

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국



(43) 국제공개일
2012년 5월 10일 (10.05.2012)

PCT

(10) 국제공개번호
WO 2012/060540 A1

- (51) 국제특허분류:
G06F 17/28 (2006.01)
- (21) 국제출원번호: PCT/KR2011/005325
- (22) 국제출원일: 2011년 7월 20일 (20.07.2011)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보:
10-2010-0109546 2010년 11월 5일 (05.11.2010) KR
- (71) 출원인 (US 을(를) 제외한 모든 지정국에 대하여): 에스케이텔레콤 주식회사 (SK TELECOM. CO., LTD) [KR/KR]; 서울특별시 중구 을지로 2가 11번지, 100-999 Seoul (KR).
- (72) 발명자: 겸
- (75) 발명자/출원인 (US 에 한하여): 황영숙 (HWANG, Young Sook) [KR/KR]; 서울특별시 성북구 돈암 1동 풍림아파트 105-502, 136-061 Seoul (KR). 김상범 (KIM, Sang-Bum) [KR/KR]; 서울특별시 노원구 상계 1동 현대 2차아파트 202동 1207호, 139-893 Seoul (KR). 윤창호 (YIN, Chang Hao) [CN/KR]; 서울특별시 성북구 정릉 2동 푸른동아 아파트 107동 1303호, 136-778 Seoul (KR). 이연수 (LEE, Yeon Su) [KR/KR]; 서울특별시 관악구 봉천 9동 관악백산블루밍아파트 205동 1904호, 151-050 Seoul (KR). 이승욱 (LEE,

Seong Wook) [KR/KR]; 서울특별시 성북구 안암동 5가 11-1 503호, 136-075 Seoul (KR). 임해창 (RIM, Hae Chang) [KR/KR]; 서울특별시 서초구 방배동 1028-1 경남아파트 3동 401호, 137-060 Seoul (KR).

(74) 대리인: 전철용 (JUN, Chul-Yong) 등; 서울특별시 중구 충무로 3가 60-1번지 극동빌딩 14층, 100-705 Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

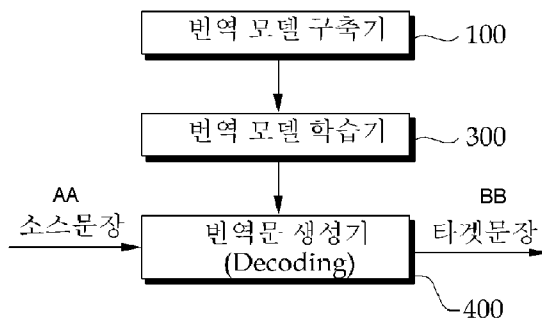
(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,

[다음 쪽 계속]

(54) Title: MACHINE TRANSLATION DEVICE AND MACHINE TRANSLATION METHOD IN WHICH A SYNTAX CONVERSION MODEL AND A VOCABULARY CONVERSION MODEL ARE COMBINED

(54) 발명의 명칭: 구문 구조 변환 모델과 어휘 변환 모델을 결합한 기계 번역 장치 및 기계 번역 방법

[Fig. 1]



100 ... Translation model constructor
300 ... Translation model learning device
400 ... Translated sentence generator (Decoding)
AA ... Source sentence
BB ... Target sentence

(57) Abstract: The present invention relates to statistical machine translation, and a machine translation device and method for supplementing shortcomings and combining advantages of conventional phrase-based statistical machine translation (SMT) and syntax-based SMT. To this end, syntax conversion knowledge and vocabulary conversion knowledge are extracted from parallel corpora to extract respective conversion probabilities and simultaneously acquire a generation probability of a target language from a single corpus, a weighted translation model is modeled by allowing each of the conversion knowledge and each of the probabilities to learn using a translation model learning device, and the modeled translation model is applied to a source sentence inputted in real time to generate a target sentence through decoding processes of a syntax converter and a vocabulary converter.

(57) 요약서: 본 발명은 통계적 기계 번역에 관한 것으로, 병렬 말뭉치로부터 구문 구조 변환 지식과 어휘 변환 지식을 추출하여 각각의 변환 확률을 추출하는 동시에 단일 말뭉치로부터 타겟 언어에 대한 생성 확률을 획득하고, 이들 각각의 변환 지식과 확률을 번역 모델 학습기를 통해 학습시켜 가중치를 둔 번역 모델을 모델링하며, 이렇게 모델링된 번역 모델을 실시간으로 입력된 소스 문장에 대해 적용

하여 구문 구조 변환기 및 어휘 변환기의 디코딩 과정을 거침으로써 타겟 문장을 생성하여, 기존의 구 기반 통계적 기계 번역(Phrase-Based SMT)과 구문 구조 기반 통계적 기계 번역(Syntax-Based SMT)의 단점을 해소하고 장점을 결합하고자 하는 기계 번역 장치 및 기계 번역 방법을 제공한다.



SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG). **공개:**

— 국제조사보고서와 함께 (조약 제 21 조(3))

명세서

발명의 명칭: 구문 구조 변환 모델과 어휘 변환 모델을 결합한 기계 번역 장치 및 기계 번역 방법

기술분야

- [1] 본 발명은 통계적 기계 번역에 관한 것으로서, 더욱 상세하게는 번역의 단계를 구문 구조 변환 과정과 어휘 변환 과정의 2단계로 모델링하고, 이 모델을 실시간으로 입력된 소스 언어 문장에 대해 적용하여 구문 구조 변환기 및 어휘 변환기의 디코딩 과정을 거침으로써 타겟 언어 문장을 생성하는 구문 구조 변환 모델과 어휘 변환 모델을 결합한 기계 번역 장치 및 기계 번역 방법에 관한 것이다.

배경기술

- [2] 자동 번역 기술은 한 언어를 다른 언어로 자동으로 전환해주는 소프트웨어적 기술을 의미한다. 이러한 기술은 20세기 중반부터 미국에서 군사적인 목적으로 연구가 시작되었으며, 지금은 세계적으로 정보 접근 범위의 확대와 휴먼 인터페이스의 혁신을 목적으로 다수의 연구소와 민간 기업에서 활발히 연구 중이다.
- [3] 자동 번역 기술의 초기 단계에서는 전문가가 수동으로 작성한 양국어(Bilingual) 사전과 한 언어를 다른 언어로 변환할 수 있는 규칙을 기반으로 발전되어 왔다. 그러나 컴퓨팅 파워의 급속한 발전이 진행된 21세기 초부터는 대량의 데이터로부터 통계적으로 번역 알고리즘을 자동으로 학습하는 통계적 기계 번역 기술에 대한 개발이 활발히 전개되고 있다.
- [4] 통계적 기계 번역(Statistical Machine Translation, SMT) 시스템은 번역 과정을 통계적으로 모델링하고 대량의 병렬 말뭉치로부터 번역 지식 및 번역 확률과, 타겟 언어에 대한 생성 확률 등을 학습하여 이를 근거로 입력된 소스 문장에 대해 가장 적합한 타겟 문장을 생성하는 것이다.
- [5] 최근의 통계적 기계 번역 시스템은 크게 구 기반 통계적 기계 번역(Phrase-Based SMT, 이하 PBSMT) 방식과 구문 구조(또는 문법) 기반 통계적 기계 번역(Syntax-Based SMT, 이하 SBSMT) 방식으로 구분할 수 있다.
- [6] PBSMT는 개별 단어 대 단어의 번역이 아닌 연속된 단어열(이하 구(phrase))을 하나의 단위로 하여 번역하는 것으로서, 구(phrase) 대 구(phrase)로의 번역 지식과 번역 확률을 학습한 후 디코딩 시 가장 확률이 높은 구(phrase) 조합을 생성하는 방법이다.
- [7] PBSMT의 가장 대표적인 모델로는 (Koehn et al., 2003; Och and Ney, 2004a) 등이 있다. 이 모델은 간단하며, 근거리 어순 변경이 잘 되고 여러 단어로 표현된 번역이 자연스럽게 수행된다는 특징이 있다. 그러나, 이 모델은 원거리 어순 변경이 잘 이루어지지 않으며, 특히 영-한 번역처럼 어순이 많이 다른 언어

쌍에서는 큰 문제가 된다. 그 이유는 PBSMT의 번역 모델에서 문장의 어순을 결정하는 문법간 변환을 명시적으로 모델링하지 않고 구(phrase)들 간의 가능한 모든 순열 중 일부만을 고려하기 때문이다.

