HMM을 이용한 한국어 품사 및 동형이의어 태깅 시스템

김동명[O1] 배영준[1] 옥철영[1] 최호섭[2] 김창환[3]
울산대학교 컴퓨터정보통신공학과[1],
한국과학기술정보연구원 정보기술개발단 정보시스템개발팀[2]
춘해대학교 보건행정과[3]
spelldm@mail.ulsan.ac.kr, {young4862, okcy}@ulsan.ac.kr, hschoe@kisti.re.kr, chkim@ch.ac.kr

Korean POS and Homonym Tagging System using HMM

Dong-myoung Kim[O1] Young-Jun Bae[1] Cheol-Young Ock[1] Ho-Soep Choi[2] Chang-Hwan Kim[3]
Dept. of Computer Engineering and Information Technology, University of Ulsan[1]
Information System Development Team, Korean Institute of Science and Technology Information[2]
Dept. of Health Administration, Choonhae College[3]

요 약

기존의 자연언어처리 연구 중 품사 태깅과 동형이의어 태깅은 별개의 문제로 취급되었다. 그로 인해 두 문제를 해결하기 위한 모델 역시 서로 다른 모델을 사용하였다. 이에 본 논문은 품사 태깅 문제와 동형이의어 태깅 문제는 모두 문맥의 정보에 의존함에 착안하여 은닉마르코프모델을 이용하여 두 가지 문제를 해결하는 시스템을 구현하였다. 제안한 시스템은 품사 및 동형이의어 태깅된 세종 말뭉치 1100만여 어절에 대해 unigram과 bigram을 추출 하였고, unigram을 이용하여 어절의 생성확률 사전을 구축하고 bigram을 이용하여 전이확률 사전을 구축하였다. 구현된 시스템의 성능 확인을 위해 비학습 말뭉치 261,360 어절에 대해 실험하였고, 실험결과 품사 태깅 99.74%, 동형이의어 태깅 97.41%, 품사 및 동형이의어 태깅 97.78%의 정확률을 보였다.

1. 서 론

한국어는 교착어(膠着語)이다. 교착어란 어간에 조사나 어미같은 기능어가 추가되어 어절을 형성하는 언어를 말하며, 한국어를 비롯해 일본어, 몽골어, 만주어와 터키어 등이 있다. 한국어는 교착어적 특성으로 인해 자연언어처리 과정에서 다른 언어에 비해 다양한 형태의 중의성이 발견된다.

한국어의 중의성은 형태적인 중의성과 의미적인 중의성으로 나눌 수 있다. 형태적인 중의성은 “나는” 과 같은 어절에서 찾아 볼 수 있다. 이 어절에 대한 형태소 분석 결과는 “나/대명사+는/보조사”, “나/동사+는/관형형어미”, “날/동사+는/관형형어미” 등으로 나타난다. 이와 같이 하나의 어절이 여러 가지 형태로 분석될 수 있으므로 이를 위한 처리가 필요하다. 다음으로 의미적인 중의성은 “배를” 과 같은 어절에서 찾을 수 있다. 이것은 “배/명사+를/목적격조사” 의 형태소 분석되지만 동형이의어 “배” 가 가지는 여러 가지 의미로 인해 중의성을 띄게 된다.

본 논문은 은닉마르코프모델(HMM : Hidden Markov Model)[1]을 이용하여 한국어의 형태적인 중의성과 의미적인 중의성을 함께 해결할 수 있는 시스템을 구축한다.

본 논문은 2장에서 관련 연구를 소개하고, 3장에서 본 논문에서 제안한 시스템에 대해 기술하며, 4장에서는 본 논문에서 제안한 시스템의 성능을 평가하고 분석한다. 5장에서는 결론과 향후 연구를 다룬다.

2. 관련 연구

이 장에서는 품사 태깅 모델과 동형이의어 태깅 모델에 대한 기존의 연구들을 살펴보고 본 논문에서 제안하는 모델에 대한 이해 및 고려되어야 할 제반 사항에 대해서 설명한다.

2.1 품사 태깅 모델

품사 태깅을 위한 모델은 크게 규칙 기반 모델과 통계 기반 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 높다는 장점이 있다. 그러나 말뭉치 구축에 시간과 노력이 많이 요구되고, 말뭉치가 일정 크기 이상

구축되어 있지 않을 경우 통계 자료 부족으로 인하여 신뢰도가 떨어진다는 단점이 있다.

