개념 기반 문서 분류를 위한 단어 애매성 해소

강원석∪ 황도삼
안동대학교 컴퓨터교육과, 영남대학교 컴퓨터공학과 & AITrc
wskang@andong.ac.kr, dshwang@ynucc.yeungnam.ac.kr

Word Ambiguity Resolution for Concept-based Text Classification

Wonseog Kang∪
Dept. of Computer Education, Andong National University
DoSam Hwang
Dept. of Computer Engineering, YeungNam University and Advanced Information Technology Research Center(AITrc)

요 약

문서 분류 시스템은 문서에 나타난 용어나 개념의 출현 정보를 이용한다. 개념 기반 문서 분류는 용어를 사용하지 않고 문서의 단어에 나타난 의미를 이용한다. 단어가 중의성을 가지는 경우 그 뜻을 정확히 가리지 않으면 문서에 출현하지 않은 의미를 이용하게 되므로 문서 분류 시스템의 성능이 저하된다. 본 논문은 개념 기반 문서분류를 위하여 단어 애매성 해소를 시도하였다. 문서에 출현된 의미 정보를 이용하여 의미들간의 공기정보를 구하고 이를 이용하여 단어의 애매성을 해소하였다. 단어의 의미정보는 시소러스 도구를 통해 획득하고 의미들간의 공기정보는 의미들간의 동시 출현 정보를 획득하여 구축하였다. 본 시스템은 문서 분류 등 자연어처리 분야에 이용할 수 있어 효용가치가 높다.

1. 서론

문서 분류 시스템은 문서에 나타난 용어나 개념의 출현 정보를 이용한다. 문서의 주제는 문서에 나타난 용어보다 의미에 더 의존하므로 개념 기반 문서 분류는 용어를 사용하지 않고 문서의 단어에 나타난 의미를 이용한다. 그런데 단어의 애매성을 해소하지 않으면 여러 의미 가운데 한 의미가 사용됨에도 불구하고 그 단어의 모든 의미가 사용되는 것으로 다루게 되므로 문서 분류 시스템의 성능이 저하된다. 본 논문은 개념 기반 문서 분류를 위하여 단어 애매성 해소를 시도하였다.

단어의 의미는 주위에 출현된 단어들에 영향을 받는다. 이와 같은 사실을 근거로 단어 애매성 해소를 위해 단어간 공기정보를 이용한 연구가 있다[1,2]. 그렇지만 단어의 의미는 함께 출현된 단어보다 단어가 가지고 있는 의미에 더 좌우된다. 본 논문은 이 사실을 토대로 의미들간의 공기정보를 이용하여 단어의 애매성을 해소하고자 하였다.

* 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

[3,4]는 의미들간의 공기정보를 이용하여 단어 애매성 해소를 하는 점에서 본 연구와 비슷하다. 그러나 본 논문은 의미를 cluster로 정의하지 않고 체계적인 트리 형태의 의미체계를 이용한다. 이 의미들은 [7]에서 개발한 시소러스 도구를 이용하여 획득한다.

본 논문의 시스템은 형태소 해석기, 시소러스 도구, 공기정보 획득기, 애매성 해소기로 구성된다. 시소러스 도구는 단어의 의미를 획득하는 도구로 의미들간의 공기 정보를 구할 수 있는 토대를 마련해준다. 공기정보 획득기는 의미들간의 공기 관계의 데이터를 획득하는 시스템이다. 그리고 애매성 해소기는 구축된 공기 정보를 이용하여 단어의 애매성을 해소하는 시스템이다. 본 논문의 2장은 정의한 의미 체계와 의미 획득도구, 의미 공기정보를 기술하고 3장은 애매성 해소 시스템을 기술한다. 그리고 4장에서 결론을 짓는다.

2. 의미 공기 정보

2.1 의미

본 시스템에 사용하는 의미는 [7]에서 사용한 의미 체계를 이용하였다. 의미 수는 총 768개로 상하위 관계로 정의되어 있다. 문서 분류에서는 명사를 추출하고 이를

이용하고 있어 본 시스템의 의미도 명사류의 의미체계에 따른다. 최상위 의미는 물체, 사건, 속성의 세 갈래로 구성된다.

