Sentence Level Classifier

Motivation

Students learn better with examples that fit their proficiency levels, but very few resources exist to…

● present students with example sentences at the appropriate level.

● automatically rate the level of a learner-written sentence.

Problem Statement

Given a sentence, automatically predict its CEFR level using a multiclass classification model.

CEFR

- Common European Framework of Reference for Languages
- put together by the Council of Europe as a way of standardising the levels of language exams in different regions
- six-point acale, from A1 (Beginner) to C2 (Mastery)

source: https://www.cambridgeenglish.org/exams-and-tests/cefr/
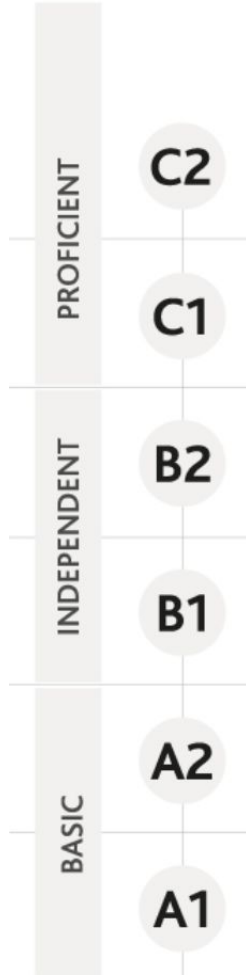
Related Work

Existing systems for English CEFR classification
- Write & Improve (Cambridge)
- Criterion (ETS)

Existing models for English CEFR classification
- on documents (Gaillat et al. 2021, Kerz et al. 2021, Yannakoudakis et al. 2018)
- on words (Sohsah et al. 2018)

Existing models for sentence level classification
- Binary sentence complexity classification (Bosco et al. 2021, Dell'Orletta et al. 2014)
- Three level CEFR classification on Arabic sentences (Khallaf and Sharoff 2021)

Research Gap

- Gap in state of the art: no six-level CEFR models available.
- Lack of CEFR-graded sentence datasets.
- Sentence classification is challenging because sentences are short and therefore less rich in features to extract and aggregate.
- The more classes there are, the more difficult it is for a model to distinguish.

Required Data

- Training a predictive model requires a substantial amount of training data.
- Required format

Sentence (text)	CEFR Level (label)
"You can sit on the chair outside and look at the bridge."	A1
"This could have a detrimental effect on our city and, to be more specific, on tourism."	C2

- However, no annotated datasets of this nature are publicly available.

Creating the dataset

- Acquire sentences from the English Vocabulary Profile (EVP).
- EVP contains words with their corresponding CEFR levels, as well as two types of example sentences, **dictionary** and **learner**.
- Aggregate both dictionary and learner examples into a dataset of size 22,654.

Core Assumption:
All example sentences are considered to have the same CEFR level as the headword.

outside · *adverb* 🔊 /ˌaʊtˈsaɪd/

─ outside (NOT IN BUILDING)

A1 not inside a building

Dictionary examples:
It's cold outside!
Go and play outside for a while.
*Shall we eat outside as the weather is so nice?*

Learner example:
You can sit on the chair outside and look [at the] bridge. (Skills for Life (Entry 1); A1; Slovak)

source: https://www.englishprofile.org/wordlists

Training the Model

- Split the data into train (80%), validation (10%), and test (10%) sets
- Convert training data to BERT embedding vectors
  - BERT (Bidirectional Encoder Representations from Transformers): a language model developed by Google researchers in 2018 and trained on over 11,000 books + Wikipedia
  - Language models can numerically represent language by transforming phrases, words, or word pieces (parts of words) into vectors
- Train the model to predict the CEFR level of any given sentence

Model Overview
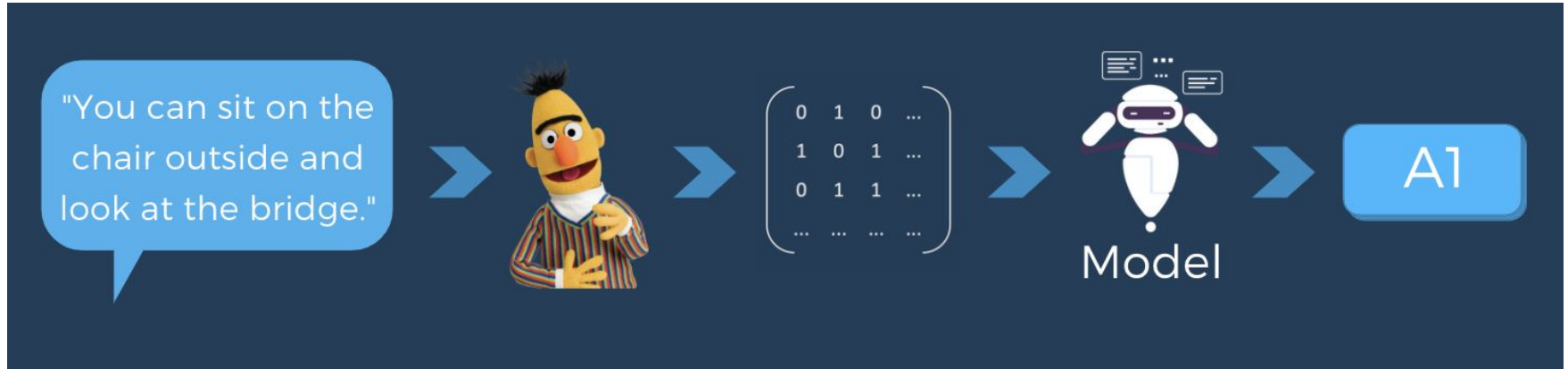
Evaluation Metrics

- **Six Level Accuracy**
  % sentences where prediction = label
- **Three Level Accuracy**
  % sentences where prediction's general level = label's general level (A, B, C)
- **Fuzzy Accuracy**
  % sentences where the deviation of the model's prediction from the label <= 1
  (i.e. label A2, prediction A1 or A2 or B2)
- **Mean Absolute Error**
  Between 0 and 5. Average amount that the prediction deviated from the label.
  - 0 = no deviation (i.e. A1, A1)
  - 1 = 1 deviation (i.e. A1, A2)
  - 2 = 2 deviations (i.e. A1, B2)

Future Work

- Use more training data from other sources (ex: Cambridge Dictionary)
- Incorporate features into the model (ex: part-of-speech tags)
- Train a classifier for phrases or documents
- Paraphrasing: generate higher/lower level versions of a given sentence