(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2012 년 1 월 19 일 (19.01.2012)

PCT

(10) 국제공개번호
WO 2012/008684 A2

(51) 국제특허분류:
G06F 17/28 (2006.01)

(21) 국제출원번호: PCT/KR2011/003977

(22) 국제출원일: 2011 년 5 월 31 일 (31.05.2011)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2010-0067635 2010 년 7 월 13 일 (13.07.2010) KR

(71) 출원인 (US 을(를) 제외한 모든 지정국에 대하여): **에스케이텔레콤 주식회사 (SK TELECOM. CO., LTD)** [KR/KR]; 서울특별시 중구 을지로 2 가 11 번지, 100-999 Seoul (KR).

(72) 발명자; 겸
(75) 발명자/출원인 (US 에 한하여): **황영숙 (HWANG, Young Sook)** [KR/KR]; 서울특별시 성북구 돈암 1 동 풍림아파트 105-502, 136-061 Seoul (KR). **김상범 (KIM, Sang-Bum)** [KR/KR]; 서울특별시 노원구 상계 1 동 현대 2 차아파트 202 동 1207 호, 139-893 Seoul (KR). **윤창호 (YIN, Chang Hao)** [CN/KR]; 서울특별시 성북구 정릉 2 동 푸른동아 아파트 107 동 1303 호, 136-778 Seoul (KR). **왕지양 (WANG, Zhiyang)** [CN/CN]; 북경시 케슈유안 싸우스 로드 6 번, 하이디안 디스트릭트, 사서함 2704 번, 10190 Beijing (CN). **리우쥔 (LIU, Qun)** [CN/CN]; 북경시 케슈유안 싸우스 로드 6 번, 하이디안 디스트릭트, 사서함 2704 번, 10190 Beijing (CN). **뤼야쥬안 (LV, Yajuan)** [CN/CN]; 북경시 케슈유안 싸우스 로드 6 번, 하이디안 디스트릭트, 사서함 2704 번, 10190 Beijing (CN).

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

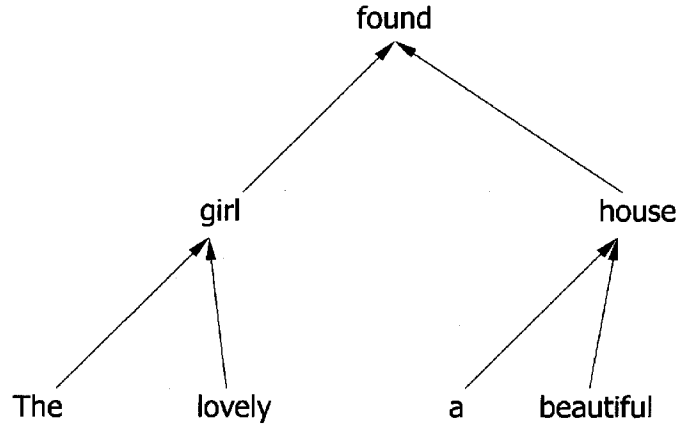
(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

[다음 쪽 계속]

(54) **Title**: METHOD AND DEVICE FOR FILTERING A TRANSLATION RULE AND GENERATING A TARGET WORD IN HIERARCHICAL-PHASE-BASED STATISTICAL MACHINE TRANSLATION

(54) **발명의 명칭** : 계층적 구문 기반의 통계적 기계 번역에서의 번역규칙 필터링과 목적단어 생성을 위한 방법 및 장치

[Fig.1]



(57) **Abstract**: The present invention relates to the field of statistical machine translation, and more particularly, to a method and device for filtering a translation rule and generating a purpose word in hierarchical-phase-based statistical machine translation, which involve filtering a translation rule using a loose well-formed dependency structure, and generating a target word with reference to a head word of a source word, such that the number of translation rules is reduced and translation performance is improved as compared to that of a hierarchical-phase-based original translation rule table.

(57) **요약서**:

[다음 쪽 계속]

WO 2012/008684 A2

MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서 없이 공개하며 보고서 접수 후 이를 별도 공개함 (규칙 48.2(g))

본 발명은 통계적 기계 번역 분야에 관한 것으로, 상세하게는 계층적 구문 기반의 통계적 기계 번역에서 느슨한 적격 의존 구조를 이용하여 번역규칙을 필터링하고 원시단어의 헤드 단어를 참조하여 목적단어를 생성함으로써 계층적 구문 기반의 원 번역규칙 테이블과 비교하여 번역규칙의 개수를 줄이면서도 번역 성능을 개선시킬 수 있는 계층적 구문 기반의 통계적 기계 번역에서의 번역규칙 필터링과 목적단어 생성을 위한 방법 및 장치에 관한 것이다.

