



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0023387
(43) 공개일자 2012년03월13일

(51) 국제특허분류(Int. Cl.)
G06F 17/21 (2006.01) G06F 17/40 (2006.01)
(21) 출원번호 10-2010-0086449
(22) 출원일자 2010년09월03일
심사청구일자 2010년09월03일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
김선호
서울특별시 양천구 목동서로 130, 목동신시가지아파트 404-805 (목동)
윤준태
서울특별시 양천구 목동서로 100, 목동아파트 313동 1003호 (목동)
(74) 대리인
특허법인충현

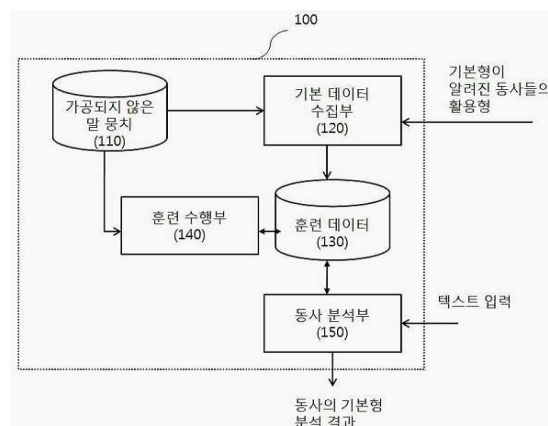
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 형태적 중의성 동사 분석 장치, 방법 및 그 기록 매체

(57) 요약

형태적 중의성이 존재하는 동사 분석 장치, 방법 및 그 기록 매체가 개시된다. 본 발명의 일 실시 예에 따른 형태적 중의성이 있는 동사 분석 장치는 레이블이 부착되지 않은 가공되지 않은 말뭉치에서 기본형이 파악된 중의성이 없는 제1동사활용과 상기 제1동사활용에 관련된 문맥 특징을 포함하는 학습 데이터를 수집하는 기본 데이터 수집부; 및 상기 학습 데이터에 기반하여 기본형과 문맥 사이의 분류자를 학습하는 훈련 수행부; 및 상기 말뭉치에서 기본형이 중의적인 제2동사활용 주변의 문맥 특징을 추출하고, 상기 학습된 분류자에 기반하여 기본형을 파악하는 동사 분석부; 기존 학습 데이터로 유도된 분류자로는 판단하기 어려운 예제를 파악하는 어려운 사례 판단부; 어려운 사례에 대한 기본형 판단을 위한 웹카운트를 수집하는 통신부; 어려운 사례를 추가하고 재학습하는 추가 훈련부를 포함한다. 본 발명의 실시 예들에 의하면, 말뭉치의 인위적인 분류 작업을 필요로 하지 않으면서도 텍스트 분석의 정확도 및 효율을 향상시킬 수 있고, 데이터베이스 구축을 포함한 텍스트 분석의 전 과정을 자동화할 수 있다.

대표도 - 도1a



(72) 발명자

박석

서울특별시 종로구 평창12길 16-6, 2동 202호 (평창동, 평창삼호빌라)

서정연

서울특별시 서초구 서초중앙로 206, 삼호가든 3차 B동 203호 (반포동)

이 발명을 지원한 국가연구개발사업

과제고유번호 핵06A1903

부처명 교육과학기술부

연구사업명 2단계 두뇌 한국(BK) 21 사업

연구과제명 PERVASIVE 컴퓨팅 기반의 의료시스템을 위한 S/W 인력 양성팀

주관기관 서강대학교 산학협력단

연구기간 2010.03.01 ~ 2011.02.28이 발명을 지원한 국가연구개발사업

과제고유번호 2009-0070211

부처명 교육과학기술부

연구사업명 기초 연구 지원 사업 여성 과학자 부문

연구과제명 자연어 처리를 이용한 바이오 텍스트 마이닝

주관기관 서강대학교 산학협력단

연구기간 2010.03.01 ~ 2011.02.28

특허청구의 범위

청구항 1

레이블이 부착되지 않은 가공되지 않은 말뭉치에서 기본형이 파악된 중의성이 없는 제1동사활용과 상기 제1동사활용에 관련된 문맥 특징을 포함하는 학습 데이터를 수집하는 기본 데이터 수집부; 및

상기 말뭉치에서 기본형이 중의적인 제2동사활용 주변의 문맥 특징을 추출하고, 상기 학습 데이터에 기반하여 상기 제2동사활용을 기본형과 매칭시키기 위한 분류자를 학습하는 훈련 수행부

를 포함하는, 형태적 중의성 동사 분석 장치.

청구항 2

제 1 항에 있어서,

입력의 텍스트가 입력되면, 텍스트에서 각각의 동사의 앞에 존재하는 문맥 특징을 추출하고 상기 분류자를 이용하여 상기 각각의 동사의 기본형을 판단하는 동사 분석부를 더 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 장치.

청구항 3

제 1 항에 있어서,

상기 훈련 수행부는

형태적 중의성이 존재하는 어절에서, 각각의 기본형과 연관되는 문맥 정보의 조건부 문맥 확률을 기준으로 상기 제2동사활용의 기본형을 판단하는 것을 특징으로 하는, 형태적 중의성 동사 분석 장치.

