(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0023640
(43) 공개일자 2021년03월04일

(51) 국제특허분류(Int. Cl.)
G06F 40/20 (2020.01)
(52) CPC특허분류
G06F 40/289 (2020.01)
G06F 40/205 (2020.01)
(21) 출원번호 10-2019-0160737
(22) 출원일자 2019년12월05일
심사청구일자 2019년12월05일
(30) 우선권주장
1020190103819 2019년08월23일 대한민국(KR)

(71) 출원인
울산대학교 산학협력단
울산광역시 남구 대학로 93(무거동)
(72) 발명자
옥철영
울산광역시 남구 신복로35번길 6, 101동 602호 (무거동, 쌍용스윗닷홈)
정충선
울산광역시 중구 우정2길 45, 106동 1204호 (우정동, 우정아이파크)
(74) 대리인
김종선, 이형석

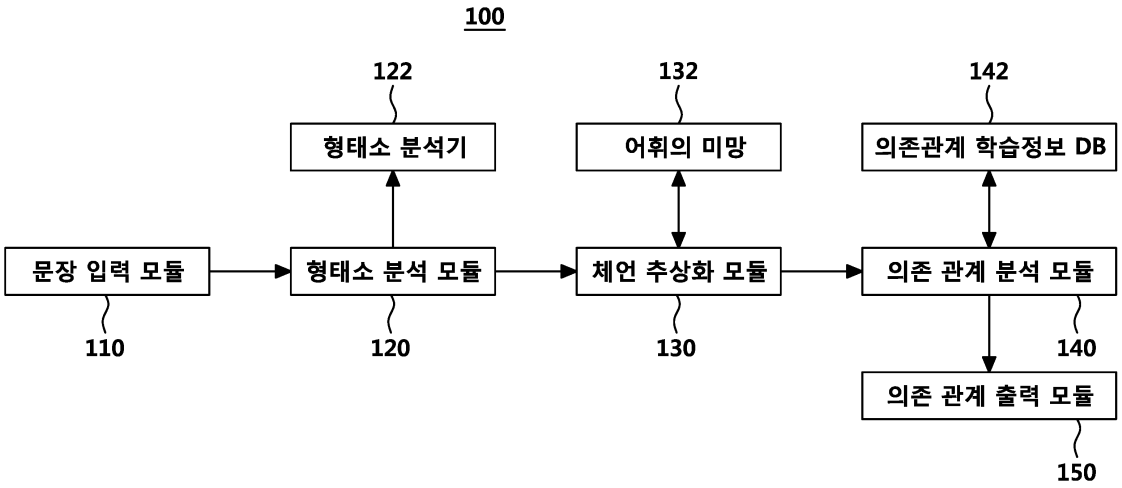
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템**

(57) 요 약

본 발명은 의미 추상화를 통해 비교적 적은 수의 문장에 대한 학습만으로 상위어가 추출되게 하여, 확장된 문장에 있어서의 의존관계를 도출할 수 있도록 하는 전이 기반의 한국어 의존관계 분석시스템에 관한 것이다. 이러한 본 발명은 입력된 문장의 형태소를 분석하는 형태소 분석모듈과, 분석된 형태소 정보에서 체언류의 표층형의 추상화된 정보를 데이터베이스로부터 획득하는 체언 추상화 모듈, 및 상기 체언의 추상화된 표층형 정보와 학습 말뭉치로부터 학습된 자질 정보를 이용하여 어절 간 의존관계를 분석하는 의존관계 분석 모듈을 포함하여 구성된다.

대 표 도 - 도1



(52) CPC특허분류
G06F 40/268 (2020.01)
G06F 40/30 (2020.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 IITP-20132001790072002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 SW컴퓨팅산업원천기술개발사업
연구과제명 컨텍스트 인지형 Deep-Symbolic 하이브리드 지능 원천 기술 개발 및 언어 지식 자원 구축
기 여 율 1/1
과제수행기관명 한국과학기술원
연구기간 2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

입력된 문장의 형태소를 분석하는 형태소 분석모듈;

분석된 형태소 정보에서 체언류의 표층형의 추상화된 정보를 데이터베이스로부터 획득하는 체언 추상화 모듈; 및

상기 체언의 추상화된 표층형 정보와 학습 말뭉치로부터 학습된 자질 정보를 이용하여 어절 간 의존관계를 분석하는 의존관계 분석 모듈을 포함하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 2

제 1 항에 있어서,

상기 형태소 분석모듈은,

상기 입력된 문장의 형태소 분석 결과와 동형이의어 분별 결과를 제공하며,

상기 형태소 분석결과는 형태소 단위로 표층형, 동형이의어 번호, 다의어 번호, 품사 태그가 부착되어 출력되고,

상기 분석결과로부터 입력 문장의 어절 단위와 형태소 단위로 표층형과 동형이의어 번호, 품사 태그 정보를 수집하는 의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템.

청구항 3

제 1 항에 있어서,

상기 체언 추상화 모듈은,

어휘 의미망을 활용하여 명사의 표층형을 추상화하여 자질로 사용하며,

상기 어휘 의미망은 명사의 계층 구조를 트리 구조로 제공하며, 상위어 정보를 통해 명사의 표층형 정보를 추상화하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 4

제 3 항에 있어서,

상기 표층형의 추상화는,

학습 말뭉치를 동형이의어로 분별하였고, 동형이의어로 분별된 정보를 기초로 어휘 의미망의 단어 계층에 따라 명사를 추상화하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 5

제 4 항에 있어서,

상기 명사의 표층형 추상화는, 표층형 자질, 추상화 자질을 포함한 어절단위 자질의 기본자질, 형태소 단위 및 음절 단위 확장 자질을 모두 학습하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 6

제 3 항에 있어서,

상기 체언 추상화 모듈은,

입력 문장의 형태소 단위의 품사 태그를 확인하여 체언류 형태소를 구분하고,

체언류 형태소에 대하여 상기 어휘 의미망에 해당 단어 정보가 있는지 확인하고,

상기 어휘 의미망에 없는 단어는 추상화하지 않고 정보가 있는 단어는 어휘 의미망의 단어 계층을 계산하고,

상기 동형이의어를 기준으로 동형이의어가 갖는 모든 다의어에 대하여 단어 계층을 모두 계산하고,

계산된 모든 다의어에 대하여 설정된 추상화 레벨만큼 상위 계층 단어로 추상화하며,

추상화된 단어 중 빈도가 가장 높은 단어를 해당 동형이의어의 추상화된 단어로 결정하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 7

제 6 항에 있어서,

상기 체언 추상화 모듈은,

추상화의 단계가 최고 단계를 초과하는 경우, 추상화 단계가 어휘 의미망의 최상위 단계 이상이 되면 root 노드 아래의 1계층 단어로 추상화하는 제1 방법과, 추상화를 하지 않고 원형을 그대로 사용하는 제2 방법 중 하나에 의해 수행되는 전이 기반 한국어 의존관계 분석 시스템.

청구항 8

제 1 항에 있어서,

상기 의존관계 분석모듈은,

한국어의 특징이 고려된 전이 기반 모델에 맞춰 입력 문장의 순서를 역순으로 입력 버퍼에 입력하고,

자질 함수를 통해 각 자질의 값을 구하여 전이(transition)을 결정하며,

상기 전이결정에 따라 상기 입력 버퍼의 모든 어절이 의존관계가 결정되어 입력 버퍼에 존재하는 어절이 없으면 의존관계 분석을 종료하고 입력 문장의 각 어절의 지배소 어절의 어절 위치를 기록하여 의존관계 결과를 출력하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 9

제 8 항에 있어서,

상기 의존관계 분석에 따른 기본 자질 성능 평가에서, 일반화된 품사 태그 자질(generalized POS)이 가장 높은 재현율을 제공하며, 표층형과 동형 이의의 번호, 품사 태그 합 자질(BPS)이 가장 높은 정확율을 제공하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 10

제 8 항에 있어서,

상기 의존관계 분석에 따른 확장 자질 성능 평가에서, 확장 자질의 재현율과 정확률은 상기 기본 자질에 비해 재현율은 높고, 정확률이 낮은 전이 기반 한국어 의존관계 분석 시스템.

청구항 11

제 8 항에 있어서,

상기 의존관계 분석에 따른 확장 자질 성능 평가에서, 형태소 자질과 음절 자질을 모두 사용한 경우가 가장 높은 의존관계 정확율을 제공하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 12

제 1 항에 있어서,

상기 의존관계 분석 모듈의 분석방법은,

한국어의 지배소 후위 원칙을 이용한 수정된 backward arc-eager 알고리즘을 사용하여 전이 기반 의존관계를 분

석하고,

상기 수정된 backward arc-eager 알고리즘은 기본 arc-eager 알고리즘에서 right-arc, shift transition은 제거되고 left-arc transition에서 stack의 어절을 제거하지 않고 buffer로 이동시키는 방법으로 분석하는 전이 기반 한국어 의존관계 분석 시스템.

