(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2014년01월28일
(11) 등록번호 10-1356417
(24) 등록일자 2014년01월17일

(51) 국제특허분류(Int. Cl.)
G06F 17/28 (2006.01) *G06F 17/27* (2006.01)
(21) 출원번호 10-2010-0109542
(22) 출원일자 2010년11월05일
심사청구일자 2011년05월12일
(65) 공개번호 10-2012-0048101
(43) 공개일자 2012년05월15일
(56) 선행기술조사문헌
KR1020060067073 A*
심보준 외 3인, “자연어 질의응답 시스템을 위한 is-a 관계 패턴의 구축과 활용,” 제16회 한글 및 한국어 정보처리 학술대회 발표자료집, vol. 16, no. 1, pp. 181-188, 2004.10.*
양성일 외 5인, “한영 자동 번역을 위한 동사구 번역패턴의 활용,” 한국정보과학회 가을 학술발표논문집, vol. 28, no. 2, pp. 178-180, 2001.10.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교 산학협력단
서울특별시 성북구 안암로 145 (안암동5가, 고려대학교안암캠퍼스)
에스케이플래닛 주식회사
경기도 성남시 분당구 판교로 264 (삼평동)
(72) 발명자
황영숙
서울특별시 성북구 북악산로 913, 풍림아파트 105동 502호 (돈암동)
김상범
서울특별시 노원구 동일로231길 86, 202동 1207호 (상계동, 현대2차아파트)
(뒷면에 계속)
(74) 대리인
전철용, 박종한, 김기효

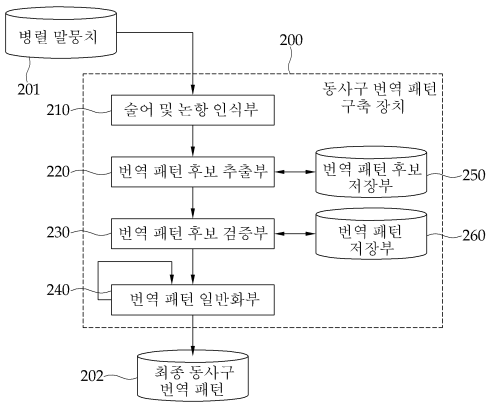
전체 청구항 수 : 총 22 항 심사관 : 임지환

(54) 발명의 명칭 **병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치 및 그 방법**

(57) 요 약

본 발명은 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치 및 그 방법에 관한 것으로서, 상세하게는 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 단어 정렬 결과를 이용하여 술어 및 논항을 인식하고 그 인식된 술어 및 논항을 이용하여 번역 패턴 후보 및 출현 빈도수를 추출한 후, 번역 패턴 후보에 대한 검증을 통해 기본 동사구 번역 패턴을 생성하고 이를 일반화시켜 일반 동사구 번역 패턴을 생성함으로써, 다양한 언어쌍에 적용 가능하고 동사구 번역 패턴 오류를 최소화할 수 있으며, 동사구 번역 패턴의 술어 및 논항의 공기 빈도와 술어의 번역 확률을 이용해 적절한 일반화 수준을 결정할 수 있다.

대 표 도 - 도2



(72) 발명자

윤창호

서울특별시 성북구 서경로 31, 푸른동아 아파트 107동 1303호 (정릉동)

이주영

서울특별시 중랑구 신내로 127, 911동 210호 (신내동, 신내아파트)

임해창

서울특별시 서초구 방배로 21 (방배동, 경남아파트) 3동 401호

특허청구의 범위

청구항 1

복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식하는 술어 및 논항 인식부;

상기 인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출하고 해당 출현 빈도수를 확인하는 번역 패턴 후보 추출부;

상기 추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 상기 추출된 번역 패턴 후보를 검증하고, 상기 검증 결과에 따라 제1 동사구 번역 패턴을 생성하는 번역 패턴 후보 검증부; 및

상기 생성된 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시키고, 상기 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들을 추출하고 상기 추출된 튜플들의 빈도수에 따라 제2 동사구 번역 패턴을 생성하여 일반화시키는 번역 패턴 일반화부

를 포함하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 2

제 1 항에 있어서,

상기 번역 패턴 후보 추출부에서 추출된 번역 패턴 후보를 저장하는 번역 패턴 후보 저장부; 및

상기 번역 패턴 후보 검증부에서 생성된 제1 동사구 번역 패턴을 저장하는 번역 패턴을 저장하는 번역 패턴 저장부

를 더 포함하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 3

제 1 항에 있어서,

상기 술어 및 논항 인식부는,

복수의 병렬 말뭉치를 입력받아 소스 문장의 구문 분석 결과를 이용하여 소스 문장의 술어 및 논항을 인식하고, 단어 정렬 결과를 이용하여 상기 인식된 소스 문장의 술어 및 논항에 대응하는 타겟 문장의 술어 및 논항을 인식하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 4

제 1 항에 있어서,

상기 번역 패턴 후보 추출부는,

서로 다른 소스 문장 및 타겟 문장에서 동일한 번역 패턴 후보를 추출하고 상기 추출된 동일한 번역 패턴 후보의 출현 빈도수를 구하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 5

제 1 항에 있어서,

상기 번역 패턴 후보 검증부는,

상기 추출된 번역 패턴 후보에 대한 원문부 패턴의 출현 빈도수를 미리 설정된 임계치와 비교하여 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 6

제 1 항에 있어서,

상기 번역 패턴 후보 검증부는,

원문부 패턴이 동일한 복수의 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 7

제 6 항에 있어서,

상기 번역 패턴 후보 검증부는,

상기 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에, 상기 대역부 패턴의 논항이 복수 개이고 상기 대역부 패턴을 포함하는 다른 대역부 패턴이 존재하는지 여부에 따라 상기 추출된 동일한 번역 패턴 후보들을 검증하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 8

제 1 항에 있어서,

상기 번역 패턴 일반화부는,

상기 생성된 제1 동사구 번역 패턴들의 논항을 구문 표지로 일반화하는 경우에 유효한 동사구 번역 패턴인지를 확인하여 일반화시켜 제2 동사구 번역 패턴을 생성하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 9

삭제

청구항 10

제 1 항에 있어서,

상기 번역 패턴 일반화부는,

상기 추출된 튜플들의 빈도수 중에서 가장 빈도수가 큰 튜플에 해당하는 논항을 제외한 다른 논항들을 일반화시켜 제2 동사구 번역 패턴을 생성하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 11

제 1 항에 있어서,

상기 번역 패턴 일반화부는,

상기 생성된 제2 동사구 번역 패턴에 대해서 번역 모호성을 분석하여 제2 동사구 번역 패턴을 결정하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 12

제 1 항에 있어서,

상기 번역 패턴 일반화부는,

상기 생성된 제2 동사구 번역 패턴에 대해서 번역 패턴 일반화를 반복적으로 수행하는 포함하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치.

청구항 13

복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식하는 술어 및 논항 인식 단계;

상기 인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출하고 해당 출현 빈도수를 확인하는 번역 패턴 후보 추출 단계;

상기 추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 상기 추출된 번역

패턴 후보를 검증하고, 상기 검증 결과에 따라 제1 동사구 번역 패턴을 생성하는 번역 패턴 후보 검증 단계; 및

상기 생성된 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시키고, 상기 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들을 추출하고 상기 추출된 튜플들의 빈도수에 따라 제2 동사구 번역 패턴을 생성하여 일반화시키는 번역 패턴 일반화 단계

를 포함하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 14

제 13 항에 있어서,

상기 술어 및 논항 인식 단계는,

복수의 병렬 말뭉치를 입력받아 소스 문장의 구문 분석 결과를 이용하여 소스 문장의 술어 및 논항을 인식하는 단계; 및

단어 정렬 결과를 이용하여 상기 인식된 소스 문장의 술어 및 논항에 대응하는 타겟 문장의 술어 및 논항을 인식하는 단계

를 포함하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 15

제 13 항에 있어서,

상기 번역 패턴 후보 추출 단계는,

서로 다른 소스 문장 및 타겟 문장에서 동일한 번역 패턴 후보를 추출하고 상기 추출된 동일한 번역 패턴 후보의 출현 빈도수를 구하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 16

제 13 항에 있어서,

상기 번역 패턴 후보 검증 단계는,

상기 추출된 번역 패턴 후보에 대한 원문부 패턴의 출현 빈도수를 미리 설정된 임계치와 비교하여 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 17

