

|                                                                                   |                                                                                                                                            |                                                    |
|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|
|  | (19) 대한민국특허청(KR)<br>(12) 공개특허공보(A)                                                                                                         | (11) 공개번호 10-2012-0018675<br>(43) 공개일자 2012년03월05일 |
| <hr/>                                                                             |                                                                                                                                            |                                                    |
| (51) 국제특허분류(Int. Cl.)<br>G06F 17/18 (2006.01) G06F 17/28 (2006.01)                | (71) 출원인<br>에스케이플래닛 주식회사<br>서울특별시 중구 을지로 65 (을지로2가)<br>고려대학교 산학협력단<br>서울 성북구 안암동5가 1                                                       |                                                    |
| (21) 출원번호 10-2010-0081663<br>(22) 출원일자 2010년08월23일<br>심사청구일자 없음                   | (72) 발명자<br>김상범<br>서울특별시 노원구 동일로231길 86, 202동 1207호<br>(상계동, 현대2차아파트)<br>윤창호<br>서울특별시 성북구 서경로 31, 푸른동아 아파트<br>107동 1303호 (정릉동)<br>(뒷면에 계속) |                                                    |
|                                                                                   | (74) 대리인<br>김기효, 전철용                                                                                                                       |                                                    |

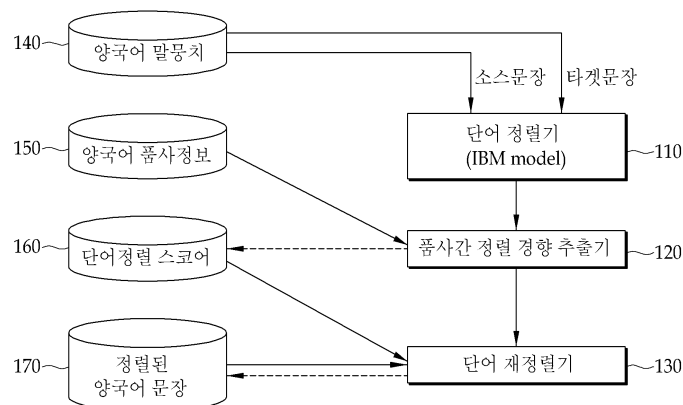
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **품사간 정렬 경향을 반영한 통계적 단어 정렬 방법 및 이를 이용한 기계 번역 장치**

### (57) 요약

본 발명은 통계적 기계 번역에 관한 것으로서, 더욱 상세하게는 소스 언어와 타겟 언어의 양국어 말뭉치로부터 서로 대응되는 단어를 추출하는 단어 정렬 시 각 단어의 품사 정보를 고려하여 품사간 관계가 존재하는 정렬 경향을 단어 정렬 스코어에 반영함으로써 수정된 단어 정렬 스코어로 다시 단어를 재정렬하여 기존 단어 정렬로부터 품사간 정렬 경향이 높은 정렬로 다시 제공함으로써 기존 단어 정렬보다 정확하고 높은 품질의 결과를 얻을 수 있는 통계적 단어 정렬 방법 및 이를 이용한 기계 번역 장치를 제공한다.

**대표도** - 도1



(72) 발명자

**황영숙**

서울특별시 성북구 북악산로 913, 풍림아파트 105  
동 502호 (돈암동)

**임해창**

서울특별시 성북구 안암로 145, 자연계캠퍼스 아산  
이학관 206호 (안암동5가, 고려대학교)

**이재희**

서울특별시 성북구 안암로 145, 자연계캠퍼스 아산  
이학관 237호 (안암동5가, 고려대학교)

## 특허청구의 범위

### 청구항 1

통계적 기계 번역 시스템에서 단어 정렬을 위한 장치로서,

번역할 대상 언어인 소스 문장과 상기 소스 문장을 원하는 언어로 번역한 타겟 문장을 수신한 양국어 말뭉치를 형태소로 세그먼트(segment)한 각 단어에 대하여 대응 가능한 소스 단어와 타겟 단어를 정렬하는 단어 정렬기;

상기 단어 정렬기를 통해 정렬된 양국어 말뭉치로부터 각 단어의 품사 정보를 적용하여 품사간 정렬 경향을 추출하는 품사간 정렬 경향 추출기;

상기 품사간 정렬 경향 추출기로부터 추출된 정렬 경향을 반영한 단어 정렬 확률과 상기 단어 정렬기를 통해 정렬된 이전 단어 정렬 확률을 비교하여 양 스코어가 서로 다르면 상기 품사간 정렬 경향 추출기로부터 추출된 정렬 경향을 반영한 단어 정렬 확률로 수정하고 수정된 단어 정렬 확률에서 가장 높은 스코어를 갖는 단어 쌍으로 수정하여 재정렬하는 단어 재정렬기

를 포함하는 것을 특징으로 하는 장치.

