전문용어 한글-한자 자동 변환*

황금하[1,2O], 배선미[1], 최기선[1]
[1]한국과학기술원 전산학과/전문용어언어공학연구센터/언어자원은행
{hgh[O], sbae, kschoi}@world.kaist.ac.kr
[2]중국 연변과학기술대학

Hangul-Hanja Transfer for Terminology

Jin-Xia Huang[1,2O], Sun-Mee Bae[1], Key-Sun Choi[1]
[1]Dept. of Computer Science, KAIST/KORTERM/BOLA
[2]Yanbian University of Science and Technology, China

요 약

기존의 한글-한자 변환에서는 문맥정보와 통계정보를 고려하지 않는 사전기반의 단어단위 변환 방법을 사용한 반면, 본 논문에서는 언어모델 및 변환모델을 이용한 문장단위의 한자 자동변환 방법을 제안하고, 사전 미등록어와 복합어의 한글-한자 변환을 위하여 단어분할을 변환의 숨김 과정으로 처리하는 통합모델을 사용하였다. 실험 결과, 전문용어의 한글-한자 변환에서 제한된 한자 데이터를 이용하여 기존의 사전기반 변환보다 나은 결과를 얻을 수 있었다.

1. 서 론

한자어 어휘에는 한자표기와 한글표기 두 가지 표기 방식이 있으며, 이 두 방식이 병용되고 있다. 한자표기는 주로 한자어의 의미애매성 해소, 특히 전문용어와 복합어의 어원 표기 및 애매성 해소를 위하여 사용되며, 이를 위하여 한글-한자 자동변환에 대한 연구가 필요한 상황이다.

한자 자동변환에는 여러가지 어려움이 있다. KSX1001에 있는 473개의 한글표기(소리값)은 4888개의 한자표기에 대응되어, 하나의 한글표기에는 적게는 한 개, 많게는 64개의 한자표기가 대응된다[1]. 한글표기된 단어는 그 어원에 따라 한자표기에 대응될 수 있을 뿐만 아니라, 고유어, 외래어에도 대응될 수 있다. 복합어와 미등록어 변환을 위한 복합어 분할작업(tokenization)도 필요하며 훈련용 한자 데이터도 거의 없는 상황이다.

본 연구은 변환모델과 언어모델에 의한 한글-한자 변환 방법을 제안하고, 사전 미등록어와 복합어의 한자 변환을 위하여 단어분할을 모델의 숨김 과정으로 처리하며, 다양한 실험을 통하여 최적의 모델 적용 방법을 찾고자 한다.

아래에서는 한자어 한글표기를 "한글표기"로, 문맥 및 의미상 한자표기로 변환하여야 할 한글표기를 "대상어"로, 변환되지 말아야 할 한글 문자/단어를 "비대상어"로 표현한다.

2. 관련 연구

기존의 한글-한자 변환 방법[2,3]에서는 입력시스템(IME: Input Method Editor)이 모든 가능한 한자후보를 제공해 주고 사용자가 원하는 후보를 선택하는 대화식 변환 방법을 사용하고 있다. 변환 대상어에 대한 인식이 없고, 모든 한글표기에 대하여 한자후보를 제공하며, 한자후보의 제공 순서는 문맥 정보와 통계정보를 고려하지 않고, 단순히 사전 등록 순서에 따르거나[2], 가장 최근에 사용된 한자후보를 우선으로 제공한다 [3]. 복합어에 대한 단어분할에서는 단순한 좌측 최장길이우선 방법을 이용한다.

한글-한자 변환과 비슷한 문제인 중국어 병음(Pinyin) 입력법은 입력된 병음열 P를 중국어 문자열 H로 변환하기 위하여 트라이그램 언어모델 $P(H)$를 적용하였고, 타이핑 모델 $P(P|H)$를 사용하여 실시간 타이핑에서의 타이핑 오류를 반영하였다. 연결된 병음열의 단어분할은 모델의 숨김 과정으로 처리한다 [4].

확률 기반 영-한 음차표기에서는 주어진 영어 단어 E에 대한 한국어 음차표기 K를 찾는 문제를 조건확률 $P(K|E)$ 문제로 간주하여 영-한 변환모델 및 영어 언어모델을 사용하거나[5, 6], 공기확률 문제 $P(E,K)$로 인식하여 영-한 변환모델 및 영어 언어모델을 이용한다[7].

