사건 어휘의 특성을 반영한 다국어 사건 연결 탐색*

이 경 순†

요 약

본 논문에서는 다국어 뉴스에 대해서 '시간' 요소와 '언어 공간' 요소를 사건 어휘의 가중치 계산에 반영하는 다국어 사건 연결 탐색하는 방법을 제안한다. 시간의 흐름과 다국어 공간상에서 어휘의 분포 특성을 어휘의 가중치로 반영하여 사건 중심 어휘에 변별력을 줌으로써 같은 사건을 다루는 문서를 탐색하도록 한다. 시간상에서 어휘가중치는 전체 시간의 모든 문서집합에서의 어휘 분포와 특정 시간의 문서집합에서의 어휘 분포를 비교함으로써 계산하고, 그 특정 시간의 어휘의 가중치로 표현한다. 두 개의 언어는 하나의 언어에서보다 더 많은 정보를 줄 수 있기 때문에, 각 언어공간에서 어휘의 중요도를 측정하고, 다국어 처리에서 다른 언어 공간에서의 정보를 참조함으로써 언어 공간에서의 참조 역할을 하도록 한다. 본 논문의 실험에서는 같은 기간의 한국어와 일본어 신문기사에 대해서 사건 연결 탐색 성능을 평가하였다. 일반적인 가중치 기법인 tfidf 가중치 기법과의 비교 평가에서, 제안 방법이 단일언어 문서쌍에 대한 사건 연결 탐색은 14.3%, 다국어 문서쌍에 대한 사건 연결 탐색에서는 16.7%의 성능 향상을 보였다. 제안한 가중치 요소에 대한 유효성을 검증하기 위해, 공간 밀집도를 측정하였는데, 같은 사건을 나타내는 문서들의 그룹에서는 높은 밀집도를 나타냈고, 서로 다른 사건을 나타내는 문서들의 그룹에서는 낮은 밀집도를 나타냈다. 이 결과를 통해서 시간과 공간 요소를 반영한 사건 어휘 가중치 방법이 단일언어 사건 연결 탐색뿐만 아니라 다국어 사건 연결 탐색에 효과적이라고 볼 수 있다.

Multilingual Story Link Detection based on Properties of Event Terms

Kyung-Soon Lee†

ABSTRACT

In this paper, we propose a novel approach which models multilingual story link detection by adapting the features such as timelines and multilingual spaces as weighting components to give distinctive weights to terms related to events. On timelines term significance is calculated by comparing term distribution of the documents on that day with that on the total document collection reported, and used to represent the document vectors on that day. Since two languages can provide more information than one language, term significance is measured on each language space and used to refer the other language space as a bridge on multilingual spaces. Evaluating the method on Korean and Japanese news articles, our method achieved 14.3% and 16.7% improvement for mono- and multi-lingual story pairs, and for multilingual story pairs, respectively. By measuring the space density, the proposed weighting components are verified with a high density of the intra-event stories and a low density of the inter-events stories. This result indicates that the proposed method is helpful for multilingual story link detection.

키워드 : 사건 연결 탐색(topic link detection), 공간 밀도(space density), 사건 어휘(event term), 사건 탐색 및 추적(topic detection and tracking), 어휘 분포(term distribution)

1. 서 론

사건 탐색 및 추적(TDT : Topic Detection and Tracking) [1] 연구는 전세계 각 나라에서 매일 보도되고 있는 신문이나 방송 뉴스 기사에서 "어떤 사건이 발생했는가?" 또는 "새로운 사건이 일어났는가?"에 대한 정보를 탐색, 추적하는 것이다. TDT의 세부 연구에는 (1) 처음으로 일어난 사건을 탐색 (new event detection), (2) 두 뉴스가 같은 사건을 다루는지를 탐색 (story link detection), (3) 같은 사건을 다루는 기사들을 탐색 (topic detection), (4) 주어진 사건에 대해서 관련된 사건을 추적(topic tractking)하는 연구가 있다.

사건 연결 탐색 (story link detection)은 임의로 선택된 두 문서가 같은 사건을 다루는지 아닌지를 결정하는 것이다. 사건 연결 탐색은 사건 탐색 (topic detection)과 사건 추적 (topic tracking)에서의 핵심 요소 기술이다. 예를 들어, 사건의 탐색은 계속 들어오는 문서에 대해서 이미 어떤 사건으로 탐색된 문서들과 연결이 되는지를 비교함으로써 처리할 수 있다.