시소러스 도구는 단어가 가지고 있는 의미를 찾아주는 도구이다. 단어의 의미를 잘 정의하더라도 의미를 찾아주는 사전이나 시소러스 도구가 없다면 의미를 이용할 수가 없다. 본 논문에서 사용한 시소러스 도구는 [7]에서 사용한 시소러스 도구로 4000여 단어에 대해 의미를 획득할 수 있는 것이다.

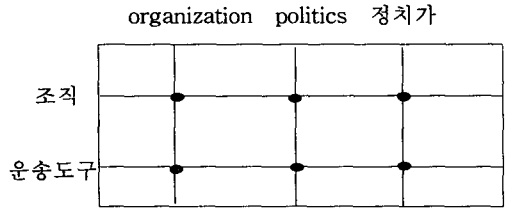
2.2 의미 공기정보

공기정보를 정의하기 위해서는 현 단어의 의미에서부터 어느 거리까지의 단어의 의미를 고려할 것인지를 결정해야 한다. 단어의 의미는 또한 문장 전체의 단어들의 의미에 의해 영향을 받을 수 있고 문서 전체의 단어들의 의미에 의해 영향받을 수 있다. 따라서 본 논문에서는 좌우 일정거리 내의 단어의 의미 이외에 문장의 단어들의 의미, 문서의 단어들의 의미에 관한 공기 정보도 정의한다.

의미의 수는 768개이므로 의미 공기 정보는 768*768의 2차원 벡터로 표현된다. 행으로는 현 단어의 의미를 표현하고 열로는 일정거리 내에 출현한 단어의 의미나 문장의 의미, 문서의 의미를 나타낸다. 한 단어에 대해 예를 들어보자.

<21세기 위원회>
대통령 직속의 기구. 정치성, 관료성을 완전히 배제한 위원회로 ....

위 문서에서 "기구"라는 단어는 "운송도구"와 "조직"이라는 의미를 가지고 있다. 공기정보를 구축할 시 2차원 벡터의 "운송도구" 의미 행과 "조직"의 의미 행, 열로는 "대통령", "직속"의 단어가 거리 N 내에 있으므로 "대통령"과 "직속"의 단어의 의미들을 획득하고 각 의미들에 대해 열 위치를 결정하여 공기 정보를 구축한다.

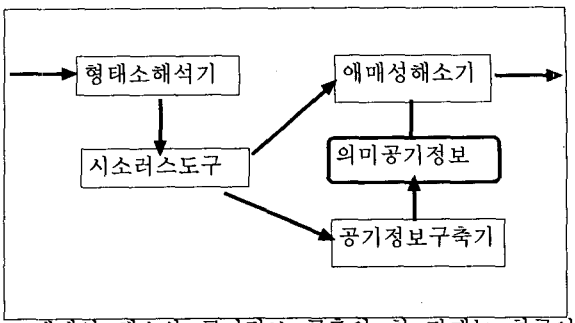


위의 예에서 운송도구에 대해서는 바르지 않은 공기정보가 구축된다. 그러나 공기정보 구축의 대상이 아주 대량의 코퍼스인 경우 보편화된 의미들의 공기관계가 구축된다. 만약 코퍼스가 크기가 작거나 크더라도 다양하고 일반적이지 않는 것이라면 올바르지 못한 의미 공기정보가 구축될 것이다. 따라서 이 방법은 일반적이고 다양한 대량의 코퍼스를 통한 의미공기정보 구축이 필수적이다.

3. 애매성 해소 시스템

3.1 시스템 개요

본 시스템은 형태소해석기, 시소러스 도구, 공기정보 구축기, 애매성 해소기로 구성된다.



애매성 해소와 공기정보 구축의 첫 단계는 한국어 형태소 해석이다. 형태소 해석은 접사 처리, 복합어 처리 등이 고려된 [7]의 한국어 형태소 해석기를 사용하였다. 이 형태소 해석기는 문서 분류에 중요한 역할을 하는 명사의 분석에 초점을 맞추었고 순수 명사 이외에 조용사 '하다'가 붙는 용언의 분석도 포함하였다.

