



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2016년03월31일

(11) 등록번호 10-1607898

(24) 등록일자 2016년03월25일

(51) 국제특허분류(Int. Cl.)

G06F 17/28 (2006.01) G06F 17/27 (2006.01)

(21) 출원번호 10-2014-0166267

(22) 출원일자 2014년11월26일

심사청구일자 2014년11월26일

(56) 선행기술조사문헌

US20130067311 A1*

US20130325759 A1*

이창기 외, 한국어 의미역 말뭉치 구축을 위한 반 자동 태깅 도구 개발, 한국정보과학회 2014 한국 컴퓨터종합학술대회 논문집, 2014.6, pp 592-594.

석미란 외, 개념 계층구조 상의 유사도를 이용한 이중 의미역의 자동변환, 한국정보과학회 2014 한국 컴퓨터종합학술대회 논문집, 2014.6, pp 1773-1775.,

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

한림대학교 산학협력단

강원도 춘천시 한림대학길 1, 한림대학교(옥천동)

(72) 발명자

김유섭

서울특별시 강남구 삼성로133길 20, 101동 811호 (청담동, 청담동 삼성1차 아파트)

석미란

서울특별시 강동구 동남로71길 41, 11동 803호 (명일동, 현대아파트)

(뒷면에 계속)

(74) 대리인

특허법인세신

전체 청구항 수 : 총 9 항

심사관 : 경연정

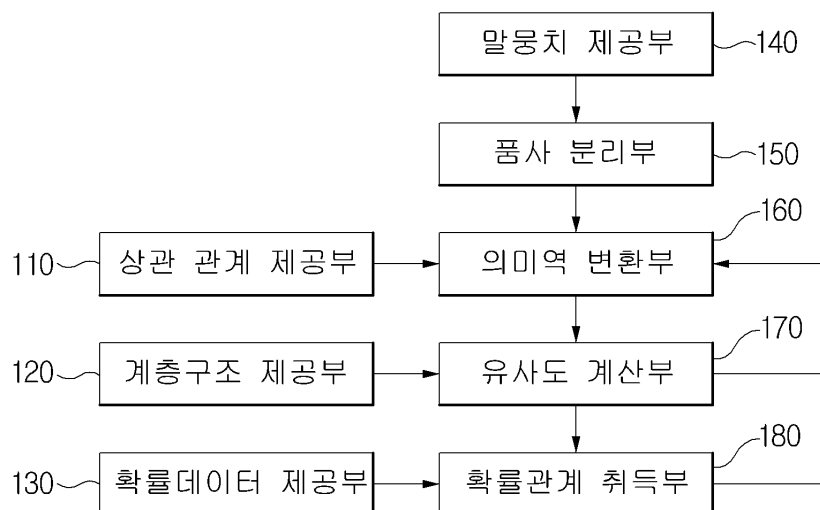
(54) 발명의 명칭 이중 의미역의 자동 변환 장치 및 방법

(57) 요약

본 발명은 이중 의미역의 자동 변환 장치에 관한 것으로, 이중 의미역들의 상관관계를 제공하는 상관관계 제공부와, 이중간의 의미역 자동 변환을 위해 구문 표지 부착 말뭉치들을 제공하는 말뭉치 제공부와, 구문 표지 부착 말뭉치로부터 단어를 분리하는 품사 분리부와, 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가

(뒷면에 계속)

대표도 - 도1



진 단어 간의 유사도를 계산하는 유사도 계산부와, 상관관계 제공부에서 제공된 이중 의미어들의 상관관계를 이용하여 품사 분리부에서 분리된 단어의 변환 대상의 의미어와 변환 후보의 의미어가 1:1로 매핑되어 있는 경우에는 변환 대상의 의미어를 매핑된 변환 후보의 의미어로 변환하고, 1:1로 매핑되어 있지 아니한 경우에는 유사도 계산부에서 계산된 유사도를 이용하여 변환 대상의 의미어를 변환 후보의 의미어들 중 어느 하나의 변환 후보의 의미어로 변환하는 의미어 변환부를 포함함으로써, 이중 의미어 간에 1:1로 매핑될 수 없는 경우에도 유사도를 이용하여 자동으로 변환할 수 있다.

(72) 발명자

서민정

강원도 춘천시 서부대성로 33, 104동 806호 (소양로2가, e편한세상 춘천)

윤영신

강원도 춘천시 영서로 2319, 302동 507호 (온의동, 금호3차아파트)

명세서

청구범위

청구항 1

이종 의미역들의 상관관계를 제공하는 상관관계 제공부와,
 이종간의 의미역 자동 변환을 위해 구문 표지 부착 말뭉치를 제공하는 말뭉치 제공부와,
 상기 말뭉치 제공부에서 제공된 구문 표지 부착 말뭉치로부터 단어를 분리하는 품사 분리부와,
 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가진 단어 간의 유사도를 계산하는 유사도 계산부와,
 상기 상관관계 제공부에서 제공된 이종 의미역들의 상관관계를 이용하여 상기 품사 분리부에서 분리된 단어의 변환 대상의 의미역과 변환 후보의 의미역이 1:1로 매핑되어 있는 경우에는 변환 대상의 의미역을 매핑된 변환 후보의 의미역으로 변환하고, 1:1로 매핑되어 있지 아니한 경우에는 상기 유사도 계산부에서 계산된 유사도를 이용하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환하는 의미역 변환부와, 및
 단어들의 계층 구조를 제공하는 계층 구조 제공부를 포함하고,
 상기 유사도 계산부는 계층 구조에 따른 단어들의 클래스 사이의 경로 길이를 구함으로써 단어들 사이의 유사도를 계산하는 것을 특징으로 하는 이종 의미역의 자동 변환 장치.

청구항 2

삭제

청구항 3

제1항에 있어서,
 상기 유사도 계산부는 계층 구조의 클래스 사이에 레벨 당 가중치를 달리하여 단어들 사이의 유사도를 계산하는 것을 특징으로 하는 이종 의미역의 자동 변환 장치.

