(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0014126
(43) 공개일자 2013년02월07일

(51) 국제특허분류(Int. Cl.)
G06F 17/28 (2006.01) *G06F 17/27* (2006.01)
(21) 출원번호 10-2011-0076064
(22) 출원일자 2011년07월29일
심사청구일자 2012년06월04일
기술이전 희망 : **기술양도**

(71) 출원인
울산대학교 산학협력단
울산 남구 무거2동 산29
(72) 발명자
옥철영
울산광역시 남구 신복로35번길 6, 101동 602호 (무거동, 쌍용스윗닷홈)
신준철
울산광역시 남구 신복로31번길 9-1, 208호 (무거동)
김홍준
울산광역시 동구 화진길 9 (방어동)
(74) 대리인
김종선

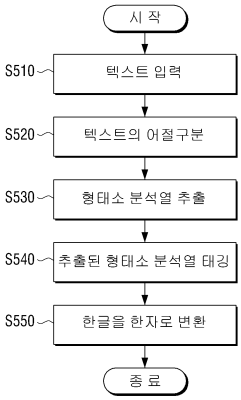
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 **한글-한자 변환 방법 및 이를 적용한 단말 장치**

(57) 요 약

한글-한자 변환 방법 및 이를 적용한 단말 장치가 제공된다. 본 한글-한자 변환 방법은, 한글로 이루어진 텍스트의 어절을 구분하고, 구분된 어절별로 적어도 하나의 형태소 분석열을 추출하며, 추출된 적어도 하나의 형태소 분석열을 태깅하고, 태깅 결과를 이용하여, 적어도 하나의 형태소 분석열을 한자로 변환한다. 이에 의해, 동형이의어가 존재하는 한글 텍스트에서도 문맥에 맞는 한글-한자 변환이 가능해 질 수 있다.

대 표 도 - 도5



특허청구의 범위

청구항 1

한글로 이루어진 텍스트의 어절을 구분하는 단계;

상기 구분된 어절별로 적어도 하나의 형태소 분석열을 추출하는 단계;

상기 추출된 적어도 하나의 형태소 분석열 및 동형이의어를 태깅하는 단계;

상기 태깅 결과를 이용하여, 상기 적어도 하나의 형태소 분석열을 한자로 변환하는 단계;를 포함하는 한글-한자 변환 방법.

청구항 2

제1항에 있어서,

상기 추출하는 단계는,

어절, 어절의 형태소별 품사 및 형태소별 품사 및 동형이의어의 빈도 정보를 포함하는 제1 데이터베이스로부터 상기 어절에 대응되는 상기 형태소 분석열을 추출하는 것을 특징으로 하는 한글-한자 변환 방법.

청구항 3

제2항에 있어서,

상기 태깅하는 단계는,

상기 추출된 적어도 하나의 형태소 분석열의 생성 확률 및 전이 확률을 이용하여 상기 적어도 하나의 형태소 분석열을 태깅하는 것을 특징으로 하는 한글-한자 변환 방법.

청구항 4

제3항에 있어서,

상기 생성 확률은,

상기 제1 데이터베이스로부터 추출되는 상기 형태소 분석열별 빈도 정보를 이용하여 결정되는 것을 특징으로 하는 한글-한자 변환 방법.

청구항 5

제3항에 있어서,

상기 전이 확률은,

제2 데이터베이스에 기 저장된 형태소별 품사의 빈도 정보를 이용하여 결정되는 것을 특징으로 하는 한글-한자 변환 방법.

청구항 6

제3항에 있어서,

상기 태깅하는 단계는,

은닉마르코프모델(HMM: Hidden Markov Model)을 이용하여 어절 간의 발생 빈도에 따라 전이 확률에 가중치를 적용하는 것을 특징으로 하는 한글-한자 변환 방법.

청구항 7

제1항에 있어서,

상기 태깅하는 단계는

비터비(Viterbi) 알고리즘을 이용하여 상기 어절에 대해 각각의 형태소 분석열을 태깅하는 것을 특징으로 하는 한글-한자 변환 방법.

청구항 8

제1항에 있어서,

한글 형태소 분석열의 태깅 결과와 한자가 대응되어 기저장되어 있는 것을 특징으로 하는 한글-한자 변환 방법.

청구항 9

한글로 이루어진 텍스트의 어절을 구분하는 어절 구분부;

상기 구분된 어절별로 적어도 하나의 형태소 분석열을 추출하는 형태소 분석열 추출부;

상기 추출된 적어도 하나의 형태소 분석열 및 동형이의어를 태깅하는 태깅부;

상기 태깅부의 태깅 결과를 이용하여, 상기 적어도 하나의 형태소 분석열을 한자로 변환하는 변환부;를 포함하는 한글-한자 변환을 위한 단말 장치.

명 세 서

기 술 분 야

[0001] 본 발명은 한글-한자 변환 방법 및 이를 적용한 단말 장치에 관한 것으로, 더욱 상세하게는, 동형이의어의 태깅 결과를 이용하여 한글을 한자로 변환하는 한글-한자 변환 방법 및 이를 적용한 단말 장치에 관한 것이다.

배 경 기 술

[0002] 한국어의 어휘는 순수한 고유어와 한자어(漢字語) 그리고 차용어(借用語)로 구성되어 있다. 한국어 어휘에 포함된 한자어는 중국과의 문화적 접촉으로 인해 우리말화된 것으로, 표준국어대사전의 등재된 507,000여개의 표제어 중 288,600 어휘(56.92%)가 한자를 가지고 있다. 비록, 한자를 사용하지 않더라도 의사소통을 하는데 특별한 문제가 없다고 해도, 전문용어나 학술용어의 대부분이 한자어로 되어 있어 한글로만 표기할 경우에는 정확한 의미를 이해하기 어렵다.

