언어 사용환경에 적응적인 영어 문맥의존 철자오류 교정 기법*

김민호○, 김경식, 권혁철
부산대학교, 전자전기컴퓨터공학과
karma@pusan.ac.kr, jinjzh1007@pusan.ac.kr, hckwon@pusan.ac.kr

Adaptive English Context-Sensitive Spelling Error Correction Techniques for Language Environments

Minho Kim○, Jingzhi Jin, Hyuk-Chul Kwon
Pusan National University, Dept. of Electrical and Computer Engineering

요 약

문서 교정기에서 문맥의존 철자오류를 교정하는 방법은 크게 규칙을 이용한 방법과 통계 정보를 이용한 방법으로 나뉜다. 한국어와 달리 영어는 오래전부터 통계 모형에 기반을 둔 문맥의존 철자오류 교정 연구가 활발히 이루어졌다. 그러나 대부분 연구가 문맥의존 철자오류 교정 문제를 특정 어휘 쌍을 이용한 분류 문제로 간주하기 때문에 실제 응용에는 한계가 있다. 또한, 대규모 말뭉치에서 추출한 통계 정보를 이용하지만, 통계 정보 자체에 오류가 있을 경우를 고려하지 않았다. 본 논문에서는 텍스트에 포함된 모든 단어에 대하여 문맥의존 철자오류 여부를 판단하고, 해당 단어가 오류일 경우 대치어를 제시하는 영어 문맥의존 철자오류 교정 기법을 제안한다. 또한, 통계 정보의 오류가 문맥의존 철자오류 교정에 미치는 영향과 오류 발생률의 변화가 철자오류 검색과 교정의 정확도와 재현율에 미치는 영향을 분석한다. 구글 웹 데이터에서 추출한 통계 정보를 바탕으로 통계 모형을 구성하고 평가를 위해 브라운 말뭉치에서 무작위로 2,000문장을 추출하여 무작위로 문맥의존 철자오류를 생성하였다. 실험결과, 문맥의존 철자오류 검색의 정확도와 재현율은 각각 98.72%, 95.79%였으며, 문맥의존 철자오류 교정의 정확도와 재현률은 각각 71.94%, 69.81%였다.

주제어: 문맥의존 철자오류, 노이지 채널 모형, 언어 모형, 문서 교정

1. 서론

문서 교정기에서 다루는 맞춤법/문법 오류는 단순 철자오류(non-word spelling error)와 문맥의존 철자오류(context-sensitive spelling error, real word spelling error)로 나눈다. 단순 철자오류는 사전(dictionary)에 등재되지 않은 어휘가 사용된 오류로서 텍스트를 형태적으로 분석하는 것만으로 쉽게 오류를 검색할 수 있다. 반면, 문맥의존 철자오류는 단어(word) 단위로 볼 때는 바르지만, 좌우 문맥을 고려했을 때 오류가 되는 것이다. 이는 문맥의 의미, 통사적 관계를 고려해야만 해당 어휘를 판단할 수 있기에 검색 난도가 매우 높다. 또한, 단순 철자오류라 하더라도 대치 가능한 후보가 여러 개일 경우 교정의 난도는 문맥의존 철자오류와 같다. 따라서 문맥의존 철자오류 교정 기법의 성능이 문서 교정기 전체의 성능을 좌우한다고 볼 수 있다.

문맥의존 철자오류 교정은 규칙을 이용한 교정 방법과 통계 정보를 이용한 교정 방법으로 나눌 수 있다. 규칙기반 교정 방법은 규칙의 구축 방법에 따라 수동/반자동/자동 생성 규칙 기반 교정으로 나눌 수 있다. 그러나 이들 모든 언어 현상을 반영하는 규칙을 만드는 것이 현실적으로 불가능하므로 실제 응용에 적용하기는 어렵다. 즉, 발생빈도가 높거나 정형화된 오류는 규칙기반 방법으로 교정할 수 있는 확률이 높으나, 입력 오류로 일어나는 비정형화된 오류교정은 규칙기반 방법만으로는 불가능하고 교정 난도가 훨씬 높다.

