



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년08월05일
(11) 등록번호 10-1646159
(24) 등록일자 2016년08월01일

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

(52) CPC특허분류
G06F 17/2785 (2013.01)
G06F 17/2705 (2013.01)

(21) 출원번호 10-2015-0057552

(22) 출원일자 2015년04월23일

심사청구일자 2015년04월23일

(56) 선행기술조사문헌

KR1020130099327 A*

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

김병수 외, 비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정, 한국정보과학회 언어공학연구회 학술발표 논문집, pp.32-39, 2006.10.

KR1020100030835 A

*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)

김용범

미합중국 워싱턴주 벨뷰시 110번째 거리 788 북동
617호

(뒷면에 계속)

(74) 대리인

유민규

전체 청구항 수 : 총 11 항

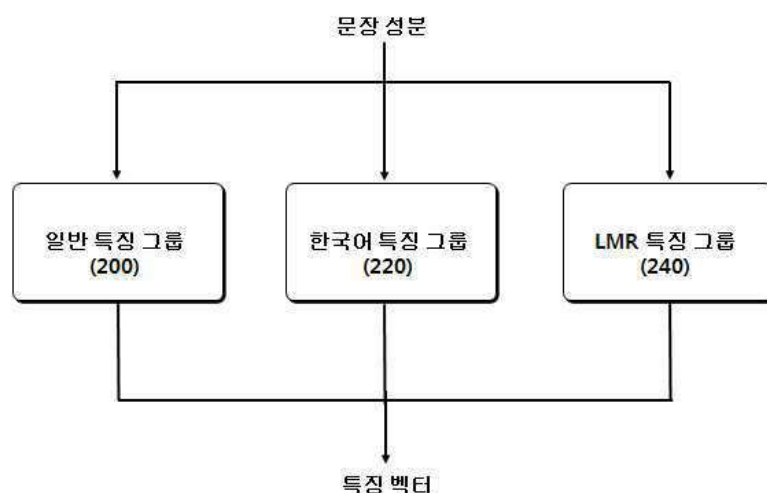
심사관 : 경연정

(54) 발명의 명칭 SRL 기반의 문장 분석 방법 및 장치

(57) 요약

SRL 기반의 문장 분석 방법 및 장치가 개시된다. SRL 기반의 문장 분석 방법은 SRL 문장 분석 장치가 일반 특징 그룹에 포함되는 복수의 제1 특징 및 제 1 언어 특징 그룹에 포함되는 복수의 제2 특징을 기반으로 문장 내의 문장 성분을 특징 벡터로 결정하는 단계와 SRL 문장 분석 장치가 특징 벡터를 기반으로 문장 성분에 대한 시멘틱 롤을 결정하는 단계를 포함하되, 복수의 제1 특징은 논항 또는 술어인 문장 성분의 문장 내에서 문법적 역할에 대한 정보를 지시하고 복수의 제2 특징은 문장 성분에 관련된 조사 또는 어미에 대한 정보를 지시할 수 있다.

대표도 - 도2



(52) CPC특허분류

G06F 17/274 (2013.01)

G06F 17/2755 (2013.01)

(72) 발명자

채희문

미합중국 위스콘신주 매디슨시 데이튼 스트리트 서
1210

벤자민 스나이더

미합중국 위스콘신주 매디슨시 데이튼 스트리트 서
1210 6395호

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

공지예외적용 : 있음

명세서

청구범위

청구항 1

SRL(semantic role labeling) 기반의 문장 분석 방법에 있어서,

SRL 문장 분석 장치가 일반 특징 그룹에 포함되는 복수의 제 1 특징 및 제 1 언어 특징 그룹에 포함되는 복수의 제 2 특징을 기반으로 문장 내의 문장 성분을 특징 벡터로 결정하는 단계; 및

상기 SRL 문장 분석 장치가 상기 특징 벡터를 기반으로 상기 문장 성분에 대한 시멘틱 롤을 결정하는 단계를 포함하되,

상기 복수의 제 1 특징은 논항(argument) 또는 술어(predicate)인 상기 문장 성분의 상기 문장 내에서 문법적 역할에 대한 정보를 지시하고,

상기 복수의 제 2 특징은 상기 문장 성분에 관련된 조사 또는 어미에 대한 정보를 지시하는 것을 특징으로 하되,

상기 제 1 언어 특징 그룹은 한국어 특징 그룹인 것을 특징으로 하고,

상기 복수의 제 1 특징은 상기 문장 성분과 상기 문장 내의 다른 논항 및 다른 술어와의 관계에 대한 정보를 더 지시하는 것을 특징으로 하는 SRL 기반의 문장 분석 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

제 1 항에 있어서,

상기 특징 벡터는 상기 문장 성분 에 대한 LMR (latent morpheme representation)을 더 포함하고,

상기 LMR은 상기 문장 성분에 대해 CCA(Canonical Correlation Analysis), 딥러닝 또는 브라운 클러스터링 중 적어도 하나를 수행하여 결정되는 것을 특징으로 하는 SRL 기반의 문장 분석 방법.

청구항 5

제 4 항에 있어서,

상기 복수의 제 1 특징은 상기 문장 성분의 어간을 지시하는 특징, 상기 문장 성분의 품사를 지시하는 특징, 상기 문장 성분의 격을 지시하는 특징을 포함하고,

상기 복수의 제 1 특징은 상기 문장 성분이 상기 논항인 경우, 상기 문장 내에서 상기 문장 성분의 좌측 및 우측에 위치한 자식격 문장 성분, 형제격 문장 성분의 어간 및 품사를 지시하는 특징을 더 포함하고,

상기 복수의 제 1 특징은 상기 문장 성분이 상기 술어인 경우, 상기 문장 성분의 부모격 문장 성분, 자식격 문장 성분의 어간 및 상기 자식격 문장 성분의 품사 집합 및 격 집합을 지시하는 특징을 포함하는 것을 특징으로 하는 SRL 기반의 문장 분석 방법.

청구항 6

제 5 항에 있어서,

상기 복수의 제 2 특징은 상기 문장 성분에 조사의 존재 여부를 지시하는 특징, 상기 조사의 언어학적 분류를

지시하는 특징, 상기 조사를 구성하는 형태소의 개수를 지시하는 특징, 상기 조사를 구성하는 각각의 형태소를 지시하는 특징을 포함하고,

상기 복수의 제 2 특징은 상기 문장 성분에 어미의 존재 여부를 지시하는 특징, 상기 어미의 언어학적 분류를 지시하는 특징, 상기 어미를 구성하는 형태소의 개수를 지시하는 특징, 상기 어미를 구성하는 형태소 각각을 지시하는 특징으로 더 포함하는 것을 특징으로 하는 SRL 기반의 문장 분석 방법.

