



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년10월14일
(11) 등록번호 10-1664278
(24) 등록일자 2016년10월04일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01)

(52) CPC특허분류
G06F 17/271 (2013.01)
G06F 17/2785 (2013.01)

(21) 출원번호 10-2015-0085480

(22) 출원일자 2015년06월16일

심사청구일자 2015년06월16일

(56) 선행기술조사문헌

정현기 외, 확장된 격틀 사건을 이용한 한국어 부사격 논항의 의미역 결정, 한국정보기술학회논문지 9(10), pp.167-176, 2011.10.*

Y.B. Kim et al., Training a Korean SRL System with Rich Morphological Features, Computational Linguistics Association Annual Meeting 논문집, pp.637-642, (2014.06.23.)*

임수종 외, 순차적 레이블링을 이용한 한국어 의미역 인식, 한국정보과학회 학술발표 논문집, pp.595-597, 2014.06

이창기 외, Structural SVM 기반의 한국어 의미역 결정, 한국정보과학회 논문지 42(2), pp.220-226, 2015.02

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

한림대학교 산학협력단

강원도 춘천시 한림대학길 1, 한림대학교(옥천동)

(72) 발명자

김유섭

서울특별시 강남구 삼성로 133길 20, 101동 811호 (청담동, 삼성1차 아파트)

석미란

서울시 강동구 동남로71길 41, 현대아파트 11-803 (명일동, 현대아파트)

(74) 대리인

유민규

전체 청구항 수 : 총 13 항

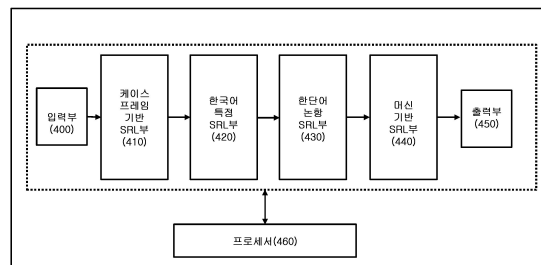
심사관 : 경연정

(54) 발명의 명칭 하이브리드 방법을 사용한SRL 기반의 문장 분석 방법 및 장치

(57) 요약

하이브리드 방법을 사용한 SRL 기반의 문장 분석 방법은 분석 대상 문장에 포함되는 복수의 논항 중 제1 논항 그룹에 포함되는 제1 논항에 대해 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계, 상기 복수의 논항 중 제2 논항 그룹에 포함되는 제2 논항에 대해 한국어 특징 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하는 단계 및 상기 복수의 논항 중 상기 제 1 논항 그룹 및 상기 제 2 논항 그룹 이외의 논항 그룹에 포함되는 적어도 하나 이상의 논항에 대해 머신(machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계를 포함할 수 있다.

대표도 - 도4



이 발명을 지원한 국가연구개발사업

과제고유번호 2010-0010612

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 일반연구자지원사업(기본연구)

연구과제명 한국어 문장의 의미 분석을 위한 언어 자원 구축 및 도구 개발

기 여 율 1/1

주관기관 한림대학교 산학협력단

연구기간 2010.05.01 ~ 2015.04.30

명세서

청구범위

청구항 1

문장 분석 장치에서 수행되는 하이브리드 방법을 사용한 SRL(semantic role labeling) 기반의 문장 분석 방법은,

분석 대상 문장에 포함되는 복수의 논항 중 제1 논항 그룹에 포함되는 제1 논항에 대해 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계;

상기 복수의 논항 중 제2 논항 그룹에 포함되는 제2 논항에 대해 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하는 단계; 및

상기 복수의 논항 중 상기 제 1 논항 그룹 및 상기 제 2 논항 그룹 이외의 논항 그룹에 포함되는 적어도 하나 이상의 논항에 대해 머신(machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계를 포함하고,

상기 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하는 단계 이후에, 상기 복수의 논항 중 제3 논항 그룹에 포함되는 제3 논항에 대해 한 단어 논항 SRL을 사용하여 시멘틱 롤을 결정하는 단계를 더 포함하되,

상기 머신 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계는, 상기 복수의 논항 중 제4 논항 그룹에 포함되는 제4 논항에 대해 2-레벨 SVM(support vector machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 것을 특징으로 하는 하이브리드 방법을 사용한 SRL 기반의 문장 분석 방법.

청구항 2

삭제

청구항 3

제1항에 있어서,

상기 제1 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제1 논항을 포함하고,

상기 제2 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가능하고 상기 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제2 논항을 포함하고,

상기 제3 논항 그룹은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 상기 제3 논항을 포함하고,

상기 제4 논항 그룹은 상기 케이스 프레임 기반 SRL, 상기 한국어 특정 접미사 구조 SRL, 상기 한 단어 논항 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가능한 나머지 논항인 상기 제4 논항을 포함하는 것을 특징으로 하는 하이브리드 방법을 사용한 SRL 기반의 문장 분석 방법.

청구항 4

제3항에 있어서,

상기 케이스 프레임 기반 SRL은 상기 제1 논항과 관련된 동사의 의미를 고려하여 프레임 파일에서 상기 의미를 가지는 상기 동사를 포함하는 예문에 포함되는 논항과 상기 동사와의 관계를 기반으로 상기 제1 논항의 시멘틱 롤을 결정하는 것을 특징으로 하는 하이브리드 방법을 사용한 SRL 기반의 문장 분석 방법.

청구항 5

제4항에 있어서,

상기 한국어 특정 접미사 구조 SRL은 한국어 특정 접미사 구조에 대한 정보를 기반으로 특정 임계값 이상의 사용 빈도(frequency)를 가지고 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 조사 또는 어미와 관련된 상기 제 2 논항의 시멘틱 롤을 상기 특정 시멘틱 롤로 결정하는 것을 특징으로 하는 하이브리드 방법을 사용한 SRL 기반

의 문장 분석 방법.

청구항 6

제5항에 있어서,

상기 한 단어 논항 SRL 은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 논항으로서 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 상기 제3 논항을 상기 특정 시멘틱 롤로 매핑하는 것을 특징으로 하는 하이브리드 방법을 사용한 SRL 기반의 문장 분석 방법.

청구항 7

제6항에 있어서,

상기 2-레벨 SVM 기반 SRL은 관련된 조사 또는 관련된 어미의 빈도가 임계 수치 이하이거나 관련된 조사 또는 관련된 어미가 특정한 시멘틱 롤로 매핑되는 임계 비율이 임계 수치 이하인 상기 제4 논항을 SVM를 기반으로 한 복수의 절차를 통해 특정 시멘틱 롤로 매핑하는 것을 특징으로 하는 하이브리드 방법을 사용한 SRL 기반의 문장 분석 방법.

청구항 8

제1항, 제3항 내지 제7항 중 어느 한 항에 따른 방법을 컴퓨터로 하여금 수행하는 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 기록매체.

청구항 9

하이브리드 방법을 사용한 SRL(semantic role labeling) 기반의 문장 분석 장치는,

분석 대상 문장에 포함되는 복수의 논항 중 제1 논항 그룹에 포함되는 제1 논항에 대해 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 케이스 프레임 기반 SRL 부;

상기 복수의 논항 중 제2 논항 그룹에 포함되는 제2 논항에 대해 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 한국어 특정 SRL 부; 및

상기 복수의 논항 중 상기 제 1 논항 그룹 및 상기 제 2 논항 그룹 이외의 논항 그룹에 포함되는 적어도 하나 이상의 논항에 대해 머신(machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 머신 기반 SRL부를 포함하고,

상기 복수의 논항 중 제3 논항 그룹에 포함되는 제3 논항에 대해 한 단어 논항 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 한 단어 논항 SRL 부를 더 포함하되,

상기 머신 기반 SRL부는, 상기 복수의 논항 중 제4 논항 그룹에 포함되는 제 4 논항에 대해 2-레벨 SVM(support vector machine) 기반 SRL을 사용하여 시멘틱롤을 결정하도록 구현되는 것을 특징으로 하는 문장 분석 장치.

청구항 10

삭제

청구항 11

제9항에 있어서,

상기 제1 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제1 논항을 포함하고,

상기 제2 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가하고 상기 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제2 논항을 포함하고,

상기 제3 논항 그룹은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 상기 제3 논항을 포함하고,

상기 제4 논항 그룹은 상기 케이스 프레임 기반 SRL, 상기 한국어 특정 접미사 구조 SRL, 상기 한 단어 논항 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가한 나머지 논항인 상기 제4 논항을 포함하는 것을 특징으로 하는 문장 분석 장치.

