(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0057742
(43) 공개일자 2012년06월07일

(51) 국제특허분류(Int. Cl.)
G06F 17/28 (2006.01) *G06F 7/76* (2006.01)
(21) 출원번호 10-2010-0117568
(22) 출원일자 2010년11월24일
심사청구일자 없음

(71) 출원인
에스케이플래닛 주식회사
서울특별시 중구 을지로 65 (을지로2가)
고려대학교 산학협력단
서울 성북구 안암동5가 1
(72) 발명자
황영숙
서울특별시 성북구 북악산로 913, 풍림아파트 105동 502호 (돈암동)
전응수
서울 성북구 안암5가 고려대학교 자연계 캠퍼스 아산이학관 237
(뒷면에 계속)
(74) 대리인
김기효, 전철용

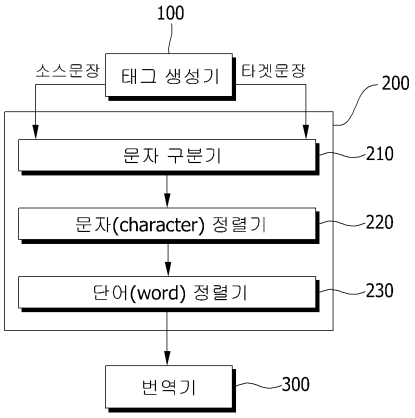
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **통계적 기계 번역에서 문자 정렬을 이용한 단어 정렬 방법 및 이를 이용한 장치**

(57) 요 약

본 발명은 소스 언어와 타겟 언어의 양국어를 포함한 병렬 말뭉치로부터 서로 대응되는 단어를 추출할 때 소스 언어 및 타겟 언어를 문자 단위로 구분하고 구분된 문자를 단어 정렬 툴에 입력하여 문자 단위로 정렬한 다음 정렬된 각 문자를 다시 단어로 매핑 하여 단어 단위로 2차 정렬을 수행함으로써 기존 단어 정렬에서 정렬되지 않았던 단어와도 정렬이 가능하여 정확성을 향상시킨 통계적 기계 번역에서 문자 정렬을 이용한 단어 정렬 방법 및 이를 이용한 장치를 제공한다.

대 표 도 - 도2



(72) 발명자

김상범

서울특별시 노원구 동일로231길 86, 202동 1207호 (상계동, 현대2차아파트)

윤창호

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

임해창

서울특별시 서초구 방배동 1028-1 경남아파트 3동 401호

특허청구의 범위

청구항 1

통계적 기계 번역 시스템에서 단어 정렬을 위한 장치로서,

번역할 대상 언어인 소스 언어 문장과 상기 소스 언어 문장을 원하는 언어로 번역하기 위한 타겟 언어 문장을 포함한 병렬 말뭉치를 문자(character) 단위로 분리(segmentation)하는 문자 구분기;

상기 문자 단위로 분리된 문장을 단어 정렬 툴(tool)에 입력하여 문자 단위로 정렬하는 문자 정렬기;

상기 문자 단위로 정렬된 말뭉치를 단어(word)와 매핑하여 단어 단위로 정렬하는 단어 정렬기

를 포함하는 것을 특징으로 하는 통계적 기계 번역에서 문자 정렬을 이용한 장치.

청구항 2

제 1 항에 있어서,

상기 문자 정렬기는

상기 문자 구분기를 통해 분리된 각 문자를 단어 정렬 툴에 입력하여 상기 각 문자에 포함된 태그 정보를 근거로 문자 정렬을 수행하는 것을 특징으로 하는 통계적 기계 번역에서 문자 정렬을 이용한 장치.

청구항 3

제 2 항에 있어서,

상기 각 문자에 포함된 태그 정보는 상기 문자를 포함하는 단어의 품사 정보, 상기 문자가 단어에서 위치하는 위치 정보를 포함한 것을 특징으로 하는 통계적 기계 번역에서 문자 정렬을 이용한 장치.

청구항 4

제 1 항에 있어서,

상기 단어 정렬기는

단어 단위로 분리된 문장과 문자 단위로 분리된 문장에 근거하여 문자의 위치 정보를 추출하고 추출한 위치 정보를 토대로 상기 문자 정렬기에서 문자 단위로 정렬된 말뭉치를 단어 단위로 매핑하는 것을 특징으로 하는 통계적 기계 번역에서 문자 정렬을 이용한 장치.

청구항 5

제 4 항에 있어서,

상기 단어 정렬기는

상기 단어 단위로 분리된 문장으로부터 해당 문자를 포함한 단어가 문장에서 위치한 위치 정보를 추출하고, 상기 문자 단위로 분리된 문장으로부터 해당 문자가 문장에서 위치한 위치 정보를 추출하는 것을 특징으로 하는 통계적 기계 번역에서 문자 정렬을 이용한 장치.

청구항 6

제 1 항에 있어서,

상기 소스 언어 문장과 타겟 소스 문장을 포함한 병렬 말뭉치에 각 문자 단위로 문자의 품사 정보 및 문자의 위치 정보를 태그로 부착하여 상기 문자 구분기로 제공하는 태그 생성기를 더 포함하는 것을 특징으로 하는 통계적 기계 번역에서 문자 정렬을 이용한 장치.