* 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2004-003-D00325).
† 정 회 원 : 전북대학교 전자정보공학부 교수
논문접수 : 2004년 10월 15일, 심사완료 : 2005년 1월 18일

사건 연결 탐색에 관한 최근의 대부분의 연구에서는 같은 사건을 나타내는 문서들에 대한 결정을 내리는데 있어서 보다 정확하게 하기 위해 문서 표현, 결정 임계치, 유사도 측정 방법 등에 중점을 두고 있다. 사건 탐색 및 추적 연구에서 대부분의 접근 방법은 기존의 문서 내용 중심적 문제인 문서클러스터링(document clustering)과 문서범주화(text categorization)에 대한 접근 방법과 별로 다르지 않다. 연구 [2]는 TDT2002 평가대회에서 사건 연결 탐색을 위해서 두 뉴스기사가 같은 사건을 다루는지의 유사도 측정을 하는데 있어서, 20여 가지의 유사도 측정 기법을 적용하여 계산하고, 그 유사도값들을 조합하여 사건 연결 탐색을 수행하였다. 연구 [3]은 뉴스기사를 표현하기위해 명사, 동사, 형용사, 복합명사 등을 추출하였고, 문서의 길이에 따른 유사도값의 차이를 줄이기 위해서, 문서 길이를 확장하는 방법을 이용하였다. 단일언어나 다국어에서 사건 연결 탐색을 위해서는 임계치에 차이를 둔 정도이다.

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

다국어 사건 탐색 및 추적(multilingual topic detection and tracking) 연구에서는 아랍어 뉴스와 영어 뉴스에 대한 사건 연결 탐색을 위해서 아랍어-영어 사전과 번역 확률을 이용한 연구가 있다[8]. 아랍어, 중국어, 영어에 대한 사건 추적에서 연구 [9]와 [10]에서는 통계사전에 기반하여 두개의 가장 좋은 대역어를 선택하고 번역 후 문서확장을 하는 방법이 기계번역시스템에 의한 하나의 대역어를 선택하는 것보다 더 좋은 성능을 보였다. 다국어 사건 탐색 기법은 사전을 이용하여 번역하거나 기계번역기를 이용하여 번역하는 등 언어 번역 과정을 거친다. 언어번역을 한 후에는 대부분 언어 중심적인 정규화 과정에 중점을 두고 있다[3, 8, 11].

본 논문에서는 다국어 사건 연결 탐색을 위해서 뉴스기사에 나타나는 어휘를 사건의 관점에서 중요도를 측정하기 위해 시간요소와 다국어 공간요소를 어휘의 가중치 측정 요소로 반영하는 방법을 제안한다. 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. 여기서 '사건 어휘'는 뉴스기사에서 다루는 사건의 핵심 역할을 하는 주요 어휘를 지칭한다. 또한, 사건은 두개 이상의 어휘로 표현되므로(예를 들어, '김일성 사망', '고베 지진', '김선일씨 피살' 등), 각 문서에서 자주 같이 발생하는 이웃 어휘들의 공기 빈도수(co-occurrence frequency)를 반영하여 가중치에 서로 영향을 주도록 하였다. 어휘 가중치로 표현된 문서벡터들에 대해서 유사도 측정을 해서 같은 사건을 다루는지를 결정하였다. 본 연구는 다음과 같은 관찰/가정에 기반해서 다국어 사건 연결 탐색에 접근하고 있다 :

- 사건 어휘 그 자체의 특성 : 사건은 "누가 언제 어디서 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 어휘의 문서에서의 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 시간동안의 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.

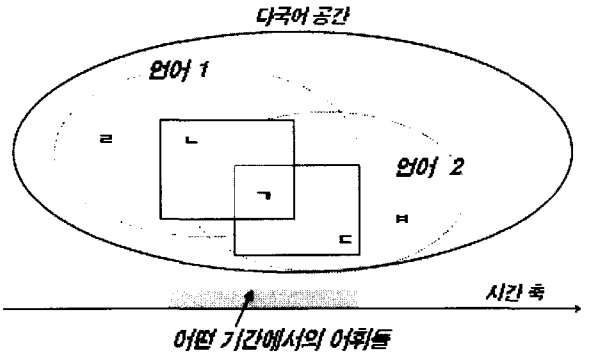
본 연구의 실험에서는 한국어와 일본어 뉴스 기사로 구성된 다국어 사건 연결 탐색 실험집합에 대해서 평가하였다.

2. 시간과 언어공간에서의 어휘 가중치를 이용한 다국어 사건 연결 탐색

본 논문에서는 사건을 기술하는 어휘들에 대해 변별적인 가중치를 부여하기 위해서 시간과 다국어 공간에서의 어휘 분포를 가중치계산의 한 요소로서 이용하였다.