청구항 4

제 1 항에 있어서,

상기 훈련 수행부에서 중의성 있는 동사의 기본형 예측 시 첫 번째로 높은 확률값으로 예측되는 제1기본형과 두 번째로 높은 확률값으로 예측되는 제2기본형에 대해, 상기 확률값의 차를 이용하여 분류의 확신도를 평가하고, 실험 데이터와 현재까지 수집된 학습 데이터 사이의 유사도를 계산한 후, 상기 확신도와 상기 유사도에 기반하여 학습 데이터를 추가할 지 여부를 결정하는 어려운 사례 판단부를 더 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 장치.

청구항 5

제 1 항에 있어서,

상기 학습된 분류자로는 처리가 어렵다고 판단된 어려운 사례의 기본형 파악을 위해, 상기 어려운 사례의 주변 문맥 특징을 이용한 질의구를 검색 엔진에 전달한 후, 상기 검색 엔진으로부터 웹 카운트를 반환받는 통신부; 및

상기 웹 카운트를 이용하여 상기 어려운 사례의 기본형을 추출하고 학습 데이터로 추가하는 방식으로 상기 분류자를 학습하는 추가 훈련 수행부를 더 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 장치.

청구항 6

레이블이 부착되지 않은 가공되지 않은 말뭉치에서 기본형을 파악할 수 있는 중의성이 없는 제1동사활용의 예제로부터 상기 기본형과 관련된 문맥 특징을 학습 데이터로 수집하는 단계;

상기 말뭉치에서 기본형이 중의적인 제2동사활용의 기본형을 상기 학습 데이터에 기반하여 판단하는 단계; 및

상기 기본형을 판단하는 과정에서 사용된 문맥 특징을 상기 제2동사활용의 기본형 추출을 위한 학습 데이터로서 추가하는 단계

를 포함하는, 형태적 중의성 동사 분석 방법.

청구항 7

제 6 항에 있어서,

임의의 텍스트가 입력되면, 텍스트에서 기본형을 구하고자 하는 동사 활용의 앞에 등장하는 어절과 기본형과의 확률을 이용하여 상기 동사 활용의 기본형을 판단하는 단계를 더 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 방법.

청구항 8

제 7 항에 있어서,

상기 동사 활용의 기본형을 판단하는 단계는

상기 동사 활용 앞의 문맥들의 조사와 내용어들을 추출하여 기본형을 판단하는 단계를 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 방법.

청구항 9

제 8 항에 있어서,

상기 동사 활용의 기본형을 판단하는 단계는

상기 동사 활용의 앞에 가장 근접하게 위치하는 2개의 어절들을 기준으로 상기 동사 활용의 기본형을 판단하는 단계인 것을 특징으로 하는, 형태적 중의성 동사 분석 방법.

청구항 10

제 6 항에 있어서,

상기 제2동사활용의 기본형을 판단하는 단계는

상기 동사 활용의 가능한 기본형들과 문맥 정보 사이의 조건부 문맥 확률을 이용한 나이브 베이지언 분류자를 이용하여 상기 제2동사활용의 기본형을 판단하는 단계인 것을 특징으로 하는, 형태적 중의성 동사 분석 방법.

청구항 11

제 6 항에 있어서,

상기 제2동사활용에 대해 첫 번째로 높은 확률값으로 예측되는 제1기본형과 두 번째로 높은 확률값으로 예측되는 제2기본형을 파악하고, 상기 첫 번째로 높은 확률값과 상기 두 번째로 높은 확률값의 차에 기반하여 상기 제2동사활용을 위한 학습 데이터의 추가 여부를 결정하는 단계를 더 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 방법.

청구항 12

제 6 항에 있어서,

상기 제2동사활용의 기본형 파악을 위해 추가되는 학습 데이터와 현재까지 수집된 학습 데이터 사이의 유사도를 계산하고, 상기 유사도에 기반하여 상기 제2동사활용을 위한 학습 데이터의 추가 여부를 결정하는 단계를 더 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 방법.

청구항 13

제 6 항에 있어서,

현재까지 수집된 학습 데이터와 그로부터 유도된 분류자로는 처리가 어렵다고 판단된 어려운 사례의 기본형 파악을 위해, 상기 어려운 사례의 주변 문맥 특징을 이용한 질의 구를 검색 엔진에 전달하는 단계; 및

상기 검색 엔진으로부터 반환되는 웹 카운트를 이용하여 상기 어려운 사례의 기본형을 추출하고 학습 데이터로 추가하는 단계를 더 포함하는 것을 특징으로 하는, 형태적 중의성 동사 분석 방법.

청구항 14

제 6 항 내지 제 13 항 중 어느 한 항의 방법을 컴퓨터에서 수행할 수 있도록 프로그램으로 기록된 기록매체.

명세서

기술 분야

[0001] 본 발명은 텍스트의 형태소 분석 및 태깅에 관한 것으로, 특히, 형태적 중의성이 있는 동사 분석 장치, 방법 및 그 기록 매체에 관한 것이다.

배경 기술

[0002] 말뭉치를 이용하는 확률 모델에 기반한 기계 학습(machine learning) 방법들은 최근 자연어 처리 문제를 해결하는 데에 성공적으로 적용되고 있다. 예를 들어, 통계적 형태소 태깅 시스템 (statistical parts-of-speech tagging system)의 경우 언어의 종류에 상관없이 약 97%의 정확도를 보이고 있다. 하지만 3%에 해당하는 에러율도 고성능의 정확도 높은 언어 처리를 요구하는 실제 응용 시스템에서는 여전히 간과할 수 없는 수치이다. 이러한 오류는 대부분 기존 태깅 시스템 내부에 고려되고 있는 제한된 문맥과 단어 정보만으로는 해결할 수 없으며, 규칙 기반의 후처리를 수행하더라도 규칙 자체가 오류가 많은 사례로부터 일반화되기 때문에 성능 향상을 달성하기 어렵다.