### 청구항 2

제 1 항에 있어서,

상기 단어 정렬기는 IBM 모델을 적용하여 단어 정렬을 수행하는 것을 특징으로 하는 장치.

### 청구항 3

제 1 항에 있어서,

상기 품사간 정렬 경향 추출기는 상기 양국어 말뭉치에서 출현한 각 단어의 품사가 타겟 품사와 정렬되는 상대 빈도 확률을 구하여 추출하는 것을 특징으로 하는 장치.

### 청구항 4

제 1 항에 있어서,

상기 품사간 정렬 경향 추출기는 소스 단어의 품사와 타겟 단어의 품사가 동일 양국어 말뭉치에 출현하였을 때 해당 단어 쌍이 정렬되는 확률을 구하여 추출하는 것을 특징으로 하는 장치.

### 청구항 5

제 4 항에 있어서,

상기 품사간 정렬 경향 추출기는 최대 가능도(Maximum Likelihood Estimation) 방식을 적용한 것을 특징으로 하는 장치.

### 청구항 6

제 1 항에 있어서,

상기 단어 재정렬기는

각 병렬 문장이 가지는 모든 단어 정렬에 대하여 수행하는 것을 특징으로 하는 장치.

### 청구항 7

통계적 기계 번역에서 단어 정렬을 위한 방법으로서,

번역할 대상 언어인 소스 문장과 상기 소스 문장을 원하는 언어로 번역한 타겟 문장을 수신한 양국어 말뭉치를 형태소로 세그먼트(segment)한 각 단어에 대하여 대응 가능한 소스 단어와 타겟 단어를 1차 정렬하는 단계;

상기 1차 정렬된 양국어 말뭉치로부터 각 단어의 품사 정보를 적용하여 품사간 정렬 경향을 추출하는 단계;

상기 추출된 품사간 정렬 경향을 상기 1차 정렬을 통해 획득한 단어 정렬 확률에 반영하여 단어 정렬 스코어를 재계산하는 단계;

상기 1차 정렬을 통해 획득한 기존 단어 쌍을 상기 재계산된 단어 정렬 스코어 중 가장 높은 스코어를 갖는 단어 쌍으로 수정하여 재정렬하는 단계

를 포함하는 것을 특징으로 하는 단어 정렬 방법.

#### 청구항 8

제 7 항에 있어서,

상기 1차 정렬은 IBM 모델을 적용하여 수행하는 것을 특징으로 하는 단어 정렬 방법.

#### 청구항 9

제 7 항에 있어서,

상기 품사간 정렬 경향을 추출하는 단계는 상기 양국어 말뭉치에서 출현한 각 단어의 품사가 타겟 품사와 정렬되는 상대 빈도 확률을 구하여 추출하는 것을 특징으로 하는 단어 정렬 방법.

#### 청구항 10

제 7 항에 있어서,

상기 품사간 정렬 경향을 추출하는 단계는 소스 단어의 품사와 타겟 단어의 품사가 동일 양국어 말뭉치에 출현하였을 때 해당 단어 쌍이 정렬되는 확률을 구하여 추출하는 것을 특징으로 하는 단어 정렬 방법.

#### 청구항 11

제 10 항에 있어서,

상기 확률은 최대 가능도(Maximum Likelihood Estimation) 방식을 적용한 것으로 하기의 수학적식을 만족하는 것을 특징으로 하는 단어 정렬 방법.

$$P(\text{align} = \text{TRUE} | \text{pos}(f), \text{pos}(e)) = \frac{\text{count}(\text{align} = \text{true} | \text{pos}(f), \text{pos}(e))}{\sum_{k \in (\text{true}, \text{false})} \text{count}(\text{align} = k, \text{pos}(f), \text{pos}(e))}$$

(pos(e): 소스 단어의 품사 정보, pos(f): 타겟 단어의 품사 정보)

#### 청구항 12

제 7 항에 있어서,

상기 재정렬하는 단계는

상기 1차 정렬을 통해 획득한 단어 정렬 확률과 상기 품사간 정렬 경향을 반영한 단어 정렬 확률을 비교하여 양 스코어가 서로 다르면 상기 품사간 정렬 경향을 반영한 단어 정렬 확률 스코어로 수정하고 수정된 스코어에서 가장 높은 스코어를 갖는 단어 쌍으로 수정하여 재정렬하는 것을 특징으로 하는 단어 정렬 방법.

#### 청구항 13

제 12 항에 있어서,

상기 재정렬하는 단계는 각 병렬 문장이 가지는 모든 단어 정렬에 대하여 수행하는 것을 특징으로 하는 단어 정렬 방법.

#### 청구항 14

제 7 항 내지 제 13항 중 어느 한 항의 과정을 실행하기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록

매체.

