좌우 어절 N-gram 및 음절 N-gram을 이용한 간섭 오타 교정 방법

손성환[O], 강승식

국민대학교, 자연어처리 연구실
ssh121@kookmin.ac.kr, sskang@kookmin.ac.kr

Interference Typo Correction Method by using Surrounding Word N-gram and Syllable N-gram

Sung-Hwan Son[O], Seung-Shik Kang
Kookmin University, Natural Language Processing Lab

요 약

스마트폰의 쿼티 자판 소프트 키보드의 버튼과 버튼 사이 좁은 간격으로 인해 사용자가 의도치 않은 간섭 오타가 발생하는 것에 주목하였다. 그리고 오타 교정의 성능은 사용자의 관점에서 얼마나 잘 오타를 교정하느냐도 중요한 부분이지만, 또한 오타가 아닌 어절을 그대로 유지하는 것이 더 중요하게 판단될 수 있다. 왜냐하면 현실적으로 오타인 어절 보다 오타가 아닌 어절이 거의 대부분을 차지하기 때문이다. 따라서 해당 관점에서 교정 방법을 바라보고 연구할 필요가 있다. 이에 맞춰 본 논문에서는 대용량 한국어 말뭉치 데이터를 가지고 확률에 기반한 한국어 간섭 오타 수정 방법에 대해 제안한다. 제안하는 방법은 목표 어절의 좌우 어절 N-gram과 어절 내 좌우 음절 N-gram 정보를 바탕으로 발생할 수 있는 간섭 오타 교정 후보들 중 가운데서 가장 적합한 후보 어절을 선택하는 방법이다.

주제어: 간섭 오타, N-gram, 스마트폰, 쿼티 자판

1. 서론

스마트폰의 보급이 점차 증가하여 그 동안 데스크탑 PC를 사용하여 진행했던 일들을 스마트폰으로 대신 처리하는 사람들이 늘어났으며[1], SNS의 발전과 호환성의 증대에 따라서 인터넷 상에서 스마트폰의 의한 데이터들이 기하급수적으로 생성되고있다.

스마트폰의 작은 크기로 인한 휴대성과 터치스크린을 통한 부드러우며 다기능 수용에 탁월한 인터페이스는 큰 장점으로 꼽히지만, 그와 동시에 입력에 있어서 터치 오류를 유발시키는 원인이기도 하다[2]. 대부분의 한국 스마트폰 운영체제인 안드로이드는 한글 입력 방식으로 천지인과 더불어 쿼티 자판을 제공하고 있는데, 특히 쿼티 자판은 버튼과 버튼 사이 간격이 좁고, 터치방식의 특성상 버튼과 버튼의 경계를 촉각으로 느낄 수 없어서 사용자가 누르려는 버튼 좌우의 버튼이 눌려지는 간섭 오타가 발생하기가 쉽다[3]. 예를 들어 '저장'이라는 단어에서 첫 음절 '저'만 보자면, 'ㅈ' 좌우의 'ㅂ', 'ㄷ'가 잘못 눌려 '버장'이나, '더장'이 되거나, 'ㅓ'가 좌우의 'ㅗ', 'ㅏ'로 잘못 눌려 '조장', '자장'이 될 수 있다. 실제로 터치스크린에 대한 연구 결과에 따르면, 터치 대상의 크기가 작아질 수록 오류가 증가하는 경향을 보였다는 결과가 있다[4]. 따라서 스마트폰으로 인한 사용자가 의도하지 않은 오타가 많이 발생하며, 이는 데이터의 잡음 요소 중 하나로 자리잡았다.

스마트폰으로 인한 오타 발생 확률 감소 및 수정에 대하여 많은 연구들이 있다. 오타 발생 확률 감소에 있어서는 입력 자판 구성 및 형태, 방식을 바꾸거나[5,6], 입력 도중에 후보 단어를 추천하는 방법[7]들이 존재한다. 이러한 방법들은 오타가 발생할 확률을 줄여주는 데 도움을 주지만, 이미 발생한 오타에 대한 교정 방법이 필요하다.