제 13 항에 있어서,

상기 번역 패턴 후보 검증 단계는,

상기 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 18

제 13 항에 있어서,

상기 번역 패턴 후보 검증 단계는,

상기 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에, 상기 대역부 패턴의 논항이 복수 개이고 상기 대역부 패턴을 포함하는 다른 대역부 패턴이 존재하는지 여부에 따라 상기 추출된 동일한 번역 패턴 후보들을 검증하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 19

제 13 항에 있어서,

상기 번역 패턴 일반화 단계는,

상기 생성된 제1 동사구 번역 패턴들의 논항을 구문 표지로 일반화하는 경우에 유효한 동사구 번역 패턴인지를 확인하여 일반화시켜 제2 동사구 번역 패턴을 생성하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 20

삭제

청구항 21

제 13 항에 있어서,

상기 번역 패턴 일반화 단계는,

상기 추출된 튜플들의 빈도수 중에서 가장 빈도수가 큰 튜플에 해당하는 논항을 제외한 다른 논항들을 일반화시켜 제2 동사구 번역 패턴을 생성하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 22

제 13 항에 있어서,

상기 번역 패턴 일반화 단계는,

상기 생성된 제2 동사구 번역 패턴에 대해서 번역 모호성을 분석하여 제2 동사구 번역 패턴을 결정하는 번역 모호성 분석 단계

를 더 포함하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 23

제 13 항에 있어서,

상기 번역 패턴 일반화 단계는,

상기 생성된 제2 동사구 번역 패턴에 대해서 상기 번역 패턴 일반화 단계를 반복적으로 수행하는 반복 수행 단계

를 더 포함하는 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법.

청구항 24

제13항 내지 제19항, 제21항 내지 제23항 중 어느 한 항에 의한 과정을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체.

명 세 서

기 술 분 야

[0001] 본 발명은 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치 및 그 방법에 관한 것으로서, 상세하게는 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 단어 정렬 결과를 이용하여 술어 및 논항을 인식하고 그 인식된 술어 및 논항을 이용하여 번역 패턴 후보 및 출현 빈도수를 추출한 후, 번역 패턴 후보에 대한 검증을 통해 기본 동사구 번역 패턴을 생성하고 이를 일반화시켜 일반 동사구 번역 패턴을 생성함으로써, 다양한 언어쌍에 적용 가능하고 동사구 번역 패턴 오류를 최소화할 수 있으며, 동사구 번역 패턴의 술어 및 논항의 공기 빈도와 술어의 번역 확률을 이용해 적절한 일반화 수준을 결정할 수 있는, 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0002] 기계 번역이란 번역 지식 및 규칙 등의 번역 자원과 알고리즘을 이용해 주어진 번역할 문장 즉, 소스 문장(Source Sentence)을 번역된 문장 즉, 타겟 문장(Target Sentence)으로 변환하는 작업을 말한다. 이러한 기계 번역 작업을 수행하는 시스템을 기계 번역 시스템이라고 한다. 번역에 있어 문장의 서술어에 해당하는 용언은 주변 문맥에 따라 그 의미를 달리 해석해야 하는 경우가 많다. 예를 들어, 영어 단어 "prevent"는 다음과 같이

목적어가 무엇인지에 따라 한국어 대역어가 달라진다. 여기서, 용언에는 영어의 경우에 동사가 포함되며, 한국어의 경우에 동사 및 형용사가 포함된다.

[0003] 예를 들면, "You cannot prevent their getting married."라는 영어 문장은 "당신은 그들이 결혼하는 것을 막을 수 없다."라는 한국어 문장으로 번역된다. 반면, 다른 예를 살펴보면, "prevent flu from spreading"라는 영어 문장은 "유행성 감기의 만연을 예방하다"라는 한국어 문장으로 번역된다.

[0004] 즉, "prevent"는 보통 "막다" 또는 "방해하다"라는 뜻으로 사용되지만, 질병을 목적어로 취하는 경우에는 "예방하다"로 번역하는 것이 더 자연스럽다. 이와 같이, 용언은 주변 문맥에 따라 번역이 달라지는 경우가 많기 때문에 개개의 용언을 번역하는 것보다 용언과 주변 단어들을 하나의 패턴으로 묶어서 번역하는 것이 용이하다. 또한, 영어-한국어와 같이 어족이 서로 다른 언어의 경우에는 주어, 서술어 및 목적어와 같은 문장 성분의 위치가 다르기 때문에 이에 대한 정보도 번역에서는 매우 중요하다.

[0005] 동사구 번역 패턴이란, "X prevent Y, X(가) Y(를) 막다", "X prevent desease, X(가) 질병(을) 예방하다"와 같이 문장의 술어와 주요 논항들에 대한 번역 정보를 패턴화한 것으로 기계 번역에서 중요한 정보로 이용된다.

[0006] 종래의 동사구 패턴을 구축하는 방법은 특정 언어 쌍을 대상으로 하고 있으며, 동사구 패턴을 구축하는 방법에 있어서도 원시 언어와 타겟 언어에 대한 동사구 패턴을 먼저 구축한 후 대역어를 부착하는 방법을 사용한다. 이에 따라, 언어 쌍에 제한이 되며 동사구 번역 패턴을 자동으로 구축할 수 있다.

[0007] 좀 더 구체적으로, 종래의 동사구 패턴 구축 방법은 영-한 번역을 위한 동사구 패턴 데이터베이스를 구축하기 위해 영-한 병렬 말뭉치뿐만 아니라 영어 말뭉치, 영어 워드넷, 그리고 대역어 정렬을 위한 사전(dictionary) 등을 이용한다. 즉, 영어 말뭉치를 이용하여 영어 동사구 패턴 데이터베이스를 사전에 구축하는데, 동사구 패턴에서 동사구의 각 논항에 대해서는 워드넷을 사용하여 논항의 의미정보를 부착하고, 병렬 말뭉치와 전자사전을 이용하여 문장을 구성하는 단어들의 대역어들을 정렬한 후, 의미벡터와 한국어 국소문맥을 이용하여 동사구 패턴을 구성하는 요소들의 대역어 중의성을 해소하면서 동사구 대역 패턴을 구축하고자 하였다.

[0008] 그러나 이러한 종래의 기술은 영어 동사구에 대한 한국어 동사구 대역 패턴만을 구축 대상으로 한다는 제한 사항이 있다. 또한, 영어 동사구 패턴의 각 구성요소에 대한 대역 정보를 부착하기 위해 영-한 대역 사전을 필요로 한다. 또한, 대역어 중의성 해소를 위해 병렬 말뭉치와 전자사전을 이용하여 휴리스틱하게 단어 정렬을 수행하고 대응되는 한국어 국소 문맥을 이용하여 의미벡터를 구성하고, 대역어 중의성 해소를 시도하는데 이로써 구축되는 한국어 대역 패턴이 불완전하다는 점등의 문제점이 있다.

발명의 내용

해결하려는 과제

[0009] 본 발명은 상기의 문제점을 해결하기 위해 창안된 것으로서, 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 단어 정렬 결과를 이용하여 술어 및 논항을 인식하고 그 인식된 술어 및 논항을 이용하여 번역 패턴 후보 및 출현 빈도수를 추출한 후, 번역 패턴 후보에 대한 검증을 통해 기본 동사구 번역 패턴을 생성하고 이를 일반화시켜 일반 동사구 번역 패턴을 생성함으로써, 다양한 언어쌍에 적용 가능하고 동사구 번역 패턴 오류를 최소화할 수 있으며, 동사구 번역 패턴의 술어 및 논항의 공기 빈도와 술어의 번역 확률을 이용해 적절한 일반화 수준을 결정할 수 있는, 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치 및 그 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0010] 이를 위하여, 본 발명의 제1 측면에 따른 장치는, 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식하는 술어 및 논항 인식부; 상기 인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출하고 해당 출현 빈도수를 확인하는 번역 패턴 후보 추출부; 상기 추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 상기 추출된 번역 패턴 후보를 검증하고, 상기 검증 결과에 따라 제1 동사구 번역 패턴을 생성하는 번역 패턴 후보 검증부; 및 상기 생성된 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시키고, 상기 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들을 추출하고 상기 추출된 튜플들의 빈도수에 따라 제2 동사구 번역 패턴을 생성하여 일반화시키는 번역 패턴 일반화부를 포함하는 것을 특징으로 한다.