최근에는 규칙 기반 모델과 통계 기반 모델을 결합하여 서로 간의 결점을 보완하는 복합적 모델을 사용하는 추세이다.

[2]에서는 규칙 기반 품사 중의성 해소 모듈을 이용해 형태소 분석된 문장에 가중치를 부여하고 중의성을 해소한 후, 어절별로 품사 중의성 해소 여부를 판단하여 중의성이 해소되지 않은 어절에 대하여 카테고리 패턴 기반의 품사 중의성 해소 모듈에서 어절 확률을 비교하여 해결한다.

[3]에서는 통계 기반 품사 태깅에서 많이 사용되는 HMM과 Viterbi 알고리즘을 이용한 품사 태깅에서 사용되는 학습 말뭉치의 오류를 검출하였다. 사람의 수작업으로 구축되는 말뭉치의 특성상 나타날 수 밖에 없는 오류에 대해서 저신뢰도 구간 검사를 통해 학습 말뭉치의 신뢰도를 높여 태깅 성능의 향상을 모색하였다.

2.2 동형이의어 태깅 모델

의미 중의성 해소(WSD : Word Sense Disambiguation)는 그 자체가 하나의 완전한 작업은 아니지만 대부분의 자연언어처리 작업에서는 반드시 필요한 작업이다. 예를들면 문서 분류를 위해 WordNet[4]이나 U-WIN[5]과 같은 어휘망을 이용할 수 있는데 이를 위해서는 동형이의어에 대한 중의성 해소가 선행되어야 한다.

[6]에서는 한국어의 동형이의어 중의성을 해소하기 위하여 사전의 뜻풀이 말뭉치에서 구축한 의미정보와 이를 적용한 베이지안 분류 모델을 이용하여 동형이의어 중의성 해소를 수행하였다.

[7]에서는 연어의 유형별로 동형이의어 중의성 해소 모형을 실험하였다. 이 실험을 통해 중의성 해소 대상 단어의 의미별 출현 비율과 중의성 해소에 결정적인 역할을 하는 단서어의 출현위치에 따라 중의성 해소 성능이 큰 차이를 보이는 것을 발견하였다. 그리고 대상 단어가 학습집단과 실험집단 안에서 충분한 수의 연어를 갖는다면 정확률 90% 수준의 높은 성능을 가져올 수 있을 것으로 보았다.

3. HMM을 이용한 한국어 품사 및 동형이의어 태깅 시스템

이 장에서는 품사 태깅과 동형이의어 태깅을 동시에 수행할 수 있는 시스템을 소개한다.

3.1 착안점

본 논문에서 제안하는 시스템의 기본 착안점은 문장에서 어절의 품사와 동형이의어는 모두 문맥 정보에 의해 결정된다는 가정에서 출발하였다.

[그림 1]에서와 같이 품사 중의성이 있는 "나는" 이라는 어절에 대한 품사 태깅 결과는 주변 문맥 정보에 의존한다.

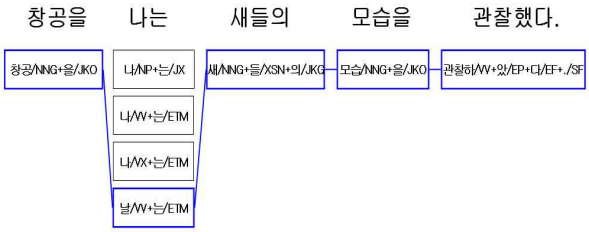

그림 30. 품사 태깅 예

[그림 2]에서는 "배를" 과 "타고" 의 품사 태깅 결과는 동일하지만 "배/NNG" 와 "타/VV"는 각각 명사와 동사인 동형이의어로써 의미적인 중의성을 가진다. 이 경우 [그림 1]의 "나는" 에서와 마찬가지로 주변 문맥 정보에 의해 각 단어의 동형이의어가 결정되는 것을 알 수 있다.

그림 31. 동형이의어 태깅의 예

[그림 1]과 [그림 2]에서와 같이 품사 태깅과 동형이의어 태깅은 모두 주변 문맥 정보에 의해 영향을 받으므로 본 논문에서는 기존의 연구에서와 달리 두 가지 과정을 하나의 모델에서 해결하도록 하였다.