[0003] 또한, 288,600여개의 한자어 어휘 중에서 84,000여개가 동형이의어이다. 이러한 동형이의어의 경우는 경우에 따라 한자를 사용함으로써 보다 정확한 의미 전달에 도움을 줄 수 있다. 예를 들어 “이 회사의 사기가 참 화려하네...” 라는 문장에서 ‘사기’ 는 표준국어대사전에 27개 동형이의어가 등재되어 있으며, 이 중에서도 史記(역사적 사실을 기록한 책), 私記(사사로운 기록), 事記(사건의 기록), 社基(회사의 기초), 社旗(회사를 상징하는 깃발) 등의 해석이 가능하다. 그렇지만 “이 회사(會社)의 사기(社旗)가 참 화려하네...” 라고 병기 혹은 변환해 두면 그 의미 전달이 보다 명확해진다.

[0004] 한편, 종래에는 한국어 형태소 태깅과 동형이의어 태깅에 대한 많은 연구들이 있었다. 구체적으로, 형태소 태깅을 위한 모델은 크게 규칙 기반 모델과 통계기반 모델이 있다. 규칙 기반 모델에서는 언어 정보를 생성 규칙의 형태로 표현하고 이를 적용하여 태깅을 수행한다. 규칙 기반 모델에서는 규칙이 적용되었을 경우에 대해서 높은 정확도로 태깅을 수행하지만, 규칙 구축에 많은 시간과 노력이 요구되며 자연언어에서 발생하는 광범위한 현상을 처리하기 어렵다는 단점이 있다. 반면 통계적 접근법은 충분한 크기의 태그 부착 말뭉치만 주어지면 태깅에 필요한 통계 정보와 추출이 용이하기 때문에 확장성이 좋고 적용 범위가 넓으며 전체적인 정확성이 높지만, 말뭉치 구축에 시간과 노력이 많이 요구되고, 말뭉치가 일정 크기 이상 구축되어 있지 않을 경우 통계 자료 부족으로 인해 신뢰도가 떨어진다는 단점이 있다. 최근에는 규칙 기반 모델과 통계 기반 모델을 결합하여 서로 간의 결점을 보완하는 복합적 모델을 사용하는 추세이다.

[0005] 또한, 동형이의어 태깅 모델은 사전의 뜻풀이 말뭉치에서 구축한 의미정보와 이를 적용한 베이지안 분류 모델을 이용하여 동형이의어를 태깅할 수 있었으며, 언어의 유형별로 동형이의어 중의성 해소 모형을 통해 중의성 해소 대상 단어의 의미별 출현 비율과 중의성 해소에 결정적인 역할을 하는 단서어의 출현위치에 따라 동형이의어를 태깅할 수 있었다.

[0006] 이렇듯 다양한 형태소 태깅 모델 및 동형이의어 태깅 모델을 이용하여, 중의적 표현의 한글을 포함하는 문맥의 해석이 가능하게 되었다.

발명의 내용

해결하려는 과제

[0007] 본 발명의 목적은 형태소 태깅 시스템을 이용하여, 한글 텍스트의 문맥에 맞도록 한글을 한자로 변환하는 한글-한자 변환 방법 및 이를 위한 단말 장치를 제공함에 있다.

과제의 해결 수단

[0008] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른, 한글-한자 변환 방법은, 한글로 이루어진 텍스트의 어절을 구분하는 단계; 상기 구분된 어절별로 적어도 하나의 형태소 분석열을 추출하는 단계; 상기 추출된 적어도 하나의 형태소 분석열 및 동형이의어를 태깅하는 단계; 및 상기 태깅 결과를 이용하여, 상기 적어도 하나의 형태소 분석열을 한자로 변환하는 단계;를 포함한다.

[0009] 그리고, 상기 추출하는 단계는, 어절, 어절의 형태소별 품사 및 형태소별 품사 및 동형이의어의 빈도 정보를 포함하는 제1 데이터베이스로부터 상기 어절에 대응되는 상기 형태소 분석열을 추출할 수 있다.

[0010] 또한, 상기 태깅하는 단계는, 상기 추출된 적어도 하나의 형태소 분석열의 생성 확률 및 전이 확률을 이용하여 상기 적어도 하나의 형태소 분석열을 태깅할 수 있다.

[0011] 그리고, 상기 생성 확률은, 상기 제1 데이터베이스로부터 추출되는 상기 형태소 분석열별 빈도 정보를 이용하여 결정될 수 있다.

[0012] 또한, 상기 전이 확률은, 제2 데이터베이스에 기 저장된 형태소별 품사의 빈도 정보를 이용하여 결정될 수 있다.

[0013] 그리고, 상기 태깅하는 단계는, 은닉마르코프모델(HMM: Hidden Markov Model)를 이용하여 어절 간의 발생 빈도에 따라 전이 확률에 가중치를 적용할 수 있다.

[0014] 또한, 상기 태깅하는 단계는 비터비(Viterbi) 알고리즘을 이용하여 상기 어절에 대해 각각의 형태소 분석열을 태깅할 수 있다.

[0015] 그리고, 한글 형태소 분석열의 태깅 결과와 한자가 대응되어 기저장될 수 있다.

[0016] 한편, 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른, 한글로 이루어진 텍스트의 어절을 구분하는 어절 구분부; 상기 구분된 어절별로 적어도 하나의 형태소 분석열을 추출하는 형태소 분석열 추출부; 상기 추출된 적어도 하나의 형태소 분석열 및 동형이의어를 태깅하는 태깅부; 상기 태깅부의 태깅 결과를 이용하여, 상기 적어도 하나의 형태소 분석열을 한자로 변환하는 변환부;를 포함한다.

발명의 효과

[0017] 상술한 바와 같은 본 발명의 다양한 실시예에 의하면, 동형이의어가 존재하는 한글 텍스트에서도 문맥에 맞는 한글-한자 변환이 가능해 질 수 있다.