청구항 4

제1항 또는 제3항에 있어서,
 조사들과 관련된 확률 관계를 제공하는 확률 테이블 제공부와,
 상기 확률 테이블 제공부를 이용하여 조사와 관련된 확률 관계를 취득하는 확률 관계 취득부를 더 포함하고,
 상기 품사 분리부는 상기 말뭉치 제공부에서 제공된 구문 표지 부착 말뭉치로부터 논항을 구분하고, 그리고 구분된 논항으로부터 단어와 조사를 분리하고,
 상기 의미역 변환부는 상기 유사도 계산부에서 계산된 유사도 및 상기 확률 관계 취득부에서 취득한 조사 관련 확률 관계를 종합하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환하는 것을 특징으로 하는 이종 의미역의 자동 변환 장치.

청구항 5

제4항에 있어서,
 상기 확률 테이블 제공부는 술어들과 관련된 확률 관계를 더 제공하고,
 상기 확률 관계 취득부는 상기 확률 테이블 제공부를 이용하여 술어와 관련된 확률 관계를 취득하고,
 상기 품사 분리부는 상기 말뭉치 제공부에서 제공된 구문 표지 부착 말뭉치로부터 논항과 술어의 의존 관계를

고려하여 논항과 술어를 분리하고,

상기 의미역 변환부는 상기 유사도 계산부에서 계산된 유사도 및 상기 확률 관계 취득부에서 취득한 조사 관련 확률 관계와 술어 관련 확률 관계를 종합하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환하는 것을 특징으로 하는 이중 의미역의 자동 변환 장치.

청구항 6

상관관계 제공부에서 이중 의미역들의 상관관계를 제공하는 단계와,

계층 구조 제공부에서 단어들의 계층 구조를 제공하는 단계와,

말뭉치 제공부에서 이중간의 의미역 자동 변환을 위해 구문 표지 부착 말뭉치가 제공되는 단계와,

품사 분리부에서 상기 구문 표지 부착 말뭉치로부터 단어를 분리하는 단계와,

의미역 변환부에서 상기 이중 의미역들의 상관관계를 이용하여 상기 분리된 단어의 변환 대상의 의미역과 변환 후보의 의미역이 1:1로 매핑되어 있는 경우에는 변환 대상의 의미역을 매핑된 변환 후보의 의미역으로 변환하는 단계와,

유사도 계산부에서 상기 분리된 단어의 변환 대상의 의미역과 변환 후보의 의미역이 1:1로 매핑되어 있지 아니한 경우에는 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가진 단어 간의 유사도를 계산하는 단계와,

상기 의미역 변환부에서 상기 유사도를 계산하는 단계의 유사도를 이용하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환하는 단계를 포함하고,

상기 유사도를 계산하는 단계는, 계층 구조에 따른 단어들의 클래스 사이의 경로 길이를 구함으로써 단어들 사이의 유사도를 계산하는 것을 특징으로 하는 이중 의미역의 자동 변환 방법.

청구항 7

제6항에 있어서,

확률 데이터 제공부에서 조사들과 관련된 확률 관계를 갖는 조사 관련 확률 테이블을 제공하는 단계와, 및

확률 관계 취득부에서 상기 조사 관련 확률 테이블을 이용하여 상기 분리된 조사와 관련된 확률 관계를 얻는 단계를 더 포함하고,

상기 단어를 구분하는 단계는 상기 구문 표지 부착 말뭉치로부터 논항을 구분하고, 그리고 구분된 논항으로부터 단어와 조사를 분리하고,

상기 변환 후보의 의미역으로 변환하는 단계는, 상기 유사도를 계산하는 단계에서 계산된 유사도와 상기 조사와 관련된 확률 관계를 얻는 단계에서 얻은 확률 관계를 종합하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환하는 것을 특징으로 하는 이중 의미역의 자동 변환 방법.

청구항 8

제7항에 있어서,

상기 확률 데이터 제공부에서 술어들과 관련된 확률 관계를 갖는 술어 관련 확률 테이블을 제공하는 단계와, 및

상기 확률 관계 취득부에서 상기 술어 관련 확률 테이블을 이용하여 상기 분리된 술어와 관련된 확률 관계를 얻는 단계를 더 포함하고,

상기 단어를 구분하는 단계는 상기 구문 표지 부착 말뭉치로부터 논항과 술어의 의존 관계를 고려하여 논항과 술어를 분리하고,

상기 변환 후보의 의미역으로 변환하는 단계는, 상기 유사도를 계산하는 단계에서 계산된 유사도와 상기 조사와 관련된 확률 관계를 얻는 단계에서 얻은 확률 관계와, 및 상기 술어와 관련된 확률 관계를 얻는 단계에서 얻은 확률 관계를 종합하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환하는 것을 특징으로 하는 이중 의미역의 자동 변환 방법.

청구항 9

제6항 내지 제8항 중 어느 한 항에 있어서,

상기 유사도를 계산하는 단계는 변환 대상이 될 의미역을 P, 2개의 변환 후보의 의미역들을 각각 S1 및 S2로 설정하고, P와 S1의 그룹을 SET 1, P와 S2의 그룹을 SET 2라고 설정하여 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가진 단어 간의 유사도를 계산하는 것을 특징으로 하는 이중 의미역의 자동 변환 방법.

청구항 10

제9항에 있어서,

상기 변환 후보의 의미역으로 변환하는 단계는, SET1의 단어들과 SET2의 단어들에서 P만을 갖는 p와의 경로 길이가 짧은 소정 개수의 단어들을 선정하고, 변환 대상의 의미역을 선정된 소정 개수의 평균 값이 작은 쪽의 변환 후보의 의미역으로 변환하는 것을 특징으로 하는 이중 의미역의 자동 변환 방법.

발명의 설명

기술 분야

[0001] 본 발명은 이중 의미역의 자동 변환 장치 및 방법에 관한 것으로, 개념 계층구조 상의 유사도를 이용한 이중 의미역의 자동 변환 장치 및 방법에 관한 것이다.