청구항 7

문자 정렬을 위한 장치로서,

번역할 대상 언어인 소스 언어 문장과 상기 소스 언어 문장을 원하는 언어로 번역하기 위한 타겟 언어 문장을 포함한 병렬 말뭉치에 각 문자 단위로 문자의 품사 정보 및 문자의 위치 정보를 태그로 부착하여 문자별로 태그를 생성하는 것을 특징으로 하는 장치.

청구항 8

통계적 기계 번역 시스템에서 단어 정렬을 위한 방법으로서,

번역할 대상 언어인 소스 언어 문장과 상기 소스 언어 문장을 원하는 언어로 번역하기 위한 타겟 언어 문장을 포함한 병렬 말뭉치를 문자(character) 단위로 분리하는 단계;

상기 문자 단위로 분리된 문장을 단어 정렬 툴(tool)에 입력하여 문자 단위로 정렬하는 단계;

상기 문자 단위로 정렬된 말뭉치를 단어(word)와 매핑하여 단어 단위로 정렬하는 단계

를 포함하는 것을 특징으로 하는 단어 정렬 방법.

청구항 9

제 8 항에 있어서,

상기 문자 단위로 분리하는 단계 이전에,

상기 소스 언어 문장과 타겟 언어 문장을 포함한 병렬 말뭉치에 각 문자 단위로 문자의 품사 정보 및 문자의 위치 정보를 태그로 부착하여 제공하는 단계를 더 포함하는 것을 특징으로 하는 단어 정렬 방법.

청구항 10

제 9 항에 있어서,

상기 문자 단위로 정렬하는 단계는,

상기 문자 단위로 분리된 각 문자를 단어 정렬 툴에 입력하여 각 문자에 포함된 태그 정보를 근거로 문자 정렬을 수행하는 것을 특징으로 하는 단어 정렬 방법.

청구항 11

제 8 항에 있어서,

상기 단어 단위로 정렬하는 단계는,

단어 단위로 분리된 문장과 문자 단위로 분리된 문장에 근거하여 문자의 위치 정보를 추출하고 추출한 위치 정보를 토대로 상기 문자 정렬기에서 문자 단위로 정렬된 말뭉치를 단어 단위로 매핑하는 것을 특징으로 하는 단어 정렬 방법.

청구항 12

제 11 항에 있어서,

상기 단어 단위로 정렬하는 단계는,

상기 단어 단위로 분리된 문장으로부터 해당 문자를 포함한 단어가 문장에서 위치한 위치 정보를 추출하고, 상기 문자 단위로 분리된 문장으로부터 해당 문자가 문장에서 위치한 위치 정보를 추출하는 것을 특징으로 하는 단어 정렬 방법.

청구항 13

제 8 항 내지 제 12 항 중 어느 한 항에 의한 과정을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

명 세 서

기 술 분 야

[0001] 본 발명은 통계적 기계 번역에 관한 것으로서, 더욱 상세하게는 소스 언어와 타겟 언어의 양국어를 포함한 병렬 말뭉치로부터 서로 대응되는 단어를 추출할 때 소스 언어 및 타겟 언어를 문자 단위로 구분하고 구분된 문자를 단어 정렬 툴에 입력하여 문자 단위로 정렬한 다음 정렬된 각 문자를 다시 단어로 매핑 하여 단어 단위로 2차 정렬을 수행함으로써 기존 단어 정렬에서 정렬되지 않았던 단어와도 정렬이 가능하도록 함으로써 단어 정렬 정확성을 향상시킨 통계적 기계 번역에서 문자 정렬을 이용한 단어 정렬 방법 및 이를 이용한 장치에 관한 것이다.

배 경 기 술

[0002] 자동 번역 기술은 한 언어를 다른 언어로 자동으로 전환해주는 소프트웨어적 기술을 의미한다. 이러한 기술은 20세기 중반부터 미국에서 군사적인 목적으로 연구가 시작되었으며, 지금은 세계적으로 정보접근범위의 확대와 휴먼인터페이스의 혁신을 목적으로 다수의 연구소와 민간기업에서 활발히 연구 중이다.

[0003] 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 모델을 자동으로 학습하는 통계적 기반 번역 기술에 대한 개발이 활발히 전개되고 있다.

[0004] 통계적 기반 번역 기술에는 기본적으로 번역할 원시 언어를 분석하여 원하는 타겟 언어로 번역하기 위해 단어 정렬을 수행하게 되는데, 단어 정렬은 병렬 말뭉치(parallel corpus)에서 서로 대응되는 단어를 찾아내는 작업이다.

[0005] 특히, 중한 말뭉치에서의 단어 정렬은 문자 전환 사전을 이용하여 중국어 단어의 문자를 한국어 문자로 전환하고 전환된 한국어를 병렬 말뭉치의 한국어 단어들과 유사도를 계산하여 유사도가 높은 단어를 정렬하는 방법을 적용하고 있다. 이 방법은 중국어와 한국어의 언어학적 규칙(구조 및 변환 규칙)을 고려하여 문자 사전과 단어 사전을 이용함으로써 규칙 기반 방식과 통계 기반 방식을 통합한 방법이라 할 수 있다.