청구항 13

제 12 항에 있어서,

상기 의존관계 분석에 따른 의존관계정확률 및 자질 정확률은 다음 식에 의해 계산되는 전이 기반 한국어 의존관계 분석 시스템.

$$\text{의존관계 정확률} = \frac{\text{정답과 같은 의존관계 결정 어절 수}}{\text{전체 어절 수}}, \quad \text{자질의 정확률} = \frac{\text{정답과 같은 } transition \text{ 결정 수}}{\text{자질의 재현 횟수}}$$

발명의 설명

기 술 분 야

[0001] 본 발명은 언어 의존관계 분석시스템에 관한 것으로, 더욱 상세하게는 의미 추상화를 통해 비교적 적은 수의 문장에 대한 학습만으로 상위어가 추출되게 하여, 확장된 문장에 있어서의 의존관계를 도출할 수 있도록 하는 전이 기반의 한국어 의존관계 분석시스템에 관한 것이다.

배 경 기 술

[0002] 국어, 특히 한국어의 의존관계 분석은 규칙 기반 연구와 학습 말뭉치를 기반한 연구 중에 학습 기반 연구가 주를 이루고 있다. 의존관계 분석을 위한 말뭉치는 어절 사이의 의존 관계 정보를 담고 있으며 어절은 형태소의 표층형과 품사 태그 정보로 표현한다.

[0003] 그러기 때문에, 학습 말뭉치의 형태소의 표층형과 품사 태그의 조합을 학습하게 되는데, 그 중 품사 태그의 경우 재현율이 높지만 품사 태그만으로 의존관계를 분석하는 것은 한계가 있다. 반면 표층형의 경우는 학습한 내용이 재현될 경우 높은 정확률을 보이지만 가짓수가 많아 어절을 구성하는 표층형의 조합을 일반화하기 어렵고 품사 태그에 비하여 낮은 학습 자질의 재현율을 보인다. 따라서 모든 조합의 표층형을 학습하기는 어렵고 학습 말뭉치에 포함된 표층형을 학습하더라도 학습 후 입력되는 테스트 문장에서 표층형이 재현될 확률은 낮다.

[0004] 이처럼 기존의 학습 기반 연구는 형태소의 표층형과 품사 태그의 조합으로 의존관계 유무를 판단하기 때문에, 어순 변화 및 필수 성분 생략 등으로 다양한 문장이 있는 한국어 문장의 의존관계를 정확히 판단하기에는 한계가 있다. 즉 한국어는 어순이 자유롭고 문장의 구성 성분의 생략이 가능하기 때문이다.

[0005] 한국어의 의존규칙을 분석했던 방법들은 종래에도 다양하게 제안되었다. 예를 들어 규칙 기반 의존관계 분석은, 비교적 짧은 문장에 대해서는 정확한 분석이 가능하지만, 문장의 길이가 길어지고 어순의 변화 및 생략이 복합되면 의존관계 정의를 위한 예외 규칙들이 필요하다. 이런 규칙은 학습 말뭉치가 없어도 정의된 규칙으로 의존관계를 결정할 수 있고 처리속도 또한 빠른 장점이 있다. 하지만 모든 문장을 정확히 분석할 수 있는 규칙을 정의하기 어렵고, 예외 규칙이 추가됨에 따라 기존 규칙과의 상관관계를 직접 판단해야 하므로 정의된 규칙의 관리가 어려운 단점이 있다.

[0006] 다른 예로 전이 기반 의존관계 분석이 있다. 전이 기반 의존관계 분석 모델은 'Joakim Nivre'에 의해 제안되었으며 비교적 단순한 알고리즘으로, 전이 기반 모델을 이용한 한국어 의존 관계 분석 연구로는 CoNLL-U(the conference on computational natural language learning) 말뭉치를 이용하여 arc-eager 알고리즘을 사용한 연구가 있다. 이 연구에서 자질(feature)의 학습 알고리즘으로 SVM(support vector machine)을 사용하여 실험하였고 2,532 문장의 말뭉치 사용하였다. 실험 결과는 사용한 말뭉치 특성상 말뭉치의 규모가 작고 오류 비율이 높으며 어절 단위의 제한된 자질 정보만을 사용하여 63.39%의 낮은 의존관계 정확률을 보여주었다.

[0007] 또 다른 예로 딥 러닝 기반의 의존관계 분석이 있다. 딥 러인 기반의 방법은 arc-eager 알고리즘의 oracle 역할로 피드 포워드(feed-forward) 신경망을 사용하고 NNLM(neural network language model) 및 Word2Vec 등의 임베딩을 활용한 연구, RNN(recurrent neural network) 모델과 LSTM(long short-term memory) 셀을 이용한 연구, end-to-end 신경망 모델을 사용하여 attention 기법을 활용하는 연구, 전이 기반 알고리즘을 신경망 셀에 결합

한 TBRU(transition based recurrent unit)을 활용한 연구, GRU(gated recurrent unit) 셀을 사용한 포인터 네트워크 모델을 사용한 연구, biaffine attention 기법을 적용한 연구 등이 있고, 이들 방법들은 한국어에서 약 90% 대의 높은 의존관계 정확률을 보이는 반면, 학습 및 처리를 위해 요구되는 시간이 기계학습 방법에 비해 크고 오류분석에 있어서 원인을 정확히 분석하기 어려워 간접적으로 모델을 수정하여 재학습 후 결과를 확인하는 경험적 오류 분석이 요구되는 단점이 있다.

발명의 내용

해결하려는 과제

[0008] 따라서 본 발명은 형태소의 표층형의 재현율 및 의존관계 정확률을 높이기 위해 기 구축된 어휘 의미망(UWordMap)을 활용하여 명사의 표층형을 추상화하여 자질로 사용하는 의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템을 제공하는 것이다.

[0009] 그리고 이러한 전이 기반 한국어 의존관계 분석 시스템은 표층형의 추상화를 위하여 학습 말뭉치를 동형이의어 분별을 하였고, 동형이의어 분별된 정보를 기반으로 어휘 의미망의 단어 계층에 따라 명사를 추상화하는 방안 및, 분석 및 학습 속도를 고려하여 전이 기반 분석 모델을 적용하였다.

과제의 해결 수단

[0010] 이와 같은 목적을 달성하기 위한 본 발명은, 입력된 문장의 형태소를 분석하는 형태소 분석모듈; 분석된 형태소 정보에서 체언류의 표층형의 추상화된 정보를 데이터베이스로부터 획득하는 체언 추상화 모듈; 및 상기 체언의 추상화된 표층형 정보와 학습 말뭉치로부터 학습된 자질 정보를 이용하여 어절 간 의존관계를 분석하는 의존관계 분석 모듈을 포함하는 의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템을 제공한다.

[0011] 상기 형태소 분석모듈은, 상기 입력된 문장의 형태소 분석 결과와 동형이의어 분별 결과를 제공하며, 상기 형태소 분석결과는 형태소 단위로 표층형, 동형이의어 번호, 다의어 번호, 품사 태그가 부착되어 출력되고, 상기 분석결과로부터 입력 문장의 어절 단위와 형태소 단위로 표층형과 동형이의어 번호, 품사 태그 정보를 수집한다.

[0012] 상기 체언 추상화 모듈은, 어휘 의미망을 활용하여 명사의 표층형을 추상화하여 자질로 사용하며, 상기 어휘 의미망은 명사의 계층 구조를 트리 구조로 제공하며, 상위어 정보를 통해 명사의 표층형 정보를 추상화한다.

[0013] 상기 표층형의 추상화는, 학습 말뭉치를 동형이의어로 분별하였고, 동형이의어로 분별된 정보를 기초로 어휘 의미망의 단어 계층에 따라 명사를 추상화한다.

[0014] 상기 명사의 표층형 추상화는, 표층형 자질, 추상화 자질을 포함한 어절단위 자질의 기본자질, 형태소 단위 및 음절 단위 확장 자질을 모두 학습한다.

[0015] 상기 체언 추상화 모듈은, 입력 문장의 형태소 단위의 품사 태그를 확인하여 체언류 형태소를 구분하고 체언류 형태소에 대하여 상기 어휘 의미망에 해당 단어 정보가 있는지 확인하고, 상기 어휘 의미망에 없는 단어는 추상화하지 않고 정보가 있는 단어는 어휘 의미망의 단어 계층을 계산하고, 상기 동형이의어를 기준으로 동형이의어가 갖는 모든 다의어에 대하여 단어 계층을 모두 계산하고, 계산된 모든 다의어에 대하여 설정된 추상화 레벨만큼 상위 계층 단어로 추상화하며, 추상화된 단어 중 빈도가 가장 높은 단어를 해당 동형이의어의 추상화된 단어로 결정하는 것을 특징으로 한다.

