의미처리 기반의 한글-한자 변환 시스템

김홍순[o*], 신준철[*], 옥철영[*]
[*]울산대학교 컴퓨터정보통신공학과
rlaghdtns2@mail.ulsan.ac.kr[o*], ducksjc@hotmail.com, okcy@ulsan.ac.kr

korean-Hanja Translation System based on Semantic Processing

Hong-Soon Kim[o*], Joon-Choul Sin[*], Cheol-Young Ok[*]
[*]Dept of Computer Engineering and Information Technology,
Ulsan University

요 약

워드프로세서에서의 한자를 가진 한글 어휘의 한자 변환 작업은 사용자에 의해 음절/단어 단위의 변환으로 많은 시간이 소요되어 효율이 떨어진다. 본 논문에서는 한글 문장의 의미처리를 통해 문맥에 맞는 한자를 자동 변환하는 시스템을 제안한다. 문맥에 맞는 한글-한자 변환을 위해서는 우선 정확한 형태소 분석 및 동형이의어 분별이 선행되어야 한다. 이를 위해 본 논문에서는 은닉마르코프모델 기반의 형태소 및 동형이의어 동시 태깅 시스템을 구현하였다. 제안한 시스템은 형태의미 세종 말뭉치 1,100만여 어절을 이용하여 unigram과 bigram을 추출 하였고, unigram을 이용하여 어절의 생성확률 사전을 구축하고 bigram을 이용하여 전이확률 학습사전을 구축하였다. 그리고 품사 및 동형이의어 태깅 후 명사를 표준국어대사전에 등재된 한자로 변환하는 시스템을 구현하였다. 구현된 시스템의 성능 확인을 위해 전체 세종 말뭉치를 문장단위로 비학습 말뭉치를 구성하여 실험하였고, 실험결과 한자를 가진 동형이의어에 대한 한자 변환에서 90.35%의 정확률을 보였다.

1. 서론

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.

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.

본 논문은 정확한 문맥 파악을 위한 한국어의 형태소 및 동형이의어 태깅 시스템을 설명하고 한글-한자 변환시스템을 제안하며, 제안한 시스템의 성능을 평가하고 분석하고, 마지막으로 결론과 향후 연구를 다룬다.

2. 관련 연구

이 장에서는 한국어 형태소 태깅과 동형이의어 태깅에 대한 기존의 연구들과 한자 변환 시스템에 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 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.

[1]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.

[2]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 고부가가치의 언어처리를 위해서는 반드시 필요한 작업이다. 예를 들면 문서 분류, 의미기반의 정보검색 등에서 WordNet[3]이나 U-WIN[4]과 같은 어휘망을 이용할 수 있는데 이를 위해서는 동형이의어에 대한 중의성 해소가 선행되어야 한다.

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

[6]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 학습 집단과 실험집단 안에서 충분한 수의 연어를 갖는다면 정확률 90% 수준의 높은 성능을 가져올 수 있을 것으로 보았다.

2.3 한글-한자 변환 시스템

한자 변환은 일반적으로 한글문서나 MS워드에서 단어를 직접 사용자가 변환시킬 한자를 보고 선택해야 하는 번거로움이 존재하고, 특히 이 경우 사용자가 한자를 모른다면 정확한 변환이 어려워진다.

[7]에서는 전문용어의 한자 변환을 위해 언어모델 및 변환모델을 이용한 문장단위의 한자 자동 변환 방법을 제안하였으며, 사전 미등록어와 복합어의 한글-한자 변환을 위하여 단어분할을 변환의 숨김 과정으로 처리하는 통합 모델을 사용하였다.

3. 한국어 태깅 시스템을 이용한 한자 자동 변환 시스템

3.1 착안점

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

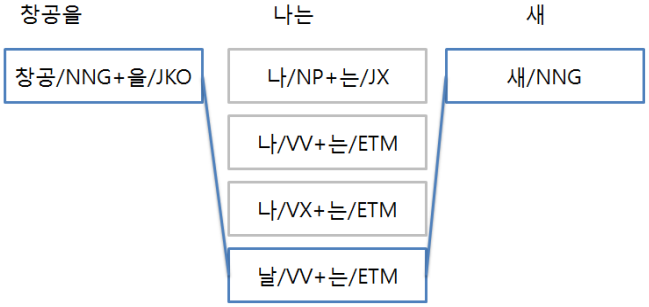

그림 1. 품사 태깅의 예

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

[그림 2]에서는 "투표로"와 "개정한다"는 품사 태깅 결과는 동일하지만 "투표/NNG"와 "개정/NNG"는 동형이의어로서 의미적인 중의성을 가진다. 이 경우 [그림 1]의 "나는"에서와 마찬가지로 주변 문맥 정보에 의해 각 단어의 동형이의어가 결정되는 것을 알 수 있다.

