

한국어 맞춤법 문법 검사기

한국어 맞춤법/문법 검사기 개발 과정[1/4]

[시제품]

- | 1990년 1월 개발 시작, 1990년 6월 1차 발표
 - 사전: 동아새국어사전을 기반으로
 - 시스템: IBM PC AT(16비트 기계)

[1판]

- | 1992년 상업화 시작
- | 1994년 (주)한글과 컴퓨터에 기술 전수

[2판]

- | 1995년 다수 어절 처리 기능 도입
- | 1996년 문체 오류 수정 규칙 도입
- | 1998년 문장 단위의 띄어쓰기 오류 수정, 구두점 오류 지원
- | 2000년 표준국어대사전[국립국어원]을 기본 사전으로 변경
- | 2001년 법원용 시스템 개발(전문 분야 시스템)

한국어 맞춤법/문법 검사기 개발 과정[2/4]

[3판]

- 2003년 전면 수정과 **재프로그램**(프로그램 단순화, 안정성 증가)
- 2004년~2005년 각종 순화용어 반영(행정순화용어 포함), 동남아어, 네덜란드 어 표기법 수용
- 2007년 화합물 명명법을 대한화학회(IUPAC에 근거)의 표준으로 바꾸고, 일부(아밀라아제/아밀레이스 따위) 널리 쓰이는 것은 기존 표기와 대한화학회 표기 모두 허용하게 함

[4판]

- 2009년부터 특수 문자 포함 어절의 처리
- 영어 맞춤법 기능 강화
- Thread Safe하게 시스템 수정
- 복합명사 띄어쓰기 일관성 유지** 기능 구현
 - 띄어쓰기와 붙여쓰기를 모두 허용하는 복합명사는 띄어쓰기 형태를 혼용해서 쓰지 않도록 함
- 2012년 교육과정평가원, 국립국어원과 함께 **교과서 검수용**으로 시스템 공동 개발
 - 교과서에 포함된 용어 수록(국사, 음악, 공업 등 10개 교과)

[5판]

- **새주소 체계**를 반영함
- 동사와 형용사에 따른 **어미 활용 처리 강화**
 - 예) 돈을 집냐? (x) 돈을 집느냐? (O)
- **인명 처리 강화**
 - 문서 내에서 인명으로 판단되면 미등록 단어라도 오류로 처리하지 않음
예) 현수 : 사전에 등록되지 않은 단어라서 분석할 수 없는 오류로 처리
예) 현수 오빠: 뒤에 호칭이 오므로 인명으로 판단하여 문서 내에서 오류로 처리하지 않음
- 2016년 **맞춤법 검사 기능을 강/약으로 구분**하여 사용자가 선택하게 함.
 - 외래어 표기, 직접 화법, 순화용어 등
예) 제도를 가지고 있는 (강한 시스템) 제도를 두고 있는/제도를 채택하고 있는 (약한 시스템) 교정하지 않음
- 2016년 외국인 한국어 학습자의 오류 처리를 반영하여 구어체 검사 성능을 향상시킴

[5판 계속]

- 2017년 국립국어원이 수정한 ‘외래어 지명 뒤의 접미사 표기’를 반영함
 - 외래어 뒤에서 띄어 썼던 가(街), 강(江), 산(山), 산맥(山脈), 섬, 성(城), 성(省), 시(市), 어(語), 왕(王), 인(人), 족(族), 주(州), 주(洲), 항(港), 해(海), 현(縣), 호(湖) 등을 붙여 쓰도록 함.
예) 카리브해/건지섬/미시시피강/고사인탄산/가르다호/고츠산맥
- 2017년 법원도서관과 함께 『법원맞춤법 자료집』과 『법률 제명 약칭 목록』 등에 근거하여 법률 문장 맞춤법 검사기를 개선함.
- 2017년부터 (주)한글과 컴퓨터에 맞춤법 검사기를 재공급함.
 - 한컴오피스용과 패키지 제품 몇 가지에 한해 적용하는 조건으로 한국어 맞춤법 문법 검사기를 재공급함.