한국어와 달리 영어는 오래전부터 통계 모형에 기반을 둔 문맥의존 철자오류 교정 연구가 활발히 이루어졌다. 그러나 대부분 연구가 문맥의존 철자오류 교정 문제를 특정 어휘 쌍을 이용한 분류 문제로 간주하기 때문에 실제 응용에는 한계가 있다. 또한, 대규모 말뭉치에서 추출한 통계 정보를 이용하지만, 통계 정보 자체에 오류가 있을 경우를 고려하지 않았다. 본 논문에서는 텍스트에 포함된 모든 단어에 대하여 문맥의존 철자오류 여부를 판단하고, 해당 단어가 오류일 경우 대치어를 제시하는 영어 문맥의존 철자오류 교정 기법을 제안한다. 또한, 통계 정보의 오류가 문맥의존 철자오류 교정에 미치는 영향과 오류 발생률의 변화가 철자오류 검색과 교정의 정확도와 재현율에 미치는 영향을 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 문맥의존 철자오류 교정의 연구현황을 분석하고, 3장에서는 본 논문에서 제안하는 전체 어절을 대상어로 한 통계적 문맥의존 철자오류의 검색 및 교정 모형과 실험방법을 설명

* 본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [R0101-15-0176, Symbolic Approach 기반 인간모사형 자가학습 지능 원천 기술 개발]

한다. 4장에서는 실험 결과를 제시한다. 마지막으로 5장에서는 결론과 향후 연구에 관해 설명한다.

2. 관련 연구

문맥의존 철자오류를 교정하는 방법은 크게 규칙을 기반으로 한 방법과 통계정보를 기반으로 한 방법으로 나눈다.

규칙을 기반으로 한 방법은 주로 사람이 규칙을 만드는 방법이다. 추가된 규칙이 많을수록 성능은 향상되지만, 이는 고도의 지식을 갖춘 전문가가 필요하고 오랜 시간을 거쳐 규칙을 추가해야 하므로 그 비용이 엄청나다. 또한, 발음 유사성에 의한 오류나 문법적 차이의 모호성에 따른 오류와 같이 정형화된 오류 유형에만 잘 동작하다. 이런 원인으로 규칙을 기반으로 한 방법은 문맥의존 철자오류의 검색과 교정에 높은 정확도(precision)를 보이지만 재현율(recall)은 낮다[1-4].

통계정보를 기반으로 한 방법은 영어를 대상으로 많이 연구되었다. 문맥의존 철자오류 문제를 어의 중의성 해소(word sense disambiguation)와 같은 문제로 간주하고 문제를 해결한다[5-9]. 이들 방법은 N-gram과 같은 언어모형에 기반을 두고 대상어와 주변 문맥에 나타난 어휘 간 의존관계를 통계적으로 계산하여 문맥의존 철자오류를 검색하고 교정하는 방법이다[7-9]. 그러나 {peace, piece}와 같이 오류 발생 유형에 따라 특정 어휘 쌍을 미리 선정하고 문제를 해결하기 때문에 실제 응용에는 한계가 있다. 또한, 대규모 말뭉치에서 추출한 통계 정보를 이용하지만, 통계 정보 자체에 오류가 있을 경우를 고려하지 않았다.

본 논문에서는 텍스트에 포함된 모든 단어에 대하여 문맥의존 철자오류 여부를 판단하고, 해당 단어가 오류일 경우 대치어를 제시하는 영어 문맥의존 철자오류 교정 기법을 제안한다. 또한, 통계 정보의 오류가 문맥의존 철자오류 교정에 미치는 영향과 오류 발생률의 변화가 철자오류 검색과 교정의 정확도와 재현율에 미치는 영향을 분석한다.

