한국어 텍스트 입력 평가를 위한 문장 데이터셋

A Sentence Dataset for Korean Text Entry Evaluations

김문정 송준상 김선준*
Munjeong Kim Junsang Song Sunjun Kim
대구경북과학기술원 대구경북과학기술원 대구경북과학기술원
DGIST DGIST DGIST
moondoong@dgist.ac.kr junid0515@dgist.ac.kr sunjun_kim@dgist.ac.kr
*교신저자

요약문

문자 입력에 대한 성능 평가는 일반적으로 실험 참가자들이 주어진 텍스트를 동일하게 입력하는 방식으로 수행된다. 이에 사용되는 텍스트는 최대한 언중의 언어 습관과 유사하여야 하나, 현재 한국어를 사용하는 언중의 특성을 반영한 문자 입력 평가용 데이터셋은 존재하지 않는다. 본 논문에서는 실제 한국어 사용자들이 데스크탑 환경에서 키보드를 이용해 작성한 대화 말뭉치를 사용하여 한국어 문장 데이터셋을 제작하였다. 또한, 문자 입력 성능평가시 외적 정합성을 높이기 위해 문장을 외운 뒤 입력하는 절차를 자주 사용하는데, 데이터셋의 문장을 외우는데 필요한 시간과 각 문장들을 얼마나 빠르고 정확하게 입력할 수 있는지에 대한 초기 실험 데이터를 제공한다. 이를 기반으로 우리는 한국어 입력 방식을 평가할 때 사용할 수 있는 하위 문장 모음들을 제시하며, 우리의 데이터셋과 데이터셋의 하위 문장 모음들이 추후 한국어 입력 방식을 평가할 때 사용되기를 기대한다.

주제어

한국어, 데이터셋, 텍스트 입력 방식

1 서론

텍스트 입력기에 대한 평가는 일반적으로 주어진 텍스트를 동일하게 입력하는 작업을 수행하고, 입력 속도와 오류율을 측정하여 평가한다. 영어권 문자 입력 방식을 평가하기 위해 널리 사용되는 대표적인 영문 데이터셋(Dataset)으로는 MacKenzie 데이터셋[1]과 Enron Mobile 데이터셋[2]이 있다.

그러나 현재까지 한국어 사용자의 특성을 반영한 한국어 데이터셋은 존재하지 않는 상황이다. 이로 인해 한국어 입력 방식을 평가하기 위해서는 임의로 발췌한 한국어 문장을 사용하거나 [3], 영어로 작성된 데이터셋을 번역하여 활용했던 사례가 있다 [4]. 이러한 방식으로 생성된 문장은 실제 한국어 언중의 특성이 반영된 문장이 아니므로, 한국어 고유의 특성을 충분히 대표하지 못할 가능성이 높다.

아울러, 문자 입력 성능평가 시에는 실험 참가자가 문장을 외운 후 입력하는 절차가 포함될 수 있다. 이 경우, 각 문장이 기억하기 쉬운지 여부를 평가해야 하지만 [2], 한국어 문장의 기억 용이성에 대한 선행 연구는 아직 존재하지 않는다. 따라서, 한국어 입력 평가의 신뢰성과 타당성을 확보하기 위해서는 기억 용이성을 고려한 데이터셋의 개발이 필수적이다. 이는 한국어 언중의 특성을 반영하고, 실험 및 평가의 외적 타당성을 높이는데 중요한 기여를 할 것이다.

본 논문에서는 사용자들이 데스크탑 환경에서 키보드를 이용해 작성한 한국어 문장으로부터 추출한 데이터셋(Dataset)을 소개한다 [5]. 이 데이터셋은 다양한 길이, 문장 부호 및 숫자가 포함된 문장들로 구성되어 있다. 텍스트 입력기 평가에 대한 기초 자료로 사용할 수 있도록, 이들 문장에 대한 기억 용이성, 각 문장에 대한 입력 속도 및 오류율을 측정하였다. 최종적으로, 기억 용이성과 형태소의 분포를 바탕으로 텍스트 입력기 평가에 사용할 수 있도록 특화하여 제작한 하위 문장 모음들을 제시한다.

