1. Wprowadzenie do ewaluacji modeli językowych

Ewaluacja modeli językowych to proces systematycznej oceny ich jakości i użyteczności w określonych zadaniach. Modele językowe (ang. Language Models, LM) są dziś stosowane w szerokim spektrum zastosowań: od tłumaczenia automatycznego, przez automatyczne streszczanie tekstów, aż po generowanie kodu czy prowadzenie dialogów z użytkownikiem.

Kluczowym wyzwaniem ewaluacji jest fakt, ze język naturalny jest niejednoznaczny — ten sam sens można wyrazić na wiele sposobów. Dlatego też żadna pojedyncza metryka nie jest w stanie w pełni uchwycić jakości generowanego tekstu. W praktyce stosuje się zestawy różnych miar, które wzajemnie się uzupełniają.

1.1. Dlaczego ewaluacja jest trudna?

Ocena jakości tekstu naturalnego natrafia na szereg trudności fundamentalnych:

- **Wieloznaczność języka:** To samo zdanie może być wyrażone na wiele sposobów, a wszystkie będą równorzędnie poprawne.
- **Subiektywność oceny:** Ludzie różnią się w ocenie tego, co jest "dobrym" tekstem — zależy to od kontekstu, kultury i indywidualnych preferencji.
- **Zależność od zadania:** Metryki przydatne w tłumaczeniu mogą nie mieć sensu w zadaniach dialogowych.
- **Koszt oceny ludzkiej:** Ocena przez ludzi jest droga, czasochłonna i trudna do skalowania — stad potrzeba metryk automatycznych.

1.2. Podział metryk

Metryki używane do oceny modeli językowych można podzielić na dwie główne kategorie:

- **Metryki automatyczne:** Obliczane algorytmicznie, bez udziału człowieka. Szybkie, powtarzalne i skalowalne, lecz często niewystarczające same w sobie.
- **Ocena ludzka (Human Evaluation):** Polega na zaangażowaniu ludzi – jako sędziów, którzy oceniają teksty według ustalonych kryteriów. Jest to złotym standardem, lecz procedura jest kosztowna i niepodatna na skalowanie.

Poniżej zostaną przedstawione kluczowe metody które koncentrują się na automatycznych metrykach podstawowych, i które stanowią fundament warsztatu badacza NLP (przetwarzania języka naturalnego).

2. BLEU — Bilingual Evaluation Understudy

2.1. Geneza i historia

BLEU został opracowany w 2002 roku przez Kishore Papineni i współpracowników z IBM Research. Praca naukowa "BLEU: a Method for Automatic Evaluation of Machine Translation" opublikowana na konferencji ACL 2002 jest jedna z najczęściej cytowanych publikacji
w historii NLP. Jej głównym celem było stworzenie automatycznej metryki, która korelowałaby z ocena ludzka w zadaniu tłumaczenia maszynowego.

Nazwa "Bilingual Evaluation Understudy" jest celowym nawiązaniem do ROUGE (patrz sekcja 3) i oznacza dosłownie "dwujęzyczny zastępnik ewaluacji" — podkreślając, ze metryka ma zastępować kosztowna ocenę ludzką.

2.2. Na czym polega BLEU?

BLEU porównuje tekst wygenerowany przez model (ang. hypothesis) z jednym lub wieloma tekstami referencyjnymi (ang. references) dostarczonymi przez człowieka. Podstawowym budulcem metryki jest pojęcie n-gramu.

> **Czym jest n-gram?**
> N-gram to kolejne n słów (tokenów) z tekstu. Unigram (1-gram) to pojedyncze słowo, bigram (2-gram) to para kolejnych słów, trigram (3-gram) to trojka słów itd. Przykład: w zdaniu "Ala ma kota" unigram to {Ala, ma, kota}, bigram to {Ala ma, ma kota}.

BLEU mierzy precyzje n-gramów: jaki ułamek n-gramów z tekstu wygenerowanego pojawia się również w tekstu referencyjnym. Standardowo oblicza się precyzje dla n = 1, 2, 3, 4, a następnie łączy się je średnia geometryczna. Kluczowym elementem jest tez kara za zbyt krótkie teksty (ang. Brevity Penalty, BP), która zapobiega oszukiwaniu metryki przez generowanie bardzo krótkich odpowiedzi.

```
BLEU = BP * exp(sum_n w_n * log P_n)
```

gdzie: BP to kara za krótkość, P_n to precyzja n-gramów, a w_n to wagi dla każdego rzędu n-gramów (standardowo równe 1/N).