3. 노이지 채널 모형에 기반을 둔 교정 모형

전체 어절을 대상어로 한 문맥의존 철자오류의 검색과 교정은 기존 연구를 기반으로 한 노이지 채널 모형으로 해석하고 확률적으로 수식 (1)의 $\widehat{D^I}$를 구하는 것이다.

$$\widehat{D^I} = \arg max_{D^I} P(D^I|D^O) \quad (1)$$

여기서 D^I는 사용자가 의도한 문서이고 D^O는 의도한 문서 D^I가 어떤 노이즈(noise)로 인하여 작성된 문서이다. 베이즈 규칙(Bayes' rule)을 적용하고 상수인 분모를 제거하면 철자오류 교정의 노이지 채널 모형인 수식 (2)를 얻을 수 있다.

$$\widehat{D^I} = \arg max_{D^I} P(D^O|D^I)P(D^I) \quad (2)$$

수식 (2)에는 두 개의 확률분포가 존재하는데, 언어모형(language model)인 $P(D^I)$와 채널 확률(channel probability)인 $P(D^O|D^I)$이다. 기존 연구에서는 언어모형으로 n-gram을 사용하거나 단어 간 조건부 확률을 사용한다. 본연 구에서는 언어 모형으로서 n-gram을 사용하며, 채널 확률을 오류 발생률로 보았다.

수식 (3)은 수식 (1)과 (2)의 D^I와 D^O를 단어의 집합으로 나타낸 것이며, 오류는 나타날 수도 있고 ($w_i \neq w_o$) 나타나지 않을 수도 있다($w_i = w_o$).

$$D^I = \{w_1^{i-1}, w_i, w_{i+1}^n\}$$
$$D^O = \{w_1^{i-1}, w_o, w_{i+1}^n\} \quad (3)$$

수식 (3)에 의해서 정확한 문서 D^I를 추정하는 디코딩 문제는 정확한 단어 w_i를 추정하는 문제, 즉 수식 (4)로 간소화(simplify)될 수 있다.

$$\widehat{w_i} = \arg max_{w_i} P(w_o|w_i) \times \prod_{\substack{k=1 \\ k \neq i}}^{n} P(w_k|w_i) \times P(w_i) \quad (4)$$

4. 실험 및 평가

4.1 실험환경

기존 연구에서는 문맥의존 철자오류의 발생 유형에 따라 미리 교정 어휘 집합(confusion set)을 선정하고, 문맥에 어울리는 어휘 집합 내 단어를 선택하는 분류 문제로 보았다. 하지만 문서 내에서 발생할 수 있는 모든 오류 유형*에 대하여 교정 어휘 집합을 생성하는 것은 불가능하므로 기존 연구 결과를 실제 응용 시스템에 적용하기에는 한계가 있다.

본 논문에서는 이런 한계점을 극복하고자 문서 내의 모든 단어를 오류 발생 단어로 판단하고, 그 단어가 오류일 경우 대치 가능한 모든 후보 단어를 생성한다. 발음 유사성에 따른 오류는 soudex를 기준으로 후보를 생성하고, 오타 오류와 띄어쓰기 오류는 edit distance를 기준으로 후보를 생성한다. 다만, 문법 오류는 어느 정도 정형화되어 있으므로 정형화된 교정 어휘 집합을 사용한다.

예를 들면, 'mail'의 경우 해당 단어가 오류라고 판단되었을 때, 생성 가능한 후보는 'nail', 'maeil',

* 문맥의존 철자오류 유형은 철자는 다르가 발음이 같거나 유사하여 발생하는 발음 유사성에 따른 오류, 오타에 의해 발생하는 오타 오류, 사용자가 문법의 차이를 정확히 알지 못해서 발생하는 문법 오류, 그리고 단어 사이의 잘못된 공백 때문에 발생하는 띄어쓰기 오류로 구분할 수 있다.

‘meil’, ‘maol’ 등이 있다. 이들 후보는 발음이 비슷하거나 오타에 의해서 발생할 수 있는 단어들이다. 다만, 실제 사전에 존재하는 단어만이 후보가 될 수 있으므로, ‘maol’과 같이 unigram 사전에 존재하지 않는 단어는 후보에서 제거된다.

Google 1T Web data 이용하여 N-gram 모형을 구축하였으며, 자료구조는 Google이 공개한 leveldb[10]를 사용하였다. 표 1은 N-gram 모형의 사전 크기와 메모리 사용량이다.