2 데이터셋 제작 과정

우리는 국립국어원의 언어정보나눔터[6]에서 배포한 '온라인 대화 말뭉치 2021 (버전 1.0)'와 '메신저 말뭉치(버전 2.0)' 2 가지 말뭉치를 사용하였다. 위 말뭉치들은 두 명 이상의 대화 참여자가 온라인 공간의 메신저 (카카오톡 등)에서 자유롭게 주고받은 대화 자료로서, 대화 참여자들의 자연스러운 언어 습관이 그대로 반영되어 있다. 본 논문에서는 데스크탑 환경에서 키보드를 사용한 사용자의 데이터셋을 구성하기 위해 2 개의 말뭉치에서 'PC' 사용자가 '2 벌식(쿼티)' 키보드를 사용한 데이터를 추출하였다.

표 1. Korean-MESSENGER 데이터셋의 특징

Name	Sentence	Morpheme	Morpheme/Sentence	Character/Morpheme	OOV	Question
Korean-MESSENGER	2700	26599	9.851	3.261	0.241%	27.7%

표 2. Korean-MESSENGER 데이터셋의 문장 예시

단어의 수	문장
1 개	반가웠어요.
2 개	진짜 비싸네요.
3 개	사실 저도 그래요.
4 개	오늘 저녁은 뭐 먹었어?
5 개	찾아보니까 우리 동네에도 하나 있네요!

메신저 말뭉치의 특성 상, 대화 참여자들은 한 문장을 온전히 작성하여 전송하지 않고 한 문장을 몇 차례 끊어서 전송하였다. 이를 문장 단위로 구성하기 위해 말뭉치를 사용자별로 구분하여 한 명의 사용자가 전송한 데이터들을 모두 이어 하나의 문자열로 구성한 뒤, 비정형적 텍스트가 포함된 문자열을 제거한 후, 남은 문자열들을 문장 단위로 나누었다.

위의 과정을 통해 최종적으로 총 2,700 개의 문장이 포함된 한국어 메신저 (Korean-MESSENGER) 데이터셋을 만들었다 [5]. 데이터셋은 1 개에서 9 개의 단어(띄어쓰기를 기준으로 0 개에서 8 개의 띄어쓰기를 의미)로 구성된 문장을 각각 300 개씩 포함한다. 각 단어 수에 따른 문장 개수를 300 개로 선정한 이유는, Enron Mobile 데이터셋[2]에서 단어 수별로 균일하게 문장을 구성한 방식과 유사한 분포를 가지도록 설계함으로써 데이터셋의 활용성을 높이고 실험 결과의 비교 가능성을 확보하고자 하기 위함이다.

문법적인 오류가 있거나 비속어 또는 이해할 수 없는 단어들을 제거하기 위해 문장들을 모두 검토하였으며, 저자 두 명이 각각 절반을 검토하고 서로의 결과를 교차 검토하였다. 데이터셋에 포함된 형태소의 수나 OOV(Out of Vocabulary)의 비율 등 데이터셋의 특징은 표 1 에서 확인할 수 있으며 데이터셋에 포함된 문장들의 예시는 표 2 에서 확인할 수 있다.

3 문장의 기억 용이성

문자 입력 실험은 한국어 사용자가 실제로 사용하는 문장을 입력하는 환경과 최대한 유사하게 구성하여 외적 타당성을 확보해야 한다. 문장 입력은 대부분의 경우 주어진 문장을 보고 따라 입력하기보다는 머리속에서 떠오르는 문장을 바로 입력하는 방식으로 문자 입력을 수행한다. 따라서 문장 입력 실험 절차에서도 이를 반영하여, 실험 참가자가 주어진 문장을 먼저 암기한 후 입력 과정에서는 해당 문장이 화면에 표시되지 않도록 설정하는 것이 외적 타당성을 높이는 데 효과적이다 [2]. 이때, 실험에 사용되는 문장은 기억하기 쉬운 문장(기억 용이성이 높은 문장)으로 선정하는 것이 중요하다.

문장의 길이와 문장을 기억하는 데 필요한 시간과의 관계를 파악하기 위해, 대학교 및 대학원 집단에서 모집된 12 명의 한국인 실험 참가자들(남성 9 명, 여성 3 명, 연령 : 24.25±2.09)을 대상으로 한국어 문장의 기억 용이성에 대한 실험을 진행하였다. 각 실험 참가자들에게는 1 개부터 9 개의 단어로 구성된 225 개의 문장들이 제공되었으며, 225 개의 문장들은 각 단어의 개수마다 25 개의 문장들로 구성되어 있다. 이 실험에서는 우리가 제작한 2,700 개의 문장들이 모두 사용되었으며, 각 문장은 1 회씩 평가되었다.