명세서

발명의 명칭: 계층적 구문 기반의 통계적 기계 번역에서의 번역규칙 필터링과 목적단어 생성을 위한 방법 및 장치

기술분야

[1] 본 발명은 통계적 기계 번역 분야에 관한 것으로, 상세하게는 계층적 구문 기반의 통계적 기계 번역에서 느슨한 적격 의존 구조를 이용하여 번역규칙을 필터링하고 원시단어의 헤드 단어를 참조하여 목적단어를 생성함으로써 계층적 구문 기반의 원 번역규칙 테이블과 비교하여 번역규칙의 개수를 줄이면서도 번역 성능을 개선시킬 수 있는 계층적 구문 기반의 통계적 기계 번역에서의 번역규칙 필터링과 목적단어 생성을 위한 방법 및 장치에 관한 것이다.

배경기술

[2] 지난 수십 년간 기계 번역 기술 분야에서 데이터 구동 방식이 매우 성공적으로 사용되어 왔다. 연산 능력을 높이고 대용량의 말뭉치(corpus)를 사용할 수 있도록 통계적 기계 번역(Statistical Machine Translation: SMT) 분야에서 많은 연구가 진행되었다. 최근의 방법은 번역모델(translation model)에 대해서 계층적 구조(hierarchical structure)를 사용하고 있다.

[3] 계층적 구문 기반(Hierarchical Phrase-Based: HPB) 모델을 예로 들어 설명한다. 계층적 방식은 여러 구문을 포함하는 구문을 찾아 서브구문(Sub-phrase)을 비 단말 부호(non-terminal symbol)로 대체한다. 여기서 비 단말 부호란 문법상 문장에 나올 수 없는 단어로서 정상적인 문법에서의 단어를 의미한다. 계층적 방식은 생성 능력이 좋고 원거리 어순 재배열(long distance reordering)이 가능하기 때문에 종래 구문기반 방식보다 강력하다. 그러나 학습용 말뭉치(training corpus)가 커지면 번역규칙(translation rule)의 개수가 급격하게 증가하여 디코딩 속도가 느려지고 메모리 용량이 증가한다. 따라서 실제 대규모 번역 작업에는 적합하지 않다.

[4] 과거부터 계층적 번역규칙 테이블(hierarchical translation rule table)을 감소시키기 위해 상당히 많은 기술이 제안되어 왔다. 일부 개발자는 원시언어 측(source side) 핵심 어구(key phrase)를 사용하여 언어학적 정보의 활용 없이 번역규칙 테이블을 필터링 했다. 어떤 개발자는 패턴 및 비 단말의 개수에 근거하여 구문 등급(syntactic class)에 번역규칙을 추가하고 여러 필터링 방식을 응용하여 번역규칙 테이블의 품질을 개선하였다.

[5] 의존 정보를 이용하는 기술은 목적언어 측(target side)의 번역규칙이 적격 의존 구조(well-formed dependency structure)이어야 한다는 제약에 따라 번역규칙 테이블의 많은 번역규칙을 제거했지만, 이러한 필터링 방식은 번역 성능을 저하시켰다. 이를 위해 종래 의존 정보를 이용하는 기술은 의존 언어 모델을 새로이 추가함으로써 성능을 향상시켰다.

[6] 통계적 기계 번역 시스템에서 번역규칙은 필수적이다. 일반적으로 양호한 규칙(good rule)의 개수가 늘어날수록 번역 성능은 좋아진다. 그러나 상술한 바와 같이 학습용 말뭉치(training corpus)가 커지면 규칙의 개수가 급격하게 증가하여 디코딩 속도가 느려지고 메모리 용량이 증가한다.

[7] SMT 분야에서 모든 번역규칙은 말뭉치로부터 자동적으로 학습된다. 그러나 모든 번역규칙이 양호한 것은 아니다. 상술한 것처럼 HPB 모델에서는, 다른 어구를 포함하는 어구를 찾아 부어구를 비 단말 신호로 대체함으로써 계층적 번역규칙을 얻는다. 이러한 번역규칙 생성 방식은 매우 단순하여 많은 번역규칙이 언어학적으로 맞지 않기 때문에 모든 번역규칙이 양호한 것은 아니다.

[8] 또한, 종래에는 제2의 단어를 도입하여 언어학적 정보(linguistic information)의 고려 없이 목적단어(target word)를 생성하였다. 더욱이 제2의 단어는 문장의 어느 부분에서 나올 수 있기 때문에 엄청난 수의 파라미터가 필요할 수 있다. 다른 방법은 최대 엔트로피 모델(maximum entropy model)을 구축하는 것이다. 최대 엔트로피 모델은 디코딩 시 번역 규칙을 선택하기 위한 풍부한 문맥정보(context information)를 결합하고 있다. 그러나 말뭉치의 크기가 커질수록 최대 엔트로피 모델도 증가하는 문제점이 있었다.