[0006] 다른 예로, 중국어-영어 말뭉치에서의 단어 정렬은 중국어 단어와 영어 단어가 1:1로 정렬되게 하기 위해서 분리(segmentation)된 중국어 단어를 더 작은 의미의 형태소로 잘라서 형태소와 영어 단어를 정렬하는 방법을 적용하고 있다. 이 방법은 정렬에 적용되는 IBM 모델이 n:1 을 정렬할 수 없으므로, 정렬이 안 되는 단어를 더 작은 형태소로 잘라서 정렬하여 정렬하지 못했던 영어 단어와도 정렬이 가능하게 한 방법으로 정렬 성능을 높였다. 그러나, 이 방법은 중국어를 모든 가능한 형태소로 잘라서 정렬하다 보니 정렬 시간이 길어지는 문제점이 있다. 또한, 정렬한 코퍼스는 문장이 아닌 터미놀로지(terminology)를 사용하여 일반 문장보다 짧다.

[0007] 또 다른 예로, 중국어-일본어 말뭉치에서의 단어 정렬 방법은 중국어와 일본어의 언어적 특성을 고려하여 정렬한 방법이 있다. 이 방법은 중국어와 일본어를 분리한 결과에서 일본어에서 사용하는 중국어 문자가 번자체일 때 그 문자를 간자체로 변경한 후 일본어와 중국어가 같은 문자끼리 정렬한다. 이렇게 정렬된 단어를 믿을 수 있는 단어 정렬로 간주하고 나머지 단어들을 자체 개발한 단어 정렬 방법으로 정렬하는 방법을 적용하고 있다.

[0008] 그런데, 상기에서 언급한 단어 정렬 방법들은 기본적으로 원어에서 목표어로의 정렬에 있어서 여러 개의 원어 단어를 한 개의 목표어의 단어로 정렬하는데 정렬이 정확하게 이루어지지 않는 문제가 있다. 특히, 중국어 문장은 문자의 열로 구성되어 문장에 단어를 구분할 수 있는 분명한 부호가 없기 때문에 중국어를 분리한 결과에 오류가 발생한다. 분리의 오류는 한 개 단어로 인식해야 하는 단어를 여러 개의 단어로 인식하는 오류가 있는데, 이는 단어 정렬 오류의 원인이 될 수 있다.

[0009] 한편, 노르웨이어-스웨덴어 번역에서는 문자를 이용하여 정렬부터 통계적 기계 번역까지 수행하는 문자 정렬 방법이 제안되었다. 이 문자 정렬 방법은 노르웨이어와 스웨덴어의 문법적, 구조적 유사성을 이용하여 문자 단위로 정렬을 수행하고 그 수행 결과를 구 기반 통계적 기계 번역에 이용하여 미등록어에 대해 잘 처리할 수 있게 하였다. 그런데, 이 문자 정렬 방법에서는 툴킷(toolkit)을 사용하여 보다 효과적인 결과를 얻고자 말뭉치 내 문장의 길이를 제한하였다. 즉, 40 문자 이상으로 구성된 문장은 모두 제거하였다. 이는 많은 문장을 사용하지 못하는 단점이 있다. 특히, 노르웨이어와 스웨덴어, 두 언어는 라틴어 계열로서 한 문장을 40 문자 이내로 제한하면 제한된 문자 40개 내에 포함되는 단어가 얼마 없다. 이에 따르면, 종래 문자 정렬 방법은 라틴어 계열 언어에 적합하지 않는 문제점이 있다.

발명의 내용

해결하려는 과제

[0010] 본 발명은 상기의 문제점을 해결하기 위해 창안된 것으로서, 소스 언어와 타겟 언어를 포함한 병렬 말뭉치로부터 서로 대응되는 단어를 추출할 때, 소스 언어 및 타겟 언어의 병렬 말뭉치로부터 단어를 문자 단위로 분할하여 각 문자가 속한 단어의 품사 정보와 문자의 위치 정보를 근거로 문자 단위로 정렬을 수행하고 정렬된 각 문자를 단어 단위로 매핑 하여 단어 단위로 재차 정렬을 수행함으로써 기존 단어 정렬에서 정렬되지 않았던 단어와도 정렬이 가능하도록 하여, 정렬의 정확성을 향상시킨 통계적 기계 번역에서 문자 정렬을 이용한 단어 정렬 방법 및 이를 이용한 장치를 제공하고자 한다.

과제의 해결 수단

[0011] 이를 위하여 본 발명의 제1 측면에 따르면, 본 발명의 장치는, 통계적 기계 번역 시스템에서 단어 정렬을 위한 장치로서, 번역할 대상 언어인 소스 언어 문장과 상기 소스 언어 문장을 원하는 언어로 번역하기 위한 타겟 소스 문장을 포함한 병렬 말뭉치를 문자(character) 단위로 분리(segmentation)하는 문자 구분기; 상기 문자 단위로 분리된 문장을 단어 정렬 툴(tool)에 입력하여 문자 단위로 정렬하는 문자 정렬기; 상기 문자 단위로 정렬된 말뭉치를 단어(word) 단위 문장과 매핑하여 단어 단위로 정렬하는 단어 정렬기를 포함하는 것을 특징으로 한다.