실험 참가자들은 'Start' 버튼을 눌러 (그림 1 왼쪽) 문장을 볼 수 있고 (그림 1 중간), 주어진 문장을 보지 않고 입력할 수 있을 만큼 외우도록 지시 받았다. 만약 충분히 외웠다고 판단되면 실험 참가자들은 'Continue' 버튼을 누르고, 문장이 사라진 뒤 입력 창으로 넘어간다 (그림 1 오른쪽). 참가자들은 외운 문장을 최대한 빠르고 정확하

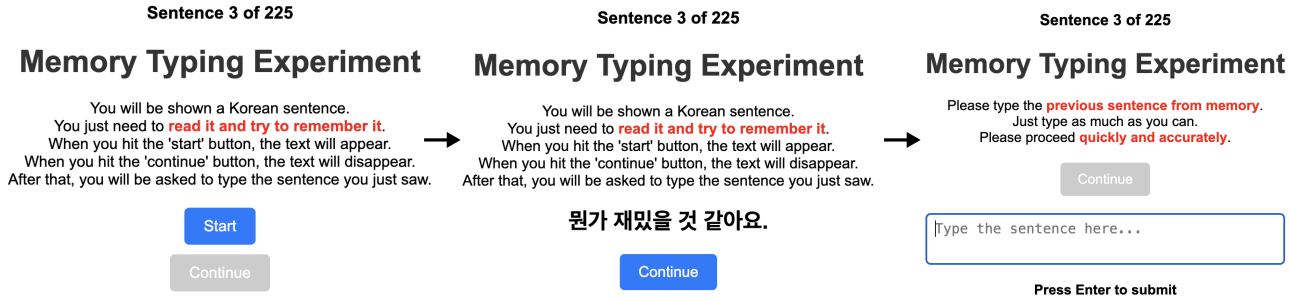

그림 1. 문장의 기억 용이성 실험의 순서. 실험 참가자들은 화면에 주어진 한국어 문장을 외운 후, 문장이 보이지 않는 상태에서 한국어 문장을 최대한 정확히 입력해야 한다.

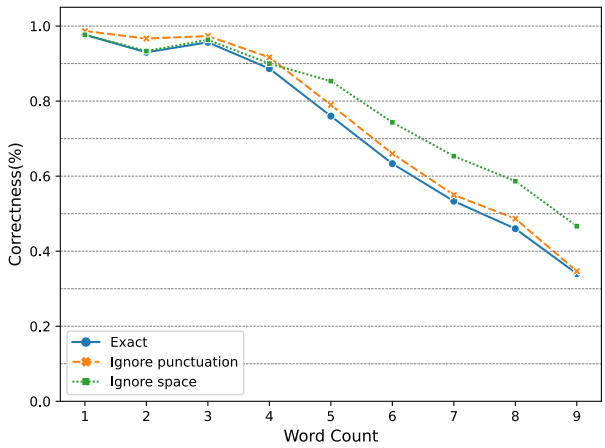

그림 2. 문장의 구성하는 단어의 개수와 정답율

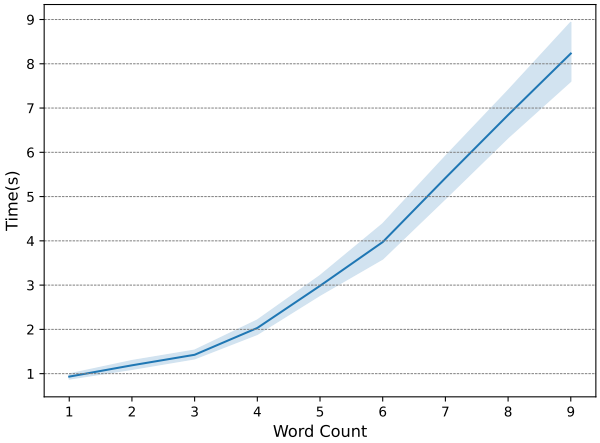

그림 3. 문장을 구성하는 단어의 개수와 문장의 기억하는데 필요한 시간

게 작성하도록 요구 받았다. 문장을 따로 필기하거나, 복사-붙여넣기 등의 행위는 금지되었다.