형태소 해석 다음으로 단어의 의미를 구하기 위하여 시소러스 도구를 사용한다. 그 결과 단어가 가질 수 있는 모든 의미가 나타난다. 시소러스 도구 개발의 어려움으로 많은 단어들이 포함되지 못한 문제점이 있다.

공기정보 구축기는 단어 애매성 해소기를 수행하기 전에 코퍼스의 문서를 분석하여 의미들간의 공기 관계 데이터를 획득하는 시스템이다. 본 시스템에서는 4 종류의 공기관계 정보를 구축하였다. 현 단어의 의미와 단어의 앞으로 거리 N 내의 단어들의 의미와의 공기정보, 현 단어의 의미와 단어의 뒤로 거리 N 내의 단어들의 의미와의 공기정보, 현 단어의 의미와 문장내의 단어들의 의미와의 공기정보, 현 단어의 의미와 문서 내의 단어들의 의미와의 공기정보이다.

3.2 단어 애매성 해소기

단어 애매성 해소기는 구축된 공기정보를 이용하여 입력 문서의 각 단어에 대한 의미를 선택한다. 그 과정을 예문서 "21세기 위원회"의 "기구"라는 단어에 대해 보자.

(1) "기구"라는 단어는 "운송도구"와 "조직"의 두 의미를 지닌다. 두 의미 중 어느 것인지를 찾기 위해 "기구" 단어 주변 N 거리내의 단어들의 의미들을 찾아 벡터로 만든다.
입력단어 공기정보벡터 $IC = (wic_1, wic_2, \ldots, wic_n)$
여기서 wic_i는 "기구" 단어 주변 N 거리내의 단어들에 나타난 의미 i의 확률 값을 나타낸다. 따라서, wic_i , $i=1,2,\ldots,n$는 0과 1 사이의 값이며, $\Sigma_i\, wic_i$ = 1이 된다. 그리고 n은 정의된 의미들의 수 768이다.

(2) 입력 문서의 공기정보벡터와 구축된 공기정보 벡터와의 유사도를 측정하여 가장 유사한 것을 선택한다. 구축된 공기정보 벡터와 유사도 비교함수는 다음과 같다.
구축된 공기정보 벡터 $NC_i = (nc_{i1}, nc_{i1}, \ldots, nc_{in})$

여기서 NC_i는 의미 i와 거리 N내에 함께 출현되는 의미 벡터이다. nc_{ij}는 의미 i와 거리 N 내에 의미 j가 출현될 확률을 나타낸다. nc_{ij} , $_{j=1,2,\ldots,n}$는 0과 1 사이의 값이며, $\Sigma_j nc_{ij} = 1$이다.

유사도 비교 함수 $SIM(IC, NC_i)$

$= 1 - H[(IC + NC_i) / 2] [H(IC) + H(NC_i)] / 2$

단, 확률벡터 $P = (w_1, w_2, \ldots, w_n)$ 에 대하여 H(P)는 P의 불확실성 정도를 나타내는 엔트로피(entropy)로써 $H(P) = \Sigma_i \; w_i \; log_2 \; w_i$로 계산된다.

유사도를 비교할 때 구축된 공기정보 벡터 가운데 모든 의미에 대한 벡터와 비교할 수도 있고 "기구" 단어가 가지는 "조직"과 "운송도구"의 두 의미에 대한 공기정보 벡터와 비교할 수도 있다. 본 논문에서는 애매성 해소하고자 하는 단어가 가지는 의미들의 공기벡터와만 비교한다.

3.3 실험 및 문제점

본 시스템에서는 공기 정보 구축을 위하여 Etriset의 27여만 단어를 대상으로 하였다. 그중 시소러스 도구가 의미를 찾는 것은 14만여 단어이다. 본 시스템의 검사를 위해 Etriset의 문서 가운데 따로 발췌하여 검사 데이터로 삼았다. 검사 데이터에서 시소러스 도구가 의미를 찾을 수 있는 단어 수는 594개이고 그 중 애매성이 있는 단어는 201개였다. 각 공기정보에 대해 실험한 결과 다음과 같은 데이터가 나왔다.