(그림 1)에 나타난 것과 같이, 어휘의 중요도는 어떤 기간동안(그날 또는 몇일)의 문서들(영역 ㄴ, 영역 ㄷ)과 전체 문서 집합 (영역 ㄹ, 영역 ㅂ)에서의 문서들에서의 어휘의 분포 비율을 측정 (영역 ㄴ과 영역 ㄹ에서의 어휘 분포 비율 측정, 또는 영역 ㄷ과 영역 ㅂ에서의 어휘 분포 비율 측정)한다. 측정된 어휘의 중요도는 해당기간 동안의 어휘의

가중치로 하여 문서들을 표현한다. 각 언어 공간마다 고유한 어휘들이 있고 분포도 다르므로, 두개 이상의 언어공간에서는 하나의 언어공간에서 보다 더 많은 정보를 제공할 수 있다. 따라서 언어 공간에 대해서 어휘의 중요도는 각 언어 공간에서 측정을 하고, 다국어공간(영역ㄱ = 영역ㄴ ∩ 영역ㄷ)에서는 다른 언어공간에서의 정보를 참조하도록 한다. 영역ㄱ에 나타난 어휘들은 다국어 공간에서 공통으로 나타나는 어휘로써 일반 어휘이거나 사건 어휘일 수 있다.



(그림 1) 시간과 다국어 공간에서의 어휘

다국어 사건 연결 탐색을 하기 위해서, 기계번역기를 이용하여 다국어 언어 공간을 하나의 공간으로 변환하였다. 언어 1에서 언어 2 또는 그 반대로, 언어 변환을 통해서 하나의 언어로 표현된 다국어 공간으로 표현하고, 각 어휘는 사건 어휘 특성에 기반해서 가중치를 계산한다. 각 문서는 어휘들의 공기관계를 측정하여 자주 같이 나타나는 어휘들의 가중치에 상호 영향을 주도록 하였다. 사건 어휘와 어휘들의 관계를 이용하여 가중치를 부여한 문서들에 대해서 두 문서의 관련도를 측정하여 같은 사건을 다루는지를 결정하였다.

2.1 다국어 사건 탐색을 위한 언어 번역

본 논문에서는 한국어와 일본어 뉴스 기사에 대해서 다룬다. 다국어 문서에 대해서 같은 사건을 다루는지를 탐색하기 위해서는 같은 언어 공간으로 변환을 해야 한다. 한국어와 일본어 뉴스 기사의 언어 공간을 하나로 하기 위해, 한국어-일본어 문서 번역기[17]를 이용하여 한국어를 일본어로 변환하였다.

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

2.2 시간상에서 어휘의 분포 변화에 따른 중요도 계산

각 뉴스기사 문서를 표현하기 위해 문서에 나타나는 어휘들에 대해서 품사 태깅을 거쳐서 명사, 고유명사, 형용사와 동사를 선택하였다. 또한, 사건을 구성하는 주요 개체를 인식하기 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, 문장에 나타난 모든 어휘들의 가능한 조합(n-gram)으로 추출하였다. 예를 들어, "북한 김일성 주석 사망"의 문장에 대해서 가능한 어휘 표현은 다음과 같다 : '북한', '김일성', '주석', '사망', '북한_김일성', '김일성_주석', '주석_사망', '북한_김일성_주석_사망', '김일성_주석_사망', '북한_김일성_주석' 등이 가능하다.

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

카이제곱은 두 사건의 독립성 여부를 판단하는 통계적 방식[16]으로, 문서 범주화에서 각 범주를 대표하는 중요한 어휘(자질)를 추출하는데 많이 이용하고 있다[15]. 문서범주화에서의 범주는 문서의 '주제'로 구분이 되는데 비해, 본 논문에서의 범주는 뉴스가 쓰여진 날짜에 해당하는 '시간' 범주로 보고, 어휘 t와 시간범주 $t0$의 독립 정도를 측정하였다. 두 사건이 독립적이라면 그 어휘는 시간상에서 중요한 영향을 미치지 않는다고 판단한다. x^2 값이 클수록 두 사건이 관련있다는 판단의 오류가 적어지므로, x^2 값이 큰 어휘를 시간범주에서 주요 사건을 나타내는 어휘로 판단하고, 그 어휘의 가중치를 높여준다.

〈표 1〉 시간상에서 중요어휘를 계산하기 위한 분할표

	어휘 t를 포함 (t∈doc)	어휘 t를 포함하지 않음 (t∉doc)	
시간 t0에 속하는 문서 (doc∈t0)	a	b	a+b
시간 t0에 속하지 않는 문서(doc∉t0)	c	d	c+d
	a+c : 문서빈도수	b+d	a+b+c+d : 전체문서수

<표 1>은 2 × 2로 된 분할표인데, t0는 시간상에서 하루에 해당하는 '그날'이고, 해당하는 그날에 보도된 뉴스기사들로 이뤄진 범주이다. 즉, '그날의 모든 뉴스 기사들'이 하

〈표 2〉 특정 시간(1994.7.10)에서 어휘들의 카이제곱 값 예.

