Koncepcja Testów Jakościowych Różnych Narzędzi OCR pod kątem zastosowań do odczytywania deklaracji podatkowych

dr hab. A. Jankowski, Profesor UWM

17 lutego 2025

Agenda

Wstęp

- Coraz więcej deklaracji podatkowych w Polsce jest odczytywanych maszynowo.
- System OCR (Optical Character Recognition) w Poltax ma kluczowe znaczenie dla efektywności przetwarzania tych deklaracji.
- Konieczne jest wypracowanie koncepcji testów jakościowych OCR, w szczególności pod kątem danych pismem ręcznym i drukowanym.

Cele Prezentacji

- Przedstawienie wieloetapowej koncepcji testowania narzędzi OCR.
- Omówienie testów na danych syntetycznych w różnych skalach.
- Wskazanie możliwych metod statystycznych (np. podejście Bayesowskie) do wyznaczania liczności próbek.
- Zdefiniowanie kluczowych miar jakościowych (metryk) testów.
- Omówienie przykładowych procedur i środowiska testowego.

Zakres Testów

- Testujemy różnych dostawców oprogramowania OCR w kontekście:
  - Rozpoznawania tekstu drukowanego (np. dane osobowe, adresowe).
  - Rozpoznawania pisma ręcznego (szczególnie podpisy, dodatkowe uwagi, komentarze).
- Rozpoznawane pola (w tym priorytety ważności):
  - Pola liczbowe (kwoty, numery identyfikacyjne, NIP, PESEL itp.).
  - Pola tekstowe (nazwiska, nazwy, adresy, nazwy urzędów).
- Weryfikacja formatu i poprawności (np. PESEL musi mieć 11 cyfr i algorytm weryfikujący prawdziwość numeru PESEL).

Założenia Testowe

- Dostęp do danych prawdziwych jest ograniczony z uwagi na ochronę danych osobowych.
- Dlatego: **wykorzystujemy dane syntetyczne**, odzwierciedlające rzeczywiste struktury formularzy podatkowych. Dane mogą być generowane ręcznie oraz automatycznie.
- Dokumenty generowane masowo (nawet setki milionów) w celu automatycznego testowania.
- **Losowe próbki** weryfikowane ręcznie przez ekspertów.
- Rozmiar próbek statystycznych ustalany m.in. metodą bayesowską.

Przegląd Fazy Testów

1. **Faza wstępna (pobieżna ocena)**
2. **Faza dokładna (szczegółowa ocena)**
3. **Faza walidacyjna (testy na większej próbie)**
4. **Faza masowa (testy wysokoskalowe)**

Faza Wstępna (Pobieżna Ocena)

- **Pretesty** Generujemy (ręcznie i automatycznie) ciągi znaków i wyrazów (np. na setki stron). Następnie traktujemy te strony jako wejście do programów OCR i automatycznie porównujemy dane wejściowe z efektami programu testowego. Na tej podstawie mamy wstępną oceną programów
- **Generowanie niewielkiej partii dokumentów** (np. kilkaset lub kilka tysięcy) zawierających:
  - Różne typy formularzy (PIT, CIT, VAT).
  - Różnorodne układy tekstu i pisma ręcznego.
- Przygotowanie skryptów i narzędzi do automatycznego rozpoznawania z formularzy.
- Szybkie porównanie wyników OCR z danymi źródłowymi (ground truth).
- Wskaźniki wstępne:
  - Ogólny poziom poprawności (accuracy).
  - Liczba pól rozpoznanych prawidłowo vs. nieprawidłowo.
- Celem jest pobieżne wstępne sprawdzenie, czy narzędzie działa w ogóle zgodnie z oczekiwaniami.

Faza Dokładna (Szczegółowa Ocena)

- Przygotowanie **większego i zróżnicowanego** zestawu dokumentów (np. kilkadziesiąt tysięcy).
- Zastosowanie **różnych wariantów pisma ręcznego** (pochyłe, niewyraźne, wielkie/małe litery).
- Uwzględnienie **błędów** wypełniania deklaracji (przekreślenia, dopiski).
- Dokładna analiza rozbieżności:
  - Analiza przypadków brzegowych (np. nietypowe formaty, brak danych).
  - Kategorie błędów (OCR nie rozpoznaje, myli litery/cyfry, błędna segmentacja pól).
- Wypracowanie **rekomendacji usprawnień** dla dostawców OCR.
- Definiowanie **precyzyjnych metryk** (np. *pole-level accuracy*, *character-level accuracy*, *field extraction rate*).

Faza Walidacyjna (Testy na Większej Próbie)

- **Losowe próbkowanie** z dużej puli (np. miliony wygenerowanych dokumentów).
- **Eksperci weryfikują wyniki uzyskane automatyczne** na próbkach statystycznie istotnych. Można tu zastosować podejście Bayesowskie.
- Wykorzystanie metod statystycznych:
  - Podejście bayesowskie do ustalania liczności próbek.
  - Wyliczanie przedziałów ufności dla wskaźników jakości.
- **Analiza porównawcza** pomiędzy różnymi rozwiązaniami OCR.
- Raportowanie i definiowanie **kryteriów akceptacji** (np. minimalna skuteczność na poziomie 95%).