오타 수정 방법에는 크게 규칙 기반 모델과 통계에 기반한 모델로 나눌 수 있다. 규칙 기반 모델은 새로운 문법이나, 신조어에 대한 추가적인 규칙을 추가함에 따라 처리 속도가 기하급수적으로 증가하는 단점이 있다. 또한 스마트폰에서 발생하는 비정형 언어에 굉장히 취약하다[8]. 반면 통계 기반 모델은 다양하고 충분한 데이터만 있다면, 보다 빠르고 융통성있게 오타를 교정하는 것이 가능하다.

따라서 본 논문에서는 문맥을 고려하기 위해 교정 어휘 대상 어절의 좌우 어절 N-gram과 좌우 음절 N-gram을 사용한 다양한 통계 기반 모델로 오타를 교정하는 것에 대해 기술할 것이다.

2. 관련 연구

1948년 Shannon의 잡음 채널 모델은 철자 오류 교정 이외에도 광범위한 문제에 성공적으로 적용되었다[9]. 1991년 Mayes, Damerau 외의 연구자들은 철자 오류 교정을 위해 '삽입', '삭제', '교체', '문자 쌍 위치 변환'이라는 4가지의 혼동 세트(confusion set)를 정의하였고[10], 2000년 Brill, Moore는 진행된 연구들을 바탕으로 베이지안 규칙을 적용하여 철자 오류 교정을 위한 향상

된 잡음 채널 모델을 제시하였다[11]. 향상된 잡음 채널 모델은 N-gram을 사용한 언어 모델을 접합시켜 문맥에 알맞게 교정하도록 하였다.

이어 2014년 김민호, 권혁철, 최성기는 편집거리가 1에 해당하는 어휘들을 '교정 어휘 쌍'으로 선정하여 교정 후보 어휘로 사용하였다[8]. 교정 대상 어절과 편집거리가 1인 교정 후보 어휘를 비교하여 오타일 확률을 계산하였으며, 대용량 한국어 말뭉치에서 구축한 어절 N-gram 데이터를 사용해 교정 어휘 대상 어절 이전 문맥과 이후 문맥을 구분하고 교정 여부 및 교정 어절을 도출해냈다.

그리고 2018년 Kirk는 터치스크린 쿼티 자판에서 잡음 채널 모델을 바탕으로 자판의 문자 버튼 간 물리적인 거리 정보를 추가함으로써 오타를 교정하는 모델에 대해 기술하였다[12].

본 논문에서는 해당 연구들을 참조하여, 스마트폰 터치스크린의 쿼티 자판에서 발생하는 간섭오타에 초점을 두고 연구를 진행했다. 따라서 대용량 한국어 데이터에서 양방향의 어절 및 음절 N-gram 데이터를 구축하고, 발생할 수 있는 교정 후보 어절을 간섭오타의 범위로 축소하였으며, 하나의 문장에서 2개 이상의 오타가 발생할 수 있다는 가정하에 교정 대상 어절의 앞 어절들과 뒷 어절들 간의 N-gram 확률로 교정 여부 및 교정 후보 어절을 선택하게 하였다. 그리고 추가로 유니그램에 존재하지 않는 어절에 한하여, 앞뒤 음절 N-gram을 사용해서 대상 어절의 교정 여부 및 교정 후보 어절을 선택하게 한 교정 모델에 대한 방법론 및 실험에 대해 기술하였다.

3. 간섭 오타 교정 모델

3.1 데이터

대용량 한국어 말뭉치로는 KCC150(Korean Consistent Corpus: 약 1억 5천만 어절)을 사용하여 앞뒤 어절 N-gram, 음절 N-gram 데이터를 구축하였다.

그리고 실험 성능을 테스트하기 위해 간섭 오타 데이터를 생성했다. 오타 데이터는 다음 규칙에 의해 생성되었다. 첫째 하나의 문장에서 n개가 무작위로 생성되나, 문장 전체 어절 개수의 25%를 넘을 수 없다. 둘째 발생되는 오타는 하나의 어절 내 무작위 하나의 음절에서 간섭 오타로 자소 단위 1개가 완성형 한글로 발생한다. 오타 생성 규칙에 따라 KCC150과 KCC150 이외의 중앙일보 신문 사회면 기사에서 최소 12어절 이상의 문장을 각각 추출하여 문장 당 3 어절의 간섭 오타를 생성하였다.