시간 t0	언어 공간	어휘	x2	t0에서 문서빈도수 a	문서 빈도수(df) a+c	t0에서 전체문서수 a+b	1년동안 전체문서수 a+b+c+d
1994년 7월 10일	한국어 뉴스	김일성	746	39	793	59	24501
		사망	661	40	933		
		권력승계	578	13	117		
		긴장고조	147	2	11		
		노조	0.63	2	480		
		협력	0.45	2	520		
	일본어 뉴스	金日成(김일성)	1937	25	139	220	96865
		死去(사거)	358	56	3057		
		北朝鮮(북조선)	817	63	1897		
		會談再開(회담재개)	192	2	9		
		討議(토의)	13	4	341		
		雇用(고용)	6.7	6	962		

나의 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 :

$$\chi^2(t,t0) = \frac{N \cdot (ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (1)$$

여기서, a는 시간 범주 $t0$에 속하는 문서들 중에서 어휘 t를 포함하는 문서의 개수를 나타낸다. b는 시간 범주 $t0$의 문서들에서 어휘 t를 포함하지 않는 문서의 개수를 나타낸다. 그러므로 a와 b를 더한 값은 $t0$ 시간(그날 하루) 동안 보도된 문서의 개수가 된다. c는 현재 보이는 모든 시간상에서의 문서집합에서 시간 범주인 $t0$를 제외한 모든 시간동안의 문서들에서 어휘 t를 포함하는 문서의 개수를 나타낸다. 즉, a는 시간범주 t0 내에서의 문서빈도수(document frequency)이고, a와 c를 더한 값 (a+c)은 전체 문서집합에서의 어휘 t의 문서빈도수이다.

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

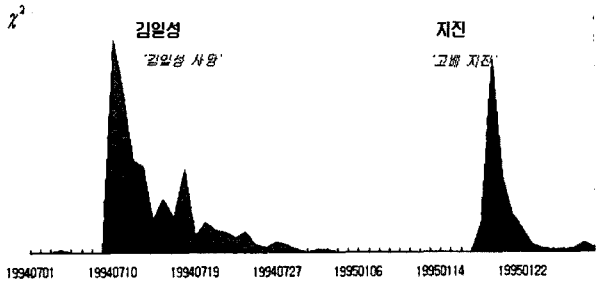
<표 2>는 통계 계산을 위한 문서인 1년동안(1994.7. -1995.6)의 문서집합에 대해서 '김일성 사망' 사건이 발생한 1994년 7월10일을 $t0$로 했을 때의 각 어휘들에 대한 카이제곱 값을 나타낸 것이다. 사건 관련 어휘인 '김일성', '사망' 등은 전체문서에서의 문서빈도수에 비해서 사건이 발생한 그날의 뉴스에서 여러 문서에 걸쳐 보도되고 있어 (한국어 뉴스에서 '김일성' : 39개 문서, '사망' : 40개 문서), 어휘들의 카이제곱 값이 높은 것을 볼 수 있다. 또한 '긴장고조', '노조' 등은 시간 $t0$ 동안 포함하고 있는 문서의 개수는 2로 같지만, 전체 문서에서의 문서빈도수가 달라서 카이제곱 값이 달라짐을 알수 있다. 실험문서집합의 특성차이로 한국어 문서집합의 크기는 작고, 일본어 문서집합의 크기는 크다. 한국어 뉴스는 한겨레신문의 지면상 보도된 기사의 일부이고, 일본어 뉴스기사는 마이니치 신문의 지면상으로 보도된 전체기사를 포함하고 있다.

시간에서의 그 어휘의 중요도는 다음과 같이 각 언어에 대해서 어휘의 카이제곱 값으로 한다.

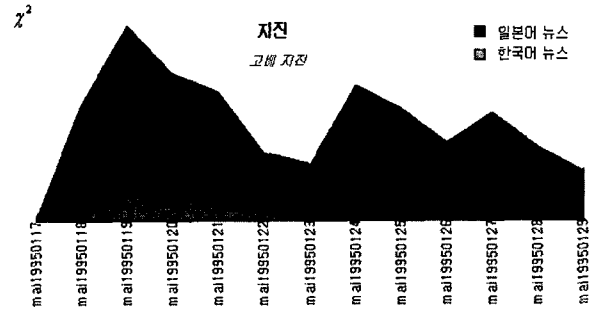
$$wTime\ (t,t0,l) = \chi^2(t,t0) \quad (2)$$

여기서, t는 어휘를 나타내고, $t0$는 '그날'에 해당하는 시간의 범주를 나타내고, l은 언어공간을 나타낸다. 그림 2는 한국어 뉴스기사에 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 기간 동안 많은 분포를 나타내서 높은 중요도를 갖게 되었다. 이와같이 사건과 관련된 어휘는 사건이 발생한 시점에서 어느 정도 시간이 흐르면서 어휘의 분포에 따른 중요도에서도 낮은 값을 갖고 있음을 알 수 있다.



(그림 2) 시간상에서 어휘의 중요도 변화



(그림 3) 다국어 공간(한국어, 일본어 공간)에서 어휘 분포 비교

2.3 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.