[0012] 본 발명의 제2 측면에 따르면, 본 발명의 문자 정렬을 위한 장치는, 번역할 대상 언어인 소스 언어 문장과 상기 소스 언어 문장을 원하는 언어로 번역하기 위한 타겟 문장을 포함한 병렬 말뭉치에 각 문자 단위로 문자의 품사 정보 및 문자의 위치 정보를 태그로 부착하여 문자 별로 태그를 생성하는 것을 특징으로 한다.

[0013] 본 발명의 제3 측면에 따르면, 본 발명의 단어 정렬 방법은, 통계적 기계 번역 시스템에서 단어 정렬을 위한 방법으로서, 번역할 대상 언어인 소스 언어 문장과 상기 소스 언어 문장을 원하는 언어로 번역하기 위한 타겟 언어 문장을 포함한 병렬 말뭉치를 문자(charater) 단위로 분리하는 단계; 상기 문자 단위로 분리된 문장을 단어 정렬 툴(tool)에 입력하여 문자 단위로 정렬하는 단계; 상기 문자 단위로 정렬된 말뭉치를 단어(word)와 매핑하여 단어 단위로 정렬하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[0014] 본 발명에 따르면, 한 방향 정렬에서 n:1 정렬이 가능해 진다. 나아가 양방향 정렬의 교집합(Intersection)을 취한 결과 1:n 정렬이 가능해 지므로 단어 단위 정렬보다 더 많은 정렬 페어(pair)를 생성함으로써 분리(segmentation)의 오류 및 형태소 분석 오류로 인해 발생하는 정렬 오류를 해결할 수 있는 효과가 있다. 또한, 한쪽 언어에서 약어를 사용하더라도 단어 정렬을 정확히 할 수 있다. 또한, 자료 부족 문제를 완화할 수 있다.

도면의 간단한 설명

[0015] 도 1은 일반 단어 정렬 알고리즘을 통해 획득한 단어 정렬 결과를 보인 도면.

도 2는 본 발명의 실시 예에 따른 기계 번역 장치를 나타낸 구성도.

도 3은 본 발명의 실시 예에 따른 기계 번역 장치의 단어 정렬 방법을 설명하기 위한 순서도.

도 4는 도 3의 문자 단위로 분리하는 과정을 구체적으로 설명하기 위한 순서도.

도 5는 종래 기술에 따른 단어 정렬 결과와 본 발명의 기술에 따른 단어 정렬 결과를 나타낸 도면.

발명을 실시하기 위한 구체적인 내용

[0016] 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.

[0017] 본 발명은 병렬 말뭉치를 이용한 통계적 기계 번역에서 소스 언어를 타겟 언어로 번역하기 위해 수행하는 단어 정렬 과정에서 정렬의 오류를 최소화하기 위한 구조를 제공한다.

[0018] 이하에서 언급하는 소스 문장(source sentence) 또는 소스 언어 문장은 번역할 대상이 되는 원시 언어의 문장

이고, 타겟 문장(target sentence) 또는 타겟 언어 문장은 소스 문장을 원하는 언어로 번역하여 출력되는 목표 언어의 문장을 의미한다.

[0019] 또한, 본 발명은 소스 언어와 타겟 언어 두 언어의 문자들 사이에 문법 또는 구조적으로 일정한 유사성이 있음을 전제로 한다. 예를 들어, 소스 언어의 문자가 형용사이면 문법적으로 이와 대응되는 타겟 언어의 문자 또한 형용사일 확률이 높다. 또한 소스 언어의 명사 문자는 타겟 언어에서 명사인 문자에 정렬될 경향이 높다.

[0020] 도 1은 일반 단어 정렬 알고리즘을 통해 획득한 단어 정렬 결과를 보인 도면이다.

[0021] 도 1에서, 실선으로 표기된 연결선은 정확하게 정렬된 경우이다. 점선은 단어 정렬을 통해 소스 언어의 단어와 타겟 언어의 단어가 대응되어야 하는데, 실제 IBM 모델을 통한 정렬 결과에서 생성하지 못한 관계를 표시한 것이다.

[0022] 도시한 예에서, 소스 언어가 중국어이고 타겟 언어가 한국어인 경우를 보면, IBM모델을 적용한 결과 중국어 문장에서의 "秘?室(secretary)" 단어는 한국어 문장에서 "비서실장(executive secretary)"과 정렬되지만, 중국어 문장에서의 "室長(director)" 단어는 "비서실장(executive secretary)" 한국어 단어에 정렬되지 못한다. 이처럼, 기존 단어 정렬에 적용되는 알고리즘 특히, IBM 모델은 n:1 대응 관계로 단어 정렬을 할 수 없다.

[0023] 따라서, 본 발명은 단어 정렬을 수행하기 전에 병렬 말뭉치를 문자 단위로 세분화하여 문자 정렬을 수행하고 다시 단어 단위로 묶음으로써 상기와 같이 단어 정렬 알고리즘을 통한 단어 정렬의 오류를 보완한다.

[0024] 이를 구현하기 위한 구성은 다음과 같다.

[0025] 도 2는 본 발명의 실시 예에 따른 기계 번역 장치를 나타낸 구성도이다.

