POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

Eksploracja i wykorzystanie korpusów porównywalnych w tłumaczeniu maszynowym

dr inż. Krzysztof Wołk
kwolk@pja.edu.pl
Katedra Multimedów

Materiały: http://tiny.cc/eqx2ty

Plan Prezentacji

- PJATK i Badania
- Systemy S2S
- Klasyfikacja, zastosowania i wady korpusów równoległych
- Metody budowy korpusów równoległych
- Poprawa jakości korpusów
- Model współczesnego języka polskiego oraz plany na przyszłość

Zagadnienia badawcze w PJATK

- Inżynieria oprogramowania
- Bazy danych
- Robotyka i systemy wieloagentowe
- **Multimedia**
- Sieci

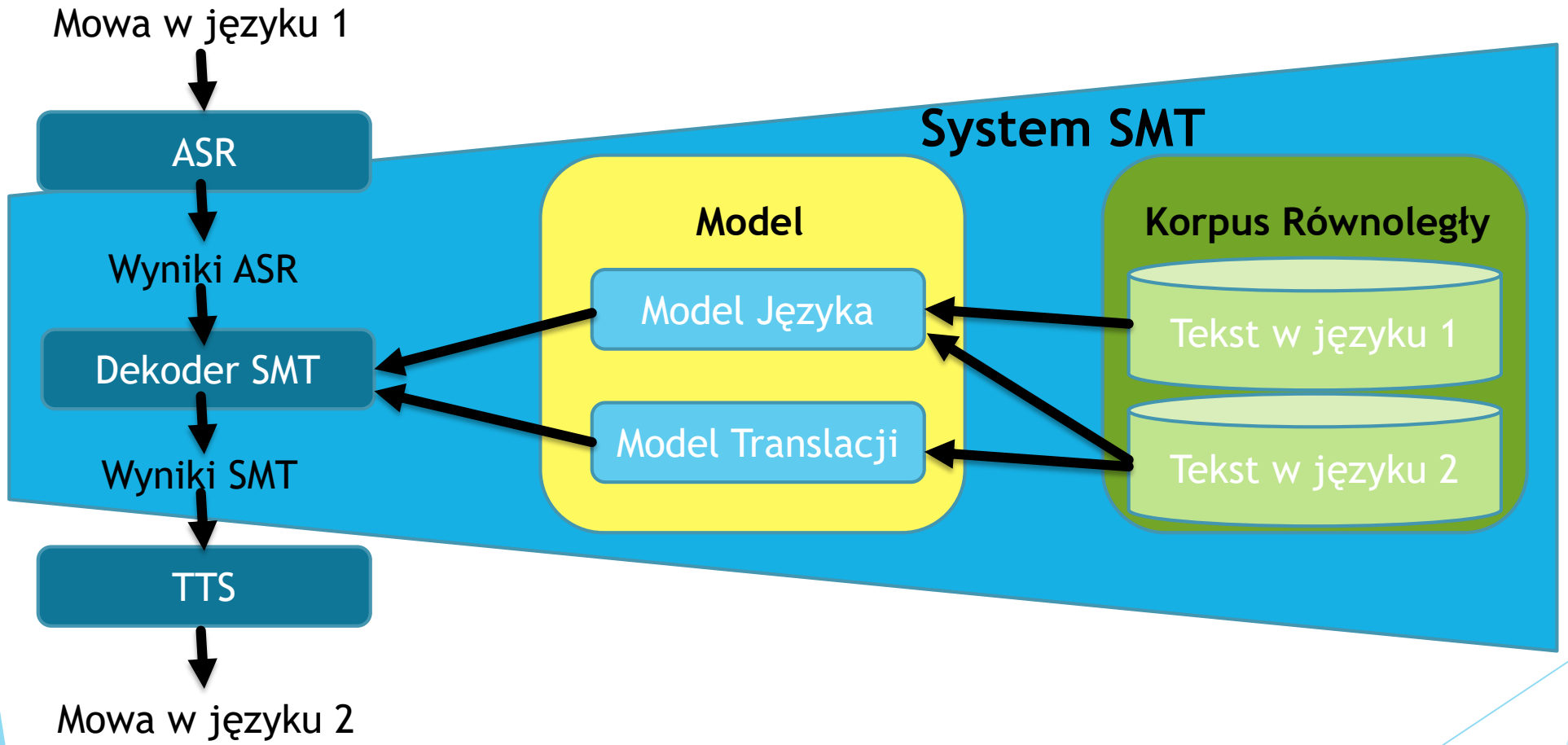
Katedra multimediów

- specjalizacja gier komputerowych
- specjalizacja grafiki 3D
- analiza muzyki (utworów i instrumentów muzycznych)
- analiza obrazów
- **technologie mowy**

Badania związane z mową

- rozpoznawanie mowy
- synteza mowy
- tłumaczenie maszynowe
- => tłumaczenie z mowy na mowę (S2S)
- identyfikacja mówców
- analiza zjawisk akustycznych i para-lingwistycznych