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

〈표 3〉 다국어공간에서 어휘의 날짜 빈도수

어 휘	한국어	일본어	영 어
인 천	261	31	3
히로시마	61	307	25
NHK	32	292	4
플로리다	17	48	249

<표 3>은 1년 동안의 뉴스기사에서 각 어휘가 나타난 날짜의 빈도수 (date frequency)를 나타낸 것이다. 1년 동안의 전체 뉴스기사에서 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. 이와 같이 그 나라와 관련이 있는 어휘들은 그 나라 뉴스의 언어 공간에서 자주 나타나는 경향을 보인다. 같은 사건을 보도하는 문서의 양에 있어서도 사건에 대한 그 나라의 관심의 정도에 따라 의존한다.

(그림 3)은 '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.

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

두개 이상의 언어 공간은 하나의 언어 공간에서 보다 더 많은 정보를 제공할 수 있기 때문에, 본 논문에서는 다국어 언어 공간을 합쳐서 어휘의 분포를 계산하지 않고, 서로 다른 언어공간에서 어휘의 분포를 각각 측정하여 중요도로 계산(*wTime(t,t0,l1), wTime(t,t0,l2), wTime(t,t0, l3), ...*) 하고, 같은 시간대에서의 다른 언어 공간에서의 가장 높은 값을 반영한다.

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

2.4 어휘의 공기관계를 반영한 가중치 계산

각 문서는 어휘들의 공기 관계를 반영하기 위해서, 어휘들의 노드들과 어휘들 사이의 공기 관계를 나타내는 간선으로 하는 공기관계 정보를 표현하였다. 공기관계에서 높은 가중치를 갖는 어휘와 높은 빈도로 같이 발생하는 어휘들은 서로에게 영향을 주도록 하여, 문서가 다루는 사건을 기술하는 사건 어휘들에 변별력을 높여주기 위한 것이다.

각 노드는 사건 어휘의 시간과 다국어공간에서의 특성에 기반하여 다음과 같이 계산한다 :

$$wnode_i = tf_i \cdot wNE_i \cdot wTimeSpace_i \quad (4)$$

여기서, $wnode_i$는 어휘 i의 가중치를 나타내는 것으로, tf_i는 어휘 빈도수, wNE_i (wNE_i = 1 또는 wNE_i = 2)는 사건의 요소에 해당하는 개체인가에 따라, <사람>, <지역>, <국가> 등에 개체 인식 단계에서 인식한 개체인 것에 대해 가중치 2의 값을 부여해서, 일반 어휘 (디폴트=1) 보다 높은 가중치를 갖도록 했다. $wTimeSpace_i$는 수식 (3)에서 계산된 값으로, 시간 $t0$와 다국어 공간 l에서 어휘 i에 대한

중요도 값을 곱하여 어휘 중요도에 반영한다.

간선은 하나의 문장에서 같이 발생하는 노드들을 표현하기 위한 것으로, 공기 관계를 측정할 때의 거리는 5로 설정하였다. 각 간선은 두 노드의 가중치와 공기 빈도수를 이용하여 다음과 같이 계산하였다 :

$$wedge_{ij}=cooc_{ij} \cdot wnode_i \cdot wnode_j \quad (5)$$

여기서, 간선의 가중치 $wedge_{ij}$는 서로 이웃하는 노드 $wnode_i$와 $wnode_j$의 가중치에 영향을 받는다. 또한 그들의 공기 빈도수 $cooc_{ij}$에 비례적으로 영향을 받는다. 사건을 나타내는 핵심 어휘들은 문서 전체에 걸쳐서 같이 발생할 것이라는 가정에 근거하여 이와 같이 계산을 했다.

최종적으로 어휘들 사이의 공기 관계를 표현한 간선의 가중치 $wedge_{ik}$를 노드의 문맥으로 반영하여, 노드의 가중치는 다음과 같이 계산된다 :

$$wnode_i' = wnode_i \cdot \alpha \sum_k wedge_{ik} \quad (6)$$

각 문서는 수식(6)에서 계산된 어휘의 가중치 벡터로 표현을 한다. 사건 연결 탐색에서 두 문서가 같은 사건을 다루는지를 측정하기 위해서 두 문서벡터에 대한 코사인계수를 계산한다. 유사도에 대한 임계치에 따라 같은 사건 또는 다른 사건을 다룬다고 판단을 한다.

3. 실험 및 평가

시간 및 다국어 공간에서 어휘의 분포를 이용하여 가중치를 계산하여 사건 연결 탐색 방법이 유효한지를 보기 위해, 한국어 뉴스기사와 일본어 뉴스기사로 구성된 다국어 테스트 컬렉션을 이용하여 평가를 하였다.