발명의 상세한 설명

기술적 과제

[9] 본 발명은 상기의 문제점을 해결하기 위해 창안된 것으로서, 본 발명의 목적은 언어의 의존관계 정보(dependency information of the bilingual languages)에 의존하는 계층적 번역규칙 테이블을 감소시키면서도 번역 성능을 향상시키는 것이다.

[10] 본 발명의 다른 목적은 추가적인 언어 모델을 사용함에 따른 시스템 복잡도를 증가시키지 않으면서 번역 성능을 더욱 향상시키는 것이다.

과제 해결 수단

[11] 이를 위하여, 본 발명의 제1 측면에 따르면, 번역규칙 필터링 방법은 느슨한 적격 의존 구조를 이용하여 원시언어 측과 목적언어 측의 계층적 구문 기반 번역규칙을 감소시키는 것을 특징으로 한다.

[12] 본 발명의 제2 측면에 따르면, 번역규칙 생성 방법은 원시언어 및 목적언어의 문장을 구성하는 단어를 정렬하는 단계와, 상기 정렬된 단어를 매트릭스로 구성하는 단계와, 상기 매트릭스에서 공통의 헤드 단어에 의존되는 단어를 묶어 어구를 생성하는 단계와, 상기 생성된 어구를 이용하여 번역규칙을 생성하는 단계를 포함하는 것을 특징으로 한다.

[13] 본 발명의 제3 측면에 따르면, 목적단어 생성 방법은 대응하는 원시단어뿐만 아니라 원시단어의 문맥적 헤드 단어에 의해 트리거 되는 것을 특징으로 한다.

[14] 본 발명의 제4 측면에 따르면, 계층적 구문기반의 통계적 기계 번역 방법은

원시언어 측과 목적언어 측에 느슨한 적격 의존 구조를 이용하여 계층적 구문기반 번역규칙을 생성하고, 상기 생성된 번역규칙을 이용하고 원시단어의 헤드 단어에 의한 트리거 방식을 적용하여 원시언어 텍스트를 목적언어 텍스트로 번역하는 것을 특징으로 한다.

[15] 본 발명의 제5 측면에 따르면, 번역규칙 생성장치는 원시언어 및 목적언어 문장으로 구성된 언어 쌍 말뭉치를 단어 정렬하는 단어 정렬기와, 상기 언어 쌍 말뭉치를 파싱 하여 느슨한 적격 의존 구조에 따라 의존 트리를 생성하는 단어 분석기와, 상기 단어 정렬된 언어 쌍 말뭉치와 의존 트리를 이용하여 번역규칙을 생성하는 번역규칙 추출기를 포함하는 것을 특징으로 한다.

[16] 본 발명의 제6 측면에 따르면, 디코더는 언어 쌍 말뭉치로부터 느슨한 적격 의존 구조에 의해 생성된 번역규칙과 단일 말뭉치로부터 생성된 언어모델을 이용하여 원시언어 텍스트를 목적언어 텍스트로 변환하는 것을 특징으로 한다.

발명의 효과

[17] 본 발명은 원시언어 측 및 목적언어 측에 모두 느슨한 적격 의존 구조 Relaxed-Well-Formed dependency structure: RWF dependency structure) 방식을 적용하여 이러한 의존 구조를 만족시키지 못하는 번역규칙은 제거함으로써 원 번역규칙 테이블로부터 약 40%의 불필요한 번역규칙을 제거하면서도 번역성능은 종래 HPB 번역 시스템에 비하여 좋아지는 효과가 있다.

[18] 또한 본 발명은 느슨한 적격 의존 구조 방식과 함께 새로운 언어 특성인 헤드 단어 트리거를 적용하여 번역 성능을 더욱 향상시킬 수 있는 효과가 있다. 특히 본 발명에 의한 언어 특성은 중국어-영어 번역 작업에 효과가 있으며 특히 대규모 말뭉치에서 효과적으로 작용한다.

도면의 간단한 설명

[19] 도 1은 의존 트리의 예를 나타낸 도면.

[20] 도 2는 원시단어와 목적단어 간의 관계를 나타낸 도면.

[21] 도 3은 본 발명에 따른 통계적 기계 번역 장치를 나타낸 도면.

발명의 실시를 위한 형태

[22] 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.

[23] 본 발명의 상세한 설명에 앞서, 공지된 구성에 대해서는 본 발명의 요지를 흐릴 수 있다고 판단되는 경우 구체적인 설명은 생략하기로 함에 유의한다.

[24] 본 발명은 원시언어 측 및 목적언어 측이 느슨한 적격 의존 구조 Relaxed-Well-Formed dependency structure: RWF dependency structure) 방식을 사용하여 이러한 의존 구조를 만족시키지 못하는 번역규칙은 제거한다. 이러한 방법을 사용하면 원 번역규칙 테이블로부터 약 40%의 불필요한 번역규칙을 제거하면서도 번역성능은 종래 HPB 번역 시스템보다 더 좋게 된다.