청구항 7

제 6 항에 있어서, 상기 SRL 문장 분석 장치가 상기 특징 벡터를 기반으로 상기 문장 성분에 대한 시멘틱 롤을 결정하는 단계는,

아래의 수학적식을 기반으로 수행되고,

$$p(y|x) = Z(x)^{-1} \prod_{i=1}^x \exp \sum_m \lambda_m f_m(y_{i-1}, y_i, x, i)$$

상기 $x = x_1, \dots, x_n$ 는 상기 문장에 포함되는 복수의 문장 성분 각각의 특징 벡터를 지시하고,

상기 $y = y_1, \dots, y_n$ 는 상기 복수의 문장 성분 각각에 대응되는 복수의 시멘틱 롤 각각을 지시하고,

상기 λ_m 는 미리 입력된 말뭉치를 기반으로 한 트레이닝을 통해 결정된 가중치이고,

상기 $f_m(y_{i-1}, y_i, x, i)$ 는 특징 함수(feature function)이고,

상기 특징 함수는 상기 문장 성분의 이전 문장 성분의 의미역 태그에서 상기 문장 성분의 의미역 태그가 나타날 제 1 확률과 상기 문장 성분의 의미역 태그가 나타날 제 2 확률에 대한 함수인 것을 특징으로 하는 SRL 기반의 문장 분석 방법.

청구항 8

제 1 항 및 제 4항 내지 제 7 항 중 어느 한 항에 따른 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

청구항 9

SRL(semantic role labeling) 기반 문장 분석 장치에 있어서,

SRL 기반 문장 분석 장치는 프로세서를 포함하고,

상기 프로세서는 일반 특징 그룹에 포함되는 복수의 제 1 특징 및 제 1 특징 그룹에 포함되는 복수의 제 2 특징을 기반으로 문장 내의 문장 성분을 특징 벡터로 결정하고,

상기 특징 벡터를 기반으로 상기 문장 성분에 대한 시멘틱 롤을 결정하도록 구현되되,

상기 복수의 제 1 특징은 논항(argument) 또는 술어(predicate)인 상기 문장 성분의 상기 문장 내에서 문법적 역할에 대한 정보를 지시하고,

상기 복수의 제 2 특징은 상기 문장 성분에 관련된 조사 또는 어미에 대한 정보를 지시하는 것을 특징으로 하되,

상기 제 1 언어 특징 그룹은 한국어 특징 그룹인 것을 특징으로 하고,

상기 복수의 제 1 특징은 상기 문장 성분과 상기 문장 내의 다른 논항 및 다른 술어와의 관계에 대한 정보를 더 지시하는 것을 특징으로 하는 SRL 기반 문장 분석 장치.

청구항 10

삭제

청구항 11

삭제

청구항 12

제 9 항에 있어서,

상기 특징 벡터는 상기 문장 성분 에 대한 LMR (latent morpheme representation)을 더 포함하고,

상기 LMR은 상기 문장 성분에 대해 CCA(Canonical Correlation Analysis), 딥러닝 또는 브라운 클러스터링 중 적어도 하나를 수행하여 결정되는 것을 특징으로 하는 SRL 기반 문장 분석 장치.

청구항 13

제 12 항에 있어서,

상기 복수의 제 1 특징은 상기 문장 성분의 어간을 지시하는 특징, 상기 문장 성분의 품사를 지시하는 특징, 상기 문장 성분의 격을 지시하는 특징을 포함하고,

상기 복수의 제 1 특징은 상기 문장 성분이 상기 논항인 경우, 상기 문장 내에서 상기 문장 성분의 좌측 및 우측에 위치한 자식격 문장 성분, 형제격 문장 성분의 어간 및 품사를 지시하는 특징을 더 포함하고,

상기 복수의 제 1 특징은 상기 문장 성분이 상기 술어인 경우, 상기 문장 성분의 부모격 문장 성분, 자식격 문장 성분의 어간 및 상기 자식격 문장 성분의 품사 집합 및 격 집합을 지시하는 특징을 포함하는 것을 특징으로 하는 SRL 기반 문장 분석 장치.

청구항 14

제 13 항에 있어서,

상기 복수의 제 2 특징은 상기 문장 성분에 조사의 존재 여부를 지시하는 특징, 상기 조사의 언어학적 분류를 지시하는 특징, 상기 조사를 구성하는 형태소의 개수를 지시하는 특징, 상기 조사를 구성하는 각각의 형태소를 지시하는 특징을 포함하고,

상기 복수의 제 2 특징은 상기 문장 성분에 어미의 존재 여부를 지시하는 특징, 상기 어미의 언어학적 분류를 지시하는 특징, 상기 어미를 구성하는 형태소의 개수를 지시하는 특징, 상기 어미를 구성하는 형태소 각각을 지시하는 특징으로 더 포함하는 것을 특징으로 하는 SRL 기반 문장 분석 장치.

청구항 15

제 14 항에 있어서,

상기 시멘틱 롤은 아래의 수학적식을 기반으로 결정되고

$$p(y|x) = Z(x)^{-1} \prod_{i=1}^x \exp \sum_m \lambda_m f_m(y_{i-1}, y_i, x, i)$$

상기 $x = x_1, \dots, x_n$ 는 상기 문장에 포함되는 복수의 문장 성분 각각의 특징 벡터를 지시하고,

상기 $y = y_1, \dots, y_n$ 는 상기 복수의 문장 성분 각각에 대응되는 복수의 시멘틱 롤 각각을 지시하고,

상기 λ_m 는 미리 입력된 말뭉치를 기반으로 한 트레이닝을 통해 결정된 가중치이고,

상기 $f_m(y_{i-1}, y_i, x, i)$ 는 특징 함수(feature function)이고,

상기 특징 함수는 상기 문장 성분의 이전 문장 성분의 의미역 태그에서 상기 문장 성분의 의미역 태그가 나타날 제 1 확률과 상기 문장 성분의 의미역 태그가 나타날 제 2 확률에 대한 함수인 것을 특징으로 하는 SRL 기반 문

장 분석 장치.

발명의 설명

기술 분야

- [0001] 본 발명은 문장 분석 방법 및 장치에 관한 것으로서, 보다 상세하게는, SRL(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 성분 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 고려함으로써, 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 SRL(semantic role labeling) 기반의 문장 분석 방법은 SRL 문장 분석 장치가 일반 특징 그룹에 포함되는 복수의 제1 특징 및 제1 언어 특징 그룹에 포함되는 복수의 제2 특징을 기반으로 문장 내의 문장 성분을 특징 벡터로 결정하는 단계와 상기 SRL 문장 분석 장치가 상기 특징 벡터를 기반으로 상기 문장 성분에 대한 시멘틱 롤을 결정하는 단계를 포함할 수 있되, 상기 복수의 제1 특징은 논항(argument) 또는 술어(predicate)인 상기 문장 성분의 상기 문장 내에서 문법적 역할에 대한 정보를 지시하고, 상기 복수의 제2 특징은 상기 문장 성분에 관련된 조사 또는 어미에 대한 정보를 지시할 수 있다.
- [0010] 한편, 상기 제1 언어 특징 그룹은 한국어 특징 그룹인 것을 특징으로 할 수 있다.
- [0011] 또한, 상기 복수의 제1 특징은 상기 문장 성분과 상기 문장 내의 다른 논항 및 다른 술어와의 관계에 대한 정

보를 더 지시할 수 있다.