3.1 실험 환경 설정

문서 집합은 한국어와 일본어 신문기사로 구성되어 있는데, 한국어는 인터넷에 보도된 뉴스기사를 수집한 것이고, 일본어는 마이니치 신문기사이다. 문서의 날짜는 1998년 1월에서 1998년 6월까지 보도된 것으로, 문서의 개수는 한국어는 40,000개, 일본어는 61,637개이다. 1998년 1월의 사건 탐색을 위해서는 그 이전의 뉴스가 필요하다. 따라서 1994년 7월에서 1995년 6월까지 보도된 뉴스를 통계 정보를 위해 이용하였다.

각 문서의 어휘들은 일본어 품사 태거 시스템인 차센(ChaSen)[12]을 이용하여 추출하였다. 한국어 문서공간에서 나타난 어휘는 193,730개이고, 일본어 문서공간에서 나타난 어휘는 353,210개였다.

매일 보도되는 뉴스 기사에 대한 사건 탐색이기 때문에, 그날 뉴스기사가 추가될 때마다, 가중치 계산에서 사용되는 문서 빈도수는 점진적으로 계산하는 점진적 문서 빈도수 (incremental document frequency)를 적용하였다. 개체인식을 위해서는 일본어 품사태거인 차센 시스템의 결과를 이용하는 개체인식 시스템인 NExT[13]을 이용하였다.

본 실험의 사건 탐색에서 다룬 사건은 <표 4>에 나타난 13개로 구성되어 있는데, 이는 TDT2 테스트 컬렉션에 포함된 사건의 일부이다. 같은 시기에 한국어, 일본어, 영어로 보도된 뉴스기사에 대한 다국어 사건 탐색을 위해 이를 이용한 것인데, 현재 본 논문에서는 한국어와 일본어에 대해서만 실험을 한 것이다. 사건은 <표 3>에 나타난 것으로 국제적인 사건에 해당되는 것들이다.

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

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

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

〈표 4〉 사건 리스트

1. Upcoming Philippine Elections
2. 1998 Winter Olympics
3. Current Conflict with Iraq
4. China Airlines Crash
5. Tornado in Florida
6. Asteroid Coming
7. Viagra Approval
8. India, A Nuclear Power
9. Israeli-Palestinian Talks(London)
10. Anti-Suharto Violence
11. Anti-Chinese Violence in Indonesia
12. Afghan Earthquake
13. Clinton-Jiang Debate

시스템의 성능 평가는 정확률, 재현률, 누락률, 오류률, 마이크로 평균 F1(micro-average F1)으로 측정하였다.

〈표 5〉 성능 평가를 위한 분할표.

	사람 정답평가 Yes	사람 정답평가 No
시스템 Yes	a	b
시스템 No	c	d

- 정확률 (Precision) : a/(a+b)
- 재현률 (Recall) : a/(a+c)
- 누락률 (Miss alarm) : c/(a+c)
- 오류률 (False alarm) : b/(b+d)
- F1 : 2×Precision×Recall/(Precision+Recall)= 2a/(2a+b+c).

TDT 연구에서 누락률과 오류률을 이용해서 사건탐색시스템이 제대로 찾지 못한 정답의 관점에서 성능 평가를 하고는 있으나, 이것이 정확률과 재현률 평가에 대해서 크게 다른 의미를 보여주고 있지는 못하다.

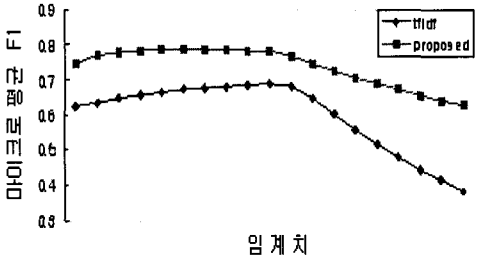
3.2 실험 결과

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.

제안한 가중치 기법에 의한 문서벡터에서는 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 연결 탐색에서 유사도 측정에서도 그 영향을 미치게 된다.

<표 6>은 단일언어 사건 연결 탐색의 실험 결과를 보여준다. 각 언어에 대해서 적용했을 때의 차이가 있는지를 비교하기 위해서, 한국어 뉴스 기사와 일본어 뉴스기사를 분리해서 같은 사건을 탐색하는 성능을 분석했다. 또한 한국어와 일본어 뉴스 기사를 모두 포함한 전체 뉴스 기사에 대한 한국어-한국어 문서쌍의 사건 연결 탐색, 한국어-일본어 사건 연결 탐색, 일본어-일본어 사건 연결 탐색 성능을 살펴보았다. 사건 연결 탐색에서 유사도에 대한 임계치를 0.005에서 0.35까지 변화시켜서 가장 좋은 성능을 보일 때의 결과이다. 본 논문에서 제안한 시간 및 다국어 공간에서 사건 어휘 분포를 고려한 가중치기법(*ewgt*)이 일반적 가중치계산기법(*tfidf*)에 비해 마이크로평균 F1에서 14.3% 성능 향상을 보였다.