## 명세서

### 기술분야

- [0001] 본 발명은 통계적 기계 번역에 관한 것으로서, 더욱 상세하게는 소스 언어와 타겟 언어의 양국어 말뭉치로부터 서로 대응되는 단어를 추출하는 단어 정렬 시 각 단어의 품사 정보를 반영하여 품사간 관계가 존재하는 정렬 경향을 단어 정렬 스코어에 적용하여 기존 단어 정렬 스코어를 수정함으로써 수정된 단어 정렬 스코어로 단어를 재정렬하여 기존 단어 정렬보다 정확하고 높은 품질의 결과를 제공하는 통계적 단어 정렬 방법 및 이를 이용한 기계 번역 장치에 관한 것이다.

### 배경기술

- [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 중에 있다.
- [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 알고리즘을 자동으로 학습하는 기술 개발이 활발히 전개되고 있다.
- [0004] 통계적으로 번역 알고리즘을 자동으로 학습하는 기술에는 기본적으로 번역할 소스 문장(source sentence)을 분석하여 타겟 문장(target sentence)으로 번역하기 위해 단어 정렬을 수행하게 되는데, 단어 정렬은 병렬 말뭉치(parallel corpus)에서 서로 대응되는 단어를 찾아내는 작업이다.
- [0005] 예를 들어, 영어 문장 "I love you"와 한글 문장 "나는 당신을 사랑합니다"에서 영어 단어와 한글 어절 각각은 다음과 같이 대응된다.
- [0006] I: 나는, love: 사랑합니다, you: 당신을
- [0007] 이와 같이 병렬 말뭉치에서 대응되는 단어를 찾는 작업은 기계 번역에서 단일어 대역어 추출, 번역 규칙 추출, 대역 구 추출, 의미 중의성 해소, 구문 분석 등에 사용되며, 기계 번역의 품질뿐만 아니라 자연어 처리 관련 많은 분야에 큰 영향을 준다.
- [0008] 단어 정렬을 수행하는 방법 중 가장 널리 사용되는 기법으로 Brown et al.,(1993)에서 제안한 IBM 모델(Model)이 있다. IBM Model은 대용량의 병렬 말뭉치에서 단어 쌍들의 공기 정보, 위치 정보 등을 고려하여 대응되는 단어를 찾는 알고리즘이다. 이 알고리즘은 대용량의 병렬 말뭉치를 이용하기 때문에 특정 언어 쌍에 국한되지 않으며 말뭉치의 교체만 이루어지면 모든 단어 쌍에서 사용할 수 있는 장점이 있다. 그러나, IBM Model은 단어의 통계학적 정보만을 사용하며 품사 정보, 구조적 역할 정보 등을 전혀 고려하고 있지 않는 단점을 가진다.
- [0009] 이러한 IBM Model의 단점을 보완하기 위해 구조적, 의미적 언어 정보를 전처리, 후처리 과정에서 말뭉치의 어순을 변경하거나 규칙을 추가하여 정렬을 하는 등의 방법들이 제안되었다.
- [0010] Hermjakob(2009)에서는 아랍어-영어 단어 정렬에서 영어 파서를 이용하여 아랍어와 유사한 어순으로 영어 문장의 어순을 변경하고 아랍어에서만 사용되는 기능어의 특징을 규칙으로 적용하였다.
- [0011] Ramanathan et al.,(2009)에서는 영어-인도 기계 번역에서 case-marker를 사용하는 인도어의 특성을 중점으로 하는 방법을 적용하였다. Lee et al.,(2006)에서는 한국어-영어 기계 번역에서 품사 정보와 한국어 형태소 분석, 문장의 어순 변경, 한국어 기능어 삭제 등의 방법을 적용하였다.
- [0012] 이런 방법들이 통계적 단어 정렬-기계 번역의 품질을 향상시키고 있지만, 특정 언어에 종속적이라는 단점이 존재한다. 이는 번역하고자 하는 단어 쌍의 변경에 유연하지 않고 단어 쌍이 변경되면 단어 정렬 및 기계 번역 시스템을 교체해야 하는 상황이 야기될 수 있다.

### 발명의 내용

## 해결하려는 과제

[0013] 본 발명은 상기의 문제점을 해결하기 위해 창안된 것으로서, 소스 언어와 타겟 언어의 양국어 말뭉치로부터 서로 대응되는 단어를 추출하는 단어 정렬 시 각 단어의 품사 정보를 반영하여 품사간 관계가 존재하는 정렬 경향을 단어 정렬 스코어에 적용하여 기존 단어 정렬 스코어를 수정으로써 수정된 단어 정렬 스코어로 다시 단어를 재정렬하여 기존 단어 정렬보다 정확하고 높은 품질의 결과를 제공하는 통계적 단어 정렬 방법 및 이를 이용한 기계 번역 장치를 제공하는 것을 목적으로 한다.