청구항 12

제11항에 있어서,

상기 케이스 프레임 기반 SRL은 상기 제1 논항과 관련된 동사의 의미를 고려하여 프레임 파일에서 상기 의미를 가지는 상기 동사를 포함하는 예문에 포함되는 논항과 상기 동사와의 관계를 기반으로 상기 제1 논항의 시멘틱 롤을 결정하는 것을 특징으로 하는 문장 분석 장치.

청구항 13

제 12 항에 있어서,

상기 한국어 특정 접미사 구조 SRL은 한국어 특정 접미사 구조에 대한 정보를 기반으로 특정 임계값 이상의 사용 빈도(frequency)를 가지고 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 조사 또는 어미와 관련된 상기 제 2 논항의 시멘틱 롤을 상기 특정 시멘틱 롤로 결정하는 것을 특징으로 하는 문장 분석 장치.

청구항 14

제13항에 있어서,

상기 한 단어 논항 SRL 은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 논항으로서 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 상기 제3 논항을 상기 특정 시멘틱 롤로 매핑하는 것을 특징으로 하는 문장 분석 장치.

청구항 15

제14항에 있어서,

상기 2-레벨 SVM 기반 SRL은 관련된 조사 또는 관련된 어미의 빈도가 임계 수치 이하이거나 관련된 조사 또는 관련된 어미가 특정한 시멘틱 롤로 매핑되는 임계 비율이 임계 수치 이하인 상기 제4 논항을 SVM를 기반으로 한 복수의 절차를 통해 특정 시멘틱 롤로 매핑하는 것을 특징으로 하는 문장 분석 장치.

발명의 설명

기술 분야

[0001] 본 발명은 문장 분석 방법 및 장치에 관한 것으로서, 보다 상세하게는, 하이브리드 방법을 사용한SRL(semantic role labeling) 기반의 문장 분석 방법 및 장치에 관한 것이다.

배경 기술

[0002] 일반적으로 자연 언어 처리에서는 형태소 분석(morphological analysis), 구문 분석(syntatic analysis), 의미 분석(semantic analysis), 담화 분석(discourse analysis)및 대화 분석(dialogue analysis) 등이 사용될 수 있다.

[0003] 현재 진행되고 있는 의미 분석 관련 연구는 크게 어의 중의성 해소(word sense disambiguation)과 의미 역할 결정(또는 의미 역 결정)(semantic role labeling, SRL)로 분류될 수 있다.

[0004] SRL은 자동적으로 문장 내에 논항(argument)와 술부(또는 동사)(predicate)를 분석하여 문장 내 성분내 시멘틱 롤 (semantic role)(또는 의미 역할, 의미역)을 부여하는 방법이다. SRL은 문장의 각 요소들의 의미 관계를 파악하는 연구 분야로써 어의 중의성 해소와 더불어 자연 언어 처리에서의 의미 분석에서 매우 중요한 위치를 차지하고 있다. 그러나 한국어의 경우에는 SRL에 필요한 언어 자원이 구축되지 못하여 연구의 진행이 매우 미진한 상황이다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) KR 10-2002-0083163

발명의 내용

해결하려는 과제

- [0006] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 본 발명은 일반 언어 특징과 함께 특정 언어의 특징을 추가적으로 고려한 하이브리드 방법을 사용함으로써, 보다 정확한 문장 분석을 가능하도록 하는 방법 및 장치를 제공하고자 한다.
- [0007] 본 발명은 한국어 특성에 따른 특징(feature)을 고려하여 SRL을 수행함으로써, 한국어에 대한 문장 분석의 정확도를 비약적으로 향상시킬 수 있는 방법 및 장치를 제공하고자 한다.
- [0008] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

- [0009] 상기한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 일 측면에 따른 하이브리드 방법을 사용한 SRL(semantic role labeling) 기반의 문장 분석 방법은 분석 대상 문장에 포함되는 복수의 논항 중 제1 논항 그룹에 포함되는 제1 논항에 대해 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계, 상기 복수의 논항 중 제2 논항 그룹에 포함되는 제2 논항에 대해 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하는 단계 및 상기 복수의 논항 중 상기 제 1 논항 그룹 및 상기 제 2 논항 그룹 이외의 논항 그룹에 포함되는 적어도 하나 이상의 논항에 대해 머신(machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계를 포함할 수 있다.
- [0010] 본 발명의 일 측면에 따른 하이브리드 방법을 사용한 SRL(semantic role labeling) 기반의 문장 분석 방법은 분석 대상 문장에 포함되는 복수의 논항 중 제1 논항 그룹에 포함되는 제1 논항에 대해 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계, 상기 복수의 논항 중 제2 논항 그룹에 포함되는 제2 논항에 대해 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하는 단계, 상기 복수의 논항 중 제3 논항 그룹에 포함되는 제3 논항에 대해 한 단어 논항 SRL을 사용하여 시멘틱 롤을 결정하는 단계와 상기 복수의 논항 중 제4 논항 그룹에 포함되는 제4 논항에 대해 2-레벨 SVM(support vector machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계를 포함할 수 있다.
- [0011] 한편, 상기 제1 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제 1 논항을 포함하고, 상기 제2 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가하고 상기 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제2 논항을 포함하고, 상기 제3 논항 그룹은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 상기 제3 논항을 포함하고, 상기 제4 논항 그룹은 상기 케이스 프레임 기반 SRL, 상기 한국어 특정 접미사 구조 SRL, 상기 한 단어 논항 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가한 나머지 논항인 상기 제4 논항을 포함할 수 있다.
- [0012] 또한, 상기 케이스 프레임 기반 SRL은 상기 제1 논항과 관련된 동사의 의미를 고려하여 프레임 파일에서 상기 의미를 가지는 상기 동사를 포함하는 예문에 포함되는 논항과 상기 동사와의 관계를 기반으로 상기 제1 논항의 시멘틱 롤을 결정할 수 있다.
- [0013] 또한, 상기 한국어 특정 접미사 구조 SRL은 한국어 특정 접미사 구조에 대한 정보를 기반으로 특정 임계값 이상의 사용 빈도(frequency)를 가지고 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 조사 또는 어미와 관련된 상기 제2 논항의 시멘틱 롤을 상기 특정 시멘틱 롤로 결정할 수 있다.
- [0014] 또한, 상기 한 단어 논항 SRL 은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 논항으로서 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 상기 제3 논항을 상기 특정 시멘틱 롤로 매핑할 수 있다.
- [0015] 또한, 상기 2-레벨 SVM 기반 SRL은 관련된 조사 또는 관련된 어미의 빈도가 임계 수치 이하이거나 관련된 조사 또는 관련된 어미가 특정한 시멘틱 롤로 매핑되는 임계 비율이 임계 수치 이하인 상기 제4 논항을 SVM를 기반으로 한 복수의 절차를 통해 특정 시멘틱 롤로 매핑할 수 있다.

