웹사전과 맞춤법 자동교정기의 논리와 개선*

정 철**

<차 례>

Ⅰ. 머리말
Ⅱ. 맞춤법 자동교정기
Ⅲ. 맞춤법 자동교정기의 논리
Ⅳ. 맞춤법 자동교정기의 문제점과 개선 방안
Ⅴ. 맞춤법 교정의 표현방식
Ⅵ. 나가며

Ⅰ. 머리말

스마트폰의 등장으로 사전 사용의 양상은 크게 달라지고 있다. 종이사전의 쇠퇴야 CD롬 사전이나 전자사전, 웹사전이 나오면서 일찌감치 예견되었지만 2008년 이후 스마트폰이 인터넷을 핸드폰으로도 연결해주어 이제는 모든 사전의 대안이 웹사전으로 통일되어가는 양상이다. PC용 웹사전이냐 스마트폰용 웹사전이냐의 차이가 있을 뿐 웹사전이라는 사실에는 변함이 없다.

* 고려대학교의 김양진 선생께서 본 발표의 토론자를 맡아 많은 조언을 주셨다. 그 중 명칭을 어휘추천기 정도로 고치는 것이 좀 더 성격에 맞지 않은가라는 의견이 있었지만, 맞춤법 오류 교정 어휘 추천기로 고치기도 어렵고 적절한 대안이 떠오르지 않아 맞춤법 자동교정기로 그냥 두었다. 본 자동교정기는 실제로 100% 교정해주지 못하며 사용자도 100% 교정되리라 기대하지는 않을 것이라 본다.

** Daum 커뮤니케이션, zepelin@daumcorp.com

이런 상황에서 웹사전은 단순히 종이사전의 인터넷 버전이 아닌 웹이라는 매체의 특성을 활용한 사전이라는 차별성을 가져가고 있으며 그 여러 변화 중 하나로 여기서는 맞춤법 자동교정기능을 소개하려 한다.

Ⅱ. 맞춤법 자동교정기

맞춤법 자동교정기(이하 자동교정기)는 말 그대로 사용자가 맞춤법에 어긋난 검색어를 입력해도 사용자의 본래 의도를 고려하여 적절한 질의어로 바꿔주는 장치를 말한다. 자동교정기는 문장단위로 처리하거나 어휘단위로 처리해줄 수 있다. 문장단위로 처리하는 것으로는 나라인포테크(부산대학교 권혁철 교수팀)가 상용화한 "한국어 맞춤법/문법 검사기"가 유명하며 이는 우리말배움터에서 사용해볼 수 있다.[1] 문장단위의 자동교정기는 문서작성시 매우 유용하게 활용할 수 있다. 따라서 워드 소프트웨어에는 맞춤법 자동교정기가 들어가 있고, 포털사이트의 메일서비스나 블로그/카페 서비스 등에서도 적용 가능하다.

본 발표에서 소개하려는 것은 문장단위가 아닌 어휘단위로 처리하는 자동교정기이다. 웹사전에 입력되는 검색어는 문장단위가 아니고 구 혹은 단어 단위이기 때문이다. 구로 입력되는 경우도 극히 일부이며 대부분은 단어 단위로 입력된다. 따라서 웹사전의 자동교정기는 단어 단위의 처리를 잘 해주어야 한다.[2]

자동교정기는 아래의 장점을 가진다.

---

1) http://urimal.cs.pusan.ac.kr/urimal_new/

2) 김양진 선생께서 맞춤법 자동교정이라는 근본 목적에 다가가기에 어휘 단위의 처리는 미흡하며 공급자 중심적인 접근이 아닌가 하는 지적을 주셨다. 옳은 지적이다. 이런 어휘 단위의 처리는 웹사전을 만드는 과정에서 손쉽게 맞춤법 오류를 개선하자는 아이디어가 떠올랐기 때문에 시작한 것이다. 하지만 웹사전에서 처리해야 하는 대상은 대부분 어휘 단위의 질의어이기 때문에 더 이상의 작업을 하는 것이 오히려 불필요한 기능일 수도 있다.