[0016] 상기 체언 추상화는, 추상화의 단계가 최고 단계를 초과하는 경우, 추상화 단계가 어휘 의미망의 최상위 단계 이상이 되면 루트(root) 노드 아래의 1계층 단어로 추상화하는 제1 방법과, 추상화를 하지 않고 원형을 그대로 사용하는 제2 방법 중 하나에 의해 수행된다.

[0017] 상기 의존관계 분석모듈은, 한국어의 특징이 고려된 전이 기반 모델에 맞춰 입력 문장의 순서를 역순으로 입력 버퍼에 입력하고, 자질 함수를 통해 각 자질의 값을 구하여 전이(transition)을 결정하며, 상기 전이결정에 따라 상기 입력 버퍼의 모든 어절이 의존관계가 결정되어 입력 버퍼에 존재하는 어절이 없으면 의존관계 분석을 종료하고 입력 문장의 각 어절의 지배소 어절의 어절 위치를 기록하여 의존관계 결과를 출력하는 것을 특징으로 한다.

[0018] 상기 의존관계 분석에 따른 기본 자질 성능 평가에서 일반화된 품사 태그 자질(generalized POS)이 가장 높은 재현율을 제공하며, 표층형과 동형 이의의 번호, 품사 태그 합 자질(BPS)이 가장 높은 정확율을 제공한다.

[0019] 상기 의존관계 분석에 따른 확장 자질 성능 평가에서, 확장 자질의 재현율과 정확률은 상기 기본 자질에 비해 재현율은 높고, 정확률이 낮다.

[0020] 상기 의존관계 분석에 따른 확장 자질 성능 평가에서, 형태소 자질과 음절 자질을 모두 사용한 경우가 가장 높은 의존관계 정확율을 제공한다.

[0021] 상기 의존관계 분석 방법은, 한국어의 지배소 후위 원칙을 이용한 수정된 backward arc-eager 알고리즘을 사용하여 전이 기반 의존관계를 분석하고, 상기 수정된 backward arc-eager 알고리즘은 기본 arc-eager 알고리즘에서 right-arc, shift transition은 제거되고 left-arc transition에서 stack의 어절을 제거하지 않고 buffer로 이동시키는 방법으로 이루어진다.

[0022] 상기 의존관계 분석에 따른 의존관계정확률 및 자질 정확률은 다음 식에 의해 계산된다.

[0023] $$\text{의존관계 정확률} = \frac{\text{정답과 같은 의존관계 결정 어절 수}}{\text{전체 어절 수}},$$

[0024] $$\text{자질의 정확률} = \frac{\text{정답과 같은 } transition \text{ 결정 수}}{\text{자질의 재현 횟수}}$$

발명의 효과

[0025] 이상과 같은 본 발명의 의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템에 따르면, 기 구축된 어휘 의미망을 활용하여 명사의 표층형을 추상화하여 자질로 사용되는 의존관계 분석방법을 제안함으로써, 형태소의 표층형의 재현율 및 의존관계 정확률을 향상시키고 있다. 또한 표층형의 추상화를 위해 학습 말뭉치를 동형이의어 분별을 하였고, 동형 이의어 분별된 정보를 기반으로 어휘 의미망의 단어 계층에 따라 명사를 추상화하였다. 또한 분석 및 학습 속도를 고려하여 전이기반 분석모델을 사용하였고, 한국어의 특징에 맞게 변형하였다.

[0026] 그 결과, 명사의 표층형 추상화를 통해 표층형 자질만을 사용하여 평가 시 최대 7.55%의 의존관계 정확률 상승을 보였고 추상화 자질을 포함한 어절 단위 자질로 기본 자질을 학습한 결과 90.43% 정확률과 1,002문장/sec 처리 속도, 1,294 문장/sec 학습 속도를 보였다. 형태소 단위 및 음절 단위 확장 자질을 모두 학습할 경우 90.75% 의존관계 분석 정확률과 631 문장/sec, 처리 속도, 562 문장/sec 학습 속도를 보였다. 또한 규칙 기반 연구와 비교하여 정확률, 학습 속도, 처리 속도 모두 향상된 결과를 확인할 수 있다.

도면의 간단한 설명

[0027] 도 1은 본 발명의 실시 예에 따른 의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템의 구성도

도 2는 본 발명을 설명하기 위해 제공된 기본적인 전이 기반의 의존관계 분석모듈(dependency parser)의 기본 구조

도 3 내지 도 11은 본 발명의 실험 및 성능평가를 나타낸 각종 지표를 그래프로 나타낸 도면들로서,

도 3은 체언 추상화 단계별 정확률 및 사용률을 보인 그래프

도 4는 '범위 초과시 1단계 추상화 방법'의 추상화 단계 별 자질 재현율을 보인 그래프

도 5는 '범위 초과시 1단계 추상화 방법'의 추상화 단계 별 의존관계 정확률을 보인 그래프

도 6은 '범위 초과시 1단계 추상화 방법'의 추상화 단계 별 자질 정확률을 보인 그래프

도 7은 '범위 초과시 원형 사용방법'의 추상화 단계별 의존관계 정확률을 보인 그래프

도 8은 '범위 초과시 원형 사용방법'의 추상화 단계별 자질 정확률을 보인 그래프

도 9는 '범위 초과시 원형 사용방법'의 추상화 단계별 자질 정확률을 보인 그래프

도 10은 학습 자질별 의존관계 정확률의 비교 그래프

도 11은 학습 자질별 학습 및 처리속도의 비교 그래프

발명을 실시하기 위한 구체적인 내용

[0028] 본 발명의 목적 및 효과, 그리고 그것들을 달성하기 위한 기술적 구성들은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시 예들을 참조하면 명확해질 것이다. 본 발명을 설명함에 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.

[0029] 그리고 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다.

[0030] 그러나 본 발명은 이하에서 개시되는 실시 예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있다. 단지 본 실시 예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0031] 이하에서는 도면에 도시한 실시 예에 기초하면서 본 발명에 대하여 더욱 상세하게 설명하기로 한다.

[0032] 본 발명에 사용되는 어휘의 정보는 표준국어대사전을 기반으로 다의어 수준의 어휘 의미망으로 구축한 울산 대학교의 'UWordMap'을 사용하며, 상기 'UWordMap'은 명사, 용언, 부사 등의 어휘들이 상호 연결된 어휘망을 말한다. 그리고 'UWordMap'은 명사의 계층 구조를 트리 구조로 담고 있어 명사의 상위어 또는 하위어 정보를 얻을 수 있다. 상위어 정보는 대상 단어보다 큰 범주의 의미를 갖는 단어이므로 상위어 정보를 통해 명사의 표층형 정보를 추상화시킬 수 있다. 트리 구조를 이용해 최상위 노드를 기준으로 몇 번째 노드로 선택함에 따라 표층형의 추상화 계층을 결정할 수 있고 현재 노드를 기준으로 몇 계층을 추상화할 것인지 결정할 수 있다.

[0033] 예를 들어, 'UWordMap'에 포함된 '사과_05'(apple)의 표층형은 다음 표 1과 같이 추상화 레벨에 따라 1계층 단어 '물건' 또는 5계층 단어 '과일' 등으로 추상화시킬 수 있다.

표 1

[0034]

단어계층	단어 표층형
0	UWIN
1	물건
2	물체
3	물질
4	열매
5	과일
6	사과

[0035] 다음 표 2는 추상화 단계에 따른 문장의 표층형이 실제 어떻게 추상화되는지 보여준 예이다. 추상화 레벨은 UWordMap의 단어 계층에서 UWIN(최상위 단어)을 기준으로 하위 단계의 계층의 수를 나타내고 원문의 단어 표층형의 레벨이 추상화 레벨보다 하위 단계의 단어일 경우 추상화 레벨의 단어로 표현하였다.

표 2

[0036]

원문	**이용료는 기본 통화료에 3분당 2백~3백원이 추가된다.**
1레벨 추상화	물건는 성질 물건에 3단위당 2단위~3단위단위이 추가된다.
2레벨 추상화	물품는 바탕 물품에 3언어단위당 2언어단위~3언어단위화폐단위이추가된다.
3레벨 추상화	재물는 본바탕 재물에 3말당 2말~3말원이 추가된다.
4레벨 추상화	돈는 근본 돈에 3단어당 2단어~3단어원이 추가된다.
5레벨 추상화	요금는 기본 요금에 3품사당 2품사~3품사원이 추가된다.

[0037] 형태소의 표층형을 추상화할 때 동일한 표층형이 다른 추상화 트리를 가질 수 있다. 이는 해당 어휘의 동형이의어 또는 다의어의 차이 때문에 표층형이 같더라도 의미적으로 다른 단어이다.

