률	0.1201	0.2440	0.1676
오류률	0.3992	0.3466	0.3734
마이크로 F_1	0.6566	0.6719	0.7665

제안한 가중치 기법에 의한 문서 벡터에서는 일부 어휘의 가중치가 다른 어휘들에 비해서 뚜렷하게 높은 가중치를 갖고 대부분은 아주 작은 값을 갖는 것을 볼 수 있었다. 문서 벡터의 어휘들의 가중치에 변별력이 있기 때문에, 이들은 두 문서의 사건 링크 탐색에서 유사도 측정에서도 그 영향을 미치게 된다.

표 4는 사건 링크 탐색의 실험 결과를 보여준다. 사건 탐색에서 유사도 값에 대한 임계치를 0.005에서 0.35까지 변화시켜서 가장 좋은 성능을 보일때의 결과이다. 본 논문에서 제안한 시간 및 다국어 공간에서 어휘 분포 특성을 고려한 방법 (proposed)이 일반적 가중치 계산 기법 (tfidf)에 비해 마이크로 평균 F1에서 14.3% 성능향상을 보였다.

표 5는 다국어 공간에서의 어휘 분포 비교를 적용한 것이 다국어 사건 탐색에서 유용했는지를 보기위한 것으로, 다국어 공간을 고려한 것이 그렇지 않은 것에 비해 14.1% 성능 향상을 보이고 있다. 이를 통해서, 다국어 사건 탐색에서 언어 공간의 참조는 효과적이다고 할 수 있다.

실험 결과를 통해서, 같은 사건을 다루는 뉴스 기사를 탐색하기 위해서, 사건 뉴스 기사에 나타나는 어휘의 시간 및 다국어 공간에서의 분포 특성을 이용하여 가중치를 계산하여 사건과 관련된 어휘의 가중치에 변별력을 줌으로써 사건 탐색에서의 관련도 계산에 영향을 주도록 한 방법이 효과적임을 볼수 있다.

5. 결 론

본 논문에서는 다국어 뉴스 기사에 대해서 시간 및 다국어 공간에서의 어휘 분포 특성을 이용하여 가중치를 적용한 방법이 한국어와 일본어 뉴스 기사에 대한 다국어 사건 링크 탐색에서 효과적임을 보았다. 이러한 결과는 뉴스 기사에서 사건을 나타내는 어휘의 빈도 분포가 사건의 발생과 전개 등 시간의 흐름에 따라 크게 변화하고 있고, 다국어 공간에서도 사건에 대한 그 나라의 관심 정도에 따라 차이가 있다고 볼 수 있겠다. 사건과 연관된 어휘의 시간과 다국어 공간에서의 특성은 사건 탐색 및 사건 추적에 적용될 수 있다.

다국어 뉴스 기사에 대한 사건 탐색 및 추적 연구의 매력은 같은 사건에 대해서 각 나라마다 그 사건을 보도하는데 있어서 그 관점이 다른 경우가 많다. 이러한 현상은 같은 나라 안에서도 신문/방송사마다 같은 사건에 대한 보도 관점이 다른 경우에 나타난다. 앞으로 계속 연구가 필요한 부분으로, 여러 나라에서 보도된 다국어 기사에서 같은 사건 문서를 탐색한 후, 그 문서들 관점을 비교 평가할 수 있다면, 어떤 사건에 대한 국가/민족/문화의 시각의 차이에 대한 정보를 제공할 수 있어, 서로 다른 국가나 문화를 이해하는데 도움이 될 것이다.

참고문헌

〔1〕 Fiscus, J., Doddington, G., Garofolo, J. and Martin, A. 1999. NIST's 1998 topic detection and tracking evaluation (TDT2). Proc. of DARPA Broadcast News Workshop.

[2] Carbonell, J., Yang, Y., Brown, R., Zhang, J. and Ma, N. 2002. New event & link detection at CMU for TDT 2002. Proc. of Topic Detection and Tracking (TDT-2002) Evaluations.
[3] Chen, Y. and Chen, H. 2002. NLP and IR approaches to monolingual and multilingual link detection. Proc. of 19th International Conference on Computational Linguistics.
[4] Fukumoto, F. and Suzuki, Y. 2000. Event tracking based on domain dependency. Proc. of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
[5] Swan, R. and Allan, J. 2000. Automatic generation of overview timelines. Proc. of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000).
[6] Eichmann, D. 2002. Tracking & detection using entities and noun phrases. Proc. of Topic Detection and Tracking (TDT-2002) Workshop.
[7] Yang, Y., Zhang, J., Carbonell, J. and Jin, C. Topic-conditioned novelty detection. Proc. of the International Conference on Knowledge Discovery and Data Mining, Edmonton (KDD 2002).
[8] Lam, W. and Huang, R. 2002. Link detection for multilingual new for the TDT2002 evaluation. Proc. of Topic Detection and Tracking (TDT-2002) Workshop.
[9] Levow, G-A. and Oard, DW. 2000. Translingual topic detection: applying lessons from the MEI project. Proc. of Topic Detection and Tracking (TDT-2000) Workshop.
[10] He, D., Park, H-R., Murray, G., Subotin, M. and Oard, DW. 2002. TDT-2002 topic tracking at Maryland: first experiments. Proc. of Topic Detection and Tracking (TDT-2002) Workshop.
[11] Leek, T., Jin, H., Sista, S. and Schwartz, R. 1999. The BBN crosslingual topic detection and tracking system. Proc. of Topic Detection and Tracking (TDT-1999) Workshop.
[12] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K. and Asahara, M. 2002. Morphological analysis system ChaSen version 2.2.9. Nara Institute of Science and Technology.
[13] Masui, F., Suzuki, N. and Hukumoto, J. 2002. Named entity extraction (NExT) for text processing development. Proc. of 8th time annual meeting of The Association for Natural Language Processing (Japan). http://www.ai.info.mie-u.ac.jp/~next/next.html

이 경 순

1990. 3~1994. 2 계명대학교 컴퓨터공학과(학사)
1995. 3~1997. 2 한국과학기술원 전자전산학과(석사)
1997. 3~2001. 8 한국과학기술원 전자전산학과(박사)
2001. 12~2003. 11 일본 국립정보학연구소(Natimal Institute of Informatics) 연구원
2004. 3~현재 전북대학교 전자정보공학부 전임강사
관심분야 : 정보검색, 지식마이닝, 자연언어처리
E-mail : selfsolee@chonbuk.ac.kr