1) 사용자의 맞춤법 오류를 고쳐줄 수 있다.
2) 사용자가 잘못 넣은 질의어를 다시한번 검색해야 하는 불편함을 해소할 수 있다.
3) 사용자가 순화대상어를 검색하면 순화어를 제시할 수 있다.

사전은 태생부터가 규범일 수밖에 없다.[3] 옳은 표현인지 확신이 서지 않을 때 찾아보는 것이 사전이기 때문이다. 따라서 사전의 표제어는 맞춤법에서 정답으로 간주할 수 있다. 정답이 명확하게 들어있다면 오답을 정답으로 연결해주는 것은 정답이 애매하거나 없는 경우에 비해 훨씬 쉬운 일이 된다. 웹사전용 자동교정기를 만든다는 생각은 여기서 출발하였다.[4]

Ⅲ. 맞춤법 자동교정기의 논리

현재 다음 국어사전에 적용된 한국어 자동 교정기는 아래와 같은 논리(logic)로 구성되어 있다.

1) 오용어 목록 비교
2) 음절 수 비교
3) 음절간 음운현상을 반영한 고빈도 패턴 적용[5]

---

3) 물론 규범사전이 아닌 기술사전에 관한 논의들이 계속 나오고 있지만 그것은 사전을 만드는 입장에서 그러한 것이지 사용자의 입장에서는 규범사전이나 기술사전 모두 규범으로 다가갈 수밖에 없다.

4) 본 발표에서 언급하는 어휘단위의 자동교정기는 모두 다음 국어사전에 적용된 것을 말하며 선행연구가 있었을지도 모르나 치밀한 확인없이 구현 자체를 목표로 삼아 진행한 것이다. 따라서 여기서 가정한 전제들에 근거가 확실하게 있다고 할 수는 없다.

5) 이 과정에서 키보드 인접도 오류 패턴을 적용하기도 한다. 예를 들어 추정, 투정, 우정, 주정이라는 단어가 있다고 할 때 *{푸정}이라는 오류가 들어오면 키보드 상에서 가장 가까운 추정을 먼저, 그 다음에 투정이나 우정을, 마지막에 가장 먼 주정이나 부정 등을 제시해주는 것이다. 하지만 사전에서 단어단위로 입력되는 경우에서는 키보드 인접도 오류가 많이 보이진 않는다.

4) 음절별 자소를 초중종성으로 나누어 비교

1) 먼저 오용어 목록 비교 단계는 관리자가 무조건 정답으로 규정한 내용과 비교하는 단계이다. 규칙으로 만들기에는 경우의 수가 적지만 반드시 적용되어야 하는 경우를 모아 처리한다. 규칙이 99%를 처리한다면 나머지 1%를 처리하는 것이 오용어 목록 단계이지만 전체적인 완성도를 높이는데 반드시 필요하다.

이 오용어 목록 비교 단계에서 순화어를 반영한다. 순화대상어와 순화어는 형태상의 유사성이 있을 수도 있지만 오뎅과 어묵처럼 형태상의 관련이 전혀 없는 것이 대부분이다. 따라서 이런 경우는 규칙을 통한 자동처리가 아닌 수동 처리를 할 수 밖에 없다.

2) 다음으로 음절 수 비교 단계는 비교 대상 어휘를 줄이는 역할을 한다. 사용자가 어휘를 잘못 입력했을 때 대체로 맞춤법을 틀리지 음절 수까지 틀리지는 않는다고 전제하는 것이다. 이는 특별히 근거를 가지고 적용한 전제는 아니며 자동 교정기에 처리 부담을 덜 주기 위한 방편이다. 표준국어대사전에 실린 표제어가 대략 50여만 개라 하므로 그 모든 것을 대상으로 할 수는 없기 때문이다. 하지만 이 처리는 여러 가지 약점을 가진다. '메뚜'를 입력하면 '메뚜기'쪽을 추천해주는 것이 더 좋을 수 있는데, '메주'가 추천될 수 있기 때문이다. '숏커트'를 입력해도 '쇼트커트'가 추천되지 못한다. 그럼에도 불구하고 대규모 입력을 처리해줘야 하는 포탈 웹사전의 특성상 이러한 논리가 포함되었으며 이는 향후 개선의 여지가 많이 있다.