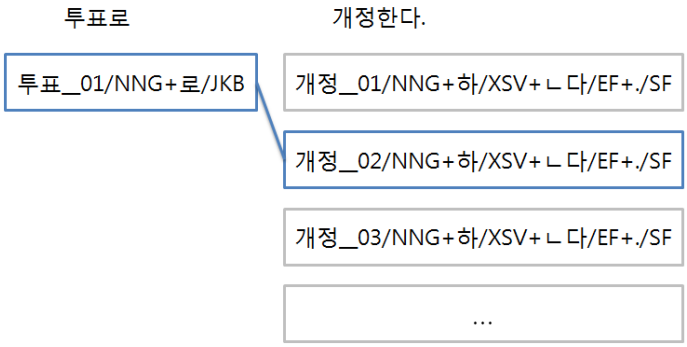

그림 2. 동형이의어 태깅의 예

[그림 3]에서는 "투표", "개정"이 문맥에 맞는 동형이의어('투표__01', '개정__02')가 결정되면 동형이의어의 해당 한자로 변환되는 것을 알 수 있다.

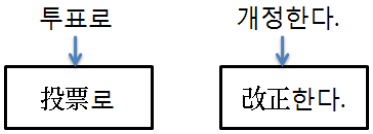

그림 3. 한자 변환의 예

[그림 1]~[그림 3]에서와 같이 한자 변환은 형태소 태깅과 동형이의어 태깅에 의해 정확한 어휘 분석 후 결정되어야 하며, 본 논문에서는 형태소 태깅 및 동형이의어 태깅 과정을 하나의 모델에서 동시에 해결하였다.

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.

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

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

$\delta_t(i)$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. 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 4]와 같은 형태로 모델링 하였다.

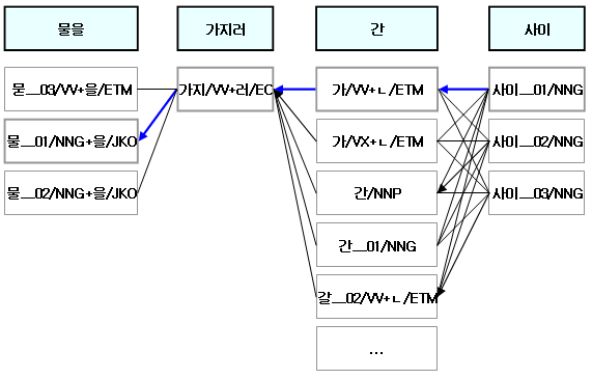

그림 4. HMM에 의한 태깅 시스템 모델링

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. [1]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 unigram에서 발견되지 않은 어절은 울산대 형태소 분석기 ICMA 모듈[8]을 이용하였다. ICMA 모듈은 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.

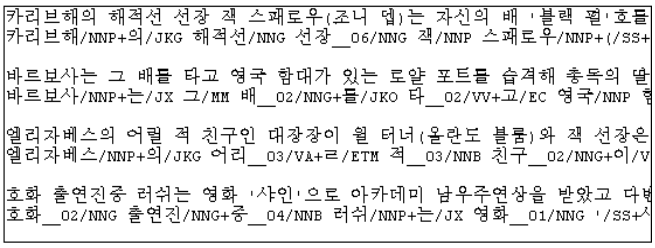

그림 5. 정제된 말뭉치의 형태

품사 태깅에 사용된 태그셋은 45개의 태그로 이루어져 있으며 동형이의어 태깅에 사용된 의미 번호는 표준국어대사전의 어깨번호와 동일한 값을 가지게 되어 결정된 명사류의 어깨번호를 확인하여 한자로 변환한다.