3.2 좌우 N-gram 교정 모델

우선 입력 문장의 각 어절의 간섭 오타를 생성한다. 생성한 간섭 오타에서 유니그램 개수가 1개 이하인 것들은 제외하여, 문장 각 어절의 교정 후보 어절들을 구축한다. 그리고 문장의 양 끝에 사용할 N-gram에 알맞게 문장 시작 토큰과 끝 토큰을 추가한다. 문장의 좌측에서부터 각 어절을 교정 대상 어절로 보고 다음 수식들을 통해 교정 여부 및 교정 후보 어절들 중에서 적합한 어절을 선택하였다.

$$Correction_{word}(C) = \frac{(P_f(C|w_{n-2},w_{n-1}) + P_b(C|w_{n+1},w_{n+2}))}{2} \quad (1)$$

식 (1)에서 C는 교정 후보 어절들(교정 대상 어절 포함) 중에서 1개의 후보를 의미한다. w는 입력 문장의 전체 어절이다. $P_f(C|w_{n-2},w_{n-1})$는 교정 대상 어절과 이전 두 어절에 대한 정방향 트라이그램 확률을 의미한다. 그리고 $P_b(C|w_{n+1},w_{n+2})$는 교정 대상 어절과 그 이후의 두 어절에 대한 역방향 트라이그램 확률을 의미한다. 따라서 교정 대상 어절의 좌우 어절들을 보고 그 확률의 평균값을 통해서 C 중 상위 어절을 선택하게 된다.

만약 정방향과 역방향 트라이그램 확률이 전부 0이라면, 같은 교정 대상 어절에 대하여 다음과 같은 바이그램 수식을 적용한다.

$$Correction_{word}(C) = \frac{(P_f(C|w_{n-1}) + P_b(C|w_{n+1}))}{2} \quad (2)$$

트라이그램을 사용할 때와 마찬가지로 이번에는 수식 (2)와 같이 바이그램에 대하여 상위 어절을 채택한다. 그러나 교정 대상 어절 및 교정 후보 어절 모두 유니그램 빈도수가 1 이하라면, 기본적인 어절 N-gram 데이터는 무용지물이다. 물론 스무딩(Smoothing) 기법을 사용할 수 있으나, 해당 경우에는 스무딩 기법도 제한되며, 좋은 결과를 기대할 수 없다. 따라서 여기서는 음절 N-gram 데이터를 사용한다. 음절 N-gram을 사용하여 C 중에서 적합한 후보를 선택하는 수식은 다음과 같다.

$$Correction_{syllable}(C) = \sum_{n=2}^{len(C)+2} \frac{(P_{sf}(c_n|c_{n-2},c_{n-1}) + P_{sb}(c_n|c_{n+1},c_{n+2}))}{2} \quad (3)$$

수식 (3)에서 c는 시작 토큰과 끝 토큰을 포함한 교정 대상 어절의 모든 음절을 의미한다. $P_{sf}(c_n|c_{n-2},c_{n-1})$는 c_n에 대한 정방향 트라이그램 확률 값이고, $P_{sb}(c_n|c_{n+1},c_{n+2})$는 c_n에 대한 역방향 트라이그램 확률 값이다. C의 모든 음절에 대하여 정방향, 역방향 트라이그램 값의 합의 평균을 모두 더하여 C의 교정 점수를 계산하였다. 그리고 계산되 교정 점수가 가장 높은 C를 선택하였다. 이에 따른 좌우 N-gram 교정 모델의 구조도는 그림 1과 같다.