표 1 N-gram 모형의 사전 크기와 메모리 사용량

	사전 크기(억)	메모리 사용량(GB)	Perplexity
Unigram	0.2	0.2	962
Bigram	3.0	3.0	170
Trigram	5.0	12.4	109

평가 데이터는 브라운 말뭉치에서 2,000문장을 무작위로 선택하였고, 정확한 문장에서 문맥의존 철자오류인 오류문장을 무작위로 약 20,000문장을 생성하여 평가 데이터를 구성하였다.

오류 교정은 검색(detection)과 교정(correction)으로 나뉜다. 오류가 정확히 검색되었더라도 많은 교정 후보 중에서 확률이 가장 높은 후보로 교정하기에 정확히 교정이 아닐 수 있다. 검색과 교정 모두 오류 교정에서 중요한 지표이다. 평가 척도는 정확도, 재현율, 그리고 F1-measure를 활용했다.

확률 추정은 최대 가능도 추정(maximum likelihood estimation; MLE)을 사용하였고 adding estimation, interpolation, backoff를 사용해 평탄화(smoothing)를 하였다.

4.2 실험결과

언어 모형에 따른 성능 변화를 관찰하고자 tri-gram 모형과 조건부 확률 모형을 서로 비교하였다. 이때 채널 확률(=오류 발생확률)은 5%로 가정하였다. 조건부 확률 모형의 경우 대상어를 중심어로 좌우 3단어를 사용하였으며, 평탄화를 위해서 KN 스무딩을 이용하였다*. 조건부 확률 모형은 오류 발생 여부 판단의 대상이 되는 중심어와 주변 문맥 간 조건부 확률로 언어 모형은 산정할 것이다.

표 2. 언어 모형에 따른 문맥의존 철자오류 교정 성능

	Trigram		조건부 확률	
	Detection	Correction	Detection	Correction
Precision	99.58%	93.43%	98.76%	95.24%
Recall	89.54%	83.65%	87.58%	84.46%

* 윈도우 사이즈에 변화를 주고자 다양한 스무딩 방법을 적용한 실험 결과 윈도우 사이즈가 좌우 3단어이고, KN 스무딩을 사용하였을 때 성능이 가장 좋았다.

표 2에서 확인할 수 있듯이 교정과 검색 모두 정확도가 재현율보다 높다는 것을 알 수 있다. 이는 일반적인 정보검색 시스템의 성능 양상과 같다고 볼 수 있다. 다만, 노이지 채널 모형의 특성상 채널 확률(=오류 발생률)을 조절함에 따라 정확도와 재현율에 변화를 둘 수 있다. 즉, 오류 발생률이 매우 낮다고 가정하면 정확도를 100%에 가깝도록 만들 수 있으며, 오류 발생률이 높다고 가정하면 정확도가 떨어지더라도 재현율이 높은 문서 교정기를 만들 수 있다. 따라서 응용 시스템의 전처리기에서 자동 대치의 목적으로 사용될 때는 정확도가 높은 문서 교정기를 구성하는 것이 유리하고, 일반 사용자에게 제공되는 워드 프로세서에 삽입된 문서 교정기의 경우에는 사용자의 피드백을 받을 수 있으므로 재현율이 높은 교정기를 구성하는 것이 유리하다.

표 2에의 결과에서 확인할 수 있는 또 다른 사실은 조건부 확률 모형을 언어 모형으로 사용하는 것이 trigram 모형을 사용한 것보다 재현율이 높다는 것이다. 그러나 이는 perplexity를 보아도 알 수 있듯이 데이터의 양이 부족할 때는 자료부족 문제 때문에 trigram의 모형의 성능이 낮을 수 있다. 따라서 데이터의 양이 충분하다면 정확도와 재현율 모두 높은 문서 교정기를 만들 수 있다.

표 3은 Google 1T Web data에서 추출한 unigram 빈도의 예이다. ‘frpm’과 같이 ‘from’의 오타로 생성될 수 있는 후보의 빈도가 높다는 것을 알 수 있다. 실제 ‘frpm’은 ‘from’의 오류일 수도 있고, 아니면 고유명사의 약어일 수도 있다.