배경 기술

[0002] 의미역 결정(Semantic Role Labeling)이란 문장 내의 술어-논항들의 의미 관계를 결정하는 과정으로, 의미역 결정을 하기 위해서는 의미역이 부착되어 있는 대량의 말뭉치가 필요하다. 현재 가장 널리 사용하는 말뭉치는 Proposition Bank(이하 'PropBank'라 함)이다.

[0003] PropBank는 술어-논항 구조를 태그해 놓은 말뭉치를 말한다. 이 말뭉치는 영어와 중국어로 위주로 구축되어 있어, 한국어 PropBank의 경우 영어와 중국어의 말뭉치에 비해 학습 자료가 적어 정확도가 적고, 한국어의 특수성이 반영되지 않아 의미역 결정에 그대로 활용하기에 문제가 있었다.

[0004] 이러한 문제점을 해소하기 위해 한국어에 특화된 세종 계획에서 정의한 의미 표지 부착 말뭉치를 사용한다. 기존의 한국어 의미역 결정 관련 연구들이 세종의 의미 체계를 주로 활용하였기 때문에 세종 의미 표지 부착 말뭉치 구축 역시 매우 중요하다.

[0005] 그러나 PropBank 의미역과 세종 의미역이 부착되어 있는 말뭉치를 구축하는 작업은 수동으로 진행되기 때문에 말뭉치 구축에는 많은 시간과 비용이 소모된다. 때문에 새로운 의미 표지 부착 말뭉치를 이미 구축되어 있는 의미 표지 부착 말뭉치를 자동으로 변환하여 구축할 수 있는 방법이 필요하다.

발명의 내용

해결하려는 과제

[0006] 상술한 문제를 해결하기 위해, 본 발명은 개념 계층구조 상의 유사도를 이용한 이중 의미역의 자동 변환 장치 및 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0007] 상술한 목적을 달성하기 위해, 본 발명의 일실시예에 따른 이중 의미역의 자동 변환 장치는, 이중 의미역들의 상관관계를 제공하는 상관관계 제공부와, 이중간의 의미역 자동 변환을 위해 구문 표지 부착 말뭉치를 제공하는 말뭉치 제공부와, 상기 말뭉치 제공부에서 제공된 구문 표지 부착 말뭉치로부터 단어를 분리하는 품사 분리부와, 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가진 단어 간의 유사도를 계산하는 유사도 계산부와, 상기 상관관계 제공부에서 제공된 이중 의미역들의 상관관계를 이용하여 상기 품사 분리부에서 분리된 단어의 변환 대상의 의미역과 변환 후보의 의미역이 1:1로 매핑되어 있는 경우에는 변환 대상의 의미역을 매핑된 변환 후보의 의미역으로 변환하고, 1:1로 매핑되어 있지 아니한 경우에는 상기 유사도 계산부에서 계산된 유사도를 이용하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로

변환하는 의미역 변환부를 제공할 수 있다.

- [0008] 상기 이중 의미역의 자동 변환 장치는 단어들의 계층 구조를 제공하는 계층 구조 제공부를 더 포함하고, 상기 유사도 계산부는 계층 구조에 따른 단어들의 클래스 사이의 경로 길이를 구함으로써 단어들 사이의 유사도를 계산할 수 있다.
- [0009] 상기 유사도 계산부는 계층 구조의 클래스 사이에 레벨 당 가중치를 달리하여 단어들 사이의 유사도를 계산할 수 있다.
- [0010] 상기 이중 의미역의 자동 변환 장치는 조사들과 관련된 확률 관계를 제공하는 확률 테이블 제공부와, 상기 확률 테이블 제공부를 이용하여 조사와 관련된 확률 관계를 취득하는 확률 관계 취득부를 더 포함하고, 상기 품사 분리부는 상기 말뭉치 제공부에서 제공된 구문 표지 부착 말뭉치로부터 논항을 구분하고, 그리고 구분된 논항으로부터 단어와 조사를 분리하고, 상기 의미역 변환부는 상기 유사도 계산부에서 계산된 유사도 및 상기 확률 관계 취득부에서 취득한 조사 관련 확률 관계를 종합하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환할 수 있다.
- [0011] 상기 확률 테이블 제공부는 술어들과 관련된 확률 관계를 더 제공하고, 상기 확률 관계 취득부는 상기 확률 테이블 제공부를 이용하여 술어와 관련된 확률 관계를 취득하고, 상기 품사 분리부는 상기 말뭉치 제공부에서 제공된 구문 표지 부착 말뭉치로부터 논항과 술어의 의존 관계를 고려하여 논항과 술어를 분리하고, 상기 의미역 변환부는 상기 유사도 계산부에서 계산된 유사도 및 상기 확률 관계 취득부에서 취득한 조사 관련 확률 관계와 술어 관련 확률 관계를 종합하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환할 수 있다.
- [0012] 본 발명의 일실시예에 따른 이중 의미역의 자동 변환 방법은, 이중 의미역들의 상관관계를 제공하는 단계와, 이중간의 의미역 자동 변환을 위해 구문 표지 부착 말뭉치가 제공되는 단계와, 상기 구문 표지 부착 말뭉치로부터 단어를 분리하는 단계와, 상기 이중 의미역들의 상관관계를 이용하여 상기 분리된 단어의 변환 대상의 의미역과 변환 후보의 의미역이 1:1로 매핑되어 있는 경우에는 변환 대상의 의미역을 매핑된 변환 후보의 의미역으로 변환하는 단계와, 상기 분리된 단어의 변환 대상의 의미역과 변환 후보의 의미역이 1:1로 매핑되어 있지 아니한 경우에는 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가진 단어 간의 유사도를 계산하는 단계와, 상기 유사도를 계산하는 단계에서 계산된 유사도를 이용하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환하는 단계를 제공함으로써, 상술한 목적을 달성할 수 있다.
- [0013] 상기 유사도를 계산하는 단계는 변환 대상이 될 의미역을 P, 2개의 변환 후보의 의미역들을 각각 S1 및 S2로 설정하고, P와 S1의 그룹을 SET 1, P와 S2의 그룹을 SET 2라고 설정하여 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가진 단어 간의 유사도를 계산할 수 있다.
- [0014] 상기 변환 후보의 의미역으로 변환하는 단계는, SET1의 단어들과 SET2의 단어들에서 P만을 갖는 p와의 경로 길이가 짧은 소정 개수의 단어들을 선정하고, 변환 대상의 의미역을 선정된 소정 개수의 평균 값이 작은 쪽의 변환 후보의 의미역으로 변환할 수 있다.