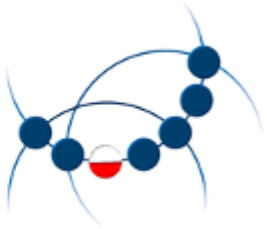
Składniki Systemu SLT

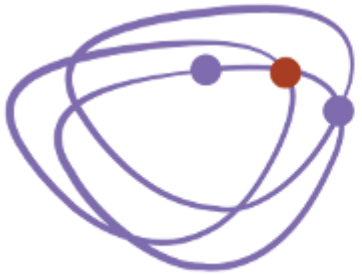


Projekty

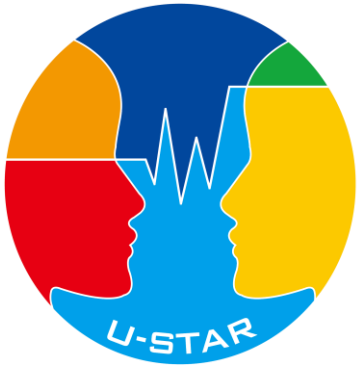
- Luna (FP6; 2006-2009) - system dialogowe
- Senat (MNiSW/NCN) - transkrypcja
- Synat (NCBiR) - transkrypcja RTV i wykładów
- EU-Bridge (FP7; 2012-2015) - tłumaczenie S2S
- CLARIN-PL (MNiSW/ERIC; 2014-2018) - zasoby dla humanistów
- U-Star (nict.go.jp) -tłumaczenie S2S

CLARIN-PL
Common Language Resources and Technology Infrastructure





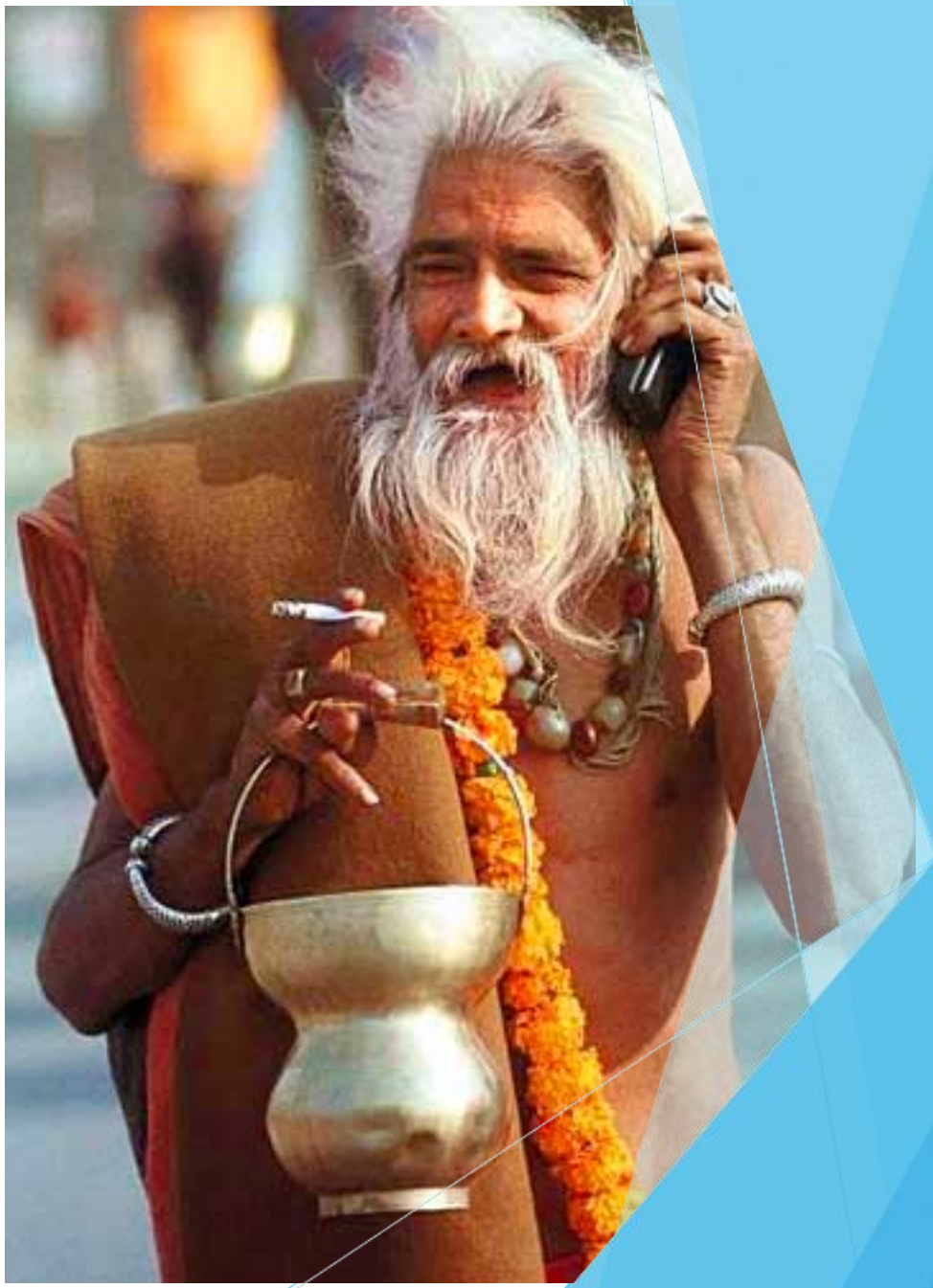
CENTRUM TECHNOLOGII JĘZYKOWYCH CLARIN-PL