3. 한자 자동 변환

본 연구에서는 한글-한자 변환에 변환모델과 언어모델을 도입하고, 사전 미등록어와 복합어의 한자 변환을 위하여 단어분할을 모델의 숨김 과정으로 처리한다. 모델 적용에서는 단어/문자 단위 모델 적용 과 변환모델 가중치 도입여부 등 사항에 대하여 실험하기로 한다. 기존방법과의 비교를 위하여 사전기반 한글-한자 변환방법을 기준선(base line) 방법으로 사용한다.

3.1 한글-한자 변환 모델

S를 한글 문자열로, S에 대응되는 한자 변환 된 문자열을 T로 표시하면, 한글-한자 변환은 최적의 T^*를 찾는 문제로 된다. 여기에서 T^*는 S에 의미적으로 가장 가까운 한자 변환 문자열로, $P(S,T)$를 최대화 시킬 수 있어야 하는데, 수식으로 표현하면 $T^*=\text{argmax}_T P(S,T)$로 된

* 이 논문은 과학기술부, 과학재단의 지원에 의하여 이루어짐.

다. 여기에서 $P(S,T)$는 조건확률 $P(T|S)$로 표현할 수 있다. 또한, 중국어 병음 입력법에서처럼, 언어모델 $P(T)$를 추가하여 보다 나은 한자 변환 결과를 얻고자 한다(식1).

$$T^* = \arg\max_T P(S,T) \approx \arg\max_T P(T \mid S)P(T) \quad (1)$$

식1 모델은 식2와 같이 전개되고, 이에 변환모델 가중치를 추가하고, 전체 식에 로그를 취할 경우, 식3으로 된다.

$$P(S,T) \approx \prod_{i=1}^{n} P(t_i \mid s_i)P(t_i \mid t_{i-1}) \quad (2)$$

$$T^* \approx \arg\max_T \sum_{i=1}^{n} \{\alpha \cdot \log(P(t_i \mid s_i)) + (1-\alpha) \cdot \log(P(t_i \mid t_{i-1}))\} \quad (3)$$

여기에서, s_i는 한글 원문 문자열에서의 i번째 단어/문자열, t_i는 한자 변환 문자열에서의 i번째 단어/문자열, t_0은 빈 문자 ε이다. 단어 분할은 주어진 모델의 승김과정으로 진행된다.

한글-한자 변환 모델로서 통계적 기계번역에서 흔히 사용하는 잡음 채널 모델(Noisy Channel model) $T^*=\arg\max_T P(S|T)P(T)$를 사용하지 않았는데, 이는 주어진 한자 변환 문자열 T에 대응되는 한글 문자열 S의 $P(S|T)$가 절대부분 1인 점을 감안하였기 때문이다.

3.1.1 모델의 단어/문자 단위 적용

식3의 제안 모델을 문자 단위로 적용할 경우 비대상어에 대한 잘못된 변환이 많아 질 수 있고, 반대로 단어 단위로 적용할 경우 사전 미등록어를 변환하지 못하는 미등록어 오류가 많아진다.

이 두 경우를 평가 하기 위하여, 식3 모델의 적용에서, 서로 다른 적용 단위를 취하여 실험하고 그 결과를 평가하였다. 단어 단위변환에서 s_i는 한글 원문에서의 i번째 단어를 뜻하고, 문자 단위 변환에서 s_i는 i번째의 한 개 이상의 문자를 포함한 문자열을 가르킨다.

3.1.2 변환모델 가중치 및 언어 모델의 적용

앞에서 언급하였듯이, 변환모델에 서로 다른 가중치를 부여하여 변환모델과 언어모델이 한자변환에 대한 영향을 살펴보았다(식3). α 값을 1로 취할 경우, 식3은 변환모델에 의한 한글-한자 변환으로, 0으로 취할 경우 언어모델에 의한 한자 변환 공식으로 된다.

단어 단위의 한자 변환에서 학습데이터의 부족 현상이 심각할 수 있기에, 언어 모델에 의한 단어 단위 한자 변환에서는 유니그램과 이진그램의 두 가지 경우를 고려하여 실험을 진행하여 이들이 한자 변환에 대한 영향을 살펴보았다.