2.3. Interpretacja wyniku

Wynik BLEU przyjmuje wartości od 0 do 1 (lub jest wyrażona w procentach: od 0 do 100). Wyższa wartość oznacza większe podobieństwo do tekstu referencyjnego. W tłumaczeniu maszynowym uznaje się w przybliżeniu następujące progi:

- **< 10:** Bardzo niska jakość, tekst praktycznie niezrozumiały.
- **10-19:** Tłumaczenie trudne w odbiorze, lecz widoczne są pewne sensowne elementy.
- **20-29:** Tłumaczenie zrozumiale, lecz z bledami.

- **30-40:** Dobra jakość, tekst zrozumiały.
- **> 40:** Wysoka jakość, bliska jakości tłumaczenia ludzkiego.

2.4. Zalety i ograniczenia

Zalety	Ograniczenia
• Prosty w obliczeniach i interpretacji • Powszechnie stosowany — łatwy do porównań • Językoznawczo niezależny • Szybki do obliczenia	• Nie uwzględnia synonimów ani parafraz • Niska korelacja z ocena ludzka w niektórych zadaniach • Nie ocenia spójności i sensowności semantycznej • Wrażliwy na tokenizację

2.5. Typowe zastosowania

- Tłumaczenie maszynowe (Machine Translation)
- Automatyczne podsumowanie tekstu
- Jako punkt bazowy w porównaniach z nowszymi metrykami

3. ROUGE — Recall-Oriented Understudy for Gisting Evaluation

3.1. Geneza i historia

ROUGE został zaproponowany przez Chin-Yew Lin z Information Sciences Institute (USC) w 2004 roku. Metryka powstała jako odpowiedz na potrzebę oceny systemów automatycznego streszczania (ang. summarization), gdzie BLEU nie sprawdzał się dobrze ze względu na swoja orientacje na precyzje.

Nazwa metryki jest gra slow — "rouge" (fr. czerwony) to również nawiązanie do nazwy BLEU (fr. niebieski), co autor podkreślił w swojej pracy. ROUGE jest dzisiaj de facto standardem w ewaluacji systemów podsumowujących.

3.2. Kluczowa różnica: Recall vs Precision

Fundamentalna różnica miedzy BLEU a ROUGE polega na tym, co jest mierzone:

> **Precyzja (Precision) vs Recall**
> Precyzja odpowiada na pytanie: 'Jaka część tego, co wygenerowaliśmy, jest poprawna?' Recall odpowiada na pytanie: 'Jaka cześć tego, co powinno znaleźć się w odpowiedzi, faktycznie się tam znalazła?' W podsumowaniu ważniejszy jest recall — chcemy mieć pewność, ze wszystkie kluczowe informacje z referencji zostały uchwycone.

3.3. Warianty ROUGE

ROUGE występuje w kilku wariantach, z których każdy akcentuje inny aspekt jakości tekstu:

ROUGE-1

Mierzy pokrycie unogramów (pojedynczych słów) miedzy tekstem generowanym a referencyjnym. Wychwytuje zawartość leksykalna — czy model użył tych samych słów co referencja.

```
ROUGE-1 = |overlap_unigramow| / |unigramy_w_referencji|
```

ROUGE-2

Mierzy pokrycie bigramow (par kolejnych słów). Jest bardziej wymagający niż ROUGE-1, gdyż sprawdza nie tylko, czy model używa tych samych słów, ale czy używa ich w tej samej kolejności.

```
ROUGE-2 = |overlap_bigramow| / |bigramy_w_referencji|
```

ROUGE-L

Opiera się na pojęciu Najdłuższego Wspólnego Podciągu (ang. Longest Common Subsequence, LCS). LCS to najdłuższy ciąg elementów, który pojawia się w obu tekstach

(niekoniecznie sąsiadujących). ROUGE-L jest bardziej elastyczny niż ROUGE-1 i ROUGE-2, gdyż nie wymaga, aby wspólne elementy występowały bezpośrednio obok siebie.

3.4. Zastosowania i kontekst

ROUGE jest standardem w ewaluacji systemów streszczania tekstu (podsumowanie). Benchmarki takie jak CNN/DailyMail czy XSum regularnie raportują wyniki w postaci trojki ROUGE-1 / ROUGE-2 / ROUGE-L. Warto pamiętać, ze mimo powszechnego stosowania, ROUGE — podobnie jak BLEU — nie jest doskonały: nie uwzględnia synonimów ani semantycznego sensu tekstu.