3.2 시스템 설계

HMM은 이중 통계적 모델로서 생성확률과 전이확률을 이용하여 최적의 상태열을 찾는다. HMM의 이와 같은 특징은 문맥 정보를 반영하기에 용이하므로 본 논문에서는 HMM을 기본 모델로 이용하였다.

HMM의 최적 상태열을 찾기 위해 Viterbi 알고리즘을 사용하였다. Viterbi에서는 관측열 $X=\{X_1,X_2,X_3,\ldots,X_T\}$가 주어져 있을 때, 이러한 관측열을 발생시키는 단일한 최적 상태열 $q=\{q_1,q_2,q_3,\ldots,q_T\}$을 찾기 위해서 식(1)과 식(2)를 정의한다.

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

$$\psi_{t+1}(j)=\arg\max_{1\le i\le N}\delta_t(i)a_{ij} \quad (2)$$

$\delta_t(i)$는 시간 t에서 첫 번째 t개의 관측과 어떤 상태 i에서 끝나는 단일 패스에서 가장 확률이 큰 최상의 스코어

를 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. Viterbi에서는 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.

Viterbi 알고리즘에서 최적 상태열을 찾기 위한 전체 과정은 다음과 같다.

1단계 : 초기화

$$\delta_1(i) = \pi_i b_i(\mathbf{x}_1) \quad (3)$$

$$\psi_1(i) = 0 \quad (4)$$

2단계 : 반복

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

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

3단계 : 종료

$$P^* = \max_{1\le i\le N} \delta_T(i) \quad (5)$$

$$q_T^* = \arg\max_{1\le i\le N} \delta_T(i) \quad (6)$$

4단계 : 최적 상태열 역추적

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1,\ldots,1 \quad (7)$$

본 논문에서는 위에 언급한 과정을 제안한 태깅 시스템에 적용하기 위하여 [그림 3]과 같은 형태로 모델링 하였다.

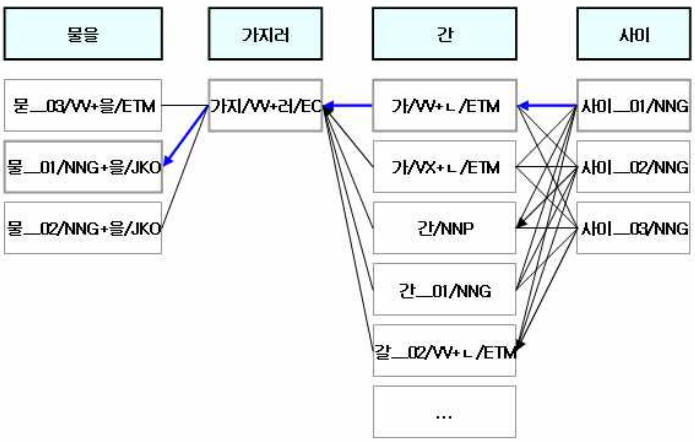

그림 3. HMM에 의한 태깅 시스템 모델링

HMM에서의 생성확률은 어절이 형태소 분석되는 경우에 대한 발생확률로 적용하였고, 어절 간의 bigram 통계를 이용한 어절의 전이확률을 모델의 상태전이확률로 사용하였다. 어절의 발생확률을 위해 추출한 unigram 데이터는 어절의 품사 정보와 의미 정보를 함께 가지고 있으므로 기분석사전의 용도로 활용하였다. [2]에서는 통계 자료 저장 문제를 들어 unigram과 카테고리 패턴을 사용하였으나 이것만으로는 어절에서 발생할 수 있는 의미 중의성 처리에 미흡하므로 bigram을 사용하였다.

본 논문에서 제안하는 시스템에는 본 연구실이 보유하고 있는 형태소 분석기 UTagger[8]를 이용하였다. UTagger는 어절사전을 기반으로 하여 빠른 형태소 분석이 가능하며 어절에 대한 빈도 정보를 갖고 있으므로 한국어 후처리 단계에서 효과적으로 이용될 수 있는 형태소 분석기이다. UTagger는 동형이의어 태깅에 대한 처리를 하지 않으며, 학습 말뭉치에서 품사 및 동형이의어 태깅 정보가 없는 경우 품사 태깅을 위한 형태소 분석 용도로 사용하였다.