[0011] 바람직하게는, 상기 번역 패턴 후보 추출부에서 추출된 번역 패턴 후보를 저장하는 번역 패턴 후보 저장부; 및 상기 번역 패턴 후보 검증부에서 생성된 제1 동사구 번역 패턴을 저장하는 번역 패턴을 저장하는 번역 패턴 저장부를 더 포함하는 것을 특징으로 한다.

[0012] 바람직하게는, 상기 술어 및 논항 인식부는, 복수의 병렬 말뭉치를 입력받아 소스 문장의 구문 분석 결과를 이용하여 소스 문장의 술어 및 논항을 인식하고, 단어 정렬 결과를 이용하여 상기 인식된 소스 문장의 술어 및 논항에 대응하는 타겟 문장의 술어 및 논항을 인식하는 것을 특징으로 한다.

[0013] 바람직하게는, 상기 번역 패턴 후보 추출부는, 서로 다른 소스 문장 및 타겟 문장에서 동일한 번역 패턴 후보를 추출하고 상기 추출된 동일한 번역 패턴 후보의 출현 빈도수를 구하는 것을 특징으로 한다.

[0014] 바람직하게는, 상기 번역 패턴 후보 검증부는, 상기 추출된 번역 패턴 후보에 대한 원문부 패턴의 출현 빈도수를 미리 설정된 임계치와 비교하여 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 것을 특징으로 한다.

[0015] 바람직하게는, 상기 번역 패턴 후보 검증부는, 원문부 패턴이 동일한 복수의 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 것을 특징으로 한다.

[0016] 바람직하게는, 상기 번역 패턴 후보 검증부는, 상기 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에, 상기 대역부 패턴의 논항이 복수 개이고 상기 대역부 패턴을 포함하는 다른 대역부 패턴이 존재하는지 여부에 따라 상기 추출된 동일한 번역 패턴 후보들을 검증하는 것을 특징으로 한다.

[0017] 바람직하게는, 상기 번역 패턴 일반화부는, 상기 생성된 제1 동사구 번역 패턴들의 논항을 구문 표지로 일반화하는 경우에 유효한 동사구 번역 패턴인지를 확인하여 일반화시켜 제2 동사구 번역 패턴을 생성하는 것을 특징으로 한다.

[0018] 바람직하게는, 상기 번역 패턴 일반화부는, 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시키고, 상기 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들과 상기 추출된 튜플들의 빈도수를 이용하여 제2 동사구 번역 패턴을 생성하는 것을 특징으로 한다.

[0019] 바람직하게는, 상기 번역 패턴 일반화부는, 상기 추출된 튜플들의 빈도수 중에서 가장 빈도수가 큰 튜플에 해당하는 논항을 제외한 다른 논항들을 일반화시켜 제2 동사구 번역 패턴을 생성하는 것을 특징으로 한다.

[0020] 바람직하게는, 상기 번역 패턴 일반화부는, 상기 생성된 제2 동사구 번역 패턴에 대해서 번역 모호성을 분석하여 제2 동사구 번역 패턴을 결정하는 것을 특징으로 한다.

[0021] 바람직하게는, 상기 번역 패턴 일반화부는, 상기 결정된 제2 동사구 번역 패턴에 대해서 상기 번역 패턴 일반화부에서의 번역 패턴 일반화를 반복적으로 수행하는 포함하는 것을 특징으로 한다.

[0022] 한편, 본 발명의 제2 측면에 따른 방법은, 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식하는 술어 및 논항 인식 단계; 상기 인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출하고 해당 출현 빈도수를 확인하는 번역 패턴 후보 추출 단계; 상기 추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 상기 추출된 번역 패턴 후보를 검증하고, 상기 검증 결과에 따라 제1 동사구 번역 패턴을 생성하는 번역 패턴 후보 검증 단계; 및 상기 생성된 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시키고, 상기 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들을 추출하고 상기 추출된 튜플들의 빈도수에 따라 제2 동사구 번역 패턴을 생성하여 일반화시키는 번역 패턴 일반화 단계를 포함하는 것을 특징으로 한다.

[0023] 바람직하게는, 상기 술어 및 논항 인식 단계는, 복수의 병렬 말뭉치를 입력받아 소스 문장의 구문 분석 결과를 이용하여 소스 문장의 술어 및 논항을 인식하는 단계; 및 단어 정렬 결과를 이용하여 상기 인식된 소스 문장의 술어 및 논항에 대응하는 타겟 문장의 술어 및 논항을 인식하는 단계를 포함하는 것을 특징으로 한다.

[0024] 바람직하게는, 상기 번역 패턴 후보 추출 단계는, 서로 다른 소스 문장 및 타겟 문장에서 동일한 번역 패턴 후보를 추출하고 상기 추출된 동일한 번역 패턴 후보의 출현 빈도수를 구하는 것을 특징으로 한다.

[0025] 바람직하게는, 상기 번역 패턴 후보 검증 단계는, 상기 추출된 번역 패턴 후보에 대한 원문부 패턴의 출현 빈도수를 미리 설정된 임계치와 비교하여 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 것을 특징으로 한

다.

[0026] 바람직하게는, 상기 번역 패턴 후보 검증 단계는, 상기 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에 해당 번역 패턴 후보를 제거하는 것을 특징으로 한다.

[0027] 바람직하게는, 상기 번역 패턴 후보 검증 단계는, 상기 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에, 상기 대역부 패턴의 논항이 복수 개이고 상기 대역부 패턴을 포함하는 다른 대역부 패턴이 존재하는지 여부에 따라 상기 추출된 동일한 번역 패턴 후보들을 검증하는 것을 특징으로 한다.

[0028] 바람직하게는, 상기 번역 패턴 일반화 단계는, 상기 생성된 제1 동사구 번역 패턴들의 논항을 구문 표지로 일반화하는 경우에 유효한 동사구 번역 패턴인지를 확인하여 일반화시켜 제2 동사구 번역 패턴을 생성하는 것을 특징으로 한다.

[0029] 바람직하게는, 상기 번역 패턴 일반화 단계는, 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시키는 패턴 그룹화 단계; 상기 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들을 추출하는 튜플 추출 단계; 및 상기 추출된 튜플들의 빈도수를 이용하여 제2 동사구 번역 패턴을 생성하는 패턴 생성 단계를 포함하는 것을 특징으로 한다.

[0030] 바람직하게는, 상기 패턴 생성 단계는, 상기 추출된 튜플들의 빈도수 중에서 가장 빈도수가 큰 튜플에 해당하는 논항을 제외한 다른 논항들을 일반화시켜 제2 동사구 번역 패턴을 생성하는 것을 특징으로 한다.

[0031] 바람직하게는, 상기 번역 패턴 일반화 단계는, 상기 생성된 제2 동사구 번역 패턴에 대해서 번역 모호성을 분석하여 제2 동사구 번역 패턴을 결정하는 번역 모호성 분석 단계를 더 포함하는 것을 특징으로 한다.

[0032] 바람직하게는, 상기 번역 패턴 일반화 단계는, 상기 결정된 제2 동사구 번역 패턴에 대해서 상기 번역 패턴 일반화 단계를 반복적으로 수행하는 반복 수행 단계를 더 포함하는 것을 특징으로 한다.

발명의 효과

[0033] 본 발명은 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 단어 정렬 결과를 이용하여 술어 및 논항을 인식하고 그 인식된 술어 및 논항을 이용하여 번역 패턴 후보 및 출현 빈도수를 추출한 후, 번역 패턴 후보에 대한 검증을 통해 기본 동사구 번역 패턴을 생성하고 이를 일반화시켜 일반 동사구 번역 패턴을 생성함으로써, 다양한 언어쌍에 적용 가능하고 동사구 번역 패턴 오류를 최소화할 수 있으며, 동사구 번역 패턴의 술어 및 논항의 공기 빈도와 술어의 번역 확률을 이용해 적절한 일반화 수준을 결정할 수 있는 효과가 있다.

[0034] 즉, 본 발명은 병렬 말뭉치를 이용한 단어, 트리 정렬 결과 및 소스 혹은 타겟 언어, 한쪽 언어에 대한 파서만을 이용하기 때문에 특정 언어 쌍에 제한되지 않고 동사구 번역 패턴을 자동으로 구축할 수 있는 효과가 있다.