## 과제의 해결 수단

[0014] 이를 위하여, 본 발명의 제1 측면에 따른 장치는, 통계적 기계 번역 시스템에서 단어 정렬을 위한 장치로서, 번역할 대상 언어인 소스 문장과 상기 소스 문장을 원하는 언어로 번역한 타겟 문장을 수신한 양국어 말뭉치를 형태소로 세그먼트(segment)한 각 단어에 대하여 대응 가능한 소스 단어와 타겟 단어를 정렬하는 단어 정렬기; 상기 단어 정렬기를 통해 정렬된 양국어 말뭉치로부터 각 단어의 품사 정보를 적용하여 품사간 정렬 경향을 추출하는 품사간 정렬 경향 추출기; 상기 품사간 정렬 경향 추출기로부터 추출된 정렬 경향을 반영한 단어 정렬 확률과 상기 단어 정렬기를 통해 정렬된 이전 단어 정렬 확률을 비교하여 양 스코어가 서로 다르면 상기 품사간 정렬 경향 추출기로부터 추출된 정렬 경향을 반영한 단어 정렬 확률로 수정하고 수정된 현 단어 정렬 확률에서 가장 높은 스코어를 갖는 단어 쌍으로 수정하여 재정렬하는 단어 재정렬기를 포함하는 것을 특징으로 한다.

[0015] 한편, 본 발명의 제2 측면에 따르면, 본 발명의 단어 정렬 방법은, 통계적 기계 번역에서 단어 정렬을 위한 방법으로서,

[0016] 번역할 대상 언어인 소스 문장과 상기 소스 문장을 원하는 언어로 번역한 타겟 문장을 수신한 양국어 말뭉치를 형태소로 세그먼트(segment)한 각 단어에 대하여 대응 가능한 소스 단어와 타겟 단어를 1차 정렬하는 단계; 상기 1차 정렬된 양국어 말뭉치로부터 각 단어의 품사 정보를 적용하여 품사간 정렬 경향을 추출하는 단계; 상기 추출된 품사간 정렬 경향을 상기 1차 정렬을 통해 획득한 단어 정렬 확률에 반영하여 단어 정렬 스코어를 재계산하는 단계; 상기 1차 정렬을 통해 획득한 기존 단어 쌍을 상기 재계산된 단어 정렬 스코어 중 가장 높은 스코어를 갖는 단어 쌍으로 수정하여 재정렬하는 단계를 포함하는 것을 특징으로 한다.

## 발명의 효과

[0017] 일반적으로 단어 정렬은 단일어 대역어 추출, 복합어 인식, 의미 중의성 해소, 번역 규칙 추출, 구 단위 대역어 추출, 자동 사전 구축 등 많은 자연어 처리 분야에서 사용된다. 본 발명은 단어의 품사간 관계를 반영하여 단어 재정렬을 수행함으로써 기존에 사용되고 있는 단어 정렬보다 정확하고 높은 품질의 단어 정렬 결과를 얻을 수 있는 효과가 있다. 이로 인해, 높은 품질의 자동 번역기, 자동 사전 구축 등을 가능하게 하는 효과를 얻을 수 있다.

## 도면의 간단한 설명

[0018] 도 1은 본 발명의 실시 예에 따른 기계 번역 장치를 나타낸 구성도.

도 2는 본 발명의 실시 예에 따른 기계 번역에 적용되는 IBM 모델을 통해 획득한 단어 정렬의 예를 보인 도면.

도 3 및 도 4는 본 발명의 실시 예에 따른 단어 정렬에서 품사간 정렬 경향에 의해 단어 정렬이 수정된 예를 보인 도면.

도 5는 본 발명의 실시 예에 따른 단어 정렬 방법을 설명하기 위한 순서도.

## 발명을 실시하기 위한 구체적인 내용

[0019] 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.

[0020] 본 발명을 설명하기에 앞서, 본 발명은 서로 다른 단어간의 정렬에서 서로 유사한 품사끼리 정렬되는 경향이 존재함을 가정한다. 이러한 가정은 반대로 전혀 유사하지 않은 품사끼리는 정렬되지 않는 경향이 존재한다고 할 수 있다. 예를 들어 한국어 품사 고유명사는 영어 품사 고유명사(proper noun)와 정렬되는 경향이 높지만, 반대

로 영어 품사 부사(adverb)와는 정렬되는 경향이 낮다. 이처럼, 정렬하고자 하는 단어 쌍의 종류에 관계없이 모든 단어에는 품사가 존재하며, 이에 따라 서로 다른 단어의 품사들 간에는 관계성이 존재한다. 이러한 관계성을 경향 또는 경향성이라 한다.