[0026] 본 발명의 실시 예에 따른 기계 번역 장치는 크게 태그 생성기(100)와 정렬기(200) 및 번역기(300)를 포함하며, 정렬기(200)는 문자 구분기(210), 문자(character) 정렬기(220), 단어(word) 정렬기(230)를 포함한다.

[0027] 태그 생성기(100)는 소스 언어 문장 및 타겟 언어 문장을 이루는 각 문자(character)에 태그를 부착한다. 태그는 문자의 특징 정보로서 다음의 두 가지를 포함한다. 첫 째는 문자를 포함한 단어의 품사 정보 태그이고, 둘째는 문자가 단어에서의 위치 정보 태그이다. 문자가 단어에서 맨 처음 문자이면 LL, 단어에서 맨 마지막 문자이면 RR, 한 문자가 한 개 단어이면 LR, 2문자 이상으로 구성된 단어에서 맨 처음 또는 마지막도 아니면 MM 태그로 구분하여 생성할 수 있다.

[0028] 여기서, 태그 생성기(100)는 정렬기(200)와 일체로 하나의 장치에 구현될 수 있지만 별도로 구비하여 소스 언어 문장 및 타겟 언어 문장에 대한 각 문자의 태그를 사전에 생성 및 관리할 수 있다.

[0029] 또한, 태그 생성기(100)는 형태소 분석을 거쳐 형태소 단위로 분리된 소스 언어 문장 및 타겟 언어 문장의 단어(형태소)에 품사 정보 및 위치 정보를 포함시킬 수 있다. 이는 각 문자를 포함하는 단어에 단어의 품사 정보와 위치 정보를 삽입함으로써 문자 단위로 구분된 말뭉치를 단어 단위로 복귀할 때 이용한다.

[0030] 문자 구분기(210)는 태그 생성기(100)로부터 소스 언어와 타겟 언어의 병렬 말뭉치를 수신하면 소스 언어 문장 또는 타겟 언어 문장을 문자 단위로 구분한다.

[0031] 이때, 문자 구분기(210)는 소스 언어 문장 또는 타겟 언어 문장을 문자 단위로 구분할 때 바이트(byte) 단위로 읽으면서 한 개의 문자가 아스키코드(ASCII)의 최대치보다 큰 지의 여부를 판단하여 문자 크기 레벨에 따라 분리(segmentation)한다. 예를 들어, 1바이트(byte)가 아스키코드의 최대치보다 크면 다음 바이트(byte)까지 합쳐서 2 바이트(byte)를 한 문자로 분리하고, 1바이트(byte)가 아스키코드의 최대치보다 작거나 같으면 1바이트(byte)를 한 문자로 분리한다. 이러한 구분 방법에 따르면, 중국어, 한국어 문장은 2바이트(byte)가 한 문자로 분리되고, 영어 또는 부호는 1바이트(byte)가 한 문자로 분리될 수 있다.

[0032] 문자(character) 정렬기(220)는 문자 구분기(210)에서 전처리 과정을 거쳐 문자로 분리된 말뭉치를 단어 정렬 알고리즘(GIZA++)에 입력 문장으로 입력하여 문자 단위로 1차 정렬을 수행한다. 이때, 단어 정렬 알고리즘을 이용한 문자 정렬은 각 문자에 부착된 태그 정보를 근거로 수행한다.

[0033] 또한, 문자 정렬기(220)는 소스 언어와 타겟 언어에 대하여 양방향으로 수행한다. 예를 들어, 소스 언어를 중국어로 타겟 언어를 한국어로 설정하여 정렬하고, 또한 반대로 소스 언어를 한국어로 타겟 언어를 중국어로 정렬하여 양방향으로 수행한다.

[0034] 또한, 본 발명의 실시 예에 따른 문자 정렬기(220)는 문자 정렬 전에 문자 단위로 분리된 말뭉치에서 정해진

개수 이상의 문자를 포함한 문장이 존재하면 해당 문장과 병렬 말뭉치에서 해당 문장과 대응하는 문장을 지워서 두 언어 말뭉치의 문장 수를 같게 한다. 그리고 두 언어 말뭉치를 유니 코드로 인코딩하는 전처리 과정을 수행한다.

[0035] 단어(word) 정렬기(230)는 문자 정렬기(220)에서 문자 단위로 정렬된 말뭉치를 단어 단위로 매핑하여 단어로 정렬한다. 단어 정렬기(230)는 단어 단위로 매핑하기 위해 단어 단위로 분리된 문장과 문자 단위로 분리된 문장에 근거하여 문자의 위치 정보를 저장한다. 이때, 한 개의 문자는 두 가지 의미의 위치 정보를 가지게 된다.

[0036] 첫째, 문자가 문장에서 어느 위치에 배열되는지에 대한 문장에서의 위치 정보이고, 둘째 그 문자를 포함한 단어가 문장에서 어느 위치에 배열되는지에 대한 문장에서의 위치 정보이다.

[0037] 이 두 가지 위치 정보에 근거하여 문자 단위로 분리된 문장을 단어 단위로 매핑하여 단어 단위로 정렬한 결과를 얻을 수 있다.

