Business needs.

Language Solutions

Translating Europe Workshop 2019

Rodzaje i zastosowanie metryk jakościowych w ocenie pracy silników tłumaczenia maszynowego

Piotr Peszyński
Warszawa, 4 października 2019

Delivered.

Plan prezentacji

Potrzeba określenia jakości/wydajności NMT

Metryki automatyczne

Metryki ręczne

Czas edycji

Oceny subiektywne

Przyszłość post-edycji

Potrzeby rynku

Coraz więcej treści wymaga przetłumaczenia

ale

liczba tłumaczy/lingwistów nie rośnie w tym samym tempie

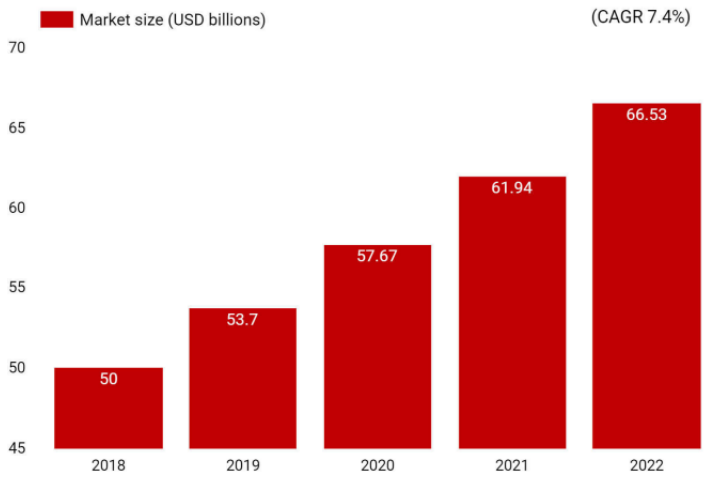
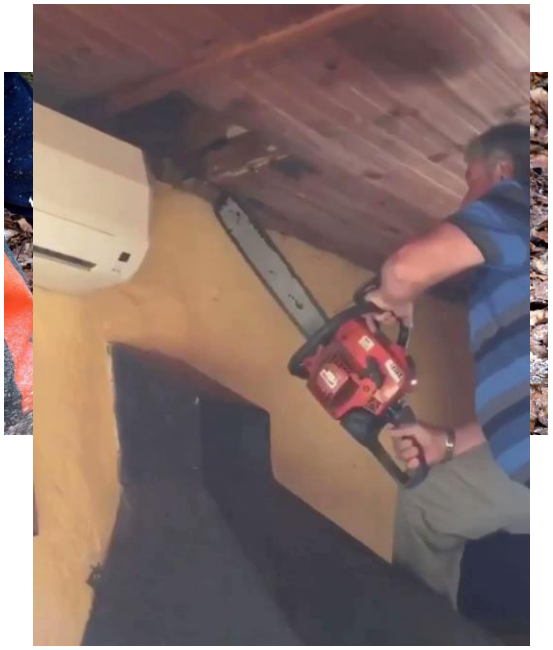


Figure 1: Industry is expected to grow at 7.4% CAGR through 2022 to USD 66.53 billion.

źródło: https://www.nimdzi.com/wp-content/uploads/2018/03/2018-Nimdzi-100-First-Edition.pdf