- [0021] 또한, 이하에서 언급하는 소스 문장(source sentence) 또는 소스 언어 문장은 번역할 대상이 되는 원 문장이고, 타겟 문장(target sentence) 또는 타겟 언어 문장은 소스 문장을 원하는 언어로 번역한 문장이다.
- [0022] 도 1은 본 발명의 실시 예에 따른 기계 번역 장치를 나타낸 구성도이다.
- [0023] 본 발명의 실시 예에 따른 기계 번역 장치(100)는 단어 정렬기(110), 품사간 정렬 경향 추출기(120), 단어 재정렬기(130), 번역에 필요한 정보를 저장한 데이터베이스로서 양국어 말뭉치 DB(140), 양국어 품사 DB(150), 단어 정렬 스코어 DB(160), 정렬된 양국어 문장 DB(170)를 포함하여 구현될 수 있다.
- [0024] 양국어 말뭉치 DB(140)는 소스 문장(source sentence)과 번역된 타겟 문장(target sentence)에 대한 병렬 말뭉치(parallel corpus)를 포함한다. 이러한 양국어 말뭉치 DB(140)는 단어 정렬기(110)로 소스 문장과 소스 문장과 대응되는 타겟 문장을 제공한다.
- [0025] 양국어 품사 DB(150)는 소스 문장과 타겟 문장의 각 단어에 대한 품사 및 원형 정보를 저장한다.
- [0026] 단어 정렬 스코어 DB(160)는 소스 문장과 타겟 문장간 각 단어에 대하여 대응 가능한 쌍(pair)과 이의 대응 확률, 대응 확률에 따라 최적으로 매핑될 수 있는 쌍 등에 대한 정보를 포함한다.
- [0027] 정렬된 양국어 문장 DB(170)는 소스 문장과 타겟 문장간 병렬 말뭉치에서 각 단어의 특성(공기 정보, 위치 정보, 품사 정보도 포함)을 고려하여 정렬된 문장을 저장한다. 이는 임시 저장도 가능하지만, 추후에 대역 사전을 갱신하기 위한 자료로 활용할 수 있다.
- [0028] 단어 정렬기(110)는 IBM 모델을 적용하여 양국어 말뭉치 DB(140)로부터 제공받은 소스 문장의 단어를 타겟 문장의 단어로 정렬한 결과를 획득하고, 또한 양국어 말뭉치 DB(140)로부터 제공받은 타겟 문장의 단어를 소스 문장의 단어로 정렬한 결과를 획득한다. 이러한 결과로 소스 언어에 적용된 단어의 특징 정보와 타겟 언어에 적용된 단어의 특징 정보에 따라 다수의 단어 정렬 결과를 1차적으로 얻는다. 얻은 결과는 정렬된 양국어 문장 DB(170)에 저장한다.
- [0029] 상기에서, 단어라 함은 형태소 단위로 세그먼트(segment)한 단어를 의미한다.
- [0030] 이처럼, 단어 정렬기(110)에서 적용되는 IBM 모델은 단어 정렬을 위한 알고리즘이다. 예컨대, IBM 모델을 통해 획득한 단어 정렬의 예를 도 2에 도시하였다.
- [0031] 도 2를 참조하면, 위 단에 표기된 영어 문장에서 visiting(Verb: VB) 단어는 한국어 형태소 매장(NNIN2), 찾(VBMA), 여성(NNIN2), 고객(NNIN2)과 정렬된다. 실제로 visiting(VB) 단어는 찾(VBMA)으로만 정렬되어야 하며, 그 외에 정렬된 쌍(pair)은 IBM Model을 통한 단어 정렬의 오류이다. 오류로 나타난 정렬 결과 중 visiting(VB): 매장(NNIN2)의 경우 영어 단어의 품사 동사(VB)와 한국어 형태소의 품사 고유명사(NNIN2)의 관계성은 정렬되지 않은 store(NN):매장(NNIN2)에서 사용된 품사 명사(NN):고유명사(NNIN2)의 관계성보다 낮다. 즉, 단어 정렬에서, 일반적인 통계학적 정보뿐만 아니라 유사한 품사들간의 관계성을 고려하는 것이 보다 정확한 단어 정렬 결과를 얻을 수 있다는 것을 생각할 수 있다. 이렇듯 품사들간의 관계성은 IBM 모델만을 사용한 단어 정렬 결과에서 추출할 수 있다.
- [0032] 다시 도 1을 참조하면, 품사간 정렬 경향 추출기(120)는 단어 정렬기(110)를 통해 정렬된 결과에서 품사들간의 관계성 즉, 품사간 정렬 경향을 추출한다. 이때,
- [0033] 추출하는 방법으로는 두 가지가 있다.
- [0034] 첫 번째 방법은, 특정 언어 품사가 병렬 말뭉치에서 출현하였을 때 해당 품사가 특정 타겟 품사로 정렬되는 확률로 측정하는 것이다. 이 방법을 수식으로 나타내면 아래와 같다.

## 수학식 1

$$P(\text{pos}(e) | \text{pos}(f)) = \frac{\text{aligncount}(\text{pos}(e); \text{pos}(f))}{\text{count}(\text{pos}(e))}$$

[0035]

[0036]  $\text{pos}(e)$ 는 소스 단어  $e$ 의 품사를 의미하며,  $\text{pos}(f)$ 는 타겟 단어  $f$ 의 품사를 의미한다. 각 단어의 품사 정보는 양국어 품사 정보 DB(150)로부터 제공받을 수 있다.