데이터 분석을 위해, 실험 프로그램은 실험 참가자들이 입력한 모든 활동을 기록하였다. 'Start' 버튼을 누른 시점부터 'Continue' 버튼을 누른 시점까지의 시간을 문장을 기억하는데 필요한 시간으로서 계산하였고, 실험 참

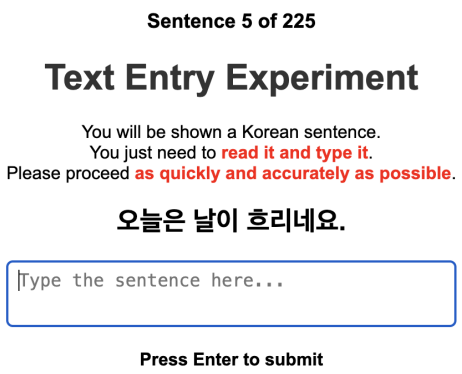

그림 4. 문장 입력 실험. 사용자들은 화면에 나타난 문장을 보고 최대한 빠르고 정확하게 입력해야 한다.

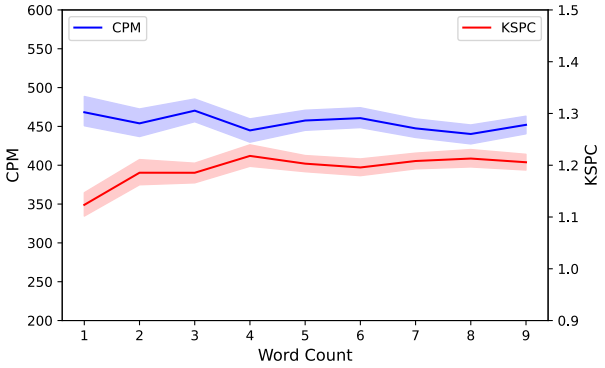

그림 5. 문장을 구성하는 단어의 개수와 CPM 및 KSPC

가자가 제시된 문장을 입력 창에서 정확히 입력한 비율을 계산하였으며, 문장 부호나 띄어쓰기와 관련된 실수를 분석하기 위해 문장 부호나 띄어쓰기를 고려하지 않았을 때 정확히 입력한 비율을 계산하였다.

실험 참가자들은 총 2,700 개의 문장 중 약 71.9%의 문장을 정확히 입력하였으며, 문장 부호를 제외한 경우에는 74.2% (2.3%p 증가), 띄어쓰기를 제외한 경우에는 78.6% (6.8%p 증가)의 정확도를 보여주었다. 또한 그림 2 에 나타난 바와 같이 단어의 개수가 많아질수록 문장을 정확히 입력하는 비율이 감소하였다. 문장을 구성하는 단어의 수가 5 개 이하인 문장들을 정확히 입력한 비율은 90.2%인 반면 문장을 구성하는 단어의 수가 6 개 이상인 문장들을 정확히 입력한 비율은 49.2%를 기록하였다.

그림 3 에는 문장의 길이와 문장을 기억하는데 소요되는 시간과의 관계가 나타나 있으며, 문장을 구성하는 단어의 개수가 많아질수록 문장을 기억하는 데 소요되는 시간이 증가함을 확인할 수 있다.

4 문장의 입력 속도와 오류율

우리는 앞선 실험에 참여한 12 명의 실험 참가자들을 대상으로 한국어 문장 입력 실험을 진행하여 입력 속도와 오류율을 파악하였다. 각 실험 참가자들은 앞선 실험과 동일한 225 개의 문장들을 제공받았으며, 화면에 나타나는 문장을 최대한 빠르고 정확하게 입력하도록 지시 받았다 (그림 4). 문장 입력 실험은 앞선 실험과는 다르게 입력해야 하는 문장이 실험 참가자들이 입력을 시작할 때부터 종료할 때 까지 화면에서 사라지지 않는다.

입력 속도는 한국에서 일반적으로 '타수'로 불리는 분당 문자 수 (Characters Per Minute, CPM) 로 계산되었다. 이 때, 문자 수는 글자 단위가 아닌 자소 단위로 세며, 한글의 자음, 모음, 공백과 문장부호가 한 단위의 문자로 계산되었다. 특히 한글의 모음 중 이중모음인 'ㅚ'나 'ㅟ'

는 2 개의 문자로 계산했다. 시간은 텍스트 입력을 위해 첫 번째 키를 누른 시점부터 입력된 텍스트를 제출 하기 전 마지막 키를 누른 시점까지 측정하였다. 모든 실험 참가자들의 입력 속도는 평균적으로 약 455 CPM 을 기록하였고, 문장을 구성하는 단어의 수에 따라 440 CPM 에서 470 CPM 사이의 변동을 보이나 별다른 경향성은 관찰할 수 없었다 (그림 5).