3) 음절간 음운현상과 고빈도 패턴에는 사용자가 소리와 철자를 혼동하는 형태와 그것에 포함되지는 않지만 자주 틀리는 형태가 포함된다.

〈표 1〉 적용 가능한 음절간 음운현상 오류와 그 교정의 예

	철자를 소리처럼 오인한 과잉교정		소리 그대로 철자를 적은 오류	
	오류형	표준형	오류형	표준형
비음화	격련 감남을녀	경련 갑남을녀	궁물 종노	국물 종로
일곱 끝소리	궂이	굳이	젓소 오지랍	젖소 오지랖
경음화	뒷굼치 콧배기	뒤꿈치 코빼기	깍따	깎다
격음화			조타 차카다	좋다 착하다
구개음화			구지	굳이
사이시옷	나무가지 꼭지점	나뭇가지 꼭짓점	댓가 촛점	대가 초점

이러한 음운현상은 아래처럼 패턴 규칙의 형태로 바꾸어 자동교정기에 적용이 가능하다.

〈표 2〉 외래어 표기 오류 관련 패턴 정리의 예]

패턴 01 : 받침 ㅍ→받침 ㅂ〔워크숖vs 워크숍〕
패턴 02 : ㄱㄲㅋ↔받침 ㄱㄱㅋ〔워크샵 vs 웍샵〕
패턴 02-1 : 그브↔받침 ㄱㅂ〔로브스터 vs 롭스터〕
패턴 03 : 받침 ㅅ + 트→트〔셋트 vs 세트〕
패턴 03-1 : 받침 ㅅ + 트→받침 ㅅ〔카펫트vs 카펫〕
패턴 03-2 : 받침 ㅅ↔트〔카펫 vs 카페트〕
패턴 03-3 : (받침 ㅅ +) ㅊ→받침 ㅅ〔도너츠vs 도넛〕
패턴 04 : 쟈재져졔쥬→자재저제주〔비젼 vs 비전〕
패턴 04-1 : 쉬→시〔인턴쉽→인턴십〕
패턴 04-2 : 휘→피, 훼→페〔휘트니스 vs 피트니스, 휀스 vs 펜스〕
패턴 05 : 챠쳐쵸→차처초추〔시츄에이션 vs 시추에이션〕
패턴 05-1 : 쟈져죠쥬→자저조주〔쟈스민 vs 자스민〕
패턴 06 : 받침 ㄴ + ㄴ→ㄴ〔런닝 vs 러닝〕
패턴 06-1 : 받침 ㄴ↔받침 ㅁ〔컨펌 vs 컴펌〕
…

하지만 이런 패턴은 서로 충돌하기도 하고 다른 부작용도 일으킬 수 있으므로 여러 테스트를 해봐야 한다. 또 패턴과 음절별 자소비교 간에 충돌도 있을 수 있다. 위에는 그런 테스트를 거치지 않은 몇 가지 예를 들었을 뿐이며 충돌을 많이 일으키는 패턴은 실제로 그런 패턴이 있다 하더라도 자동교정기에서 빼는 것이 옳을 수도 있다. 현재 다음 국어사전에 *{맞닥들이다}와 *{맞딱들이다}를 입력하면 맞닥뜨리다가 아닌 맞아들이다를 추천해준다. 이런 경우는 패턴이 점수보다 우선적으로 적용되지 못한 부작용이라 할 수 있다.