[0037] 위 수학식 1은 소스 언어의 품사가 병렬 말뭉치에서 출현할 때 특정 타겟 언어의 품사와 정렬되는 확률을 측정하는 것이다. 즉, 품사들간의 상대 빈도를 측정하는 것이다.  $\text{count}$ 는 특정 품사가 병렬 말뭉치 내에서 출현한 빈도를,  $\text{aligncount}$ 는 병렬 말뭉치 내에서 특정 두 품사가 정렬된 빈도를 나타내는 함수이다.

[0038] 위 수학식 1과 유사하게, 타겟 단어  $f$ 의 품사가 병렬 말뭉치에서 출현할 때, 특정 소스 단어  $e$ 의 품사와 정렬되는 상대 빈도는 아래의 수학식 2와 같이 측정할 수 있다.

## 수학식 2

$$P(\text{pos}(e) | \text{pos}(f)) = \frac{\text{aligncount}(\text{pos}(f); \text{pos}(e))}{\text{count}(\text{pos}(f))}$$

[0039]

[0040] 이와 같이 구한 상대 빈도를 IBM Model을 통하여 얻어지는 정렬 대응 확률과 결합하여, 아래의 수학식 3과 같이 품사간 정렬 경향이 반영된 단어 정렬 스코어를 계산한다.

## 수학식 3

$$\text{Score}(f, e) = P_{\text{IBM}}(f, e) \times P(\text{pos}(e) | \text{pos}(f)) \times P(\text{pos}(f) | \text{pos}(e))$$

[0041]

[0042] 여기서,  $P_{\text{IBM}}$ 은 IBM 모델에서 계산한 두 단어간의 정렬 대응 확률이다.

[0043] 위 수학식 3을 통해 계산된 단어 정렬 스코어는 단어 정렬 스코어 DB(160)로 전달하여 저장 및 갱신할 수 있다.

[0044] 두 번째 방법으로는, 소스 언어의 품사가 타겟 언어의 품사와 동일한 병렬 문장에서 나타났을 때, 해당 품사 쌍이 정렬되는 확률로 측정하는 것으로 최대 가능도(Maximum Likelihood Estimation) 방법에 의해 아래와 같이 구할 수 있다. 최대 가능도는 어떤 모집단의 모수(population parameter)에 관한 가장 적절한 추정치(Estimate)를 구하는 방법이다.

## 수학식 4

$$P(\text{align} = \text{TRUE} | \text{pos}(f), \text{pos}(e)) = \frac{\text{count}(\text{align} = \text{true} | \text{pos}(f), \text{pos}(e))}{\sum_{k \in (\text{true}, \text{false})} \text{count}(\text{align} = k, \text{pos}(f), \text{pos}(e))}$$

[0045]

[0046] 위 수학식 4에서 구한 품사간 정렬 경향성을 IBM 모델을 통해 얻은 정렬 대응 확률과 선형 결합(linear combination)하면 아래의 수학식 5를 얻을 수 있다.

## 수학식 5

$$\text{Score}(f;e) = \lambda P_{\text{IBM}}(f|e) + (1-\lambda) P(\text{align} = \text{TRUE} | \text{pos}(f), \text{pos}(e))$$

[0047]

[0048] 위 수학식 5를 통해 계산된 결과를 단어 정렬 스코어로서 추출하며, 추출한 단어 정렬 스코어는 단어 정렬 스코어 DB(160)에 전달하여 저장 및 갱신한다.

[0049] 단어 재정렬기(130)는 상기의 품사간 정렬 경향 추출기(120)를 통해 추출된 단어 정렬 스코어를 기반으로 IBM 모델의 단어 정렬 결과를 수정함으로써 단어 재정렬한다.

[0050] 단어 정렬을 수정하는 방법(rull)은 다음과 같다.

[0051] 1) 병렬 문장이 가지는 모든 정렬 링크에 대해 수행한다.

[0052] 2) 현재 정렬된 링크  $f_1 : e_m$  의  $\text{Score}(f_1, e_m)$  보다 높은  $\text{Score}(f_1, e_k)$  를 가지는 k번째의 소스 단어 중에서 가장 높은 Score를 가지는 단어로 정렬 링크를 수정한다.  $f_1 : e_m \rightarrow f_1 : e_k$

[0053] 3) 이 과정을 모든 병렬 문장에 대해 반복한다.

[0054] 도 3 및 도 4에 도시한 예를 참조하면,  $e_1$  내지  $e_4$  는 소스 단어이고,  $f_1$  내지  $f_4$  타겟 단어를 의미하며 소스 단어와 타겟 단어간 연결된 선은 소스 단어와 타겟 단어간 대응됨을 나타내는 정렬선을 의미한다.

[0055] 먼저 도 3에 도시한 바와 같이, IBM 모델의 단어 정렬 결과가 존재할 때 각 단어 쌍의 정렬 확률은 아래의 표 1과 같이 나타날 수 있다.