	앞N범주	뒤N범주	문장범주	문서범주
성공률	45%	41%	46.2%	46.2%
형태소해석실패 제외한 경우 성공률	60%	53%	62%	62%

46%의 성공률은 바람직하지 않다. 그렇지만 개념기반 문서분류[7]에는 필요하지 않는 의미를 배제할 수 있으면 좋으므로 낮은 성공률이라도 보탬이 된다. 문서분류에 적용시 일정 한계치 이상의 유사도를 얻는 의미를 단어의 의미로 사용하면 된다.

애매성 해소가 실패한 경우를 분석하여 다음과 같은 사실을 알아내었다.

(1) 형태소 해석 실패 : "중, 일어, 가지, 개" 등의 단어가 실패를 많이 했다. 이 단어들은 다른 단어보다 빈번히 출현하는 단어이다. "중"의 경우 가운데라는 뜻인데 명사로 스님의 뜻으로 해석이 되었고, "일어"는 "일어나다" 동사인데 형태소 해석에서 일어(일본어)로 해석하였고 "가지"는 "가지다" 동사인데 가지(식물)로 해석이 되었고 "개"는 개수를 뜻하는 것인데 명사 개(동물)로 해석이 되었다. 형태소 해석의 실패를 제외한 경우의 성공률은 62%까지 나왔다.

(2) 시소러스 도구의 개선 : 현재 4000여 단어에 대해 의미를 획득할 수 있다. 더 많은 단어에 대해 의미를 획득할 수 있다면 더 보편적인 의미 관계 정보를 얻을 수 있을 것으로 보인다.

(3) 구문, 격의미 해석의 부족 : 단어의 의미는 구문 관계, 격의미 관계 등의 정보로 해석할 수 있으나 본 연구에서는 이것이 생략되었다. 따라서 이에 대한 보완이 필요하다.

4. 결 론

의미를 사용하는 문서 분류의 성능을 향상하기 위해 단어의 애매성 해소방법으로 의미간 공기정보를 사용하였다. 시스템을 실험한 결과 형태소 해석 실패를 제외한 경우 62%의 성공률을 얻었다. 이 성공률은 구문, 격의미 해석이 포함되지 않은 공기 관계의 정보만을 이용한 것이므로 개선의 여지가 많다. 앞으로 구문, 격의미 해석과 의미공기정보 이용의 복합적인 애매성 해소 시스템을 구축하여 보다 나은 성공률을 도모할 것이다. 그리고 시소러스 도구의 확장을 통해 더 보편적인 의미간의 공기정보를 획득하여 의미 공기정보 이용의 성공률의 향상을 꾀할 것이다.

5. 참고 문헌

[1]. 김봉섭, 이종혁, 이근배, "말뭉치를 기반으로 한 한국어 명사의 의미 중의성 해소," 한국정보과학회 학술대회 논문집, 24권 2호, 1997.

[2]. 문경희, 이종혁, 김정인, 양기주, "일한 기계번역 시스템 : 연어 패턴을 이용한 어휘 다의성 해소," 정보과학회 논문지(B), 25권 8호, 1998.

[3]. 이호, 백대호, 임해창, "분류 정보를 이용한 단어 의미 중의성 해결," 한국정보과학회 논문지(B) 24권 7호, 1997.

[4]. Brown, P.F. , Pietra V.J.D., DeSouza P.V., Lai, J.C. and Mercer R.L., "Class-based N-gram models of natural language," CL, Vol. 18, No.4, 1992.

[5]. 조정미, 조영환, 김길창, "코퍼스와 사전을 이용한 한국어 동사 의미 분별," 한국정보과학회 학술대회 논문집, 24권 2호, 1997.

[6]. 양재형, 심광섭, "시소러스와 하위범주화 사전을 이용한 격모호성 해결," 정보과학회 논문지(B), 26권 9호, 1999.

[7]. 강원석, 강현규, "시소러스 도구를 이용한 실시간 개념기반 문서분류 시스템," 한국정보과학회 논문지, 26권 1호, 1999.

[8]. 강원석, 황도삼, 최기선, "의미의 상하위 정보를 이용한 웹문서 분류 시스템", 제 11회 한글 및 한국어 정보처리학술대회 논문집, 한국정보과학회 및 한국인지과학회, 1999.10.