[0012] 또한, 상기 특징 벡터는 상기 문장 성분 에 대한 LMR (latent morpheme representation)을 더 포함하고, 상기 LMR은 상기 문장 성분에 대해 CCA, 딥러닝 또는 브라운 클러스터링 중 적어도 하나를 수행하여 결정될 수 있다.

[0013] 또한, 상기 복수의 제1 특징은 상기 문장 성분의 어간을 지시하는 특징, 상기 문장 성분의 품사를 지시하는 특징, 상기 문장 성분의 격을 지시하는 특징을 포함하고, 상기 복수의 제1 특징은 상기 문장 성분이 상기 논항인 경우, 상기 문장 내에서 상기 문장 성분의 좌측 및 우측에 위치한 자식격 문장 성분, 형제격 문장 성분의 어간 및 품사를 지시하는 특징을 더 포함하고, 상기 복수의 제1 특징은 상기 문장 성분이 상기 술어인 경우, 상기 문장 성분의 부모격 문장 성분, 자식격 문장 성분의 어간 및 상기 자식격 문장 성분의 품사 집합 및 격 집합을 지시하는 특징을 포함할 수 있다.

[0014] 또한, 상기 복수의 제2 특징은 상기 문장 성분에 조사의 존재 여부를 지시하는 특징, 상기 조사의 언어학적 분류를 지시하는 특징, 상기 조사를 구성하는 형태소의 개수를 지시하는 특징, 상기 조사를 구성하는 각각의 형태소를 지시하는 특징을 포함하고, 상기 복수의 제2 특징은 상기 문장 성분에 어미의 존재 여부를 지시하는 특징, 상기 어미의 언어학적 분류를 지시하는 특징, 상기 어미를 구성하는 형태소의 개수를 지시하는 특징, 상기 어미를 구성하는 형태소 각각을 지시하는 특징으로 더 포함할 수 있다.

[0015] 또한, , 상기 SRL 문장 분석 장치가 상기 특징 벡터를 기반으로 상기 문장 성분에 대한 시멘틱 롤을 결정하는 단계는,

[0016] 아래의 수학적식을 기반으로 수행되고,
$$p(y|x) = Z(x)^{-1} \prod_{i=1}^x \exp \sum_m \lambda_m f_m(y_{i-1}, y_i, x, i)$$
, 상기 $x = x_1, \dots, x_n$

는 상기 문장에 포함되는 복수의 문장 성분 각각의 특징 벡터를 지시하고, 상기 $y = y_1, \dots, y_n$ 는 상기 복수의 문장 성분 각각에 대응되는 복수의 시멘틱 롤 각각을 지시하고, 상기 λ_m 는 미리 입력된 말뭉치를 기반으로 한 트레이닝을 통해 결정된 가중치이고, 상기 $f_m(y_{i-1}, y_i, x, i)$ 는 특징 함수(feature function)이고, 상기 특징 함수는 상기 문장 성분의 이전 문장 성분의 의미역 태그에서 상기 문장 성분의 의미역 태그가 나타날 제1 확률과 상기 문장 성분의 의미역 태그가 나타날 제2 확률에 대한 함수일 수 있다.

[0017] 본 발명의 다른 측면에 따른 SRL(semantic role labeling) 기반 문장 분석 장치에 있어서, SRL 기반 문장 분석 장치는 프로세서를 포함하고, 상기 프로세서는 일반 특징 그룹에 포함되는 복수의 제1 특징 및 제 1 언어 특징 그룹에 포함되는 복수의 제2 특징을 기반으로 문장 내의 문장 성분을 특징 벡터로 결정하고, 상기 특징 벡터를 기반으로 상기 문장 성분에 대한 시멘틱 롤을 결정하도록 구현될 수 있되, 상기 복수의 제1 특징은 논항(argument) 또는 술어(predicate)인 상기 문장 성분의 상기 문장 내에서 문법적 역할에 대한 정보를 지시하고, 상기 복수의 제2 특징은 상기 문장 성분에 관련된 조사 또는 어미에 대한 정보를 지시할 수 있다.

[0018] 한편, 상기 제 1 언어 특징 그룹은 한국어 특징 그룹인 것을 특징으로 할 수 있다.

[0019] 또한, 상기 복수의 제 1 특징은 상기 문장 성분과 상기 문장 내의 다른 논항 및 다른 술어와의 관계에 대한 정보를 더 지시할 수 있다.

[0020] 또한, 상기 특징 벡터는 상기 문장 성분 에 대한 LMR (latent morpheme representation)을 더 포함하고, 상기 LMR은 상기 문장 성분에 대해 CCA, 딥러닝 또는 브라운 클러스터링 중 적어도 하나를 수행하여 결정될 수 있다.

[0021] 또한, 상기 복수의 제1 특징은 상기 문장 성분의 어간을 지시하는 특징, 상기 문장 성분의 품사를 지시하는 특징, 상기 문장 성분의 격을 지시하는 특징을 포함하고, 상기 복수의 제1 특징은 상기 문장 성분이 상기 논항인 경우, 상기 문장 내에서 상기 문장 성분의 좌측 및 우측에 위치한 자식격 문장 성분, 형제격 문장 성분의 어간 및 품사를 지시하는 특징을 더 포함하고, 상기 복수의 제1 특징은 상기 문장 성분이 상기 술어인 경우, 상기 문장 성분의 부모격 문장 성분, 자식격 문장 성분의 어간 및 상기 자식격 문장 성분의 품사 집합 및 격 집합을 지시하는 특징을 포함할 수 있다.

[0022] 또한, 상기 복수의 제2 특징은 상기 문장 성분에 조사의 존재 여부를 지시하는 특징, 상기 조사의 언어학적 분류를 지시하는 특징, 상기 조사를 구성하는 형태소의 개수를 지시하는 특징, 상기 조사를 구성하는 각각의 형태소를 지시하는 특징을 포함하고, 상기 복수의 제2 특징은 상기 문장 성분에 어미의 존재 여부를 지시하는 특징,

상기 어미의 언어학적 분류를 지시하는 특징, 상기 어미를 구성하는 형태소의 개수를 지시하는 특징, 상기 어미를 구성하는 형태소 각각을 지시하는 특징으로 더 포함할 수 있다.