VoiceTra4U

Różnorodność językowa

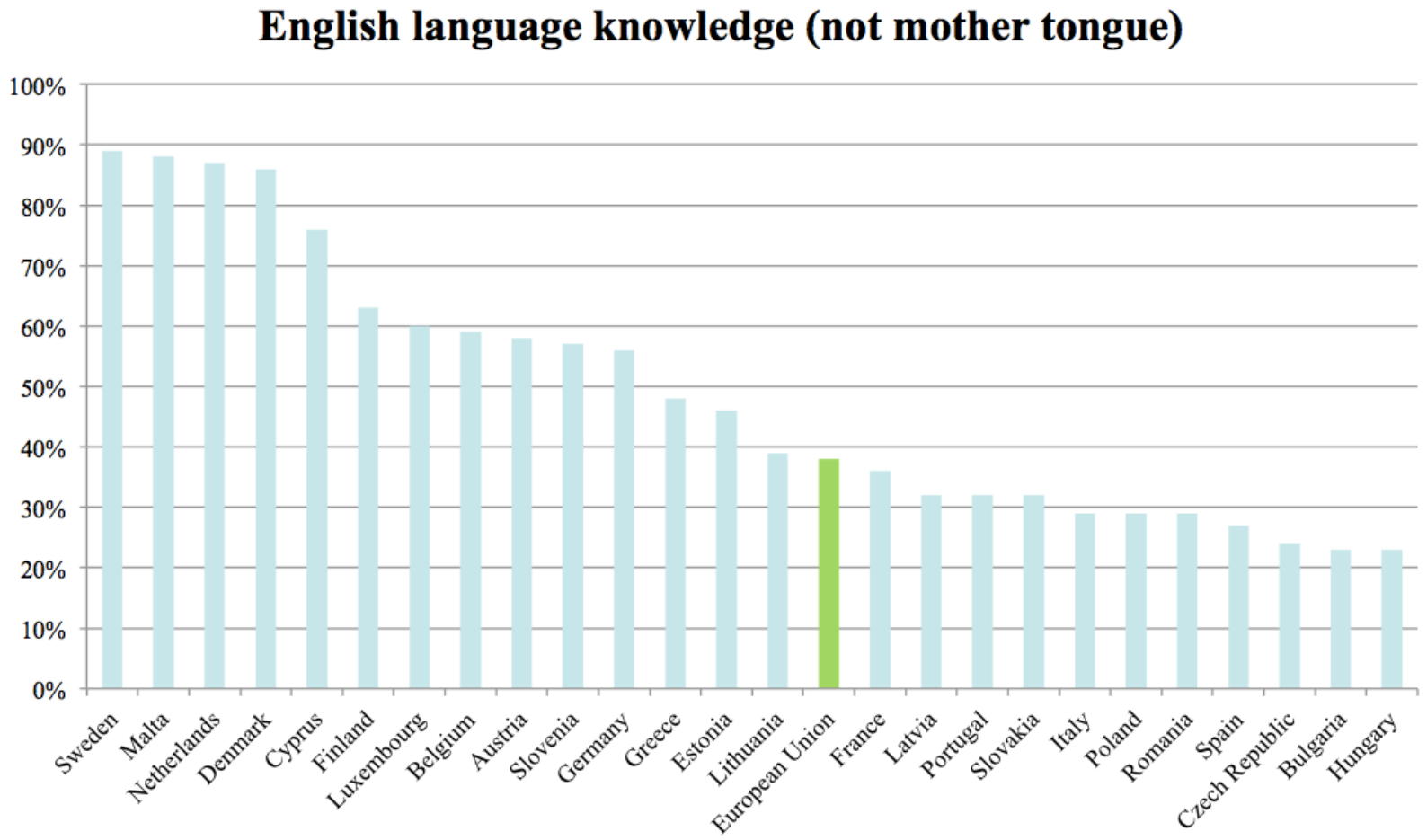
- 7.4 miliarda mieszkańców
- 6 miliardów komórek
- 2 miliardy smartphone’ów
  - 4 miliardy - 2018
  - 90% do 2020
- Tylko 1 miliard PC