[0035] 본 발명은 대량의 병렬 말뭉치에서 다수의 동사구 번역 패턴을 자동으로 생성함으로써 기계번역의 성능을 향상시킬 수 있으며, 동사구 번역 패턴을 구축하기 위한 비용 및 시간을 절감할 수 있는 효과가 있다.

[0036] 더 나아가, 본 발명은 사람의 개입을 최소화할 수 있기 때문에 언어 전문가 없이도 양질의 번역 지식을 구축할 수 있는 효과가 있다.

도면의 간단한 설명

[0037] 도 1 은 본 발명에 적용되는 동사구 번역 패턴에 대한 일실시예 예시도,

도 2 는 본 발명에 따른 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치의 일실시예 구성도,

도 3 은 본 발명에 따른 도 2의 술어 및 논항 인식부에서의 술어 및 논항 인식에 대한 일실시예 예시도,

도 4 는 본 발명에 따른 도 2의 번역 패턴 일반화부에서의 번역 패턴 일반화에 대한 일실시예 예시도,

도 5 는 본 발명에 따른 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법에 대한 일실시예 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0038] 이하, 첨부된 도면을 참조하여 본 발명에 따른 실시 예를 상세하게 설명한다. 본 발명의 구성 및 그에 따른 작용 효과는 이하의 상세한 설명을 통해 명확하게 이해될 것이다. 본 발명의 상세한 설명에 앞서, 동일한 구성요

소에 대해서는 다른 도면 상에 표시되더라도 가능한 동일한 부호로 표시하며, 공지된 구성에 대해서는 본 발명의 요지를 흐릴 수 있다고 판단되는 경우 구체적인 설명은 생략하기로 함에 유의한다.

[0039] 도 1 은 본 발명에 적용되는 동사구 번역 패턴에 대한 일실시예 예시도이다.

[0040] 본 발명의 설명에 앞서 하나의 병렬 문장에서 추출되는 동사구 번역 패턴의 모습과 용어를 설명하기로 한다.

[0041] 도 1에 도시된 바와 같이, 병렬 문장(110)은 "Bush will have the ceremony on January 20 of next year"인 소스 문장과 "부시는 내년 1월 20일 취임식을 갖는다"인 타겟 문장으로 이루어져 있다.

[0042] 본 발명은 이러한 복수의 병렬 문장(110)으로부터 동사구 패턴을 구축할 수 있다. 본 발명에서 구축하는 동사구 패턴은 크게 기본 동사구 번역 패턴(120)과 일반 동사구 번역 패턴(130)으로 나눌 수 있다. 하나의 동사구 번역 패턴은 원문부 패턴(121)과 대역부 패턴(122)의 쌍(Pair)으로 이루어진다. 도 2에 도시된 "1:[bush] 2:[V have] 3:[ceremony] 4:[on january]"는 원문부 패턴(121), "1:[부시 case=는] 4:[1월] 3:[취임식 case=을] 2:[V 갖]"은 대역부 패턴(122)에 해당한다. 동사구 번역 패턴을 간단히 표현하기 위해 앞으로는 임의의 동사구 패턴은 "P=<LHS, RHS>"와 같이 나타내기로 한다. 이때 LHS(Left Hand Side)는 원문부 패턴(121), RHS(Right Hand Side)는 대역부 패턴(122)에 해당한다. 원문부 패턴(121), 대역부 패턴(122)을 통칭해서 동사구 패턴으로 부르기로 하며, 이것은 동사구 번역 패턴과는 구별된다.

[0043] 도 1에 도시된 일례에서는 병렬 문장(110)의 소스 문장은 영어이고, 타겟 문장은 한국어인 경우를 예로 사용하였으나, 본 발명에 따른 동사구 번역 패턴 구축 장치의 구문 분석 결과와 병렬 말뭉치가 존재하는 모든 언어 쌍에 대해 적용 가능하다.

[0044] 영어 동사구 패턴에 해당하는 원문부 패턴(LHS)(121)은 하나의 술어와 복수 개의 논항으로 이루어진다. 일례에서 원문부 패턴(121)의 술어는 "have"이며, "bush", "ceremony", "on January"는 모두 논항에 해당한다. 대역부 패턴(RHS)(122)는 원문부 술어에 대응되는 대역부 술어(예컨대, "갖")와 원문부 논항에 대응되는 대역부 논항 (예컨대, "부시", "1월", "취임식")으로 이루어진다. 각 술어와 논항은 서로 대응관계를 알 수 있도록 "숫자:"와 같은 상호참조 번호를 표시하기로 한다.

[0045] 기본 동사구 번역 패턴(120)은 번역 패턴의 모든 논항이 어휘로 표현된 패턴을 의미한다. 일반 동사구 번역 패턴(130)은 기본 동사구 번역 패턴(120)에서 일부 논항을 명사구(NP), 동사구(VP), 부사구(ADVP) 등과 같은 구문 표지(phrase label)로 표시한 것을 뜻한다. 여기서, 구문 표지에서의 "동사구"와 "동사구 번역 패턴"의 동사구는 서로 다른 개념이다. 구문 표지를 이용해 표현한 일반 동사구 번역 패턴은 기본 동사구 번역 패턴에 비해 좀 더 많은 문장을 번역 할 수 있다. 예를 들어, 그림 1의 기본 동사구 번역 패턴은 주어가 "bush"이고, 부사구가 "on January"인 문장에 대해서만 적용 가능하지만, 일반 동사구 번역 패턴은 주어가 임의의 명사구이고, 전치사 on 다음에 어떤 명사구가 오는 모든 문장에 대해 적용 가능하다.

[0046] 전술된 용어는 본 발명에서 사용하는 동사구 번역 패턴에 대한 간단한 정의이다. 하지만, 이것은 동사구 번역 패턴을 정의하는 다양한 방법 중의 하나이며, 일반적으로 동사구 패턴이라고 부를 수 있는 형태는 본 발명을 적용하여 구축할 수 있다.

[0047] 도 2 는 본 발명에 따른 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 장치의 일실시예 구성도이다.

[0048] 도 2에 도시된 바와 같이, 도 2 은 본 발명에 따른 동사구 번역 패턴 구축 장치(200)는 술어 및 논항 인식부(210), 번역 패턴 후보 추출부(220), 번역 패턴 후보 검증부(230), 번역 패턴 일반화부(240), 번역 패턴 후보 저장부(250) 및 번역 패턴 저장부(260)를 포함한다.

[0049] 본 발명에 따른 동사구 번역 패턴 구축 장치(200)는 복수의 병렬 말뭉치(201)를 입력받는다. 병렬 말뭉치(201)는 소스 문장과 타겟 문장으로 이루어져 있다.

[0050] 술어 및 논항 인식부(210)는 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식한다. 술어 및 논항 인식부(210)는 복수의 병렬 말뭉치를 입력받아 소스 문장의 구문 분석 결과를 이용하여 소스 문장의 술어 및 논항을 인식하고, 단어 정렬 결과를 이용하여 상기 인식된 소스 문장의 술어 및 논항에 대응하는 타겟 문장의 술어 및 논항을 인식한다.

[0051] 번역 패턴 후보 추출부(220)는 상기 인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출하고 해당 출현 빈도수를 확인한다. 번역 패턴 후보 추출부(220)는

서로 다른 소스 문장 및 타겟 문장에서 동일한 번역 패턴 후보를 추출하고 상기 추출된 동일한 번역 패턴 후보의 출현 빈도수를 구한다.

[0052] 번역 패턴 후보 검증부(230)는 상기 추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 상기 추출된 번역 패턴 후보를 검증하고, 상기 검증 결과에 따라 제1 동사구 번역 패턴을 생성한다. 번역 패턴 후보 검증부(230)는 상기 추출된 번역 패턴 후보에 대한 원문부 패턴의 출현 빈도수를 미리 설정된 임계치와 비교하여 임계치 이하인 경우에 해당 번역 패턴 후보를 제거한다. 또한, 번역 패턴 후보 검증부(230)는 원문부 패턴이 동일한 복수의 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에 해당 번역 패턴 후보를 제거한다. 번역 패턴 후보 검증부(230)는 상기 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에, 상기 대역부 패턴의 논항이 복수 개이고 상기 대역부 패턴을 포함하는 다른 대역부 패턴이 존재하는지 여부에 따라 상기 추출된 동일한 번역 패턴 후보들을 검증한다.