- [0016] 본 발명의 다른 측면에 따른 하이브리드 방법을 사용한 SRL(semantic role labeling) 기반의 문장 분석 장치는 분석 대상 문장에 포함되는 복수의 논항 중 제1 논항 그룹에 포함되는 제1 논항에 대해 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 케이스 프레임 기반 SRL 부, 상기 복수의 논항 중 제2 논항 그룹에 포함되는 제2 논항에 대해 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 한국어 특정 SRL 부 및 상기 복수의 논항 중 상기 제 1 논항 그룹 및 상기 제 2 논항 그룹 이외의 논항 그룹에 포함되는 적어도 하나 이상의 논항에 대해 머신(machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 머신 기반 SRL 부를 포함할 수 있다.
- [0017] 본 발명의 다른 측면에 따른 문장 분석 장치는 분석 대상 문장에 포함되는 복수의 논항 중 제1 논항 그룹에 포함되는 제1 논항에 대해 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 케이스 프레임 기반 SRL 부, 상기 복수의 논항 중 제2 논항 그룹에 포함되는 제2 논항에 대해 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 한국어 특정 SRL 부, 상기 복수의 논항 중 제3 논항 그룹에 포함되는 제3 논항에 대해 한 단어 논항 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 한 단어 논항 SRL 부와 상기 복수의 논항 중 제4 논항 그룹에 포함되는 제4 논항에 대해 2-레벨 SVM(support vector machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하도록 구현되는 머신 기반 SRL 부를 포함할 수 있다.
- [0018] 한편, 상기 제1 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제1 논항을 포함하고, 상기 제2 논항 그룹은 상기 케이스 프레임 기반 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가능하고 상기 한국어 특정 접미사 구조 SRL을 사용하여 시멘틱 롤에 대한 결정이 가능한 상기 제2 논항을 포함하고, 상기 제3 논항 그룹은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 상기 제3 논항을 포함하고, 상기 제4 논항 그룹은 상기 케이스 프레임 기반 SRL, 상기 한국어 특정 접미사 구조 SRL, 상기 한 단어 논항 SRL을 사용하여 시멘틱 롤에 대한 결정이 불가능한 나머지 논항인 상기 제4 논항을 포함할 수 있다.
- [0019] 또한, 상기 케이스 프레임 기반 SRL은 상기 제1 논항과 관련된 동사의 의미를 고려하여 프레임 파일에서 상기 의미를 가지는 상기 동사를 포함하는 예문에 포함되는 논항과 상기 동사와의 관계를 기반으로 상기 제1 논항의 시멘틱 롤을 결정할 수 있다.
- [0020] 또한, 상기 한국어 특정 접미사 구조 SRL은 한국어 특정 접미사 구조에 대한 정보를 기반으로 특정 임계값 이상의 사용 빈도(frequency)를 가지고 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 조사 또는 어미와 관련된 상기 제2 논항의 시멘틱 롤을 상기 특정 시멘틱 롤로 결정할 수 있다.
- [0021] 또한, 상기 한 단어 논항 SRL 은 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 논항으로서 임계 비율 이상으로 특정 시멘틱 롤로 매핑되는 상기 제3 논항을 상기 특정 시멘틱 롤로 매핑할 수 있다.
- [0022] 또한, 상기 2-레벨 SVM 기반 SRL은 관련된 조사 또는 관련된 어미의 빈도가 임계 수치 이하이거나 관련된 조사 또는 관련된 어미가 특정한 시멘틱 롤로 매핑되는 임계 비율이 임계 수치 이하인 상기 제4 논항을 SVM를 기반으로 한 복수의 절차를 통해 특정 시멘틱 롤로 매핑할 수 있다.

발명의 효과

- [0023] 전술한 본원의 과제 해결 수단에 의하면, 본 발명은 일반 언어 특징과 함께 특정 언어의 특징을 추가적으로 고려한 하이브리드 방법을 사용함으로써 보다 정확한 분장 분석을 가능하도록 하는 방법 및 장치를 제공할 수 있다.
- [0024] 본 발명은 한국어 특성에 따른 특징(feature)을 고려하여 SRL을 수행함으로써, 한국어에 대한 문장 분석의 정확도를 비약적으로 향상시킬 수 있는 방법 및 장치를 제공할 수 있다.

도면의 간단한 설명

- [0025] 도 1은 영어 문장에 적용된 SRL을 나타낸 개념도이다.
- 도 2는 한국어의 특징을 나타내기 위한 개념도이다.
- 도 3은 본 발명의 실시예에 따른 SRL 방법을 나타낸 순서도이다.
- 도 4는 본 발명의 실시예에 따른 하이브리드 방법을 사용한 시멘틱 롤 결정 장치를 나타낸 개념도이다.

발명을 실시하기 위한 구체적인 내용

- [0026] 아래에서는 첨부한 도면을 참조하여 본원이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본원의 실시예를 상세히 설명한다. 그러나 본원은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본원을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0027] 본원 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다.
- [0028] 본원 명세서 전체에서, 어떤 부재가 다른 부재 “상에” 위치하고 있다고 할 때, 이는 어떤 부재가 다른 부재에 접해 있는 경우뿐 아니라 두 부재 사이에 또 다른 부재가 존재하는 경우도 포함한다.
- [0029] 본원 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함" 한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성 요소를 더 포함할 수 있는 것을 의미한다.
- [0030] 본원 명세서 전체에서 사용되는 정도의 용어 "약", "실질적으로" 등은 언급된 의미에 고유한 제조 및 물질 허용 오차가 제시될 때 그 수치에서 또는 그 수치에 근접한 의미로 사용되고, 본원의 이해를 돕기 위해 정확하거나 절대적인 수치가 언급된 개시 내용을 비양심적인 침해자가 부당하게 이용하는 것을 방지하기 위해 사용된다. 본원 명세서 전체에서 사용되는 정도의 용어 "~(하는) 단계" 또는 "~의 단계"는 "~를 위한 단계"를 의미하지 않는다.
- [0031] 본 명세서에 있어서 '부(部)'란, 하드웨어에 의해 실현되는 유닛(unit), 소프트웨어에 의해 실현되는 유닛, 양방을 이용하여 실현되는 유닛을 포함한다. 또한, 1개의 유닛이 2개 이상의 하드웨어를 이용하여 실현되어도 되고, 2개 이상의 유닛이 1개의 하드웨어에 의해 실현되어도 된다.
- [0032] 본 명세서 있어서 단말, 장치 또는 디바이스가 수행하는 것으로 기술된 동작이나 기능 중 일부는 해당 단말, 장치 또는 디바이스와 연결된 서버에서 대신 수행될 수도 있다. 이와 마찬가지로, 서버가 수행하는 것으로 기술된 동작이나 기능 중 일부도 해당 서버와 연결된 단말, 장치 또는 디바이스에서 수행될 수도 있다. 이하 첨부된 도면을 참고하여 본원의 일 실시예를 상세히 설명하기로 한다.
- [0033] SRL(semantic role labeling)은 자동적으로 문장 내에 논항(argument)과 술어(또는 동사(verb))(predicate)를 분석하여 문장 내 성분에 시멘틱 롤 (semantic role)(또는 의미 역할, 의미역)을 부여하는 방법이다. 다른 표현으로 SRL은 동사와 논항의 의미 관계를 밝히기 위해 사용될 수 있다.
- [0034] 한국어는 풍부한 형태론적 구조(morphological structure)를 가지고 있는 교착어이다. 이러한 교착어는 형태소(morpheme)의 결합에 의해 하나의 의미를 나타내고, 어근(stem)에 형태소가 어떠한 조합으로 결합되어있느냐에 따라 다른 의미를 가질 수 있다.
- [0035] SRL은 문장에서 주어진 술어(또는 동사)의 논항을 식별하고, 논항에 시멘틱 롤(의미 역할, 의미 역)을 부여하는 것이다. 논항은 어떠한 술어가 절이나 문장을 만들기 위해 필요로 하는 필수적인 의미 성분들을 말한다.
- [0036] 아래의 표 1은 PropBank에 의해 정의된 17가지의 시멘틱 롤을 나타낸 표이다.

[0037] <표 1>

Role(역할)	Definition(정의)
ARG0	Agent
ARG1	Patient
ARG2	Start point/Benefactive
ARG3	Ending point
ARGM-ADV	Adverbial
ARGM-CAU	Cause
ARGM-CND	Condition
ARGM-DIR	Direction
ARGM-DIS	Discourse
ARGM-EXT	Extent
ARGM-INS	Instrument
ARGM-LOC	Locative
ARGM-MNR	Manner
ARGM-NEG	Negation
ARGM-PRD	Predication
ARGM-PRP	Purpose
ARGM-TMP	Temporal

[0038]

[0039] 표 1을 참조하면, ARG 번호(ARG0~ARG3)은 문장에서 중심적인 역할을 하는 논항에 대응된다. ARGM은 동사가 선택적으로 취할 수 있는 논항으로 동사를 수식해주는 논항에 대응될 수 있다.

[0040] 구체적으로 ARG0은 행위자의 역할을 하는 논항(또는 동사를 수행하는 주체의 역할을 하는 논항)에 부여될 수 있다. ARG1은 피동자(patient)의 역할을 하는 논항에 부여될 수 있다. ARG2는 시작점(~로부터), 동사의 수혜자의 역할을 하는 논항에 부여된다. ARG3은 종료점(~까지)의 역할을 하는 논항에 부여될 수 있다.

[0041] ARGM-ADV는 부사적 어구(adverbial), ARGM-CAU는 원인(cause), ARGM-CND는 조건(condition), ARGM-DIR은 방향(direction), ARGM-DIS는 담화(discourse) 표지(well, on the other hand 처럼 구어에서 대화 내용의 이동을 나타내는 어구), ARGM-LOC은 장소(location), ARGM-MNR은 태도(manner), ARGM-MOD는 범조 동사(modal verb)(또는 조동사), AM-NEG은 부정 표지(negation marker), ARGM-PRD는 술어(predication) 또는 자격(~로서), ARGM-PRP는 목적(purpose), ARGM-TMP는 시간(temporal) 역할을 하는 논항에 부여될 수 있다.