- Automatyczne podsumowanie (ekstrakcyjna i abstrakcyjna)
- Ewaluacja systemów question answering
- Ocena systemów dialogowych

4. METEOR — Metric for Evaluation of Translation with Explicit ORdering

4.1. Geneza i historia

METEOR został opracowany w 2005 roku przez Satanjeeva Banerjeego i Alon Lavriego na Carnegie Mellon University. Motywacja do stworzenia metryki było udokumentowane słabe dopasowanie BLEU do ocen ludzkich w niektórych językach i zadaniach. METEOR był wielokrotnie rozbudowywany — jego nowsze wersje włączają moduł dopasowania synonimów oparty na WordNecie oraz wsparcie dla wielu języków.

4.2. Co odróżnia METEOR od BLEU i ROUGE?

METEOR wprowadza kilka istotnych innowacji względem swoich poprzedników:

- **Stemming:** METEOR normalizuje słowa do ich rdzeni (np. 'biegnie', 'biegnąć', 'bieg' traktowane są jako ta sama forma). Dzięki temu odmiana gramatyczna nie powoduje karania modelu.
- **Synonimy:** METEOR korzysta ze słowników synonimów (np. WordNet dla języka angielskiego), aby rozpoznać, ze słowa o tym samym znaczeniu są równoważne. BLEU i ROUGE tego nie robią.
- **Kolejność słów (penalty za permutacje):** METEOR wprowadza kare za przestawienia słów — im bardziej kolejność słów w generowanym tekście odbiega od referencji, tym wynik jest niższy. Dzięki temu metryka bardziej faworyzuje teksty o naturalnym szyku zdania.
- **Harmoniczna średnia precision i recall:** METEOR łączy precyzje i recall w sposób, który mocniej premiuje recall.

4.3. Wzor i interpretacja

```
METEOR = F_mean * (1 - Penalty)
```

gdzie F_mean to ważona średnia harmoniczna precyzji i recallu (z większa waga dla recallu), a Penalty to kara za nieciągłość (permutacje) dopasowanych elementów.

Ważne

METEOR lepiej koreluje z ludzka ocena tłumaczenia niż BLEU, co potwierdzają liczne badania. Niestety ma on pewne ograniczenia praktyczne: jego dostępność dla języków innych niż angielski jest ograniczona, a obliczenia są bardziej złożone niż w przypadku BLEU.

4.4. Zastosowania

- Tłumaczenie maszynowe — szczególnie jako uzupełnienie BLEU
- Zadania generowania języka naturalnego

- Ewaluacja systemów streszczających (jako alternatywa dla ROUGE)

5. BERTScore — semantyczna metryka oparta na osadzeniach

5.1. Geneza i historia

BERTScore został zaproponowany w 2019 roku przez Tianyi Zhang i współpracowników (Cornell University i Microsoft Research) w pracy "BERTScore: Evaluating Text Generation with BERT". Pojawienie się tej metryki było bezpośrednio związane z rewolucja, jaka model BERT (Bidirectional Encoder Representations from Transformers, Google, 2018) wywołał w przetwarzaniu języka naturalnego.

BERTScore reprezentuje zasadnicza zmianę paradygmatu: zamiast porównywać teksty na poziomie leksykalnym (dokładne dopasowanie słów czy n-gramow), ocenia ich podobieństwo semantyczne — czyli to, czy teksty maja zbliżone znaczenie, nawet jeśli używają innych słów.

5.2. Czym sa embeddingi i dlaczego są ważne?

Embeddingi (osadzenia wektorowe)

Embedding to numeryczna reprezentacja słowa (lub całego zdania) w postaci wektora liczb rzeczywistych w przestrzeni wielowymiarowej. Modele takie jak BERT uczą się tworzyć takie reprezentacje w taki sposób, ze słowa o podobnym znaczeniu maja zbliżone wektory. Np. 'pies' i 'kundelek' będą miały wektory blisko siebie, a 'pies' i 'samochód' — daleko od siebie. Semantyczne podobieństwo tekstów można więc mierzyć jako odległość ich embeddingowych reprezentacji.