독보적인 한국어 맞춤법/문법 검사기

I 한국어 맞춤법 문법 검사기가 이 분야에서 독보적인 이유

- 한국어 NLP분야에서도 맞춤법/문법 교정은 한국어의 교착어 특성과 검증 자료 구축의 어려움 등으로 처리가 어려운 분야임
 - 본 맞춤법 검사기는 형태소 분석, 태깅, 구문 분석, 대용량 자료 처리 등에서 국내 최고의 한국어 정보처리의 핵심 기술을 보유하고 있는 부산대학교 인공지능연구실과의 기술 이전과 협력을 통해 검사기 성능을 유지하고 있음.
- 맞춤법 검사기에 들어가는 규칙과 단어의 사전 정보 등은 자연언어처리와 전산언어학에 대한 오랜 경험과 언어처리 지식이 필요함. 즉, 개발자의 오랜 연구 경험, 언어처리 지식, 문제 해결 감각이 고도로 요구되는 시스템임.
 - 본 맞춤법 검사기는 개발자 본인이 부분문장 분석 규칙, 충돌 방지, 규칙 검증을 매일 수행하고 있음.
 - 문제 해결 시 일관성 있는 처리로 시스템의 성능을 일정하게 유지할 수 있음.
- 새로운 용어, 신조어, 고유명사, 외래어 표기법 발표 등이 계속 발생하므로 꾸준히 개선시키는 노력이 필요함
- 짧은 기간에 가시적 성과를 요구하는 업체에서는 개발이 어려움

한국어 맞춤법/문법 검사기 장점

상세한 도움말 제공

- 맞춤법 오류에 따라 다양한 오류 정보를 제공하여 사용자의 신뢰를 높이고, 맞춤법을 스스로 배울 수 있게 함.

맞춤법 교정 강도 조절 기능 제공

- 사용자가 목적이나 자신의 맞춤법 실력에 따라 시스템의 성능을 선택할 수 있음.

사투리, 비속어, 인터넷 신조어 등을 표준어 또는 순화된 표현으로 바꿔줌

- 예)