Różnorodność językowa

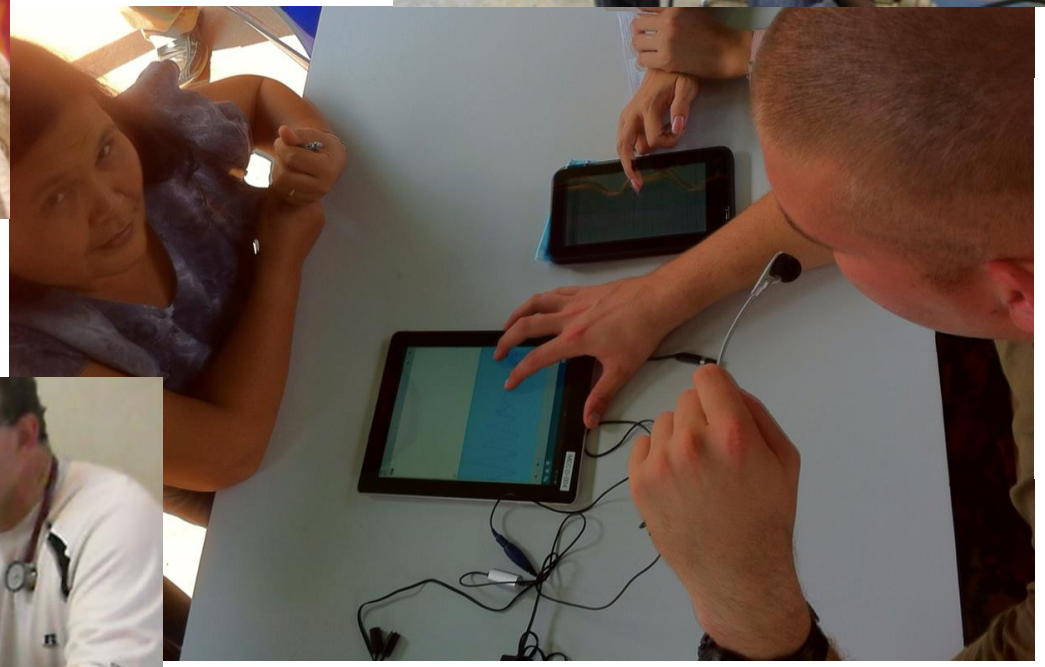
Ponad 6,000 języków na świecie

14.1% Mandaryński
5.85% Hiszpański
5.52% Angielski



https://en.wikipedia.org/wiki/English_language_in_Europe

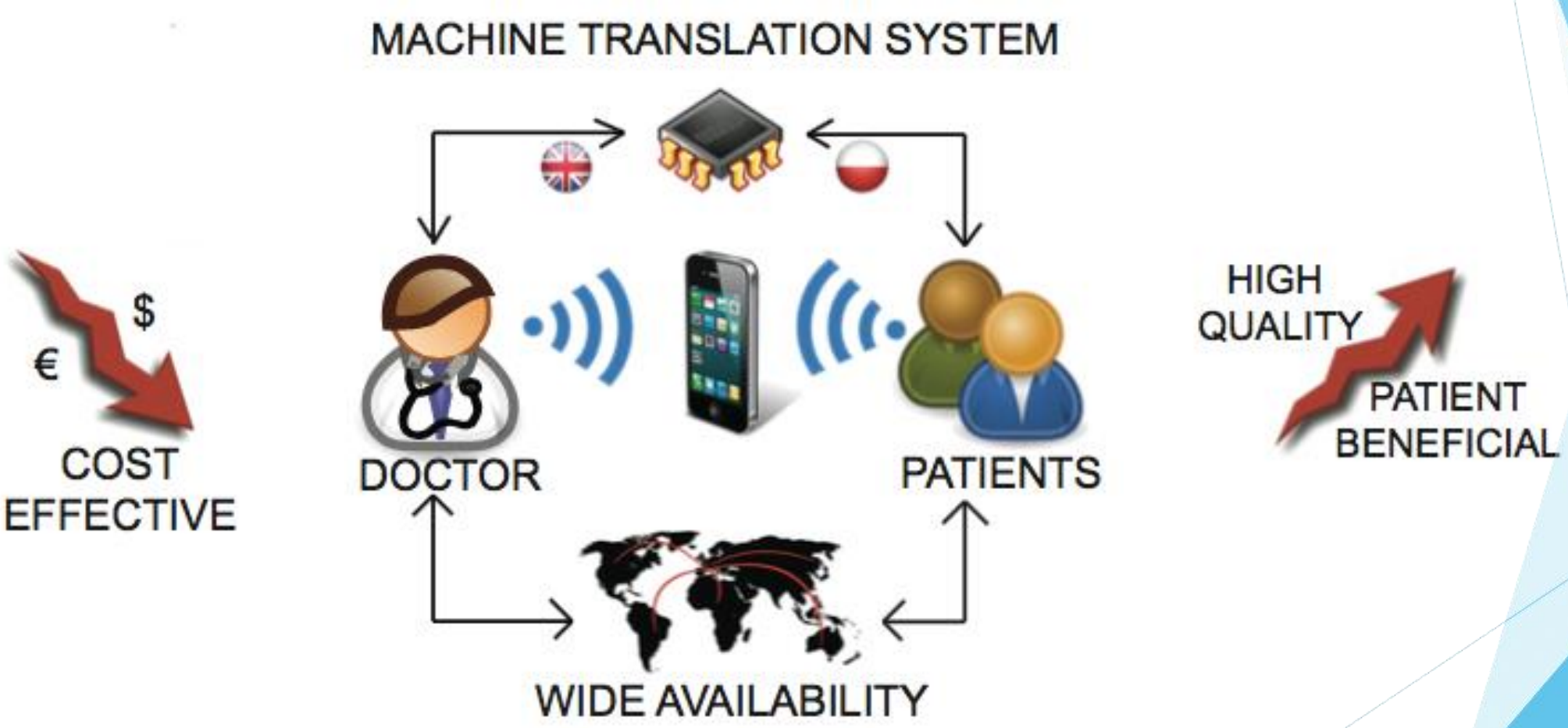
Znaczenie



WARSZAWSKI UNIWERSYTET MEDYCZNY

Polish Telemedicine Society
Polskie Towarzystwo Telemedycyny i e-Zdrowia

Odpowiedź w technologii



Obszar zainteresowań

Google

Microsoft

amazon

~~GENERAL PURPOSE~~

„If the **baby** does not like the **milk**, boil **it**“

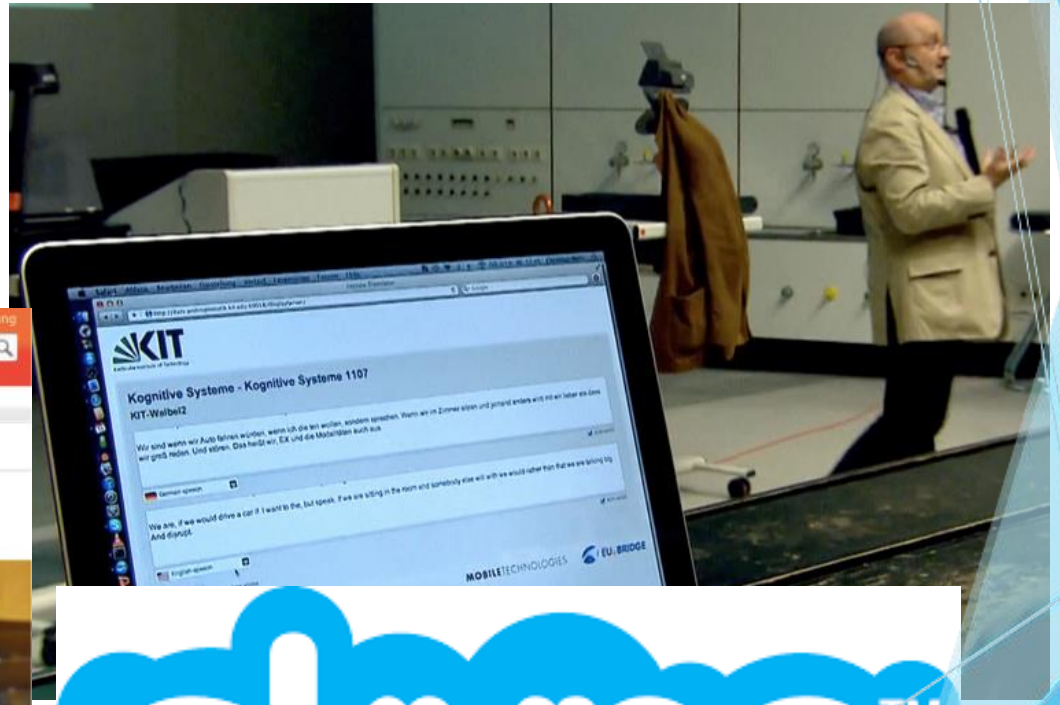
Systemy wewnętrzne

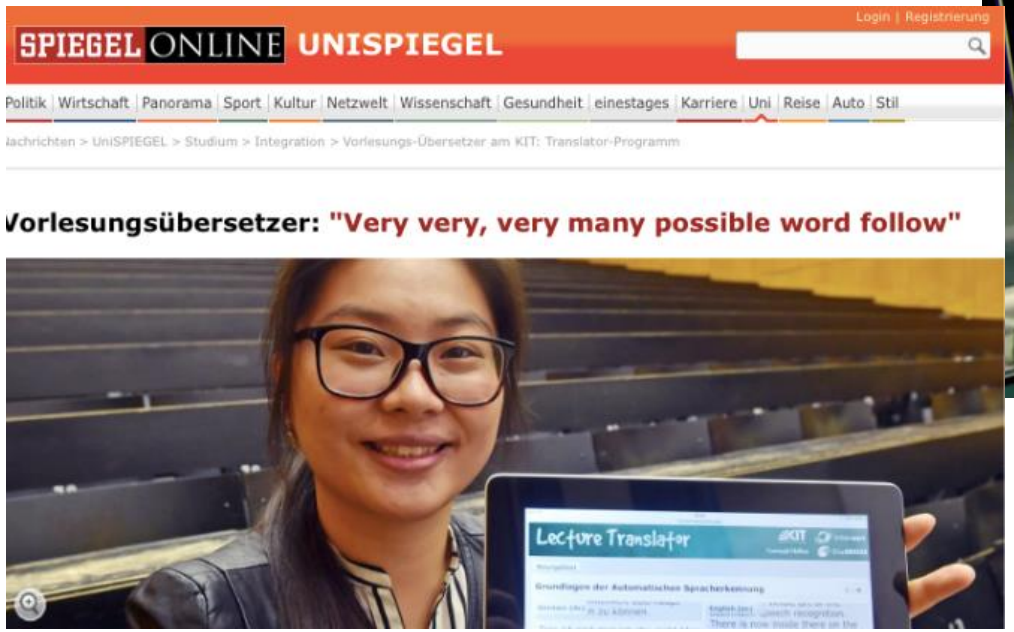
Karlsruhe Institute of Technology

- Tłumaczenie wykładów

EUROPEAN PARLIAMENT

- Obrady





skype™

Mowa VS pisany text

Pisanie:
- Większa precyzja i szczegóły
- Bardziej bezosobowy
- Bardziej formalny
- Bardziej złożone struktury
- Interpunkcja
- 20 słów

Mowa:
- Mniej formalna
- Mniej dokładna
- Zawiera powtórzenia i pauzy
- Przekazuje intencje i ton poza znaczeniem tylko słów
- Pod wpływem dialektów
- 15 słów

RUIZ, Nicholas; FEDERICO, Marcello. Complexity of Spoken Versus Written Language for Machine Translation. In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. 2014. p. 173-180.