[0003] 이러한 형태소 분석 오류 중에서도 동사-동사 형태적 중의성(verb-verb morphological ambiguity)에 기인한 오류는 한국어의 형태소 분석 및 태깅에서 가장 까다롭고 중요한 문제로 인식되고 있다. 해당 중의성 오류는 동사의 활용(conjugation) 시 발생하는 형태적 변이에 따른 것으로써 활용이란 동사 변화를 의미하며 한 단어가 문장 내에서 문법적 기능을 표시하기 위해 어미를 여러 가지 형태로 붙여서 사용하는 일을 의미한다. 한국어의 경우, 어간과 어미가 결합하는 과정에서 형태적인 변이가 발생하는 경우가 빈번하다.

[0004] 일반적으로 검색, 번역, 마이닝 등 자연어 처리 응용 시스템들은 문장이나 텍스트로부터 해당 단어들의 원형(base form)과 해당 단어의 품사를 찾는 형태소 분석 및 태깅 단계를 가장 기본적인 프로세스로 활용하고 있는데, 여기서 동사-동사 형태적 중의성이란 텍스트에 등장한 형태적 변이로부터 해당 단어의 원형을 찾아내는 데 있어서 그 가능한 원형의 해가 중의적임을 의미한다. 예를 들어 "파는"이라는 단어 (어절)가 등장하였을 때 그 어간의 원형은 "팔다"와 "파다"의 두 가지가 가능하며 올바른 형태소 원형은 해당 문장의 주변 문맥(context)이나 문서의 주제(topic)에 의존하여 한가지로 결정된다.

[0005] 이와 같이 형태적으로 중의성을 유발하는 동사들의 해당 원형을 찾아주는 문제를 이하, 동사-동사 형태적 중의성 해결 (verb-verb morphological disambiguation), VVMD로 줄여 표기하고, 동사-동사 형태적 중의성 (verb-verb morphological ambiguity)은 줄여 VVMA로 표기한다.

[0006] 또한, 활용 시 형태적 중의성을 유발하는 동사들은 대부분 의미(sense)적인 측면에서 다의어에 속하는 경우가 많아 텍스트상에 빈번하게 등장하는 동사들이므로 그 올바른 처리가 실제 언어 처리 성능 향상에 미치는 영향이 크다.

발명의 내용

해결하려는 과제

[0007] 본 발명이 이루고자 하는 첫 번째 기술적 과제는 동사 원형 인식을 위해 적용되는 기계 학습 기법에 필요한 학습 말뭉치를 사람의 수작업을 통한 분류나 클래스 (레이블) 정보 부착(annotation) 과정 없이, 레이블이 부착되지 않은 일반 말뭉치 (unlabeled corpus)와 웹 카운트를 이용하여, 형태적 중의성이 있는 동사를 자동으로 분석하여 형태소 분석 성능을 향상시킬 수 있는 동사 분석 장치를 제공하는데 있다.

[0008] 본 발명이 이루고자 하는 두 번째 기술적 과제는 상기의 형태적 중의성이 있는 동사 분석 장치에 적용되는 효과적인 동사 분석을 위한 기계 학습 방법을 제공하는 데 있다.

[0009] 본 발명이 이루고자 하는 세 번째 기술적 과제는 상기의 형태적 중의성이 있는 동사 분석 방법을 컴퓨터에서 실행시키기 위한 프로그램을 컴퓨터로 읽을 수 있는 기록매체를 제공하는 데 있다.

과제의 해결 수단

- [0010] 상기의 첫 번째 기술적 과제를 이루기 위하여, 본 발명의 일 실시 예에 따른 한국어 동사 분석 장치는 주변 문맥 (context) 정보를 이용하여 그 원형을 찾아낸다. 동사 원형 학습을 위해 사용되는 주변 문맥 데이터는 해당 동사들이 실제 중의성 없이 한가지 경우의 원형 도출이 가능한 활용 (conjugation)들의 예제들과 그 주변 문맥들에서 추출된다. 본 발명의 일 실시 예에 따른 형태적 중의성이 있는 동사 분석 장치는 중의성 있는 문제를 중의성 없는 동일한 영역의 문제로 매핑하는데, 중의성이 발생하지 않는 동사 활용형들을 이용하면 레이블이 부착되지 않은 가공되지 않은 말뭉치(unlabeled raw corpus)들로부터도 동사들의 클래스, 즉 원형에 대한 정보를 미리 알 수 있으므로 클래스 정보가 이미 부착된 학습 말뭉치처럼 이용할 수 있다.
- [0011] 상기의 두 번째 기술적 과제를 이루기 위하여, 본 발명의 일 실시 예에 따른 형태적 중의성이 있는 동사 분석 방법은 동사 원형 집합 V 를 정의하는 단계; 활용(conjugation) 시 형태적 중의성이 발견되는 동사 활용(어절) 집합 E_A 와 해당 원형들의 활용 시 중의성이 나타나지 않는 활용(어절) 집합 E_u 를 정의하는 단계; E_u 집합의 예제들로부터 관련된 문장을 추출하여 학습 데이터 D_{train} 를 구축하는 기본 데이터 수집 단계; 수집 데이터들로부터 각 동사 원형과 함께 나타나는 주변 문맥을 학습 자질 (learning feature)로 사용하여 동사 원형 클래스를 할당하는 분류자(classifier)를 기계 학습 방법에 따라 구현하는 단계; 및 E_A 어절이 포함된 문장들로부터 실험 데이터 집합 D_{test} 를 구축하는 단계를 포함한다.
- [0012] 본 발명의 다른 실시 예에 따른 형태적 중의성이 있는 동사 분석 방법은 현재까지의 학습 데이터로부터 유도된 분류자를 이용하여 실험 데이터에 대해 분류하기 힘든 어려운 예제 (hard example) 집합 D_{hard} 를 추출하는 선택적 샘플링 (selective sampling) 단계; D_{hard} 를 위한 해당 정보 추출을 위하여 web count를 구하고 D_{hard} 를 D_{train} 에 포함시켜 분류자를 재학습하는 단계를 더 포함한다.