[0042] 도 1은 영어 문장에 적용된SRL을 나타낸 개념도이다.

[0043] 도 1을 참조하면, The luxury auto maker last year sold 1,214 cars in the U.S. 라는 문장에 대한 SRL이 제시된다.

[0044] 문장 내에서 ‘The luxury auto maker(100)’는 동사 ‘sold(120)’의 주체로서 ARG0으로 분류될 수 있다. 또한, 문장 내에서 ‘1,214 cars(130)’는 ‘sold(120)’의 대상으로서 ARG1으로 분류될 수 있다. 또한, 문장 내에서 ‘last year(110)’는 시점을 나타내는 논항으로서 ARGM-TMP로 분류되고, 문장 내에서 ‘in the U.S(140)’는 장소를 나타내는 논항으로서 ARGM-LOC로 분류될 수 있다.

[0045] 한국어의 경우, 한국어 버전 프로포지션 뱅크(Proposition Bank)의 프레임 파일의 부족으로 인해 케이스 프레임 기반 방법(Case frame based metho)으로 SRL할 수 있는 논항(argument)의 수가 매우 적고, 기계 학습(machine learning) 알고리즘을 이용해 SRL했을 때도 만족할 만한 성능을 얻을 수 없었다. 따라서 본 논문에서는 한국어 고유의 특징을 살린 방법론을 사용하였다.

[0046] 이하, 본 발명의 실시예에서는 국문 문장의 특성을 고려하여 국문 문장에 포함되는 문장 성분들에 시멘틱 롤을 부여하는 방법을 개시한다.

[0047] 도 2는 한국어의 특징을 나타내기 위한 개념도이다.

[0048] 한국어에서는 접미사(suffixes)인 조사(Josa)와 어미(Eomi)가 SRL에 중요한 역할을 한다.

[0049] 어미는 술어(predicate(verb or adjective))의 끝에 붙어 다양하게 활용되는 낱말의 부분일 수 있다. 조사는 명사, 대명사, 수사를 통틀어 일컫는 문법 용어인 체언(indeclinable parts of speech in Korean grammar)에 붙어 단어와 단어 사이의 문법적인 관계를 나타내거나 뜻을 첨가해주는 역할을 하기 위한 부분일 수 있다.

[0050] 도 2를 참조하면, “Beacause he believes me, I believe him, too(그가 나를 믿으므로, 나도 그를 믿는다

(Geuga nareul mideumeuro, nado geureul midneunda))” 라는 문장이 개시되어 있다. 이러한 문장에서 술어는 believe(mid)일 수 있다. 체언은 he(Gue), me(na), I(na), him(gue)일 수 있다.

[0051] 본 발명의 실시예에서는 케이스 프레임 사전(case frame dictionary)을 이용한 SRL 방법, 한국어 고유의 특징을 이용한 SRL 방법, 기계 학습(machine learning) 알고리즘을 이용한 SRL 방법 등을 혼합한 하이브리드 방법 기반의 한국어 문장의 SRL 방법이 개시된다.

[0052] 이하, 한국어의 특징인 접미사(조사, 어미) 및 보조 용언(auxiliary predicate)의 구조 및 접미사/보조 용언과 관련된 논항의 시멘틱 롤과의 매핑 관계가 개시된다.

[0053] 한국어의 조사는 크게 격조사와 접속 조사, 보조사로 나뉘 수 있다. 조사가 붙는 단어가 문장 안에서 갖는 문법적 기능을 나타내면 격조사이고, 격조사는 주격 조사, 목적격 조사, 부사격 조사, 보격 조사, 관형격 조사, 호격 조사 등으로 나뉘 수 있다. 접속 조사는 두 단어를 같은 자격으로 이어주기 위한 역할을 할 수 있고, 보조사는 뜻을 가지고 있는 조사이다. 예를 들어, 도 2를 참조하면, ‘가(ga)’는 주격을 나타내는 격조사이고, ‘를(reul)’은 목적격을 나타내는 격조사이고, ‘도(do)’는 추가의 뜻을 나타내는 보조사일 수 있다.

[0054] 이러한 조사들 중 논항을 특정 시멘틱 롤로 일정 비율(예를 들어, 80%) 이상 매핑시키는 조사들이 존재할 수 있다. 아래의 표 2는 관련 논항을 일정 비율(예를 들어, 80%) 이상으로 특정 시멘틱 롤로 매핑시키는 조사를 개시한다.

[0055] <표 2>

Josa	Role	%	Frequency
Reul	ARG1	97.5	2597
Eul	ARG1	97.1	4195
E daeha eo	ARGM-ADV	95.9	73
Man	ARG1	89.4	104
Buteo	ARGM-TMP	89.1	92
Cheoreom	ARGM-EXT	85.6	97
Eseo	ARGM-LOC	80.8	1069

[0056]

[0057] 예를 들어, 를(reul), 을(eul)과 같은 조사는 격조사로서 목적격 조사일 수 있다. 이러한 를(reul), 을(eul)과 같은 목적격 조사와 연계된 논항의 시멘틱 롤은 97.5%, 97.1%의 비율로 피동자(patient)의 역할(ARG1)일 수 있다.

[0058] ‘에 대하여(E_daeha_eo)’와 연계된 논항의 시멘틱 롤은 95.9%의 비율로 부사적 어구(adverbial)의 역할(ARGM-ADV)일 수 있다.

[0059] ‘만(Man)’과 연계된 논항의 시멘틱 롤은 89.4%의 비율로 피동자의 역할(ARG1)일 수 있다.

[0060] ‘부터(Buteo)’와 연계된 논항의 시멘틱 롤은 89.1%의 비율로 시간(temporal) 역할(ARGM-TMP)일 수 있다.

[0061] ‘처럼(Cheoreom)’과 연계된 논항의 시멘틱 롤은 85.6%의 비율로 정도(extent)의 역할(ARGM-EXT)일 수 있다.

[0062] ‘에서(Eseo)’와 연계된 논항의 시멘틱 롤은 80.8%의 비율로 위치(location)의 역할(ARGM-LOC)일 수 있다.

[0063] 이하에서는 어미의 분류 체계 및 어미와 어미와 관련된 논항 간의 시멘틱 롤의 매핑 관계가 개시된다.

[0064] <표 3>

Eomi	Pre-final	Honorific	
		Modesty	
		Tense	Future: Past
	Final	Closing	Normal Interrogative Imperative Petitionary Exclamatory
		Connection	Coordinate Subordinate
		Transmutation	Nominal Determinative Adverbial

[0065]

[0066] 표 3은 어미(Eomi)의 분류 체계를 개시한다.

[0067] 선 어말 어미(Prefinal Eomi)는 어미 중에서 어말에 놓이지 못하고 어근과 어말 어미(Final Eomi) 사이에 존재하는 어미이다. 선 어말 어미는 높임(Honorific), 공손(Modesty), 시제(Tense) 등을 나타낼 수 있다.

[0068] 어말 어미 중 문장을 끝맺지 않고 이어 주는 구실을 하는 어미를 연결 어미(Connection Eomi)라고 하며, 문장을 끝맺는 구실을 하는 어미는 종결 어미(Closing Eomi), 문장의 성격을 관형사나 명사처럼 바꾸는 구실을 하는 어미를 전성 어미(Transmutation Eomi)라고 한다.

[0069] 예를 들어, 도 2에서 ‘으므로(eumeuro)’는 까닭이나 근거를 나타내는 연결 어미이고, ‘는다(neunda)’는 현재 사건이나 사실을 서술하는 뜻을 나타내는 종결 어미이다.

[0070] 이러한 어미들 중 논항을 특정 시멘틱 롤로 일정 비율 이상 매핑시키는 어미들이 존재할 수 있다. 아래의 표 4는 어미와 어미 관련 논항 사이의 시멘틱 롤 매핑 비율을 개시한다.