제안한 시스템의 학습을 위해 "21세기 세종 계획 형태의미 분석 말뭉치" 중 11,100,293개 어절을 이용하였다. 학습 말뭉치는 unigram과 bigram을 추출하기 용이하도록 [그림 5]와 같은 형태로 정제하여 사용하였다.

```
카리브해의 해적선 선장 잭 스패로우(조니 뎁)는 자신의 배 '블랙 펄'호를
카리브해/NNP+의/JKG 해적선/NNG 선장__06/NNG 잭/NNP 스패로우/NNP+(/SS+

바르보사는 그 배를 타고 영국 함대가 있는 토알 포트를 습격해 총독의 딸
바르보사/NNP+는/JX 그/MM 배__02/NNG+를/JKO 타__02/VV+고/EC 영국/NNP

엘리자베스의 어릴 적 친구인 대장장이 윌 터너(올란도 블룸)와 잭 선장은
엘리자베스/NNP+의/JKG 어리__03/VA+ㄹ/ETM 적__03/NNB 친구__02/NNG+이/V

호화 출연진중 러쉬는 영화 '샤인'으로 아카데미 남우주연상을 받았고 다
호화__02/NNG 출연진/NNG+중__04/NNB 러쉬/NNP+는/JX 영화__01/NNG '/SS+
```

그림 5. 정제된 말뭉치의 형태

품사 태깅에 사용된 태그셋은 45개의 태그로 이루어져 있으며 동형이의어 태깅에 사용된 의미 번호는 표준국어대사전의 어깨번호와 동일한 값을 가진다.

한국어의 특성상 어절 간 조합에서 태깅 결과의 경우의 수가 너무 크기 때문에 bigram의 정보만으로는 신뢰성 있는 전이확률을 얻기 힘들다. 이것을 보완하기 위해 추가적으로 가중치를 적용하는 전이확률모듈을 구현하였다. 전이확률모듈은 정보검색 분야의 TF/IDF의 원리를 응용하여 어절 간의 발생 빈도에 따른 가중치를 적용하는 모듈이다.

3.3 시스템 구축

3.2절의 설계를 바탕으로 하여 시스템을 구현하면 [그림 5]와 같은 구조를 갖게 된다. 입력 텍스트에 대해 문장 단위 태깅을 수행한다. 각각의 문장이 포함하는 어절들은 Tokenizer에 의해 나뉘어지며, 어절의 생성확률은 어절 unigram 사전과 UTagger에 의해 결정되는 확률값을 갖는다. 어절 간의 전이확률은 어절 bigram 사전과 전이확률모듈에 의해 확률값을 갖게되며, 이 어절들의 생성확률과 전이확률을 Viterbi 알고리즘에 적용하여 최적의 태그열을 찾는다.

4. 시스템 성능 평가 및 분석

4.1 성능 평가를 위한 데이터

시스템의 태깅 결과를 확인하기 위하여 “세종 말뭉치” 중 학습에 사용되지 않은 말뭉치를 이용하였다. 평가를 위한 말뭉치는 품사 및 동형이의어 태그 부착 말뭉치를 사용하였다.

표 1. 시스템 성능 평가를 위한 정답 말뭉치

	총 어절 수	동형이의어 태깅이 필요한 어절 수
품사 및 동형이의어 태그 말뭉치	261,360 어절	101,844 어절

[표 1]은 평가에 사용된 말뭉치의 총 어절 수와 그 중 동형이의어 태깅이 필요한 어절 수를 나타낸다.

4.2 전체 성능 실험 및 결과

성능 실험 결과는 [표 2]에 나타난 것과 같다. [표 2]의 결과에서 정확률는 어절 단위의 정확률을 나타낸다.

$$\text{정확률} = \frac{\text{올바르게태깅된어절수}}{\text{정답말뭉치의어절수}} \times 100\,(\%)$$

표 2. 전체 성능 실험 결과

	품사 태깅 결과	동형이의어 태깅 결과	품사 및 동형이의어 태깅 결과
정확률	99.74 %	97.41 %	97.78 %

실험 결과 [표 2]의 품사 태깅 결과는 정답 말뭉치에서 품사 태깅만의 정확률을 보여주고, 동형이의어 태깅 결과는 동형이의어 태깅이 필요한 어절에서의 태깅 정확률을 보여주며 품사 및 동형이의어 태깅 결과는 어절의 태깅 결과가 품사와 동형이의어 모두 일치 하였을 경우의 정확률을 나타낸다.