[25] 종래의 적격 의존 구조는 목적언어 측에서만 적용된 반면 본 발명에 의한 느슨한 적격 의존 구조는 원시언어 측과 목적언어 측 양쪽에 모두 적용된다는 점이 다르다.

[26] 이러한 느슨한 적격 의존 구조에 근거하여, 본 발명은 또한 새로운 언어 특성을 도입하여 번역 성능을 향상시킨다. 종래의 구문 기반 SMT 모델에는 IBM 모델 1에 기반한 어휘적 번역확률(lexical translation probability) p(e|f)이 있다. 즉, 목적단어 e는 원시단어 f에 의해 트리거(trigger)된다.

[27] 그러나 직감적으로 목적단어 e의 생성은 원시단어 f에만 관여되는 것이 아니라 때로는 원시언어 측의 다른 문맥 단어에 의해 트리거 될 수 있다. 여기서 단어 f의 의존 에지(dependency edge)(f -> f')가 목적단어 e를 생성한다고 가정한다. 이를 헤드 단어 트리거(head word trigger)라고 부른다.

[28] 따라서 하나의 언어에서 2개의 단어는 다른 단어에 의해 한 단어의 의미를 트리거 시킨다. 이것은 목적단어에 대한 보다 복잡하고 좋은 선택을 제공한다. 이러한 의존관계 특성은 중국어-영어 번역 작업에서 효과가 있으며 특히 대규모 말뭉치에서 효과적으로 작용한다.

[29] 이와 같이 의존 에지를 조건으로 채용하는 본 발명은 문맥 정보를 분석하는 종래의 방식과 전혀 다르다.

[30] 도 1은 의존 트리의 예를 나타낸 것이다. 도 1에서 단어 'found'가 트리의 루트(root)가 된다.

[31] 일부 기계 번역 개발자는 적격 의존 구조(well-formed dependency structure)를 제안하여 계층적 번역규칙 테이블(hierarchical translation rule table)을 필터링 한다. 적격 의존 구조는 하나의 루트를 가진 의존 트리(single-rooted dependency tree)이거나 형제관계의 트리 세트(a set of sibling trees)가 될 수 있다. 목적언어 측은 적격 의존 구조이어야 한다는 제약에 따라 많은 번역규칙들이 버려지기 때문에 번역 성능이 떨어지게 된다.

[32] 본 발명에서는 적격 의존 구조를 확장하여 이른바 느슨한 적격 의존 구조(relaxed well-formed dependency structure)를 제안하여 계층적 번역규칙 테이블을 필터링 한다.

[33] 문장 S=w_1w_2 ... w_n이 있다고 가정한다. 이때 d_1d_2 ... d_n는 S 문장의 각 단어에 대한 부모 단어(parent word) 위치를 나타낸다. 예를 들어, d_3=4라는 것은 w_3이 w_4에 의존한다는 것을 의미한다. 만약 w_i가 루트라면 d_i=-1라고 정의한다.

[34] 형식적으로 의존 구조 w_i ... w_j는 느슨한 적격 의존 구조이다. 여기서 $h \notin [i, j]$ 이고, 모든 단어 w_i ... w_j는 직접 또는 간접적으로 w_h 또는 -1에 의존된다. 여기서 h=-1이다. 만약 아래의 조건을 만족시킨다면, 느슨한 적격 의존 구조는 그 정의로부터 적격 의존 구조를 포함할 수 있다.

[35]

$$d\_h \notin [i, j]$$

[36] $$\forall k \in [i, j], d\_k \in [i, j] \; or \; d\_k = h$$

[37] 느슨한 적격 의존 구조는 헤드 단어(head word)가 아닌 복수의 단어로 이루어진 집합으로 구성될 수 있으며, 복수의 단어는 공통의 헤드 단어에 의존될 수 있다. 헤드 단어는 각 단어의 부모 단어에 해당한다.

[38] 느슨한 적격 의존 구조에서는 서브 루트의 모든 자식 단어(children word)가 완전할 필요는 없다. 도 1의 의존 트리를 참조하여, 적격 구조를 제외하면 "girl found a beautiful house"를추출할 수 있다. 따라서 수식어 "the lovely"가 "the cute"로 바뀌면, 이 규칙이 작동된다.

[39] [표 1]

[40]

System	Rule table size
HPB	30,152,090
RWF	19,610,255
WF	7,742,031

[41] 표 1은 FBIS 상에 여러 제약이 적용된 경우 번역규칙 테이블의 사이즈를 나타낸다. FBIS 말뭉치는 6.9M의 중국어 단어와 8.9M 영어 단어를 가진 239K 문장 쌍을 포함한다.