국립국어원의 외래어 표기법과 맞춤법 수정안을 빠르게 반영하고 있음

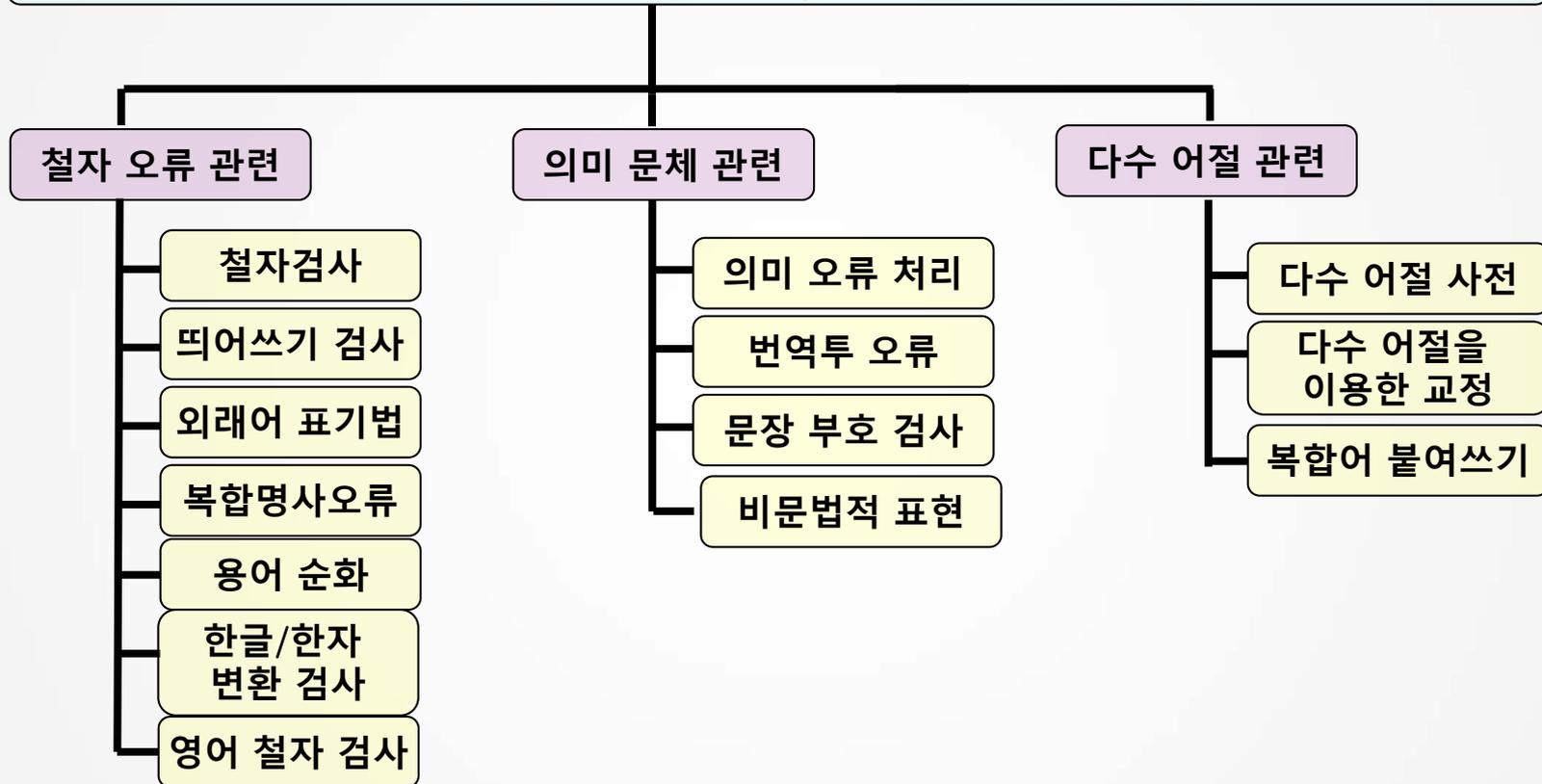
- 예)

낮은 시스템 요구 사항

- 현재 우리말배움터에서 제공하는 온라인 한국어 맞춤법/문법 검사기도 개인용 데스크탑에서도 운용될 정도로 메모리 요구량 등이 적음. (최대 150MB 정도)
- 대용량의 사전을 고밀도로 압축하는 기술, 방대한 규칙을 효과적으로 적용하는 알고리즘 등 고도의 한국어 처리 기술이 응축되어 있으므로 효율적인 컴퓨팅 자원 활용이 가능함.

한국어 맞춤법/문법 검사기 - 기능

한국어 맞춤법/문법 검사기



특징1. 상세한 도움말

- I 52,000여 개의 도움말을 오류 유형에 따라 구별하여 제공
 - ▶ 오류 유형에 따라 맞춤법 조항과 연계한 도움말을 제공하고 있음.
 - ▶ 도움말을 통하여 사용자가 맞춤법 관련 지식을 습득할 수 있도록 함.
 - ▶ 이를 통해 맞춤법에 대한 이해도를 높일 수 있음.

I 도움말 예

- ▶ 어미 관련 도움말

예1) 한글맞춤법 제30항에 따르면 순우리말로 된 합성어로서 앞 말이 모음으로 끝나고 뒷말의 첫소리 모음 앞에서 'ㄴ' 소리가 덧나면 사이시옷을 넣어야 합니다.

(예) 두렛일, 뒷일, 베갯잇, 나뭇잎

예2) '무엇을 하라고 하기에'라는 뜻의 어미로는 '으라기에'를 사용해야 합니다.