발명의 효과

- [0015] 상술한 구성에 의해, 본 발명은 이중 의미역 간에 1:1로 매핑될 수 없는 경우에도 유사도를 이용하여 자동으로 변환할 수 있다.
- [0016] 본 발명은 또한, 조사와 술어 관련 확률 테이블을 이용함으로써 자동 변환시 변환 확률을 높일 수 있다.

도면의 간단한 설명

- [0017] 도 1은 본 발명의 일실시예에 따른 이중 의미역의 자동 변환 장치의 블록도를 도시하는 도면이다.
- 도 2는 PropBank 의미역과 세종 의미역의 상관관계를 도시하는 도면이다.
- 도 3은 유사도 계산을 위해 이용되는 한국어 명사편의 계층 구조를 나타내는 도면이다.
- 도 4는 본 발명의 일실시예에 따른 이중 의미역의 자동 변환 방법의 흐름도를 도시하는 도면이다.
- 도 5는 도 4에 도시된 의미역 간의 유사도를 계산하기 위한 흐름도를 도시하는 도면이다.
- 도 6은 논항의 조사와 논항과 같이 쓰이는 술어에 따라 의미역이 완전히 다르게 변환되는 예를 도시하는 도면이다.

다.

도 7은 본 발명의 다른 실시예에 따른 이중 의미역의 자동 변환 방법의 흐름도를 도시하는 도면이다.

도 8은 도 7의 실시예에 사용되는 술어-논항 관계의 예를 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0018] 이하, 첨부된 도면을 참조하여 본 발명에 따른 이중 의미역의 자동 변환 장치 및 방법의 바람직한 실시예를 설명한다. 참고로, 아래에서 본 발명을 설명함에 있어서, 본 발명의 구성요소를 지칭하는 용어들은 각각의 구성요소들의 기능을 고려하여 명명된 것이므로, 본 발명의 기술적 구성요소를 한정하는 의미로 이해되어서는 안 될 것이다.
- [0019] 도 1은 본 발명의 일 실시예에 따른 이중 의미역의 자동 변환 장치의 블록도를 도시하는 도면이다.
- [0020] 도 1에 도시된 바와 같이, 이중 의미역의 자동 변환 장치는 상관관계 제공부(110), 계층 구조 제공부(120), 확률 테이블 제공부(130), 말뭉치 제공부(140), 품사 분리부(150), 의미역 변환부(160), 유사도 계산부(170) 및 확률 관계 취득부(180)를 포함할 수 있다.
- [0021] 상관관계 제공부(110)는 이중 의미역들의 상관관계를 제공하고, 계층 구조 제공부(120)는 단어들의 계층 구조를 제공하고, 확률 테이블 제공부(130)는 조사들과 관련된 확률 관계 및 술어들과 관련된 확률 관계를 제공한다.
- [0022] 말뭉치 제공부(140)는 이중간의 의미역 자동 변환을 위해 구문 표지 부착 말뭉치를 제공한다. 품사 분리부(150)는 말뭉치 제공부(140)에서 제공된 구문 표지 부착 말뭉치로부터 논항과 술어를 구분하고, 그리고 구분된 논항으로부터 단어와 조사를 분리한다.
- [0023] 유사도 계산부(170)는 변환 대상이 될 의미역을 가진 단어와 각 변환 후보의 의미역을 가진 단어 간의 유사도를 계산한다. 이를 위해, 유사도 계산부(170)는 계층 구조에 따른 단어들의 클래스 사이의 경로 길이를 구함으로써 단어들 사이의 유사도를 계산할 수 있으며, 또한 계층 구조의 클래스 사이에 레벨 당 가중치를 달리하여 단어들 사이의 유사도를 계산할 수 있다.
- [0024] 확률 관계 취득부(180)는 확률 테이블 제공부(130)를 이용하여 조사와 관련된 확률 관계 및 술어와 관련된 확률 관계를 취득할 수 있다.
- [0025] 의미역 변환부(160)는 상관관계 제공부(110)에서 제공된 이중 의미역들의 상관관계를 이용하여 품사 분리부(150)에서 분리된 단어의 변환 대상의 의미역과 변환 후보의 의미역이 1:1로 매핑되어 있는 경우에는 변환 대상의 의미역을 매핑된 변환 후보의 의미역으로 변환하고, 1:1로 매핑되어 있지 아니한 경우에는 의미역 변환부(160)는 유사도 계산부(170)에서 계산된 유사도 및 확률 관계 취득부(180)에서 취득한 조사 관련 확률 관계와 술어 관련 확률 관계를 종합하여 변환 대상의 의미역을 변환 후보의 의미역들 중 어느 하나의 변환 후보의 의미역으로 변환할 수 있다.
- [0026] 도 2는 PropBank 의미역과 세종 의미역의 상관관계를 도시하는 도면이다.
- [0027] 도 2에 도시되어 있는 바와 같이, 방향을 뜻하는 PropBank의 M-DIR 의미역과 세종의 DIR 의미역과 같이 1:1 비슷한 확률로 매핑되는 경우가 있다. 그러나 ARG(시작점)과 같이 PropBank 의미역과 세종 의미역 사이에는 중의성이 존재하는 경우도 있다. 이러한 PropBank 의미역은 중의성을 가진 2개 이상의 세종 의미역을 가지게 된다.
- [0028] 도 2에 도시된 PropBank 의미역과 세종 의미역의 상관관계를 이용하여, ARG0, ARG1, M-CAU와 같이 세종 후보 의미역 1:1인 경우에는 바로 상관관계의 의미역을 결정하고, ARG2, M-ADV, M-CND와 같이 후보 의미역이 여러 개인 경우, 변환 대상이 되는 PropBank 의미역을 가진 단어와 세종 후보 의미역을 가진 단어들 간의 유사도를 계산하여 의미역을 결정할 수 있다.
- [0029] 의미역 간의 유사도를 계산하기 위해, BOLA(Bank of language resources)의 한국어 개념기반 어휘 의미망인 코어넷을 사용하였다. 코어넷은 총 2,938개의 계층적 개념과 92,448개의 어휘 의미가 연결되어 있다. 본 실시예에서는 CBL1(한국어 명사편)을 이용했다.
- [0030] 도 3은 유사도 계산을 위해 이용되는 한국어 명사편의 계층 구조를 나타내는 도면이다.
- [0031] 도 3에 도시된 한국어 명사편의 계층 구조로 각 숫자는 개념 체계 내에서의 위치에 관한 정보, 즉 상위 개념과 단계 정보를 제공한다. 개념 번호의 자릿수는 단계를 알려주며, 마지막 자릿수를 제거한 번호가 상위 개념에 해