[42] 표 1에서 HPB는 기본적인 계층적 구문 기반 모델을 나타내고, RWF는 느슨한 적격 의존 구조가 적용된 모델을 나타내고 WF는 적격 의존 구조가 적용된 모델을 나타낸다. 표 1에 표시된 것처럼, 번역규칙 테이블의 사이즈가 HPB, RWF, WF 순으로 작아짐을 알 수 있다.

[43] RWF는 원 번역규칙 테이블의 약 35%를 필터링 하였고, WF는 원 번역규칙 테이블의 74%를 제거했다. RWF는 WF에 비하여 추가적으로 39%를 추출했다. 추가된 번역규칙은 언어학적으로 맞는 것이다.

[44] 로그 선형 모델(log-linear model)에 적용되는 헤드 단어에 의한 트리거의 특성은 트리거 기반 방식(trigger-based approach)에 근거한다.

[45] 종래의 구문 기반 SMT 시스템에서 원시단어 f는 목적단어 e와 정렬되는데, IBM 모델 1에 따르면 어휘적 번역확률(lexical translation probability)은 p(elf)이다. 그러나 의존 관계 측면에서 목적단어 e의 생성은 정렬된 원시단어 f에 의해 트리거 될 뿐만 아니라 f의 헤드단어 f'와 연관된다. 따라서 어휘적 번역확률은 p(elf -> f')가 되며, 목적단어에 대한 더욱 섬세한 어휘적 선택이 가능하다.

[46] 도 2는 원시단어와 목적단어의 관계를 나타낸 것이다. 도 2에서 실선의 화살표는 자식(f)에서 부모(f')로 의존관계를 나타낸다. 목적단어 e는 원시단어 f와 그의 헤드 단어 f'에 의해 트리거 된다. 즉 어휘적 번역확률은 p(elf -> f')이다.

[47] 특별히 번역확률은 최대 가능성 방식(Maximum Likelihood: MLE)에 의해 계산될 수 있다.

[48]
$$p(e \mid f->f') = \frac{count(e, f->f')}{\sum_{e'} count(e', f->f')}$$

[49] 어구 쌍 $\overline{f}$, $\overline{e}$와 단어 정렬 a 및 원시문장 d_1^J 의 의존관계(J는 원시문장의 길이, I는 목적문장의 길이)가 주어진다.

[50] 따라서 어휘적 번역확률 분포 p(elf -> f')가 주어지면, 어구 쌍 $\overline{f}$, $\overline{e}$ 의 특성 값을 아래와 같이 계산한다.

[51]
$$p(\overline{e} \mid \overline{f}, d_1^J, a) = \prod_{i=1}^{I} \frac{1}{|\{j \mid (j,i) \in a\}|} \sum_{\forall (j,i) \in a} p(e_i \mid f_j -> f_{d_j})$$

[52] $p(\overline{e} \mid \overline{f}, d_1^J, a)$ 를 구하면, 유사한 방식으로 $p(\overline{f} \mid \overline{e}, d_1^I, a)$ 를 구할 수 있다.

[53] d_1^J는 목적언어 측의 의존관계를 나타낸다. 이러한 새로운 특성은 어휘적 가중치(lexical weighting)와 같이 로그 선형 모델(log-linear model)에 추가된다.

[54] [표 2]

[55]

System	Dev02	Test04	Test05
HPB	0.3473	0.3386	0.3206
RWF	0.3539	**0.3485****	**0.3228**
RWF+Tri	0.3540	**0.3607****	**0.3339***

[56] 표 2는 GQ 말뭉치의 결과를 나타낸다. GQ는 LDC 말뭉치로부터 수동 선택된 것이다. GQ는 41M 중국어 단어와 48M 영어 단어를 가진 1.5M 문장 쌍을 포함한다. 상기 FBIS는 GQ의 부분집합이다.

[57] 여기서 Tri는 양측에서의 특성 헤드 단어 트리거를 의미한다. * 또는 **는 기본보다 좋다는 것이다.

[58] 표 2에서, 기본적인 추출 방식에 따라 GQ 말뭉치로부터 152M 번역규칙을 생성한다. 만약 RWF 구조를 사용하여 양측을 제약하면, 번역규칙의 개수가 87M가 되며 이는 전체 번역규칙의 43%가 제거된 것이다.

[59] 표 2로부터 새로운 특성이 2개의 다른 테스트(Test04, Test05) 상에 동작한다.

Test04 상에서 이득은 +2.21% BLEU이고, Test05 상에서 이득은 +1.33% 이다. 번역품질은 격 불감(case-insensitive) BLEU metric을 이용하여 평가한다. RWF 구조만을 이용한 경우Test05 상에서 기본과 동일한 성능을 나타내고 Test04에서는 +0.99% 이득을 나타낸다.

[60] 도 3은 본 발명에 따른 통계적 기계 번역 장치의 내부 구성을 나타낸 것이다. 통계적 기계 번역 장치는 크게 학습(training) 부분과 디코딩(decoding) 부분으로 구성된다.