3.1.3 언어 데이터의 이용

데이터 부재 문제 해결을 위하여 한국어 사전에서 한글-한자 어휘쌍을 추출하여 한자어 사전을 구성하고, 이를 코퍼스로 간주하여 사전 데이터 D를 구성한다. 또한, 실험데이터와 같은 분야의 전문용어 한글-한자 어휘쌍을 사용자 데이터 U로 사용하고, 한자 어휘와 중국어 어휘를 코드 변환으로 매핑한 후 한자 데이터 대신 중국어 데이터 C를 이용한다. 이는 중국어 어휘의 출현빈도와 공기정보는 이에 대응되는 한자 어휘의 관련 정보를 반영한다는 가정에 기반한 것이다.

3.2 대상어 품사 제한

대상어의 정확한 인식을 위하여 한국어 형태소 분석기를 사용한다. 또한 대상어 품사의 서로 다른 제한이 한글-한자 변환에 주는 영향을 살펴보기 위하여, 우선 한자 변환 대상어를 체언(n)으로 제한하는 경우와 체언, 용언(p), 수식언(m), 독립언(i), 접사(x)로 제한하는 두 가지 경우를 고려한다.

3.3 사전에 의한 한자어 생성

이 방법은 본 연구의 제안 방법과 비교하기 위한 기준선 방법으로, 주어진 대상어를 한자어 사전에서 검색하여 사전 등록 순서가 첫 번째인 한자후보를 변환결과로 출력한다.

여기에서 사전 크기가 한글-한자 변환에 주는 영향을 살펴보기 위하여 "단일사전"과 "통합사전" 두 가지 사전으로 실험을 진행하였다. 단일사전에는 약 5만개의 한글-한자 단어 엔트리가, 통합사전에는 약 28만개의 한글-한자 단어 엔트리가 있다.

4. 실험 및 토론

본 장에서는 3장에서 제안한 한자 변환 모델과, 모델의 단어/문자단위 적용, 변환모델 가중치 등 서로 다른 고려사항에 대한 실험 및 결과를 설명하고 토론을 가지고자 한다.

4.1 평가 방법

대용량실험에 대한 평가를 지원하기 위하여 자동 평가 프로그램을 구축하였으며 이미 사람에 의하여 정확하게 변환된 한자어 자료를 표준 결과로 사용한다. 평가기준에는 IME에서 일반적으로 사용되는 문자단위 정확도(accuracy) 외에, 사용자 각도에서의 가독성(readability) 평가를 위하여 단어/문장단위 정확도, 단어단위 정확률(precision), 재현률(recall), Dice-coefficient에 의한 유사도 기반 평가 기준 등을 사용했다.

4.2 단어 단위 변환

4.2.1 사전에 의한 변환 및 중국어 데이터 평가

단어 단위 변환 실험에는 단일 사전에 의한 변환(Dic), 통합사전에 의한 변환(BigDic), 유니그램 및 이진그램에 의한 변환 실험이 포함된다(표1). 여기에서의 유니그램과 이진그램은 중국어 데이터 C를 이용하여 추출한 것이다. 실험집합(test set)은 90개 전문용어(180개 형태소)이고, 평가는 사람에 의하여 진행된 단어단위 정확률 재현률 평가이다.

표 1. 재현률, 정확률 및 F-측정치

	Dic	BigDic	unigram	bigram
P	57.1%	50.0%	78.6%	78.6%
R	25.7%	44.0%	70.6%	70.6%
F1	35.4%	46.8%	74.4%	74.4%

단일 사전과 비교시, 통합사전에 의한 변환에서는 사전 확충에 의한 "미등록어 오류"의 감소로 재현률이 큰 폭으로 향상 된 반면, 한자 후보의 증가로 정확률은 오히려 내려 오는 것을 볼 수 있다. F1측정치로 볼 때 큰 사전이 작은 사전보다 나은 결과를 보여준다.

통합사전과 유니그램의 비교에서는 정확률 재현률 모두 큰 폭으로 향상되는 것을 볼 수 있다. 이는 통계정보와 중국어 데이터는 한글-한자 변환에 도움됨을 설명한다. 다만 이진그램이 유니그램에 비해 성능 향상이 없는데, 이는 단어 단위 언어 모델을 이용한 한글-한자 변환에서 데이터 희귀성 문제가 여전히 심각함을 보여준다.