당한다. 도 3에서 자연현상을 나타내는 1223은 3단계이며 상위 개념으로는 일<추상>을 나타내는 122가 있고, 더 상위인 12는 추상을 나타낸다.

- [0032] 본 실시예에서는 각 클래스간의 경로의 길이를 측정함으로써 유사도를 계산하였다. 자연현상과 추상물(정신) 사이의 유사도 계산을 위해 경로의 길이를 측정해 보면, 1223과 122 사이의 에지, 122와 12 사이의 에지, 12와 121 사이의 에지, 121과 1211 사이의 에지로 총 경로의 길이가 4인 것을 알 수 있다.
- [0033] '문명' 단어와 '봄빛' 단어의 유사도를 계산한다고 했을 때, '문명'은 추상물(정신)에 포함되는 단어이고, '봄빛'은 자연현상에 포함되는 단어이므로, 1211과 1223 사이의 경로의 길이를 구함으로써 두 단어 사이의 유사도를 계산할 수 있다. 경로의 길이가 짧을수록 유사도가 높다.
- [0034] 도 4는 본 발명의 일실시예에 따른 이중 의미역의 자동 변환 방법의 흐름도를 도시하는 도면이다.
- [0035] 자동 변환 장치에 PropBank 의미 표지 부착 말뭉치가 입력된다(S402). 자동 변환 장치는 입력된 PropBank 말뭉치에서 변환 대상의 의미역을 선정한다(S404).
- [0036] 자동 변환 장치는 도 2에 도시된 PropBank 의미역과 세종 의미역의 상관관계를 이용하여 변환 대상의 의미역이 세종 의미역과 1:1 의미역 관계인지를 확인한다(S406). 자동 변환 장치는 변환 대상의 의미역이 세종 의미역과 1:1 의미역 관계인 경우 상관관계의 의미역으로 변환한다(S408).
- [0037] 자동 변환 장치는 변환 대상의 의미역이 세종 의미역과 1:1 의미역 관계가 아닌 경우이면 의미역 간의 유사도를 계산한다(S410). 그리고 자동 변환 장치는 변환 대상이 될 의미역을 가진 단어와 후보 의미역을 가진 단어 간의 유사도를 이용하여 어느 하나의 변환 후보 의미역으로 변환한다(S412).
- [0038] 도 5는 도 4에 도시된 의미역 간의 유사도를 계산하기 위한 흐름도를 도시하는 도면이다.
- [0039] 자동 변환 장치는 먼저, 변환 대상이 될 PropBank 의미역을 P, 변환 후보 세종 의미역 2개를 각각 S1, S2로 설정하고, P와 S1의 그룹을 SET 1, P와 S2의 그룹을 SET 2라고 설정한다(S502). 자동 변환 장치는 또한, P만 가진 것을 p로 설정한다(S504).
- [0040] 자동 변환 장치는 p, SET1, SET2로 단어들을 의미별 명사 계층 구조에서 찾고 SET1의 단어들을 G1, SET2의 단어들을 G2로 묶는다(S506). 자동 변환 장치는 G1, G2에 속한 단어들의 개념 번호(클래스)를 찾고, p의 개념번호도 찾는다(S508).
- [0041] 자동 변환 장치는 p의 개념번호와 G1의 개념번호를 비교하고, G2 또한 p와 유사도를 비교한다(S510). 그리고 자동 변환 장치는 G1과 G2에서 p와의 거리가 짧은 단어 5개를 선정하고, 선정된 5개의 평균값이 작은 쪽으로 PropBank 의미역을 세종 의미역으로 변환한다(S512).
- [0042] 본 실시예에서는 트리 형식의 된 명사 클래스간의 레벨 당 가중치를 두어 유사도를 계산할 수 있다. 계층 구조를 반영하지 않고 유사도 계산을 했을 경우, 1234와 123사이의 에지, 123과 1232사이의 에지, 1232와 12325사이의 에지를 계산하여 4라는 값을 가질 수 있다. 클래스의 숫자가 1의 자리로 갈수록 더 작은 클래스를 나타내므로, 두 명사가 좀 더 유사하다는 것을 알 수 있다. 이를 이용하여 트리 구조로 된 명사 클래스의 레벨 당 유사도 값을 더 적게 주는 방식을 사용하였다. 가장 큰 클래스가 다를 경우 1을, 하위 클래스로 갈수록 유사도 값을 1/2로 줄여가며 계산을 할 수 있다. 따라서 레벨 3과 레벨 4 각각의 가중치 0.125와 0.0625를 합쳐 0.188이란 유사도를 가질 수 있다.
- [0043] 도 4 및 도 5에서는 자동 변환을 위하여 의미역 변환 대상이 되는 논항의 의미 유사도를 계산하였다. 이미 구축된 데이터에서 변환 대상 논항과 의미상으로 유사한 단어들의 의미역을 찾아 이로 변환하고자 하였다. 그러나 이러한 경우에는 논항 그 자체만 고려할 뿐, 그 논항의 문맥 정보를 고려하지 못한다. 따라서 논항의 의미 유사도뿐만 아니라 논항과 의미상으로 밀접한 관계가 있을 것으로 추정되는 논항의 조사 및 술어도 고려한 자동 변환 방법을 더 설명한다.
- [0044] 도 6은 논항의 조사와 논항과 같이 쓰이는 술어에 따라 의미역이 완전히 다르게 변환되는 예를 도시하는 도면이다.
- [0045] 일반 문장은 논항과 술어로 구성된다. 논항이 주격 또는 목적격으로 쓰이는 경우, 주어, 목적어를 나타내는 PropBank의 ARG0, ARG1, 세종의 AGT, THM으로 대부분 1:1로 맵핑된다. 그러나 부사격의 경우, 논항의 조사와 술어에 따라 의미역이 완전히 다르게 부착된다.