(그림 4)는 사건 연결 탐색의 결정에서 임계치에 따른 성능의 변화를 나타낸다. 전체 임계치에 대해서 제안한 방법이 tfidf 방법을 능가함을 볼 수 있다.



(그림 4) 임계치에 따른 성능 변화

<표 7>은 다국어 공간에서의 어휘 분포 비교를 적용한 것이 다국어 사건 탐색에서 유용했는지를 보기 위한 것으로, 한국어-일본어 뉴스기사 쌍에 대해서 같은 사건을 다루는지를 탐색하는 실험을 하였다. 다국어 공간을 고려한 것 (수식 (3)을 적용)이 마이크로 평균 F1에서 0.766을 나타냈고, 다국어 공간을 고려하지 않은 것 (수식 (3)을 적용하지 않음)이 0.6719를 나타내서, 다국어 요소를 적용함으로써 14.1% 성능 향상을 보이고 있다. 이러한 결과를 통해서 다국어 공간에서 어휘 분포의 차이를 나타내는 정보를 참조하여 반영하는 것은 효과적이다고 할 수 있다.

〈표 6〉 사건 연결 탐색 성능 비교표 7

	단일 언어 사건 탐색				다국어 뉴스 기사 (한국어-한국어, 일본어-일본어, 한국어-일본어 사건 연결 탐색)	
	한국어 뉴스기사 (한국어-한국어 사건 연결 탐색)		일본어 뉴스기사 (일본어-일본어 사건 연결 탐색)			
	tfidf	ewgt	tfidf	ewgt	tfidf	ewgt
정 확 률	0.3865	0.4240	0.2899	0.3313	0.3025	0.3559
재 현 률	0.8506	0.9042	0.9808	0.9131	0.9657	0.8970
누 락 률	0.1494	0.0958	0.0192	0.0869	0.0343	0.1030
오 류 률	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 7〉 다국어 공간의 적용에 따른 교차언어 사건 연결 탐색 성능 비교

	한국어-일본어 교차언어 뉴스기사 쌍		
	tfidf	다국어공간 적용 않음	다국어공간 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 F1	0.6566	0.6719	0.7665

실험 결과를 통해서, 같은 사건을 다루는 뉴스기사를 탐색하기 위해서, 사건 뉴스기사에 나타나는 어휘의 시간 및 다국어 공간에서의 분포 특성을 이용하여 가중치를 계산하여 사건과 관련된 어휘의 가중치에 변별력을 줌으로써 사건 탐색에서의 관련도 계산에 영향을 주도록 한 것이 효과적임을 볼 수 있다.

3.3 실험 결과 검증

제안한 방법의 가중치 계산의 성능 결과를 검증하기 위해, 문서공간밀도(document space density)를 측정하였다. 색인 성능과 문서공간밀도 사이의 상호관계 분석에서, 연구 [14]는 클러스터 된 공간에서 각 클러스터 내부적(intra-cluster)으로는 밀집되어 있고, 클러스터들 사이의 거리(inter-cluster)는 먼 형태로 표현되어 있을 때 가장 좋은 검색 성능을 갖는 것을 보였다.

정보검색의 색인성능평가에서 이용된 공간밀도 측정방법은 사건 연결 탐색에서 제안한 방법이 같은 사건 속하는 문서들의 그룹인 클러스터 내부를 더 밀도가 높게 만들고, 서로 다른 사건그룹들 사이의 거리를 더 느슨하게 만들 수 있는지를 측정하는데 이용할 수 있다. 따라서 사건 클러스터 내부 공간밀도가 높고, 사건 클러스터 사이의 밀도가 낮도록 하는 방법은 보다 더 좋은 성능을 낼 수 있다고 볼 수 있다.

문서 m개로 구성된 사건 클러스터가 K개 주어졌을 때, 클러스터의 중심 벡터 C_k는 클러스터에 속하는 문서들의 가중치의 평균으로 정의하고, 전체문서집합에 대한 전체 중심벡터 C_{main}은 전체 N개의 문서의 가중치의 평균으로 정의한다.

$$C_K = \frac{1}{M}\sum_{i \in E_k} d_i \tag{7}$$

$$C_{main} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{8}$$

각 클러스터의 중심벡터와 전체문서집합의 중심벡터를 문서공간밀도를 다음과 같이 측정한다. 공간밀도 계산에서 유사도 계산을 하는데 코사인 계수 측정을 이용하였다.