5.3. Jak dziala BERTScore?

BERTScore oblicza podobieństwo na poziomie tokenów (słów) miedzy tekstem generowanym a tekstem referencyjnym. Algorytm działa w następujących krokach:

1. Każde słowo w obu tekstach jest przetwarzane przez BERT, który generuje dla niego wektor embeddingowy (z uwzględnieniem kontekstu — to samo słowo w różnych kontekstach dostanie różny embedding).
2. Dla każdego tokena w tekście generowanym znajduje się token z referencji, do którego jest on najbardziej podobny (mierząc podobieństwo cosinusowe wektorów).
3. Na tej podstawie oblicza się semantyczna precyzje i recall, a następnie łączy w F-score.

```
BERTScore_F1 = 2 * (BERTScore_Precision * BERTScore_Recall) /
               (BERTScore_Precision + BERTScore_Recall)
```

5.4. Odporność na parafrazowanie

Kluczowa zaleta BERTScore jest odporność na parafrazowanie. Przykład:

Referencja	Hipoteza	Komentarz
The cat sat on the mat.	*A feline rested upon the rug.*	BERTScore: wysoki (synonymy). BLEU: niski (brak n-gram overlap).

5.5. Zalety i ograniczenia BERTScore

- **Zaleta:** Odporność na synonimy i parafrazy — znacznie lepiej niż BLEU i ROUGE.
- **Zaleta:** Wysoka korelacja z ocena ludzka, szczególnie dla zróżnicowanych prac generatywnych.
- **Zaleta:** Wielojęzyczność — dostępne modele BERT dla wielu języków.
- **Ograniczenie:** Wymaga zasobów obliczeniowych — uruchomienie modelu BERT jest znacznie kosztowniejsze niż obliczenie BLEU.
- **Ograniczenie:** Wyniki są trudniejsze do interpretacji — nie jest oczywiste, co oznacza konkretna wartość BERTScore.
- **Ograniczenie:** Jakość metryki zależna od jakości używanego modelu BERT.

6. Perplexity — miara niepewności modelu językowego

6.1. Czym jest perplexity?

Perplexity (pol. perplesja lub miara dezorientacji) to metryka wywodząca się z teorii informacji, stosowana do oceny modeli probabilistycznych — w szczególności modeli językowych.
W odróżnieniu od poprzednich metryk, perplexity NIE porównuje tekstu modelu z tekstem referencyjnym. Mierzy natomiast, jak dobrze model "rozumie" (tj. prognozuje) zadany tekst.

> **Intuicja za perplexity**
> Wyobraź sobie, ze model językowy to system zgadujący. Perplexity mówi nam, ile mniej więcej równo prawdopodobnych opcji (kolejnych słów) model bierze pod uwagę w każdym kroku generowania. Im niższa wartość, tym model jest bardziej pewny i lepiej skalibrowany. Perplexity = 1 oznaczałoby model doskonały, który zawsze idealnie przewiduje następne słowo.

6.2. Definicja matematyczna

Matematycznie, perplexity jest eksponentem entropii krzyżowej (cross-entropy) modelu na danym zbiorze testowym. Dla sekwencji słów w_1, w_2, …, w_N:

```
PP(W) = P(w1, w2, ..., wN)^(-1/N)
```

Gdzie P(w1, w2, ..., wN) to łączne prawdopodobieństwo sekwencji według modelu. Równoważna definicja przez entropie krzyżowa:

```
PP(W) = exp(H(W)) = exp(-1/N * sum log P(wi | w1...w(i-1)))
```

W praktyce im model jest lepiej wytrenowany na zbiorze danych zbliżonymi do tekstu testowego, tym niższe perplexity uzyska.

6.3. Interpretacja

Perplexity nie ma jednej ustalonej skali interpretacji — wszystko zależy od zadania i dziedziny:

- **Niskie perplexity:** Model jest pewny i dobrze prognozuje tekst. Typowo: modele GPT na tekstach podobnych do danych treningowych osiągają PP < 20-30.
- **Wysokie perplexity:** Model jest niepewny. Może oznaczać, ze tekst testowy pochodzi z innej domeny niż dane treningowe, lub ze model jest słabej jakości.
- **Porównanie modeli:** Perplexity ma sens przede wszystkim przy porównywaniu modeli na tym samym zbiorze testowym. Bezwzględna wartość perplexity nie jest przenośna miedzy różnymi zadaniami.