- [8] 따라서, 최근에는 문법기반 구문 구조(Syntax)의 변환을 모델링하는 방법에 대해 많이 연구되고 있으며 이를 SBSMT라고 한다. SBSMT는 구문 구조 변환 지식을 학습하기 위해 병렬 말뭉치에서 두 언어의 대응하는 구문 트리(tree)로부터 트리 대 트리간, 혹은 트리 대 문자열간의 변환 지식 및 확률을 학습한다. 이 SBSMT는 PBSMT에 비해 원거리 어순 변경이 잘 되고, 불연속적인 구에 대해 번역이 잘 수행된다는 특징이 있다. 그러나 SBSMT는 파서(parser)의 성능에 크게 의존적이며 번역 지식이 문법적인 구 단위로 한정되기 때문에 사용할 번역 지식 자체가 매우 적어지게 된다. 이에 따라, 사용할 번역 지식이 없는 경우 연속되는 단어열의 번역이 단순한 단어 대 단어로의 번역이 되거나 연결되는 단어와 어울리지 않는 부자연스러운 번역이 되는 문제가 있다. 대표적인 방법으로는 (Galley et al., 2004, 2006), (Lavie et al., 2008), (Yamada and Knight, Gildea et al.)등이 있다.

- [9] 이처럼, 종래의 통계적 기계 번역 기술은 PBSMT 방식의 경우 연속되는 단어 번역의 자연스러움은 향상되었지만 원거리 어순 변경에 실패하여 전혀 다른 문장이 되는 문제점이 있다. SBSMT 방식의 경우는 생성된 타겟 문장의 어순은 맞지만 번역 지식 부족으로 단순한 단어 대 단어 번역으로 이루어져 번역이 자연스럽지 못한 문제점이 있다.

발명의 상세한 설명

기술적 과제

- [10] 본 발명은 상기의 문제점을 해결하기 위해 창안된 것으로서, 병렬 말뭉치로부터 구문 구조 변환 지식과 어휘 변환 지식을 추출하여 각각의 변환 확률을 추출하는 동시에 단일 말뭉치로부터 타겟 언어에 대한 생성 확률을 획득하고, 이들 각각의 변환 지식과 확률을 번역 모델 학습기를 통해 학습시켜 가중치를 둔 번역 모델을 모델링하며, 이렇게 모델링된 번역 모델을 실시간으로 입력된 소스 문장에 대해 적용하여 구문 구조 변환기 및 어휘 변환기의 디코딩 과정을 거침으로써 타겟 문장을 생성하여, 기존의 구 기반 통계적 기계 번역(Phrase-Based SMT)과 구문 구조 기반 통계적 기계 번역(Syntax-Based SMT)의 단점을 해소하고 장점은 결합하고자 하는 기계 번역 장치 및 기계 번역 방법을 제공하고자 한다.

과제 해결 수단

- [11] 이를 위하여 본 발명의 제1 측면에 따르면, 본 발명의 통계적 기계 번역 장치는, 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 타겟 문장의 구문 구조 변환 지식과 어휘 변환 지식을 추출하고 추출된 각각의 지식에 대하여 변환 확률을 계산하는 번역 모델

구축기; 상기 번역 모델 구축기를 통해 추출한 각각의 번역 지식과 변환 확률을 학습하여 구문 구조 변환 모델과 어휘 변환 모델을 생성하는 번역 모델 학습기; 및 실시간으로 입력되는 소스 문장에 대하여 상기 번역 모델 학습기를 통해 학습시킨 구문 구조 변환 모델과 어휘 변환 모델을 적용하여 상기 소스 문장을 타겟 문장으로 디코딩하는 번역문 생성기를 포함하는 것을 특징으로 한다.

- [12] 본 발명의 제2 측면에 따르면, 본 발명의 번역 모델 구축 장치는, 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 타겟 문장에 대한 구문 구조 변환 지식을 추출하고 추출된 지식에 대해 변환 확률을 계산하는 구문 구조 변환 지식 추출기와, 상기 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 어휘 변환 지식을 추출하고 추출된 지식에 대해 변환 확률을 계산하는 어휘 변환 지식 추출기를 포함하는 것을 특징으로 한다.
- [13] 본 발명의 제3 측면에 따르면, 본 발명의 번역문 생성 장치는, 실시간으로 입력되는 소스 문장에 대하여 구문 분석하고 분석한 소스 문장의 구문 구조로부터 타겟 문장의 구문 구조 변환 지식을 추출하고 추출된 지식에 대해 변환 확률을 학습시킨 구문 구조 변환기; 상기 구문 구조 변환기를 통해 추출된 타겟 문장의 구문 구조에 제약 조건을 둔 어휘 변환 모델을 기반으로 타겟 어휘열을 생성하는 어휘 변환기; 상기 구문 구조 변환기를 통해 추출한 변환 확률에서 상기 어휘 변환기를 통해 추출한 타겟 어휘열의 생성 확률을 곱합한 후 가장 확률이 높은 타겟 어휘열을 번역문으로 생성하는 확률 계산기를 포함하는 것을 특징으로 한다.
- [14] 본 발명의 제4 측면에 따르면, 본 발명의 번역 모델 구축 방법은, (a) 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 타겟 문장에 대한 구문 구조 변환 지식을 추출하는 단계; (b) 상기 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 어휘 변환 지식을 추출하는 단계; (c) 상기 구문 구조 변환 지식 및 어휘 변환 지식에 대한 변환 확률을 각각 계산하고 각 변환 확률에 대하여 가중치를 학습시키는 단계를 포함하여 것을 특징으로 한다.
- [15] 본 발명의 제5 측면에 따르면, 본 발명의 기계 번역 방법은 (a) 실시간으로 입력되는 소스 문장에 대하여 구문 분석하고 분석한 소스 문장의 구문 구조로부터 타겟 문장의 구문 구조 변환 지식 및 변환 확률을 추출하는 단계; (b) 상기 타겟 문장의 구문 구조 변환 지식으로부터 추출한 타겟 문장의 구문 구조에 제약 조건을 둔 어휘 변환 모델을 기반으로 타겟 어휘열을 생성하는 단계; (c) 상기 타겟 문장의 구문 구조 변환 확률과 상기 타겟 어휘열의 생성 확률을 곱합하여 확률이 높은 타겟 어휘열을 번역문으로 생성하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

- [16] 본 발명에 따르면, 기존의 어휘 기반 번역 모델과 구문 구조 기반 번역 모델의 단점은 해소하고 장점은 결합할 수 있다.
- [17] 즉, 본 발명은 구문 구조 변환 모델을 통해 원거리 단어 재배열의 성능을 향상시킬 수 있으며, 이와 동시에 대량의 비-구문 구조 변환 지식(non-syntactic translation knowledge)을 사용하므로 문맥에 맞는 적절한 번역이 수행되어 번역문의 적절성 및 자연스러움을 향상시킬 수 있다.