Faza Masowa (Testy Wysokoskalowe)

- Wykorzystanie **olbrzymich zbiorów** danych (setki milionów syntetycznych dokumentów).
- Automatyczne przetwarzanie OCR na klastrze serwerów / w chmurze.
- **Monitoring błędów** i anomalii w czasie rzeczywistym.
- **Próbkowanie** jakości wyników w **różnych punktach czasowych** (np. co 10 mln dokumentów).
- Monitorowanie i porównanie parametrów wydajnościowych (liczba jednoczesnych wątków i czas przetwarzania).
- Podsumowanie końcowe:
  - Wyliczenie globalnego wskaźnika poprawności.
  - Określenie powtarzalności i wydajności w skali masowej.

Generowanie Danych Syntetycznych

- Wykorzystanie szablonów formularzy podatkowych (PIT, CIT, VAT).
- Losowe wypełnianie pól:
  - Tekst drukowany (nazwiska, adresy, NIP, PESEL).
  - Pismo ręczne generowane np. przy użyciu technologii sztucznej inteligencji (AI) do symulacji odręcznego pisma.
- Dodawanie **szumów, artefaktów druku**, odkształceń, nierówności skanowania.
- Utrudnienia i błędy charakterystyczne dla rzeczywistości:
  - Nadpisywanie pól.
  - Zamazywanie (np. odciski palców).
  - Słaba jakość druku.

Przykładowe Metryki Jakości OCR

- **Character Error Rate (CER)** – % błędnych znaków.
- **Word Error Rate (WER)** – % błędnych słów (używane częściej w językach ciągłych, np. rozbudowane teksty).
- **Field Accuracy** – % poprawnie rozpoznanych pól (np. imię, nazwisko).
- **Field Completeness** – stosunek pól uzupełnionych do wszystkich pól wymaganych.
- **False Positive Rate (FPR) i False Negative Rate (FNR)** w kontekście rozpoznania obecności pola.
- **Czas przetwarzania / liczba wątków** (wydajność).

Automatyzacja Testów

- Budowa **zautomatyzowanego pipeline'u**:
  1. Generacja dokumentów (syntetycznych).
  2. Skanowanie (lub symulacja skanowania) – tworzenie obrazów.
  3. Przekazywanie do narzędzia OCR.
  4. Zbieranie wyników rozpoznania.
  5. Porównanie wyników z *ground truth*.
  6. Generowanie raportów.
- **Losowe próbkowania** – wybór dokumentów do ręcznej weryfikacji przez ekspertów.
- System powiadomień o **istotnych odchyleniach** (np. skokowy wzrost błędów).

Statystyka w Testach Jakości OCR

- Celem jest oszacowanie **wiarygodności wyników** na podstawie próbek.
- Metody:
  - **Klasyczna statystyka** (testy hipotez, przedziały ufności).
  - **Podejście bayesowskie** – aktualizacja wiedzy na temat jakości OCR w miarę napływu danych.
- Parametry do ustalenia:
  - Wielkość próby (np. z wzoru na margines błędu).
  - Poziom ufności (np. 95%).

Szacowanie Liczności Próbek

- **Przedział ufności** dla wskaźnika (np. dokładności OCR) obliczany np. ze wzoru:

$$n \geq \left(\frac{z_{\alpha/2} \sqrt{p(1-p)}}{E} \right)^2$$

gdzie:

- $z_{\alpha/2}$ – wartość krytyczna (dla 95% $\approx 1{,}96$),
- p – spodziewany poziom dokładności,
- E – zakładany margines błędu.

- W podejściu **bayesowskim** można dodatkowo:
  - Użyć rozkładu a priori na p (np. Beta).
  - Aktualizować rozkład *posteriori* po każdej partii testów.

Porównanie Dostawców OCR

- Po zgromadzeniu wyników z **fazy walidacyjnej** i **faz masowych**:
  - Tworzymy **ranking** pod kątem najważniejszych metryk (accuracy, FPR, FNR, wydajność).
  - Analizujemy **kategorie błędów** specyficzne dla każdego dostawcy (np. część gubi polskie znaki diakrytyczne).
- Oceniamy **elastyczność** w dostosowywaniu do nowych rodzajów dokumentów.
- Sprawdzamy **skalowalność** rozwiązania (jak radzi sobie z rosnącą liczbą dokumentów).
- Wskazujemy **koszty licencyjne** vs. osiągane wyniki jakościowe.

Rekomendacje

1. Wdrożenie **cyklicznych testów regresyjnych** w oparciu o dane syntetyczne.
2. Wdrożenie **technik walidacyjnych** poprawności wprowadznych danych.
3. Automatyzacja procesu **zbierania i raportowania błędów** w środowisku produkcyjnym.
4. Rozbudowa modeli OCR o **moduły ML/AI** uczące się nowych wzorców pisma ręcznego.
5. **Standaryzacja formularzy** (jeśli to możliwe) w celu poprawienia rozpoznawalności.
6. Wypracowanie minimalnych **kryteriów akceptacji** (np. 95% Field Accuracy).

Podsumowanie

- Wieloetapowe testy (od pobieżnych do masowych) pozwalają rzetelnie ocenić narzędzia OCR i zoptymalizować koszty testów
- Wykorzystanie **automatycznie generowanych danych syntetycznych** umożliwia testy na skalę niemożliwą do osiągnięcia w przypadku danych rzeczywistych.
- **Losowe próbkowanie** i **metody statystyczne** zapewniają miarodajną ocenę jakości.
- Wprowadzenie **cyklicznych testów regresyjnych** pozwoli utrzymać wysoką jakość w dłuższej perspektywie.
- Końcowym celem jest **rzetelna i skalowalna** automatyzacja rozpoznawania deklaracji podatkowych w systemie Poltax.

Dziękuję za uwagę!

Pytania?