4) 마지막으로 음절별 자소 비교는 음절 내의 첫 글자, 가운데 글자, 끝 글자를 나누어 점수를 비교하는 것이다. 이를 통해 예를 들어 '마딱뜨리다'를 넣었을 때의 ㄷ, ㄸ, ㅌ과 같이 인접한 동일 계열간의 자소 교체는 1점 감점을, 동일 계열이 아닌 자소 교체는 2점 감점을 준다면 위치별 감점은 아래처럼 줄 수 있다. 따라서 감점이 가장 적은 맞닥뜨리다가 최우선 추천 후보가 될 수 있다.

〈표 3〉 음절내 자소 교체로 인한 오류도 계산의 예

	마	딱	뜨	리	다	감점
맞닥뜨리다	2	1	0	0	0	3
마사뜨리다	0	4	0	0	0	4
다닥뜨리다	4	1	0	0	0	5
다닥트리다	4	1	1	0	0	6

Ⅳ. 맞춤법 자동교정기의 문제점과 개선 방안

다음 국어사전에 적용된 자동교정기는 초보적인 형태로, 성능 개선의 여지가 많이 있다. 아직 적용하지 못했으나 운영하면서 얻은 아이디어들

을 여기에 소개한다.

먼저 비교대상어를 확장할 필요가 있다. 사람들이 용언의 활용형에서 많이 틀리는데 사전 표제어에는 기본형만 있으므로 자동 교정기가 제대로 정답 추천을 해줄 수 없다. 따라서 비교대상어에 고빈도의 용언 활용태를 넣어야 한다. 특히 불구동사처럼 실제 말뭉치에서 기본형이 거의 사용되지 않는 경우 뿐 아니라 전체 활용형 중에서 두세 가지 형태가 해당 어휘 빈도의 대부분을 차지하는 경우는 비교대상인 정답군에 반드시 포함해야 한다. 이는 세종말뭉치를 가공해서 충분히 얻어낼 수 있는 자료이다.[6)]

불필요한 비교대상어를 목록에서 빼는 것도 전체적인 성능을 개선하는데 도움을 줄 수 있다. 구지는 동음이의어가 표준에 11개나 실려있지만 현대어에서 거의 안쓰이는 것들 뿐이다. 실제로 말뭉치에서 *{구지}를 찾아보면 굳이를 잘못 적은 것이 대부분을 차지한다. 이런 어휘들은 사전에 실려있는 것이야 옳겠으나 자동 교정기 내의 비교대상어 목록에서는 빼고 오용어 목록 비교단계에서 처리해주는 것이 합리적이다.

음절 수의 일치에서 벗어날 필요도 있다. 특히 외래어의 경우 장음을 살리느냐 아니냐, 마지막 자음을 어떻게 발음하느냐에 따라 음절 수가 달라질 수 있기 때문이다. '플룻'을 넣으면 현재는 음절 수에 걸려서 '플롯'이 추천되고 있는데 사실 이것은 '플루트'를 찾은 것이다. 이를 위해서는 외국어에서 일어나는 음운현상도 일부 자동 교정기에 반영해야 한다.[7)]

---

6) 세종말뭉치에서 묻다 활용형을 찾아보면 아래와 같다.

_(때가) 묻다 : 묻은([세]253), 묻어([세]154), 묻지([세]21)
_(때를) 묻히다 : 묻혀([세]48), 묻힌([세]32), 묻히고([세]18), 묻히지([세]11)
_(쓰레기를) 묻다 : 묻고([세]98), 묻어([세]85), 묻었다([세]59), 묻은([세]58), 묻는([세]25), 묻을([세]20), 묻으며([세]15)
_(쓰레기가) 묻히다 : 묻혀([세]262), 묻힌([세]89), 묻힐([세]28), 묻혔다([세]23), 묻히고([세]22), 묻혀버렸다([세]18), 묻히는([세]15)

가지수로 보면 형태는 16가지이지만 빈도수로 따지면 묻은, 묻혀, 묻어, 묻힌, 묻고의 5가지가 전체의 80%를 차지한다. 즉 이 5개의 활용형을 정답 세트에 넣어두면 묻다 활용형 관련 맞춤법 오류는 대부분 잡을 수 있는 것이다. 이 5개에서 벗어나는 오류형은 *{묻었다}, *{묻어있다}, *{묻혀서} 등이 발견되었다.