4.2.2 대상어 품사 제한 및 사전 사용자 중국어 데이터 평가

본 실험은 우선 사전 데이터 D의 유니그램을 이용한 한글-한자 변환에서, 대상어 품사를 체언으로 제한한 경우와, 체언, 용언, 수식언, 독립언, 접사로 확장하는 경우를 비교하여, 대상어 품사 제한이 실험에 주는 영향을 살핀다. 다음 사전 데이터 D, 사용자 데이터 U, 중국어 데이터 C가 한글-한자 변환에 주는 영향을 평가하는데 그 목적

이 있다. 실험집합은 5127개 전문용어(12786개 형태소, 형태소당 한자 후보수 4.67)이고, 사용자 데이터U는 실험집합과 같은 데이터이다. 대용량 실험 평가를 지원하기 위하여 Dice-coefficient 기반의 유사도 평가기준을 사용한다.

표2에서 행 TS는 한자 변환 대상어를 대상으로한 유사도 계산결과이고, WS는 전체 실험 데이터를 대상으로한 유사도 계산 결과이다. 칼럼 D-N과 D가 보여주는 바와 같이, 전문용어 한자 변환에서, 대상어 제한을 확대한 경우인 D는 체언으로 제한한 D-N보다 나은 결과를 보인다.

표2. 대상어 제한 및 언어 데이터 평가

	D-N	D	U	C	DC	DU	UC	DUC
TS	0.71	0.75	0.81	0.72	0.75	0.82	0.82	0.81
WS	0.76	0.80	0.85	0.77	0.90	0.85	0.85	0.85

다음은 데이터 평가인데, 닫힌 실험(close test)이기에 당연한 결과이지만, 사용자 데이터 U는 사전 데이터 D 보다 나은 결과를 보였다. 또한 사전 데이터 D는 중국어 데이터 C보다 나은 결과를 보이고, 두 개 이상 데이터의 공동 사용에서는 표2의 D⇔C, D⇔DC, DU⇔UC, DU⇔DUC 비교에서 볼 수 있듯이, 중국어 데이터는 사전데이터로 어느정도 대체할 수 있다. 중국어 데이터가 뉴스분야로서 실험데이터와 일치하지 않기 때문일 수도 있다. 일단 이런 결론에 따라 이후 실험에서 중국어 데이터를 더 이상 사용하지 않았다.

4.2.3 단어단위 모델에서의 가중치 평가

단어단위 모델 평가에서는 α에 서로 다른 가중치를 부여함으로 언어모델에 의한 변환(α=0), 통합 모델에 의한 변환(α=0.5), 변환모델에 의한 변환(α=1)을 평가하였다. 본 실험에서 사용된 데이터는 4.2.2의 실험과 같은 실험집합으로, 사전데이터와 사용자데이터를 이용한 닫힌 실험이고, 자동 평가 방법을 이용하였다. 표3이 보여주는 바, 변환모델이 가장 좋은 결과를 보여주고 있다.

표3. 단어 단위 제안 모델 평가

	α=0	α=0.5	α=1
P	78.6%	76.52%	84.80%
R	70.6%	70.70%	77.31%
F1	74.4%	73.4%	80.8%

4.3 문자 단위 변환

단어단위 모델 변환은 앞에서 최적결과를 보였던 변환모델에 의한 단어단위 변환(α=1)을 비교대상으로 삼는다. 문자단위 변환 실험에서는 제안된 모델에 대하여 변환모델 가중치 α에 대하여 서로 다른 값을 부여함으로 최적 모델을 찾는다. 실험집합은 1000개의 전문용어(2727개 단어 포함, 한자어 당 한자후보 수는 3.9), 사용 데이터로는 사전 데이터 D와, 실험집합과 같은 분야의 12,000개 전문용어 사용자 데이터 U를 사용하였고, 전체실험은 열린 실험(open test)이다.

표4의 첫번째 칼럼은 평가기준이고, CA, WA, SA는 각기 문자/단어/문장단위 정확도를, WS는 전체 문서 대상 유사도 기반 평가를, F1은 정확률 재현률에 기반한 F1측정치 등 평가 기준을 나타낸다. 첫번째 행은 변환방법과 사용데이터로, 기준선 방법인 사전기반변환 Dic외에, 모두 모델 기반 변환이다. W가 부착된 것은 단어단위모델 적용, 없는 것은 문자단위 모델적용, D는 사전데이터, U는 사용자 데이터, C는 중국어데이터, 이 뒤의 숫자는 변환모델 가중치 α를 표시한다.