[0038] 문자 단위로 분리된 중국어 문장을 $C_{j'} = C_1C_2C_3\ldots C_{J'-2}C_{J'-1}C_{J'}$, 한국어 문장을 $K_{i'} = K_1K_2K_3\ldots\ldots K_{I'-2}K_{I'-1}K_{I'}$라 하면 중국어 문자 위치 정보 및 한국어 문자 위치 정보는 일예로 $C_{j'j} = C_{11}C_{21}C_{32}\ldots\ldots C_{J'-2J-1}C_{J'-1J}C_{J'J}$, $K_{i'i} = K_{11}K_{21}K_{32}\ldots\ldots K_{I'-2I-1}K_{I'-1I}K_{I'I}$와 같이 표현할 수 있다.

[0039] 상기의 예시에서, 문자 정렬 결과 중국어 문장에서 3번째 문자와 한국어 문장에서 3번째 문자가 정렬되고 위의 위치 정보에 근거하여 단어의 정렬 정보를 찾는다면 중국어 문장의 3번째 문자는 2번째 단어에 해당하고, 한국어 문장의 3번째 문자는 1번째 단어에 해당함을 알 수 있다. 따라서, 중국어 문장의 2번째 단어와 한국어 문장의 1번째 단어가 정렬된다는 것을 생성할 수 있다.

[0040] 번역기(300)는 단어 정렬기(230)를 통해 정렬된 정보를 토대로 통계적 기계 번역을 수행한다.

[0041] 도 3은 본 발명의 실시 예에 따른 기계 번역 장치의 단어 정렬 방법을 설명하기 위한 순서도로, 이를 참조하여 통계적 기계 번역 시스템에서의 단어 정렬을 수행하는 방법을 설명한다.

[0042] 먼저, 태그 생성기(100)로부터 각 문자에 태그가 부착된 소스 문장 및 타겟 문장을 포함한 병렬 말뭉치를 수신하면 문자 구분기(210)에서는 태그가 부착된 소스 문장 및 태그가 부착된 타겟 문장으로부터 한 개의 문자(character)를 추출하여 아스키 코드(ASCII)와 비교한다(S100, S110).

[0043] 비교 결과, 한 개의 문자가 아스키 코드의 최대치보다 큰 경우 또는 작거나 같은 경우 해당 레벨에 따라 소스 문장 및 타겟 문장의 문자 단위를 다르게 분리한다(S120, S130). 이는 소스 문장 또는 타겟 문장에서 문자 정보로 인식할 수 있는 단위를 판독하기 위한 과정으로 구체적인 방법은 도 4와 같다.

[0044] 즉, 도 4에 도시한 것처럼 한 개의 문자가 아스키 코드의 최대치보다 크면 한 개의 문자 단위를 다음 바이트까지 합쳐서 2바이트 단위로 분리한다(S132, S134).

[0045] 또한, 한 개의 문자가 아스키 코드의 최대치보다 작거나 같으면 한 개의 문자 단위를 1바이트로 분리한다(S136, S138).

[0046] 이후, 문자 정렬기(220)에서 문자 단위로 분리된 말뭉치를 문자 구분기(210)로부터 수신하면 이를 정렬 알고리즘 GIZA++에 입력 문장으로 입력하여 문자 단위로 정렬한다(S140, S150).

[0047] 이때, 문자 단위의 정렬은 문자 단위로 분리된 각 문자의 태그 정보, 특히 품사 태그 정보에 근거하여 정렬할 수 있다.

[0048] 이후, 단어 정렬기(230)에서 문자 단위로 정렬된 말뭉치를 단어 단위로 매핑하여 단어 단위로 정렬한다(S160). 단어 단위로 매핑하는 방법은 단어 단위로 분리된 문장과 문자 단위로 분리된 문장을 토대로 문자의 위치 정보를 추출함으로써 가능하다. 문자 단위로 분리된 문장으로부터 해당 문자가 문장에서 어느 위치에 배열되어 있는지에 대한 위치 정보를 알 수 있고, 단어 단위로 분리된 문장으로부터는 해당 문자를 포함한 단어

가 문장에서 어느 위치에 배열되어 있는지에 대한 위치 정보를 알 수 있다.

[0049] 이렇게 문자 및 단어 정렬된 결과는 도 5의 (b)와 같다.

[0050] 문자 정렬을 통해 (b)의 B에 나타낸 바와 같이 소스 문장의 문자와 타겟 문장의 문자가 일대일로 정렬되고 이후 단어 정렬을 수행하면 C에 나타낸 것처럼 소스 문장의 두 단어가 타겟 문장의 한 단어와 정렬되는 n:1 관계로 정렬됨을 알 수 있다.

[0051] 만약, B 단계의 문자 정렬을 수행하지 않고 바로 단어 정렬을 수행하면 도 5의 (a)에 도시한 바와 같이 타겟 문장의 단어 "지역민심"과 정렬되어야 할 소스 문장의 단어 "民心(민심)"이 정렬되지 않는다.

[0052] 이렇게 다량의 문자 정렬 정보를 이용하여 단어 정렬 정보를 추출하면 중복되는 정렬 정보들이 포함될 수 있으므로 중복되는 정렬 정보를 지우면 문자 정렬 결과에서 정확한 단어 정렬 결과를 얻을 수 있다.