6.4. Zastosowania i ograniczenia

Zastosowania

- Ewaluacja modeli językowych (language models) podczas trenowania
- Porównanie modeli na tym samym zbiorze testowym
- Ocena kalibracji modelu — jak bardzo jest pewny swoich prognoz
- Wykrywanie danych spoza domeny (out-of-domain detection)

Kluczowe ograniczenia

- **Nie mierzy jakości faktycznej odpowiedzi:** Model może mieć niskie perplexity, a mimo to generować błędne lub nieużyteczne teksty.
- **Nie jest porównywalny miedzy różnymi modelami z różnymi tokenizatorami:** Jeśli dwa modele używają innych schematów tokenizacji, ich perplexity nie są wprost porównywalnie.
- **Przeadaptowanie (overfitting):** Model może mieć niskie perplexity na danych treningowych, ale wysokie na nowych danych — perplexity mierzy się więc zawsze na zbiorze testowym.

7. Tabela porównawcza metryk

Poniżej zestawiono omawiane metryki w formie zbiorczej tabeli. Pozwala ona na szybkie porównanie kluczowych właściwości każdej z miar:

Metryka	Podstawa	Typ pomiaru	Parafra-zowanie	Glowne zastosowanie	Kore-lacja z ocena ludzka
BLEU	n-gramy	Precyzja	Niska	Tłumaczenie	Średnia
ROUGE	n-gramy	Recall	Niska	Podsumowanie	Średnia
METEOR	n-gramy+	F-score	Średnia	Tłumaczenie	Wysoka
BERTScore	Embeddingi	Podobieństwo	Wysoka	NLG ogólnie	Wysoka
Perplexity	Prawdop.	Pewność modelu	N/D	Język, LM	Pośrednia

7.1. Kiedy używać której metryki?

Wybór metryki powinien być zawsze podyktowany specyfika zadania. Poniżej kilka praktycznych wskazówek:

- **Tłumaczenie maszynowe:** Standardem jest BLEU, jednak warto uzupełnić go METEOR (lepiej koreluje z ocena ludzka) oraz BERTScore (odporny na parafrazy).
- **Automatyczne podsumowanie:** ROUGE-1, ROUGE-2 i ROUGE-L są standardem. BERTScore jest dobrym uzupełnieniem, szczególnie dla streszczeń abstrakcyjnych.
- **Modele językowe i pretrenowanie:** Perplexity jest kluczowa miara podczas trenowania. Niska perplexity na zbiorze walidacyjnym to dobry sygnał kalibracji modelu.
- **Ocena ogólna systemów NLG:** Kombinacja BERTScore + ROUGE daje dobre pokrycie. Tam gdzie to możliwe, warto uzupełnić ocena ludzka.

Złota zasada ewaluacji

Żadna pojedyncza metryka nie zastępuje oceny ludzkiej. W profesjonalnych badaniach stosuje sie zazwyczaj kilka metryk automatycznych łącznych z przynajmniej cząstkowymi ocenami ludzkimi, aby uzyskać pełny obraz jakości modelu. Wyniki należy zawsze interpretować w kontekście konkretnego zadania i danych testowych.

8. Podsumowanie

Ewaluacja modeli językowych jest procesem wielowymiarowym, wymagającym stosowania różnych narzędzi w zależności od zadania i celów badania. W niniejszym materiale przedstawiono pięć kluczowych metryk automatycznych:

- **BLEU** — klasyczna metryka oparta na precyzji n-gramow. Prosta, szybka i powszechna w tłumaczeniu maszynowym, lecz nieuwzględniająca znaczenia semantycznego.
- **ROUGE** — rodzina metryk opartych na recallu n-gramow. Standard w automatycznej summaryzacji. Warianty ROUGE-1, ROUGE-2 i ROUGE-L wychwytują różne aspekty pokrycia treści.
- **METEOR** — rozszerzenie koncepcji n-gramowych o stemming, synonimy i kare za permutacje. Wykazuje lepsza korelacje z ocena ludzka niz BLEU.
- **BERTScore** — metryka semantyczna oparta na embeddingach BERT. Odporna na parafrazy i synonimy — mierzy sens, nie tylko dokładne dopasowanie leksykalne.
- **Perplexity** — miara kalibracji modelu językowego. Ocenia, jak pewnie model prognozuje tekst; im niższa, tym lepiej skalibrowany model.

Zrozumienie tych metryk, ich zalet i ograniczeń, jest kluczowe dla rzetelnej oceny systemów AI. Dobre opanowanie tej wiedzy umożliwia świadomy wybór narzędzi ewaluacyjnych i krytyczna interpretacje wyników eksperymentów.