[0023] 또한, 상기 시멘틱 롤은 아래의 수학적식을 기반으로 결정되고,

$$p(y|x) = Z(x)^{-1} \prod_{i=1}^x \exp \sum_m \lambda_m f_m(y_{i-1}, y_i, x, i)$$

, 상기 $x = x_1, \dots, x_n$ 는 상기 문장에 포함되는 복수의 문장

성분 각각의 특징 벡터를 지시하고, 상기 $y = y_1, \dots, y_n$ 는 상기 복수의 문장 성분 각각에 대응되는 복수의 시멘틱 롤 각각을 지시하고, 상기 λ_m 는 미리 입력된 말뭉치를 기반으로 한 트레이닝을 통해 결정된 가중치이고, 상기 $f_m(y_{i-1}, y_i, x, i)$ 는 특징 함수(feature function)이고, 상기 특징 함수는 상기 문장 성분의 이전 문장 성분의 의미역 태그에서 상기 문장 성분의 의미역 태그가 나타날 제1 확률과 상기 문장 성분의 의미역 태그가 나타날 제2 확률에 대한 함수일 수 있다.

발명의 효과

[0024] 전술한 본원의 과제 해결 수단에 의하면, 본 발명은 일반 언어 특징과 함께 특정 언어의 특징을 추가적으로 고려함으로써, 보다 정확한 분장 분석을 가능하도록 하는 방법 및 장치를 제공할 수 있다.

[0025] 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

[0026] 도 1은 영어 문장에 적용된 SRL을 나타낸 개념도이다.

도 2는 본 발명의 실시예에 따른 SRL을 위해 특징(feature) 기반의 국문 문장 분석 방법을 나타낸 개념도이다.

도 3은 본 발명의 실시예에 따른 SRL을 위한 특징 기반의 국문 문장 분석 방법을 나타낸 순서도이다.

도 4는 본 발명의 실시예에 따른 SRL을 위한 특징 기반의 국문 문장 분석 장치를 나타낸 개념도이다.

발명을 실시하기 위한 구체적인 내용

[0027] 후술하는 본 발명에 대한 상세한 설명은, 본 발명이 실시될 수 있는 특정 실시예를 예시로서 도시하는 첨부 도면을 참조한다. 이들 실시예는 당업자가 본 발명을 실시할 수 있기에 충분하도록 상세히 설명된다. 본 발명의 다양한 실시예는 서로 다르지만 상호 배타적일 필요는 없음이 이해되어야 한다. 예를 들어, 여기에 기재되어 있는 특정 형상, 구조 및 특성은 일 실시예와 관련하여 본 발명의 정신 및 범위를 벗어나지 않으면서 다른 실시예로 구현될 수 있다. 또한, 각각의 개시된 실시예 내의 개별 구성요소의 위치 또는 배치는 본 발명의 정신 및 범위를 벗어나지 않으면서 변경될 수 있음이 이해되어야 한다. 따라서, 후술하는 상세한 설명은 한정적인 의미로서 취하려는 것이 아니며, 본 발명의 범위는, 적절하게 설명된다면, 그 청구항들이 주장하는 것과 균등한 모든 범위와 더불어 첨부된 청구항에 의해서만 한정된다. 도면에서 유사한 참조 부호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭한다.

[0028] 이하, 도면들을 참조하여 본 발명의 바람직한 실시예들을 보다 상세하게 설명하기로 한다.

[0029] SRL(semantic role labeling)은 자동적으로 문장 내에 논항(argument)와 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.

[0030] 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.

[0031] SRL은 문장에서 주어진 술어(또는 동사)의 논항을 식별하고, 논항에 시멘틱 롤(의미 역할, 의미역)을 부여하는 것이다. 논항은 어떠한 술어가 절이나 문장을 만들기 위해 필요로 하는 필수적인 의미 성분들을 말한다.

[0032] 아래의 표 1은 PropBank에 의해 정의된 17가지의 시멘틱 롤을 나타낸 표이다.

[0033] <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

[0034]

[0035] 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.

[0036] 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.

[0037] ARGM-ADV는 부사적 어구(adverbial), ARGM-CAU는 원인(cause), ARGM-DIR은 방향(direction), ARGM-는 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 동자(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.

[0038] 도 1은 영어 문장에 적용된SRL을 나타낸 개념도이다.

[0039] 도 1을 참조하면, The luxury auto maker last year sold 1,214 cars in the U.S. 라는 문장에 대한 SRL이 제시된다.

[0040] 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.

[0041] 이하, 본 발명의 실시예에서는 국문 문장의 특성을 고려하여 국문 문장에 포함되는 문장 성분들에 시멘틱 롤을 부여하는 방법을 개시한다.

[0042] 도 2는 본 발명의 실시예에 따른 SRL을 위해 특징(feature) 기반의 국문 문장 분석 방법을 나타낸 개념도이다.

[0043] 도 2를 참조하면, 본 발명의 실시예에 따른 특징 기반의 국문 문장 분석은 3가지의 특징 그룹을 기반으로 수행될 수 있다.

[0044] 국문 문장 분석을 위한 3가지의 특징 그룹은 일반 특징 그룹(200), 한국어 특징 그룹(220), LMR(latent morpheme representation) 그룹(240)을 포함할 수 있다.

[0045] 일반 특징 그룹(200), 한국어 특징 그룹(220) 및 LMR 특징 그룹(240)은 아래의 표 2, 표 3 및 표 4 각각과 같은 하위 항목(또는 특징)을 포함할 수 있다. 표 2는 일반 특징 그룹(200)의 하위 항목(또는 일반 특징 그룹(200)에 포함되는 특징), 표 3은 한국어 특징 그룹(220)의 하위 항목(또는 한국어 특징 그룹(220)에 포함되는 특징), 표

4는 LMR 특징 그룹(240)의 하위 항목(또는 LMR 특징 그룹(240)에 포함되는 특징)을 포함할 수 있다.

[0046]

<표 2>

Feature(특징)	Description(설명)
A-Stem, P-Stem	Stem of an argument and a predicate 논항과 술어의 어간
A-POS_Lv1, P-POS_Lv1	Coarse-grained POS of A-Stem and P-Stem 개략적으로 분류된 논항과 술어의 어간의 품사
A-POS_Lv2, P-POS_Lv2	Fine-grained POS of A-Stem and P-Stem 정교하게 분류된 논항과 술어의 어간의 품사
A-Case, P-Case	Case of A-Stem and P-Stem 논항과 술어의 격
A-LeftmostChildStem	Stem of the leftmost child of an argument 논항의 최좌측에 위치한 자식격의 어간
A-LeftSiblingStem	Stem of the left sibling of an argument 논항의 좌측에 위치한 형제격의 어간
A-LeftSiblingPOS_Lv1	Coarse-grained POS of A-LeftSiblingStem 개략적으로 결정된 A-LeftSiblingStem의 품사
A-LeftSiblingPOS_Lv2	Fine-grained POS of A-LeftSiblingStem 정교하게 결정된 A-LeftSiblingStem의 품사
A-RightSiblingPOS_Lv1	Coarse-grained POS of a stem of the right sibling of an argument 개략적으로 결정된 논항의 오른쪽 형제격의 어간의 품사
A-RightSiblingPOS_Lv2	Fine-grained POS of a stem of the right sibling of an argument 정교하게 결정된 논항의 오른쪽 형제격의 어간의 품사
P-ParentStem	Stem of a parent of a predicate 술어의 부모격의 어간
P-ChildStemSet	Set of stems of children of a predicate 술어의 자식격의 어간의 집합
P-ChildPOSSet_Lv1	Set of coarse POS of P-ChildStemSet P-ChildStemSet의 개략적으로 결정된 품사의 집합
P-ChildCaseSet	Set of cases of P-childStemSet P-childStemSet의 격의 집합

