

집단 지성¹에 의한 알타이 언어 문서화

DOCUMENTATION OF ALTAIC LANGUAGES WITH COLLECTIVE INTELLIGENCE

김민규

서울대학교 자연과학대학 화학부
o0o0o@snu.ac.kr; 010-8984-9831

프로젝트의 정의

처리되지 않은 방대한 양의 음성 자료를 게임화²하여 불특정의 일반인들의 집단 지성으로 문서화한다.

프로젝트의 배경

- 기존의 절멸 위기 언어의 연구는 [언어 채록]—[음성 전사]—[자료 분석] 단계를 거친다.
- 음성 자료를 전사하고 정리하는 단계에서 많은 시간과 전문 인력을 필요로 한다.
- ASK REAL 프로젝트로 얻은 방대한 양의 알타이 언어 자료가 연구되지 못한 채 남아있다.

프로젝트의 목표

- 알타이 언어 자료의 문서화 및 오픈 데이터베이스 구축 (언어학적 목표)
- 전사 게임을 이용한 절멸 위기 언어의 문서화 플랫폼 개발 (컴퓨터 공학적 목표)

결과물의 개념도



연구자

- 채록된 언어 자료 트리밍, 태깅, 업로드
- 참여자의 전사 결과물(후보군)에서 정답 선택
→ 정답자 추가 점수 부여
- 전사에 필요한 시간 단축, 대중의 인식 확인
- 전사 자료 검색, 구조화 및 코퍼스 구축

참여자

- 음성학 기본 지식을 위한 튜토리얼
- 새로운 단어를 전사할 때마다 점수 획득
- 유사한 전사 결과가 많을수록 고득점
- 점수가 높아질수록 다양한 인센티브 부여
→ 타 언어에 대한 접근 권한, 고난이도 자료…

¹ 집단 지성(集團知性; Collective Intelligence): 다수의 개체들이 서로 협력하거나 경쟁하는 과정을 통하여 얻게 된 집단의 지적 능력을 의미하며, 이는 개체의 지적 능력을 넘어서는 힘을 발휘한다는 것이다.

² 게임화(-化; Gamification) : 게임에서 흔히 볼 수 있는 재미·보상·경쟁 등의 요소를 다른 분야에 적용하는 기법이다. ‘사람들은 재미를 느끼면 어떠한 활동이든 기꺼이 한다’는 재미 이론을 핵심으로 삼고 있다.

전사 결과와 점수 부여 방식의 예

user	product	points	incentives (by MED)	incentives (by researcher)
user001	gis <u>an</u>	+1	+3	✗
user002	gis <u>ən</u>	+1	+5	✓ +1
user003	gis <u>ən</u>	+1	+5	✓ +1
user004	kisan	+1	+3	✗
user005	gis <u>ən</u>	+1	+4	✗

관련 연구

- Foldit — 2008년 워싱턴대학에서 개발한 폴드잇은 아미노산 연결 해독 게임이다. 게이머는 단백질의 3차원 구조를 흔들거나 구부리는 등 직접 움직여가며 효율적인 형태로 바꿔야 한다. 폴드잇은 게이머 수만 명의 참여 덕분에 3주 만에 정확한 단백질 구조 모델을 파악해 냈다.
- reCAPTCHA — 오래전에 제작된 종이책들을 텍스트화하기 위해 OCR 프로그램을 사용하는데, 낙서나 얼룩, 헤짐 등의 방해요소만 있어도 OCR 프로그램은 텍스트를 제대로 인식하지 못해 노동력과 인건비가 많이 들어간다. 이를 위해 CAPTCHA를 입력하는 많은 사용자들의 힘을 빌리는 것이 바로 리캡처이다.
- Duolingo — 사용자들이 무료로 언어를 배우는 동시에 크라우드 소싱의 방식으로 언어를 번역할 수 있는 플랫폼이다. 듀오링고 교육 방식의 큰 특징 중 하나는 게이미피케이션의 도입이다.
- Phonemica — 포네미카 프로젝트는 중국어의 각종 방언을 사용자가 직접 녹음하여 올리고, 그에 따른 한자 표기, 표준어 번역, 로마자화, 번역 등을 제공함으로써, 거대한 코퍼스를 구축하는 프로젝트이다.
- Forvo — 여러 언어로 된 단어들의 발음을 사용자가 직접 녹음하여 올리는 크라우드소싱 음성 자료 채록 서비스. 외국어 학습자는 이 서비스를 통해 모르는 단어의 발음을 다양한 원어민들로부터 들어볼 수 있다.