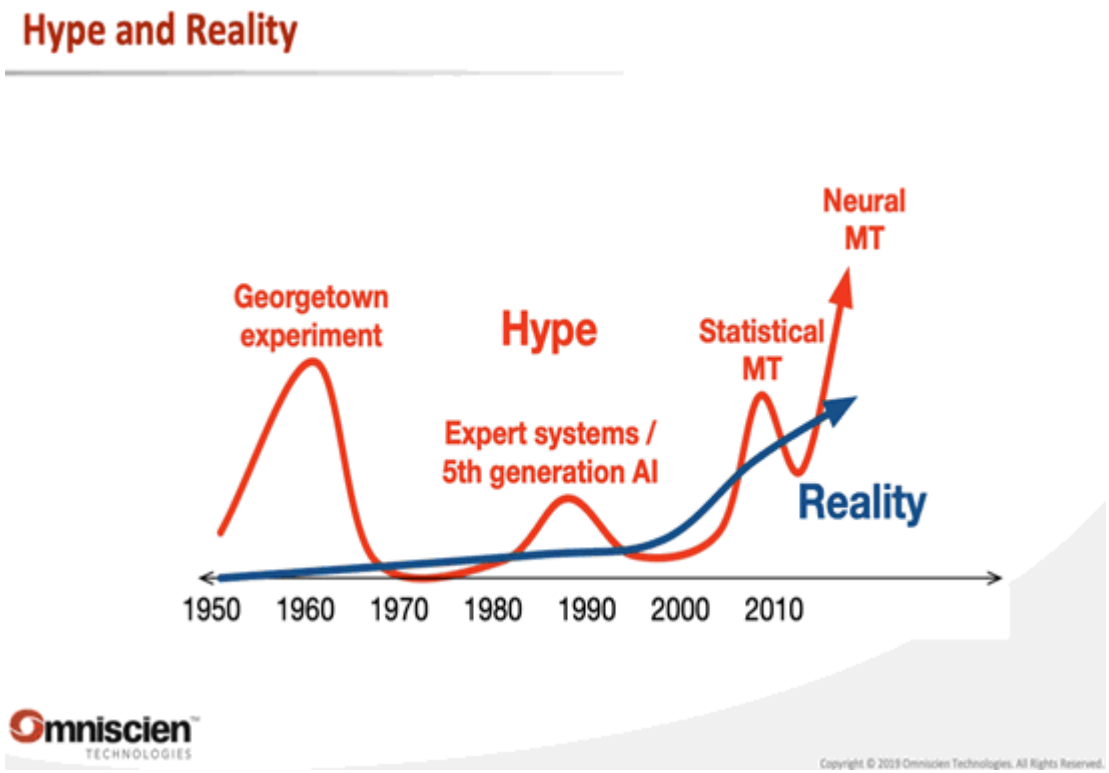
Zwiększanie produktywności

Słowniki
Edytory tekstu



MT?

Rzeczywistość MT



Source: Advances in MT: What Is Exciting and Shows Promise Ahead? Dion Wiggins & Philipp Koehn Omniscien Technologies

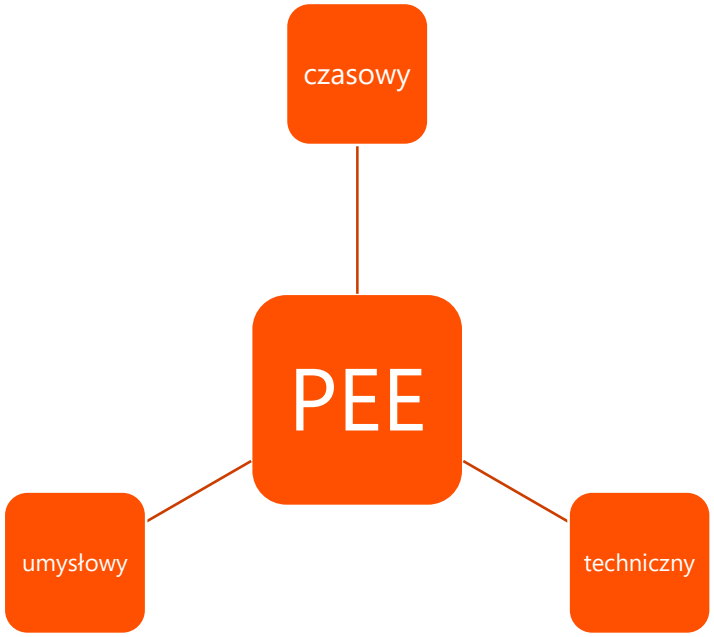
A co myślą tłumacze?

https://willrobotstakemyjob.com/

Jakie wyzwania przynosi NMT

- Często wystarczająca jakość dla zastosowań „masowych" – gdzie jakość, ani 100% wierność nie jest istotna.
- Bardzo obiecujące jako wsparcie w zastosowaniach profesjonalnych, ale:
  - płynne tłumaczenie utrudnia wykrywanie błędów
  - nieprzewidywalna jakość na poziomie segmentu
  - zdolności słowotwórcze
  - kontekst na poziomie zdania/segmentu
  - problemy ze spójnością terminologiczną
  - lepsze rezultaty dla podobnych języków i języków z ubogą fleksją
  - kluczowe znaczenie odpowiedniego doboru dokumentu do silnika oraz przygotowanie formalne (pre-processing)
  - silniki neuronowe nie są identyczne