또한 오류율은 문자 당 키스트로크(Keystrokes Per Character, KSPC)로 계산되었다. KSPC 는 텍스트를 입력할 때 Shift 키와 Backspace 키를 제외한 모든 키스트로크의 수를 입력해야 할 문장의 문자 수로 나눈 값이다. 모든 실험 참가자들의 문장 별 KSPC 는 평균적으로 약 1.19 KSPC 를 기록하였으며, 문장 길이에 대한 KSPC 의 평균값은 문장 길이가 1 단어에서 4 단어로 증가하는 양상을 보였지만, 그 이후는 약 1.20 KSPC 에서 비교적 일정하게 유지되었다 (그림 5).

5 텍스트 입력기 평가용 하위 문장 모음 제시

텍스트 입력기 평가용으로 사용하는 문장 모음은 기억의 용이성, 원본 글뭉치와의 통계적 유사성 등 몇 가지 요구조건을 만족해야 한다. 우리는 아래 3 개의 분류 방식에 따라 하위 문장 모음들을 제시한다. 특히 통계적 유사성을 만족하기 위해 위의 2700 개 문장 세트로부터 Paek and Hsu[7]의 방법론을 따라 표본 추출되었다.

5.1 기억 용이성

기억 용이성 실험의 결과로부터, 기억하기 쉬운 문장 모음 5 개를 생성하였다 (mem1-5 로 명명). 이를 위해, 3 초 내에 기억할 수 있는 문장을 목표로 삼았다. 각 단어의 개수에 대하여 문장들을 외우는데 필요한 시간의 평균을 계산한 결과, 단어의 개수가 2 개에서 5 개 사이일 때 평균적으로 1 초에서 3 초 사이의 시간이 소요되었다 (각각 1.19 초, 1.43 초, 2.03 초, 2.98 초). 단어의 개수가 2 개에서 5 개 사이의 문장들 중 실험 참여자들이 기억 용이성 실험 시 올바르게 입력한 문장들을 추출한 후 각 단어의 개수에 대해서 10 개씩 총 40 개의 문장으로 이루어진 문장 모음 5 개를 제작하였다. 이 문장 모음들은 외우기 쉬운 문장이 필요한 텍스트 입력 평가에 사용하기 적합하다.

5.2 형태소의 분포

Paek and Hsu [7]의 방법론을 따르려면 단어 (word) 단위의 분석이 필요하다, 그러나, 언어의 유형론적 분류에서 고립어로 분류되는 영어와 달리 교착어로 분류되는 한국어는 각자 고유한 의미를 가지는 형태소 (morpheme) 가 병렬적으로 결합되어 (공백으로 구분되는) 하나의 단어를 만들게 된다. 이에 우리는 데이터셋의 언어 습관을 대표하는 하위 문장 모음들을 제작하기 위해 형태소 분석기[8]를 사용하여 단어를 형태소 단위로 쪼갠 뒤, 형태소 단위의 n-그램(n-gram) 세트를 무작위로 샘플링(Sampling) 하여 전체 데이터셋과 가장 유사한 바이그램(bi-gram) 분포를 가지는 모음을 선택하는 절차를 이용하였다 [7]. 우리는 전체 Korean-MESSENGER 데이터셋의 바이그램 분포와 비슷한 40 개, 80 개, 160 개 및 320 개의 문장으로 구성된 문장 모음 5 개를 생성하였다 (각각 bi40, bi80, bi160, bi320 으로 명명). 이 문장 모음들은 1 개에서 9 개 사이의 단어들로 구성되어 있으며, 각 문장 모음들은 3.63, 2.98, 2.31, 1.69 의 엔트로피를 기록했다. 온라인 메신저의 글자 분포를 대표하는 문장 모음들이 필요한 경우 이 문장 모음들을 사용하는 것을 권장한다.

5.3 기억 용이성 & 형태소의 분포

또한 우리는 기억하기 쉬운 5 개의 문장 모음에 포함된 200 개의 문장에서, Korean-MESSENGER 데이터셋과 바이그램 분포가 가장 비슷한 40 개의 문장을 추가로 선정하였다 (mem-bi 로 명명). 이 문장 모음은 3.88 의 엔트로피를 보였다. 이 문장 모음은 기억하기 쉬우면서 온라인 메신저의 글자 분포를 대표하는 문장 모음들이 필요한 경우에 사용하기 적합하다.