(예) 먹으래기에 (x) → 먹으라기에 (o) / 먹으래길래 (x) → 먹으라기에 (o)

- ▶ 의미 문체 관련 도움말

예1) '부치다'는 편지나 물건 등을 보내거나 회부함을 뜻하고, 붙이다'는 서로 맞닿아서 떨어지지 않게 하거나 달게함을 의미합니다. '있는 위에 더 붙게 하다'를 뜻하는 동사는 접두사 '덧'과 '붙이다'가 결합한 '덧붙이다'가 옳습니다.

예2) '흔하다'가 바르지 않은지요? '부족한 점을 들추어내거나 말하다.'의 뜻이면 바릅니다.

특징2. 다양한 사전

I 실제 사용하는 어휘를 중심으로 등록하여 높은 정확도를 얻을 수 있음

단어 사전	2,426,000여 개	체언이나 용언이 되는 단어의 사전 (법률, 언론, 지명, 인명 등의 특성화 사전도 포함되어 있음.)
다수 어절 사전	1,278,400여 개	외래어 인명 및 다수 어절로 이루어진 관용적 표현 등 예) 클리프 리처드, 향강에 떠도는 소문 등
복합 명사 사전	18,900개	의미 상으로 복합 명사 구성을 허용하지 않는 규칙
어미 사전 <small>[선어말어미제외]</small>	48,000여 개	~다, ~고 등과 같은 어미의 사전
조사 사전	10,100여 개	~은, ~이 등과 같은 조사의 사전

- 이 외 응용 분야에 따라, 사용자 사전도 추가할 수 있음.
- 맞춤법/문법 검사를 위해 정제한 단어들을 사용하며, 최근 등장하거나 빈번히 쓰이는 시사용어와 신조어들도 꾸준히 추가하고 있음
- 뜻이나 발음이 유사한 단어 때문에 자주 틀리는 표현은 규칙으로 처리하여 오류를 분석하므로 더욱 성능이 우수함.
- 일반인이 자주 틀리는 맞춤법, 외래어 표기, 입력 오류, 띄어쓰기/붙여쓰기가 있는 단어나 어절들은 별도의 사전들과 규칙으로 처리하고 있음

특징3. 다수어절 처리

- | 주로 외국 사람 이름, 관용적 표기가 다수 어절로 이루어진 경우, 이를 허용하거나, 교정할 수 있도록 하는 방법임.
- | 본 시스템에서 사용하는 방법 - 다수 어절 사전을 별도로 구축하여 교정함.
 - o 처리하기 어려운 많은 단어들을 정확하게 처리하여 높은 정확도를 제공
 - o 다수 어절을 처리하는 루틴을 따로 구현하여 효과적으로 교정하고 있음

예제1) 클리프 리처드:

다수 어절 사전에 등록해서 허용하게 함. (인명)

예제2) 항강에 떠도는 소문

다수 어절 적용 전: 항강 → 한강/ 항간 으로 교정

다수 어절 적용 후: 항간에 떠도는 소문(관용적 표현)

예제3) 할리퀸 코스튬

다수 어절 적용 전: 코스튬 → 복장/의상/분장으로 순화함.

다수 어절 적용 후: 할리퀸_코스튬 (허용)

특징4. 규칙으로 의미적 오류를 처리함.

- 비슷한 발음이나 의미 때문에 서로 혼동하여 잘못 사용하는 오류가 많음.
예) 받치다 vs. 바치다: 세금을 받치다. -> 세금을 바치다.
예) 부치다 vs. 붙이다: 포스터를 담벼락에 부치고 -> 포스트를 담벼락에 붙이고
- 현재 약 32,000여 개의 규칙으로 의미적 오류 처리하고 있음.
- 아래와 같은 정규화된 표현을 규칙으로 변환 입력하여, 의미적인 오류를 교정함.

[흠한 ==> 흔한] {동물, 식물, 140-156, 160-165, 173-178, 낙엽, 암, 환자, 돈, 동성애, 동성애자, 경우, 예, 사례, 사건}{+조사}(부사) [흠하(다) ==> 흔하(다)]
도움말) '흔하다'가 바르지 않은지요? '부족한 점을 들추어내거나 말하다.'의 뜻이면 바릅니다.