발명의 효과

- [0013] 본 발명의 실시 예들에 의하면, 말뭉치의 인위적인 분류 작업 없이 교사 학습(supervised learning)의 효과를 가질 수 있는 동사 분석 방법을 제시하여 텍스트 분석의 정확도 및 효율을 향상시킬 수 있다. 또한 현재까지 학습된 데이터로부터 유도된 동사 원형 분류자를 새로운 데이터에 적용할 때, 기존에 이미 학습된 데이터와 다르거나 분류자의 결정이 확실하지 않은 어려운 예제들에 대해서만 기존 학습 데이터 집합에 추가하면서 학습 세트를 늘려나가는 점진적 학습 방법(incremental learning)과 웹 데이터를 함께 고려할 수 있는 방법을 제시하여 동사 분석 전 과정을 자동화할 수 있다.

도면의 간단한 설명

- [0014] 도 1a는 본 발명의 일 실시 예에 따른 형태적 중의성 동사 분석 장치의 도표이다.
 도 1b는 본 발명의 다른 실시 예에 따른 형태적 중의성 동사 분석 장치의 도표이다.
 도 2는 본 발명의 일 실시 예에 따른 형태적 중의성 동사 분석 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0015] 이하에서는 도면을 참조하여 본 발명의 바람직한 실시 예를 설명하기로 한다. 그러나, 다음에 예시하는 본 발명의 실시 예는 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 다음에 상술하는 실시 예에 한정되는 것은 아니다.
- [0016] 동사의 활용형은 시제, 상, 기타 문법적 카테고리에 의해 달라질 수 있는데, 특히 동사의 어미에서 그 변화가 두드러진다. 표 1은 다양한 어미를 가진 동사들을 보여준다.

표 1

단어	형태소 분석
갔다	가(동사)+았(과거 시제 선어말 어미)+ 다(종결형 어미)
가시었다	가(동사)+시었(존칭 과거 시제 선어말 어미) +다(종결형 어미)
간다	가(동사) +ㄴ(현재 시제 선어말 어미) +다(종결형 어미)
가는	가(동사) +는(관형사형 전성어미)
가기	가(동사) +기(명사형 전성어미)
가고	가(동사) +고(대등적 연결 어미)

표 2는 분석이 형태적 중의성을 보이는 동사 활용형들의 예와 그것의 가능한 형태소 분석의 예, 해당 원형의 중의성 없는 활용형들을 보여준다.

표 2

중의성있는 어절	기본 형	가능한 형태소 분석 결과 Analysis	Unambiguous conjugations of baseforms
들었, 들어, 들었고, 들어서, 들어야...	듣, 들	듣(hear,listen,work)/VB+았(past ending/PE) 들(hold,raise,give,cut,take,eat,clear,cast,be dyed,join,catch,feel,contain,suffer,break into,accommodate,ripe)/VB+ 였t/PE	듣고, 들고 듣는, 들은 ...
물어, 물었, 물어야, 물어서...	묻, 물	묻(ask,blame,charge,bury,be stained)/VB+았/PE 물(bite,pay,spoil,hold,stick to)/VB+ 였/PE	묻고, 물고 묻는, 무는
걸어, 걸었, 걸어서, 걸어야, 걸었고....	걷, 걸	걷(roll,walk,gather,remove,go out)/VB+ 였/PE 걸(hang,bet,risk,speak,call,expect,trip,sue,cast,set)/VB+ 였/PE	걷고, 걸고 걷기만, 걸기만 걷는게, 걷는게 ...
길어, 길었, 길어서, 길어야...	깎, 길	깎(draw,pump)/VB+았/PE 길(long)/VB+ 였/PE	깎게, 길게 깎고, 길고 깎는지, 길지 깎는다고, 길다고...
끄는, 끄내고, 끄다면, 끄는...	끄, 끌	끄(put out,turn out,extinguish,break,repay)/VB+는(adnominal ending)/ENRR1 끌(drag,lead,attract,prolong,install,pull out)/VB+ 는/ENTR1	끄고, 끌고 끄게, 끌게 끄도록, 끌도록 껴야, 끌어야...
까는, 까내고, 까는, 까다면...	까, 깎	까(peel,hatch,deduct,strike)/VB+는/ENTR1 깎(spread out,lower)+ 는/ENTR1	까게, 깎게, 까고, 깎고 까기만, 깎기만 까야, 깎아야...
쓰는, 쓰니...	쓰, 쓸	쓰(use,write,employ,hire,spend,speak,wear,bitter)/VB+는/ENTR1 쓸(sweep)/VB+는/ENTR1	쓰고, 쓸고 쓰게, 쓸게 쓰도록, 쓸도록 쓰기만, 쓸기만...
파는, 판다고, 판다는, 파는데, 파지...	파, 팔	파(dig,carve,remove,investigate)/VB+ 는 팔(sell,turn away,trade on)/VB+ 는	파고, 팔고 파게, 팔게, 파면서, 팔면서 파려고, 팔려고...
나는, 나니, 날니다, 난다는 ...	나, 날	나(grow,born,sprout,occur,produce,come out,smell)/VB+는 날(fly,soar)/VB+ 는	나고, 날고 나게, 날게 나도록, 날도록 나려고, 날려고...

