Post-editing and rescoring of automatic speech recognition results

Post-edycja i ponowna ocena wyników rozpoznawania mowy

Danijel Koržinek

<danijelk@pja.edu.pl>

18. czerwca 2020

Cel zadania

Celem zadania jest stworzenie rozwiązania konwertującego sekwencję wyrazów wygenerowaną za pomocą pewnego systemu rozpoznawania mowy (ASR) w inną sekwencję która lepiej odzwierciedla wymówioną wypowiedź.

Motywacja

- Typowy (nie-E2E) przepływ ASR:
  - audio $\rightarrow$ model akustyczny $\rightarrow$ G2P $\rightarrow$ model języka $\rightarrow$ dekoder $\rightarrow$ tekst
  - model języka to zwykle 3- lub 4-gramowy statystyczny LM z ograniczonym słownikiem
- E2E rozmywa granice pomiędzy poszczególnymi etapami przetwarzania
  - ale generuje jeszcze mniej przewidywalne wyniki
- ASR zazwyczaj dostarcza wyniki w następujących postaciach:
  - 1-best
  - N-best (około 100 lub 1000)
  - krata (ang. lattice)
- “Re-scoring” to typowe rozwiązania do radzenia z przetwarzaniem danych audio
  - używa mały i szybki LM do wygenerowania wyniku w pierwszym przebiegu
  - ponowna ocena kraty w celu wybrania lepszej sekwencji na wyjściu

Sugerowane podejście

- “Re-scoring” modelem statystycznym
- “Re-scoring” przez NNLM
- SMT
- modele seq-to-seq
- BERT, GPT, Transformer, etc...

Dane wejściowe

- 1-best
  - każde nagranie zawiera jedną najlepsza transkrypcję wypowiedzi
- n-best
  - każde nagranie zawiera do 100 najlepszych hipotez transkrypcji
- krata
  - każde nagranie zawiera listę krawędzi tworzących graf kraty
  - każda linia grafu zawiera następujące pola: węzeł początkowy, węzeł końcowy, wyraz, waga modelu języka, waga modelu akustycznego, lista stanów fonetycznych

Ewaluacja

$$WER = \frac{N_{del} + N_{sub} + N_{ins}}{N_{ref}}$$

- Na podstawie pakietu NIST Sclite

Kontakt

danijel@pja.edu.pl