우선 기준선 방법인 사전기반 변환방법Dic를 다른 모델 기반 방법과 비교시, 모델이 나은 결과를 보여준다. 다음 사용자데이터가 한글-한자 변환에 중요한 역할을 하는 것을 볼 수 있는데, 칼럼 Dw1⇔DUCw1, D5⇔DU5의 비교에서 확인 된다. 또 문자단위 변환 모델에서, DU0에서 DU1(α=0 ~1.0)까지의 칼럼을 비교하면, DU1, 즉 변환모델에서 가장 좋은 결과를 보이는 것을 볼 수 있다. 또한 DUCw1⇔DUx의 비교결과, 문자 단위 모델이 대체적으로 단어단위 모델보다 나은 결과를 보여주고 있다.

표4. 문자단위 변환 실험 및 전체 모델별 비교 평가

%	Dic	Dw1	DUCw1	D5	DU0	DU2	DU5	DU8	DU1
CA	62.9	69.1	75.0	73.1	81.0	89.3	90.2	91.0	91.4
WA	49.9	73.8	75.3	64.6	72.4	77.1	82.3	82.1	81.4
SA	18.8	43.4	51.2	34.7	48.2	67.0	67.5	67.1	68.1
WS	68.4	75.5	79.7	77.9	82.5	90.4	91.2	91.7	92.:
F1	39.0	65.6	69.7	51.2	60.8	75.7	75.9	75.9	76.2

특기 할 것은 Dw1⇔D5의 비교이다. 사전 데이터만 사용할 경우, 문자단위 모델은 단어단위 모델에 비해 문자단위 정확도 CA에서 나은 결과를 보이나, 문장단위 정확도 SA와 정확률/재현률을 나타내는 F1-측정치에서는 반대된 결과를 보이고 있다. 오류분석 결과, 사용자 데이터가 충분할 경우 문자단위 모델이 단어단위 모델보다 나은 결과를 보이는 반면(DUx⇔DUCw1), 사용자 데이터가 부족할 경우, 문자단위 모델은 많아진 잡음으로 더 많은 단어 내 부분 변환 오류를 나타내어, 단어 단위 모델 적용보다 못하다는 결론을 얻게 되었다.

5. 결론 및 향후 연구

본 논문은 한자 자동변환에 통계적 방법을 도입하였고, 단어/문자단위 모델 적용, 변환모델 가중치 도입, 대상어 품사제한 등 여러가지 고려사항에 대한 실험을 진행하였다. 전문용어 어원표기를 위한 한자 자동변환 실험에서 기준선 방법인 사전기반 방법과 비하여 본 연구의 제안 방법이 뚜렷한 진보를 보였다.

본 논문의 실험은 전문용어 분야에만 제한되어있으나 일반분야에서도 같은 방법의 적용이 가능하다. 향후 일반 분야의 문장단위 자동변환을 위하여 학습데이터와 실험데이터를 준비하고 제안된 모델이 일반분야에서의 한자 변환에 대한 평가를 진행하여야 한다. 또한 중국어 데이터뿐만 아니라 일본어 데이터 활용 가능성에 대한 연구도 필요하다.

참고문헌

[1] 김경석. 2003. KSX1001에 있는 한자의 소리값, http://asadal.cs.pusan.ac.kr/hangeul/code/ksx1001-name-hj-stat-v03.txt

[2] "한글97", 한글과 컴퓨터 사

[3] "한글입력기(IME2002)", Microsoft사

[4] Zheng Chen & Kai-Fu Lee. 2000. A New Statistical Approach To Chinese Pinyin Input. The 38th Annual Meeting of the Association for Computational Linguistics

[5] 이재성. 1999. 다국어 정보검색을 위한 영-한 음차표기 및 복원 모델, 박사학위논문, 한국과학기술원 전산학과

[6] 김정재, 이재성, 최기선. 1999. 신경망을 이용한 발음단위기반 자동 영한 음차표기 모델", 1999년 한국인지과학회 춘계학술대회

[7] SungYoung Jung, SungLim Hong & EunOk Paek. 2000. An English to Korean Transliteration Model of Extended Markov Window. 18th International Conference on Computational Linguistics