6 논의 및 한계

Korean-MESSENGER 데이터셋은 한국어 입력 방식의 특성을 평가하는 연구를 수행하는 데 있어서 높은 외부 타당성을 제공한다. 우리가 제공하는 데이터셋에 포함된 문장들은 실제 한국어 사용자들이 데스크탑 환경에서 키보드를 사용해 작성한 일상적인 대화이다. 기존에 임의의 한국어 문장을 사용하거나 영어 데이터셋을 한국어로 번역하여 텍스트 입력 평가를 진행했던 한계를 보완하며, 한국어 사용자들이 데스크탑 환경에서 키보드를 이용해 실제로 작성한 한국어 문장들로 이루어진 최초의 한국어 데이터셋이라는 점에 의의가 있다.

또한 우리는 실험실 환경에서 인간 대상 실험을 통해 한국어 문장의 기억 용이성, 입력 속도와 오류율에 대한 실험적 데이터를 제공하였다. 실험실 환경은 실험을 진행할 때 주어진 문장을 따로 기록하지 못하게 하거나 텍스트 입력 과정 중간에 쉬지 못하도록 통제도록 하여 높은 실험 통제력을 보여줄 수 있었다. 다만 크라우드 소싱 (Crowd Sourcing) 과 같은 방법을 이용해 더 많은 실험 참가자를 모집하였다면 하나의 문장에 대한 더 많은 반복 실험이 가능했을 것이다.

우리의 데이터셋은 문장부호와 숫자를 포함한 2,700 개의 문장으로 구성되어 있고, 각 문장들의 기억 용이성과 형태소의 분포를 바탕으로 제작된 하위 모음들도 제시하였다. 데이터셋은 2,700 개 각각의 문장에 대한 기억 용이성, 입력 속도와 오류율을 담은 메타데이터를 포함하므로 텍스트 입력의 종단적 평가가 가능하며 특정 텍스트 요구 사항을 충족하는 모음들을 생성하기에 충분한 데이터이다.

7 결론

우리가 제시하는 Korean-MESSENGER 데이터셋[5]은 키보드를 이용한 텍스트 입력 평가를 위한 한국어 데이터셋으로서, 실제 한국어 사용자들이 데스크탑 환경에서 키보드를 이용해 온라인 메신저에서 나눈 대화를 기반으로 구성되었다. 인간 대상 실험을 통하여 데이터셋에 포함된 문장의 기억 용이성에 대한 실험적 데이터를 수집하였고, 각 문장이 얼마나 빠르고 정확하게 입력될 수 있는지를 조사하였다. 또한 기억 용이성과 형태소 분포를 기준으로 샘플링한 하위 문장 모음들도 제시하였다. 본 데이터셋과 데이터셋에 대한 실험적 결과가 다른 연구자들이 한국어 기반 텍스트 입력 평가를 진행하는 데 유용하게 활용되기를 바란다.

(본 연구에서 소개하는 Korean-MESSENGER 데이터셋은 학술 및 비상업적 연구 목적으로 사용할 수 있으며, 데이터셋에 대한 세부 정보 및 활용 방법은 페이지[5]를 참고하시기 바랍니다.)

사사의 글

본 연구성과물은 2024 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.RS-2024-00395955). 또한 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.RS-2023-00211872, No.RS-2023-00223062).

참고 문헌

1. MacKenzie, I. S. and Soukoreff, R. W. Phrase sets for evaluating text entry techniques. In CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03). pp. 754–755. 2003.
2. Vertanen, K. and Kristensson, P. O. A versatile dataset for text entry evaluations based on genuine mobile emails. In Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11). pp. 295–298. 2011.
3. Song, Y., Liza, S. J., and Oakley, I. Typing on the Edge: Korean Text Entry on a Smartwatch Using a Side Mounted Input Surface. 한국 HCI 학회 학술대회 (HCI Korea '16). pp. 223–225. 2016.
4. Ilinkin, I. and Kim, S. Evaluation of text entry methods for Korean mobile phones, a user study. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10). pp. 2023–2026. 2010.
5. Munjeong Kim. Korean-MESSENGER. https://github.com/munjkim/Korean-MESSENGER. Dec 2, 2024.
6. 언어정보나눔터. 모두의 말뭉치. https://kli.korean.go.kr/corpus/main/requestMain.do?lang=ko. Dec 1, 2024.
7. Paek, T. and Hsu, B. J. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11). pp. 2477–2480. 2011.
8. Park, E. L. and Cho, S. KoNLPy: Korean natural language processing in Python. Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology. 2014.