도면의 간단한 설명

- [18] 도 1은 본 발명의 실시 예에 따른 기계 번역 장치의 구성을 나타낸 도면.
- [19] 도 2는 본 발명의 실시 예에 따른 기계 번역 장치가 구문 구조 변환 모델과 어휘 변환 모델로 분리된 구성을 나타낸 도면.
- [20] 도 3은 본 발명의 실시 예에 따른 번역문 생성기의 구성을 나타낸 도면.
- [21] 도 4는 도 2에 도시한 구문 구조 변환 지식 추출기의 세부 구성을 나타낸 도면.
- [22] 도 5는 본 발명의 실시 예에 따른 번역문 생성기의 어휘 변환기를 나타낸 도면.
- [23] 도 6은 본 발명의 실시 예에 따른 구문 구조 변환 모델과 어휘 변환 모델의 차이를 설명하기 위한 예시 도면.
- [24] 도 7은 본 발명의 실시 예에 따른 기계 번역 방법을 전체적으로 설명하는 흐름도.
- [25] 도 8은 본 발명의 실시 예에 따른 기계 번역을 위해 번역 모델을 구축하는 방법을 설명하는 흐름도.
- [26] <도면의 주요 부분에 대한 부호의 설명>
- [27] 100: 번역 모델 구축기 110: 구문 구조 변환 지식 추출기
- [28] 120: 구문 구조 변환 지식DB 130: 어휘 변환 지식 추출기
- [29] 140: 어휘 변환 지식DB 150: 언어 모델 생성기
- [30] 160: 언어 모델 300: 번역 모델 학습기
- [31] 400: 번역문 생성기 410: 구문 구조 변환기
- [32] 420: 어휘 변환기 111: 문장 선별기
- [33] 113: 소스 트리 생성기 115: 트리 노드 정렬기
- [34] 117: 트리 변환 지식 추출기 119: 확률 계산기
- [35] 421: 자질 추출기 422: 번역 옵션 생성
- [36] 423: 번역 옵션 제약기 424: 가설 탐색
- [37] 425: 번역 디스톨션 제약기

발명의 실시를 위한 형태

- [38] 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 구성요소에 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.

- [39] 이하에서 언급하는 소스 문장(source sentence) 또는 소스 언어 문장은 번역할 대상이 되는 원시 언어의 문장이고, 타겟 문장(target sentence) 또는 타겟 언어 문장은 소스 문장을 원하는 언어로 번역하여 출력되는 목표 언어의 문장을 의미한다.
- [40] 도 1은 본 발명의 실시 예에 따른 기계 번역 장치의 구성을 나타낸 도면이고, 도 2는 본 발명의 실시 예에 따른 기계 번역 장치가 구문 구조 변환 모델과 어휘 변환 모델로 분리된 구성을 나타낸 도면이다.
- [41] 도 1 및 도 2를 참조하면, 본 발명의 실시 예에 따른 기계 번역 장치는 크게 번역 모델 구축기(100), 번역 모델 학습기(300), 디코더(400)를 포함한다.
- [42] 번역 모델 구축기(100)는 소스 언어 및 타겟 언어로 이루어진 병렬 말뭉치로부터 구문 구조 변환 지식과 어휘 변환 지식을 추출하고 각각의 변환 확률을 계산하는 구문 구조 변환 지식 추출기(110) 및 어휘 변환 지식 추출기(130)를 포함한다. 추출한 구문 구조 변환 지식과 어휘 변환 지식은 각 구문 구조 변환 지식 DB(120)와 어휘 변환 지식 DB(140)에 저장한다.
- [43] 또한, 번역 모델 구축기(100)는 타겟 언어로 이루어진 단일 말뭉치로부터 언어의 생성 확률을 추출하는 언어 모델 생성기(150)를 포함한다. 추출한 언어의 생성 확률은 언어 모델 DB(160)에 저장하여 언어 모델을 구축한다.
- [44] 번역 모델 학습기(300)는 번역 모델 구축기(100)를 통해 추출된 각 자질의 가중치를 학습한다. 이러한 번역 모델 학습기(300)는 (Bertoldi, Haddow, Fouet, 2009)의 최소 오차 학습(Minimum error train training (MERT))의 방법을 동일하게 사용할 수 있다.
- [45] 번역문 생성기(400)는 두 단계의 디코딩(decoding) 과정으로 구성된다. 즉, 1단계는 구문 구조 변환 과정이고 2단계는 어휘 변환 과정이다. 각 단계는 도 3에 도시한 구문 구조 변환기(410)와 어휘 변환기(420) 및 확률계산기(430)에서 수행된다.
- [46] 즉, 1단계를 수행하는 구문 구조 변환기(410)는 확률이 높은 N개의 타겟 구문 구조를 생성하고, 이 결과를 기반으로 2단계의 어휘 변환기(420)에서는 가능한 타겟 어휘열 중 가장 확률이 높은 것을 M개 선택하게 된다. 확률계산기(430)는 최종적으로 구문 구조 변환기(410)에서 생성한 N개의 확률과 어휘 변환기(420)에서 생성한 M개의 확률을 결합한 후 가장 확률이 높은 타겟 어휘열을 번역문으로 출력한다.
- [47] 따라서, 본 발명의 실시 예에 따른 기계 번역 장치는 소스 언어와 타겟 언어로 이루어지는 병렬 말뭉치로부터 구문 구조 변환 지식과 어휘 변환 지식 및 각각의 변환 확률을 추출하고 이와 동시에 단일 말뭉치로부터 타겟 언어의 생성 확률을 획득하여, 이들 각각의 변환 지식과 확률들을 학습시켜 가중치를 둔 번역 모델을 모델링한다. 그리고, 이렇게 모델링이 완성된 번역 모델을 실시간으로 입력되는

소스 문장에 대해 적용하여 구문 구조 변환기(410) 및 어휘 변환기(420)의 두 디코딩 과정을 거침으로써 최종 타겟 문장을 생성한다.

[48] 이러한 기계 번역 장치의 구성요소에 대하여 구체적으로 설명한다.

[49] 번역 모델 구축기(100)는 도 2에 도시한 바와 같이 번역 지식과 번역 확률을 추출하는 모듈이다.

[50] 특히, 본 발명의 실시 예에 따른 번역 모델 구축기(100)는 다음과 같은 통계적 번역 모델에 기반한다. 문장 e 는 구문 구조 $S(e)$ 와 어휘열 $L(e)$ 로 구성된다고 가정하면, 가능한 모든 구조 $S(e)$ 를 사용하여 수학적 1과 같이 나타낼 수 있다.

[51] 수학적 1

$$e = \sum_{S(e)} \langle S(e), L(e) \rangle$$

[52] 이때, 소스 문장 f 를 타겟 문장 e 로 번역하는 것을 아래의 수학적 2와 같이 정의할 수 있다. 즉, 소스 문장 f 가 타겟 문장 e 로 번역되는 확률 중 최대 확률을 갖는 문장 $\hat{e}$ 를 번역문으로 한다.

[53] 수학적 2

$$\hat{e} = \arg \max_e p(e | f)$$

[54] 본 발명에서는 위의 기본 번역 모델 개념을 아래의 수학적 3과 같이 구문 구조 변환 모델과 어휘 변환 모델로 분리한다.

[55] 수학적 3

$$\begin{aligned} p(e | f) &= \sum_{S(e)} p(S(e), L(e) | S(f), L(f)) \\ &= \sum_{S(e)} p(S(e) | S(f), L(f)) \times p(L(e) | S(e), S(f), L(f)) \end{aligned}$$

[56] 즉, 수학적 3에서 $p(S(e) | S(f), L(f))$ 부분은 구문 구조 변환 모델을 나타내며, $p(L(e) | S(e), S(f), L(f))$ 부분은 어휘 변환 모델을 나타낸다.

[57] 구문 구조 변환 모델은 소스 문장의 어휘열($L(f)$)과 소스 문장의 구문 구조($S(f)$)가 주어졌을 때 어떠한 타겟 문장의 구문 구조($S(e)$)가 나타날 확률을 의미한다. 어휘 변환 모델은 소스 문장의 어휘열($L(f)$)과 구문 구조($S(f)$) 및 타겟 문장의 구문 구조($S(e)$)가 주어졌을 때 특정 타겟 문장의 어휘열($L(e)$)이 나타날 확률을 의미한다. 이때, 소스 문장의 구문 구조($S(f)$)는 1-best 결과를 추출하는 분석기(parser)를 사용하면 유일하게 결정될 수 있다.

[58] 이의 구문 구조 변환 모델을 구체적으로 전개하면 다음과 같다.