[0071] <표 4>

Eomi	Class	Role	%	Frequency
Go	Coordinate Connection	ARGM-DIS	97.1	3353
Gi-ddae mun e	Subordinate Connection	ARGM-CAU	96.6	174
Ge	Adverbial Transmutation	ARGM-MNR	86.9	236
Eot	Past tense Prefinal	ARGM-DIS	75.7	1335
Myeon	Subordinate Connection	ARGM-CND	72.1	677
Si	Honorific Prefinal	ARGM-DIS	64.7	51
Da-go	Subordinate Connection	ARG1	58.8	337

[0072]

[0073] 표 4를 참조하면, ‘고(go)’는 등위 연결 어미(예를 들어, and, or, but 등)이고, ‘고’와 연계된 논항의 시멘틱 롤의 97.1%는 ARGM-DIS(담화(discourse) 표지 역할 논항)일 수 있다.

[0074] ‘기 때문에(Gi-ddae mun e)’는 종속 연결 어미이고, ‘기 때문에’와 연계된 논항의 시멘틱 롤의 96.6%는 ARGM-CAU(원인 역할 논항)일 수 있다.

[0075] ‘게(Ge)’는 부사 전성 어미이고, ‘게’와 연계된 논항의 시멘틱 롤의 86.9%는 ARGM-MNR(태도 역할 논항)일 수 있다.

[0076] ‘었(Eot)’는 과거 시제 선 어말 어미이고, ‘었’와 연계된 논항의 시멘틱 롤의 75.7%는 ARGM-MNR(태도 역할 논항)일 수 있다.

[0077] ‘면(Myeon)’는 종속 연결 어미이고, ‘시’와 연계된 논항의 시멘틱 롤의 72.1%는 ARGM-CND(조건(condition))일 수 있다.

[0078] ‘시(Si)’는 높임 선 어말 어미이고, ‘시’와 연계된 논항의 시멘틱 롤의 64.7%는 ARGM-DIS일 수 있다.

[0079] ‘다고(Da-go)’는 종속 연결 어미이고, ‘다고’와 연계된 논항의 시멘틱 롤의 58.8%는 ARG1일 수 있다.

[0080] 이하에서는 보조 용언(auxiliary predicate)(또는 보조 술부) 및 보조 용언과 관련된 논항 간의 시멘틱 롤의 매핑 관계가 개시된다.

[0081] 보조 용언은 다른 술어의 뒤에 기대어 그 말의 뜻을 도와주는 술어이다. 본 용언은 문장에서 제거하면 그 문장이 성립하지 않으나, 보조 용언은 보조 용언을 제거해도 문장이 성립할 수 있다.

[0082] 아래의 표 5는 보조 용언을 예시적으로 나타낸 표이다.

<표 5>

Meaning	Auxiliary predicate	Example
Causative	Ha-da	Geu-ri-ge Ha-da (Make him draw)
Denial	A-ni-ha-da	Meok-ji A-ni-ha-da (Do not eat)
Progress	It-da	Nae-ri-go It-da (Be coming down)
Close	Nae-da	Gyeon-di-eo Nae-da (Carry through)
Service	Ju-da	Do-wa Ju-da (Lend a hand)
Hope	Sip-da	Bo-go Sip-da (Want to see)

‘하다(Hada)’는 사역(causative)의 의미로 사용될 수 있다. 예를 들어, ‘하다(Hada)’는 그리게 하다(Geu-ri-ge Ha-da(Make him draw))와 같이 특정한 동작을 시킨다는 의미를 더하기 위해 사용될 수 있다.

‘아니하다(A-ni-ha-da)’는 부인(denial)의 의미로 사용될 수 있다. 예를 들어, ‘아니하다(A-ni-ha-da)’는 먹지 아니하다(Meok-ji A-ni-ha-da(Do not eat))와 같이 특정한 동작을 하지 않는다는 의미를 더하기 위해 사용될 수 있다.

‘있다(It-da)’는 진행(progress)의 의미로 사용될 수 있다. 예를 들어, ‘있다(It-da)’는 내리고 있다(Nae-ri-go It-da(Be coming down))와 같이 특정한 동작에 진행의 의미를 더하기 위해 사용될 수 있다.

‘내다(Nae-da)’는 종료(close)의 의미로 사용될 수 있다. 예를 들어, ‘내다(Nae-da)’는 견디어 내다(Gyeon-di-eo Nae-da(carry through))와 같이 특정한 동작에 종로의 의미를 더하기 위해 사용될 수 있다.

‘주다(Ju-da)’는 서비스(service)의 의미로 사용될 수 있다. 예를 들어, ‘주다(Ju-da)’는 도와 주다(Do-wa Ju-da(Lend a hand))와 같이 특정한 동작에 서비스의 의미를 더하기 위해 사용될 수 있다.

‘싶다(Sip-da)’는 소망,기원(hope)의 의미로 사용될 수 있다. 예를 들어, ‘싶다(Sip-da)’는 보고 싶다(Bo-go Sip-da(Want to see))와 같이 특정한 동작에 소망, 기원의 의미를 더하기 위해 사용될 수 있다.

아래의 표 6은 보조 용언과 보조 용언 관련 논항 사이의 시멘틱 롤 매핑 비율을 개시한다.

<표 6>

Auxiliary predicate	Role	%	Frequency
Gi Wiha	ARGM-PRP	81	124
Eo It	ARGM-DIS	78	27
Ge Ha	ARGM-DIS	71	97
Go It	ARGM-DIS	67	402
Giman Ha	ARGM-CND	50	12

표 6을 참조하면, ‘기 위하(Gi-Wiha)’와 연계된 논항의 시멘틱 롤의 81%는 ARGM-PRP(목적)일 수 있다.

‘여 있(Eo-It)’과 연계된 논항의 시멘틱 롤의 78%는 ARGM-DIS(담화(discourse) 표지 역할 논항)일 수 있다.

‘게 하(Ge-Ha)’와 연계된 논항의 시멘틱 롤의 71%는 ARGM-DIS일 수 있다.

‘고 있(Go-It)’과 연계된 논항의 시멘틱 롤의 67%는 ARGM-DIS일 수 있다.

‘기만 하(Giman-Ha)’와 연계된 논항의 시멘틱 롤의 50%는 ARGM-CND(조건)일 수 있다.

- [0099] 이하, 위와 같은 조사, 어미, 보조 용언 각각이 특정 시멘틱 롤과 매핑되는 결과를 기반으로 한 한국어 문장에 대한 SRL 방법이 개시된다.
- [0100] 도 3은 본 발명의 실시예에 따른 SRL 방법을 나타낸 순서도이다.
- [0101] 도 3에서는 분석 대상 문장에 포함되는 논항의 시멘틱 롤을 결정하는 방법이 개시된다.
- [0102] 도 3을 참조하면, 분석 대상 문장에 대한 케이스 프레임 기반 SRL(case frame based SRL)이 수행된다(단계 S300).
- [0103] 케이스 프레임 기반 SRL은 다른 표현으로 격틀 사건을 이용하는 방법이라고 할 수도 있다. 케이스 프레임 기반 SRL은 격틀(case frame)이라는 언어 지식을 이용하는 방법으로, 격틀 사전에 기술된 격틀과 선택 제약 정보(selectional restriction) 등을 이용하여 분석 대상 문장에 포함된 논항에 적합한 시멘틱 롤을 할당하는 방법이다.
- [0104] 본 발명의 실시예에 따르면, 우선적으로 분석 대상 문장에 포함된 논항 중 케이스 프레임이 존재하는 논항에 대하여 한국어 버전 프로포지션 뱅크(proposition bank)의 프레임 파일에서 프레임 정보가 추출되고 자동으로 시멘틱 롤이 결정될 수 있다. 프레임 파일(또는 케이스 프레임 파일)은 특정 동사의 여러 의미 및 여러 의미를 가진 특정 동사와 논항과의 관계, 해당 논항의 시멘틱 롤에 대한 정보를 포함할 수 있다.
- [0105] 먼저, 프레임 파일에서 동사는 복수의 의미(sense)를 가질 수 있다. 따라서, 분석 대상 논항과 관련된 동사의 의미를 고려하여 프레임 파일에서 동사의 적절한 의미가 선택되어야 한다. 동사의 의미는 논항의 통사적, 의미적 특징에 따라 분류될 수 있다. 따라서, 본 발명의 실시예에 따르면, 시멘틱 롤을 부여하고자 하는 말뭉치(corpus)의 논항(예를 들어, 명사)들과 프레임 파일에서 동사의 복수의 의미 각각에 따른 예문에 있는 논항(예를 들어, 명사)들 사이의 시멘틱 유사도(semantic similarity)가 결정될 수 있다. 시멘틱 유사도가 가장 높은 예문이 선택되고, 해당 예문에 포함된 특정 의미를 가진 동사와 논항과의 관계, 논항의 시멘틱 롤에 대한 정보를 기반으로 분석 대상 문장에 포함되는 논항의 시멘틱 롤이 결정될 수 있다.
- [0106] 한국어 특정 접미사 구조를 사용한 SRL이 수행된다(단계 S310).
- [0107] 프레임 파일이 부족하거나, 프레임 파일이 있어도, 단계 S300을 통해 시멘틱 롤을 결정하지 못한 논항은 한국어 특정 접미사 구조(조사, 어미)에 대한 정보를 기반으로 SRL을 수행할 수 있다. 단계 S310에 따른 한국어 특정 접미사 구조를 사용한 SRL은 일정 임계값 이상의 사용 빈도(frequency)를 가지고 일정 임계 비율 이상 특정 롤로 매핑되는 조사 및 어미에 대하여 적용될 수 있다.
- [0108] 관련된 조사가 있는 논항은 전술한 표 2와 같은 조사와 시멘틱 롤 간의 관계를 고려하여 시멘틱 롤을 결정할 수 있다.
- [0109] 관련된 어미가 있는 논항은 전술한 표 4와 같은 어미와 시멘틱 롤 간의 관계 및 아래의 표 7과 같은 서로 다른 타입의 술어구(different types of predicate phrase)를 고려하여 시멘틱 롤을 결정할 수 있다.
- [0110] <표 7>