예) 흠한 사례는 아니지만, → 흔한 사례는 아니지만,
뺑소니 사건이 흠한 일은 아니다. → 뺑소니 사건이 흔한 일은 아니다.

- 정규화된 표현과 의미 정보를 이용하면, 구분하기 어려운 것도 제어할 수 있음.
예1) 이 **검사기 밖에** 못 하는 일이다. → 이 **검사기밖에** 못 하는 일이다.: 조사이므로 붙여 써야 함.
예2) 이 일은 내 **능력밖에** 있어서 → 이 일은 내 **능력 밖에** 있어서 : 밖(명사)+에(조사) 이므로 띄어 써야 함.

특징5. 인명처리와 복합 명사 일관성

I 인명 처리 강화

- 문서 내에 있는 단어가 인명으로 판단되면 오류로 처리하지 않음.
- 분석되지 않는 명사 뒤에 인명 뒤에 붙는 직위, 호칭, 조사가 올 경우, 인명으로 판단하여 처리함.
- 대본처럼 특정 패턴이 있는 경우 등장 인물을 인명으로 처리할 수도 있음.

예) 현수 : 사전에 등록되지 않은 단어라서 분석할 수 없는 오류로 처리

예) 현수 오빠: 뒤에 호칭이 오므로 인명으로 판단하여 문서 내에서 오류로 처리하지 않음

- 인명인지 아닌지의 판단에 따라 띄어쓰기 등도 달라짐.

예) 이기준으로 볼 때 → 이 기준으로 볼 때 (교정)

예) 이기준 교수는 : 바름

I 복합명사 띄어쓰기 일관성 유지 기능 구현

- 복합명사는 띄어 써도 되고 붙여 써도 되지만, 한 문서 내에서 혼용해서 쓸 경우 일관성에 문제가 생김
- 한 문서 내에서 같은 복합명사가 붙여 쓴 형태와 띄어 쓴 형태가 혼용되면, 먼저 나타난 형태로 교정하여 일관성을 유지시켜 줌

특징6. 복합명사 의미 제약

- 명사 좌우를 의미 단위로 분석하여, 의미적으로 제약을 둠. 현재 18,900여 개의 규칙이 있음.
- 잘못된 복합명사로의 분석을 줄여 검사기의 정확도를 높일 수 있고, 잘못된 교정 결과 생성을 미리 방지할 수 있음. 의미 단위로 띄어 쓰므로 뜻을 명확히 할 수 있음.

예) 아파트당첨 → 아파트_당첨

복합명사 의미 제약 전: 아파트(명사) + 당첨(명사)로 분석하여 허용

복합명사 의미 제약 후: '아파트_당첨'으로 띄어쓰게 함.

예) 미국회의 → 미국_회의

복합명사 의미 제약 전: 미국(명사) + 회의(명사)로 분석하여 허용

복합명사 의미 제약 후: '미국_회의'로 띄어 쓰게 함.

특징7. 맞춤법 검사의 강도 조절이 가능함

한국어 맞춤법 문법 검사기는 바른 언어 사용을 위해서 순화대상 용어와 영어/일어 번역 투 문장 교정을 원칙적으로 함.

그러나, 순화 대상 용어와 번역투 문장 을 어느 정도는 허용하자는 의견이 많아서 강도를 선택해서 검사를 적용할 수 있게 하였음

예제	약한 검사	강한 검사
인테리어	교정하지 않음	실내장식
스캔들	교정하지 않음	물의/추문/좋지 않은 소문
제도를 가지고 있는	교정하지 않음	제도를 두고 있는 / 제도를 채택하고 있는

특징8. 강화된 띄어쓰기

- ❑ 띄어쓰기가 거의 없고, 과도하게 축약된 문장도 띄어쓰기가 가능함.