표 1

[0056]

| 단어 쌍        | 정렬 확률 |
|-------------|-------|
| $f_1 : e_1$ | 0.4   |
| $f_1 : e_2$ | 0.1   |
| $f_4 : e_3$ | 0.3   |
| $f_4 : e_4$ | 0.2   |

[0057] 위 표 1에서 정렬 확률이 높은 것으로 쌍을 이루면  $f_1$  은  $e_1$  과 정렬되고,  $f_4$  는  $e_3$  과 정렬된다. 이 정렬 확률에 품사간 정렬 경향 추출기(120)에 추출한 단어 정렬 스코어를 반영하면 도 4와 같이 수정된다.

[0058] 즉, 도 4를 참조하면 품사간 정렬 경향을 반영한 정렬 확률은 아래의 표 2와 같이 나타나며, 이에 따라 1차 단어 정렬된  $f_1 : e_1$  링크는  $f_1 : e_2$  로 수정되고,  $f_4 : e_3$  링크는  $f_4 : e_4$  링크로 수정된다.

표 2

[0059]

| 단어 쌍        | 정렬 확률 |
|-------------|-------|
| $f_1 : e_1$ | 0.1   |
| $f_1 : e_2$ | 0.3   |
| $f_4 : e_3$ | 0.05  |
| $f_4 : e_4$ | 0.2   |

[0060] 이렇게 수정(재정렬)된 양국어 쌍은 정렬된 양국어 문장 DB(170)에 저장하여 추후에 대역 사전을 갱신하기 위한 자료로 활용할 수 있다.

[0061] 이렇게 구성되는 기계 번역 장치(100)를 이용하여 단어 정렬에 의해 번역하는 방법에 대하여 도 5를 참조하면 설명한다.

[0062] 먼저, 양국어 말뭉치 DB로부터 수신한 소스 문장 또는 타겟 문장에 IBM 모델을 적용하여 각 단어에 대하여 대응

되는 쌍을 찾아 단어 정렬한다(S110).

IBM 모델은 단어의 공기 정보 및 위치 정보 등의 특성을 고려하여 정렬 확률이 큰 링크로 정렬한다.

이후, 정렬된 양국어 말뭉치로부터 품사간 정렬되는 경향을 추출한다(S120). 추출 방법은 양국어 말뭉치에서 출현한 각 단어의 품사가 타겟 품사와 정렬되는 상대 빈도 확률을 구하여 추출하는 방법, 또는 소스 문장의 단어 품사와 타겟 문장의 단어 품사가 동일 양국어 말뭉치에 출현하였을 때 해당 단어 쌍이 정렬되는 확률을 구하여 추출하는 방법 중 어느 하나를 적용할 수 있다.

이후, 추출한 품사간 정렬 경향을 IBM 모델을 통해 획득한 단어 정렬 확률에 반영하여 단어 정렬 스코어를 재계산한다(S130).

이후, 재계산된 단어 정렬 스코어를 기반으로 양국어 문장을 재정렬한 후 번역을 수행한다(S140).

재정렬하는 방법으로는, 품사간 정렬 경향을 반영한 단어 정렬 스코어와 현재 정렬된 링크의 스코어 즉, IBM 모델을 통해 정렬된 링크의 스코어를 비교하여 양 스코어가 서로 다르면 품사간 정렬 경향을 반영한 단어 정렬 스코어로 수정하고 수정된 스코어에서 가장 높은 스코어를 갖는 단어 쌍으로 수정하여 재정렬하는 방식을 기본적으로 한다. 그리고, 이러한 방식을 병렬 문장이 가지는 모든 정렬 링크에 대해 수행하며, 크게는 모든 병렬 문장에 대해 반복한다.

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.

다양한 재생장치는 PC, 노트북, 휴대용 단말 등일 수 있다.

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.

이 경우, 컴퓨터로 읽을 수 있는 기록 매체에 기록한 프로그램은, 앞서 설명한 바와 같이 번역할 대상 언어인 소스 문장과 상기 소스 문장을 원하는 언어로 번역한 타겟 문장을 수신한 양국어 말뭉치를 형태소로 세그먼트(segment)한 각 단어에 대하여 대응 가능한 소스 단어와 타겟 단어를 1차 정렬하는 과정과, 1차 정렬된 양국어 말뭉치로부터 각 단어의 품사 정보를 적용하여 품사간 정렬 경향을 추출하는 과정과, 상기 과정에서 추출된 품사간 정렬 경향을 상기 1차 정렬을 통해 획득한 단어 정렬 확률에 반영하여 단어 정렬 스코어를 재계산하는 과정, 1차 정렬을 통해 획득한 기존 단어 쌍을 상기 재계산된 단어 정렬 스코어 중 가장 높은 스코어를 갖는 단어 쌍으로 수정하여 재정렬하는 과정을 포함하여 실행될 수 있다.