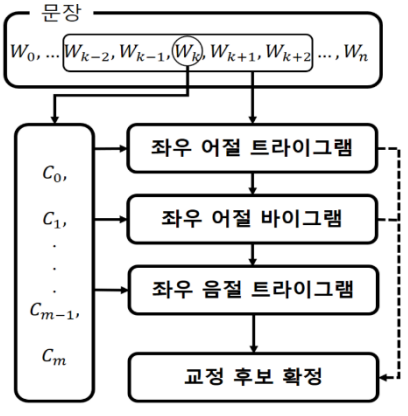

그림 1 좌우 N-gram 모델 구조도

4. 실험 및 결과

좌우 N-gram 교정 모델을 통해 3.1에서 언급한 2가지의 간섭 오타 데이터에 대하여 실험을 진행하였다. 표 1은 2가지 간섭 오타 데이터에 대한 정보를 정리한 것이다.

표 1. 오타 데이터

	문장수	어절수	오타수
KCC150	3,400	47,913	10,200
테스트 데이터	3,400	58,529	10,200

표 1의 테스트 데이터는 중앙일보의 사회면 기사에서 수집한 문장들을 모아놓은 데이터로 KCC150과 중복되지 않는 데이터이다. 우선 KCC150 간섭 오타 데이터에 대해 실험을 진행한 결과는 표 2와 같다.

표 2. KCC150 간섭 오타 데이터에 대한 실험결과

모델	정확도	재현율	정밀도	F1 점수
Word_bi	0.98908	0.97020	0.97835	0.97426
Word_bi + Syllable_tri	*0.99267*	*0.98686*	*0.97890*	*0.98286*
Word_bi+ tri	0.98915	0.97049	0.97836	0.97441
Word_bi+ tri+ Syllable_tri	**0.99274**	**0.98716**	**0.97890**	**0.98301**

그리고 동일한 모델들로 테스트 데이터의 간섭 오타 데이터에서 실험을 진행한 결과는 표 3과 같다.

표 3. 테스트 데이터 간섭 오타 데이터에 대한 실험결과

모델	정확도	재현율	정밀도	F1 점수
Word_bi	*0.98250*	*0.95157*	*0.94822*	*0.94989*
Word_bi + Syllable_tri	0.97757	0.97127	0.90665	0.93785
Word_bi+ tri	**0.98259**	0.95118	**0.94904**	**0.95011**
Word_bi+ tri+ Syllable_tri	0.97781	**0.97167**	0.90752	0.93850

실험결과에 따르면 N-gram 데이터를 구축하는 데 사용된 KCC150의 간섭 오타 데이터의 경우 좌우 어절 바이그램과 트라이그램 그리고 좌우 음절 트라이그램을 같이 사용하는 것이 대체적으로 가장 좋은 성능을 나타냈다. 그러나 좌우 어절 바이그램과 좌우 음절 트라이그램을 사용한 모델과 큰 차이를 보이지는 않았다.

그리고 테스트 데이터의 간섭 오타 데이터로 실험한 경우에는 KCC150과는 다른 양상을 보였다. 좌우 음절 트라이그램을 사용한 모델들은 재현율(Recall)의 관점에서는 전반적으로 상승했지만, 정밀도 관점에서 전반적으로 급격하게 성능이 감소했다. 그 이유는 좌우 음절 트라이그램은 오타인 미등록어(OOV)를 더 잘 수정했지만, 오타가 아닌 미등록어를 수정해버리는 거짓 양성 문제가 더 크게 발생했기 때문이다. 특히 좌우 음절 바이그램까지 합친 모델의 경우 더 심화되었다. 이것은 결국 정확도가 상대적으로 낮아지는 결과를 불러왔다. 따라서 좌우 음절 트라이그램을 사용하지 않는 모델들이 정밀도, 정확도 및 F1 점수에서 더 높은 값을 가지는 것을 확인하였다. 따라서 표 3에서 좌우 음절 트라이그램을 사용하지 않는 좌우 어절 바이그램과 트라이그램 알고리즘 모델이 상대적으로 가장 좋은 성능을 보여주었다. 그러나 좌우 어절 바이그램만 사용한 모델과 큰 격차는 보이지 않았다.