표 3. 태깅 오류 유형

오류 유형	정답 예시 오류 예시
품사 태깅 O, 동형이의어 태깅 X	시청__03/NNG+,/SP 시청__01/NNG+,/SP
품사 태깅 X, 동형이의어 태깅 불필요	다르/VA+ㄴ/ETM 다른/MM
품사 태깅 X, 동형이의어 태깅 X	1/SN+시간__04/NNG 1/SN+시간/NNB

위의 [표 3]은 제안한 시스템으로 태깅한 결과에서 나타난 오류를 유형별로 빈도가 높은 순으로 나열 하였다. 오류 유형에서 “품사 태깅 O”와 “품사 태깅 X”는 품사 태깅이 올바른 경우와 잘못된 경우를 나타내고, “동형이의어 태깅 X”는 동형이의어 태깅이 잘못된 경우를 나타내며 “동형이의어 태깅 불필요”는 어절에서 동형이의어가 나타나지 않은 경우를 나타낸다.

가장 빈도가 높은 오류 유형은 어절의 품사 태깅은 일치하지만 동형이의어 태깅이 일치 하지 않는 경우로 나타났다. 품사 태깅이 잘못되었을 경우 동형이의어 태깅이 필요한 경우에 대해 동형이의어 태깅 역시 잘못 태깅되는 결과가 나타났으며, 이 결과를 통해 품사 태깅과 동형이의어 태깅이 상관관계가 있음을 확인하였다.

4.3 동형이의어 태깅 성능 비교

동형이의어 태깅의 성능을 판단하기 위해 [6]의 ‘실험 1’에서 사용한 9개의 동형이의어에 대해서 정확률을 비교하였다.

표 4. 동형이의어 태깅 성능 정확률 비교

동형이의어	[6] 모델			제안한 모델		
	의미	정확한태깅 문장수 / 전체문장수	정확률	의미	정확한태깅 문장수 / 전체문장수	정확률
기관	몸	14 / 17	93.14	몸	6 / 12	94.59
	장치	1 / 2		장치	0 / 0	
	조직	175 / 185		조직	309 / 321	
기구	장치	20 / 24	68.03	장치	18 / 21	93.94
	조직	63 / 98		조직	75 / 78	
눈	신체부위	412 / 431	90.02	신체부위	426 / 441	95.51
	식물	1 / 1		식물	0 / 0	
	기상현상	47 / 79		기상현상	21 / 27	
다리	교각	5 / 21	78.48	교각	27 / 30	97.56
	신체부위	57 / 58		신체부위	93 / 93	
병	그릇	1 / 12	92.02	그릇	60 / 75	91.80
	질병	149 / 151		질병	108 / 108	
배	과일	1 / 6	75.68	과일	6 / 12	87.10
	운송수단	80 / 92		운송수단	27 / 27	
	신체부위	31 / 50		신체부위	48 / 54	
비	청소도구	0 / 1	82.68	청소도구	0 / 0	100.00
	기상현상	79 / 86		기상현상	159 / 159	
	비석	2 / 10		비석	0 / 0	
	비율	0 / 1		비율	3 / 3	
신	신발	1 / 2	87.43	기분	18 / 18	95.00
	종교	326 / 372		종교	39 / 42	
차	운송수단	28 / 46	56.70	운송수단	114 / 114	100.00
	음료	17 / 39		음료	15 / 15	
	차이	10 / 12		차이	3 / 3	
평균	84.63			95.28		

[6]모델에서 사용된 정확률은 NPH와 거리 가중치가 적용된 정확률이다. [표 4]의 의미는 각 단어가 나타내

는 동형이의어의 의미이며, 문장수는 실험에서 각각의 동형이의어가 나타난 문장의 개수를 나타낸다.