활용 시 형태적 중의성을 보이는 동사 원형의 개수 자체는 많지 않지만 (12개 정도) 그 활용형들은 상당한 부분을 차지하며 해당 동사들은 다의적인 여러 의미로 쓰여 한국어 텍스트에 자주 등장하는 중요한 단어들이 많다.

VVMA는 형태적 중의성이란 할지라도 동사의 의미 중의성 해결과 관련이 깊다. 일반적으로 주어진 문맥이나 담화(discourse) 하에서는 단어들은 단 하나의 의미만을 표현하고 있기 때문에 (Yarowsky 1995) 주변 문맥의 단어들은 해당 단어의 의미 중의성 해결에 단서를 제공해 줄 수 있다.

또한, 모든 동사 활용형이 다 형태적 중의성을 내포하고 있지는 않다. 예를 들어 "파는"의 경우는 중의성이 존

재하지만 "팔고" 나 "파고" 활용의 경우 그 원형이 "팔다"와 "파다"로 중의성이 존재하지 않는다. 따라서, 본 발명의 일 실시 예는 중의성 없이 나타난 동사 활용형으로부터 원형 도출이 가능한 예제들의 주변 문맥을 추출하여 중의성이 나타난 활용형 문제 해결에 이용한다.

[0023] 본 발명의 일 실시 예에서는 학습 데이터로서, 레이블이 되어 있지않은 말뭉치로부터 중의성 없는 활용형들의 주위 문맥과 동사 원형을 학습시킨 후, 중의성이 발생하는 활용형들의 주위 문맥을 보고 기존 학습한 문맥과 비슷한 원형 클래스를 할당한다. 즉, 중의성 없는 활용형은 기본형이 불명확한 단어들을 위한 학습 데이터로 사용될 수 있다. 이 과정에서 각각의 기본형과 함께 등장하는 문맥 정보가 수집될 수 있다.

[0024] 도 1a는 본 발명의 일 실시 예에 따른 형태적 중의성이 있는 동사 분석 장치(100)의 블록도이다.

[0025] 기본 데이터 수집부(120)는 가공되지 않은 말뭉치(110)에서 기본형을 알 수 있는 동사 활용형 (이하 '제1동사활용')과 관련된 문맥 특징을 포함하는 학습 데이터를 수집한다. 이 과정에서 기본적인 학습 데이터(seed examples)가 생성된다. 고려되는 문맥 특징은 중의성이 발생하는 어절 앞에 나타나는 두 어절과 두 어절 각각의 내용어 (content words)와 조사정보가 문맥 특징 벡터로 사용되어 학습 데이터(130)를 구성한다.

[0026] 훈련 수행부(140)는 학습 데이터(130)에 기반한 기계 학습 방법을 이용하여 각 기본형을 판단하기 위하여 문맥 정보와 동사 원형 간의 조건부 문맥 확률을 구한다. 훈련 수행부(140)는 기본형과 함께 사용될 확률이 가장 높은 문맥 정보를 기준으로 기본형을 판단하기 위한 분류자를 학습한다. 본 발명의 일 실시 예에서는 재학습이 편리한 Naïve Bayesian 학습 방법을 이용한다.

[0027] 표 3은 기본형 "듣"과 "들"을 구분하기 위해 사용되는 학습 데이터에서 문맥 특징 자질(feature)의 예를 보여준다.

표 3

[0028]

내용어 (word ₁)	내용어 (word ₂)	Word ₁	Word ₂	격조자 (word ₁)	격조사 (word ₂)	Class (output)
제작비	많이	제작비가	많이	가(주격조사)		들
나	이야기	내	이야기를		를	듣
같	이야기	같은	이야기를		를	들
점	이유	점을	이유로	을	로	들
급격히	줄	급격히	줄어			들

[0029] "Class"는 이러한 표 3의 문맥 특징 자질을 사용하는 경우에 출력되는 기본형을 의미한다.

[0030] 동사 분석부(150)는 말뭉치(110)에 포함된 기본형이 중의적인 동사 활용형(이하 '제2동사활용')의 기본형을 판단을 위한 부분으로 훈련 수행부(140)에서 학습된 분류자를 이용하여 새로운 예제에 대한 기본형을 알아낸다. 임의의 텍스트가 입력되면, 동사 분석부(150)는 입력된 텍스트에서 각각의 동사의 앞에 존재하는 주변 단어들로부터 조사, 단어, 내용어 등 학습에 필요한 벡터를 추출한다.

[0031] 중의적 형태를 가진 동사 활용의 기본형을 판단하는 것은 텍스트 의미 판독에서 중요한 역할을 한다. 따라서 도 1a의 동사형 분석장치(100)는 검색 엔진, 번역기 등에 응용될 수 있다.

[0032] 도 1b는 본 발명의 다른 실시 예에 따른 형태적 중의성이 있는 동사 분석 장치의 블록도이다.

[0033] 기본 데이터 수집부(120)는 가공되지 않은 말뭉치(110)에서 제1동사활용과 관련된 문맥 특징을 포함하는 학습 데이터(130)를 수집하고 훈련 수행부(140)는 학습 데이터(130)에 기반하여 문맥에 따른 기본형 분류자를 학습하며 동사 분석부(150)에서는 말뭉치(110)에 포함된 제2동사활용의 기본형을 판단한다.