[0053] 한편, 본 발명에서 제안한 방법을 소스 문장에서 타겟 문장으로 정렬하는 단 방향이 아닌 타겟 문장에서 소스 문장으로 정렬하는 양방향 정렬을 수행할 수 있다. 이 경우 양방향 정렬의 교집합(intersection)을 취하면 단어 단위 정렬보다 더 많은 정렬을 생성할 수 있다.

[0054] 한편, 본 발명은 이상에서 설명한 단어 정렬 방법을 소프트웨어적인 프로그램으로 구현하여 컴퓨터로 읽을 수 있는 소정 기록 매체에 기록해 둠으로써 다양한 재생 장치에 적용할 수 있다.

[0055] 다양한 재생 장치는 PC, 노트북, 휴대용 단말 등일 수 있다.

[0056] 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.

[0057] 이 경우, 컴퓨터로 읽을 수 있는 기록 매체에 기록한 프로그램은, 앞서 설명한 바와 같이 번역할 대상 언어인 소스 언어 문장과 상기 소스 언어 문장을 원하는 언어로 번역하기 위한 타겟 언어 문장을 포함한 병렬 말뭉치를 문자(charater) 단위로 분리하는 과정과, 상기 문자 단위로 분리된 문장을 단어 정렬 툴(tool)에 입력하여 문자 단위로 정렬하는 과정과, 상기 문자 단위로 정렬된 말뭉치를 단어(word)와 매핑하여 단어 단위로 정렬하는 과정을 포함하여 실행될 수 있다.

[0058] 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

[0059] 병렬 말뭉치를 이용한 통계적 기계 번역 시스템에서, 종래 기술에 따른 단어 정렬은 한 개의 원시언어 단어를 여러 개의 목표 언어 단어로 정렬할 수 없어서 양국어 단어간 정렬이 정확하게 이루어지지 못하는 문제가 있었으나, 본 발명은 문자 단위로 분리하여 문자를 정렬하고 다시 단어와 매핑하여 단어 단위로 정렬함으로써 기존 단어 정렬에서 정렬되지 않았던 단어와도 정렬이 가능하도록 하여 정확성을 향상시킬 수 있다. 이는 한 방향 정렬에서 n:1 정렬이 가능해 지고 나아가 양방향 정렬을 수행하면 1:n 정렬이 가능해 지므로, 기존 단어 정렬보다 더 많은 페어(pair)를 생성함으로써 분리의 오류 및 형태소 분석 오류로 인해 발생하는 정렬 오류를 해결할 수 있다. 또한, 한쪽 언어에서 약어를 사용하더라도 단어 정렬을 정확히 할 수 있고 자료 부족 문제를 완화할 수 있어, 자동 번역기 또는 자동 사전의 구축에 있어서 번역의 품질을 향상시킬 수 있다.

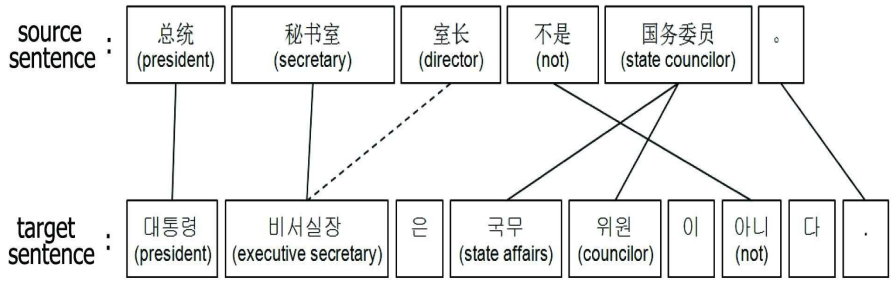
부호의 설명

[0060]

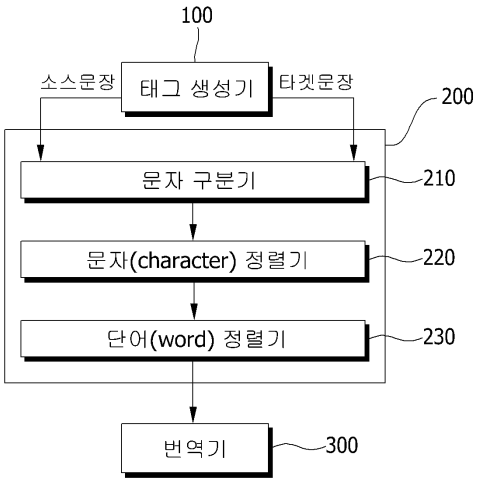
100: 태그 생성기	200: 정렬기
210: 문자 구분기	220: 문자 정렬기
230: 단어 정렬기	300: 번역기