Problemy z językiem mówionym

Chat, Posty, SMS, ...

- Slang internetowy: HBD, Happy B-Day, LOL, PM ME, LMK, CU, IMHO,
- Segmentacja: "Post this ❤ ❤ on ❤ ❤ ❤ ❤ anybody's wall ❤ ❤ that ❤ ❤ made ❤ ❤ you ❤ ❤ smile", "A D D M E"
- Normalizacja tekstu: "HBD 2 u", "Luv ya"
- ???: "ya people eih m7d4 mn el member m3ah t7leeeeel 2aly 2"
- Hasztagi: "#PmOrAddMe"
- Błędy w pisowni: "Tank you", "Didnt"
- Emotikony: "... together:)", "=)", "-_-", "♡", "ʘ_ʘ"

•"lol-jah I want hr to be like that.she is cool you knw , inapota. She is going like mother like doghtr."

Rodzaje korpusów wielo-językcznych

- ✓ Równoległe - Parallel Corpora
- ✓ Zaszumione Równoległe - Noisy-Parallel Corpora
- ✓ Porównywalne - Comparable Corpora (e.g. independently edited Wikipedia articles)
- ✓ Prawie-Porównywalne - Quasi-Comparable Corpora

Problemy z Korpusami Równoległymi

- ✓Zbyt mała ilość danych
- ✓Jakość danych
- ✓Słabej jakości tłumaczenia

Zdanie 1: ABCD

ABABCD.

ABCD.ABCD.

Zdanie 2: WXYZ

ABCBCD.

ABWXYZCD.

np. w TED 10% błędy w pisowni, 17% błędy dopasowania, 6% błędne tłumaczenia

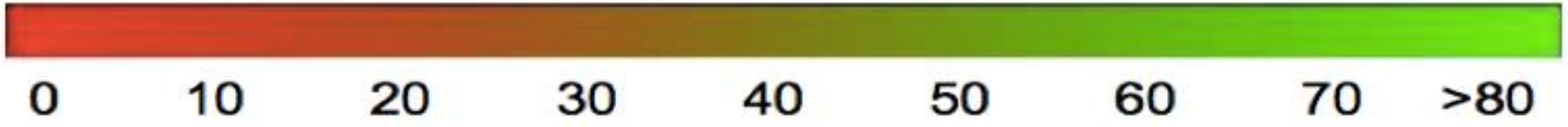
“ WOŁK, Krzysztof; MARASEK, Krzysztof. Polish-English Speech Statistical Machine Translation Systems for the IWSLT 2013.

Zastosowania korpusów równoległych

- ✓ Tłumaczenie maszynowe
- ✓ Leksykologia
- ✓ Ekwiwalencja językowa
- ✓ Badania naukowe lingwistyczne oraz NLP
- ✓ Lingwistyka kontrastywna
- ✓ Pamięci tłumaczeniowe
- ✓ I wiele innych zastosowań

Metody Ewaluacji

- BLEU – Bilingual Evaluation Understudy



- NIST – National Institute of Standards and Technology
- METEOR – Metric for Evaluation of Translation with Explicit Ordering
- TER
- Wilcoxon – test istotności

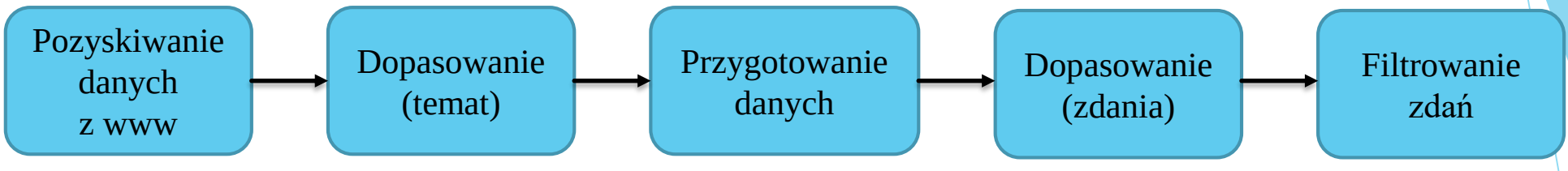
“ LAVIE, Alon. Evaluating the output of machine translation systems. *AMTA Tutorial*. 2010.