[0053] 번역 패턴 일반화부(240)는 상기 생성된 제1 동사구 번역 패턴들을 일반화시켜 제2 동사구 번역 패턴을 생성한다. 번역 패턴 일반화부(240)는 상기 생성된 제1 동사구 번역 패턴들의 논항을 구문 표지로 일반화하는 경우에 유효한 동사구 번역 패턴인지를 확인하여 일반화시켜 제2 동사구 번역 패턴을 생성한다. 번역 패턴 일반화부(240)는 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시키고, 상기 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들과 상기 추출된 튜플들의 빈도수를 이용하여 제2 동사구 번역 패턴을 생성한다. 번역 패턴 일반화부(240)는 상기 추출된 튜플들의 빈도수 중에서 가장 빈도수가 큰 튜플에 해당하는 논항을 제외한 다른 논항들을 일반화시켜 제2 동사구 번역 패턴을 생성한다. 번역 패턴 일반화부(240)는 상기 생성된 제2 동사구 번역 패턴에 대해서 번역 모호성을 분석하여 제2 동사구 번역 패턴을 결정한다. 번역 패턴 일반화부(240)는 상기 결정된 제2 동사구 번역 패턴에 대해서 상기 번역 패턴 일반화부에서의 번역 패턴 일반화를 반복적으로 수행한다.

[0054] 번역 패턴 후보 저장부(250)는 상기 번역 패턴 후보 추출부에서 추출된 번역 패턴 후보를 저장한다.

[0055] 번역 패턴 저장부(260)는 상기 번역 패턴 후보 검증부에서 생성된 제1 동사구 번역 패턴을 저장한다.

[0056] 도 3 은 본 발명에 따른 도 2의 술어 및 논항 인식부에서의 술어 및 논항 인식에 대한 일실시예 예시도이다.

[0057] 도 3에 도시된 바와 같이, 술어 및 논항 인식부(210)는 입력된 병렬 말뭉치(201)의 소스 문장을 분석하고, 소스 문장(310)의 구문 분석 결과(301)와 소스 문장 및 타겟 문장(320) 간의 단어 정렬 결과를 이용하여 술어 및 논항 인식한다. 술어 및 논항 인식부(210)는 술어 및 논항 정보를 함께 출력해주는 구문 분석기를 이용하여 술어 및 논항을 찾을 수 있다. 술어 및 논항 인식부(210)는 소스 문장에서 술어와 논항을 인식한 후에 단어 정렬 결과를 이용해 술어와 논항에 대응되는 타겟 단어를 찾는다. 구문 분석된 문장에서 술어와 논항을 인식하는 것은 본 발명에 적용되는 실시예 이외에 다양한 방법이 존재하므로 특정 인식 방법에 한정되지 않는다.

[0058] 이때, 언어적 차이 및 의역 등으로 인해 소스 문장(310)의 술어와 대응되는 부분이 타겟 문장(320)에서는 술어가 아닌 경우가 발생할 수 있다. 예를 들어, 소스 문장(310)이 "He succeeded in sailing the world"라는 영어 문장이고, 타겟 문장(320)이 "그는 세계 일주 항해에 성공했다"라는 한국어 번역 문장인 경우를 살펴보기로 한다. 영어 문장과 그에 대한 한국어 번역 문장에서 영어 쪽 술어인 "sailing"은 한국어 번역 문장에서는 "항해"라는 명사로 번역된다. 따라서 영어 쪽 동사구 패턴에 해당하는 부분이 한국어에서는 동사구가 아닌 경우가 생길 수 있다.

[0059] 하지만, 이러한 경우에도 소스 문장(310)의 동사구 패턴을 하나의 단위로 묶어서 번역하는 것이 더 자연스러운 번역 결과를 생성할 수 있다. 그러므로 본 발명에 따른 동사구 번역 패턴을 구축하는 것이 유리하다. 따라서 본 발명에서 소스 문장(310)의 동사구 패턴(원문부 패턴)과 그에 대한 타겟 문장(320)의 번역 패턴을 동사구 번역 패턴으로 간주하기로 한다.

[0060] 한편, 번역 패턴 후보 추출부(220)는 술어 및 논항 인식부(210)에서 인식된 소스 문장(310)의 술어 및 논항들과 타겟 문장(320)의 술어 및 논항들을 이용해 각각 원문부 패턴, 대역부 패턴 및 기본 동사구 번역 패턴을 추출한다. 여기서, 생성한 기본 동사구 번역 패턴은 아직 오류 검증을 거치지 않았으므로, 번역 패턴 후보 저장부(150)에 저장된다. 서로 다른 병렬 말뭉치(201)의 병렬 문장에서 동일한 번역 패턴 후보가 추출될 수 있으므로, 하나의 동사구 번역 패턴은 전체 병렬 말뭉치에서 여러 번 추출될 수 있다. 번역 패턴 후보 추출부(220)는 번역 패턴 후보를 추출할 때, 전체 병렬 말뭉치(201)에서 해당 번역 패턴 후보가 추출되는 추출 빈도수를 확인하고, 번역 패턴 후보 저장부(150)는 그 확인된 추출 빈도수를 함께 저장할 수 있다.

[0061] 번역 패턴 후보 검증부(230)는 번역 패턴 후보 추출부(220)에서 모두 추출된 번역 패턴 후보를 검증한다. 여기서, 번역 패턴 후보 검증은 구문 분석, 단어 정렬 오류로 인해 잘못 추출된 동사구 번역 패턴을 제외하기 위한 것이다.

[0062] 번역 패턴 후보의 검증 과정을 구체적으로 살펴보면, 번역 패턴 후보 검증부(230)는 번역 패턴 후보 저장부(250)에 저장된 모든 번역 패턴 후보에 대해 번역 패턴의 원문부 패턴 및 원문부 패턴(LHS)의 출현 빈도를 구한다. 만약, 원문부 패턴의 출현 빈도가 임계값 이하인 경우, 번역 패턴 후보 검증부(230)는 잘못 추출된 번역 패턴으로 간주하여 패턴 후보 저장부(250)에서 제거한다.

[0063] 번역 패턴 후보 검증부(230)는 원문부 패턴이 동일한 여러 개의 번역 패턴 후보들 중에서 대역부 패턴(RHS)의 빈도가 일정 횟수 이하인 것은 번역 패턴 후보 저장부(250)에서 제거한다. 이때, 번역 패턴 후보 검증부(230)는 대역부 패턴의 논항의 개수가 n개인 번역 패턴은 해당 대역부 패턴을 포함하는 번역 패턴이 존재하는지 여부를 확인한다. 확인 결과, 번역 패턴 후보 검증부(230)는 이러한 번역 패턴이 존재할 경우에 해당 번역 패턴을 제거하지 않는다. 예를 들어, 원문부 패턴은 동일하고 대역부 패턴은 "P1 = < … , 1:[부시 case=는] 3:[취임식 case=을] 2:[V 갖]>" 및 "P2 = < … , 1:[부시 case=는] 4:[1월] 3:[취임식 case=을] 2:[V 갖]>"와 같이 번역 패턴 후보 P1, P2가 있는 경우를 살펴보면 다음과 같다. 번역 패턴 후보 P1의 대역부 패턴의 출현 빈도가 어떤 임계값 이하인 경우라 하더라도, P1의 대역부 패턴은 P2의 대역부 패턴에 포함된다. 그러므로 번역 패턴 후보 검증부(230)는 P2의 대역부 패턴의 출현 빈도가 임계값 이상일 경우 P1을 제거하지 않는다. 이는 번역 과정에서 어떤 논항이 생략되는 현상을 고려하기 번역 패턴 후보를 검증하기 위함이다.

[0064] 상기와 같은 번역 패턴 후보의 검증 과정을 통해, 번역 패턴 후보 검증부(230)는 번역 패턴 후보 저장부(250)에서 오류 가능성이 큰 번역 패턴 후보들을 제거한다. 그리고 번역 패턴 후보 검증부(230)는 제거되지 않고 남은 번역 패턴 후보들을 기본 동사구 번역 패턴으로 간주하고, 번역 패턴 저장부(260)에 저장시킨다.