[0046] 논항의 조사와 논항과 같이 쓰이는 술어에 따라 의미역이 완전히 다르게 변환되는 예가 도 6에 도시되어 있다. 명사 “학교”와 술어 “가다”에 조사 “로”가 붙은 경우, PropBank와 세종에서 도착점을 나타내는 ARG3, GOL로 맵핑된다. “에”가 붙은 경우, 장소를 나타내는 LOC, LOC가 맵핑된다. “에서”가 붙은 경우에는 출발점을 나타내는 ARG2, SRC로 맵핑된다.

[0047] 이처럼 의미역 결정을 하는데 조사와 술어에 따라 의미역이 완전히 다르게 결정되므로, 변환 대상이 될 의미역을 가진 단어와 후보 의미역을 가진 단어 간의 유사도를 계산하여 어느 의미역으로 변환될 지를 결정하는 것으로는 정확성이 떨어질 수 있다. 따라서 본 실시예에서는 정확한 의미역 결정을 위해 각 조사와 술어에 대한 정보를 가진 결합 확률 테이블을 구축한다.

[0048] 테이블의 데이터를 작성하기 위해 본 실시예에서는 한국전자통신연구원의 구문 표지 부착 말뭉치를 사용하여, 약 12,000개 문장 데이터의 술어-논항에 대해 의미역을 부착하였다. 의미역을 부착한 후, 확률 테이블은 동사는 4,414개의 테이블로 구축되고, 조사는 241개의 테이블로 구축되었다. 각 테이블은 세종 의미역들이 12,000개의 데이터에서 쓰인 정보를 저장하고 있다.

[0049] 표 1은 조사 “에”와 “같이”의 정보를 가진 확률 테이블의 부분이다.

표 1

조사	세종 의미역	확률
에	LOC	0.43
	GOL	0.26
같이	FFF	0.56
	FNS	0.06
	COM	0.17

[0051] 표 1에서, “에”는 세종 의미역에서 장소를 나타내는 LOC, 도착점을 나타내는 GOL으로 사용되었다. 따라서 조사 “에”는 장소로 많이 쓰였고, 장소에 따라 도착점인 경우에 많이 쓰였다는 것을 알 수 있다. 조사 “같이”는 영향주를 나타내는 EFF, 결과 상태를 나타내는 FNS, 동반주를 나타내는 COM으로 쓰였는데, 영향주의 의미로 주로 쓰였다는 것을 알 수 있다.

[0052] 표 2는 동사 “근거하다”와 “보다”의 확률 테이블에서 상위 의미역을 나타낸 것이다.

표 2

술어	세종 의미역	확률
근거하다	THM	0.33
	LOC	0.67
보다	AGT	0.12
	THM	0.51
	LOC	0.14

[0054] 표 2에서, “근거하다”는 세종 의미역에서 장소를 나타내는 LOC를 많이 사용했다. “보다”의 경우, 행위주를 나타내는 AGT, 대상을 나타내는 THM, 장소를 나타내는 LOC로 맵핑되었고, 이 중 대상이라는 의미로 많이 사용했다는 것을 알 수 있다. 따라서 확률테이블은 논항 외, 조사, 술어에 따라 의미역이 다르게 쓰일 수 있고, 문맥상 논항의 의미역 변환에 필요하다.

[0055] 도 7은 본 발명의 다른 실시예에 따른 이중 의미역의 자동 변환 방법의 흐름도를 도시하는 도면이고, 도 8은 도 7의 실시예에 사용되는 술어-논항 관계의 예를 도시하는 도면이다.

[0056] 본 실시예에서는 술어-논항의 관계를 파악하기 위해 한국전자통신연구원의 구문 표지 부착 말뭉치를 문장 데이터로 이용한다. 설명을 위해 예제 "때문에 사실주의라는 말이 적용되려면 19세기의 고전적 사실주의에서와 같이 오늘의 우리의 견지에서 보아서도 현실에 직접 근거했다고 볼 수 있는 경우에 한해서 적용되어야 한다."를 사용한다.

[0057] 자동 변환 장치에 상술한 예제의 한국전자통신연구원의 구문 표지 부착 말뭉치가 입력된다(S702). 도 8에 도시된 바와 같이, 논항과 술어의 의존 관계는 2개의 인덱스로 표현되는데, 각 줄마다 첫 번째 인덱스는 어절의 순

번이고, 두 번째 인덱스는 의존하고 있는 어절의 순번을 가리킨다. 각각의 줄에서 두 번째 인덱스와 첫 번째 인덱스가 같으면 의존 관계가 있다고 표현한다. 예를 들면, 논항 “때문에”, “적용되려면”, “한하여”는 술어 “적용되어야 한다.”와 관계가 있다. 이런 논항의 의미역을 변환하기 위해, 자동 변환 장치는 논항과 술어의 의존관계를 고려하여 논항과 술어로 구분한다(S704).

[0058] 표 3은 논항과 술어의 의존 관계를 고려하여 구분한 결과이다.

표 3

[0059]

논항	의미역	세종 후보 의미역
말+이	ARG1	THM
견지+에서	LOC	LOC
현실+에	ARG2	FNS, GOL, LOC
사실주의+에서와 같이	ADV	EFF, CNT
경우+에	ARG2	FNS, GOL, LOC

[0060] 자동 변환 장치는 논항과 술어의 의존 관계를 고려하여 구분된 논항을 명사 + 조사로 분리한다(S706).