Użyte narzędzia

- Moses

- Półautomatyczna korekta
- Adaptacja parametrów treningu systemu SMT

Łańcuch narzędzi



Nr	Hunalign		Filtrowanie		Nr	Hunalign		Filtrowanie	
	TAK	NIE	TAK	NIE		TAK	NIE	TAK	NIE
1	109	130	18	0	11	8	325	0	0
2	6	527	25	2	12	2	256	0	0
3	17	24	0	0	13	70	414	1	0
4	4	302	1	0	14	3	182	0	0
5	6	211	1	0	15	3	285	0	0
6	1	498	0	0	16	111	108	0	0
7	16	440	0	0	17	7	98	0	0
8	1	221	0	0	18	1	202	0	0
9	51	62	0	0	19	21	192	0	0
10	127	245	0	0	20	2	1078	0	0

Wołk K., Marasek K., “A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation.”, Advances in Intelligent Systems and Computing volume 275, p.107-114, Publisher: Springer, ISSN 2194-5357, ISBN 978-3-319-05950-1

Wołk K., Marasek K., „Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs.”, Procedia Technology, 18, Elsevier, p.126-132, 2014

Inne narzędzia dopoasowujące

Experimental Results

Aligner	Score
Proposed Method	98.94
Bleualign	96.89
Hunalign	97.85
ABBYY Aligner	84.00
Wordfast Aligner	81.25
Unitex Aligner	80.65

Experimental Results

Aligner	Lines
Human Translation	1005
Proposed Method	1005
Bleualign	974
Hunalign	982
ABBYY Aligner	866
Wordfast Aligner	843
Unitex Aligner	838

Aligner	**BLEU**	**NIST**	**MET**	**TER**	**% of correctness**
Human Translation	100	15	100	0	100
Proposed Method	98,91	13,81	99,11	1,38	98
Bleualign	91,62	13,84	95,27	9,19	92
Hunalign	93,10	14,10	96,93	6,68	94
ABBYY Aligner	79,48	12,13	90,14	20,01	83
Wordfast Aligner	85,64	12,81	93,31	14,33	88
Unitex Aligner	82,20	12,20	92,72	16,39	86

Wydobywanie przez Analogie

Ilość zdań w korpusie treningowym	**3,800,000**
Ilość odnalezionych analogii	8,128
Ilość wydobytych zdań równoległych	114,107
Zdanie równoległe bez duplikatów	64,080

We współpracy z Emilią Rejmund
erejmund@pjwstk.edu.pl

PL-EN	**BLEU**
TED Bazowy	19.69
Korpus z analogii	16.44
EN-PL	**BLEU**
TED Bazowy	9.97
Korpus z analogii	9.74

Proste analogie np. **A ma się do B jak C do D**.

Wyliczane na podstawie **dystansu Levenshtein'a**
np. dist(A,B)=dist(C,D) i dist(A,C)=dist(B,D)

Wołk K., Rejmund E., Marasek K., „Harvesting comparable corpora and mining them for equivalent bilingual sentences using statistical classification and analogy-based heuristics", Proceedings of 22nd International Symposium ISMIS 2015 Lyon, France, October 21-23,2015, Foundations of Inteligent System, p. 433-441, ISSN 0302-9743,

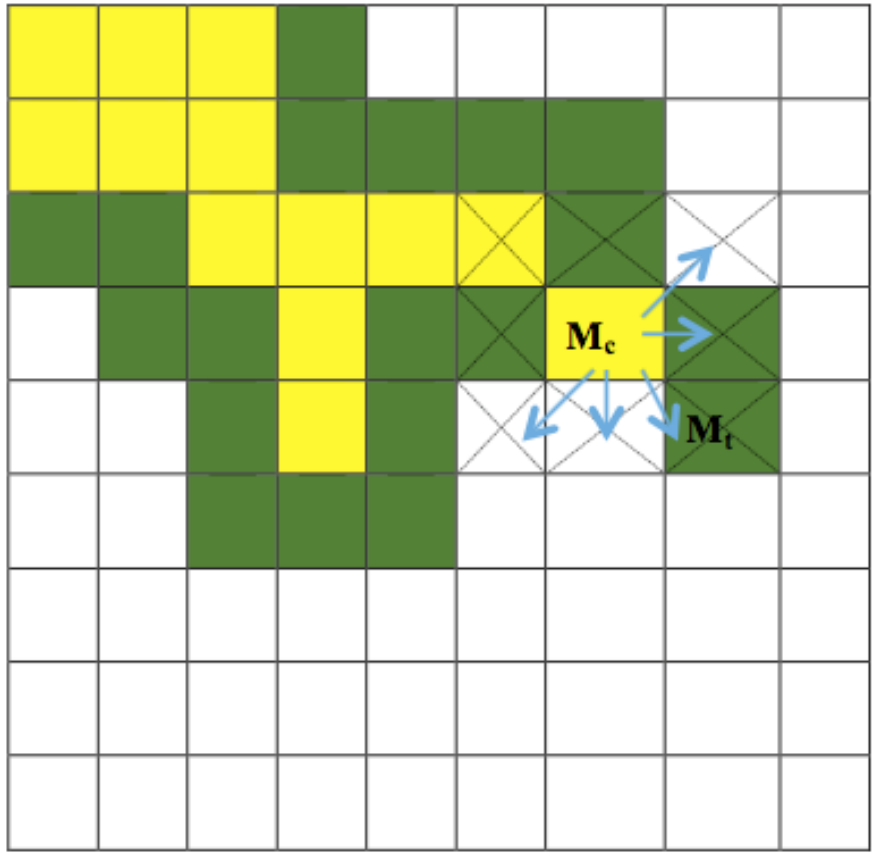
Eksploracja Korpusów Porównywalnych z WIKI