예)

❶ 나모시같은거길고통넓은바지사가려구!시원하게 → 나 모시 같은 거 길고 통 넓은 바지 사 가려고! 시원하게

❷ ㅋㅋㅋ나열등감땀에죽어버리고싶어ㅋㅋ시험만아니면학교는재밌음ㅋㅋ중딩보다백배

→ ㅋㅋㅋ 나 열등감 때문에 죽어버리고 싶어. ㅋㅋ 시험만 아니면 학교는 재밌음 ㅋㅋ 중학생보다 백배

- ❑ 신문 기사 등에서 추출한 대용량의 말뭉치로 N-gram 통계 사전을 구축하여 띄어쓰기 교정의 정확도를 높임.

특징9. 풍부한 전문용어

여러 기관과의 공동 개발로 전문용어 보완

- 헌법재판소: 법률 용어
- 법원도서관: 법률 관련 용어
- 한국전자통신연구원: 인터넷에서 사용되는 통신체
- 삼성종합기술원: SMS문자의 약어 및 이모티콘
- 연합뉴스: 언론사에서 사용하는 겹말 오류, 외래어 표기 오류 등
- 산림청: 산림청의 임업용어사전과 임업순화사전
- 문화체육관광부(국립국어원): 순화용어, 외래어표기법
- 교육과정평가원: 교과서에 수록된 과목별 전문용어

보유한 전문용어 양

- 법률용어: 13,500단어 예) 진료수가, 회계보고 등에 관한 예산회계법시행 특례규정
- 과학용어(원소, 분자, 화합물, 유기물 명): 14,200단어 예) 카드늄 →카드뮴, 티옉산 →싸이옉산
- 기타 전문용어: 57,700단어 예) 사이클링히트, 상보성화합물반도체, 부진정연대채무
- 국내 6만여 개의 코스닥 상장기업명 및 법인 이름 예)이레전자산업, KTC텔레콤
- 지도 정보를 이용한 상호 및 지명 추가(533,000여 개)

순화용어와 순화표현을 수시로 반영함.

국립국어원의 외래어심의 자료 반영함.

특징10. 외래어와 신조어·통신체 교정

I 최신 외래어 표기법 적용

- 정비된 외래어 표기법 반영(2020년 1월분까지 반영) – 외래어 인명 포함
- 화학용어 지원
 - 2008.10.09일 기준으로 국립국어원 화학용어 표기법이 대학화학회 명명법을 따름
 - 니트렌디핀 → 나이트렌디핀, 폴리아미드 → 폴리아마이드, 프로필렌글라이콜 → 프로필렌글리콜
- 동남아 언어, 러시아 어, 포르투갈 어와 네덜란드 어 표기법에 따른 고유명사와 오류 유형 추가
 - 도스토예프스키 → 도스토옙스키, 호치민 → 호찌민

I 신조어와 통신체 교정

- “인터넷 언어 전처리 정제 API개발”과제(한국전자통신연구원)를 통해 통신체 교정 기능을 개선함.
 - 하삼(X) → 하세요.(O)
 - 잘지냈찌? (X) → 잘 지냈어? (O)
 - 듣보잡(X) → 듣지도 보지도 못한 잡것/잡것(O)
 - 오나전(X) → 완전(O)
 - KIN(X) → 짜증 남/ 꺼져라(O)

특징11: User Configurable API 제공을 통한 높은 활용성

■ User API 를 통해 활용성을 극대화 할 수 있음.

- ▶ 사전 선택: 목적에 맞게 사전을 선택할 수 있어서, 시스템 경량화, 실행시간 메모리 사용량과 실행시간 감소의 효과를 얻을 수 있음.
- ▶ 교정 방법 선택: 일반 교정을 제외한 다섯 가지 교정 기능을 옵션으로 선택할 수 있음
 - 의미문체 교정: 문장의 앞 뒤 문맥 또는 의미에 따라 교정
 - 다수어절 교정: 다수 어절로 이루어진 한 단어의 오용어 교정
 - 복합명사: 명사 사이의 의미 관계 제약으로 복합명사 허용 여부 판단
 - 띄어쓰기만 교정: 다른 오류는 무시하고 띄어쓰기 오류만 교정
 - 강한 규칙/약한 규칙: 순화용어, 문체 오류, 외래어 허용 등을 강하게 또는 약하게 검사할 수 있음.
- ▶ 교정 결과 선택: 오류 종류, 오류 유형, 밑줄 색, 도움말 번호 등의 정보를 이용하여 교정 결과를 선택할 수 있음.