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

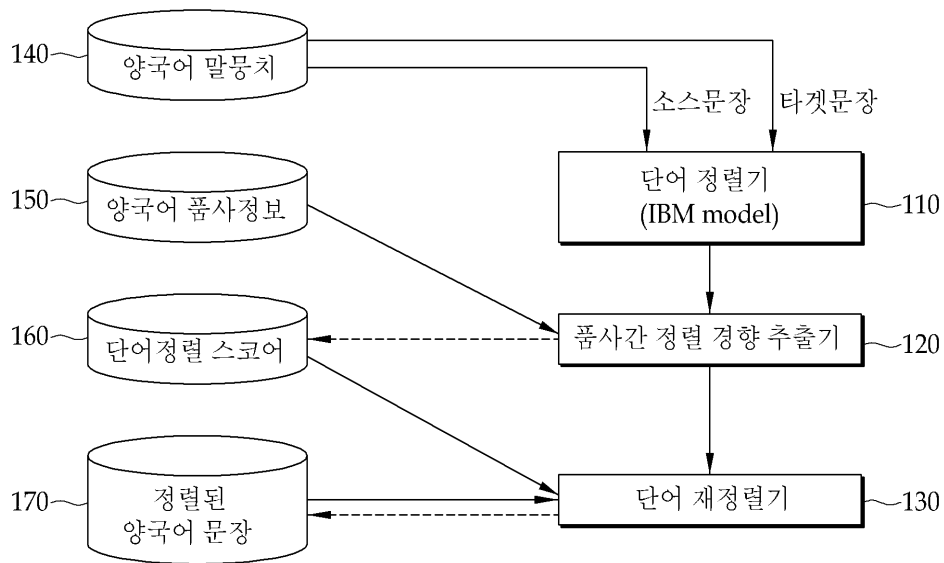
종래 기술에 따른 단어 정렬 및 통계적 기계 번역은 IBM 모델을 통해 단어의 공기 정보 및 위치 정보와 같은 특성만 고려하여 수행하였으나, 본 발명은 각 단어의 품사 정보를 반영하여 단어들의 품사간 의존하는 경향이 높은 경우까지 고려함으로써 기존에 사용하고 있는 IBM 모델로부터 획득한 단어 정렬의 오류를 해소함으로써 정확하고 높은 품질의 단어 정렬 결과를 얻을 수 있으며, 나아가 자동 번역기 또는 자동 사전의 구축에 있어서 번역의 품질을 향상시킬 수 있다.

## 부호의 설명

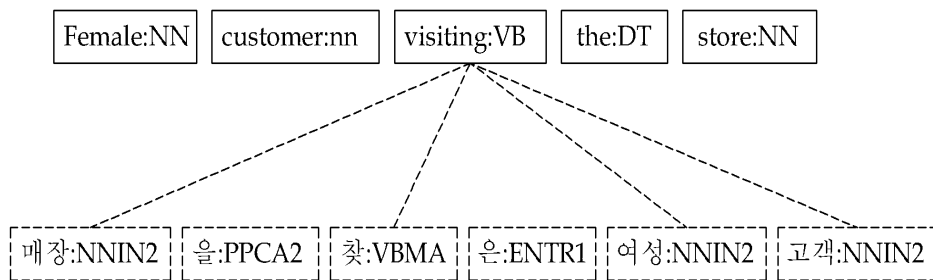
|                    |                    |
|--------------------|--------------------|
| 100: 기계 번역 장치      | 110: 단어 정렬기        |
| 120: 품사간 정렬 경향 추출기 | 130: 단어 재정렬기       |
| 140: 양국어 말뭉치 DB    | 150: 양국어 품사 정보 DB  |
| 160: 단어 정렬 스코어 DB  | 170: 정렬된 양국어 문장 DB |

도면

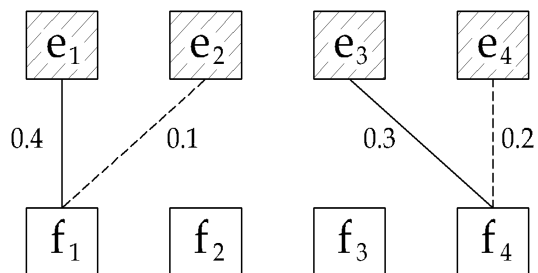
도면1



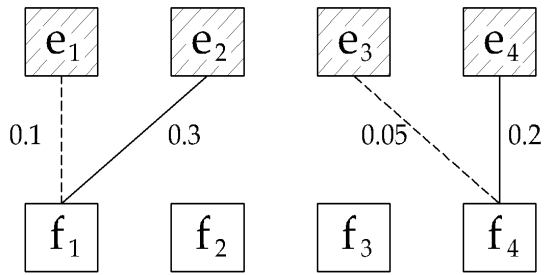
도면2



도면3



도면4



도면5