- 사건 클러스터 내부의 밀도(Intra-event density : DensityIntraC) : 하나의 클러스터내에서 클러스터 중심 C_k와 클러스터에 속하는 각 문서 d_i와 유사도를 계산해서 평균한다. 공간밀도 비율 계산에서 요소 x로 한다.

$$DensityIntraC = \frac{1}{N}\sum_{k=1}^{K}\sum_{i \in E_k} d_i \cdot C_k \tag{9}$$

- 사건 클러스터 사이의 밀도(Inter-event density : DensityInterByCmain) : 전체 중심벡터를 이용하여 계산. 전체중심벡터 Cmain과 각 클러스터 중심 C_k와의 유사도를 평균한다.

$$DensityInt\ erByCmain = \frac{1}{K}\sum_{k=1}^{K} C_{main} \cdot C_k \tag{10}$$

- 사건 클러스터 사이의 밀도(Inter-event density : *DensiyInterC*) : 클러스터 중심들의 쌍(C_i와 C_j)에 대해서 유사도를 계산해서 평균한다. 공간밀도 비율 계산에서 요소 y로 한다.

$$DensityInt\ erC = \frac{1}{K(K-1)}\sum_{i=1}^{K}\sum_{\substack{j=1 \\ i \neq j}}^{K} C_i \cdot C_j \tag{11}$$

- 공간밀도 비율(Space density ratio) : 전체 문서공간 밀도를 측정. y/x로 계산한다.

〈표 8〉 가중치 기법에 따른 클러스터 내부와 클러스터 사이의 공간 밀도 변화 비교

	클러스터내부 밀도 (x)		클러스터사이의 밀도 : 전체문서 중심 이용		클러스터사이의 밀도 : 클러스터 중심 이용 (y)		공간밀도 비율 (y/x)	
	tfidf	ewgt	tfidf	ewgt	tfidf	ewgt	tfidf	ewgt
한국어 문서	0.371	0.551	0.302	0.245	0.060	0.046	0.06/0.371 = 0.161	0.046/0.551 = 0.084 -47.73%
일본어 문서	0.303	0.481	0.302	0.262	0.076	0.050	0.076/0.303 = 0.251	0.050/0.481 = 0.105 -58.30%
다국어 문서	0.229	0.482	0.312	0.264	0.077	0.049	0.077/0.22 = 0.258	0.049/0.482 = 0.101 -60.87%

본 연구에서 제안한 가중치 기법이 한국어, 일본어, 다국어 문서에서의 모든 밀도 측정에서 tfidf 가중치기법 보다 낮은 값을 갖고, 클러스터들 사이의 거리는 각 클러스터 내부의 문서들 사이의 거리보다 더 큰 값을 갖는다. <표 8>에 나타난 것과 같이, 제안 가중치 기법은 사건클러스터내부의 밀도를 최대화 시키고, 사건 클러스터 사이의 밀도를 최소화시킴으로써 공간밀도를 줄였다. 따라서 제안된 가중치 기법에 의한 <표 6>과 <표 7>에서의 성능 향상은 문서 공간에서 감소된 밀도와 연관이 있음을 보여준다고 할 수 있다.

4. 결 론

본 논문에서는 다국어 뉴스기사에 대해서 시간 및 다국어 공간에서의 어휘 분포 특성을 이용하여 가중치를 적용한 방법이 한국어와 일본어 뉴스 기사에 대한 다국어 사건 연결 탐색에서 효과적임을 보았다. 이러한 결과는 뉴스기사에서 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 뉴스기사에 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/방송사 마다 같은 사건에 대한 보도 관점이 다른 경우에 나타난다. 앞으로 계속 연구가 필요한 부분으로, 여러 나라에서 보도된 다국어 뉴스기사에서 탐색된 어떤 사건에서, 그 문서들 관점을 서로 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.

본 논문에서는 단순히 tfidf 가중치 기법과 제안 방법의 성능을 비교하였는데, 향후 연구로 은닉 변수 모델(latent variable model)과 은닉 의미 커널(latent semantic kernel)과 같은 방식과의 비교 분석을 통해서 다국어 뉴스에서의 사건 어휘 추출에 대한 모델 개선이 필요하다.

참 고 문 헌

[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(In Japanese). http : //www.ai.info.mie-u. ac.jp/~next/

[14] Salton, G., Wong, A. and Yang, C.S. 1975. A vector

space model for automatic indexing. Communications of the ACM, 18(11).

[15] Yang, Y., Pedersen J.P. 1997. A Comparative Study on Feature Selection in Text Categorization Proceedings of the Fourteenth International Conference on Machine Learning(ICML'97).

[16] Devore, J. L. 1995. Probability and Statistics for Engineering and the Sciences. Morgan Kaufmann Publishers, Inc., 4th edition.

[17] ChangshinSoft. 2001. ezTrans Korean-to-Japanese/ Japanese-to-Korean machine translation system.

이 경 순

e-mail : selfsolee@chonbuk.ac.kr

1994년 계명대학교 컴퓨터공학과(학사)

1997년 한국과학기술원 전자전산학 (공학석사)

2001년 한국과학기술원 전자전산학 (공학박사)

2001년~2003년 일본 국립정보학연구소(National Institute of Informatics) 연구원

2004년~현재 전북대학교 전자정보공학부 전임강사

관심분야 : 정보검색, 지식 마이닝, 자연언어처리