[0065] 전술된 바와 같이, 일반 동사구 번역 패턴은 기본 동사구 번역 패턴에서 일부 논항 또는 모든 논항을 구문 표지로 일반화한 것을 의미한다. 번역 패턴 일반화부(240)는 번역 패턴 후보 검증부(230)에서 검증된 기본 동사구 번역 패턴을 일반화시킨다. 동사구 번역 패턴을 일반화하는 것은 구축된 번역 패턴의 적용률을 높이기 위함이다. 하지만, 과도한 일반화는 번역 패턴의 번역 모호성을 떨어뜨리게 된다. 따라서, 도 4를 참조하여 번역 패턴 일반화부(240)에서 수행되는 번역 패턴 일반화 과정을 살펴보기로 한다.

[0066] 도 4 는 본 발명에 따른 도 2의 번역 패턴 일반화부에서의 번역 패턴 일반화에 대한 일실시예 예시도이다.

[0067] 도 4에 도시된 바와 같이, "take participant in party" 및 "파티:[에] 참석하", "take participant in meeting" 및 "미팅:[에] 참석하", "take rabbit in trap" 및 "토끼:[를] 덫:[으로[잡"이 포함된 기본 동사구 번역 패턴들(410)이 있는 경우를 살펴보기로 한다.

[0068] 번역 패턴 일반화부(240)는 "take participant in party"와 "take participant in meeting"로부터 "take participant in NP"와 같이 한 개의 논항만을 적절하게 일반화할 수 있다. 이러한 적절한 일반화(420)는 이에 해당하는 다양한 문장들을 "~에 참석하다"로 번역하는데 도움을 줄 수 있다.

[0069] 하지만, "take NP in NP"와 같이 두 개의 논항 모두를 일반화하는 과도한 일반화(430)의 경우에는 "take"를 "참석하다"로 번역해야 하는지 또는 "잡다"로 번역해야 할지 알 수 없는 번역 모호성이 발생하게 된다.

[0070] 결과적으로 동사구 번역 패턴의 유용성이 감소하게 된다. 따라서 번역 패턴 일반화부(240)는 논항을 하나씩 구문 표지로 일반화했을 때, 여전히 유용한 패턴인지를 판단하여 일반 동사구 번역 패턴을 생성한다.

[0071] 일반 동사구 번역 패턴 생성 과정을 살펴보면 다음과 같다.

[0072] 번역 패턴 일반화부(240)는 기본 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 하나의 그룹으로 묶는다. 이를 "Gv"라 한다. 예를 들어, 기본 동사구 번역 패턴들 중에서 원문부 패턴의 술어가 "take"인 패턴들은 "Gtake" 그룹에 속하게 된다.

[0073] 그리고 번역 패턴 일반화부(240)는 동일한 그룹에 속하는 모든 패턴에 대해서 "원문부 술어, 원문부 논항, 대역부 술어"가 포함된 튜플(Tuple)을 추출하고 각 튜플의 빈도를 구한다.

[0074] 예를 들어, 도 4에 도시된 기본 동사구 번역 패턴들(410)을 일반 동사구 번역 패턴들로 일반화시키기 위해, 생성되는 튜플을 살펴보면 다음과 같다.

[0075] 번역 패턴 일반화부(240)는 동일한 그룹 중에서 "take, participant, 참석하"인 튜플을 추출한다. 여기서, 튜플

의 빈도는 2가 된다. 번역 패턴 일반화부(240)는 동일한 그룹 중에서 "take, in party, 참석하"인 튜플을 추출한다. 여기서, 튜플의 빈도는 1가 된다. 번역 패턴 일반화부(240)는 동일한 그룹 중에서 "take, in meeting, 참석하"인 튜플을 추출한다. 여기서, 튜플의 빈도는 1가 된다. 번역 패턴 일반화부(240)는 동일한 그룹 중에서 "take, rabbit, 잡"인 튜플을 추출한다. 여기서, 튜플의 빈도는 1가 된다. 번역 패턴 일반화부(240)는 동일한 그룹 중에서 "take, in trap, 잡"인 튜플을 추출한다. 여기서, 튜플의 빈도는 1가 된다.

[0076] 번역 패턴 일반화부(240)는 추출된 튜플들 중에서 가장 빈도가 높은 튜플에 해당하는 논항을 제외한 다른 논항들을 하나씩 일반화한다. 여기서, "take, participant, 참석하"인 튜플이 가장 높은 빈도를 차지하므로, 번역 패턴 일반화부(240)는 "participant"를 제외한 다른 논항을 먼저 일반화를 수행하여 일반 동사구 번역 패턴 후보를 생성한다.

[0077] 번역 패턴 일반화부(240)는 생성된 번역 패턴 후보에 대해 번역 모호성을 조사하고, 번역 모호성이 없는 번역 패턴 후보는 올바른 일반 동사구 번역 패턴으로 간주한다.

[0078] 번역 패턴 일반화부(240)는 새로이 만들어진 일반 동사구 번역 패턴들에 대해 2 내지 4번 번역 패턴 일반화 과정을 반복하면서 더욱 일반화된 동사구 번역 패턴을 생성할 수 있다.

[0079] 본 발명에서 동사구 번역 패턴 일반화는 구문 표지로 일반화는 경우로 한정되지 않는다. 의미 온톨로지와 같은 언어 자원이 주어질 경우, 본 발명을 이용해 의미 코드 등으로 일반화하는 것도 가능하다.

[0080] 도 5 는 본 발명에 따른 병렬 말뭉치를 이용한 동사구 번역 패턴 구축 방법에 대한 일실시예 흐름도이다.

[0081] 술어 및 논항 인식부(210)는 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식한다(502). 술어 및 논항 인식부(210)는 복수의 병렬 말뭉치를 입력받아 소스 문장의 구문 분석 결과를 이용하여 소스 문장의 술어 및 논항을 인식하고, 단어 정렬 결과를 이용하여 상기 인식된 소스 문장의 술어 및 논항에 대응하는 타겟 문장의 술어 및 논항을 인식한다.

[0082] 번역 패턴 후보 추출부(220)는 술어 및 논항 인식부(210)에서 인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출한다(504).

[0083] 번역 패턴 후보 추출부(220)는 추출된 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보에 대해서 해당 출현 빈도수를 확인한다(504). 번역 패턴 후보 추출부(220)는 서로 다른 소스 문장 및 타겟 문장에서 동일한 번역 패턴 후보를 추출하고 그 추출된 동일한 번역 패턴 후보의 출현 빈도수를 구한다.

[0084] 번역 패턴 후보 검증부(230)는 번역 패턴 후보 추출부(220)에서 추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 상기 추출된 번역 패턴 후보를 검증한다(508).

[0085] 번역 패턴 후보 검증부(230)는 검증 결과에 따라 제1 동사구 번역 패턴을 생성한다(510). 여기서, 검증 과정을 구체적으로 살펴보면, 번역 패턴 후보 검증부(230)는 번역 패턴 후보에 대한 원문부 패턴의 출현 빈도수를 미리 설정된 임계치와 비교하여 임계치 이하인 경우에 해당 번역 패턴 후보를 제거한다. 그리고 번역 패턴 후보 검증부(230)는 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에 해당 번역 패턴 후보를 제거한다. 번역 패턴 후보 검증부(230)는 추출된 동일한 번역 패턴 후보들 중에서 대역부 패턴의 출현 빈도수가 미리 설정된 임계치 이하인 경우에, 대역부 패턴의 논항이 복수 개이고 대역부 패턴을 포함하는 다른 대역부 패턴이 존재하는지 여부에 따라 동일한 번역 패턴 후보들을 검증할 수 있다.