Case 1	Radix+Final
Case 2	Radix+Prefinal+Final
Case 3	Radix+Auxiliary predicate+Final
Case 4	Radix+Auxiliary predicate+Prefinal+Final

- [0111]
- [0112] 표 7을 참조하면, 어미가 있는 논항은 4가지의 케이스로 분류될 수 있다.
- [0113] Case 1은 논항과 관련된 술어(predicate) 뒤에 바로 어말 어미(final eomi)가 붙은 경우이다. Case 1에서는 어말 어미와 관련된 시멘틱 롤로 논항이 라벨링(labeling)될 수 있다.
- [0114] *Case 2는 논항과 관련된 술어 뒤에 선 어말 어미(prefinal Eomi)가 붙고 그 뒤에 어말 어미가 붙은 경우이다. 이러한 경우, 표 4에서 선 어말 어미 및 어말 어미 중 더 높은 시멘틱 롤 매핑 정확도를 가지는 어미와 매핑되는 시멘틱 롤로 논항의 시멘틱 롤이 결정될 수 있다.
- [0115] Case 3과 Case 4는 술부 뒤에 보조 용언(또는 보조 술부(auxiliary predicate))가 붙고 보조 용언의 뒤에 어미가 붙은 경우이다. 이러한 경우, 전술한 표 4 및 표 6에서 해당 보조 용언과 어미 중 더 높은 시멘틱 롤 매핑

정확도를 가지는 보조 용어 또는 어미와 매핑되는 시멘틱 롤로 논항의 시멘틱 롤이 결정될 수 있다.

[0116] 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 논항의 시멘틱 롤을 결정한다(단계 S320).

[0117] 조사가 있는 논항과 마찬가지로 수동 SRL을 기반으로 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 논항(이하, 한 단어 논항)이 80% 이상 특정 시멘틱 롤로 결정되는 경우, 한 단어 논항이 자동 SRL을 기반으로 해당 시멘틱 롤로 결정될 수 있다.

[0118] <표 8>

Arguments	Meaning	Roles	%	Frequency
Geu-reo-na	But	ARGM-DIS	100	271
Da	All	ARGM-EXT	100	98
An	Not	ARGM-LOC	98	45
Tto	Also	ARGM-DIS	97	221
Gyeol-guk	Ultimately	ARGM-DIS	94	71
Jal	Well	ARGM-MNR	90	156

[0119]

[0120] 표 8을 참조하면, 조사와 어미가 없으며 한 단어로 이루어진 한 단어 논항 중 80% 이상 특정 시멘틱 롤로 매핑되는 논항이 개시된다.

[0121] 표 8을 참조하면, ‘그러나(Geu-reo-na)(But)’는 수동 SRL을 기반으로 100% ARGM-DIS와 매핑될 수 있다.

[0122] ‘다(Da)(All)’는 수동 SRL을 기반으로 100% ARGM-EXT와 매핑될 수 있다.

[0123] ‘안(An)(Not)’는 수동 SRL을 기반으로 98% ARGM-LOC와 매핑될 수 있다.

[0124] ‘또(Tto)(Also)’는 수동 SRL을 기반으로 97% ARGM-DIS와 매핑될 수 있다.

[0125] ‘결국(Gyeol-guk)(Ultimately)’은 수동 SRL을 기반으로 94% ARGM-DIS와 매핑될 수 있다.

[0126] ‘잘(Jal)(Well)’은 수동 SRL을 기반으로 90% ARGM-MNR과 매핑될 수 있다.

[0127] 2-레벨 SVM(support vector machine)을 기반으로 나머지 논항의 시멘틱 롤을 결정한다(단계 S330).

[0128] 단계 S300 내지 단계 S320을 통해 시멘틱 롤을 결정하지 못한 나머지 논항에 대하여 2-레벨 SVM이 수행될 수 있다.

[0129] 2-level SVM은 논항에 조사나 어미가 붙어있으나, 분석을 수행한 조사나 어미의 빈도(frequency)가 임계 수치(예를 들어, 5) 이하이거나 분석을 수행한 조사나 어미가 특정한 시멘틱 롤로 매핑되는 확률이 임계 수치(예를 들어, 80%) 이하인 경우, 논항의 시멘틱 롤의 결정을 위해 사용될 수 있다.

[0130] 2-level SVM에서는 먼저, 첫번째 레벨(level1)로 SVM-Light를 사용하여 해당 논항이 ARGN인지 ARGM인지 여부를 결정(예측)할 수 있다. 출력 데이터를 예측 결과대로 분류하고 각각을 두번째 레벨의 멀티클래스 SVM의 테스트 데이터(또는 입력 데이터)로 사용할 수 있다.

[0131] 두번째 레벨(level2)로 멀티클래스 SVM이 사용되고 멀티클래스 SVM은 첫번째 레벨(level1)의 SVM-Light를 기반으로 ARGN으로 예측한 데이터를 ARG0, ARG1, ARG2, ARG3의 4가지 클래스로 다시 세부적으로 분류할 수 있다.

[0132] 또한, 두번째 레벨(level2)로 사용되는 멀티클래스 SVM은 첫번째 레벨(level1)의 SVM-Light를 기반으로 ARGM으로 예측한 데이터를 ARGM-DIS, ARGM-MNR, ARGM-TMP, ARGM-LOC, ARGM-EXT, ARGM-CAU, ARGM-CND, ARGM-ADV, ARGM-INS, ARGM-PRP, ARGM-PRD, ARGM-DIR, ARGM-NEG의 13가지 클래스로 세부적으로 다시 분류할 수 있다.

[0133] SVM-Light 및 멀티캐스트 SVM은 한국어 특정 특징(Korean specific feature)를 사용하여 논항에 대한 시멘틱 롤을 결정할 수 있다. 아래의 표 9는 한국어 특정 특징을 나타낸다.

[0134] <표 9>

Feature(특징)	Description(설명)
A-JosaExist	If 1, Josa exists in an argument, otherwise 0 조사가 존재하는 경우, 1, 그 외의 경우, 0
A-JosaClass	Linguistic classification of Josa 조사의 언어학적 분류
A-JosaLength	Number of Morphemes consisting of Josa 조사를 구성하는 형태소의 개수
A-JosaMorphemes	Each morpheme consisting of Josa 조사를 구성하는 각각의 형태소
A-JosaIdentity	Josa of a argument 논항의 조사
A-EomiExist	If 1, Eomi exists in an argument, otherwise 0 논항에 어미가 존재하는 경우, 1, 그 외의 경우, 0
A-EomiClass_Lv1	Linguistic classification of Eomi 어미의 언어학적 분류
A-EomiClass_Lv2	Another linguistic classification of Eomi 어미의 다른 언어학적 분류
A-EomiLength	Number of morphemes consisting of Eomi 어미를 포함하는 형태소의 개수
A-EomiMorphemes	Each morpheme consisting of Eomi 어미를 포함하는 형태소
A-EomiIdentity	Eomi of an argument 논항의 어미

[0135]

[0136] 표 9를 참조하면, 한국어 특정 특징은 아래와 같다. 이하에서, 예시적으로, "이 책은 성종 때 한글로 번역 출판된 이래 영조 5년 때까지 여러 번 중간됐으며 정조 20년에는 이룬 행실도와 합쳐져 오륜행실도란 책이 이루어졌다"라는 문장을 기반으로 설명될 수 있다.