[0038] 따라서 본 실시 예서는 동형이의어 구분이 된 말뭉치를 사용하였고 테스트 문장에 동형이의어 태그가 없을 경우 형태소 분석기(UTagger)를 이용하여 부착된 동형이의어 태그를 이용하였다. 하기 표 3과 같이 표층형이 같은 사과(apple)와 사과(apology)를 구분하여 추상화할 경우 추상화 계층에 따라 다른 단어로 추상화된다.

표 3

[0039]

단어계층	단어 표층형	
0	UWIN	UWIN
1	물건	행위
2	물체	기원
3	물질	사과_08(apology)
4	열매	
5	과일	
6	사과_05(apple)	

[0040] 본 발명에 적용되는 학습 자질에 대해 살펴본다. 학습 자질은 기본 학습자질과 확장 학습자질로 구분할 수 있다.

[0041] 기본 학습 자질로는 형태소의 품사 태그, 표층형의 조합을 사용하고 스택(stack) 및 버퍼(buffer)의 탑(top) 어절의 인접한 어절과 상대 거리 등의 정보를 자질로 사용하였고, 어절 단위를 기준으로 학습하는 자질을 선택하였다. 학습 자질 중 일반화된 품사 태그는 재현율이 낮은 품사 조합을 일반화하여 재현율을 높이기 위하여 체언류의 품사를 일반화하고 접미사와 어근, 기호 중 일부를 생략하였다. 또한 장거리 의존관계(long term distance) 판단을 위해 대상 어절 사이의 동사의 수를 학습하였고 학습 대상 어절을 stack 및 buffer의 top 어절과 함께 top 어절을 기준으로 어절의 위치가 바로 이전 또는 바로 다음 어절, stack top 어절의 지배소 어절의 품사 태그를 학습하였다.

[0042] 대상 어절의 범위를 제한한 이유는 참조될 노드의 범위를 넓혀 더 많은 노드의 정보를 참조할수록 학습 정보에 노이즈가 발생하여 범위를 제한한 경우보다 낮은 의존관계 정확률을 보였기 때문이다. 기본 자질을 arc(의존관계 설정)로 설정하였고, 이렇게 한 이유는 모든 학습 자질이 재현되지 않았을 경우 확률적으로 지배 거리 확률에 의해 인접한 노드와 의존관계를 맺는 것이 정확률이 높기 때문이다.

[0043] 다음 표 4는 기본 학습자질과 일반화 품사종류의 예를 나타낸 것이다.

표 4

[0044]

기존 자질 요소	**[stack top word, buffer top word]**: generalized POS tag, body, BSP(body +sense number +POS tag), abstracted body, stack first morpheme BSP + buffer POS, stack POS + count of verb between buffer top word. **[stack previous word, stack top word]**, **[stack top word, stack next word]**, **[stack top word, stack governor word]**, **[buffer previous word, buffer top word]** : POS default feature (arc)
일반화된 POS	NNB, NNP, NP, NR, JKV, XPN, XSN, XSA, XR, SO, SE, SH, SL, SN, SW, SS (NNG로 일반화 후 중복 제거)
body:형태소의 표층형 sense number: 동형이의어 번호 POS: 형태로 품사 태그 BSP: 형태소의 표층형 + 동형이의어 번호 + 품사 태그	

[0045] 확장 학습 자질는 어절 단위가 아닌 형태소 단위와 음절 단위 자질의 조합이다. 확장 학습 자질로 기본 자질에서 품사 태그의 일반화와 표층형의 추상화를 통해 손실된 정보를 보완하기 위해 원형 그대로의 자질을 학습하였다. 확장 자질은 가능한 모든 조합을 학습하고 어절 단위의 형태소의 수 또는 어절의 수로 평균값을 계산하여 어절 단위의 확률을 계산하였다.

[0046] 표 5에 확장 학습 자질의 예를 나타냈다.

표 5

[0047]

확장 자질 요소	[**stack top word, buffer top word**]: POS combination, morpheme POS with word POS combination, body combination, morpheme body with word body combination, BSP combination, morpheme BSP with word BSP combination, syllable combination with morpheme POS
body:형태소의 표층형 sense number: 동형이의어 번호 POS: 형태로 품사 태그 BSP: 형태소의 표층형 + 동형이의어 번호 + 품사 태그	

[0048] 도 1은 본 발명의 실시 예에 따른 의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템의 구성도이다. 이를 보면, 분석 시스템(100)은 문장 입력 모듈(110), 형태소 분석 모듈(120), 체언 추상화 모듈(130), 의존관계 분석 모듈(140) 및 의존관계 출력모듈(150)을 포함하는 구성이다.

[0049] 본 실시 예에서 문장 입력 모듈(110)은 의존관계 분석을 위한 일련의 문장이 입력되는 모듈로서, 다양한 매체를 이용할 수 있을 것이다.

[0050] 본 실시 예에서 형태소 분석 모듈(120)은, 형태소 분석기(122) 예컨대, 'UTagger'를 이용하여 입력 문장의 형태소 분석 결과와 동형이의어 분별 결과를 분석한다. 이러한 분석 결과의 예는 표 6과 같을 수 있다.

표 6

[0051]

입력 문장	아름다운 영희의 동생을 보았다.
형태소 분석결과	아름답__000001/VA+ㄴ/ETM 영희/NNP+의/JKG 동생__010001/NNG+을/JKO 보__010101/VV+았/EP+다/EF+./SF

[0052] 이를 보면, 형태소 분석 모듈(120)은 형태소 단위로 표층형, 동형이의의 번호, 다의어번호, 품사 태그가 부착되어 출력됨을 알 수 있고, 따라서 입력 문장의 어절 단위와 형태소 단위로 표층형과 동형이의어 번호, 품사 태그 정보를 수집하게 된다.

[0053] 본 실시 예에서 체언 추상화 모듈(130)은, 어휘 의미망(132), 예컨대 울산대학교의 'UWordMap'의 정보를 이용하여 체언류의 표층형의 추상화한 정보를 획득한다. 입력 문장의 형태소 단위의 품사 태그를 확인하여 체언류 형태소를 구분하고 체언류 형태소에 대하여 'UWordMap'에 해당 단어 정보가 있는지 확인하다.

[0054] 이러한 확인 결과에 따라 'UWordMap'에 속하지 않는 단어에 대해서는 추상화하지 않고 정보가 있는 단어에 대하여 'UWordMap'의 단어 계층을 계산한다. 반면 해당 단어가 있으면, 동형이의어를 기준으로 추상화를 수행하므로, 동형이의어가 갖는 모든 다의어에 대하여 단어 계층을 모두 계산한다. 이는 하나의 동형이의어가 2개 이상의 다의어를 갖을 수 있기 때문이다. 그리고 모든 다의어에 대하여 설정된 추상화 레벨만큼 상위 계층 단어로 추상화하고 추상화된 단어 중 빈도가 가장 높은 단어를 해당 동형이의어의 추상화된 단어로 결정한다.

[0055] 의존관계 분석 모듈(140)은, 의존관계 학습정보 db(142)을 이용하여 상기 체언 추상화 모듈(130)에서 체언의 추상화된 표층형 정보와 학습 말뭉치로부터 학습된 자질 정보를 이용하여 의존관계를 결정한다. 그리고 의존관계 분석 모듈(140)의 분석 과정은, 한국어의 특징이 고려된 전이 기반 모델에 맞춰 입력 문장의 순서를 역순으로 입력 버퍼(buffer)에 넣고, 자질 함수를 통해 각 자질의 값을 구하여 전이(transition)을 결정한다. 그리고 전이(transition) 결정에 따라 버퍼의 모든 어절이 의존관계가 결정되어 버퍼에 존재하는 어절이 없으면 의존관계 분석을 종료하고 입력 문장의 각 어절의 지배소 어절의 어절 위치를 기록하여 의존관계 결과를 출력한다.

[0056] 도 2는 본 발명을 설명하기 위해 제공된 기본적인 전이 기반의 의존관계 분석모듈(dependency parser)(140)의 기본 구조이다. 도시된 바와 같이, 스택(stack)(144) 및 입력버퍼(input buffer)(145)가 구비되며, 오라클 함수로부터 전이(transition)을 수행하도록 파서(146)를 포함하고 있다. 그리고 각 단계의 전이를 결정하는 오라클 함수는 하기 수학식 1을 통해 학습 자질별 가중치가 반영된 확률의 합으로 결정한다.