Jakość a użyteczność NMT



- Miary automatyczne
- Ocena ręczna
- Czas (słowo/godz.)
- Ocena subiektywna

Metryki automatyczne

- Oparte na analizie porównawczej hipotezy silnika MT i tłumaczenia referencyjnego (optymalnie kilku wersji)
- Badają zgodność zbitek słów (n-gramów) lub liczbę zmian (wstawienia, usunięcia, przeniesienia)
- Podstawowe narzędzia w pracach badawczych nad rozwojem silników
- Dają wygodny wynik liczbowy

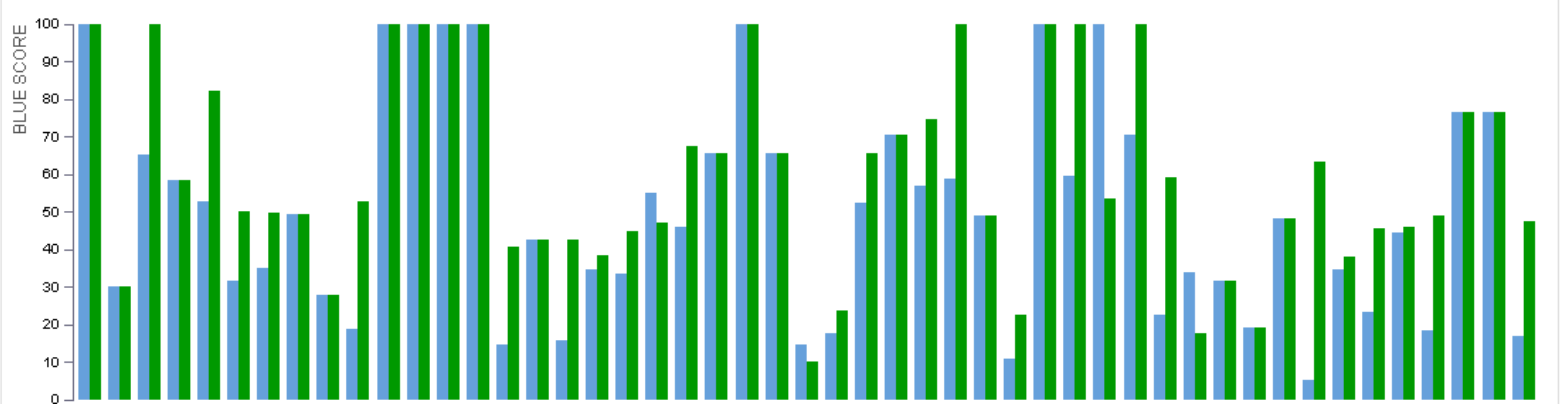
BLEU	TER
NIST	WER
METEOR	Edit-distance
LEPOR	
F-Measure	

BLEU (Bilingual Evaluation Understudy)

- Najpopularniejsza metryka
- Mocno krytykowana
- Niezależna od języka
- Polega na zliczeniu „pasujących" zestawów słów (n-gramów).
- Wykorzystuje referencyjny zestaw testowy, który:
  - powinien być wysokiej jakości
  - zawierać alternatywne tłumaczenia (optymalnie)
  - powinien być z tej samej dziedziny jak korpus szkoleniowy
  - nie powinien zawierać istotnej liczby liczb, dat itp.
  - obejmować min. 1000 segmentów
  - nie może być częścią korpusu szkoleniowego

Przykłady BLEU

BLEU:	43.70				44.05			
Precision x brevity:	45.46 x 96.13				44.81 x 98.31			
Type	1-gram	2-gram	3-gram	4-gram	1-gram	2-gram	3-gram	4-gram
Individual	69.10	49.89	39.08	31.72	69.24	50.00	38.47	30.26
Cumulative	66.42	56.44	49.28	43.70	68.07	57.84	50.20	44.05



METEOR (Metric for Evaluation of Translation with Explicit ORdering)