[0047]

[0048]

<표 3>

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 an 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

[0049]

[0050] <표 4>

Feature(특징)	Description(설명)
A-StemRepr	Stem representation of an argument 논항의 어근 표현
A-JosaRepr	Josa representation of an argument 논항의 조사 표현
A-EomiRepr	Eomi representation of an argument 논항의 어미 표현

[0051]

[0052] 본 발명의 실시예에 따르면, 문장에 포함된 문장 성분들은 표2 내지 표 4에 개시된 특징들을 기반으로 하나의 특징 벡터로 표현될 수 있다.

[0053] 이하, 3개의 특징 그룹에 포함된 복수의 특징 각각에 대해 구체적으로 개시한다.

[0054] 우선, 일반 특징 그룹(200)에 포함되는 특징에 대해 개시한다.

[0055] 일반 특징 그룹(200)에서 Stem은 어간을 의미할 수 있다. 한국어의 과 같이 굴절하는 단어는 한 단어의 개념적 의미를 갖는 어간과 문법적 기능을 표시하는 어미로 구성되는데, 어미는 문법적 기능에 따라 변화하지만 어간은 한 단어의 개념적 의미를 나타내는 것이기 때문에 변화하지 않는 고정된 요소일 수 있다.

[0056] POS_Lv1은 제1 레벨의 개략적인 분류로 논항 또는 술어의 1차적으로 분류된 품사(명사, 동사, 형용사, 부사 등)를 의미할 수 있다.

[0057] POS_Lv2는 제2 레벨의 정교한 분류로 논항 또는 술어의 2차적으로 분류된 품사(용언불가능 보통명사, 지시시간 부사, 성상 형용사 등)을 의미할 수 있다.

[0058] Case는 격으로서 주격, 목적격, 부사격 등과 같은 논항 또는 술어의 문장 내의 격을 의미할 수 있다.

[0059] 이하에서는 예시적으로 “ 이 책은 성종 때 한글로 번역 출판된 이래 영조 5년 때까지 여러 번 중간됐으며 정조 20년에는 이륜 행실도와 합쳐져 오륜행실도란 책이 이루어졌다” 라는 문장을 기반으로 각 특징 그룹에 포함된 복수의 특징이 개시된다.

[0060] 위의 예시적인 문장은 아래의 표 5와 같이 분류될 수 있다.

[0061] <표 5>

논항 또는 술어 번호	의존 번호	
1	2	이[지시관형사]
2	9	책[용언불가능보통명사]+은[일반보조사] (주격)
3	6	성중_때[지시시간부사]
4	6	한글[용언불가능보통명사]+로[부사격조사]
5	6	번역[용언가능보통명사]
6	9	출판되[일반동사]+L_이래[종속연결어미]
7	9	영조_5_년_때[용언불가능보통명사]+까지[부사격조사]
8	9	여러_번[용언불가능보통명사]
9	15	중간되[일반동사]+였[과거시제선어말어미]+으며[대등연결어미]
10	15	정조_20_년[용언불가능보통명사]+에는[부사격조사]
11	12	이론행실도[서적고유명사]+와[부사격조사]
12	15	합쳐지[일반동사]+어[종속연결어미]
13	14	오론행실도[서적고유명사]+란[관형격조사]
14	15	책[용언불가능보통명사]+이[주격조사]
15	0	이루어지[일반동사]+였[과거시제선어말어미]+다[평서형종결어미]+.[문미기호]

[0062]

[0063] A-Stem은 동사에 붙어 있는 논항을 지시할 수 있다. 예시 문장에서 책, 성중_때, 한글, 번역, 여러_번, 이론행실도가 A-Stem일 수 있다.

[0064] P-Stem은 문장에 있는 일반 동사(또는 술어)를 지시할 수 있다. 예시 문장에서 출판되, 중간되, 합쳐지, 이루어지가 P-Stem일 수 있다.

[0065] A-POS_Lv1은 논항의 첫 번째 level 품사(명사, 부사, 형용사 등)를 지시할 수 있다. 예시 문장에서 ‘책’의 A-POS_Lv1은 명사이고 ‘성중_때’의 A-POS_Lv1은 부사일 수 있다.

[0066] A-POS_Lv2은 논항의 두 번째 level 품사(예를 들어, 용언불가능보통명사, 지시시간부사, 성상형용사 등)를 지시할 수 있다. 예시 문장에서 ‘책’의 A-POS_Lv2는 용언불가능보통명사이고 ‘성중_때’의 A-POS_Lv2는 지시시간부사일 수 있다.

[0067] P-POS_Lv1은 술어의 첫 번째 level 품사(동사 등)를 지시할 수 있다. 예시 문장에서 ‘출판되’의 P-POS_Lv1은 동사일 수 있다.

[0068] P-POS_Lv2는 술어의 두 번째 level 품사(예를 들어, 일반 동사 등)를 지시할 수 있다. 예시 문장에서 ‘출판되’의 P-POS_Lv2은 일반 동사일 수 있다.

[0069] A-Case는 논항의 격(예를 들어, 주격, 목적격, 부사격 등)을 지시할 수 있다. 예시 문장에서 ‘책’의 A-Case는 주격일 수 있다.

[0070] P-Case는 술어의 격을 지시할 수 있다.

[0071] A-LeftSiblingStem은 현재 논항 이전에 나타난 모든 논항들 각각을 지시할 수 있다. 예시 문장에서 ‘책’의 이전 논항은 ‘이’일 수 있다.

[0072] A-LeftmostChildStem은 현재 논항 이전의 논항 중 의존 번호가 현재 논항의 번호인 어간(stem) 중 현재 논항의 번호와 가장 가까운 어간(stem)을 지시할 수 있다. 예시 문장에서 현재 논항이 ‘책’일 때 ‘책’의 이전 논항인 ‘이’의 의존 번호는 2로서 현재 ‘책’의 번호일 수 있다. ‘이’는 현재 논항인 책과 가장 가까운 어간이므로 ‘책’의 A-LeftmostChildStem은 ‘이’일 수 있다.