도면의 간단한 설명

[0018] 도 1은 본 발명의 일 실시예에 따른, 한글-한자 변환을 위한 단말 장치의 블럭도,

도 2는 본 발명의 일 실시예에 따른, 제1 데이터베이스 내의 정보를 예시한 도면,

도 3은 본 발명의 일 실시예에 따른, 제2 데이터베이스 내의 정보를 예시한 도면,

도 4는 본 발명의 일 실시예에 따른, 품사 및 동형이의어 태깅의 모델링을 나타낸 도면, 그리고,

도 5는 본 발명의 일 실시예에 따른, 한글-한자 변환 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0019] 이하에서는 도면을 참조하여, 본 발명에 대해 더욱 상세히 설명하도록 한다.

[0020] 도 1은 본 발명의 일 실시예에 따른, 한글-한자 변환을 위한 단말 장치(100)의 블럭도이다. 도 1에 도시된 바와 같이, 한글-한자 변환을 위한 단말 장치(100)는 어절 구분부(110), 형태소 분석열 추출부(120), 형태소 분석부

(130), 태깅부(140), 제1 데이터베이스(150), 제2 데이터베이스(160) 및 변환부(170)를 포함한다.

[0021] 어절 구분부(110)는 입력된 텍스트의 어절(語節)을 구분한다. 어절은 텍스트를 구성하고 있는 각각의 마디를 의미한다. 예를 들면, 어절 구분부(110)에 "나는 오전에 공부를 열심히 한다."라는 문장이 입력되었다면, 어절 구분부(110)에서는 이 문장을 5개의 어절 즉, "나는", "오전에", "공부를", "열심히", "한다"로 구분한다.

[0022] 이때, 어절 구분부(110)에 입력되는 텍스트는 어떤 사상이나 감정의 표현이 완결된 것을 의미하나, 표현이 완결되지 않은 단어(單語) 문장, 및 주어를 갖추지 않은 명령문과 같은 예외의 경우도 문장에 해당한다. 예를 들면, 어절 구분부(110)에는 "나는 오전에 공부를 열심히 한다", 및 "나는 참 행복한 인간이야"와 같이 주어와 서술어를 모두 갖춘 문장이 입력될 수 있고, 혹은 "물을 가지러 간 사이", 및 "발명을 실시하기 위한 구체적인 내용"과 같이 그 표현이 완결되지 않은 단어 문장이 입력될 수도 있다.

[0023] 형태소 분석열 추출부(120)는 어절 구분부(110)에 의해 구분된 어절별로 적어도 하나의 형태소 분석열을 추출한다. 형태소 분석열 추출부(120)는 형태소 분석열 추출을 위해 기생성된 제1 데이터베이스(150) 및 제2 데이터베이스(160)로부터 해당 형태소 분석열을 추출할 수 있다.

[0024] 이때, 제1 데이터베이스(150)는 어절, 그 어절의 형태소 분석열, 및 그 형태소 분석열의 빈도 정보를 포함한다. 또한, 제2 데이터베이스(160)는 인접하는 두 어절의 조합에 의한 어절 군, 그 어절 군의 형태소별 분석열, 및 그 형태소별 분석열의 빈도 정보를 포함한다. 제1 데이터베이스(150) 및 제2 데이터베이스(160)를 이용하여 형태소 분석열을 추출하는 방법에 대하여는 도 2 및 도 3를 참조하여 상세히 설명한다.

[0025] 도 2는 본 발명의 일 실시예에 따른 제1 데이터베이스(150) 내의 정보를 예시한 도면이다. 도 2에 도시된 바와 같이, 제1 데이터베이스(150)는 어절, 그 어절의 형태소 분석열, 및 그 형태소 분석열의 빈도 정보가 테이블 형태로 저장된 정보를 포함한다. 본 실시예에서는 학습 말뭉치에 포함되어 있는 어절을 이용하여 형태소 분석열 추출부(120)에서 해당 정보를 추출하기 용이한 형태로 정제하여 제1 데이터베이스(150)를 구성한다. 바람직하게, "21세기 세종계획 형태 의미 분석 말뭉치"를 이용하여 제1 데이터베이스(150)를 구성할 수 있다.

[0026] 도 2에 예시된 바와 같이, "한"이란 어절에 대하여 14개 혹은 그 이상의 형태소 분석열이 존재한다. 형태소 분석열은 "형태소/품사기호"의 형태로 구성될 수 있으며, 경우에 따라서는 "형태소__의미번호/품사기호"의 형태로 구성될 수 있다. 품사기호는 품사 태깅에 사용되는 것이고, 의미번호는 동형이의어 태깅에 사용된다. 여기서, 의미번호는 표준국어대사전의 어깨번호와 동일한 값을 갖도록 구성하였다. 또한, 하나의 어절에 대한 복수의 형태소 분석열은 그 각각의 출현 빈도가 다르며, 그 출현 빈도를 반영한 것이 빈도 정보이다.

[0027] 본 실시예에서 사용한 품사기호는 일반적으로 사용되는 품사 기호로, 표 3과 같다.

표 1

[0028]

대분류	소분류	세분류
체언	명사 NN	일반명사 NNG
		고유명사 NNP
		의존명사 NNB
	대명사 NP	대명사 NP
	수사 NR	수사 NR
용언	동사 VV	동사 VV
	형용사 VA	형용사 VA
	보조용언 VX	보조용언 VX
	지정사 VC	긍정지정사 VCP
		부정지정사 VCN
수식언	관형사 MM	
	부사 MA	일반부사 MAG
		접속부사 MAJ
독립언	감탄사 IC	감탄사 IC

관계언	격조사 JK	주격조사 JKS
		보격조사 JKC
		관형격조사 JKG
		목적격조사 JKO
		부사격조사 JKB
		호격조사 JKV
		인용격조사 JKQ
	보조사 JX	보조사 JX
	접속조사 JC	접속조사 JC
의존형태	어미 E	선어말어미 EP
		종결어미 EF
		연결어미 EC
		명사형전성어미 ETN
		관형형전성어미 ETM
	접두사 XP	체언접두사 XPN
	접미사 XS	명사파생접미사 XSN
		동사파생접미사 XSV
		형용사파생접미사 XSA
	어근 XR	어근 XR
기호	마침표, 물음표, 느낌표	SF
	쉼표, 가운뎃점, 콜론, 빗금	SP
	따옴표, 괄호표, 줄표	SS
	줄임표	SE
	붙임표(물결, 숨김, 빠짐)	SO
	외국어	SL
	한자	SH
	기타 기호(논리수학기호, 화폐기호 등)	SW
	명사추정범주	NF
	용언추정범주	NV
	숫자	SN
	분석불능범주	NA

[0029] 표 1에 예시한 바와 같이, 품사 기호를 "세종 말뭉치"에서 사용하는 품사 기호와 일치시킴으로써, 학습의 효율을 향상시킬 수 있다.