[0086] 번역 패턴 일반화부(240)는 번역 패턴 후보 검증부(230)에서 생성된 제1 동사구 번역 패턴들을 일반화시켜 제2 동사구 번역 패턴을 생성한다. 번역 패턴 일반화부(240)는 번역 패턴 후보 검증부(230)에서 생성된 제1 동사구 번역 패턴들의 논항을 구문 표지로 일반화하는 경우에 유효한 동사구 번역 패턴인지를 확인하여 일반화시켜 제2 동사구 번역 패턴을 생성할 수 있다. 여기서, 번역 패턴 일반화 과정을 구체적으로 살펴보면, 번역 패턴 일반화부(240)는 제1 동사구 번역 패턴에서 원문부 패턴의 술어가 동일한 패턴을 그룹화시킨다. 그리고 번역 패턴 일반화부(240)는 동일한 그룹에 속하는 모든 패턴에 대해서 튜플들을 추출한다. 이어서, 번역 패턴 일반화부(240)는 추출된 튜플들의 빈도수를 이용하여 제2 동사구 번역 패턴을 생성한다. 또한, 번역 패턴 일반화부(240)는 추출된 튜플들의 빈도수 중에서 가장 빈도수가 큰 튜플에 해당하는 논항을 제외한 다른 논항들을 일반화시켜 제2 동사구 번역 패턴을 생성할 수 있다. 번역 패턴 일반화부(240)는 생성된 제2 동사구 번역 패턴에 대해서 번역 모호성을 분석하여 제2 동사구 번역 패턴을 결정할 수 있다. 번역 패턴 일반화부(240)는 결정된 제2 동사구 번역 패턴에 대해서 번역 패턴 일반화 과정을 반복적으로 수행할 수 있다.

[0087] 전술된 바와 같이, 본 발명에 따른 동사구 번역 패턴 구축 장치(200)는 병렬 말뭉치와 구문 분석 정보를 이용해 동사구 번역 패턴을 구축한다.

[0088] 동사구 번역 패턴 구축 장치(200)는 병렬 말뭉치상의 각 단어의 공기 빈도만을 이용하여 자동으로 단어 정렬을 수행하기 때문에 사전과 같은 부가적인 언어 자원을 필요로 하지 않으며, 대상으로 하는 두 언어 중에 한쪽 언어의 파싱 정보를 사용하여 트리 정렬을 수행할 수 있다. 동사구 번역 패턴 구축 장치(200)는 그 결과로부터 동사구 패턴을 추출하므로 대역어 중의성 해소의 문제를 자동으로 해결함은 물론, 원시 언어의 가능한 모든 동사구 패턴에 대한 타겟 언어의 대응 패턴이 자동으로 완성하여 추출할 수 있다.

[0089] 동사구 번역 패턴 구축 장치(200)는 언어 독립성 측면에서 소스 문장에 대한 구문 분석 결과와 단어 정렬 결과만을 사용하므로 다양한 언어쌍(Language Pair)에 적용 가능하다.

[0090] 동사구 번역 패턴 오류 최소화를 위해, 동사구 번역 패턴 구축 장치(200)는 자동으로 패턴을 구축하는 과정에서 잘못 생성된 번역 패턴을 자동으로 제거한다. 본 발명은 대량의 병렬 말뭉치에서 동사구 번역 패턴 후보들을 추출하고, 각 패턴 후보들의 출현 빈도를 구한다. 동사구 번역 패턴 구축 장치(200)는 그런 다음 번역 패턴을 구성하는 각 요소의 출현 빈도와 번역 확률을 이용해 잘못 생성됐을 법한 동사구 번역 패턴을 제거하고 남은 후보들만을 올바른 동사구 번역 패턴으로 간주한다.

[0091] 동사구 번역 패턴 구축 장치(200)는 술어-논항 공기 빈도 및 술어 대역 확률을 이용한 번역 패턴 일반화 수준을 결정할 수 있다. 자연어 처리에서 패턴을 일반화하는 것은 패턴의 적용률을 향상시키는 반면, 패턴의 모호성을 증가시키는 문제가 있다. 이것은 동사구 번역 패턴의 경우에도 동일하다. 어떤 동사구 번역 패턴을 일반화할 경우 좀 더 많은 문장에 적용될 수 있지만, 지나친 일반화는 번역 모호성을 가중시킨다. 따라서 동사구 번역 패턴 구축 장치(200)는 번역 패턴의 술어와 논항의 공기 빈도(Co-occurrence Frequency)와 술어의 번역 확률을 이용해 적절한 일반화 수준을 결정할 수 있다.

[0092] 한편, 본 발명은 상기 동사구 번역 패턴 구축 방법을 소프트웨어적인 프로그램으로 구현하여 컴퓨터로 읽을 수 있는 소정 기록매체에 기록해 둠으로써 다양한 재생장치에 적용할 수 있다.

[0093] 다양한 재생장치는 PC, 노트북, 휴대용 단말 등일 수 있다.

[0094] 예컨대, 기록매체는 각 재생장치의 내장형으로 하드디스크, 플래시 메모리, RAM, ROM 등이거나, 외장형으로 CD-R, CD-RW와 같은 광디스크, 콤팩트 플래시 카드, 스마트 미디어, 메모리 스틱, 멀티미디어 카드일 수 있다.

[0095] 이 경우, 컴퓨터로 읽을 수 있는 기록매체에 기록한 프로그램은, 앞서 설명한 바와 같이, 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식하는 술어 및 논항 인식 과정과, 상기 인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출하고 해당 출현 빈도수를 확인하는 번역 패턴 후보 추출 과정과, 상기 추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 상기 추출된 번역 패턴 후보를 검증하고, 상기 검증 결과에 따라 제1 동사구 번역 패턴을 생성하는 번역 패턴 후보 검증 과정과, 상기 생성된 제1 동사구 번역 패턴들을 일반화시켜 제2 동사구 번역 패턴을 생성하는 번역 패턴 일반화 과정을 포함하여 실행될 수 있다.

[0096] 이상의 설명은 본 발명을 예시적으로 설명한 것에 불과하며, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 본 발명의 기술적 사상에서 벗어나지 않는 범위에서 다양한 변형이 가능할 것이다. 따라서 본 발명의 명세서에 개시된 실시 예들은 본 발명을 한정하는 것이 아니다. 본 발명의 범위는 아래의 특허청구범위에 의해 해석되어야 하며, 그와 균등한 범위 내에 있는 모든 기술도 본 발명의 범위에 포함되는 것으로 해석해야 할 것이다.

산업상 이용가능성

[0097] 본 발명은 복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 단어 정렬 결과를 이용하여 술어 및 논항을 인식하고 그 인식된 술어 및 논항을 이용하여 번역 패턴 후보 및 출현 빈도수를 추출한 후, 번역 패턴 후보에 대한 검증을 통해 기본 동사구 번역 패턴을 생성하고 이를 일반화시켜 일반 동사구 번역 패턴을 생성함으로써, 다양한 언어쌍에 적용 가능하고 동사구 번역 패턴 오류를 최소화할 수 있으며, 동사구 번역 패턴의 술어 및 논항의 공기 빈도와 술어의 번역 확률을 이용해 적절한 일반화 수준을 결정할 수 있다.