- [0073] A-LeftSiblingPOS_Lv1은 A-LeftSibling의 첫 번째 level 품사를 지시할 수 있다. 예시 문장에서 ‘책’의 LeftSibling인 ‘이’의 POS_Lv1은 관형사일 수 있다. 즉, ‘책’의 A-LeftSiblingPOS_Lv1은 관형사일 수 있다.
- [0074] A-LeftSiblingPOS_Lv2은 A-LeftSibling의 두 번째 level 품사를 지시할 수 있다. 예시 문장에서 ‘책’의 LeftSibling인 ‘이’의 POS_Lv2는 지시관형사일 수 있다. 즉, ‘책’의 A-LeftSiblingPOS_Lv2는 지시관형사일 수 있다.
- [0075] A-RightSiblingPOS_Lv1은 A-RightSibling의 첫 번째 level 품사를 지시할 수 있다. 예시 문장에서 ‘책’의 RightSibling은 ‘출판되’, ‘영조_5_년_때’, ‘여러_번’일 수 있고, ‘출판되’, ‘영조_5_년_때’, ‘여러_번’의 POS_Lv1은 동사, 명사일 수 있다. 따라서, ‘책’의 A-RightSiblingPOS_Lv1은 동사, 명사일 수 있다.
- [0076] A-RightSiblingPOS_Lv2은 A-RightSibling의 두 번째 level 품사를 지시할 수 있다. 따라서, ‘책’의 RightSibling은 ‘출판되’, ‘영조_5_년_때’, ‘여러_번’일 수 있고, ‘출판되’, ‘영조_5_년_때’, ‘여러_번’의 POS_Lv2은 일반동사, 용언불가능보통명사일 수 있다. 따라서, ‘책’의 A-RightSiblingPOS_Lv1은 일반동사, 용언불가능보통명사일 수 있다.
- [0077] P-ParentStem은 현재 술어의 의존 번호가 자신의 번호인 술어를 지시할 수 있다. 예시문장에서 ‘중간되’의 의존번호 15로 ‘중간되’의 P-ParentStem은 ‘이루어지’일 수 있다.
- [0078] P-ChildPOSSet_Lv1은 P-Child의 POS_Lv1의 집합 POS_Lv1₁, POS_Lv1₂, POS_Lv1₃을 지시할 수 있다. 예시문장에서 ‘중간되’의 P-child는 ‘책’, ‘출판되’, ‘영조_5년_때’, ‘여러_번’일 수 있고, ‘중간되’의 POS_Lv1 set은 동사, 명사일 수 있다.
- [0079] P-ChildCaseSet은 P-ChildCase의 집합(예를 들어, 부사격, 주격)을 지시할 수 있다. 예시 문장에서 ‘중간되’의 P-child는 ‘책’, ‘출판되’, ‘영조_5년_때’, ‘여러_번’일 수 있고, P-child case는case는 주격일 수 있다.
- [0080] 다음으로 한국어 특징 그룹(220)의 하위 항목(또는 한국어 특징 그룹(220)에 포함되는 특징)이 개시된다.
- [0081] 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.
- [0082] A-JosaClass는 조사의 8개의 class중 하나를 지시할 수 있다. 예시 문장에서 조사가 ‘은’일 경우, ‘은’의 A-JosaClass는 일반보조사일 수 있다.
- [0083] 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.
- [0084] 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.
- [0085] A-JosaIdentity는 논항의 조사를 지시할 수 있다. 논항 ‘책’의 A-JosaIdentity는 ‘은’일 수 있다.
- [0086] 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
- [0087] A-EomiClass_Lv1은 어미의 14개의 클래스 중 하나를 지시할 수 있다.
- [0088] A-EomiClass_Lv2는 어미의 4개의 클래스 중 하나를 지시할 수 있다.
- [0089] A-EomiLength는 어미의 형태소의 수를 지시할 수 있다.
- [0090] A-EomiMorphemes는 어미의 형태소 각각을 지시할 수 있다.
- [0091] A-EomiIdentity는 논항의 어미를 지시할 수 있다.
- [0092] 다음으로 LMR 특징 그룹(240)의 하위 항목(또는 LMR 특징 그룹(240)에 포함되는 특징)이 개시된다.
- [0093] LMR 특징 그룹(240)에서는 1) CCA(Canonical Correlation Analysis)를 기반으로 한 선형 연속 표현linear

continuous representation, 2) Deep learning을 기반으로 한 비선형 연속 표현, 3) 브라운 클러스터링(brown clustering)을 기반으로 한 불연속 표현으로 논항의 어근, 조사, 어미가 표현될 수 있다.

[0094] A-StemRepr은 논항의 어근 표현으로 CCA, Deep, Brown 방법으로 압축적으로 표현된 형태일 수 있다.

[0095] A-JosaRepr은 논항의 조사 표현으로 CCA, Deep, Brown 방법으로 압축적으로 표현된 형태일 수 있다.

[0096] A-EomiRepr은 논항의 어미 표현으로 CCA, Deep, Brown 방법으로 압축적으로 표현된 형태일 수 있다.

[0097] 본 발명의 실시예에 따르면, 특정 문장의 문장 성분은 일반 특징 그룹(200)의 하위 항목(또는 일반 특징 그룹(200)에 포함되는 특징), 한국어 특징 그룹(220)의 하위 항목(또는 한국어 특징 그룹(220)에 포함되는 특징), LMR 특징 그룹(240)의 하위 항목(또는 LMR 특징 그룹(240)에 포함되는 특징)을 기반으로 특징 벡터로 표현될 수 있다.

[0098] 특징 벡터로 표현된 문장 성분은 CRF(conditional random field) 모델(또는 CRF 함수)을 기반으로 시멘틱 롤(semantic role)(또는 의미 역할, 의미역)을 할당받을 수 있다. 즉, 특징 벡터로 표현된 문장 성분에 대응되는 전술한 표 1에서 개시한 PropBank에 의해 정의된 17가지의 시멘틱 롤 중 하나의 시멘틱 롤이 CRF 모델을 기반으로 획득될 수 있다. 이하 본 발명의 실시예에서는 설명의 편의상 CRF 모델을 기반으로 PropBank에 의해 정의된 17가지의 시멘틱 롤을 할당하는 방법이 개시되나, 다른 모델을 기반으로 다른 방식으로 정의된 시멘틱 롤이 문장 성분에 할당될 수도 있다.

[0099] CRFs는 기계 학습을 기반으로 입력 데이터 열과, 클래스로 분류된 열의 조건부 확률을 계산할 수 있다.

[0100] 입력 데이터 x , 입력된 열에 대응되는 y 열을 각각 $x = \langle x_1, x_2, \dots, x_n \rangle$, $y = \langle y_1, y_2, \dots, y_n \rangle$ 라 하고 $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$ 가 주어졌을 때 CRF는 아래의 수식과 같이 표현될 수 있다.