수학식 1

[0057] $$f(x) = \left(\frac{x_{dependency\ count}}{x_{total\ action\ count}} - \frac{x_{half\ of\ total\ action\ count}}{x_{total\ action\ count}}\right) * log_{10}(x_{total\ action\ count} + 1 + \alpha)$$

[0058] 여기서, $\alpha > 0$ and $\alpha = minimum\ constant$,

[0059] $$Y_{oracle} = \tanh\left(\sum_{i=1}^{n} \omega_i f(x_i)\right) \quad \begin{matrix} if\ y \geq 0\ dependency\ relationship \\ else\ no\ dependency \end{matrix}$$

[0060] 자질 함수 f(x)는 학습 말뭉치에서 학습된 자질의 전체 action 중 arc action의 비율에 전체 action의 절반 값을 감소시켜 0 ~ 1의 비율 범주를 0.5 ~ -0.5의 범주로 변화하였다. 따라서 arc count의 비율이 양수일 경우 arc action의 확률이 reduce action의 확률 보다 높고 음수일 경우 반대가 된다. 그리고 action의 빈도수에 따라 낮은 빈도의 확률 값은 자질들의 확률의 합에서 낮은 영향을 주기 위해 전체 action count에 로그를 취한 값을 곱하여 전이 확률을 계산하였다. 로그함수를 취할 때는 부호의 역전을 방지하기 위해 1 이상의 최솟값 α을 합한다.

[0061] 그리고 action의 판단은 오라클 함수의 결과값으로부터 각 자질의 가중치를 곱한 자질의 확률 값을 쌍곡선 탄젠트(hyperbolic tangent) 함수로 정규화하여 최종 값을 -1 ~ 1의 범주로 변화시켰다. 결과값이 양수일 경우 left arc action을 통해 의존관계를 결정하고 음수일 경우 reduce action으로 의존관계 아님을 결정하였다.

[0062] 각 자질 별 가중치는 학습 말뭉치의 문장 순서를 임의의 순서로 변경하고 반복 학습하여 결정한다. 각 자질의 가중치 변경 값은 하기 수학식 2와 같이 정답 값 CA(0.5 또는 -0.5)와 oracle 결과값의 오차 값 E를 구하고 각 자질의 현재 가중치에 오차 값과 oracle 결과값의 역함수, 자질 결과의 값, 학습 비율 α를 곱하여 결정하여 오차에 대한 각 가중치의 변경 값을 결정하였다.

수학식 2

[0063] $$E = CA - \frac{Y_{oracle}}{2}$$

[0064] $$\omega_{n+1} = \omega_n + E * \tanh^{-1}(Y_{oracle}) * f(x_n) * \alpha$$

[0065] 한편, 본 실시 예에는 한국어의 지배소 후위 원칙을 고려한 수정된 backward arc-eager 알고리즘을 사용하여 전이 기반 의존관계를 분석하였다. 따라서 표 7과 같이 기본 arc-eager 알고리즘에서 right-arc, shift transition이 제거되고 left-arc transition에서 stack의 어절을 제거하지 않고 buffer로 이동시킨다.

표 7

[0066]

Start state	$([ROOT], [n, \ldots, 1], \{\})$
Final state	$(S, [\,], A)$
Transitions	
$\text{Left} - \text{Arc}_l^e$	$:= (\sigma\|i, j\|\beta, A) \Rightarrow (\sigma\|i\|j, \beta, A \cup \{(i, l, j)\})$
Reduce	$:= (\sigma\|i, \beta, A) \Rightarrow (\sigma, \beta, A)$
Preconditions	
$\text{Left} - \text{Arc}_l^e$	$\neg\exists k \exists l'[(k, l', j) \in A]$
Reduce	$\exists k \exists l[(k, l, i) \in A]$

[0067] 그리고 상기 backward arc-eager 알고리즘을 사용하기 위해서는 입력 문장 $x = (w_0, w_1, \ldots w_{n-1}, w_n)$ 을

$x' = (w_n, w_{n-1}, \ldots, w_1, w_0)$ 와 같이 역순으로 변경해야 할 것이고, 표 8과 같이 나타낼 수 있다.

표 8

[0068]

입력 버퍼				입력 버퍼		
[0]	[1]	[2]		[0]	[1]	[2]
나는	밥을	먹었다.	⇒	먹었다.	밥을	나는

[0069] 다음에는 본 발명의 성능 평가에 대하여 살펴본다. 성능 평가는 표층형 추상화 레벨에 따른 학습 자질의 재현율과 정확률을 분석하였고, 기본자질 성능, 확장자질 성능 등을 분석하였다.

[0070] 성능 평가를 위해 본 실시 예에서, 실험 및 학습에 사용된 말뭉치는 21세기 세종 계획 구문 분석 말뭉치를 의존 문법으로 변환한 말뭉치를 사용하였고 동형이의어 태그를 추가로 부착하여 사용하였다. 또한 의존관계 라벨은 사용하지 않고 구문 분석 시 의존관계의 분석만을 진행하였다. 따라서 정확률 평가에서는 UAS(unlabeled attachment score) 평가만 수행한다. 의존관계 분석의 정확률 외에 학습 자질에 따른 학습 속도와 테스트 속도 또한 함께 평가하여 의존관계 정확률과 속도를 함께 성능 평가하였다.

[0071] 또 상기 말뭉치는 띄어쓰기 오류로 인한 어절 수가 불일치하는 문장과 교차 지배관계가 있는 문장, 문장 중간에 자기 순환 지배관계 어절이 포함된 문장 등을 대상에서 제외하였다. 정제된 문장은 57,265문장으로 총 561,645 어절로 구성되어 있다. 성능 평가는 10-fold cross-validation 방법으로 전체 말뭉치의 10%를 테스트 대상으로 하고 나머지를 학습하는 과정을 반복하였다. 의존관계 정확률은 하기 수학식 3과 같이 전체 말뭉치에서 올바르게 분석된 의존관계 어절 수를 전체 어절 수로 나누어 계산하였다.

수학식 3

[0072]
$$\text{의존관계 정확률} = \frac{\text{정답과 같은 의존관계 결정 어절 수}}{\text{전체 어절 수}}$$

[0073] 그리고 자질의 정확률은 하기 수학식 4와 같이 의존관계 결정에 자질이 사용될 경우 자질의 의존관계 정확률을 판단하기 위해 정답 의존관계 결정을 위한 transition과 같은 자질의 transition 결정 수를 자질의 재현 횟수로 나누어 계산하였다.

수학식 4

[0074]
$$\text{자질의 정확률} = \frac{\text{정답과 같은 } transition \text{ 결정 수}}{\text{자질의 재현 횟수}}$$

[0075] 한편, 실험을 수행한 시스템 환경은 표 9와 같다.

표 9

[0076]

운영체제	Windows 10
CPU	Intel®Core™i7-5820K @ 3.30 GHz
RAM	32 GB
개발 언어 및 도구	C#, .Net Framework 4.5.2, Visual Studio 2017

[0077] 이와 같은 시스템을 통한 성능 실험결과에 대해 설명한다.

[0078] 첫 번째로, 어휘의 의미적 추상화 성능 실험결과이다. 학습 말뭉치에서 학습 자질로 표층형의 정보만을 학습하여 의존관계 분석을 수행하면, 표층형 추상화 레벨에 따른 학습 자질의 재현율과 정확율을 비교할 수 있다.

[0079] 비교결과, 자질의 사용률에 따라 의존관계 분석 정확률이 비례적으로 증가하였고 추상화 정도에 따라 반비례 관계로 자질 정확률은 하락하였음을 알 수 있었다. 예를 들어 표 10과 같이 체언의 표층형을 UWordMap의 단위 계

층 1레벨로 추상화할 경우 손실된 정보로 인해 표층형 자질의 정확률은 2.92% 감소하지만 자질의 재현율이 36.75% 상승하여 결과적으로 의존관계 정확률은 7.55%가 향상되었다.

표 10

[0080]

	의존관계 정확률	표층형 자질 재현율	자질의 정확률
원문 표층형	73.95%	18.33%	98.26%
abstraction 1 level	81.50%	55.08%	93.34%
△	+7.55%	+36.75%	-2.92%

[0081] 도 3은 본 발명의 실험 결과에서 체언 추상화 단계별 정확률 및 사용률의 변화를 나타낸 도면이다. 이를 보면, 추상화 단계가 진행될 수록 자질 정확률은 다소 올라가지만 의존관계 정확률과 자질 재현율은 감소하고 있음을 알 수 있다.

[0082] 체언의 추상화 방법은 UWordMap의 단어 계층을 기준으로 상대적으로 추상화하였고 추상화 단계가 최고 단계를 초과하는 경우 처리 방법에 따라 두 가지 방법으로 나눌 수 있다. 만약 추상화 단계가 UWordMap의 최상위 단계 이상이 될 경우 root 노드(UWIN) 아래의 1계층 단어로 추상화하는 '범위 초과 시 1단계 추상화 방법'과 추상화를 하지 않고 원형을 그대로 사용하는 '범위 초과 시 원형 사용 방법'이 있다.

[0083] 추상화 단계 처리 방법에 따른 의존관계 정확률 및 자질의 재현율은 범위를 초과할 경우, 도 4 내지 도 6에 도시한 바와 같이 최상위 아래 단어로 추상화하는 방법은 추상화 단계에 따라 자질의 재현율과 의존관계 정확률은 비례 관계이고 자질 정확률은 반비례 관계를 보였다.