[표 4]의 비교 결과를 보면 [6]의 '실험 1'에서는 선정한 9개의 동형이의어에 대하여 평균 84.63%의 정확률을 보인 반면, 본 논문에서 제안한 모델에서는 평균 95.28%의 정확률을 보였다. 서로 사용된 문장이나 어절의 개수에서 차이가 있기 때문에 정확한 비교 평가라고 하기에는 부족하다. 그러나 [표 2]의 전체 성능 실험 결과에서 나타난 동형이의어 태깅 결과와 [표 3]의 결과로 비춰 볼 때 본 논문에서 제안한 시스템은 [6]에서 제안한 모델에 비해 전체적으로 뛰어난 동형이의어 태깅 성능을 나타내었다. 이 성능 비교 결과 동형이의어 태깅 시 공기 정보를 이용한 베이지안 모델보다 문맥 정보를 이용한 HMM이 좀 더 좋은 성능을 나타냄을 알 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 품사 태깅 문제와 동형이의어 태깅 문제는 모두 문맥 정보에 의존함에 착안하여 은닉마르코프 모델만을 이용하여 두 가지 문제를 해결하는 시스템을 구현하였다. 기존의 연구들의 대부분은 품사 태깅과 동형이의어 태깅을 별개의 과정으로 취급하였다. 그로 인해 문제를 해결하는 모델 역시 별개의 모델을 사용하여 왔으나 본 논문에서는 품사 및 동형이의어 태깅된 말뭉치와 HMM만을 사용한 시스템을 구현하여 품사 태깅과 동형이의어 태깅의 문제를 해결하였다.

시스템의 학습을 위해 품사 및 동형이의어 태깅된 학습 말뭉치를 사용하여 어절의 생성확률을 위한 unigram 사전을 구축하였다고, unigram 사전에 미등록된 어절을 처리하기 위하여 UTagger를 이용하였다. 어절 간의 전이확률을 구하기 위해 bigram 사전을 구축하였으며, 말뭉치의 태깅 오류 및 희소 어절에 대한 신뢰도를 유지하기 위하여 TF/IDF를 응용한 전이확률모듈을 두어 전이확률의 가중치를 설정하였다.

제안한 시스템의 성능 평가를 위해 학습에 사용되지 않은 품사 및 동형이의어 태깅된 말뭉치 261,360어절을 사용하여 태깅한 결과 품사 태깅 99.73%, 동형이의어 태깅 97.41%, 품사 및 동형이의어 태깅 97.78%의 정확률을 나타내었다. 또한, 동형이의어 태깅 시 [6]에서 제안한 모델보다 뛰어난 성능을 보였다.

현재 특정 어휘와 어휘 사이의 공기 관계 및 문맥 관계 등을 제안한 시스템에 추가하여 생성확률 및 전이확률에 가중치를 주어 시스템의 성능을 향상시키는 방법에 대해 연구를 수행하고 있다.

앞으로 대용량 통계 자료의 저장소 문제를 해결하기 연구를 진행한다면 더욱 확장성 있는 시스템을 구축할 수 있을 것으로 예상된다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다.(IITA-2008-(C1090-0801-0039))

참고 문헌

[1] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.

[2] 황명진, 강미영, 권혁철, "규칙과 어절 확률을 이용한 혼합 품사 태깅 모델", 한국정보과학회 가을 학술발표논문집, 제33권, 제2호(B), 11-15페이지, 2006.

[3] 설용수, 김동주, 김규상, 김한우, "말뭉치 오류를 고려한 HMM 한국어 품사 태깅 시스템", 한국 컴퓨터정보과학회 하계학술발표논문집, 제15권, 제1호, 2007.

[4] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller, "Introduction to WordNet : An On-line Lexical Database", 1993.

[5] 최호섭, "대규모 사용자 어휘지능망 구축과 활용", 울산대학교 박사학위논문, 2007.

[6] 김준수, 최호섭, 옥철영, "가중치를 이용한 통계 기반 한국어 동형이의어 분별 모델", 한국정보과학회논문지, 제30권, 제11호, 2003.

[7] 정영미, 이용구, "정보검색 성능 향상을 위한 단어 중의성 해소 모형에 관한 연구", 한국정보관리학회지, 제22권, 제2호, 125-145페이지, 2005.

[8] 김재한, 옥철영, "어절 사전을 이용한 한국어 형태소 분석", 한국정보과학회 학술발표논문집, 제21권, 제1호, 813-816페이지, 1994.

[9] 허정, 옥철영, "사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 한국정보과학회논문지(소프트웨어 및 응용), 제28권, 제9호, 2001.

[10] 이동훈, 강미영, 황명진, 권혁철, "규칙과 비감독 학습 기반 통계정보를 이용한 품사 태깅 시스템", 한국컴퓨터종합학술대회 논문집, 제32권, 제1호(B), 2005.

[11] 김진동, 임희석, 임해창, "Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 한국정보과학회논문지(B), 제24권, 제12호, 1997.

[12] 강미영, "한국어의 언어적 특징을 고려한 범주 패턴 기반 통계적 태깅 모델", 부산대학교 박사학위논문, 2007.