[59] 수학식 4

$$\begin{aligned}
 & p(S(e) | S(f), L(f)) \\
 &= \prod_i p(s_i(e_i) | S(f), L(f)) \\
 &= \prod_i p(s_i(e) | s_1(f), s_2(f), \dots, s_J(f), L(f)) \\
 &= \prod_i p(d_i | L(f)) \\
 &= \prod_i p(d_i | l_i(f))
 \end{aligned}$$

[60] 위 수학식 4에서 s_i 는 S (구문 구조)의 i 번째 서브 트리를 의미하며, i 에 의해 소스 구문 구조의 서브 트리과 타겟 구문 구조의 서브 트리가 대응될 수 있다. L_i 는 i 번째 서브 트리에 해당하는 문장의 부분 문자열을 의미한다. d_i 는 $s_i(f)$ 에서 $s_i(e)$ 로의 유도(derivation)를 의미한다. 이때, 수학식 4는 서브 트리간 독립적이며 각 유도 간 및 서브 트리의 외부 어휘열에 대한 독립을 가정한다.

[61] 어휘 변환 모델 부분은 일 예로 아래의 수학식 5, 6과 같이 자질함수 F_1, F_2 에 따라 전개할 수 있다.

[62] 수학식 5

$$\begin{aligned}
 & p(L(e) | S(e), S(f), L(f)) \\
 &= p(L(e) | F_1(S(e), S(f), L(f))) \\
 &= p(L(e) | Pos(e), Dst, L(f))
 \end{aligned}$$

[63] 수학식 6

$$\begin{aligned}
 & p(L(e) | S(e), S(f), L(f)) \\
 &= p(L(e) | F_2(S(e), S(f), L(f))) \\
 &= p(L(e) | Dst, L(f))
 \end{aligned}$$

- [64] 위 수학적식 5에서 F_1 은 타겟 문장의 구문 구조($(S(e))$)와 소스 문장의 구문 구조($(S(f))$) 및 구문 구조간 정렬 정보로부터 타겟 문장의 품사열($Pos(e)$)과, 소스 문장에 속한 단어의 번역 순서(Dst) 제한만을 자질로서 출력하는 함수이다.
- [65] 수학적식 6에서 F_2 는 F_1 보다 완화된 제약 조건으로 소스 문장에 속한 단어의 번역 순서(Dst)만을 자질로 출력하는 함수이다. F_1, F_2 의 예는 도 6에 도시되어 있다. 도 6은 F_1 는 타겟 문장의 품사열 ($Pos(e)$), 소스 문장의 어휘열($L(f)$), 소스 문장에 속한 단어의 타겟 문장에서의 배열순서 (Dst)에 대한 자질을 조건으로 하여 타겟 문장의 품사열 정보에 기반한 어휘열 생성의 일례를 보여주는데 반해, F_2 는 소스 문장의 어휘열($L(f)$)과 소스 문장에 속한 단어의 타겟 문장에서의 배열순서 (Dst) 자질만을 조건으로 하여 타겟 문장의 어휘열이 생성되는 일례를 보여준다.
- [66] 또한, 어휘 변환 모델은 기존 PBSMT에서 사용하는 것처럼 상기 변환 확률 외에도 여러 가지 자질을 사용하는 로그 선형(log-linear) 방식으로 최종 계산할 수 있다.
- [67] 수학적식 7

$$p(L(e) | Dst, L(f))$$

$$= \exp \sum_i^n \lambda_i h_i(L(e), Dst, L(f))$$

- [68] 이렇게 문법 변환 모델을 만들기 위한 본 발명의 구문 구조 변환 지식 추출기(110)는 구체적으로 도 4에 도시한 구성을 포함한다.
- [69] 구문 구조 변환 지식 추출기(110)는 위 수학적식 4에서 사용되는 각각의 유도(derivation) d_i 의 지식과 확률을 추출하기 위한 것이다. 즉, 종래의 구문 구조 변환 지식을 추출하기 위한 기술은 단어 정렬 정보를 기반으로 소스 언어의 파스 트리(phrase tree)와 타겟 언어의 파스 트리의 노드 정렬을 수행한 후 정렬된 노드를 중심으로 변환 지식을 추출하였다. 그런데, 이러한 종래의 방식은 소스 언어, 타겟 언어의 모두에 대해 모두 파서(parser)를 사용하여야 하며 파서 성능에 매우 종속적인 방법으로 번역 지식으로의 오류 전파 문제가 발생할 수 있다.
- [70] 그러나, 본 발명의 구문 구조 변환 지식 추출기(110)는 도 4에서 도시한 것처럼 구문 구조 변환 지식을 추출하기 위해 소스 말뭉치만 구문 분석을 수행하며, 문장 선별기(111)에서 구문 분석된 소스 말뭉치와 단어 정렬 정보를 선별하고 이를 소스 트리 생성기(113)의 병합(Merge), 삭제(Deletion), 분리/첨부(Detach/Attach), 순서변경(Reordering), 삽입(Insertion), 범위(Spanning)의 6가지 연산을 거침으로써 타겟 트리를 생성한다. 그리고, 트리 노드 정렬기(115)는 소스 트리 생성기(113)에서 생성한 소스 트리과 타겟 트리를

중심으로 노드 정렬을 파악한 후 트리 변환 지식 추출기(117)를 통해 각 노드에서 서브 트리의 변환 지식을 추출한다. 이러한 추출을 모든 문장에 수행한다. 그러면, 확률 계산기(119)는 추출된 지식에 대해 변환 확률 및 생성 확률을 통계적으로 계산한다.

- [71] 다시 도 2를 참조하면, 번역 모델 구축기(100)에서의 어휘 변환 지식 추출기(130)는 위의 수학적식 5 및 6을 통해 수행될 수 있다. 여기서, 구(*phrase*) 번역은 단어의 번역 순서(*Dst*)와 독립적 임을 가정하고 문장의 위치를 고려한 번역 확률은 별도로 계산하지 않는다. 단, 수학적식 5에서 나타난 품사열(*Pos(e)*)을 제약하기 위해 각 어휘 변환 지식은 타겟 어휘열의 품사열이 별도로 추출되어야 하고 확률도 별도로 계산되어야 한다.
- [72] 이에 따라, 본 발명에서의 어휘 변환 지식 추출기(130)는 기존의 구 기반 통계적 기계 번역(*PBSMT*)에서 사용하는 방법이면 어느 것이나 가능하며 일반적으로 위의 수학적식 7에서 사용될 수 있는 변환 확률, 언어 모델 확률 등이 추가적으로 이 단계에서 계산될 수 있다.
- [73] 언어 모델 생성기(150)는 *SRILM* 언어 모델링 툴킷(*SRILM language modeling toolkit*), *IRST* 언어 모델링 툴킷(*IRST language modeling toolkit*) 등의 외부 모듈을 사용하여 구축할 수 있으며, *N*-그램(*n-gram*) 언어 모델 결과이면 어느 것이나 사용 가능하다.
- [74] 번역 모델 학습기(300)는 위 수학적식 7에서의 각 자질에 대한 가중치를 학습하여, 구문 구조 변환 모델과 어휘 변환 모델을 생성한다.
- [75] 번역문 생성기(400)의 구문 구조 변환기(410)는 입력 문장을 구문 분석하고 구문 구조 변환 지식 추출기(110)에서 학습한 구문 구조 변환 모델을 참조하여 점수가 높은 최대 *N*개의 타겟 트리를 생성하는 모듈이다.
- [76] 이 구문 구조 변환기(410)는 입력된 소스 문장의 구문 분석 결과에 대해 가장 높은 확률로 생성된 소스 트리를 하위 노드부터 순회하면서 적용 가능한 규칙을 저장한 후, 점수(*score*)가 높은 *N*개의 유도 집합(*derivation set*)을 찾기 위해 빔 서치(*Beam Search*)를 수행한다. 이때 점수는 변환 확률과 생성 확률의 곱으로 계산되며 상호 규칙 간에는 독립을 가정한다.
- [77] 어휘 변환기(420)는 구문 구조 변환기(410)에서 생성된 타겟 문장의 구문 구조에 제약 조건을 두어, 어휘 변환 지식 추출기(130)에서 학습한 어휘 변환 지식을 기반으로 타겟 어휘열을 생성하는 모듈이다.
- [78] 타겟 문장의 구문 구조 제약은 위에서 언급한 수학적식 5와 6에서와 같이 소스 문장의 어휘열/구문 구조, 타겟 문장의 품사열의 제약 조건을 포함한 자질 함수를 통해 수행될 수 있다. 이러한 제약은 수학적식 7에서 번역 옵션(*translation option*) 자질함수, 디스톨션(*distortion*: 순서 재배열) 자질함수로 반영된다. 이러한 자질함수 값은 구 기반 디코딩 시 구 옵션(*phrase option*) 제약과 디스톨션(*distortion*) 제약으로 직접 반영할 수도 있다. 이때, 자질함수 F_1 과 F_2 는 시스템의 옵션으로 선택할 수 있으며 구 테이블(*phrase table*)의 크기에 따라

사용자가 선택할 수 있도록 한다.