[0084] 그리고 '범위 초과 시 원형 사용 방법'은 도 7 내지 도 9에 도시한 바와 같이 추상화 레벨이 증가함에 따라 최상위 단계를 초과하여 추상화되지 않는 단어가 증가하고 있어 의존 관계 정확률 및 자질 재현율은 상승 후 감소하는 패턴을 보인다.

[0085] 두 번째, 기본 자질 성능 평가이다.

[0086] 기본 자질을 이용하여 의존관계 정확률, 처리 속도, 학습 속도 등을 계산하고 10-fold cross validation을 통한 평균값으로 성능을 평가한다. 표 11과 같이 규칙 기반 방법으로 구축된 의존관계 분석기를 대상으로 의존관계 정확률과 처리 속도를 함께 비교하였다. 두 모델 모두 안정된 정확률과 처리 속도를 보였고 규칙 기반 모델보다 학습 기반의 제안 모델이 의존관계 정확률과 처리 속도 모두 향상된 것을 보여준다.

표 11

[0087]

10-fold index	전이 기반 제안 시스템			규칙 기반 분석기	
	정확률 (%)	처리속도(sec)	학습속도(sec)	정확율(%)	처리속도(sec)
1	90.62	5.507	41.110	88.23	14.653
2	90.26	7.439	39.731	88.20	12.103
3	90.37	5.589	40.606	88.43	9.321
4	90.54	6.671	41.612	88.21	11.336
5	90.57	5.314	39.235	88.66	11.488
6	90.27	5.282	38.335	88.50	11.240
7	90.32	5.363	40.556	88.53	11.139
8	90.37	5.362	38.830	88.07	9.073
9	90.50	5.267	40.315	88.13	11.234
10	90.44	5.386	38.029	88.48	11.348
평균	90.43	5.718	39.836	88.34	11.293
△	+2.09	-5.575 (-97.5%)			

[0088] 그리고 자질 별로 재현율과 정확률을 비교하면 표 12와 같이 일반화된 품사 태그 자질(generalized POS)이 가장 높은 재현율을 보였고 표층형과 동형이의어 번호, 품사 태그 합 자질(BSP)이 가장 높은 정확률을 보여준다.

표 12

[0089]

자질 종류	재현율	정확율
generalized POS	**98.87%**	91.49%
body	14.97%	94.64%
abstracted body	19.14%	94.10%
BSP	14.71%	**94.96%**
stack first morpheme BSP + buffer POS	20.63%	94.05%
stack POS + count of verb between buffer top.	98.76%	91.69%
(stack prev, stack top) POS	97.33%	91.76%
(stack top, stack next) POS	82.81%	90.87%
(stack top, stack gov) POS	83.48%	90.95%
(buffer prev, buffer top) POS	79.67%	91.82%

[0090] 세 번째, 확장자질 성능 평가에 대해 살펴본다.

[0091] 확장 자질은 기본 자질에 확장 자질을 추가하여 성능 평가를 수행하였다. 확장 자질은 형태소 단위 조합과 음절 단위 조합 각각을 비교하고 모든 조합 자질을 사용하여 성능을 평가하였다. 형태소 단위 조합 자질, 음절 단위 조합 자질 모두 정확률에 향상이 있었으나 향상률은 크지 않았다. 확장 자질의 조합은 가능한 모든 조합에 대하여 학습하였다. 형태소 단위 또는 음절 단위 자질의 자질 함수의 값은 모든 조합의 결과의 값을 합하여 buffer top 어절의 자질 수로 나누어 어절의 자질 함수의 값을 계산하였다.

[0092] 다음 수학식 5는 확장 자질의 어절 자질 계산 수학식이고, 표 13은 확장 자질 사용에 따른 성능 비교 표이다.

수학식 5

[0093]

$$\text{어절의 자질 값} = \frac{\text{전체 조합의 자질 함수의 결과값의 합}}{\text{buffer 어절의 형태소 수(또는 음절 수)}}$$

표 13

[0094]

10-fold index	기본 자질 + 형태소 단위 조합 자질			기본 자질 + 음절 단위 조합 자질		
	정확율(%)	처리속도 (sec)	학습속도 (sec)	정확율(%)	처리속도 (sec)	학습속도 (sec)
1	90.98	8.49	73.38	90.90	6.00	48.49
2	90.57	8.32	80.05	90.52	5.65	50.42
3	90.63	8.52	78.99	90.54	5.57	47.78
4	90.86	8.37	80.63	90.77	5.51	49.54
5	90.80	8.47	83.12	90.75	5.60	47.95
6	90.54	8.83	81.09	90.49	5.60	48.33
7	90.63	8.54	81.55	90.62	5.68	46.45
8	90.65	8.60	82.48	90.56	5.69	50.13
9	90.83	8.68	80.44	90.69	7.71	48.45
10	90.76	11.11	77.90	90.70	5.77	49.12
평균	90.73	8.79	80.46	90.65	5.88	48.67
△	+0.30	+3.08sec (+53.8%)	+40.63sec (+101.98%)	+0.23%	+0.16sec (+2.78%)	+8.83sec (+22.16%)

[0095] 반면, 표 14의 확장 자질별 성능을 참고하여 확장자질의 재현율과 정확율을 살펴보면, 기본 자질에 비해 높은 재현율을 보이지만 정확율은 떨어짐을 알 수 있다.

표 14

[0096]

자질 종류	재현율(%)	정확율(%)
syllable combination with morpheme POS	90.30	82.13
morpheme POS combination	99.99	82.19
morpheme POS with word POS	98.58	82.87
morpheme body combination	71.42	82.52
morpheme body with word body	71.42	88.62
morpheme BSP combination	68.84	90.94
morpheme BSP with word BSP	14.82	90.93
morpheme BSP with word POS	83.56	86.05

[0097] 그리고 표 15의 통합자질 사용 성능을 살펴보면, 형태소 자질과 음절 자질을 모두 사용할 경우 가장 높은 의존관계 정확률을 보이나 향상률을 높지 않게 나타났다.

표 15

[0098]

10-fold index	기본 자질 + 형태소 조합 자질 + 음절 자질		
	의존관계 정확율(%)	처리속도(sec)	학습속도(sec)
1	91.01	9.37	94.00
2	90.59	9.00	94.68
3	90.62	9.05	91.16
4	90.87	8.88	90.93
5	90.80	9.05	91.14
6	90.58	8.96	90.13
7	90.71	8.93	91.18
8	90.69	9.10	90.96
9	90.84	9.06	91.13
10	90.83	9.33	91.24
평균	90.75	9.07	91.66
△	+0.33%	+3.35sec(+58.66%)	+51.82sec(+130.08%)

[0099] 표 16을 자질 사용에 다른 성능 비교결과이다. 자질 선택에 따른 의존관계 정확률, 학습 및 처리 속도를 비교하여 보았다. 기본 자질은 정확률은 가장 낮았지만 속도 면에서 가장 짧은 시간을 소요하였다. 가장 높은 정확률은 모든 자질을 사용한 통합 자질이고 속도는 가장 느리다는 것을 알 수 있다.

표 16

[0100]

사용 자질	의존관계 정확율(%)	처리속도(sec)	학습속도(sec)
기본자질	90.46	5.72	39.84
기본+음절자질	90.65	5.88	48.67
기본+형태소 자질	90.73	8.79	80.46
통합자질	90.75	9.07	91.66

[0101] 상기 표 16의 사용 자질에 따른 정확률, 처리속도, 학습속도의 비교를 위한 그래프는 도 10 및 도 11에 도시하였다.

[0102] 이와 같이 본 발명은 어휘 의미망을 활용하여 명사의 표층형을 추상화한 다음 자질로 사용하는 의존관계 분석 시스템을 제안하여, 종래 의존관계 분석방법들에 비해 정확률, 학습속도, 처리속도 등을 개선하였음을 알 수 있다. 그리고 이러한 본 발명은 자동 번역기, 챗봇, 자동 Q&A 등 다양한 프로그램에 응용할 수 있을 것이다.

[0103] 이상과 같이 본 발명의 도시된 실시 예를 참고하여 설명하고 있으나, 이는 예시적인 것들에 불과하며, 본 발명이 속하는 기술 분야의 통상의 지식을 가진 자라면 본 발명의 요지 및 범위에 벗어나지 않으면서도 다양한 변형, 변경 및 균등한 타 실시 예들이 가능하다는 것을 명백하게 알 수 있을 것이다. 따라서 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적인 사상에 의해 정해져야 할 것이다.