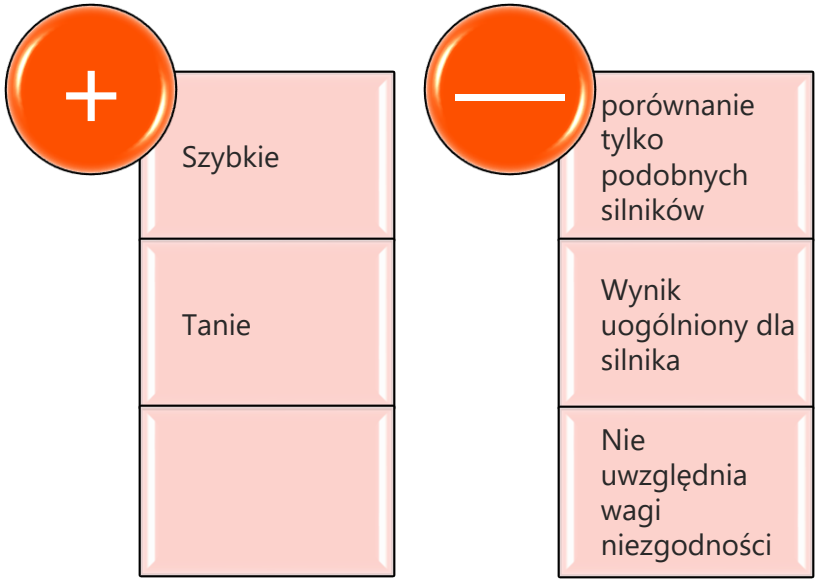
- Uważana za lepiej skorelowaną z osądem ludzkim
- Wykonuje mapowanie uni-gramów, a następnie grupuje uni-gramy
- Wymaga użycia zasobów językowych, a więc i większych nakładów niż np. metryka BLEU
- Uwzględnia dodatkowe cechy językowe, takie jak:
  - synonimy
  - formy słownikowe
  - rdzenie słów

TER (Translation Error Rate)

- Zlicza liczbę zmian potrzebnych do ujednolicenia hipotezy MT z tłumaczeniem referencyjnym względem sumarycznej liczby słów
- W przeciwieństwie do BLEU, czy METEOR im mniejsza wartość tym lepiej
- Lepiej obrazuje aspekt „techniczny” PEE, ale nie uwzględnia aspektu czasowego

Miary automatyczne, wnioski

- Wyniki bardzo niejednoznaczne w kontekście przyspieszenia pracy tłumacza
- Oceniają bliskość hipotezy do tłumaczenia referencyjnego



Metryki manualne

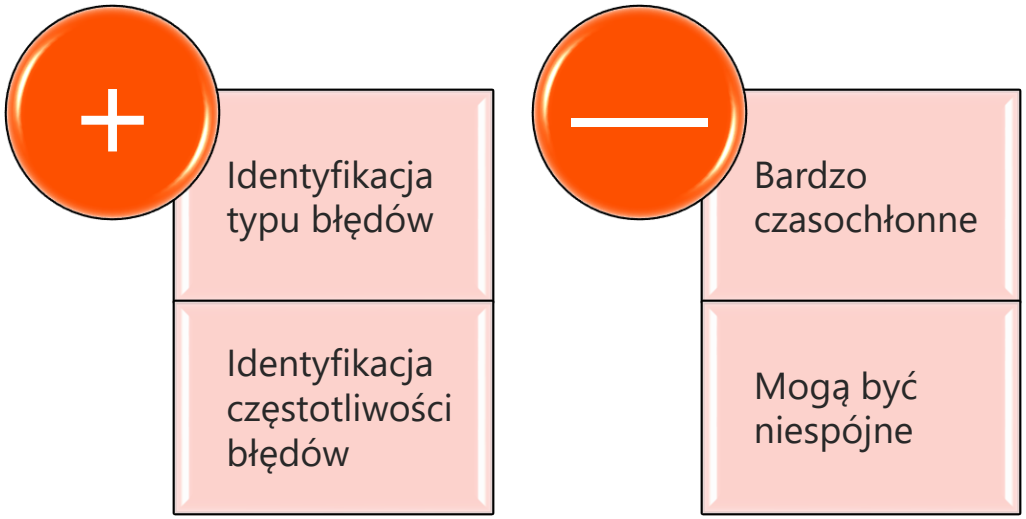
- Modele LQA (Language Quality Model)
- Mają zastosowanie do tłumaczeń ludzkich i maszynowych
- Dają bardzo szczegółowy obraz jakości tłumaczenia z rozbiciem na konkretne kategorie błędów

MQM – Multidimensional Quality Metric

- Bardzo rozbudowany katalog potencjalnych błędów
- Przewiduje ocenę również jakości tekstu źródłowego
- Faktycznie użyte kategorie błędów można dostosować do potrzeb