[0137] A-JosaExist은 논항에 조사가 붙어있는지 여부를 지시할 수 있다. 예를 들어, 논항에 조사가 붙어 있으면 A-JosaExist는 1이고 아니면 0일 수 있다. 예시 문장에서 ‘책’의 ‘은’이라는 조사가 붙어 있으므로 ‘책’의 A-JosaExist는 1일 수 있다.

[0138] A-JosaClass는 조사의 8개의 class중 하나를 지시할 수 있다. 예시 문장에서 조사가 ‘은’일 경우, ‘은’의 A-JosaClass는 일반보조사일 수 있다.

[0139] A-JosaLength는 조사의 형태소의 수를 지시할 수 있다. 예시 문장에서 조사가 ‘은’일 경우, ‘은’의 A-JosaLength는 1이고, 조사가, ‘뿐_아니_라’일 경우, ‘뿐_아니_라’의 A-JosaLength는 3일 수 있다.

[0140] A-JosaMorphemes는 조사의 형태소 각각을 지시할 수 있다. 조사가 ‘은’일 경우 ‘은’의 A-JosaMorphemes는 형태소인 ‘은’이고, 조사가 ‘뿐 아니라’인 경우, ‘뿐 아니라’의 A-JosaMorphemes는 형태소인 ‘뿐’, ‘아니’, ‘라’일 수 있다.

[0141] A-JosaIdentity는 논항의 조사를 지시할 수 있다. 논항 ‘책’의 A-JosaIdentity는 ‘은’일 수 있다.

[0142] A-EomiExist는 논항에 어미의 존재 여부를 지시할 수 있다. 논항에 어미가 존재하는 경우, A-EomiExist는 1이고, 논항에 어미가 존재하지 않는 경우, A-EomiExist는 0일 수 있다

[0143] A-EomiClass_Lv1은 어미의 14개의 클래스 중 하나를 지시할 수 있다.

[0144] A-EomiClass_Lv2는 어미의 4개의 클래스 중 하나를 지시할 수 있다. 이 때, 어미의 4개의 클래스는 선어말(prefinal), 종결(closing), 연결(connection) 및 전성(transmutation)일 수 있다.

[0145] A-EomiLength는 어미의 형태소의 수를 지시할 수 있다.

[0146] A-EomiMorphemes는 어미의 형태소 각각을 지시할 수 있다.

[0147] EomiIdentity는 논항의 어미를 지시할 수 있다.

[0148] 도 3을 통해 설명된 하이브리드 방법을 사용한 SRL(semantic role labeling) 기반의 문장 분석 방법은 단계 S300, S310 및 S330만을 포함할 수도 있다. 이 경우, 단계 S330은 복수의 논항 중 제 1 논항 그룹 및 제 2 논항 그룹 이외의 논항 그룹에 포함되는 적어도 하나 이상의 논항에 대해 머신(machine) 기반 SRL을 사용하여 시멘틱 롤을 결정하는 단계일 수 있다.

- [0149] 도 4는 본 발명의 실시예에 따른 하이브리드 방법을 사용한 시멘틱 롤 결정 장치를 나타낸 개념도이다.
- [0150] 도 4를 참조하면, 하이브리드 방법을 사용한 시멘틱 롤 결정 장치는 입력부(400), 케이스프레임 기반 SRL 부(410), 한국어 특정 SRL 부(420), 한 단어 논항 SRL 부(430), 머신 기반 SRL 부(440), 출력부(450) 및 프로세서(460)를 포함한다. 하이브리드 방법을 사용한 시멘틱 롤 결정 장치는 전술한 하이브리드 방법을 사용한 시멘틱 롤 결정 방법을 수행하기 위해 구현될 수 있다. 예를 들어, 각 구성부는 아래의 동작을 수행하기 위해 구현될 수 있다.
- [0151] 본 발명의 일 실시예에 따르면, 하이브리드 방법을 사용한 시멘틱 롤 결정 장치는 입력부(400), 케이스프레임 기반 SRL 부(410), 한국어 특정 SRL 부(420), 머신 기반 SRL 부(440), 출력부(450) 및 프로세서(460)를 포함할 수 있다. 이 경우, 머신 기반 SRL 부(440)는 복수의 논항 중 제 1 논항 그룹 및 제 2 논항 그룹 이외의 논항 그룹에 포함되는 적어도 하나 이상의 논항에 대해 머신(machine) 기반 SRL을 사용하여 시멘틱 롤을 결정할 수 있다.
- [0152] 입력부(400)는 분석 대상 문장의 입력을 위해 구현될 수 있다.
- [0153] 케이스프레임 기반 SRL 부(410)는 케이스 프레임 기반 SRL을 수행하기 위해 구현될 수 있다. 케이스프레임 기반 SRL 부(410)는 분석 대상 논항과 관련된 동사의 의미를 고려하여 프레임 파일에서 동사의 적절한 의미가 선택하고, 해당 의미의 동사를 포함하는 예문에 포함되는 논항과 동사와의 관계를 기반으로 분석 대상 문장에 포함되는 논항의 시멘틱 롤을 결정할 수 있다.
- [0154] 한국어 특정 SRL 부(420)는 케이스프레임 기반 SRL 부에 의해 시멘틱 롤이 결정되지 않은 나머지 논항에 대하여 한국어 특정 접미사 구조(조사, 어미)에 대한 정보를 기반으로 한 SRL을 수행하기 위해 구현될 수 있다. 한국어 특정 SRL 부(420)는 일정 임계값 이상의 사용 빈도(frequency)를 가지고 일정 임계 비율 이상 특정 롤로 매핑되는 조사 및 어미와 관련된 논항에 대한 시멘틱 롤을 결정할 수 있다.
- [0155] 한 단어 논항 SRL 부(430)는 케이스프레임 기반 SRL 부(410)/한국어 특정 SRL 부(420)에 의해 시멘틱 롤이 결정되지 않은 한 단어 논항에 대한 SRL을 수행하기 위해 구현될 수 있다. 한 단어 논항 SRL 부(430)는 조사와 어미를 모두 포함하지 않고 하나의 단어로 이루어진 논항(이하, 한 단어 논항)이 80% 이상 특정 시멘틱 롤로 결정되는 경우, 해당 한 단어 논항은 자동 SRL을 기반으로 특정 시멘틱 롤로 결정될 수 있다.
- [0156] 머신 기반 SRL 부(440)는 2-level SVM은 논항에 조사나 어미가 붙어있으나, 분석을 수행한 조사나 어미의 빈도(frequency)가 임계 수치(예를 들어, 5) 이하이거나 분석을 수행한 조사나 어미가 특정한 시멘틱 롤로 매핑되는 확률이 임계 수치(예를 들어, 80%) 이하인 경우, 논항의 시멘틱 롤의 결정을 위해 사용될 수 있다. 머신 기반 SRL 부(440)는 케이스프레임 기반 SRL 부(410)/한국어 특정 SRL 부(420)/한 단어 논항 SRL 부(430)에 의해 시멘틱 롤을 결정하지 못한 나머지 논항에 대한 시멘틱 롤을 결정할 수 있다.
- [0157] 출력부(450)는 케이스프레임 기반 SRL 부(410), 한국어 특정 SRL 부(420), 한 단어 논항 SRL 부(430), 2-레벨 SVM 기반 SRL 부(440) 기반으로 시멘틱 롤을 결정한 문장 분석 결과에 대한 정보를 출력하기 위해 구현될 수 있다.
- [0158] 프로세서(460)는 입력부(400), 케이스프레임 기반 SRL 부(410), 한국어 특정 SRL 부(420), 한 단어 논항 SRL 부(430), 2-레벨 SVM 기반 SRL 부(440), 출력부(450)의 동작을 제어하기 위해 구현될 수 있다.
- [0159] 이하에서는 본 발명의 실시예에 따른 하이브리드 방법을 사용한 시멘틱 롤 결정 방법이 사용되는 경우, 분석 과정 및 분석 결과가 개시된다.
- [0160] ETRI의 corpus 10000개의 문장을 사용하여 총 78086개의 논항을 획득할 수 있었다. 78086개의 논항 중 약 80%인 62862개의 논항이 훈련 데이터(train data)로 사용되고, 약 20%인 15224개의 논항이 테스트 데이터(test data)로 사용될 수 있다.
- [0161] 케이스 프레임 기반 방법을 기반으로 15224개의 논항 중 4468개(약 29%)의 논항에 대한 시멘틱 롤이 할당될 수 있다. 4468개의 논항 중 4021개의 논항(90%)이 실제 시멘틱 롤과 일치하는 결과를 가졌다.
- [0162] 케이스 프레임 기반 방법을 기반으로 SRL을 하지 못한 나머지 10756개의 논항 중 특정 시멘틱 롤로 80% 이상 매핑되는 조사와 관련된 논항은 1721개였고, 실제 시멘틱롤과 일치한 논항은 1451개로 82%의 정확도를 보였다.
- [0163] 또한, 어미가 있고 빈도가 5 이상인 논항은 3415개였고, 3415개 중 2686개가 실제 시멘틱 롤과 일치하여 79%의 정확도를 보였다.