부호의 설명

[0104] 110: 문장 입력 모듈

120: 형태소 분석 모듈

122: 형태소 분석기

130: 체언 추상화 모듈

132: 어휘 의미망

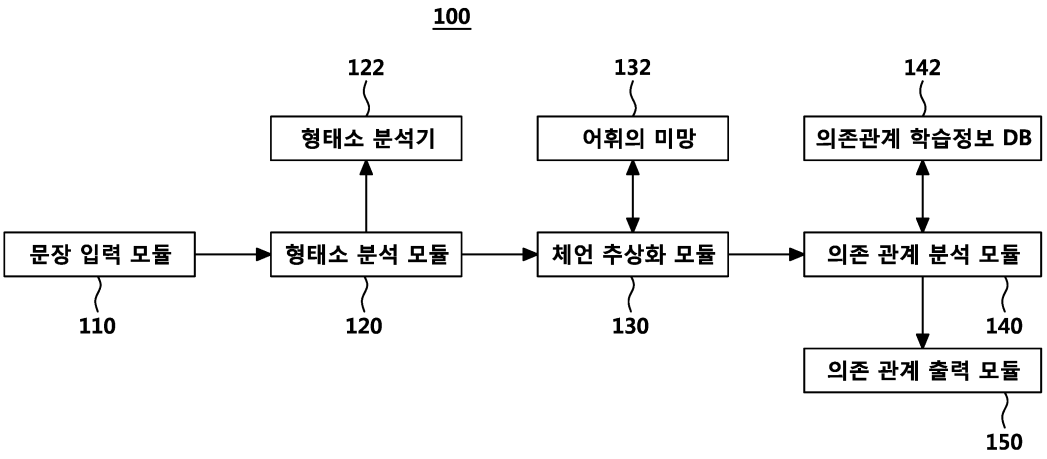
140: 의존 관계 분석 모듈

142: 의존 관계 학습정보 DB

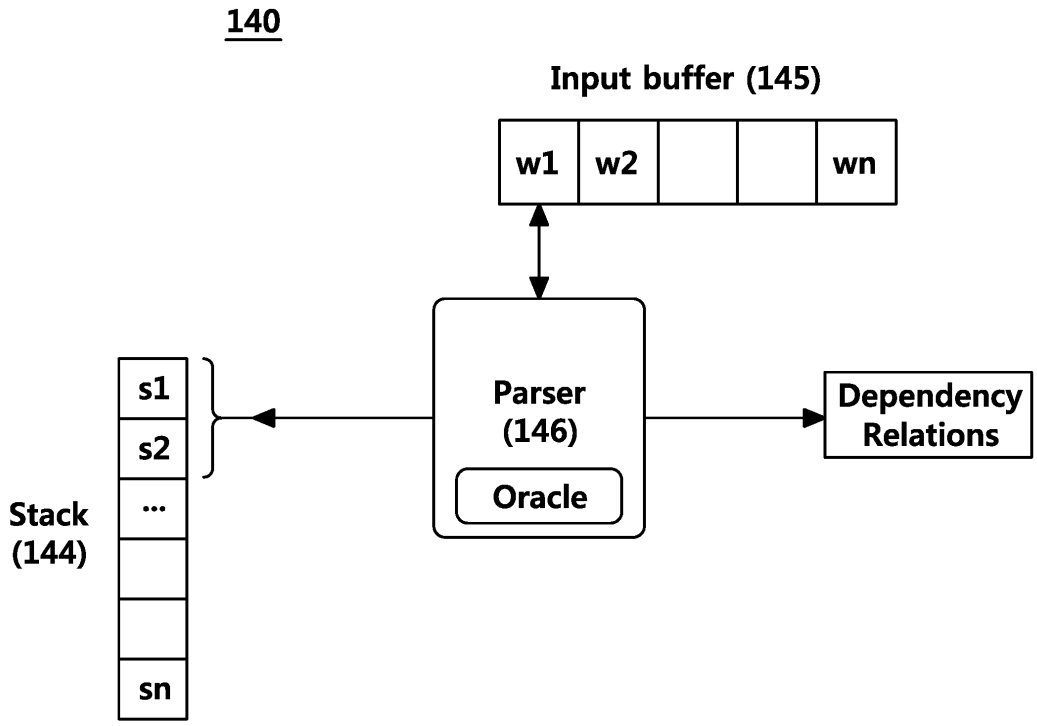
150: 의존 관계 출력 모듈

도면

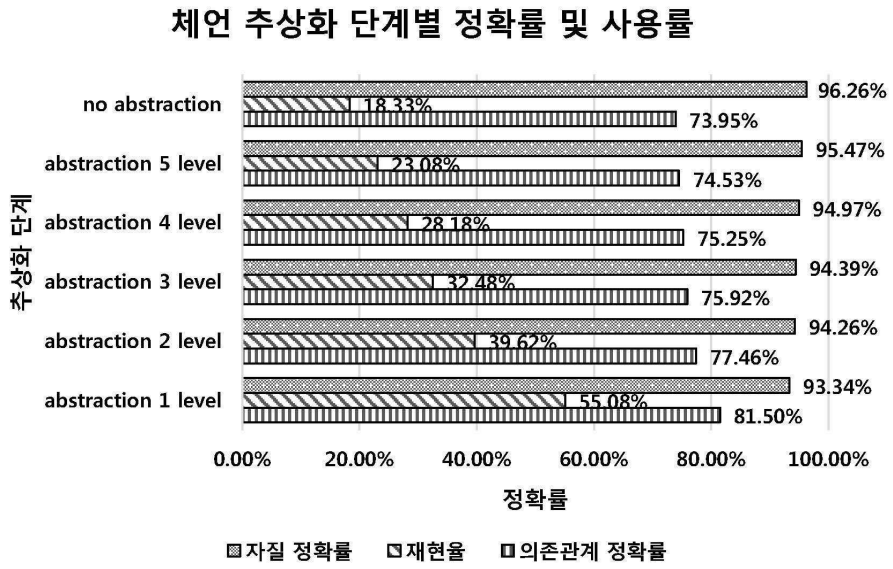
도면1



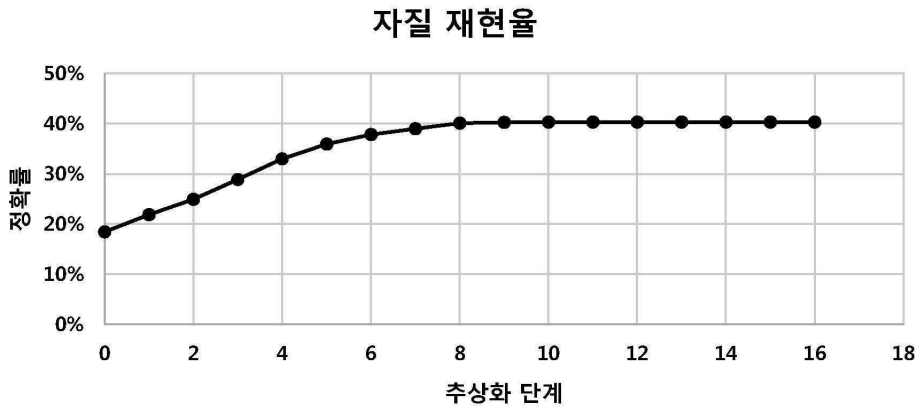
도면2


140
Input buffer (145)
w1
w2
wn
Parser
(146)
Oracle
Dependency
Relations
Stack
(144)
s1
s2
...
sn

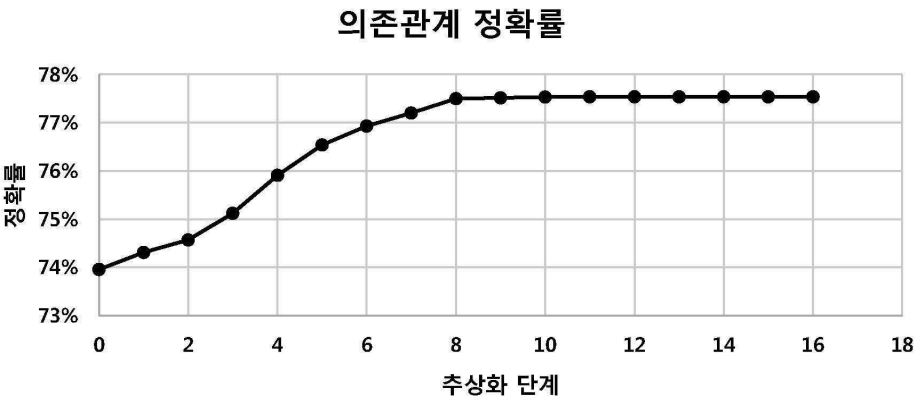
도면3


체언 추상화 단계별 정확률 및 사용률
추상화 단계
no abstraction
abstraction 5 level
abstraction 4 level
abstraction 3 level
abstraction 2 level
abstraction 1 level
18.33%
96.26%
73.95%
23.08%
95.47%
74.53%
28.18%
94.97%
75.25%
32.48%
94.39%
75.92%
39.62%
94.26%
77.46%
55.08%
93.34%
81.50%
0.00%
20.00%
40.00%
60.00%
80.00%
100.00%
정확률
자질 정확률
재현율
의존관계 정확률