[79] 특히, 본 발명에 따른 어휘 변환기(420)의 디코딩 과정은 도 5와 같다.

[80] 여기서 종래 기술과 다른 구성은 자질 추출기(421), 번역 옵션 제약기(423), 번역 디스톨션 제약기(425), M-Best 추적기(427)로 각각을 설명하면 다음과 같다.

[81] 자질 추출기(Feature Extractor: 421)는 어휘 변환기(420)로 입력된 소스 문장의 어휘열, 소스 문장의 구문 구조, 타겟 문장의 구문 구조를 기반으로 자질을 추출한다.

[82] 번역 옵션 제약기(Translation option constraint: 423)는 자질 추출기(421)에 의해 생성된 품사열(Pos(e))과 번역 순서(Dst)로부터 번역 옵션을 명시적으로 제약할 수 있다. 명시적으로 가지치기(pruning)를 하지 않는 경우에는 기존의 방식대로 생성된 번역 옵션에 매우 낮은 확률값을 부여한다. 여기서의 번역 옵션은 어휘 번역의 단위가 될 수 있는 가능한 모든 구(phrase)들을 의미한다.

[83] 번역 디스톨션 제약기(distortion constraint: 425)는 종래 기술에 따르면 정해진 순서 없이 랜덤하게 번역 순서를 정하였으나, 본 발명에서는 자질 추출기(421)를 통해 얻은 어순 제약을 통해 명시적으로 번역 순서를 정하거나 또는 확률적으로 번역 순서를 정하여 제약한다.

[84] 예를 들어, 번역 옵션을 랜덤하게 선택하는 것이 아니라 타겟 오더에 의해 순서를 정하도록 명시적 제약을 두거나 또는 타겟 위치에 위반하는 순서의 경우 매우 낮은 확률을 갖도록 확률적 제약을 둘 수 있다.

[85] M-Best 추적기(427)는 가설 공간 탐색(424) 후 가장 점수가 높은 M개의 가설을 선택하기 위한 모듈로서, 탐색이 끝난 지점에서 역으로 추적하여 N개의 어휘열(derivation set)을 출력한다. 기존에는 최종 1개의 가설만을 선택하였으나, 수학적 3에서 볼 수 있듯이 동일한 e 를 생성하는 모든 가능한 경우에 대해 확률의 결합이 필요하므로 본 발명에서는 M개의 가설이 선택된다.

[86] 마지막으로 N개의 구문 구조에서 생성된 각각의 M개의 가설에 대해 동일한 문장 e 를 생성하는 것에 대해 확률을 결합한 후 가장 확률이 높은 타겟 어휘열을 번역문으로 선택하여 출력한다.

[87] 도 7은 본 발명의 실시 예에 따른 기계 번역 방법을 전체적으로 설명하는 흐름도이다.

[88] 먼저 소스 문장이 입력되면, 번역문 생성기의 구문 구조 변환기는 실시간으로 입력되는 소스 문장에 대하여 구문 분석하여 타겟 문장에 대한 구문 구조 변환 지식과 구문 구조 변환 확률을 추출한다(S100, S110). 타겟 문장에 대한 구문 구조 변환 지식과 구문 구조 변환 확률은 사전에 학습시킨 번역 모델 즉, 구문 구조 변환 모델을 통해 추출이 가능하다.

[89] 이후, 추출한 구문 구조 변환 지식으로부터 타겟 어휘열을 생성한다(S120). 타겟 어휘열을 사전에 학습시킨 번역 모델 즉, 어휘 변환 모델을 통해 추출이 가능하다.

[90] 이를 테면, 번역문 생성기의 구문 구조 변환기로부터 입력되는 소스 문장의

단어 정렬 정보, 구문 분석 정보를 기반으로 타겟 문장의 자질(feature)을 추출하고, 자질에 제약할 번역 옵션을 설정하며, 설정한 번역 옵션을 자질에 반영하여 가설을 탐색한다. 탐색 후 가설에 대하여 동일한 타겟 어휘열을 생성하는 확률을 계산하여 확률이 높은 타겟 어휘열을 선택하는 방식을 이용한다. 여기서, 번역 옵션은 품사열, 번역 어순 등의 조건을 포함한다.

[91] 이후, 구문 구조 변환 확률과 타겟 어휘열의 생성 확률을 결합하여 확률이 높은 타겟 어휘열을 번역문으로 생성한다(S130, S140).

[92] 도 8은 본 발명의 실시 예에 따른 기계 번역을 위해 번역 모델을 구축하는 방법을 설명하는 흐름도이다.

[93] 먼저, 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 타겟 문장에 대한 구문 구조 변환 지식을 추출한다(S200). 구문 구조 변환 지식을 추출하는 방법으로는 앞서 설명한 바와 같이 소스 문장의 단어 정렬 정보와 타겟 문장의 단어 정렬 정보를 이용하여 타겟 문장의 구문 트리를 생성하고, 생성한 구문 트리와 소스 문장의 구문 분석에 따른 소스 트리를 중심으로 노드를 정렬한 다음 각 노드에서 서브 트리의 변환 지식을 추출하는 과정을 포함한다.

[94] 이후, 병렬 말뭉치로부터 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 어휘 변환 지식을 추출한다(S210).

[95] 이후, 구문 구조 변환 지식 및 어휘 변환 지식에 대한 변환 확률을 각각 계산하고 각 변환 확률에 대하여 가중치를 학습시킴으로써 구문 구조 변환 모델과 어휘 변환 모델을 각각 형성한다(S220, S230).

[96] 한편, 본 발명은 이상에서 설명한 단어 정렬 방법을 소프트웨어적인 프로그램으로 구현하여 컴퓨터로 읽을 수 있는 소정 기록 매체에 기록해 둬으로써 다양한 재생 장치에 적용할 수 있다.

[97] 다양한 재생 장치는 PC, 노트북, 휴대용 단말 등일 수 있다.

[98] 예컨대, 기록 매체는 각 재생 장치의 내장형으로 하드 디스크, 플래시 메모리, RAM, ROM 등이거나, 외장형으로 CD-R, CD-RW와 같은 광디스크, 콤팩트 플래시 카드, 스마트 미디어, 메모리 스틱, 멀티미디어 카드일 수 있다.

[99] 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

[100] 종래의 구 기반 또는 구문 구조 기반의 통계적 기계 번역은 원거리 어순 변경에 실패하여 전혀 다른 문장이 되거나 또는 타겟 문장의 어순은 맞지만 번역 지식

부족으로 단순한 단어 대 단어 번역으로 이루어지는 문제점이 있으나, 본 발명은 소스 언어 및 타겟 언어의 병렬 말뭉치로부터 구문 구조 변환 및 어휘 변환의 두 모델링을 수행하고 이를 기반으로 디코딩함으로써 구 기반의 통계적 기계 번역과 구문 구조 기반의 통계적 기계 번역의 단점은 해소하고 장점을 결합할 수 있으며, 이로 인해 문맥에 맞는 적절한 번역이 수행되어 번역문의 적절성 및 자연스러움을 향상시킬 수 있다.