또 패턴을 확장해서 적용할 필요가 있다. 지금은 한국어에서 나타나는 비음화, 경음화, 격음화 등의 음운규칙이 반영되어 있지만 외래어에서 나타나는 패턴들을 좀 더 보완할 수 있다. 물론 이를 획일적으로 적용할 수 없으며 테스트해보고 부작용이 있는지 확인해봐야 한다. 또 패턴으로 적용하는 것이 나은지 그냥 오용어 목록에 넣어 예외처리하는 것이 나은지도 확인할 필요가 있다.

기계학습(machine learning) 결과물을 적용하는 것도 필요하다. 자동교정기는 점수가 비슷하면 맞춤법 교정 결과를 여러 개 내놓는다.

〈그림 1〉 '수방씨'의 검색실패 화면

이처럼 자동교정기에 *{수방씨}를 넣으면 가장 정답에 가까운 수박씨를 먼저 제시해주고 이후 수방기, 수방사, 무방비 등 다음 순위의 답들을 제시한다. 이 경우 수박씨가 다행히 가장 좋은 결과였지만 사용자는 수박씨보다 무방비를 더 원했을 경우도 분명히 있다. 자동교정기는 자동으로 추천해주기 때문에 항상 오류의 요인들을 가지고 있기 때문이다. 그때 사용자는 무방비를 선택할 것이고 이는 사전 서버에 기록된다. 이 기록이 쌓여서 많은 이들이 무방비를 더 원했다고 판단되면 그 때는 정답으로 무방비를 제시하고 수박씨는 뒤에 보여주는 것이 옳다.

마지막으로 오용어 목록을 지속적으로 개선해야 한다. 규칙은 비교적

7) 이러한 음절수 일치조건이 들어간 이유는 전체적인 성능저하를 막고, 논리를 단순하게 하기 위한 것이다. 추후 음절 수 일치조건을 빼면 분명 성능은 더 나아지게 할 수 있으리라 생각되지만 중간에 또 다른 부작용이 끼어들 수도 있어 상당한 테스트를 필요로 할 것이다.

많은 경우의 수를 해결할 때 유용하지만 모든 것을 규칙으로 처리하면 전체적인 처리 속도는 저하된다. 규칙으로 어느 정도의 보완이 끝났을 때에 완성도를 높이는 방안은 바로 오용어 목록을 계속 보완하는 것이며 이런 오용어 보완에 가장 좋은 방법은 사용자가 사전을 검색했던 검색어 목록을 꾸준히 검토하는 것이다.

Ⅴ. 맞춤법 교정의 표현방식

인터넷 활용량이 늘어날수록 중요해진 것은 사용자 인터페이스이다. 얼마나 사용자가 쉽게 쓸 수 있도록 서비스를 구성하는가가 서비스의 기능이 어떤가보다 더 중요해진 시대에 우린 살고 있다. 따라서 맞춤법 교정기능도 중요하지만 그 교정된 결과를 어떻게 보여지는가도 중요하다.

앞서 수박씨의 사례에서 보여준 것처럼 자동교정기의 결과물은 일단 정답에 가장 가까운 사전 항목을 먼저 제시해주고 다른 단어를 원했을지 모르는 경우를 대비해 후순위의 후보를 제시해주는 두 가지의 단계로 이루어져있다. 이 외에 해결할 수 없는 경우에 대해 보완이 필요하다.