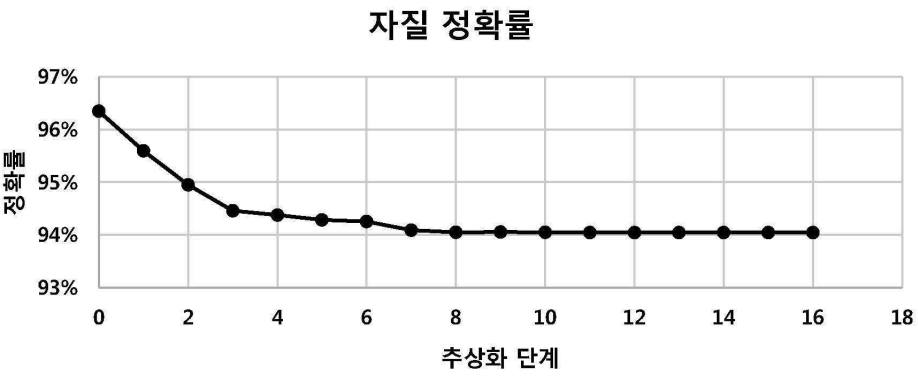
도면4


자질 재현율
50%
40%
30%
20%
10%
0%
정확률
0
2
4
6
8
10
12
14
16
18
추상화 단계

도면5


의존관계 정확률
78%
77%
76%
75%
74%
73%
정확률
0
2
4
6
8
10
12
14
16
18
추상화 단계

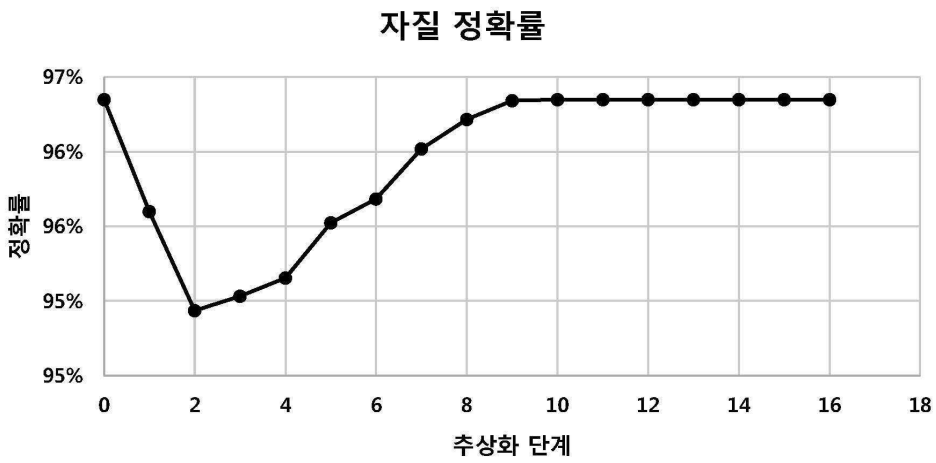
도면6


자질 정확률
97%
96%
95%
94%
93%
정확률
0
2
4
6
8
10
12
14
16
18
추상화 단계

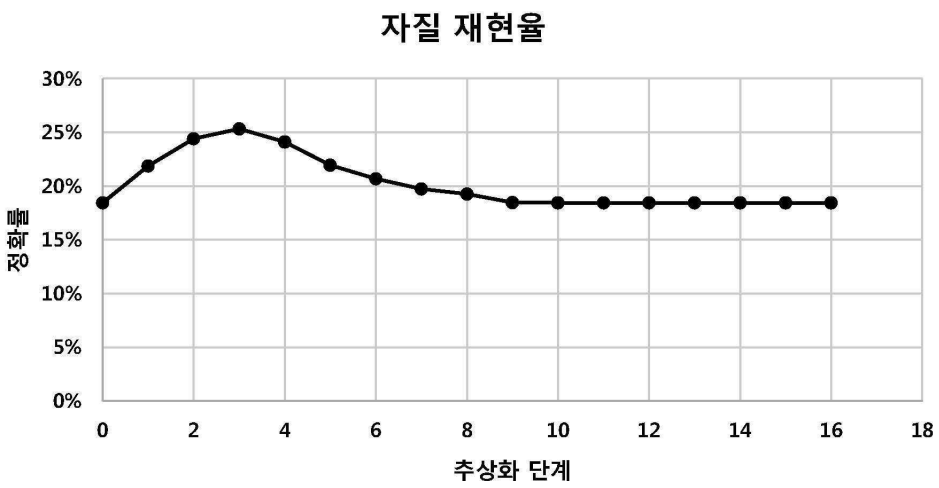
도면7


의존관계 정확률
74.80%
74.60%
74.40%
74.20%
74.00%
73.80%
정확률
0
2
4
6
8
10
12
14
16
18
추상화 단계

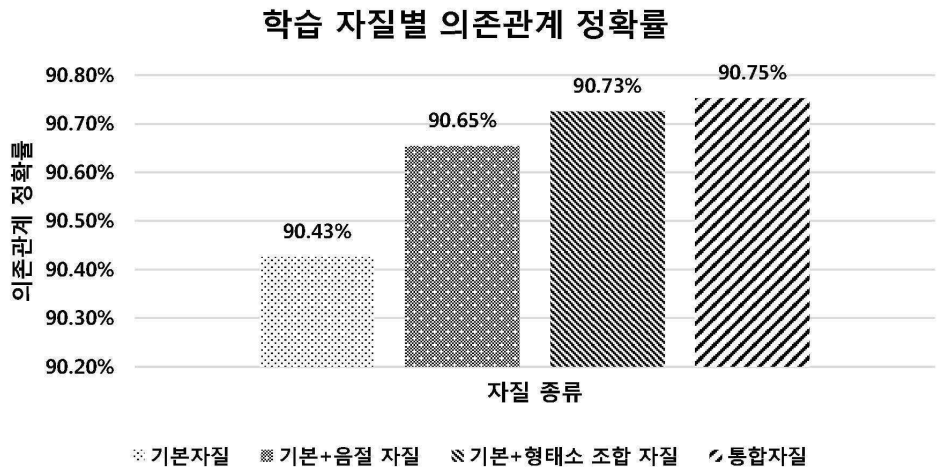
도면8


자질 정확률
97%
96%
96%
95%
95%
정확률
0
2
4
6
8
10
12
14
16
18
추상화 단계

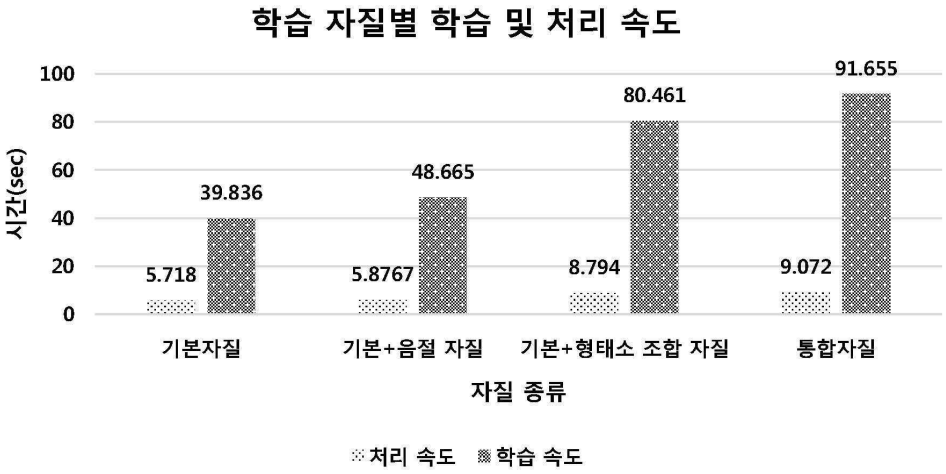
도면9


자질 재현율
30%
25%
20%
15%
10%
5%
0%
정확률
0
2
4
6
8
10
12
14
16
18
추상화 단계

도면10


학습 자질별 의존관계 정확률
90.80%
90.70%
90.60%
90.50%
90.40%
90.30%
90.20%
의존관계 정확률
90.43%
90.65%
90.73%
90.75%
자질 종류
기본자질
기본+음절 자질
기본+형태소 조합 자질
통합자질

도면11


학습 자질별 학습 및 처리 속도
100
80
60
40
20
0
시간(sec)
5.718
39.836
5.8767
48.665
8.794
80.461
9.072
91.655
기본자질
기본+음절 자질
기본+형태소 조합 자질
통합자질
자질 종류
처리 속도
학습 속도