[0061] 자동 변환 장치는 도 2에 도시된 PropBank 의미역과 세종 의미역의 상관관계를 이용하여 변환 대상의 의미역이 세종 의미역과 1:1 의미역 관계인지를 확인한다(S708). 자동 변환 장치는 변환 대상의 의미역이 세종 의미역과 1:1 의미역 관계인 경우 상관관계에 있는 의미역으로 변환한다(S710).

[0062] 자동 변환 장치는 변환 대상의 의미역이 세종 의미역과 1:1 의미역 관계가 아닌 경우에는 변환 대상이 될 의미역을 가진 단어와 후보 의미역을 가진 단어 간의 유사도를 계산한다(S712). 자동 변환 장치는 또한, 조사와 관련된 확률 테이블의 확률 관계를 이용하여 조사와 관련된 확률 값들을 얻는다(S714). 자동 변환 장치는 또한, 논항과 의존 관계에 있는 술어와 관련된 확률 테이블의 확률 관계를 이용하여 술어와 관련된 확률 값들을 얻는다(S716).

[0063] 자동 변환 장치는 단어 간의 유사도, 조사와 관련된 확률 값들 및 술어와 관련된 확률 값들을 종합하여 변환 대상이 될 의미역을 가진 단어를 후보 의미역들 중 어느 의미역으로 변환될 지를 결정한다(S718).

[0064] 표 4는 도 7의 흐름도에 따른 의미역 변환의 예를 설명한다.

표 4

[0065]

논항	PropBank	세종	술어	합계
현실+에	ARG2	GOL	근거하다	0.58
현실+에	ARG2	FNS	근거하다	0.30
현실+에	ARG2	LOC	근거하다	0.71
사실주의+에서와 같이	ADV	EFF	보다	0.53
사실주의+에서와 같이	ADV	CNT	보다	0.50
경우+에	ARG2	GOL	한하다	0.43
경우+에	ARG2	FNS	한하다	0.10
경우+에	ARG2	LOC	한하다	0.84

[0066] 표 4를 보면, 현실+에”가 ARG2(LOC), “사실주의+에서와 같이”는 ADV(EFF), “경우+에”는 ARG2(LOC)로 가장 높은 수치를 보여 각 세종 의미역으로 PropBank 의미역을 변환한다.

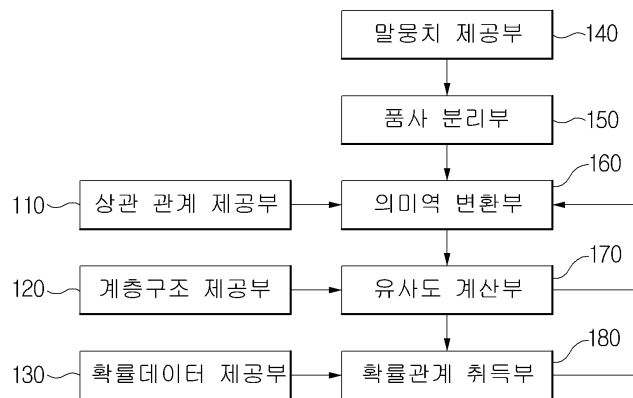
[0067] 본 발명의 보호 범위는 이하 특허청구범위에 의하여 해석되어야 마땅할 것이다. 또한, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것인 바, 본 발명과 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

- [0068]
- | | |
|-----------------|----------------|
| 110: 상관관계 제공부 | 120: 계층 구조 제공부 |
| 130: 확률 테이블 제공부 | 140: 말뭉치 제공부 |
| 150: 품사 분리부 | 160: 의미역 변환부 |
| 170: 유사도 계산부 | 180: 확률 관계 취득부 |