[0030] 도 3은 본 발명의 일 실시예에 따른, 제2 데이터베이스(160) 내의 정보를 예시한 도면이다. 도 3을 참조하면, 제2 데이터베이스(160)는 어절 군, 그 어절 군의 형태소 분석열, 및 그 형태소 분석열의 빈도 정보가 테이블 형태로 저장된 정보를 포함한다. 본 실시예에서는 학습 말뭉치에 포함되어 있는 어절 군을 이용하여 형태소 분석열 추출부(120)에서 해당 정보를 추출하기 용이한 형태로 정제하여 제2 데이터베이스(160)를 구성한다. 바람직하게, "21세기 세종계획 형태 의미 분석 말뭉치"를 이용하여 제2 데이터베이스(160)를 구성할 수 있다.

[0031] 도 3에 예시한 바와 같이, 제2 데이터베이스(160)에는 두 어절의 조합에 의한 어절 군의 다양한 형태에 따른 어절 군의 형태소 분석열 및 그 형태소 분석열의 빈도 정보가 저장되어 있다. 도시한 바와 같이, 어절 군의 형태소 분석열은 "형태소__의미번호/품사기호"와 같은 형태로 구성된다.

[0032] 도 2에서 설명한 제1 데이터베이스(150)와 마찬가지로, 제2 데이터베이스(160)에 저장된 어절 군의 형태소 분석열에 사용된 의미번호는 표준국어대사전의 어깨번호와 동일한 값을 갖으며, 품사 기호는 표 3에 예시한 바와 같다.

[0033] 다시 도 1에 대해 설명하면. 형태소 분석부(130)는 형태소 분석열 추출부(120)에서 제1 데이터베이스(150)로부터 어절에 해당하는 형태소 분석열을 추출할 때, 해당 형태소 분석열이 존재하지 않을 경우에만 동작하는 것으로, 해당 어절의 형태소를 분석하고, 그 분석 결과에 따라 형태소 분석열을 생성한다. 이때, 형태소 분석부(130)는 UTagger 프로그램을 이용할 수 있다. UTagger 프로그램은 어절 사전을 기반으로 하여 빠른 형태소 분석이 가능하고, 어절에 대한 빈도 정보를 갖고 있으므로, 한국어의 후처리 단계에서 효과적으로 이용될 수 있는 형태소 분석기이다.

[0034] 태깅부(140)는 형태소 분석열 추출부(120)에서 추출된 형태소 분석열의 생성 확률 및 전이 확률을 산출하고, 산출된 생성 확률 및 전이 확률을 이용하여 각 어절에 대해 최적의 형태소 분석열 및 동형이의어를 태깅한다.

[0035] 여기서, 생성 확률은 제1 데이터베이스(150)로부터 추출되는 형태소 분석열별 빈도 정보를 이용하여 결정되며, 전이 확률은 제2 데이터베이스(160)에 기 저장된 형태소별 품사 및 동형이의어의 빈도 정보를 이용하여 결정될 수 있다.

[0036] 이때, 태깅부(140)는 은닉마르코프모델(HMM: Hidden Markov Model)을 이용하여 어절 간의 발생 빈도에 따라 전이 확률에 가중치를 적용할 수 있다. 구체적으로, 태깅부(140)는 기결정된 전이 확률 중, 어절 군의 전체 빈도 정보 대비 해당 어절 군의 형태소 분석열의 빈도 정보가 기설정된 임계값 이하이면, 해당 어절 군의 형태소 분석열의 전이 확률을 제거할 수 있다. 이는 해당 어절 군의 출현 빈도가 극히 적을 경우에는 이를 말뭉치의 오류로 판단하고, 그 어절 군의 형태소 분석열은 최종 형태소 분석열로 태깅되지 않도록 하기 위한 것이다.

[0037] 구체적으로, 한국어의 특성상 어절 간의 조합(어절 군)에서 태깅 결과의 경우의 수가 너무 많기 때문에, 제2 데이터베이스(160)의 정보만으로는 신뢰성 있는 전이확률을 얻을 수 없다. 이를 보완하기 위하여, 추가적으로 의미없는 정보를 제거하여 학습 말뭉치의 신뢰성을 높이기 위하여 태깅부(140)에서 해당 어절 군의 출현 빈도가 극히 적은 형태소 분석열의 전이 확률을 제거할 수 있다. 이때, 확률 결정부(160)에서는 정보검색 분야의 TF/IDF(Term Frequency/Inverse Document Frequency) 원리를 응용할 수 있다. 하기의 표 2 및 표 3을 참조하여 설명한다.

표 2

[0038]

A 어절 + B 어절
A1 + B1 태깅 : 1000회 출현
A1 + B2 태깅 : 500회 출현
A1 + B3 태깅 : 300회 출현
A2 + B1 태깅 : 1회 출현
…

표 3

[0039]

C 어절 + D 어절
C1 + D1 태깅 : 10회 출현
C1 + D2 태깅 : 5회 출현
C1 + D3 태깅 : 3회 출현
C2 + D1 태깅 : 1회 출현
…

[0040] 표 2을 참조하면, 어절 군이 "A 어절 + B 어절"일 경우, "A2 + B1"의 형태는 다른 형태에 비해 현저히 낮은 출현율을 보인다. 이런 경우의 대부분은 말뭉치의 오류에 해당한다.