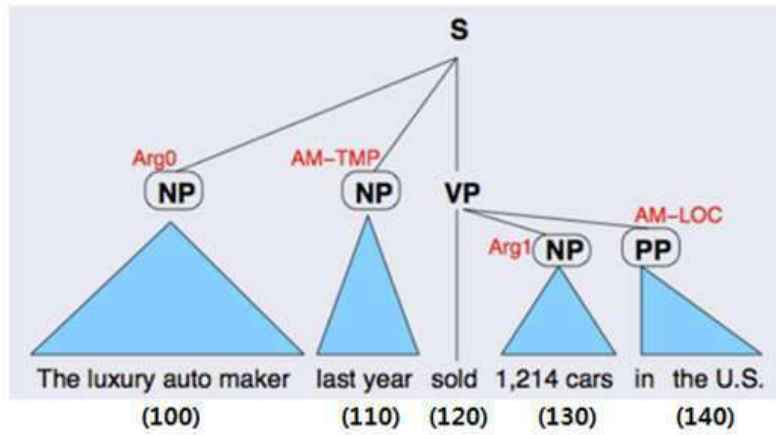
- [0164] 조사와 어미가 모두 없고 한 단어로 이루어진 논항 중 80% 이상 특정 시멘틱 롤로 매핑되는 논항은 1217개였고, 1161개가 실제 시멘틱롤과 일치하여 95%의 정확도를 보였다.
- [0165] 위에서 기술한 방법들로 시멘틱 롤이 부여된 10821개의 논항을 제외한 4403개의 논항에 대해 2레벨 SVM이 수행한 결과, 레벨 1에서 81.79%의 정확도를 얻었고, 레벨 1에서 ARGN으로 예측한 데이터를 멀티캐스트 SVM한 결과 48.72%의 정확도를 보였고, 레벨 2에서 ARGN으로 예측한 데이터를 멀티캐스트 SVM한 결과 38%의 정확도를 보였다.
- [0166] 본 발명의 실시예에 따른 하이브리드(Hybrid) 방법을 사용하여 한국어 문장에 포함된 논항에 대한 SRL을 수행하는 경우, 기존의 SRL 방법보다 높은 정확도로 논항에 대한 시멘틱 롤 라벨링이 수행될 수 있다.
- [0167] 이와 같은 하이브리드 방법을 사용한 SRL 기반의 문장 분석 방법은 애플리케이션으로 구현되거나 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다.
- [0168] 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.
- [0169] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD 와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.
- [0170] 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0171] 이상에서는 실시예들을 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.
- [0172] 전술한 본원의 설명은 예시를 위한 것이며, 본원이 속하는 기술분야의 통상의 지식을 가진 자는 본원의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.
- [0173] 본원의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본원의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

- [0174] 400: 입력부
- 410: 케이스 프레임 기반 SRL부
- 420: 한국어 특정 SRL부
- 430: 한 단어 논항 SRL부
- 440: 머신 기반 SRL부
- 450: 출력부
- 460: 프로세서

도면

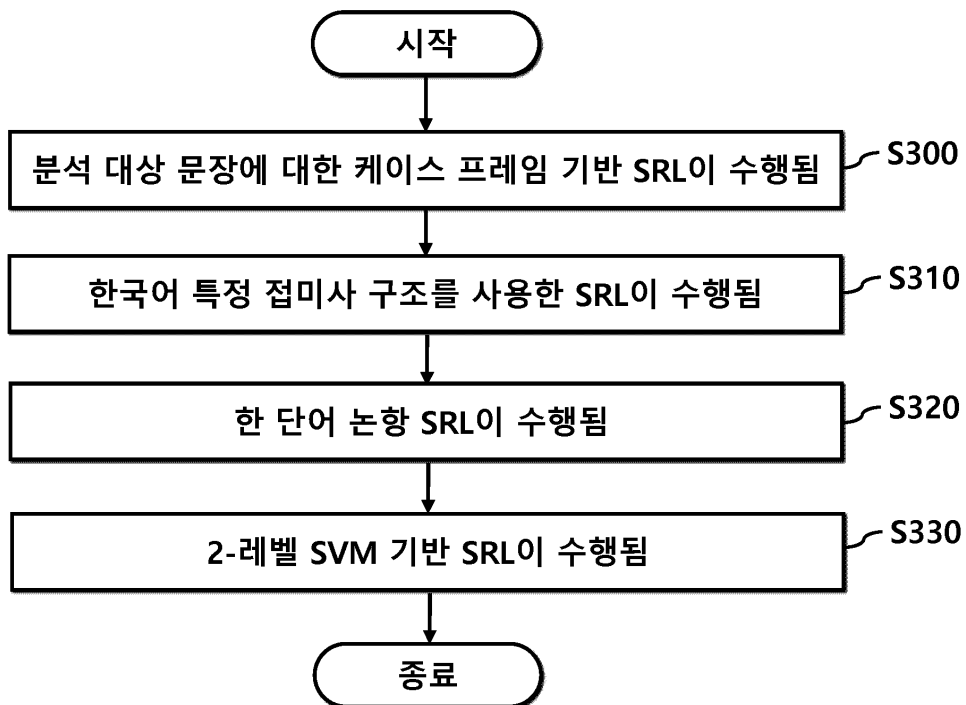
도면1



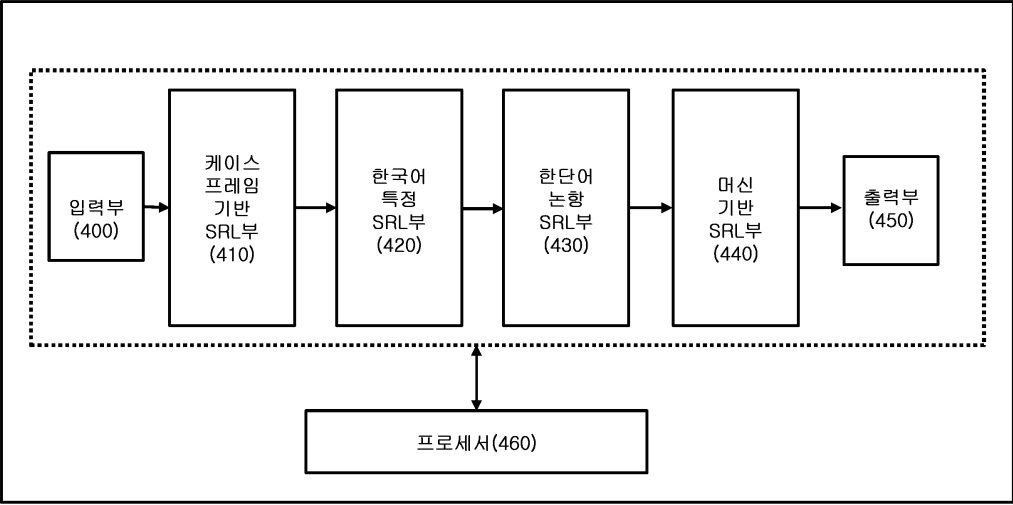
도면2

Because he believes me, I believe him too.
Geuga nareul midneuro, nado geureul midneunda.

도면3



도면4



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구범위 제8항

【변경전】

~ 기록된 기록매체.

【변경후】

~ 기록된 컴퓨터로 관독 가능한 기록매체.