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

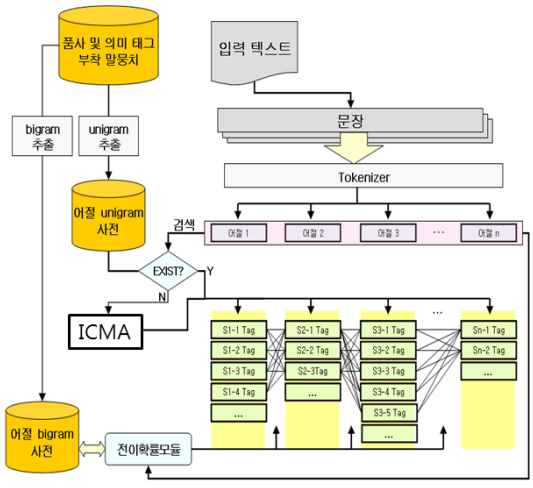

그림 6. 전체 시스템 구조

3.2절의 설계를 바탕으로 하여 시스템을 구현하면 [그림 6]과 같은 구조를 갖게 된다. 입력 텍스트에 대해 문장 단위 태깅을 수행한다. 각각의 문장이 포함하는 어절들은 Tokenizer에 의해 나뉘어지며, 어절의 생성확률은 어절 unigram 사전과 ICMA 모듈에 의해 결정되는 확률값을 갖는다. 어절 간의 전이확률은 어절 bigram 사전과 전이

확률모듈에 의해 확률값을 갖게되며, 이 어절들의 생성확률과 전이확률을 Viterbi 알고리즘에 적용하여 최적의 태그열을 찾는다. 태깅을 모두 수행한 후에 결정된 품사와 어깨번호를 이용하여 한자로 변환하게 된다.

4. 시스템 성능 평가 및 분석

4.1 성능 평가를 위한 데이터

시스템의 태깅 결과를 확인하기 위하여 "세종 말뭉치"의 각 파일 안의 문장의 90%를 학습하고 나머지를 평가하는 방법으로 총 10번의 성능을 평가 하였다.

표 1. 평가를 실시한 말뭉치의 각 어절 수

	1	2	…	10
학습	10,004,759	10,007,910	…	10,005,225
평가	1,112,866	1,109,715	…	1,112,400
한자	362,996	360,268	…	356,388

[표 1]은 태깅 시스템을 이용하기 위해 학습한 어절의 수와 평가를 실시한 어절의 수, 그중 한자로 변환되는 어절의 수를 나타낸다.

표 2. 시스템 성능 평가를 위한 말뭉치

	세종 말뭉치	한자
어절 수	11,117,625	3,615,992

[표 2]는 평가에 사용된 세종 말뭉치의 총 어절 수와 그 중 한자로 변환된 어절 수를 나타낸다.

4.2 전체 성능 실험 및 결과

성능 실험 결과는 태깅 시스템의 동형이의어와 품사 태깅 결과를 나타내고, 전체 태깅 결과의 정확률과 한자 변환의 정확률을 나타내었다.

표 3. 태깅 시스템 평가 결과

	동형이의어	형태소
정확률	94.62%	92.96%

[표 3]은 태깅 시스템에서 동형이의어와 형태소의 정답 비율을 나타낸 것이다. 총 11,117,625어절 중에서 동형의어가 포함된 어절은 6,546,929개이고, 이중 정답 어절수는 5,990,148개였고, 형태소 정답 어절의 수는 10,323,483개로 나타났다.

표 4. 전체 성능 실험 결과

	태깅 결과	한자 변환 결과
정확률	91.82	90.35

성능 실험 결과는 [표 4]에 나타난 것과 같다. 태깅 결과 정답과 일치한 어절은 10,208,461개였고, 한자 변환 결과는 3,375,396개의 어절이 정확히 변환되었다.

5. 결론 및 향후 연구

본 논문에서는 한자 변환 시스템을 구현하기 위해 한국어 태깅 시스템을 이용하여 품사 태깅 문제와 동형이의어 태깅 문제를 해결하도록 구현하였다. 기존의 연구들의 대부분은 한자 처리와 품사&동형이의어 태깅을 별개의 과정으로 취급하였다. 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 ICMA 모듈을 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 세종 말뭉치의 11,117,625어절을 사용하였으며, 태깅 결과 91.82%, 한자 변환 결과 90.35%의 정확률을 나타내었다.

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.

참고 문헌

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

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

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

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

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

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

[7] 황금하, 배선미, 최기선, "전문용어 한글-한자 자동 변환", 한국정보과학회 춘계학술발표논문집, 31권 1호, 2004.

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

[9] 박원병, 김재훈, "유한상태변환기만을 이용한 한국어 형태소 분석 및 품사 태깅", 한국정보처리학회 학술대회논문집, 165-168페이지, 2006.