[0041] 반면, 표 3를 참조하면, 어절 군이 "C 어절 + D 어절"일 경우, 표 2에서의 "A2 + B1"와 같이 "C2 + D1"이 1회 출현하였다. 하지만, "C 어절 + D 어절"의 경우는 전체적으로 그 출현 빈도가 낮으므로, 1회 출현한 "C2 + D1"도 유용한 정보가 될 수 있다.

[0042] 태깅부(140)는 상기와 같은 원리를 이용하여 말뭉치의 오류로 판단되는 어절 군의 형태소 분석열의 전이 확률을 제거한다. 기설정된 임계값이 1%라고 가정하면, 태깅부(140)는 어절 군의 빈도 정보에 따라 전체 조합의 빈도의 합에서 1% 이하의 형태소 분석열의 전이 확률을 제거할 수 있다.

[0043] 또한, 태깅부(140)는 최적의 형태소 분석열을 태깅하기 위하여 은닉 마르코프 모델(Hidden Markov Model : HMM)을 이용한다. 이때, 은닉 마르코프 모델은 관측 불가능한 프로세스를 관측 가능한 다른 프로세스를 통해 추정하는 이중 통계적 모델로, 생성 확률과 전이 확률을 이용하여 최적의 상태열을 찾는다.

[0044] 특히, 태깅부(140)는 은닉 마르코프 모델의 최적 상태열을 찾기 위해 비터비(Viterbi) 알고리즘을 사용한다.

[0045] 구체적으로, 태깅부(140)는 관측열 $X=\{X_1, X_2, X_3, \cdots, X_T\}$가 주어져 있을 EO, 이러한 관측열을 발생시키기 위한

단일한 최적 상태열 q={q_1, q_2, q_3, …, q_T}을 찾기 위해서 아래와 같이 수학식 1 및 수학식 2를 정의한다.

수학식 1

[0046] $$\delta_{t+1}(j) = \max_{1 \le i \le N} \delta_t(i) a_{ij} b_j(\mathbf{x}_{t+1})$$

수학식 2

[0047] $$\psi_{t+1}(j) = \arg\max_{1 \le i \le N} \delta_t(i) a_{ij}$$

[0048] $\delta_t(i)$는 시간 t에서 첫 번째 t개의 관측과 어떤 상태 i에서 끝나는 단일 패스에서 가장 확률이 큰 최상의 스코어를 의미한다. a_{ij}는 시간 t의 어떤 상태 i에서 시간 t+1의 어떤 상태 j로의 전이확률이다. $b_j(\mathbf{x}_{t+1})$는 시간 t+1에서의 어떤 상태 j의 생성확률이다. 비터비 알고리즘에서는 시간 t+1에서 j에 대하여 $\delta_{t+1}(j)$를 최대화하는 상태의 트랙을 배열 $\psi_{t+1}(j)$에 저장하고 역추적(Back Tracking)을 통하여 최적 상태열을 탐색한다.

[0049] 비터비 알고리즘에서 최적 상태열을 찾기 위하여 아래와 같이, 네 가지 과정을 거친다.

[0050] 수학식 3은 초기화 과정을 위한 수식으로, 문장의 첫번째 어절의 생성 확률을 적용한 것이다. 문장의 첫번째 어절은 이전 어절이 존재하지 않기 때문에, 전이 확률을 반영할 필요가 없다.

수학식 3

[0051] $$\delta_1(i) = \pi_i b_i(X_1)$$

[0052] $$\psi_1(i) = 0$$

[0053] 여기서, $\delta_1(i)$는 첫번째 어절의 형태소 분석열의 최대 생성 확률, π_i는 현재 어절에 대하여 기설정된 값, $b_i(X1)$은 첫번째 어절의 형태소 분석열의 생성 확률, $\psi_1(i)$는 첫번째 어절의 형태소 분석열의 생성 확률 및 전이 확률의 곱이다.

[0054] 수학식 4는 반복 과정을 위한 수식으로, t+1번째 어절의 생성 확률과 t번째 어절에서 t+1번째 어절로의 전이 확률을 적용한 것이다. 어절 군에서, 앞의 어절에 해당하는 형태소 분석열은 i가 되고, 뒤의 어절에 해당하는 형태소 분석열은 j가 된다. 여기서, i는 1 이상 N 이하의 수이다.

수학식 4

[0055] $$\delta_{t+1}(j) = \max \delta_t(i)\, a_{ij}\, b_j(X_{t+1}), t = 1, \ldots, T-1$$

[0056] $$\psi_{t+1}(j)=\arg\max \delta_t(i)\, a_{ij}$$

[0057] 여기서, $\delta_{t+1}(j)$는 생성 확률 및 전이 확률이 최대인 값, $\delta_t(i)$는 t 어절에서 i 형태소 분석열과 j 형태소 분석열간의 최대의 전이 확률, a_{ij}는 t 어절에서 i 형태소 분석열로부터 j 형태소 분석열로의 전이 확률, $b_j(X_{t+1})$은 t+1 어절에서 j 형태소 분석열의 생성 확률, $\psi_{t+1}(j)$는 $\delta_{t+1}(j)$의 최대 값이다.

[0058] 수학식 5는 종료 과정을 위한 수식으로, 문장의 마지막 어절의 생성 확률을 적용한 것이다. 마지막 어절은 뒤에 오는 어절이 없기 때문에, 뒤에 오는 어절과의 전이 확률을 반영할 필요가 없다.

수학식 5

[0059] $$P=\max \delta_T(i)$$

[0060] $$q_T=\arg\max \delta_T(i)$$

[0061] 여기서, P는 마지막 어절의 생성 확률과 전이 확률이 최대인 값, q_T는 P가 최대인 값이다.