자동교정기로 해결할 수 없는 경우 오용어 목록을 보완하여 정답을 제시해주는데 그 중 사용자를 교육하기 위해 더 자상하게 표현해줄 필요도 있다. 대표적인 경우가 고빈도의 맞춤법 오류와 순화대상어 제시이다.

고빈도의 맞춤법 오류의 예로는 *{구지}나 *{주꾸미} 같은 것이 있다. 언젠가 이 고빈도 오류들이 표준어의 지위를 차지하게 될 수도 있지만 지금은 아니므로 이런 경우는 명시적으로 고쳐줄 필요가 있다. 필요하면 추가적인 문법설명 등을 달 수도 있을 것이다.

〈그림 2〉 '구지'를 입력했을 때 보조설명이 나오는 희망화면 예시

일본식 표현이나 만연한 외래어를 우리말에 가까운 형태로 고쳐 표기하려는 언어순화는 꾸준히 있어왔으며 그 결과물은 국립국어원의 순화어 자료집이나 우리말 다듬기 홈페이지에서 찾아볼 수 있다.[8] 이러한 순화 결과물은 최대한 널리 알려 규범으로 작용하게 해야 하며 그에 적합한 것

8) http://www.malteo.net/, 순화어 사용이 얼마나 바람직한가에 대해서는 여러 가지 이견이 있을 수 있다. 홈페이지도 누리집이라는 순화어로 대체할 수 있겠지만 나는 별로 그러고 싶지는 않은 것이 솔직한 마음이다. 홈페이지를 누리집이 대체하는 것이 좋은가, 노가다를 노동자나 노역이라는 말이 대체할 수 있는가에 대해 회의적이기 때문이다.

이 바로 웹사전 결과에서 그때그때 순화어를 제시하는 방식이다. 역시 여력이 닿는다면 왜 그 순화어를 쓰는 것이 좋은가를 더 제시하여 순화어 제시의 효과를 높일 수도 있다.[9)]

Ⅵ. 나가며

앞서 언급한 것처럼 사전은 태생적으로 규범적인 도구이며 인터넷 시대를 맞이하여 좀 더 효과적인 교육도구로 변해가고 있다. 또 한편으로는 종이사전 판매고가 점차 감소하여 사전의 재생산 기반이 약해지고 있는 모순적인 상황에 놓여있기도 하다.

이 시점에서 사전은 두 가지 전략을 모두 취할 수밖에 없다. 하나는 사전의 경쟁자인 인터넷 문서(카페, 블로그, 홈페이지 등)들과의 차별점을 강조하기 위해 종합 언어 학습도구로서의 컨텐츠 확대와 신뢰도 강화이고 또 다른 하나는 사전이 재생산될 수 있도록 사회적으로 이슈를 제기하여 자본을 확보하는 것이다.

종합 언어학습 도구로서의 사전의 성격은 앞으로 충분히 발전될 여지가 있다. 역순사전이나 갈래사전 등 종이사전에서도 구현된 사전적인 특성이 아직 웹사전에 반영되지 못한 것도 있을 뿐 아니라 한국어 의미망이 좀 더 체계적으로 구축되고 방대한 유의어 반의어군을 정리하면 기존의 사전과는 혁신적으로 다른 웹사전이 나올 수 있을 것이다. 맞춤법 자동교정기는 그 작은 일부에 불과하다.*

---

9) 사실 순화어를 제시하는 것은 맞춤법 교정이라는 본래 취지에서는 벗어나는 일이다. 이를 김양진 선생도 지적하셨으나 여기서 굳이 순화어 제시까지 언급한 것은 이 기능이 웹사전의 교육적 효과를 위한 장치이기 때문이다. 나가는 부분에서 재차 언급했지만 웹사전은 언어학습 도구로서 발전될 여지가 많이 있다.

* 본 논문은 2010. 10. 31. 투고되었으며, 2010. 11. 5. 심사가 시작되어 2010. 11. 29. 심사가 종료되었음.