부호의 설명

[0098] 200: 동사구 번역 패턴 구축 장치 210: 술어 및 논항 인식부

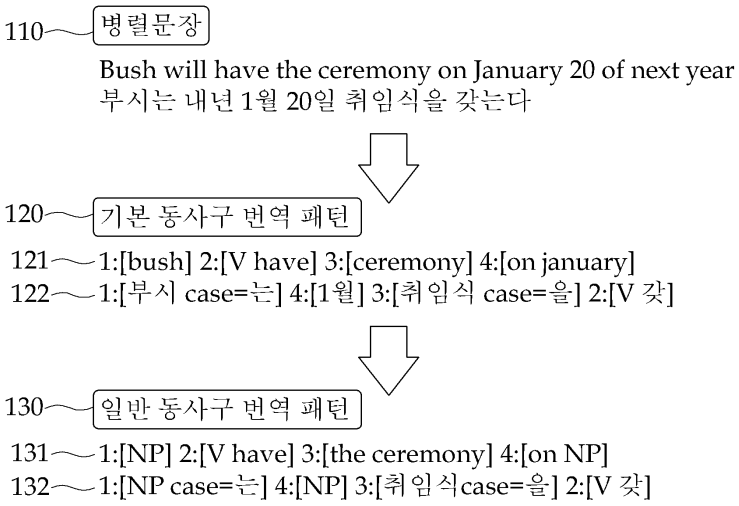
220: 번역 패턴 후보 추출부 230: 번역 패턴 후보 검증부

240: 번역 패턴 일반화부 250: 번역 패턴 후보 저장부

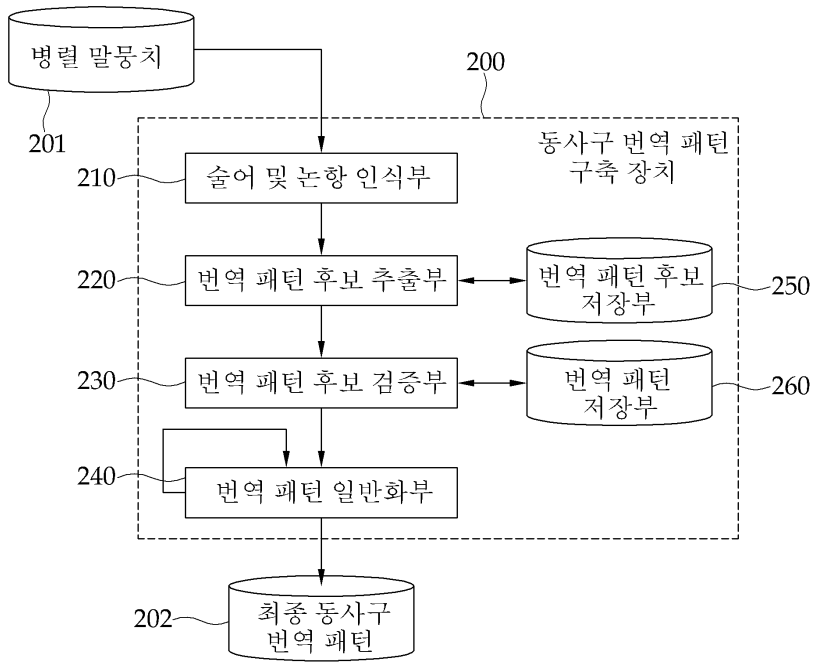
260; 번역 패턴 저장부

도면

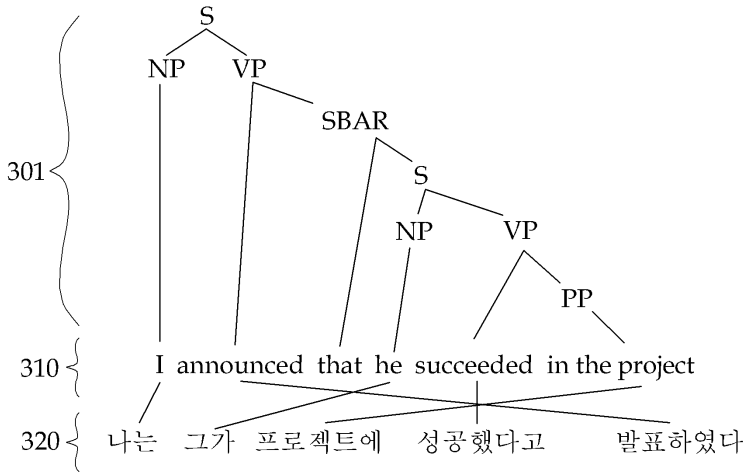
도면1



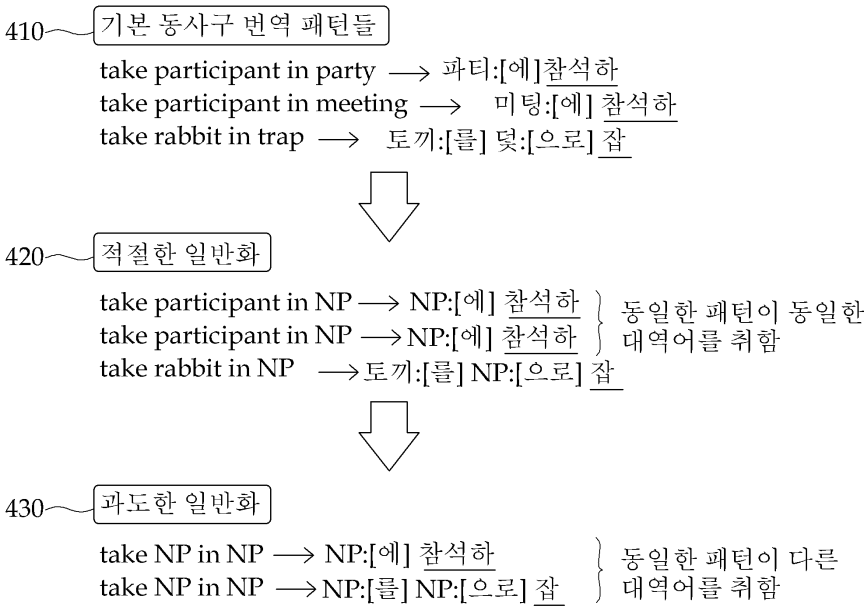
도면2



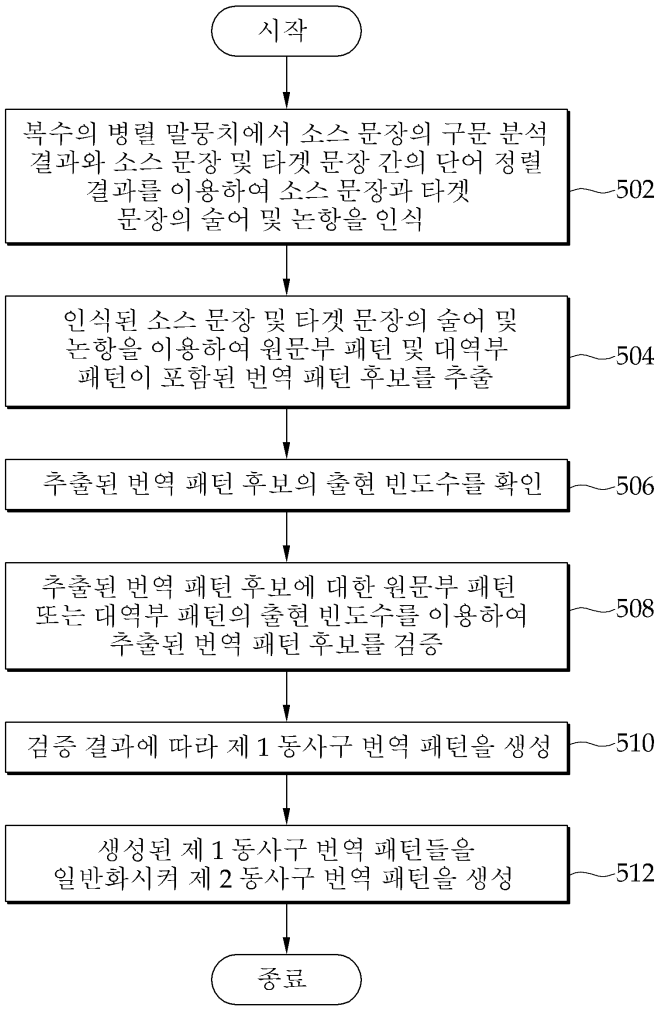
도면3


S
NP
VP
SBAR
S
NP
VP
PP
301
310
I announced that he succeeded in the project
320
나는 그가 프로젝트에 성공했다고 발표하였다

도면4


410
기본 동사구 번역 패턴들
take participant in party ⟶ 파티:[에]참석하
take participant in meeting ⟶ 미팅:[에] 참석하
take rabbit in trap ⟶ 토끼:[를] 덫:[으로] 잡
420
적절한 일반화
take participant in NP ⟶ NP:[에] 참석하
take participant in NP ⟶ NP:[에] 참석하
동일한 패턴이 동일한 대역어를 취함
take rabbit in NP ⟶ 토끼:[를] NP:[으로] 잡
430
과도한 일반화
take NP in NP ⟶ NP:[에] 참석하
take NP in NP ⟶ NP:[를] NP:[으로] 잡
동일한 패턴이 다른 대역어를 취함

도면5


시작
복수의 병렬 말뭉치에서 소스 문장의 구문 분석 결과와 소스 문장 및 타겟 문장 간의 단어 정렬 결과를 이용하여 소스 문장과 타겟 문장의 술어 및 논항을 인식
502
인식된 소스 문장 및 타겟 문장의 술어 및 논항을 이용하여 원문부 패턴 및 대역부 패턴이 포함된 번역 패턴 후보를 추출
504
추출된 번역 패턴 후보의 출현 빈도수를 확인
506
추출된 번역 패턴 후보에 대한 원문부 패턴 또는 대역부 패턴의 출현 빈도수를 이용하여 추출된 번역 패턴 후보를 검증
508
검증 결과에 따라 제 1 동사구 번역 패턴을 생성
510
생성된 제 1 동사구 번역 패턴들을 일반화시켜 제 2 동사구 번역 패턴을 생성
512
종료