[0062] 수학식 6은 최적의 형태소 분석열의 역추적(Back Tracking) 과정을 위한 수식으로, 문장의 마지막 어절에서 첫 번째 어절까지 계산된 값을 역추적하여 최적의 형태소 분석열을 태깅하는 것이다.

수학식 6

[0063] $$q_t=\psi_{t+1}(q_{t+1}),t=T-1,\ldots,1$$

[0064] 이하에서는 도 4를 참조하여, 도 1에서 설명한 품사 및 동형이의어 태깅 방법에 대해 구체적으로 설명하도록 한다. 도 4는 본 발명의 일 실시예에 다른 품사 및 동형이의어 태깅의 모델링을 나타낸 도면이다.

[0065] 도 4를 참조하면, "물을 가지러 간 사이"라는 문장이 입력되고, 어절 구분부(110)는 입력된 문장의 어절을 구분한다. 이에 의해, 입력된 문장은 4개의 어절로 구분되었다. 즉, 어절 구분부(110)는 "물을", "가지러", "간", "사이"의 4개 어절로 구분한다.

[0066] 이후, 형태소 분석열 추출부(120)는 어절 구분부(110)에 의해 구분된 각 어절에 대하여 제1 데이터베이스(150)로부터 해당 형태소 분석열을 추출한다. 이 결과에 따라, "물을" 어절에는 "묻_03/VV+을/ETM", "물_01/NNG+을/JKO", 및 "물_02/NNG+을/JKO"로 3개의 형태소 분석열이 추출된다. 도시한 바와 같이, 형태소 분석열은 "형태소_의미번호/품사"의 형태로 구성되어 있다.

[0067] 또한, 형태소 분석열 추출부(120)는 제1 데이터베이스(150)로부터 해당 형태소 분석열과 그 빈도 정보를 함께 추출한다. 도시한 바와 같이, "묻_03/VV+을/ETM"의 형태소 분석열은 0.14, "물_01/NNG+을/JKO"의 형태소 분석열은 0.85, 및 "물_02/NNG+을/JKO"의 형태소 분석열은 0.01의 출현 빈도를 나타낸다.

[0068] 태깅부(140)는 각 어절별 형태소 분석열이 갖는 빈도 정보를 이용하여 생성 확률을 결정하고, 어절 군의 형태소 분석열이 갖는 빈도 정보를 이용하여 전이 확률을 결정한다.

[0069] 또한, 태깅부(140)는 비터비 알고리즘을 이용하여 각 어절별로 최적의 하나의 형태소 분석열을 태깅한다. "물을"의 어절에서 각 형태소 분석열의 빈도 정보는 0.14, 0.85, 및 0.01로 "물_01/NNG+을/JKO" 형태소 분석열의 생성 확률이 최고값을 가진다. "가지러"의 어절에서는 가능한 형태소 분석열이 하나뿐이므로, 비터비 알고리즘이 "물을" 어절에서 "가지러" 어절까지 진행되었을 때, "가지/VV+러/EC"가 얻는 $\psi_{t+1}(j)$ 값은 0.85가 된다.

[0070] 동일한 방법으로, "사이" 어절까지 진행되었을 때, "사이" 어절에서 가장 높은 $\Psi_{t+1}(j)$ 값을 가지는 형태는 0.52 값을 가지는 "사이_01/NNG"가 된다. 이후, 확률 결정부(160)는 역추적을 통해 문장의 마지막부터 처음까지 최고의 $\Psi_{t+1}(j)$ 값을 가지는 각 어절별 형태소 분석열을 태깅하게 된다.

[0071] 변환부(170)는 상술한 품사 및 동형이의어 태깅 결과를 이용하여, 적어도 하나의 형태소 분석열을 한자로 변환한다. 구체적으로, 저장부(미도시)에 기 저장된 태깅 결과에 따른 품사 또는 어깨번호를 이용하여, 변환부(170)는 형태소 분석열을 한자로 변환한다.

[0072] 예를 들어, 저장부에 "개정_01"은 "改定"와 매칭되어 저장되고, "개정_02"는 "改正"와 매칭되어 저장되고, "개정_03"은 "改訂"와 매칭되어 저장되어 있다고 가정한다. 이때, "투표로 개정한다."라는 텍스트의 태깅 결과, "투표", "개정"이 문맥에 맞는 동형이의어인 "투표_01", "개정_02"로 결정되면, 변환부(170)는 태깅된 결과인 "투표_01", "개정_02"와 대응되도록 "투표"를 "投票"로 변환하고, "개정"을 "改正"으로 변환한다.

[0073] 이때, 변환부(170)는 텍스트에 포함된 한자로 변환 가능한 모든 형태소 분석열을 한자로 변환할 수 있고, 사용자에 선택에 의해 특정 형태소 분석열만을 한자로 변환할 수 있다.

[0074] 상술한 바와 같은 단말 장치(100)에 의해 동형이의어가 존재하는 한글 텍스트에서도 문맥에 맞는 한글-한자 변환이 가능해 질 수 있다.

[0075] 이하에서는 도 5를 참조하여, 단말 장치(100)의 한글-한자 변환 방법을 설명하기로 한다.

[0076] 도 5는 본 발명의 일 실시예에 따른, 한글-한자 변환 방법을 설명하기 위한 흐름도이다.

[0077] 우선, 단말 장치(100)에 텍스트가 입력된다(S510). 이때, 텍스트는 어떤 사상이나 감정의 표현이 완결된 것을 의미하나, 표현이 완결되지 않은 단어(單語) 문장, 및 주어를 갖추지 않은 명령문과 같은 예외의 경우도 문장 역시 포함한다.

[0078] 그리고, 단말 장치(100)는 입력된 텍스트의 어절을 구분한다(S520). 예를 들어, "나는 오전에 공부를 열심히 한다."라는 문장이 입력되었다면, 단말 장치(110)는 이 텍스트를 5개의 어절 즉, "나는", "오전에", "공부를", "열심히", "한다"로 구분한다.