표 4는 후보 생성을 위해 적용한 사전에 따른 성능 비교이다. 즉, 단순 철자오류를 후보에서 제거하기 위해 Google 1T web data에서 구성한 unigram 사전과 일반 사

표 3. Google 1T Web data의 오류 예

순번	Unigram	빈도
…		
891448	e250	8,931
891449	emich.edu	8,931
891450	enkindle	8,931
891451	epaulet	8,931
891452	**frpm**	**8,931**
891453	getf	8,931
…		

표 4. 후보 생성을 위해 적용한 사전에 따른 성능

	Unigram 사전		옥스퍼드 사전	
	Detection	Correction	Detection	Correction
Precision	99.58%	93.43%	98.72%	95.79%
Recall	89.54%	83.65%	71.94%	69.81%

전 중 어느 것을 사용하는 것이 유리한지를 판단하기 위해서이다.

표 4에서 알 수 있듯이 문맥의존 철자오류 검색에서는 unigram 사전을 이용하는 것이 유리하며, 교정에서는 옥스퍼드 사전과 같은 일반 사전을 이용하는 것이 유리하다는 것을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 n-gram 모형을 노이지 채널 모형의 언어 모형으로 사용하고, 채널 확률인 문맥의존 철자오류 발생률을 사용자 적응형 문맥의존 철자오류의 검색과 교정을 위한 장치로 활용했다. 본 논문에서 제안한 문맥의존 철자오류 교정 기법은 95% 이상 정확도, 70% 이상 재현율의 성능을 보였다. 이는 특정 어휘 쌍을 사용한 기존 연구와 비슷하거나 더 높은 수준의 성능을 보임으로서 본 논문에서 제안한 문서 교정 기법을 실제 응용에도 적용할 수 있음을 보인다.

향후 연구에서는 다양한 평탄화 방법을 적용하여 성능을 높이고, 실제 응용을 위해 메모리와 실행 속도 최적화를 진행할 예정이다. 또한, 더욱 정확한 성능을 평가하려면 오류어와 대치어가 분석된 말뭉치가 필요하기에, 평가용 말뭉치를 구축할 예정이다.

참고문헌

[1] 김민호, 최현수, 권혁철, 윤애선. “한국어 어휘의미망을 이용한 문맥 철자 오류 규정규칙의 일반화,” *한국정보과학회 학술발표논문집*, 653-655. (2013)

[2] 최현수, 윤애선, 권혁철. “조사제약 조건의 완화에 의한 문맥의존 철자오류 교정의 재현율 향상 방식,” *정보과학회논문지: 소프트웨어 및 응용*, 41(3), 249-256. (2014)

[3] 최현수, 윤애선, 권혁철. “통합적 방식을 이용한 한국어 문맥의존 철자오류 교정규칙의 재현율 향상,” *한국정보과학회 학술발표논문집*, 577-579. (2014)

[4] 최현수, 권혁철, 윤애선. “동적 윈도우를 갖는 조건부확률 모델을 이용한 한국어 문맥의존 철자오류 교정 규칙의 재현율 향상,” *한국정보과학회 학술발표논문집*, 420-422. (2014)

[5] Golding, Andrew R. and Dan Roth and J. Moon. “A Winnow-Based Approach to Context-Sensitive Spelling correction,” *Machine Learning*, Vol. 34, 107-130. (1998)

[6] 김민호, 권혁철, 최성기. “어절 N-gram을 이용한 문맥의존 철자오류 교정,” *정보과학회논문지*, 41(12), 1081-1089. (2014)

[7] A. Islam and D. Inkpen. “Semantic text similarity using corpus-based word similarity and string similarity,” *ACM Transactions on Knowledge Discovery from Data*, Vol. 2, No.2, 1-25. (2008)

[8] A. Islam and D.Inkpen. “Real-Word Spelling Correction using Google Web 1T 3-gram,” *Proc. of International Conference on Natural Language Processing and knowledge Engineering*, Vol. 3, 1241-1249. (2009)

[9] W.-O. Amber, G. Hirst, and A. Budanitsky. “Real-word spelling correction with trigrams: a reconsideration of the Mays, Damerau, and Mercer model,” *Pro. of 9th International Conference on Intelligent Text Processing and Computational Linguistics*, Vol. 4919, 605-616. (2008)

[10] LevelDb [Online]. Available: http://leveldb.org (accessed 2015, Mar. 07)