청구범위

- [청구항 1] 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보 및 구문분석 정보를 이용하여 타겟 문장의 구문 구조 변환 지식 및 어휘 변환 지식을 추출하고 추출된 각각의 지식에 대하여 변환 확률을 계산하는 번역 모델 구축기;
상기 번역 모델 구축기를 통해 추출한 각각의 번역 지식과 변환 확률을 학습하여 구문 구조 변환 모델과 어휘 변환 모델을 생성하는 번역 모델 학습기; 및
실시간으로 입력되는 소스 문장에 대하여 상기 번역 모델 학습기를 통해 학습시킨 구문 구조 변환 모델과 어휘 변환 모델을 적용하여 상기 소스 문장을 타겟 문장으로 디코딩하는 번역문 생성기
를 포함하는 것을 특징으로 하는 통계적 기계 번역 장치.
- [청구항 2] 제 1 항에 있어서,
상기 번역 모델 구축기는,
상기 타겟 문장의 단일 말뭉치로부터 타겟 어휘열의 생성 확률을 추출하여 언어 모델을 생성하는 언어 모델 생성기
를 더 포함하는 것을 특징으로 하는 통계적 기계 번역 장치.
- [청구항 3] 제 1 항에 있어서,
상기 번역문 생성기는
상기 구문 구조 변환 과정과 어휘 변환 과정에 의해 출력된 복수의 구문 구조 변환 확률과 어휘 변환 확률을 결합하여 확률이 높은 타겟 어휘열을 최종 번역문으로 생성하는 것을 특징으로 하는 통계적 기계 번역 장치.
- [청구항 4] 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 타겟 문장에 대한 구문 구조 변환 지식을 추출하고 추출된 지식에 대해 변환 확률을 계산하는 구문 구조 변환 지식 추출기와,
상기 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 어휘 변환 지식을 추출하고 추출된 지식에 대해 변환 확률을 계산하는 어휘 변환 지식 추출기
를 포함하는 것을 특징으로 하는 번역 모델 구축 장치.
- [청구항 5] 제 4 항에 있어서,
상기 구문 구조 변환 추출기는,
상기 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장의 단어 정렬 정보를 이용하여 타겟 문장의 구문 트리(타겟 트리)를 생성하는

트리 생성기;

상기 트리 생성기에서 생성된 타겟 트리와 상기 소스 문장의 구문 분석 정보에 따른 소스 트리를 중심으로 노드를 정렬하는 트리 노드 정렬기;

각 노드에서 서브 트리의 변환 지식을 추출하는 트리 변환 지식 추출기;

상기 트리 변환 지식 추출기에서 추출된 지식에 대하여 변환 확률을 통계적으로 계산하는 확률 계산기

를 포함하는 것을 특징으로 하는 번역 모델 구축 장치.

[청구항 6]

제 4 항에 있어서,

상기 어휘 변환 추출기는

상기 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보에 특정 제약 조건을 정의해 둔 자질 함수를

적용하여 상기 어휘 변환 확률을 추출하는 것을 특징으로 하는 번역 모델 구축 장치.

[청구항 7]

제 6 항에 있어서,

상기 자질 함수는

상기 타겟 문장의 구문 구조와 소스 문장의 구문 구조 및 상호 구문 구조간 정렬 정보로부터 타겟 문장의 품사열과, 소스 문장에 속한 단어의 번역 순서를 제약하여 이를 자질로서 출력하는 함수인 것을 특징으로 하는 번역 모델 구축 장치.

[청구항 8]

제 6 항에 있어서,

상기 자질 함수는

상기 타겟 문장의 구문 구조와 소스 문장의 구문 구조 및 상호 구문 구조간 정렬 정보로부터 소스 문장에 속한 단어의 번역 순서만을 자질로 출력하는 함수인 것을 특징으로 하는 번역 모델 구축 장치.

[청구항 9]

실시간으로 입력되는 소스 문장에 대하여 구문 분석하고 분석한 소스 문장의 구문 구조로부터 타겟 문장의 구문 구조 변환 지식을 추출하고 추출된 지식에 대해 변환 확률을 학습시킨 구문 구조 변환기;

상기 구문 구조 변환기를 통해 추출된 타겟 문장의 구문 구조에 제약 조건을 둔 어휘 변환 모델을 기반으로 타겟 어휘열을 생성하는 어휘 변환기;

상기 구문 구조 변환기를 통해 학습시킨 변환 확률에서 상기 어휘 변환기를 통해 생성한 타겟 어휘열의 생성 확률을 결합한 후 가장 확률이 높은 타겟 어휘열을 번역문으로 생성하는 확률 계산기를 포함하는 것을 특징으로 하는 번역문 생성장치.

[청구항 10]

제 9 항에 있어서,

상기 어휘 변환기는,
 상기 구문 구조 변환기로부터 입력되는 소스 문장의 단어 정렬 정보, 구문 분석 정보, 타겟 문장의 구문 분석 정보를 기반으로 자질(Feature)을 추출하는 자질 추출기;
 상기 자질 추출기에 의해 추출된 품사열과 번역 순서로부터 번역 옵션을 제약하는 번역 옵션 제약기;
 상기 자질 추출기에 의해 추출된 어순 제약을 통해 번역 순서를 재배열하는 번역 디스톨션 제약기;
 상기 자질 추출기를 통해 추출된 자질에 상기 번역 옵션 제약기 및 상기 번역 디스톨션 제약기를 통해 정한 제약 조건을 반영하여 가설을 탐색하는 가설 탐색기;
 상기 가설 탐색기에서 탐색한 가설에 대하여 동일한 타겟 어휘열을 생성하는 것에 대한 확률을 계산하여 상기 확률이 높은 타겟 어휘열을 선택하는 M-Best 추적기
 를 포함하는 것을 특징으로 하는 번역문 생성장치.

[청구항 11]

(a) 실시간으로 입력되는 소스 문장에 대하여 구문 분석하고 분석한 소스 문장의 구문 구조로부터 타겟 문장의 구문 구조 변환 지식 및 변환 확률을 추출하는 단계;
 (b) 상기 타겟 문장의 구문 구조 변환 지식으로부터 추출한 타겟 문장의 구문 구조에 제약 조건을 둔 어휘 변환 모델을 기반으로 타겟 어휘열을 생성하는 단계;
 (c) 상기 타겟 문장의 구문 구조 변환 확률과 상기 타겟 어휘열의 생성 확률을 결합하여 확률이 높은 타겟 어휘열을 번역문으로 생성하는 단계
 를 포함하는 것을 특징으로 하는 기계 번역 방법.

[청구항 12]

제 11 항에 있어서,
 상기 (b) 단계는,
 상기 구문 구조 변환기로부터 입력되는 소스 문장의 단어 정렬 정보, 구문 분석 정보, 타겟 문장의 구문 분석 정보를 기반으로 자질(Feature)을 추출하는 단계;
 상기 자질에 제약할 번역 옵션을 설정하는 단계;
 상기 자질에 상기 번역 옵션을 반영하여 가설을 탐색하는 단계;
 탐색된 가설에 대하여 동일한 타겟 어휘열을 생성하는 확률을 계산하여 확률이 높은 타겟 어휘열을 선택하는 단계
 를 포함하는 것을 특징으로 하는 기계 번역 방법.

[청구항 13]

제 11 항에 있어서,
 상기 번역 옵션은 품사열, 번역 어순 중 어느 하나 이상을 선택적으로 포함하는 것을 특징으로 하는 기계 번역 방법.

[청구항 14]

(a) 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 타겟 문장에 대한 구문 구조 변환 지식을 추출하는 단계;
 (b) 상기 복수의 병렬 말뭉치에서 소스 문장과 타겟 문장 사이의 단어정렬 정보와 소스 문장의 구문분석 정보를 이용하여 어휘 변환 지식을 추출하는 단계;
 (c) 상기 구문 구조 변환 지식 및 어휘 변환 지식에 대한 변환 확률을 각각 계산하고 각 변환 확률에 대하여 가중치를 학습시키는 단계
 를 포함하여 번역 모델을 구축한 것을 특징으로 하는 번역 모델 구축 방법.