[0079] 그리고, 단말 장치(100)는 구분된 어절별로 적어도 하나의 형태소 분석열을 추출한다(S530). 이때, 단말 장치(100)는 형태소 분석열 추출을 위해 기생성된 제1 데이터베이스(150) 및 제2 데이터베이스(160)로부터 해당 형태소 분석열을 추출할 수 있다. 여기서, 제1 데이터베이스(150)는 어절, 그 어절의 형태소 분석열, 및 그 형태소 분석열의 빈도 정보를 포함한다. 또한, 제2 데이터베이스(160)는 인접하는 두 어절의 조합에 의한 어절 군, 그 어절 군의 형태소별 분석열, 및 그 형태소별 분석열의 빈도 정보를 포함한다.

[0080] 그리고, 단말 장치(100)는 추출된 형태소 분석열의 생성 확률 및 전이 확률을 산출하고, 산출된 생성 확률 및 전이 확률을 이용하여 각 어절에 대해 최적의 형태소 분석열을 태깅한다(S540). 여기서, 생성 확률은 제1 데이터베이스(150)로부터 추출되는 형태소 분석열별 빈도 정보를 이용하여 결정되며, 전이 확률은 제2 데이터베이스(160)에 기 저장된 형태소별 품사의 빈도 정보를 이용하여 결정될 수 있다.

[0081] 이때, 기결정된 전이 확률 중, 어절 군의 전체 빈도 정보 대비 해당 어절 군의 형태소 분석열의 빈도 정보가 기설정된 임계값 이하이면, 해당 어절 군의 형태소 분석열의 전이 확률을 제거할 수 있다.

[0082] 그리고, 단말 장치(100)는 최적의 형태소 분석열을 태깅하기 위하여 은닉 마르코프 모델(Hidden Markov Model : HMM)을 이용한다. 이때, 은닉 마르코프 모델은 관측 불가능한 프로세스를 관측 가능한 다른 프로세스를 통해 추정하는 이중 통계적 모델로, 생성 확률과 전이 확률을 이용하여 최적의 상태열을 찾는다. 특히, 단말 장치(100)는 은닉 마르코프 모델의 최적 상태열을 찾기 위해 비터비(Viterbi) 알고리즘을 사용한다. 비터비 알고리즘에 대해서는 도 1에서 상술한 바와 같이,

[0083] 그리고, 단말 장치(100)는 태깅 결과를 이용하여, 적어도 하나의 형태소 분석열을 한자로 변환한다(S550). 이때, 단말 장치(100)는 기 저장된 태깅 결과에 따른 품사 또는 어깨번호를 이용하여, 형태소 분석열을 한자로 변환할 수 있다.

[0084] 상술한 바와 같은 한자-한글 변환 방법에 의해 동형이의어가 존재하는 한글 텍스트에서도 문맥에 맞는 한글-한자 변환이 가능해 질 수 있다.

[0085] 한편, 상술한 제1 데이터베이스(150) 및 제2 데이터베이스(160)가 단말 장치(100)에 포함된 것으로 상정하였으

나, 이는 일 실시예에 불과할 뿐, 외부에서 네트워크로 연결될 수도 있다.

[0086] 또한, 상술한 태깅 방법은 비터비 알고리즘을 사용하여, 동형이의어를 태깅하였으나, 이는 일 실시예에 불과할 뿐 다른 동형이의어 태깅 방법에 의해 동형이의어를 태깅할 수 있다.

[0087] 이상에서는 본 발명의 바람직한 실시예에 대하여 도시하고 설명하였지만, 본 발명은 상술한 특정의 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진자에 의해 다양한 변형실시가 가능한 것은 물론이고, 이러한 변형 실시예들은 본 발명의 기술적 사상이나 전망으로부터 개별적으로 이해되어서는 안될 것이다.