Metryki ręczne, wnioski

- Bardzo przydatne ze względu na możliwości szkoleniowe
- Podobnie jak metryki automatyczne nie umożliwiają jednoznacznego wytyczenia stopnia usprawnienia pracy
- Trwają prace nad rozwiązaniami umożliwiającymi automatyczną klasyfikację błędów



Pomiar czasu (słowo/godzinę)

- Pozornie bardzo bezpośrednie wyrażenie użyteczności MT
- Uwzględniają czas potrzebny na analizę hipotezy MT
- Niestety, nacechowane bardzo dużą zmiennością:
  - doświadczenie lingwisty
  - znajomość narzędzia
  -

Subject	Text	Seq.	MT	Words/h	Quality
FR-1	A	1	No	520.37	4.00
FR-1	B	2	No	630.82	5.50
FR-1	C	3	Yes	909.88	5.00
FR-1	D	4	Yes	602.56	5.00
FR-2	A	1	Yes	987.00	4.50
FR-2	B	2	Yes	1237.13	3.50
FR-2	C	3	No	682.64	4.00
FR-2	D	4	No	505.40	4.50
Average TM-ONLY			No	584.81	4.50
Average POST-EDIT			Yes	934.14	4.50
Difference (%)				59.74	0.00

(a) DE–FR

Subject	Text	Seq.	MT	Words/h	Quality
IT-1	A	1	No	389.41	4.00
IT-1	B	2	No	398.71	4.00
IT-1	C	3	Yes	647.87	4.50
IT-1	D	4	Yes	393.14	4.00
IT-2	A	1	Yes	401.19	5.50
IT-2	B	2	Yes	536.09	5.50
IT-2	C	3	No	553.00	5.50
IT-2	D	4	No	469.56	5.50
Average TM-ONLY			No	452.67	4.75
Average POST-EDIT			Yes	494.57	4.88
Difference (%)				9.26	0.13

(b) DE–IT

Table 2: Experimental conditions and results: the number of target words produced per hour (Words/h) and averaged overall impression scores (Quality) as assigned by two expert raters per translation.

Language Solutions

Źródło: Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain Samuel Läubli Chantal Amrhein Patrick Düggelin Beatriz Gonzalez Alena Zwahlen Martin Volk

Ocena subiektywna

- Subiektywna ocena stopnia przyspieszenia pracy
- Całościowe podejście, które uwzględnia wszystkie czynniki (w tym potencjalne uprzedzenia, bądź entuzjazm do MT)
- Wnioski wyciągnąć można dopiero po pewnym czasie
- Zindywidualizowane – zespół i zastosowanie

Co to oznacza, dla tłumacza

- Większość tych miar działa na zasadzie uśrednienia ogólnych wniosków na temat silnika i przeniesienie ich na bieżące zlecenia
- Najbardziej miarodajne wyniki w warunkach produkcyjnych można uzyskać dopiero po pewnym czasie
- Również metody, które pozornie uwzględniają najbliżej wysiłek związany z post-edycją zastosować dopiero po zakończeniu tłumaczenia (edit-distance, czas pracy)

Co na horyzoncie?

- Algorytmy silników NMT są nadal intensywnie rozwijane:
  - Adaptacja NMT
  - Szacowanie jakości (QE)
  - Automatyczny post-editing
  - Dostosowanie terminologii
  - Kontekst na poziomie dokumentu
  - Automatyczny dobór najlepszego silnika na poziomie segmentu
- Wykorzystanie sieci neuronowych w branży tłumaczeniowej nie musi być ograniczone tylko do tłumaczenia maszynowego
  - Automatyczna klasyfikacja dokumentów
  - Automatyczne przypisywanie najlepszych tłumaczy do zleceń
  - Rozpoznawanie mowy
  - Fuzzy uplifting/repair (z wykorzystaniem MT)
  - Paraphrasing TM?

Podsumowując

- MT może przyspieszyć pracę
- Używane metryki niekoniecznie pozwalają ustalić w jakim stopniu
- W praktyce zwiększenie produktywności wyrażone redukcją WWC jest negocjowalne – rozpoznawane bojem
- Nowe rozwiązania obiecują zmiany i w ocenie produktywności i samej praktyce post-edycji
- Warto zacząć myśleć w kategoriach czasu nie jednostek rozliczeniowych
- MT nie jest konkurencją dla tłumacza, tylko narzędziem

Dziękuję za uwagę