[청구항 15]

제 14 항에 있어서,
 상기 (a) 단계는,
 상기 복수의 병렬 말뭉치에서 소스 문장의 단어 정렬 정보와 타겟 문장의 단어 정렬 정보를 이용하여 타겟 문장의 구문 트리를 생성하는 단계;
 상기 타겟 문장의 구문 트리와 상기 소스 문장의 구문 분석 정보에 따른 소스 트리를 중심으로 노드를 정렬하는 단계;
 각 노드에서 서브 트리의 변환 지식을 추출하는 단계
 를 포함하는 것을 특징으로 하는 번역 모델 구축 방법.

[청구항 16]

제 14 항에 있어서,
 상기 (c) 단계에서 어휘 변환 확률을 계산하는 단계는,
 상기 소스 문장의 단어 정렬 정보와 타겟 문장의 단어 정렬 정보 및 타겟 문장의 구문 분석 정보에 특정 제약 조건을 정의해 둔 자질 함수를 적용하여 추출하는 것을 특징으로 하는 번역 모델 구축 방법.

[청구항 17]

제 16 항에 있어서,
 상기 자질 함수는
 상기 타겟 문장의 구문 구조와 소스 문장의 구문 구조 및 상호 구문 구조간 정렬 정보로부터 타겟 문장의 품사열과, 소스 문장에 속한 단어의 번역 순서를 제약하여 이를 자질로서 출력하는 함수를 이용한 것을 특징으로 하는 번역 모델 구축 방법.

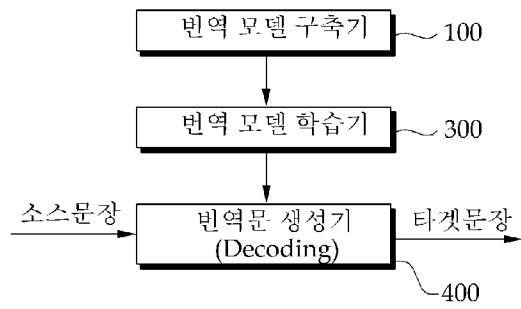
[청구항 18]

제 16 항에 있어서,
 상기 자질 함수는
 상기 타겟 문장의 구문 구조와 소스 문장의 구문 구조 및 상호 구문 구조간 정렬 정보로부터 소스 문장에 속한 단어의 번역 순서만을 자질로 출력하는 함수를 이용한 것을 특징으로 하는 번역 모델 구축 방법.

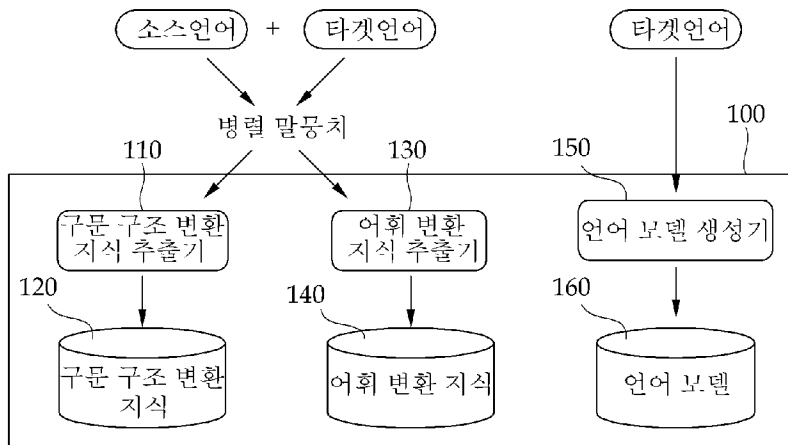
[청구항 19]

제 11 항 내지 제 18 항 중 어느 한 항에 의한 과정을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

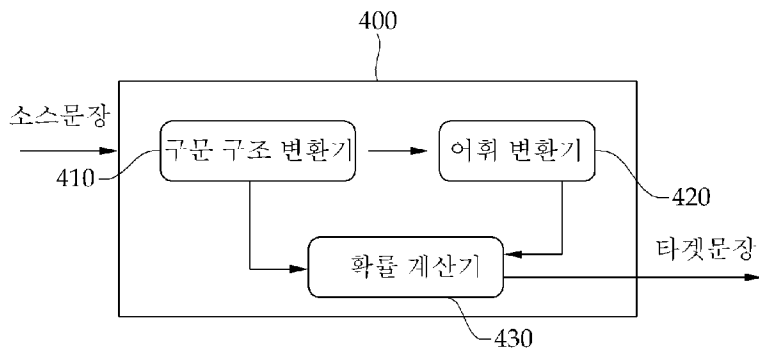
[Fig. 1]



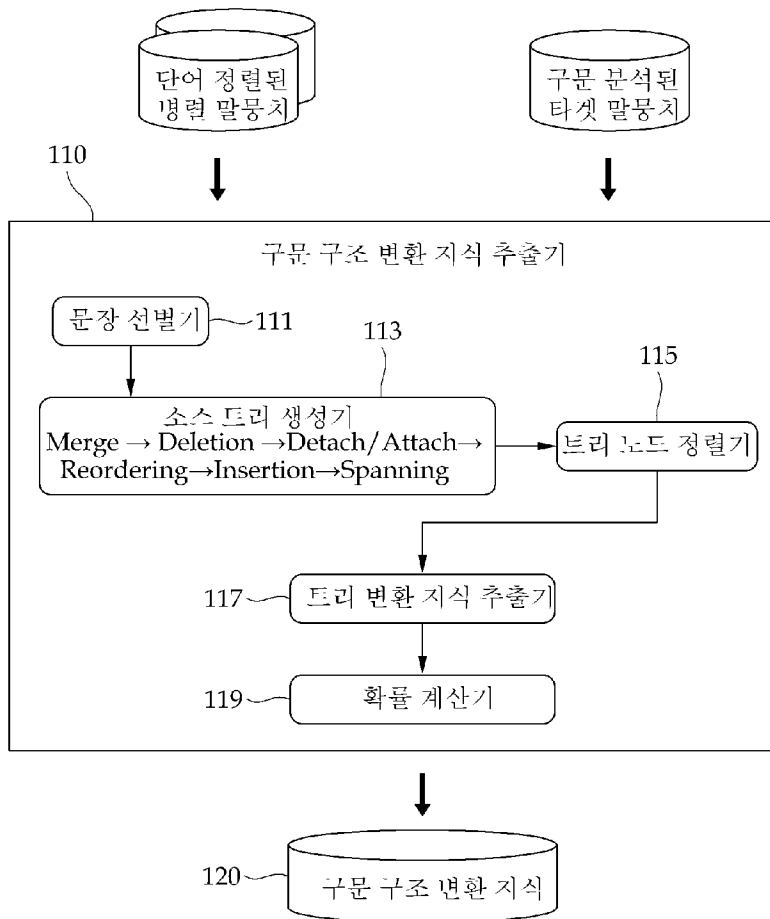
[Fig. 2]



