

Michał Junczyk
Poznań, 11.06.2024r.

Zastosowanie metod zarządzania zbiorami nagrań mowy do oceny jakości systemów automatycznego rozpoznawania mowy dla języka polskiego.

Celem rozprawy była poprawa użyteczności dostępnych zbiorów danych mowy oraz ich wykorzystanie do oceny jakości systemów ASR dla języka polskiego. Pierwsze zadanie polegało na rozwiązaniu problemów związanych z dostępnością i interoperacyjnością tych zbiorów. W tym celu stworzono przekrojowy katalog polskich danych mowy ASR, obejmujący 53 zbiory opisane za pomocą 66 atrybutów. Katalog ten umożliwił zidentyfikowanie zbiorów nagrań mowy dostępnych na otwartych licencjach. Wybrane zbiory danych zostały zorganizowane w sposób ułatwiający ich wykorzystanie przez praktyków zajmujących się rozwojem systemów ASR. Zorganizowane zbiory danych zostały nazwane "BIGOS (*Benchmark Intended Grouping of Open Speech*)" oraz "PELCRA for BIGOS", gdzie PELCRA to nazwa grupy badawczej z Uniwersytetu Łódzkiego, która zgodziła się udostępnić swoje zbiory na potrzeby badania i otwartego wyzwania dla społeczności. Zbiory łącznie zawierają ponad 800 godzin nagrań i niemal 400 000 nagrań od 5 000 mówców. Wybrane nagrania z obu uporządkowanych zbiorów zostały wykorzystane do oceny jakości 25 modeli ASR na 24 podzbiorach o zróżnicowanych charakterystykach, co było największym tego typu badaniem dla języka polskiego. W celu zwiększenia wiarygodności analiz oraz ułatwienia replikacji wyników stworzono i udostępniono system do przeprowadzania testów, oraz analizy wyników. Przeprowadzone prace miały także na celu promocję standardowych metod oceny jakości systemów ASR dla języka polskiego. Cel ten zrealizowano poprzez publiczne udostępnienie wyników badań i organizację otwartego konkursu w ramach programu PolEval.

Poniżej opisano strukturę rozprawy doktorskiej.

Rozdział 1 wprowadza cele i pytania badawcze. Przedstawione jest tło problemu z podkreśleniem znaczenia zbiorów danych mowy w procesie rozwoju i oceny jakości systemów ASR. Następnie opisano wyzwania dot. zarządzania zbiorami danych oraz w praktykach testowania systemów ASR dla języka polskiego. Zdefiniowane cele badawcze obejmują poprawę dostępności i interoperacyjności istniejących zbiorów danych mowy dla jęz. polskiego, przygotowanie przekrojowych i łatwych w użyciu zbiorów danych mowy z użyciem zasobów dostępnych publicznie na otwartych licencjach, opracowanie procedur systematycznej oceny jakości (ang. benchmarkingu) systemów ASR oraz promocję opracowanych zbiorów danych i narzędzi w społeczności ASR w Polsce.

Rozdział 2 zawiera szczegółowy przegląd literatury, omawiający dotychczasowe prace nad oceną jakości systemów ASR oraz zarządzaniem zbiorami nagrań mowy. Opisane zostały wyzwania związane z zarządzaniem i oceną jakości systemów ASR. Przeanalizowano również metody i narzędzia do zarządzania danymi i testowania jakości systemów ASR stosowane w międzynarodowym środowisku badawczym i przemysłowym. W celu pogłębienia uzasadnienia potrzeby standaryzacji metod oceny jakości systemów ASR dla języka polskiego przytoczone zostały przykłady dobrych praktyk testowania z innych dziedzin uczenia

maszynowego. Rozdział ten opisuje także wspólne wyzwania w procedurach oceny jakości uczenia maszynowego, tworząc podstawy do zrozumienia luk i możliwości w zarządzaniu danymi i ocenie jakości polskich systemów ASR.

Rozdział 3 opisuje metodologię badawczą używaną w badaniu. Najpierw opisano proces zbierania danych oraz atrybuty użyte do klasyfikacji istniejących polskich zbiorów danych. Następnie opisano założenia przyjęte podczas projektowania zorganizowanego zbioru, oryginalne źródła, z których wykorzystano nagrania, transkrypcje i metadane oraz proces organizacji zbioru. Opisano również metodykę przeglądu publicznie raportowanych testów porównawczych systemów ASR dla jęz. polskiego. Rozdział obejmuje również opis narzędzi stworzonych do oceny jakości, w tym definicje metryk oceny, protokołów oraz opis testowanych systemów. Na koniec opisana jest organizacja otwartego konkursu, stworzonego w celu promocji i zachęcenia społeczności do korzystania z przygotowanych zbiorów danych.

Rozdział 4 rozpoczyna się od przedstawienia wyników przeglądu polskich zbiorów danych mowy ASR. Następnie szczegółowo opisano zorganizowane zbiory danych za pomocą zestawu standardowych metryk. Metryki te zostały zainspirowane praktykami stosowanymi w przodujących ośrodkach przemysłowych zajmujących się tematyką rozpoznawania mowy. Reprezentują one cechy zbiorów danych, takie jak długość nagrania, tempo mowy czy liczba unikalnych słów i zostały uzyskane za pomocą narzędzi opracowanych w ramach prac badawczych. Następnie przedstawiono wyniki przeglądu dot. testów porównawczych w celu oceny stanu wiedzy nt. możliwości technologii ASR i procedur testowych. W kolejnej sekcji przedstawiono wyniki oceny jakości systemów ASR dla różnych scenariuszy testowych m.in. porównanie jakości systemów na wszystkich zbiorach, na specyficznych podzbiorach, systemów otwartych i komercyjnych, w funkcji długości nagrania, dla różnych grup społeczno-demograficznych. Przeprowadzona analiza demonstruje przydatność zorganizowanego zbioru oraz skuteczność zaprojektowanego systemu w ujawnianiu różnic pomiędzy ocenianymi systemami ASR wynikających z cech testowanych modeli, oraz danych testowych.

Rozdział 5 omawia rezultaty badań. Mocne i słabe strony ocenianych systemów zostały zinterpretowane z uwzględnieniem wiedzy nt. cech zbiorów danych testowych. Dyskusja obejmuje również potencjalny wpływ zaproponowanych procedur oceny jakości na stosowane praktyki w nauce i przemyśle. Omówiono również potencjalne ulepszenia dot. zbiorów danych i praktyk oceny jakości do rozważenia w przyszłości. Rozdział kończy się podkreśleniem znaczenia standaryzacji metod oceny i zaangażowania społeczności w dalszych rozwój technologii ASR w Polsce.

Niniejsza praca wnosi kilka kluczowych wkładów do dziedziny polskiego ASR:

- Przegląd i katalog polskich zbiorów danych mowy ASR w celu zwiększenia ich dostępności i użyteczności.
- Zwiększona dostępność oryginalnych zbiorów danych poprzez dobrze zorganizowany katalog, z uznaniem oryginalnych autorów.

- Organizacja zbiorów do oceny jakości systemów ASR (BIGOS i PELCRA) w wygodnym formacie.
- Przegląd publicznie dostępnych testów porównania jakości systemów ASR dla jęz. polskiego
- Stworzenie systemu do systematycznej oceny jakości systemów ASR.
- Organizacja otwartego wyzwania dla społeczności ASR w Polsce
- Ulepszenie praktyk dot. dokumentacji i analizy w celu lepszego zrozumienia jakości systemów ASR i oraz zbiorów danych używanych do oceny.

Michał Junczyk