도면

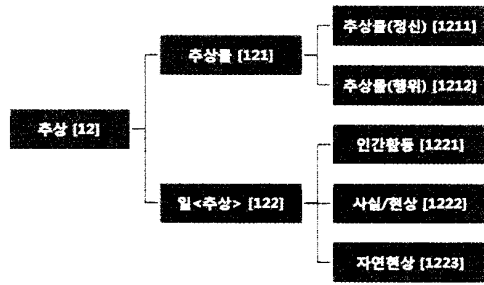
도면1



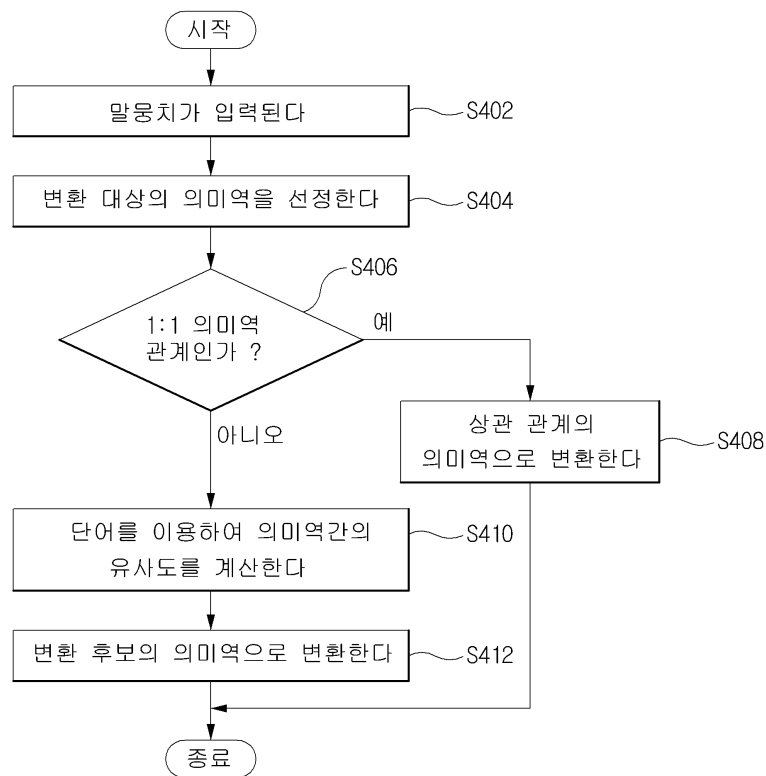
도면2

PropBank 의미역	세종 의미역
ARG0(주체자)	AGT(행위주)
ARG1(수동자)	THM(대상)
ARG2(시작점)	FNS(결과상태)
	GOL(도착점)
	LOC(장소)
ARG3(끝점)	GOL(도착점)
M-ADV(부사적 어구)	EFF(영향주)
M-CAU(원인)	CNT(내용)
M-CND(조건)	EFF(영향주)
M-DIR(방향)	CNT(내용)
	DIR(방향)
M-DIS(문장의 연결)	EFF(영향주)
	CNT(내용)
M-EXT(크기)	CRT(기준치)
	EFF(영향주)
M-INS(도구)	INS(도구)
M-LOC(장소)	LOC(장소)
M-MNR(방법)	INS(도구)
M-PRD(술어의 자격)	INS(도구)
	EFF(영향주)
M-PRP(목적)	PUR(목적)
M-TMP(시간)	LOC(장소)

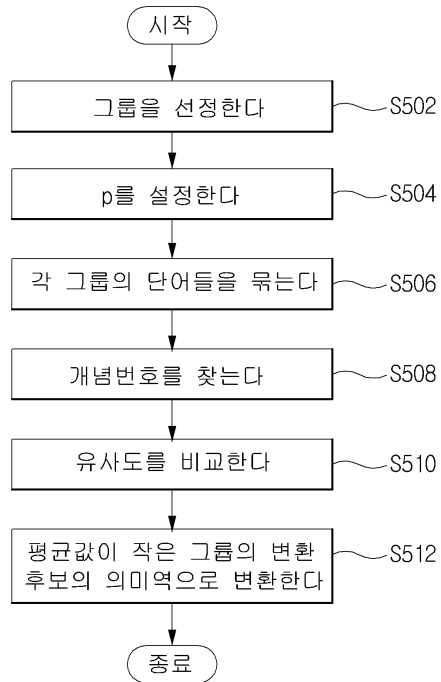
도면3



도면4



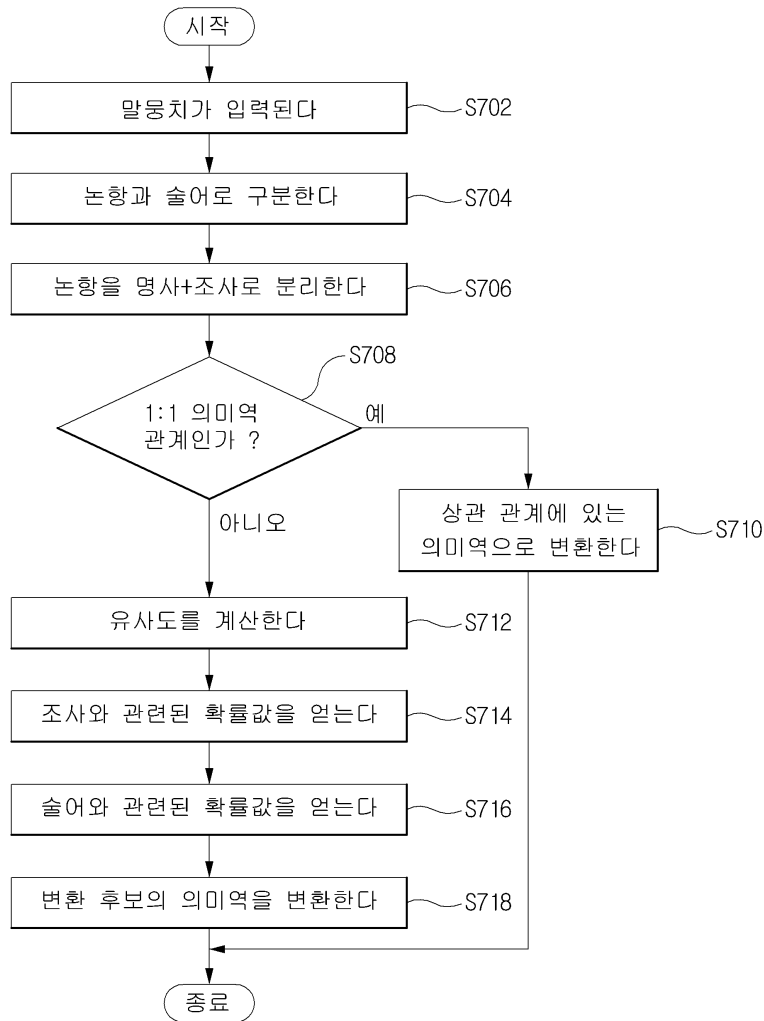
도면5



도면6

학교[용언불가능보통명사]+로[부사격조사]	ARG3(GOL)
학교[용언불가능보통명사]+에[부사격조사]	LOC(LOC)
학교[용언불가능보통명사]+에서[부사격조사]	ARG2(SRC)

도면7



도면8

1	18	때문에[문장접속부사]
2	3	사실주의[용언불가능보통명사]+라는[관형격조사]
3	4	말[용언가능보통명사]+이[주격조사]
4	18	적용되[일반동사]+려면[종속연결어미]
5	7	19_세기[용언불가능보통명사]+의[관형격조사]
6	7	고전적[성상관형사]
7	15	사실주의[용언불가능보통명사]+에서와_같_이[부사격조사]
8	9	오늘[용언불가능보통명사]+의[관형격조사]
9	10	우리[인칭대명사]+의[관형격조사]
10	11	견지[용언가능보통명사]+에서[부사격조사]
11	15	보[일반동사]+아서도[종속연결어미]
12	14	현실[용언불가능보통명사]+에[부사격조사]
13	14	직접[성상상태부사]
14	15	근거하[일반동사]+있[과거시제선어말어미]+다고[종속연결어미]
15	16	보[일반동사]+르_수_있[기타보조용언]+는[관형사형전성어미] (동격)
16	17	경우[용언불가능보통명사]+에[부사격조사]
17	18	한하[일반동사]+어서[종속연결어미]
18	0	적용되[일반동사]+어야_하[기타보조용언]+느_다[평서형종결어미]+(문미기호)