[0034] 일반적으로, 학습 데이터의 확장으로 보다 많은 문맥 정보를 수집할수록 동사 분석의 정확도가 향상될 수 있다. 또한 학습 도메인이 바뀌거나 새로운 단어들의 등장으로 새로운 학습 데이터 확보가 필요한 경우도 많다. 이때, 문맥 정보를 추출하기 위한 한 방법으로 웹 문서를 이용하는 방법을 생각해 볼 수 있다. 하지만 웹 페이지들은 다운로드 하기에는 지나치게 크다. 표 4는 중의성을 내포하고 있는 단어 "까는"과 이를 해결하기 위해 사용될 수 있는 중의성이 없는 활용형의 한 가지인 "까고", "팔고"의 웹 카운트의 예를 보여준다. "Daterange"는 특정 기간에 공개된 문헌들로 한정하여 결과를 출력하는 것으로 표 4는 한 달 기간 내의 문서로 검색의 범위를 한정

한 결과이다.

표 4

Word	Total count	Daterange counts
까고	492,000	70,500
깔고	1,200,000	122,000
까는	1,020,000	156,000
사고	37,000,000	246,000
살고	9,200,000	247,000
사는	21,500,000	248,000

따라서, 어려운 사례 판단부(160)는 선택적 샘플링을 통해 모든 사례에 대해 학습 데이터를 확장하지 않고, 현 학습된 분류자가 분류하기 힘든 사례에 대해서만 선택적으로 학습 데이터를 확장한다. 이에 따라 본 발명의 다른 실시 예는 반복적인 데이터 추가를 피하고 불필요하게 데이터의 크기가 커지는 것을 방지하면서도 기본형 예측의 정확도는 높일 수 있다.

이하에서는 제2동사활용에 대해 주변 문맥 추출하여 기존 학습된 분류자를 이용하여 첫 번째로 높은 확률값으로 예측되는 제1기본형과 두 번째로 높은 확률값으로 예측되는 제2기본형을 파악한다. 어려운 사례 판단부(160)는 첫 번째로 높은 확률값과 두 번째로 높은 확률값의 차를 계산한다. 어려운 사례 판단부(160)는 이 확신도(Confidence) 확률값에 기반하여 해당 제2동사활용에 대한 새로운 학습 데이터를 추가로 수집할 지 여부를 결정하도록 구성될 수 있다. 이러한 확률값은 수학적 식 1과 같은 형태로 계산될 수 있다.

수학적 식 1

$$p(c | \mathbf{x}) = \frac{p(c)p(x_1, x_2 \cdots x_n | c_j)}{p(\mathbf{x})}$$

$$= \frac{p(c)}{p(\mathbf{x})} \prod_{i=1}^n p(x_i | c) \propto p(c) \prod_{i=1}^n p(x_i | c)$$

$$Confidence(\mathbf{x}) = \frac{p(c_i | \mathbf{x}) - p(c_j | \mathbf{x})}{p(c_i | \mathbf{x})}$$

여기서, $p(c|\mathbf{x})$ 는 사례 $\mathbf{x}$ 가 카테고리(기본형) c 로 예측될 확률이고, $\mathbf{x}$ 는 앞서 설명한 문맥 자질 집합(context feature set)이다. 해당 p 값을 예측하기 위해서는 Naïve Bayesian 분류자를 사용한 예이다.

수학적 식 1에서 c_i 는 분석되지 않은 단어의 문맥 특징 벡터에 해당하는 $\mathbf{x}$ 와 가장 근접한 카테고리(기본형)이고, c_j 는 $\mathbf{x}$ 와 두 번째로 근접한 카테고리(기본형)로 가정하면 학습된 분류자가 해당 문맥을 가지고 분류할 때의 확신도는 수학적 식 1의 Confidence 값에 해당된다.

수학적 식 1에서 보듯이, 이들 두 확률의 차이가 클수록 해당 데이터에 대한 분류자의 확신도는 커진다. 본 발명의 일 실시 예에서는 분류의 신뢰가 낮은 사례(동사)를 어려운 사례(hard examples)로 정의하여 이를 중심으로 학습 데이터를 확장한다.

또한, 어려운 사례 판단부(160)는 추가될 학습 데이터와 현재까지 수집된 학습 데이터 사이의 유사도에 기반하여 학습 데이터의 추가 여부를 결정할 수도 있다. 이를 다양성(diversity) 값으로 정의하여 수학적 식 2와 같이 계산할 수 있다. 이 다양성 값은 유사도에 반비례하는 값으로서, 이 값을 반영하는 경우, 지금까지 학습된 기존 데이터와 다른 형태의 예제에 대해서는 높은 값을 부여하여 어려운 사례로 판단할 수 있다.

수학식 2

$$diversity(t_i) = 1 - \max_{s_j \in S} \frac{K(t_i, s_j)}{\sqrt{K(t_i, t_i)K(s_j, s_j)}}$$

$$K(t_i, s_j) = t_i \cdot s_j, \cos \theta = \frac{t_i \cdot s_j}{\|t_i\| \cdot \|s_j\|}$$

[0043]

[0044]

여기서, t_i 는 학습 데이터에 추가될 사례를 의미한다. K 는 입력된 두 사례 사이의 유사도를 계산하는 함수이다. S 가 기존 학습에 사용된 사례들의 집합이라면, s_j 는 기존에 학습 데이터에 존재하는 사례들 중 t_i 와 가장 유사한 사례를 의미하는데, 다양성 값은 t_i 와 s_j 사이의 유사도를 반영한 값으로 기존 학습 데이터와 현재 추가될 데이터가 얼마나 비슷한지를 파악한다.