<초록>

웹사전과 맞춤법 자동교정기의 논리와 개선

정 철

매체의 변화에 따라 웹사전의 사용이 증가하여 이제는 웹사전이 커다란 비중을 차지하게 되었다. 웹사전은 웹이라는 매체의 특성을 활용한 사전이라는 점에서 종이 사전과 차별화되는데, 이러한 차별점으로 들 수 있는 웹사전의 한 가지 특징이 맞춤법 자동교정기능이다.

맞춤법 자동교정기는 사용자가 맞춤법에 어긋난 검색어를 입력한 경우 사용자의 의도를 고려하여 적절하게 바꾸어 주는 장치로, 어휘 단위와 문장 단위 모두에서 가능하다. 본고에서는 웹사전에서의 맞춤법 자동교정기를 다루므로 어휘단위에 주목하였고, 다음(Daum) 국어사전에 적용된 논리를 중심으로 논의를 진행하였다.

다음 국어사전에 적용된 한국어 자동 교정기는 '오용어 목록 비교', '음절 수 비교', '음절간 음운현상을 반영한 고빈도 패턴 적용', '음절별 자소를 초중종성으로 나누어 비교'의 논리로 구성되어 있는데, 개별 논리에 한계가 있을 뿐 아니라 논리가 서로 상충되기도 하는 문제가 존재한다.

아직 적용되지 않았으나 다음 국어사전을 운영하면서 아이디어를 얻은 개선점으로, 용언 활용태와 같은 비교대상어 확장, 불필요한 비교대상어 제거, 음절수 일치의 탈피, 외래어에서의 음운변동 등 패턴의 확장, 기계학습(machine learning) 결과물 적용, 오용어 목록의 지속적인 개선 등이 있다.

맞춤법 교정 기능 자체를 개선하는 것에 더하여 사용자가 쉽게 사용할 수 있도록 맞춤법 교정 결과를 보여주는 것도 중요하다. 현재는 정답에 가장 가까운 사전 항목을 먼저 제시하고 이어서 후순위 후보를 제시하

는 두 단계이나, 고빈도 맞춤법 오류와 순화대상어 제시 등은 추가적인 문법 설명 및 이유를 함께 제시하는 것과 같이 사용자 교육의 측면에서 더 상세하게 표현해 줄 필요가 있다.

【핵심어】 웹사전, 맞춤법 자동교정, 어휘단위, 사용자 인터페이스, 고빈도 맞춤법 오류 어휘

<Abstract>

The Logic and Web Dictionary and It's Refinement of Automatic Correction of Spelling Orthography

Jeong, Cheol

The purpose of this thesis is to explain the logic and web dictionary, and how to refine the automatic correction of spelling orthography. In these days, so many people depend on the web dictionary much more and the web dictionary get a great roles in the life along with the change of media. The most big difference between paper dictionary and web dictionary is the function of automatic correction in spelling orthography. The automatic correction is the devices to correct the words input by the users considering user's intention in the level of words or sentences. This thesis mainly focused on the lexical level applied in Daum web Korean dictionary. Daum web Korean Dictionary consist of lists based on the comparison of misused words, syllables, the patterns of high frequence words and phoneme analysed by initial, medial and final positions. Actually there are some problems because it has some limitations and contradiction in the respective logics. I'd like to suggest some ideas to correct and improve this problems and limitations through enlarging patterns of phonemic fluctuation, compared words with verbal inflection, eliminating unavailable words, applying the result of machine learning and improvement of misused words, etc. It is also important to show the result corrected spelling orthography so that the user can use the function more easily. At present, the service has two steps which present the word list searched by the user first and then suggest the candidate words next. In the future we need to develop the service system to

present the high frequence words list of mistakes or error with additional explanation to educate the users effectively.

【Key words】 web dictionary, spelling orthography, automatic correction, phonemic syllables, high frequence words list of mistakes or error