특징12: 사용자 맞춤형 사전 구성 기능

I 사용자 맞춤형 사전 구성 기능

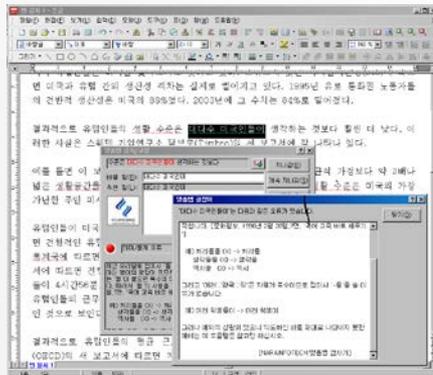
- ▶ 검사기는 일반 사용자를 대상으로 하기 때문에 제한적인 곳에서 사용하는 전문용어나 인명, 특수 표현들을 모두 허용할 수는 없음.
- ▶ 이에 대한 대책으로 사용자가 개인적으로 허용하고 싶은 단어를 맞춤형 사전에 등록하여 사용할 수 있도록 기능을 제공함.
- ▶ 단어 허용뿐만 아니라 사용자가 원하는 대치어로 교정할 수도 있음.

특징13. 다양한 인터페이스 제공

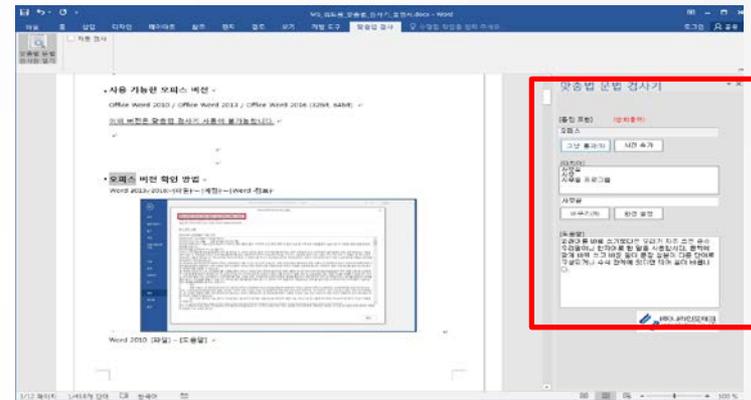
I 다양한 인터페이스를 제공하므로 간단한 API로도 다른 프로그램에 응용할 수 있음

● 아래아한글, 워드 패드, 워드, 웹서비스 등

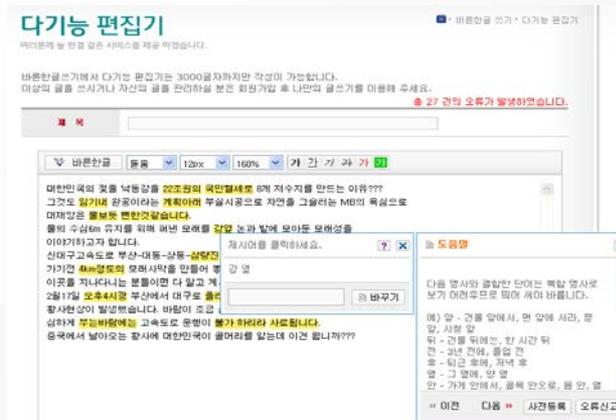
• 적용 사례 1: 한글용 검사기 (한컴 오피스 2018이하 버전에 적용)