[0045]

본 발명의 다른 실시 예에서는 다양성 값이 가장 큰 사례 즉 기존 학습 데이터와 상이한(유사도가 가장 작은) 예제를 어려운 사례로 정의하여 학습 데이터를 확장하도록 하였다. 실험 데이터의 약 3%에 대해 확신도와 다양성 값을 조합하여 어려운 사례로 추출한다.

[0046]

표 5는 "까"와 "깔"의 분석에서 추출된 어려운 사례들의 예와 이들의 관련 문맥 특징의 예를 보여준다.

표 5

[0047]

hard examples	Context feature	Correct Category
보안 프로그램을 까는 그녀를 면전에서 까고 팬티엄 4에서	보안 프로그램 보안 프로그램을 josal=을 그녀 면전 그녀를 면전에서 josal=를 josal=에서 팬티엄 4 팬티엄 4에서 josal=에서	깔(install) 까(slate) 깔(install)

[0048]

통신부(170)는 어려운 사례 판단부 (160)을 통해 어려운 사례로 판단되어 제2동사활용을 위한 학습 데이터를 추가하는 경우, 제2동사활용 주위의 문맥 특징과 가능한 기본형을 결합한 구를 질의어로 네트워크 또는 인터넷으로 연결된 검색 엔진에 입력한다. 통신부(170)가 상기 검색 엔진으로부터 웹 카운트를 반환받으면, 추가 훈련 수행부(180)에 전달하여 현재 사례(어려운 사례)에 대한 기본형을 찾게 한다.

[0049]

표 6은 어려운 사례들의 문맥 특징을 이용한 질의와 이러한 질의에 따라 검색엔진 구글(Google)이 반환한 웹 카운트를 보여준다.

표 6

Query (Feature + Decision)	Google counts
“보안* 깔고”	7
“보안* 까고”	0
“보안* 깔게”	0
“보안* 까게”	0
“프로그램* 깔고”	1,510
“프로그램* 까고”	2
“프로그램* 깔게”	52
“프로그램* 까게”	0
“프로그램을 깔고”	1,140
“프로그램을 까고”	1
“프로그램을 깔게”	28
“프로그램을 까게”	0
“을 깔고”	29,500
“을 까고”	3,820
“을 깔게”	436
“을 까게”	126

[0050]

[0051]

표 6의 웹 카운트는 이전 3 개월간에 공개된 문서에 한정하여 계산된 값이다. 웹 카운트 추출을 위해 사용된 질의(Query)는 어려운 사례의 문맥 특징(예를 들어 "프로그램")과 이와 관련하여 사용되는 중의성이 없는 동사 활용형(예를 들어 "깔고")이 결합된 구로 구성된다. 즉, 표 6의 "깔고", "까고", "까게", "깔게"와 같은 용언의 활용형들은 "까"나 "깔"을 분석하기 위한 예로 사용된다. 와일드 카드(*)가 포함된 질의어 구 "프로그램* 깔고"에는 "프로그램을 깔고", "프로그램만 깔고", "프로그램만 무조건 깔고" 등의 다양한 어구들이 상기 질의와 매칭될 수 있다. 즉, "깔", "깔게", "깔기만", "깔고" 등의 활용형을 포함한 예들은 해당 기본형의 문맥을 훈련하기 위해 사용될 수 있다.

[0052]

추가 훈련 수행부(180)는 어려운 사례에 대한 학습 데이터를 확장한다. 추가 훈련 수행부(180)는 통신부(170)를 통해 수신된 웹 카운트를 이용하여 제2동사활용을 위한 학습 데이터를 생성하고 데이터베이스(130)에 추가한다.

[0053]

한편, 도 1b의 분석 장치(100)도 도 1a에서와 마찬가지로 동사 분석부(150)를 포함할 수 있다.

[0054]

도 2는 본 발명의 일 실시 예에 따른 형태적 중의성이 존재하는 동사 분석 방법의 흐름도이다.

[0055]

먼저, 가공되지 않은 말뭉치에서 제1동사활용의 활용형과 관련된 문맥 특징인 학습 데이터를 수집한다(S210).

[0056]

다음, 학습 데이터에 기반하여 말뭉치에 포함된 제2동사활용의 기본형을 판단한다(S220). 보다 구체적으로, 이 과정(S220)은 제2동사활용과 연결되어 사용될 확률이 가장 높은 문맥 정보를 기준으로 제2동사활용의 기본형을 판단하는 과정일 수 있다.

[0057]

그리고 제2동사활용의 기본형을 판단하는 과정에서 사용된 문맥 특징을 제2동사활용을 위한 학습 데이터로서 추가한다(S230).

[0058]

마지막으로, 임의의 텍스트가 입력되면, 텍스트에서 각각의 동사의 앞에 존재하는 단어들을 학습 데이터와 비교하여 텍스트에 포함된 동사의 기본형을 판단한다(S240). 이 과정(S240)에서 사용되는 단어들은 앞의 두 어절 및 조사와, 그 내용어로서, 각각의 동사와 관련된 문맥 단어들이다. 일 예로, 이 과정(S240)은 각각의 동사의 앞에 가장 근접하게 위치하는 명사 2개를 추출하여 학습 데이터와 비교함으로써 각각의 동사의 기본형을 판단하는 과정일 수 있다.