[0101] 본 발명의 실시예에서는 하나의 문장을 구성하는 문장 성분의 특징 벡터가 $x = x_1, \dots, x_n$ 이고, 각각의 문장 성분에 대응되는 시멘틱 롤이 $y = y_1, \dots, y_n$ 일 수 있다. 다른 표현으로 Input X 는 문장을 구성하는 단어의 집합일 수 있고, Output Y 는 각 단어에 대응하는 의미역 태그일 수 있다. x_1 에 대응하는 의미역 태그는 y_1 일 수 있다.

[0102] 전술한 바와 같이 각각의 문장 성분은 어간, 접미사 형태소(suffix morpheme)를 포함할 수 있다. 시멘틱 롤은 표1에서 개시한 {NONE, ARG0, ..., ARG-TMP} 집합에서 결정될 수 있다. NONE은 의미역을 찾지 못한 경우를 의미할 수 있다.

[0103] <수학식 1>

$$p(y|x) = Z(x)^{-1} \prod_{i=1}^x \exp \sum_m \lambda_m f_m(y_{i-1}, y_i, x, i)$$

[0104] 수학식 1에서 $f_m(y_{i-1}, y_i, x, i)$ 는 특징 함수(feature function)일 수 있다. 특징 함수는 이전 문장 성분(또는 단어)의 의미역 태그에서 현재 문장 성분(또는 단어)의 의미역 태그가 나타날 확률(전이 확률)과 현재 단어의 의미역 태그가 나타날 확률(emission 확률)을 포함할 수 있다.

[0105] 즉, CRF 함수는 트레이닝에 의해 트레이닝되어 결정된 가중치(λ_m)를 기반으로 결정될 수 있다. 트레이닝을 통해 결정된 가중치를 기반으로 특정 문장 성분에 대해 가장 높은 확률을 가지는 시멘틱 롤이 결정될 수 있다.

[0106] 본 발명의 실시예에서는 위와 같은 3가지 특징 그룹에 기반하여 입력된 문장 성분에 대한 시멘틱 롤을 결정함으로써 보다 정확도가 높게 문장 성분에 대한 시멘틱 롤을 결정할 수 있다.

[0107] 이하에서는 본 발명의 실시예에 따른 SRL 기반의 문장 분석 방법이 사용된 경우, F1-스코어 결과가 개시된다.

[0108] 아래의 표 6은 말뭉치(corpus)로서 PKPB(penn Korean PropBank), 동아 뉴스 기사 말뭉치(이하, 설정 말뭉치)를 사용하여 테스트를 수행한 결과이다.

[0110] <표 6>

Corpus	Gen	Gen+Kor	Gen+Kor+LMR			
			CCA	Deep	Brown	All
PKPB	64.83	75.17%	75.51%	75.43%	75.55%	75.54%
설정 말뭉치	66.88	80.33%	80.88%	80.84%	80.77%	81.07%
PKPB + 설정말뭉치	64.86	78.61%	79.32%	79.44%	78.91%	79.20%

[0111]

[0112] 첫번째 경우는 일반 특징 그룹(Gen)의 하위 항목만으로 문장 성분에 대한 시멘틱 물을 탐색한 경우이다. 이때 정확도를 측정하는F1 스코어는 PKPB: 64.83%, 설정 말뭉치: 64.83%, PKPB+설정 말뭉치:64.83%를 획득하였다.

[0113] 두번째 경우는 일반 특징 그룹(Gen)의 하위 항목 및 한국어 특징 그룹(Kor)의 하위 항목만으로 문장 성분에 대한 시멘틱 물을 탐색한 경우이다. 이때 정확도를 측정하는F1 스코어는 PKPB: 75.17%, 설정 말뭉치: 80.33%, PKPB+설정말뭉치: 78.61%를 획득하였다. 한국어 특징 그룹의 하위 항목을 사용한 경우, F1 스코어의 증가를 확인할 수 있다. 즉, 본 발명의 실시예에 따른 한국어 특징 그룹이 사용되는 경우, 문장 성분에 대한 시멘틱 물 분석의 정확도가 증가할 수 있다.

[0114] 세번째 경우는 일반 특징 그룹(Gen)의 하위 항목, 한국어 특징 그룹(Kor)의 하위 항목 및 LMR 특징 그룹(LMR)의 하위 항목을 모두 사용하여 문장 성분에 대한 시멘틱 물을 탐색한 경우이다. LMR 특징 그룹에서1) CCA를 기반으로 한 경우, 2) Deep learning을 기반으로 한 경우, 3) 브라운 클러스터링을 기반으로 한 경우, 4) CCA, Deep learning 및 브라운 클러스터링을 모두 사용한 경우 각각에 대한 F1 스코어가 개시되어 있다. 4) CCA, Deep learning 및 브라운 클러스터링을 모두 사용한 경우 최대 F1 스코어 81.07%를 획득하였다.

[0115] 도 3은 본 발명의 실시예에 따른 SRL을 위한 특징 기반의 국문 문장 분석 방법을 나타낸 순서도이다.

[0116] 도 3에서는 전술한 특징 기반 국문 문장 분석 방법이 개시된다.

[0117] 도 3을 참조하면, CRF를 트레이닝 하여 가중치를 결정한다(단계 S300).

[0118] 문장 성분에 대한 시멘틱 물을 부여하기 위한CRF의 가중치가 트레이닝을 기반으로 결정될 수 있다. 가중치를 결정하기 위한 다양한 말뭉치(corpus)가 SRL 기반 문장 분석 장치로 입력될 수 있다. CRF 모델의 가중치는 입력된 말뭉치를 기반으로 트레이닝 되어 결정될 수 있다. 트레이닝을 위해 시멘틱 물을 할당받은 문장을 포함하는 말뭉치가 CRF에 입력되고, CRF는 입력된 말뭉치를 기반으로 가중치를 결정할 수 있다.

[0119] 문장을 문장 성분으로 분할한다(단계 S310).

[0120] SRL 문장 분석 장치는 문장 성분에 대한 시멘틱 물 분석을 위해 문장을 개별 문장 성분 단위로 분할할 수 있다.

[0121] 3개의 특징 그룹에 포함된 복수의 특징 각각을 기반으로 문장 성분을 특징 벡터로 표현한다(단계 S320).

[0122] 문장 성분은 전술한 3개의 일반 특징 그룹, 한국어 특징 그룹, LMR 그룹 각각에 포함되는 하위 항목(또는 특징)을 기반으로 특징 벡터로 분석될 수 있다.

[0123] 문장 내 문장 성분에 대한 시멘틱 물을 결정한다(단계 S330).