• 적용 사례 2: MS Word 용



• 적용 사례 3: 다기능 편집기



특징14. 올바른 한글 사용을 이끔

I 올바른 한글 사용 보급을 위한 홈페이지 운영

- ① 웹을 통해 맞춤법 검사기를 무료 제공하고있음
- ② 우리말 배움터
<http://urimal.cs.pusan.ac.kr>
- ③ 사용자의 요구를 반영하여 검사기 내용 복사 및 대치어 수정 기능 등도 최근 추가함

I 현재까지 우리말 배움터 이용자 수:약 77백만 명

I 일 이용자 수: 3만명 이상 (이력서, 자기소개서, 블로그, 소셜, 수규모 회사 업무용에 주로 사용)



검사기 성능 평가

- 부산대학교 인공지능연구실에서 한국어 맞춤법 문법 검사기 성능평가를 TTA(한국정보통신기술협회)에 의뢰함 (2019.10)
- 문어체와 구어체를 대상으로 각각 오류 검색과 오류 교정의 성능을 평가함.
- 오류 검색과 오류 교정의 F1 score 90%를 목표성능으로 하였고, 이를 PASS 하였다는 통보를 받았음. (현재(20200121) 최종 보고서는 전달받지 못하여, 결과 확인용으로 받은 자료만 아래에 첨부하였음)

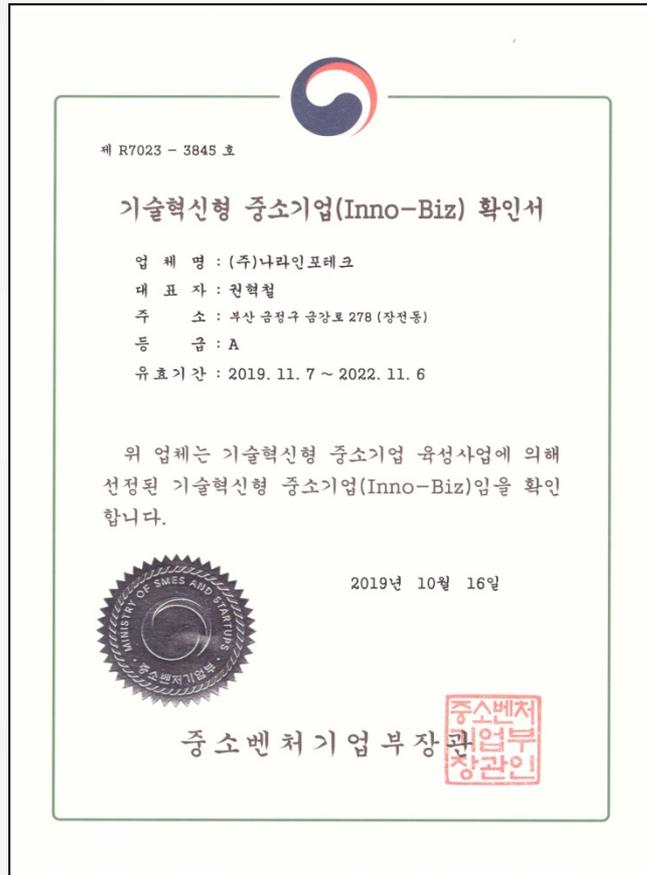
* 참고: F1 score: $2 * (\text{정확도} * \text{재현율}) / (\text{정확도} + \text{재현율})$

부산대학교 산학협력단에서 요청한 “한국어 맞춤법/문법 검사기 v9.0”에 대한 시험 결과는 다음과 같다.

시험항목	시험결과
1. 문어체 오류 F1 score 측정	P
2. 구어체 오류 F1 score 측정	P

특허 및 인증서 [1/3]

I 기술혁신형 중소기업(INNO-BIZ) 확인서(2019.10)



I 연구소 인정서(202001자)



특허 및 인증서 [2/3]

한국어 맞춤법 검사기 및 검사 방법에 관한 특허(2008.01) | 프로그램 등록증 (2005.06)



Good Software (GS) 인증서(2006.01)