[61] 학습 부분의 동작을 간단히 설명하면, 먼저 언어 쌍 말뭉치(bilingual corpus)를 이루고 있는 원시언어(source)와 목적언어(target)를 단어 정렬(word alignment)하고 각각을 분석(parsing)하여 의존 트리(dependency tree)를 생성한다. 원시언어와 목적언어에 대한 의존 트리는 본 발명에 의한 느슨한 적격 의존 구조를 이용하여 생성된다. 단어 정렬된 언어 쌍 말뭉치와 각각의 의존 트리는 번역규칙 추출기(rule extractor)로 입력되고 번역규칙 추출기는 번역규칙 세트를 생성한다. 본 발명에 의한 번역규칙 추출기에 의해 생성된 번역규칙 테이블은 기본적인 HPB 시스템의 번역규칙 테이블보다 사이즈가 작다.

[62] 단일 말뭉치(monolingual corpus)는 목적언어에 해당하는 것으로 언어 모델 학습을 거친 후 N-gram 분석 방식을 통해 N-gram 언어모델이 생성된다. 여기서 N-gram은 인접한 N개의 음절을 말한다. 예를 들어, '잡학사전'에서 2-gram은 잡학, 학사, 사전이다.

[63] 디코딩 부분의 동작을 간단히 살펴보면, 원시언어 텍스트는 전 처리된 후 디코더로 입력되고 디코더는 번역규칙 세트와 N-gram 언어모델을 이용하여 목적언어 텍스트를 생성하게 된다. 디코더는 본 발명에 의한 느슨한 적격 의존 구조에 의해 생성된 번역규칙 테이블을 이용하고 헤드 언어 트리거를 적용하여 목적언어 텍스트를 생성한다. 따라서 본 발명에 의한 디코더는 번역성능을 향상시킬 수 있다.

[64] 한편, 본 발명은 상기 번역규칙 필터링과 목적단어 생성을 위한 방법을 소프트웨어적인 프로그램으로 구현하여 컴퓨터로 읽을 수 있는 소정 기록매체에 기록해 둠으로써 다양한 재생장치에 적용할 수 있다. 다양한 재생장치는 PC, 노트북, 휴대용 단말 등일 수 있다.

[65] 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.

[66] 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

[67] 본 발명은 원시언어 측 및 목적언어 측에 모두 느슨한 적격 의존 구조 방식을

적용함으로써 원 번역규칙 테이블의 사이즈를 줄이면서도 번역성능은 종래 HPB 번역 시스템에 비해 좋아지며 또한 이러한 느슨한 적격 의존 구조 방식과 함께 새로운 언어 특성인 헤드 단어 트리거를 적용하게 되면 번역 성능을 더욱 향상시킬 수 있기 때문에 계층적 구문 기반의 통계적 기계 번역 분야에서 널리 사용될 수 있다.

청구범위

[청구항 1] 느슨한 적격 의존 구조를 이용하여 원시언어 측과 목적언어 측의 계층적 구문 기반 번역규칙을 감소시키는 것을 특징으로 하는 번역규칙 필터링 방법.

[청구항 2] 제1항에 있어서,
상기 느슨한 적격 의존 구조는 w_i ... w_j이고 하기의 조건을 만족하는 것을 특징으로 하는 번역규칙 필터링 방법.

(1) $d\_h \notin [i, j]$

(2) $\forall k \in [i, j], d\_k \in [i, j] \, or \, d\_k = h$

[청구항 3] 제1항에 있어서,
상기 느슨한 적격 의존 구조는 헤드 단어가 아닌 복수의 단어로 이루어진 집합을 포함하는 것을 특징으로 번역규칙 필터링 방법.

[청구항 4] 제3항에 있어서,
상기 집합을 이루는 복수의 단어는 공통의 헤드 단어에 의존되는 것을 특징으로 하는 번역규칙 필터링 방법.

[청구항 5] 원시언어 및 목적언어의 문장을 구성하는 단어를 정렬하는 단계와,
상기 정렬된 단어를 매트릭스로 구성하는 단계와,
상기 매트릭스에서 공통의 헤드 단어에 의존되는 단어를 묶어 어구를 생성하는 단계와,
상기 생성된 어구를 이용하여 번역규칙을 생성하는 단계를 포함하는 것을 특징으로 하는 번역규칙 생성 방법.

[청구항 6] 제5항에 있어서,
상기 생성된 어구를 구성하는 단어는 헤드 단어가 아닌 것을 특징으로 하는 번역규칙 생성 방법.

[청구항 7] 목적단어의 통계적 생성은 대응하는 원시단어뿐만 아니라 원시단어의 문맥적 헤드 단어에 의해 트리거 되는 것을 특징으로 하는 목적단어 생성 방법.