[0124] SRL 문장 분석 장치는 문장 성분의 특징 벡터를 입력받고 이전 문장 성분의 의미역 태그에서 현재 문장 성분의 의미역 태그가 나타날 확률(전이 확률)과 현재 단어의 의미역 태그가 나타날 확률(emission 확률)을 기반으로 문장 성분의 시멘틱 물을 결정할 수 있다.

[0125] 도 4는 본 발명의 실시예에 따른 SRL을 위한 특징 기반의 국문 문장 분석 장치를 나타낸 개념도이다.

[0126] 도 4에서는 전술한 특징 기반 국문 문장 분석 방법을 수행하는 SRL 기반 문장 분석 장치가 개시된다.

[0127] 도 4를 참조하면, SRL 기반 문장 분석 장치는 가중치 결정부(400), 문장 성분 결정부(410), 특징 벡터 결정부(420), 시멘틱 물 결정부(430) 및 프로세서(440)를 포함할 수 있다.

[0128] 가중치 결정부(400)는 CRF를 트레이닝 하여 가중치를 결정하기 위해 구현될 수 있다. 가중치 결정부(400)는 문장 성분에 대한 시멘틱 물을 부여하기 위한CRF의 가중치를 트레이닝을 기반으로 결정할 수 있다. 가중치를 결정하기 위한 다양한 말뭉치(corpus)가 SRL 기반 문장 분석 장치로 입력될 수 있다. CRF 모델의 가중치는 입력된

말뭉치를 기반으로 트레이닝 되어 결정될 수 있다.

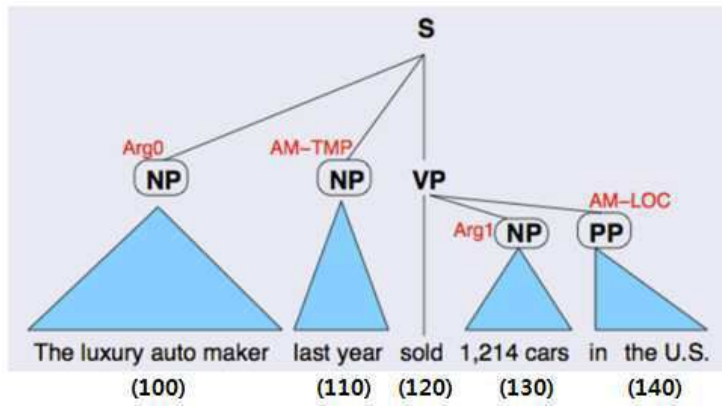
- [0129] 문장 성분 결정부(410)는 문장을 문장 성분으로 분할하기 위해 구현될 수 있다.
- [0130] 문장 성분 결정부(410)는 문장 성분에 대한 시멘틱 룰 분석을 위해 문장을 개별 문장 성분 단위로 분할하기 위해 구현될 수 있다.
- [0131] 특징 벡터 결정부(420)는 3개의 특징 그룹에 포함된 복수의 특징 각각을 기반으로 문장 성분을 특징 벡터로 표현하기 위해 구현될 수 있다. 문장 성분은 전술한 3개의 일반 특징 그룹, 한국어 특징 그룹, LMR 그룹 각각에 포함되는 하위 항목(또는 특징)을 기반으로 특징 벡터로 분석될 수 있다.
- [0132] 시멘틱 룰 결정부(430)는 문장 내 문장 성분에 대한 시멘틱 룰을 결정하기 위해 구현될 수 있다. 시멘틱 룰 결정부(430)는 문장 성분의 특징 벡터를 입력받고 이전 문장 성분의 의미역 태그에서 현재 문장 성분의 의미역 태그가 나타날 확률(전이 확률)과 현재 단어의 의미역 태그가 나타날 확률(emission 확률)을 기반으로 문장 성분의 시멘틱 룰을 결정할 수 있다.
- [0133] 프로세서(440)는 가중치 결정부(400), 문장 성분 결정부(410), 특징 벡터 결정부(420), 시멘틱 룰 결정부(430)의 동작을 제어하기 위해 구현될 수 있다.
- [0134] 이와 같은 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.
- [0135] 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.
- [0136] 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.
- [0137] 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.
- [0138] 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.

부호의 설명

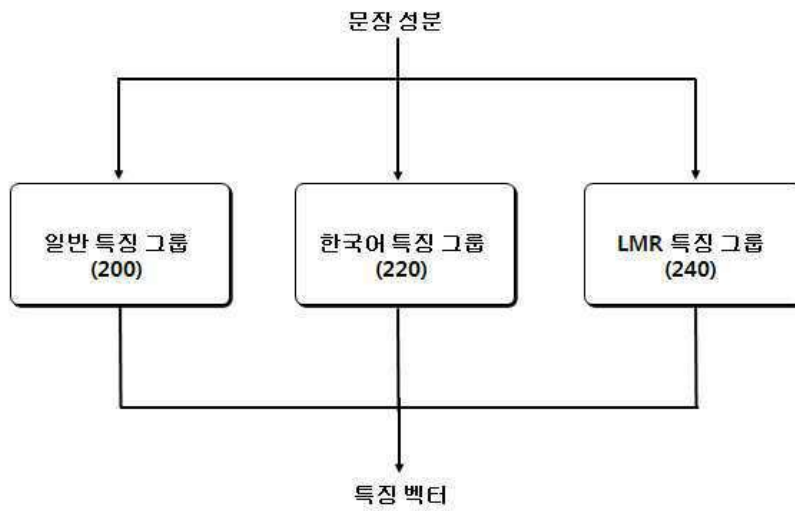
- [0139] 400: 가중치 결정부
- 410: 문장 성분 결정부
- 420: 특징 벡터 결정부
- 430: 시멘틱 룰 결정부
- 440: 프로세서

도면

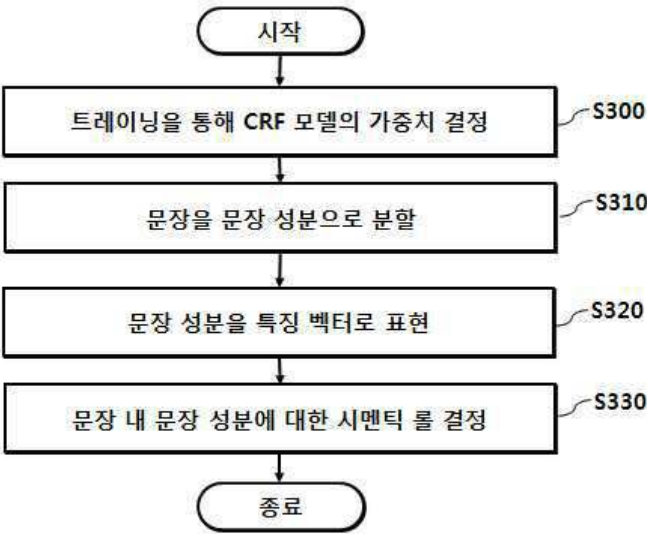
도면1



도면2



도면3



도면4