부호의 설명

[0088] 110: 어절 구분부 120: 형태소 분석열 추출부

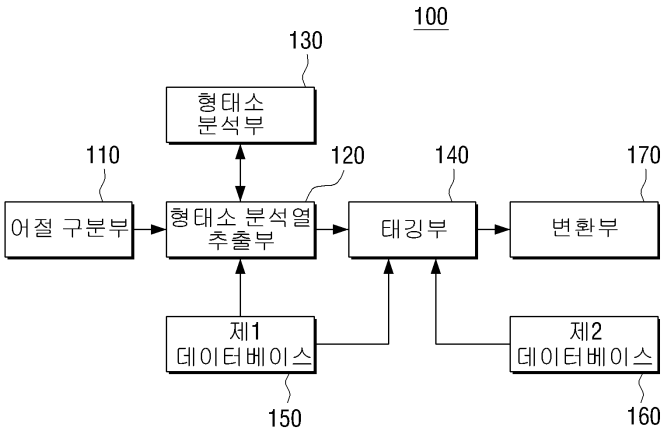
130: 형태소 분석부 140: 태깅부

150: 제1 데이터베이스 160: 제2 데이터베이스

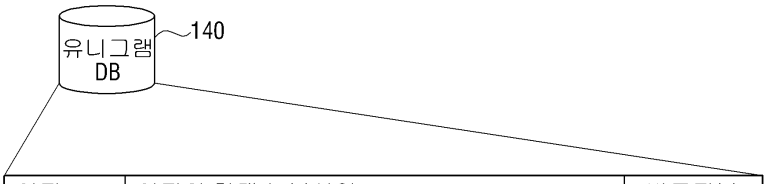
170: 변환부

도면

도면1

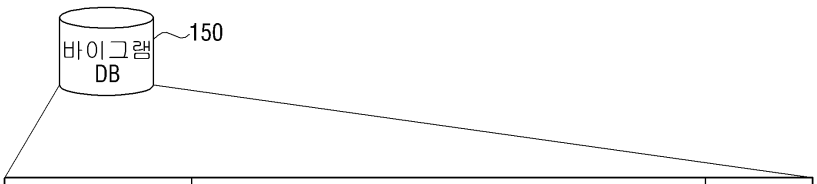


도면2



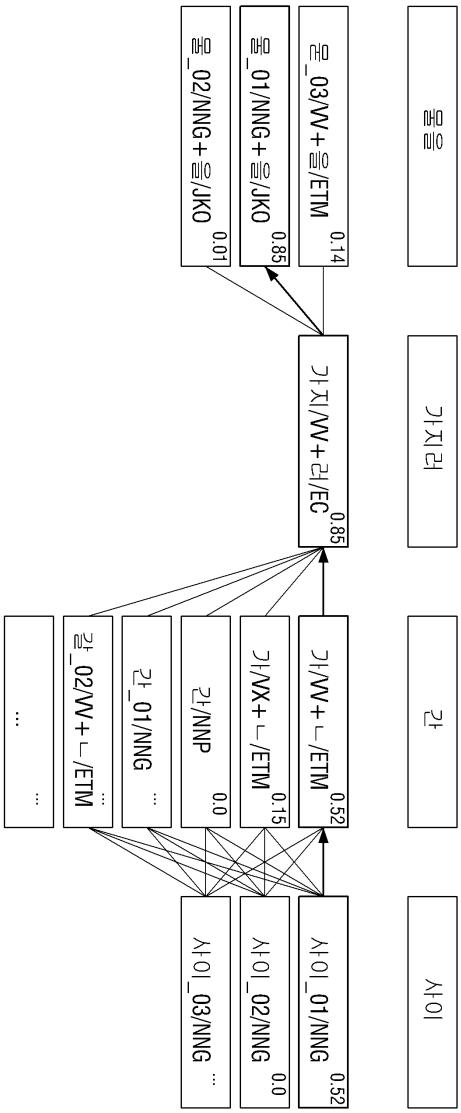
어절	어절의 형태소 분석열	빈도정보
한	하/VX+/ETM	5
한	한/NNB	1
한	한/MMM	1
한	한/XPN	2
한	하/VV+ㄴ/ETM+」/SS	1
한	하/XSV+ㄴ/ETM	2
한	하/XSV+ㄴ/ETM	4
한	하/VV+/ETM	8
한	한_ _05/NNG	33
한	한/NNP	117
한	하/VV+ㄴ/ETM	6766
한	한/MM	28981
한	하/VX+ㄴ/ETM	1836
한	한_ _06/NNG	1098

도면3

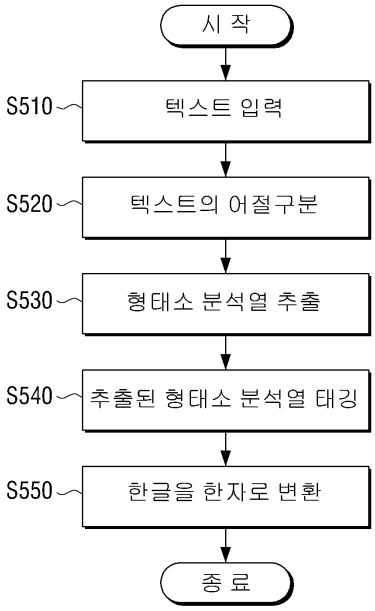


어절 군	어절 군의 형태소 분석열	빈도정보
수 있느냐는	수_ _02/NNB 있/VV+느냐는/ETM	47
수 있느냐'는	수_ _02/NNB 있/VV+느냐/EC+'/SS+는/ETM	4
수 있느냐다	수_ _02/NNB 있/VV+느냐+이/VCP+다/EF+./SF	2
수 있느냐도	수_ _02/NNB 있/VV+느냐/EC+도/JX	1
수 있느냐라는	수_ _02/NNB 있/VV+느냐라는/ETM	1
수 있느냐라는	수_ _02/NNB 있/VV+느냐/EF+이/VCP+라는/ETM	1
수 있느냐로	수_ _02/NNB 있/VV+느냐/EC+로/JKB	1
수 있느냐를	수_ _02/NNB 있/VV+느냐/EC+를/JKO	6
수 있느냐며	수_ _02/NNB 있/VV+느냐며/EC	1
수 있느냐면서	수_ _02/NNB 있/VV+느냐면서/EC	1
수 있느냐'면서	수_ _02/NNB 있/VV+느냐/EC+'/SS+면서/EC	1
수 있느냐에	수_ _02/NNB 있/VV+느냐/EC+에/JKB	10
수 있느냐의	수_ _02/NNB 있/VV+느냐/EC+의/JKB	11
수 있느냐이다.	수_ _02/NNB 있/VV+느냐/EC+이/VCP+다/EF+/SF	2
수 있느니	수_ _02/NNB 있/VV+느니/EC	1
수 있느니..."	수_ _02/NNB 있/VV+느니/EC+../SE+../SE+,/SF+'/S	1
수 있느니라.	수_ _02/NNB 있/VV+느니라/EF+/SF	1
수 있느니라."	수_ _02/NNB 있/VV+느니라/EF+/SF+'/SS	1
수 있는	수_ _02/NNB 있/VV+는/ETM	11278
수 있는	수_ _02/NNB 있/VX+는/ETM	6

도면4


물을
가지러
간
사이
굴_03/VV+을/ETM 0.14
물_01/NNG+을/JKO 0.85
물_02/NNG+을/JKO 0.01
가지/VV+러/EC 0.85
가/VV+ㄴ/ETM 0.52
가/VX+ㄴ/ETM 0.15
간/NNP 0.0
간_01/NNG ...
갈_02/VV+ㄴ/ETM ...
... ...
사이_01/NNG 0.52
사이_02/NNG 0.0
사이_03/NNG ...

도면5


시 작
S510
텍스트 입력
S520
텍스트의 어절구분
S530
형태소 분석열 추출
S540
추출된 형태소 분석열 태깅
S550
한글을 한자로 변환
종 료