[청구항 8] 제7항에 있어서,
상기 원시단어의 문맥적 헤드 단어에 의해 트리거 되어 의존 에지의 조건 하에서 목적단어를 생성하는 것을 특징으로 하는 목적단어 생성 방법.

[청구항 9] 제7항에 있어서,
상기 헤드 단어에 의한 트리거는 로그 선형 모델에 통합되는 것을 특징으로 하는 목적단어 생성 방법.

[청구항 10] 원시언어 측과 목적언어 측에 느슨한 적격 의존 구조를 이용하여 계층적 구문 기반 번역규칙을 생성하고,
상기 생성된 번역규칙을 이용하고 원시단어의 헤드 단어에 의한 트리거 방식을 적용하여 원시언어 텍스트를 목적언어 텍스트로 번역하는 것을 특징으로 하는 계층적 구문 기반의 통계적 기계 번역 방법.

[청구항 11] 제10항에 있어서,
상기 느슨한 적격 의존 구조는 헤드 단어가 아닌 복수의 단어로 이루어진 집합으로 포함하며, 상기 복수의 단어는 공통의 헤드 단어에 의존되는 것을 특징으로 하는 계층적 구문 기반의 통계적 기계 번역 방법.

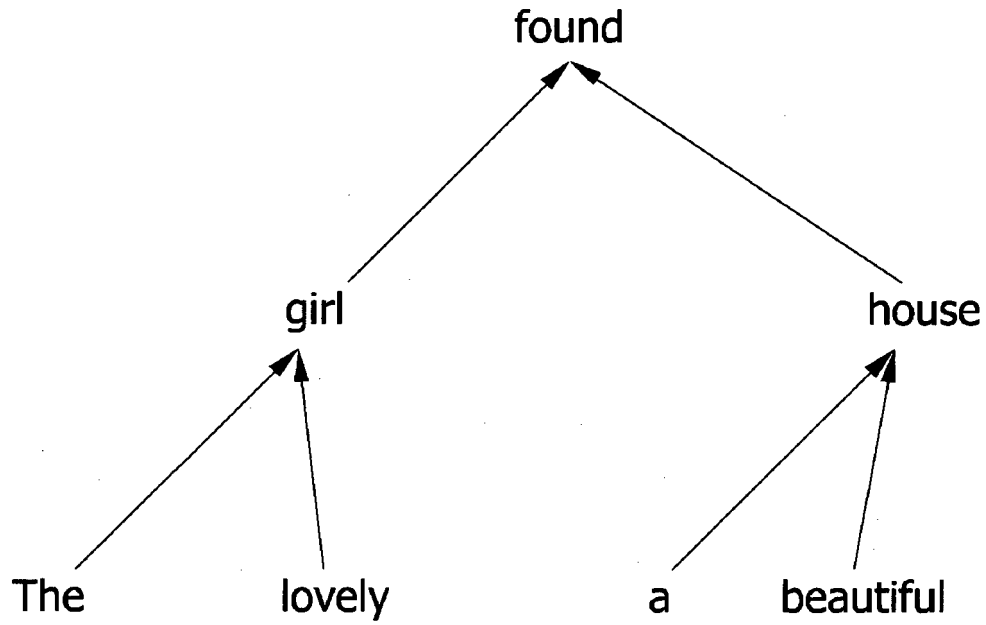
[청구항 12] 원시언어 및 목적언어 문장으로 구성된 언어 쌍 말뭉치를 단어 정렬하는 단어 정렬기와,
상기 언어 쌍 말뭉치를 파싱 하여 느슨한 적격 의존 구조에 따라 의존 트리를 생성하는 단어 분석기와,
상기 단어 정렬된 언어 쌍 말뭉치와 의존 트리를 이용하여 번역규칙을 생성하는 번역규칙 추출기를 포함하는 것을 특징으로 하는 번역규칙 생성장치.

[청구항 13] 언어 쌍 말뭉치로부터 느슨한 적격 의존 구조에 의해 생성된 번역규칙과 단일 말뭉치로부터 생성된 언어모델을 이용하여 원시언어 텍스트를 목적언어 텍스트로 변환하는 것을 특징으로 하는 디코더.

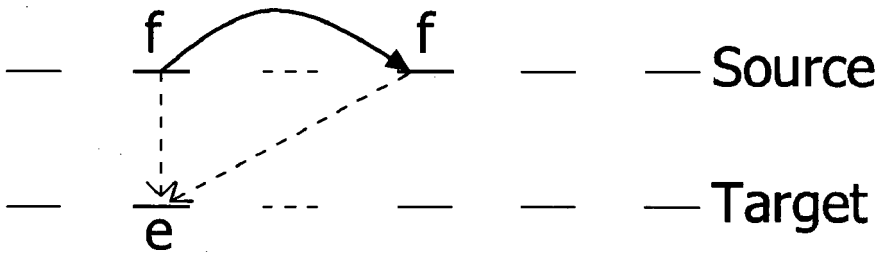
[청구항 14] 제14항에 있어서,
상기 목적언어 텍스트를 구성하는 목적단어는 상기 원시언어를 구성하는 원시단어 및 그 원시단어의 헤드단어에 트리거 되어 생성되는 것을 특징으로 하는 디코더.