도면

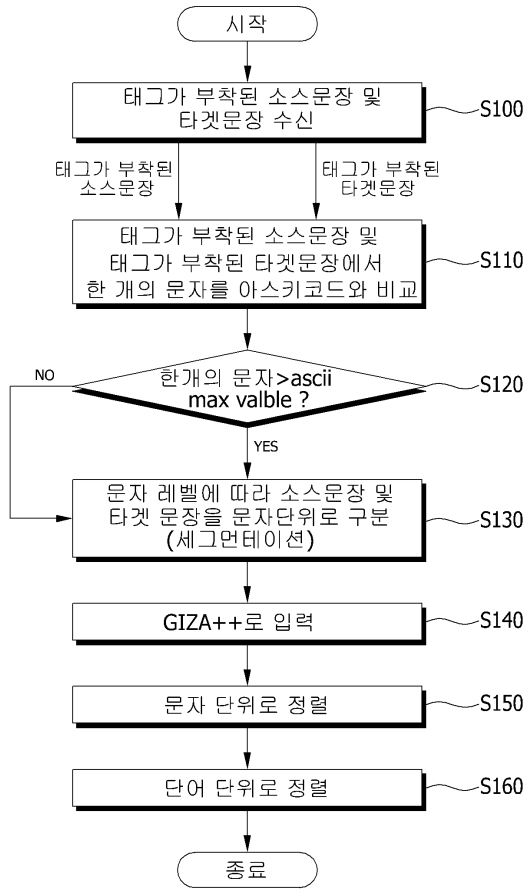
도면1


source sentence :
总统 (president)
秘书室 (secretary)
室长 (director)
不是 (not)
国务委员 (state councilor)
。
target sentence :
대통령 (president)
비서실장 (executive secretary)
은
국무 (state affairs)
위원 (councilor)
이
아니 (not)
다
.

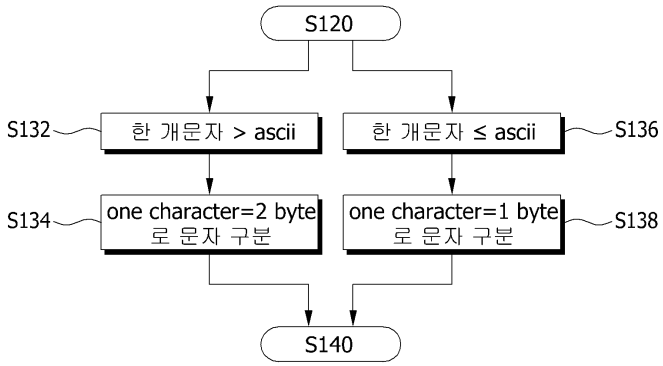
도면2


100
소스문장
태그 생성기
타겟문장
200
문자 구분기
210
문자(character) 정렬기
220
단어(word) 정렬기
230
번역기
300

도면3

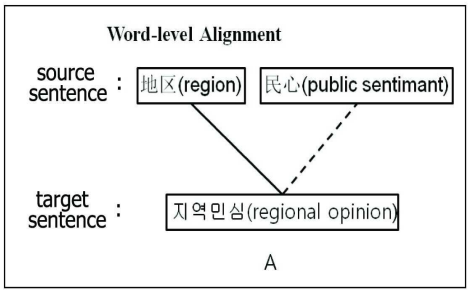

시작
태그가 부착된 소스문장 및 타겟문장 수신
S100
태그가 부착된 소스문장
태그가 부착된 타겟문장
태그가 부착된 소스문장 및 태그가 부착된 타겟문장에서 한 개의 문자를 아스키코드와 비교
S110
NO
한개의 문자>ascii max valble ?
S120
YES
문자 레벨에 따라 소스문장 및 타겟 문장을 문자단위로 구분 (세그먼테이션)
S130
GIZA++로 입력
S140
문자 단위로 정렬
S150
단어 단위로 정렬
S160
종료

도면4

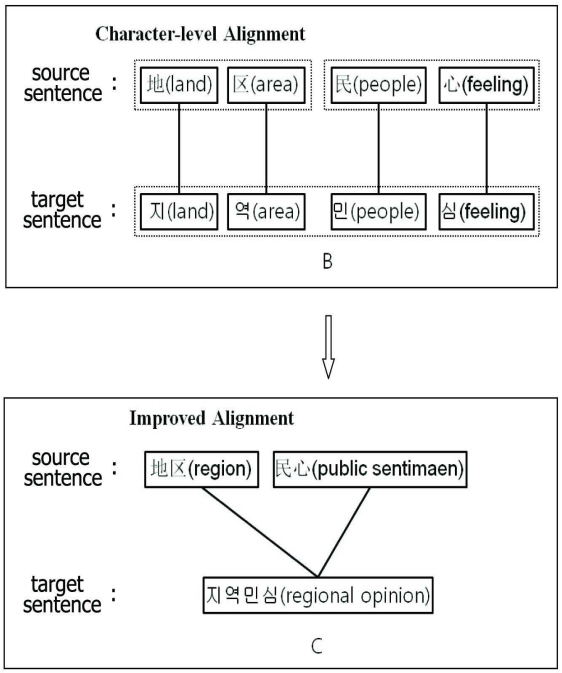

S120
S132
한 개문자 > ascii
한 개문자 ≤ ascii
S136
S134
one character=2 byte 로 문자 구분
one character=1 byte 로 문자 구분
S138
S140

도면5

(a) 종래 :


Word-level Alignment
source sentence :
地区(region)
民心(public sentimant)
target sentence :
지역민심(regional opinion)
A

(b) 본 발명 :


Character-level Alignment
source sentence :
地(land)
区(area)
民(people)
心(feeling)
target sentence :
지(land)
역(area)
민(people)
심(feeling)
B
Improved Alignment
source sentence :
地区(region)
民心(public sentimaen)
target sentence :
지역민심(regional opinion)
C