5. 결론

제안하는 교정 방법들의 공통적인 오류는 대부분 복합명사 또는 대명사인 미등록어('장승포가축병원', '사사카와평화재단' 과 같은 경우)나, 교정한 어절이 정답 어절과는 다르지만, 결과가 문법적으로는 맞는 경우('드러나지 **않는** 옷을'이 정타일 때, '드러나지 **않은** 옷을'로 수정한 경우, '경호원 **중** 최소한 3명이 숨졌다고'가 정타일 때, '경호원 **등** 최소한 3명이 숨졌다고'로 수정한 경우 등)가 대부분이였다. 그리고 정밀도(Precision)에 있어서 미등록어가 적으면 좌우 음절 트라이그램이 효과를 발휘하지만, 많으면 오히려 거짓 양성 문제에 의해서 정밀도가 감소하는 역효과가 발생하는 부분이 있었다. 즉, 미등록어가 적은 데이터에서는 좌우 음절 트라이그램을 사용한 모델이 좋은 성능을, 미등록어가 많은 데이터에서는 좌우 음절 트라이그램을 사용하지 않은 모델이 더 좋은 성능을 기록한 것이다. 이는 미등록어 교정에 대한 접근이 오타가 아닌 미등록어를 얼만큼 잘 판단하고, 확실한 경우에만 교정하는 방식에 있음을 보여주고 있다.

따라서 앞으로의 연구방향은 미등록어에 대한 고찰과 오타 여부에 대한 세밀한 판단 방식에 대한 연구가 더 필요하다. 또한 딥러닝 모델을 통해 교정 후보 어절에서 선택하도록 학습하는 것도 좋은 성능을 기대할 수 있을 것이다.

감사의 글

이 논문은 2017 년 정부 (과학기술정보통신부) 의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017M3C4A7068186)

참고문헌

[1] M. Sarwar and T. R. Soomro, “Impact of Samartphone's on Society,” European J. of Scientific Research, vol.98, no.2, pp. 216-226, Mar. 2013.

[2] 김보람, 김태일, 임영재, 정의승, “스마트폰 터치스크린에서의 작은 터치키의 사용성 연구”, Journal of the Korean Institute of Industrial Engineers, 제38권, 제2호, 80-88쪽, 2012년

[3] 고석훈, “인접-오타를 이용한 소프트 키보드의 동적 적응 연구,” 멀티미디어학회논문지, 제21권, 제11호, 1263-1270쪽, 2018년 11월

[4] N. Henze, E. Rukzio, and S. Boll, “100,000,000 Taps : Analysis and Improvement of Touch Performance in the Large,” Proc. of the 13th Int. Conf. on Human Computer Interaction with Mobile Devices and Services, Stokholm, Sweden, pp.133-142, Aug. 2011.
[5] I.S. MacKenzie and S.X. Zhang, “The Design and Evaluation of a High-Performance Soft Keyboard,” Proceeding of the SIGCHI Conference on Human Factors in Computing Systems, pp.25-31, 1999.
[6] S. Zhai, M. Hunter, and B.A. Smith, “The Metropolis Keyboard: An Exploration of Quantitative Techniques for Virtual Keyboard Design,” Proceeding of the 5th International Conference on Intelligent User Interfaces, pp.119-128, 2000.
[7] T. Stocky, A. Faaborg and H. Lieberman, “A Commonsense Approach to Predictive Text Entry,” CHI'04 Extended Abstracts on Human Factors in Computing Systems, pp.1163-1166, April. 2004.
[8] 김민호, 권혁철, 최성기, “어절 n-gram을 이용한 문맥의존 철자오류 교정”, 정보과학학회논문지, 제 14권, 제12호, 1081-1080쪽, 2014년 12월
[9] C. E. Shannon, “A Mathematical Theory of Communication,” Bell system technical journal, vol.27, no.3, pp.379-423, 1948.
[10] E. Mayes and F. Damerau, et al. “Context Based Spelling Correction,” Information Processing and Management, vol.27, no.5, pp.517-522, 1991.
[11] E. Brill and R. C. Moore, “An Improved Error Model for Noisy Channel Spelling Correction,” Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp.286-293, October. 2000.
[12] A. Kirk, “Improving the Accuracy of Mobile Touchscreen QWERTY Keyboard,” Electronic Theses and Dissertations, October. 2018.