[청구항 15] 제1항 내지 제11항 중 어느 한 항에 의한 과정을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

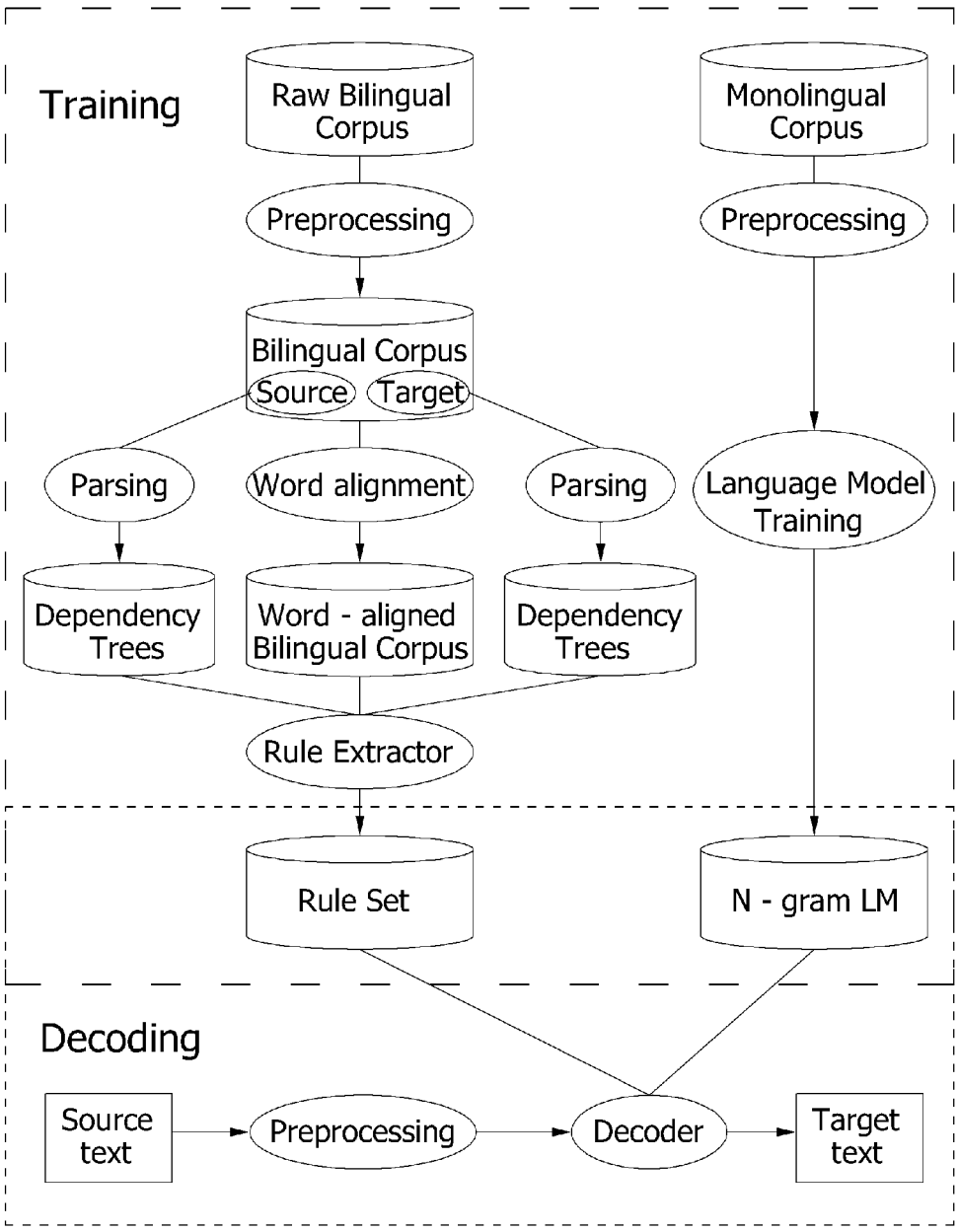
[Fig.1]



[Fig.2]



[Fig. 3]


Training
Raw Bilingual Corpus
Monolingual Corpus
Preprocessing
Preprocessing
Bilingual Corpus
Source
Target
Parsing
Word alignment
Parsing
Language Model Training
Dependency Trees
Word - aligned Bilingual Corpus
Dependency Trees
Rule Extractor
Rule Set
N - gram LM
Decoding
Source text
Preprocessing
Decoder
Target text