[0059]

본 발명은 소프트웨어를 통해 실행될 수 있다. 바람직하게는, 본 발명의 실시 예들에 따른 형태적 중의성이 존재하는 동사 분석 방법을 컴퓨터에서 실행시키기 위한 프로그램을 컴퓨터로 읽을 수 있는 기록매체에 기록하여 제공할 수 있다. 소프트웨어로 실행될 때, 본 발명의 구성 수단들은 필요한 작업을 실행하는 코드 세그먼트들이

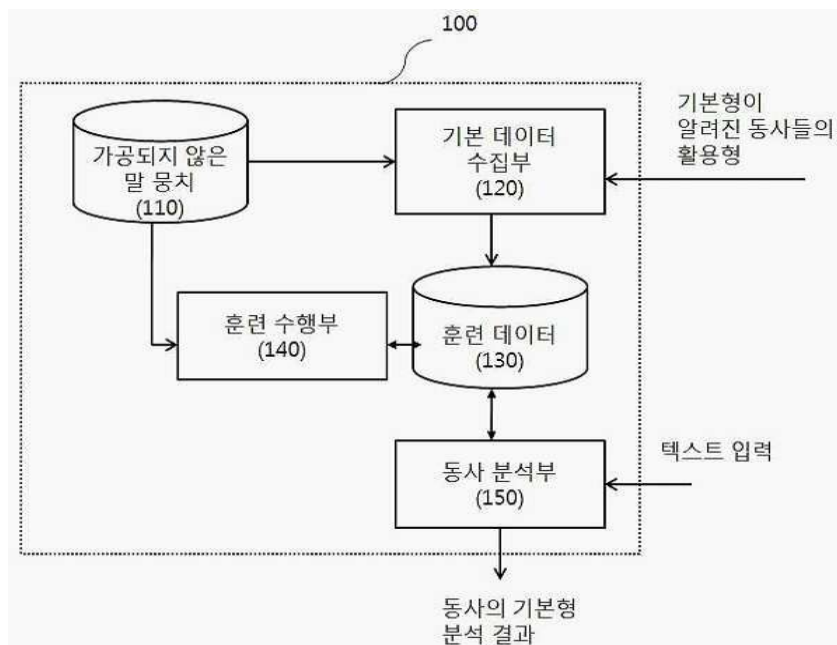
다. 프로그램 또는 코드 세그먼트들은 프로세서 판독 가능 매체에 저장되거나 전송 매체 또는 통신망에서 반송파와 결합된 컴퓨터 데이터 신호에 의하여 전송될 수 있다.

[0060] 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 장치의 예로는 ROM, RAM, CD-ROM, DVD±ROM, DVD-RAM, 자기 테이프, 플로피 디스크, 하드 디스크(hard disk), 광데이터 저장장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 장치에 분산되어 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

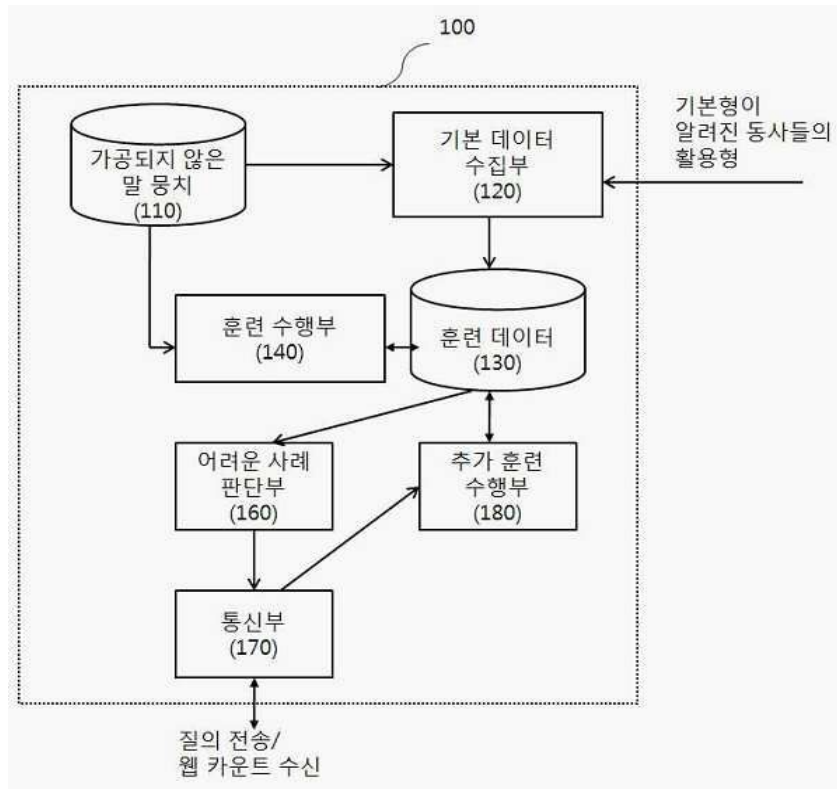
[0061] 본 발명은 도면에 도시된 실시 예들을 참고로 하여 설명하였으나 이는 예시적인 것에 불과하며 당해 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 실시 예의 변형이 가능하다는 점을 이해할 것이다. 그리고, 이와 같은 변형은 본 발명의 기술적 보호범위 내에 있다고 보아야 한다. 따라서, 본 발명의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해서 정해져야 할 것이다.

도면

도면1a



도면1b



도면2