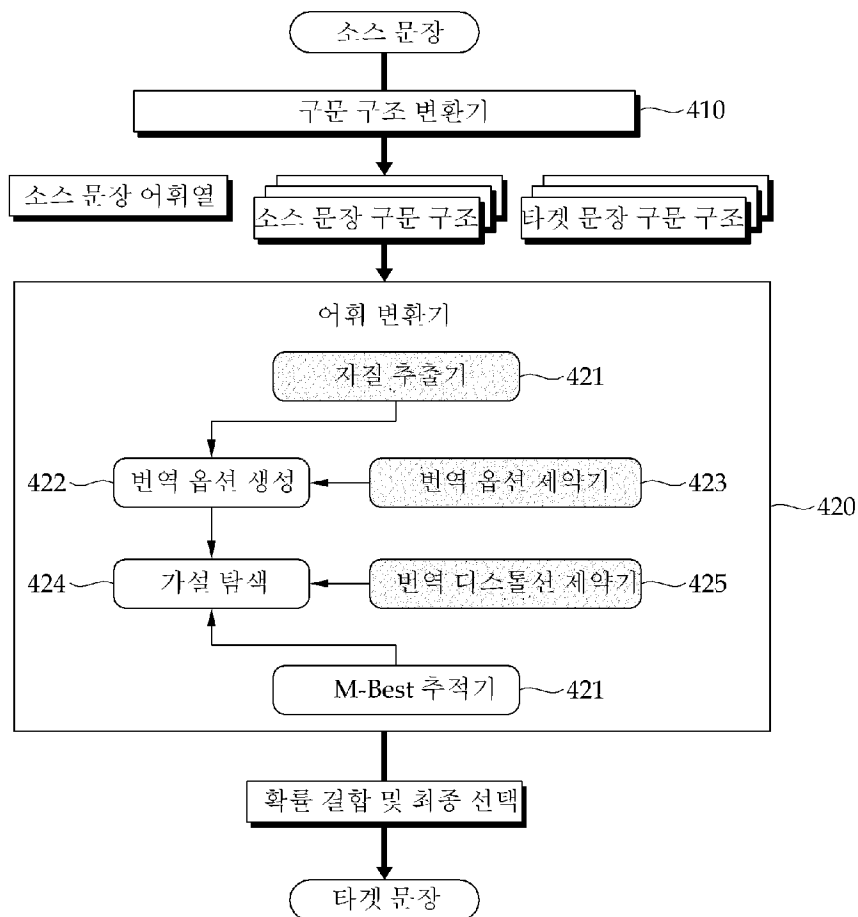
[Fig. 3]



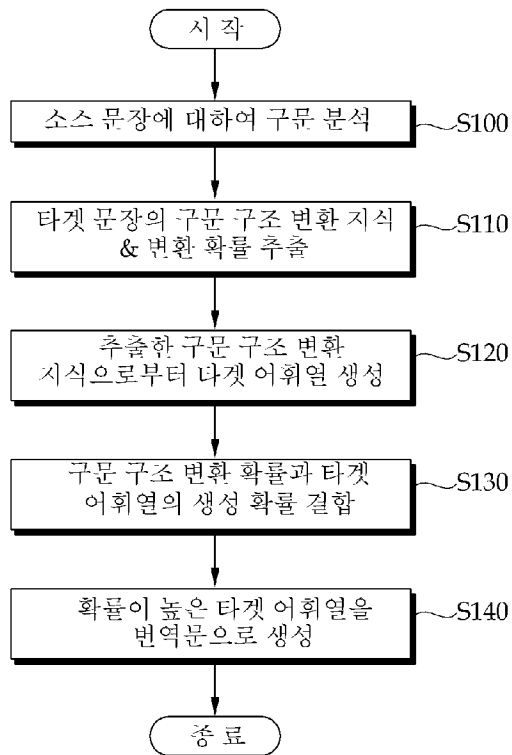
[Fig. 4]



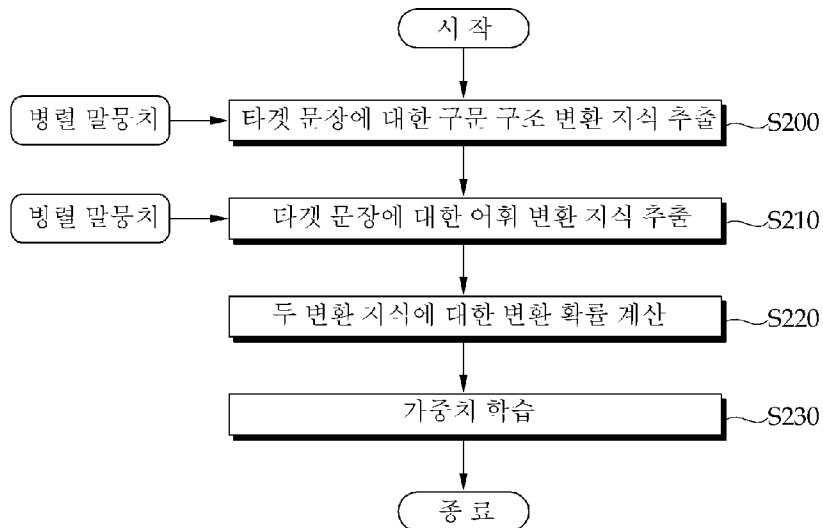
[Fig. 5]



[Fig. 7]



[Fig. 8]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2011/005325**A. CLASSIFICATION OF SUBJECT MATTER****G06F 17/28(2006.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 17/28

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean Utility models and applications for Utility models: IPC as above

Japanese Utility models and applications for Utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: corpus, translation, statistics, conversion

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Yeong-Suk Hwang, Introduction of Statistics-based Machine Translation Technology and Research Trends, Journal of Institute of Language and Information Studies, Vol.25, pages 89-114, 30 April 2010	1-5,9,11,13-15,19
A	See page 91, line 6 - page 110, line 18.	6-8,10,12,16-18
A	S.H. Kim et al., Electronics and Telecommunications Trends of ETRI, Vol.25 No.3 pages 28-39, 30 June 2010	1-19
	See page 29, line 3 - page 34, line 17.	
A	Kim, Un-Oe et al., Electronics and Telecommunications Trends of ETRI, Vol.23 No.1, pages 89 - 97, 28 February 2008	1-19
	See page 90, line 2 - page 92, line 1.	
A	KR 10-2005-0064574 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 29 June 2005	1-19
	See page 3, line 3 - page 5, line 2.	



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

16 FEBRUARY 2012 (16.02.2012)

Date of mailing of the international search report

20 FEBRUARY 2012 (20.02.2012)

Name and mailing address of the ISA/KR



Korean Intellectual Property Office
Government Complex-Daejeon, 139 Seonsa-ro, Daejeon 302-701,
Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2011/005325

Patent document cited in search report	Publication date	Patent family member	Publication date
KR 10-2005-0064574 A	29.06.2005	NONE	

A. 발명이 속하는 기술분류(국제특허분류(IPC)) G06F 17/28(2006.01)i		
B. 조사된 분야 조사된 최소문헌(국제특허분류를 기재) G06F 17/28 조사된 기술분야에 속하는 최소문헌 이외의 문헌 한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우)) eKOMPASS(특허청 내부 검색시스템) & 키워드: 말뭉치, 번역, 통계, 변환		
C. 관련 문헌		
카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
X	황영숙, 통계기반 기계번역 기술의 소개 및 최신 연구동향 분석, 연세대학교 언어정보연구원 정기간행물, 제25권, pages 89-114, 2010.04.30.	1-5, 9, 11, 13-15, 19
A	페이지 91, 라인 6 - 페이지 110, 라인 18 참조.	6-8, 10, 12, 16-18
A	김승희외, 한국전자통신연구원 전자통신동향분석 제25권 제3호 pages 28-39, 2010.06.30. 페이지 29, 라인 3 - 페이지 34, 라인 17 참조.	1-19
A	김운외, 한국전자통신연구원 전자통신동향분석 제23권 제1호, pages 89 - 97, 2008.02.28. 페이지 90, 라인 2 - 페이지 92, 라인 1 참조.	1-19
A	KR 10-2005-0064574 A (한국전자통신연구원) 2005.06.29. 페이지 3, 라인 3 - 페이지 5, 라인 2 참조.	1-19
☐ 추가 문헌이 C(계속)에 기재되어 있습니다. ☒ 대응특허에 관한 별지를 참조하십시오.		
* 인용된 문헌의 특별 카테고리: “A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌 “E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌 “L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌 “O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌 “P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌 “T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌 “X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신구성 또는 진보성이 없는 것으로 본다. “Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다. “&” 동일한 대응특허문헌에 속하는 문헌		
국제조사의 실제 완료일 2012년 02월 16일 (16.02.2012)		국제조사보고서 발송일 2012년 02월 20일 (20.02.2012)
ISA/KR의 명칭 및 우편주소  대한민국 특허청 (302-701) 대전광역시 서구 청사로 189, 정부대전청사 팩스 번호 82-42-472-7140		심사관 박상현 전화번호 82-42-481-8263 

국제조사보고서에서
인용된 특허문헌

공개일

대응특허문헌

공개일

KR 10-2005-0064574 A

2005.06.29

없음