- Yalign
- Reimplementacja mechanizmu Yalign
  - Parser
  - Wielowątkowość
  - Needleman-Wunsch
  - Akceleracja GPU
  - Tuning klasyfikatora
  - Wikipedia - Adaptacje

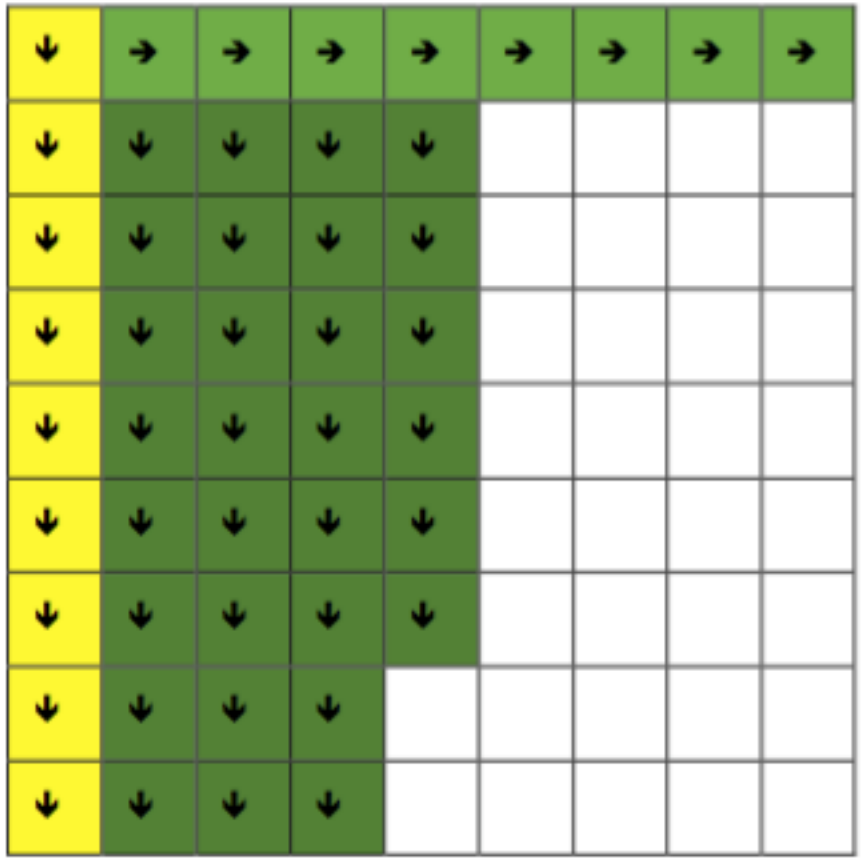
Polski artykuł – średnio 379 słów
Angielski artykuł – średnio 590 słów

Polska Wikipedia - 1,047,423 artykułów
Angielska Wikipedia - 4,524,017 artykułów

Zmiana Algorytmu Dopasowywania



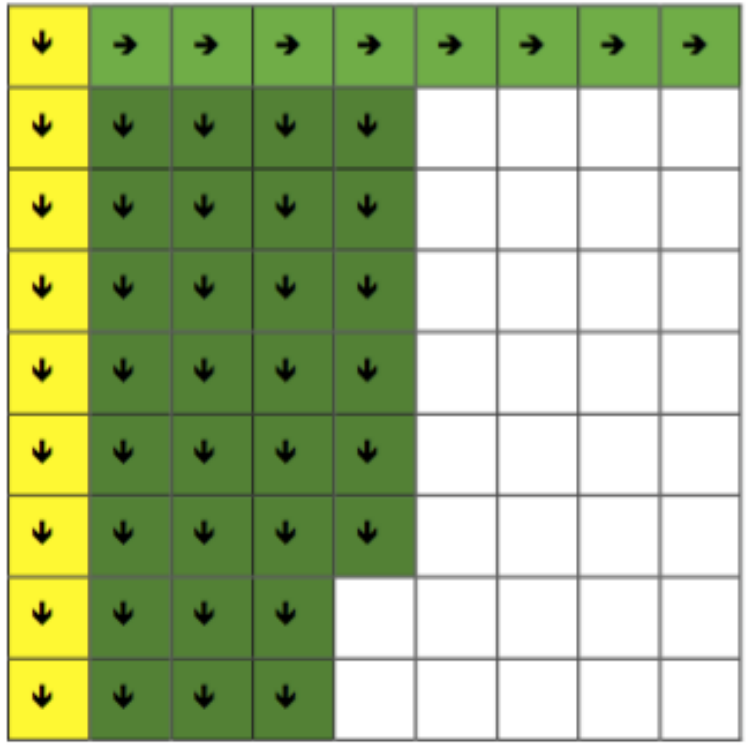
A*



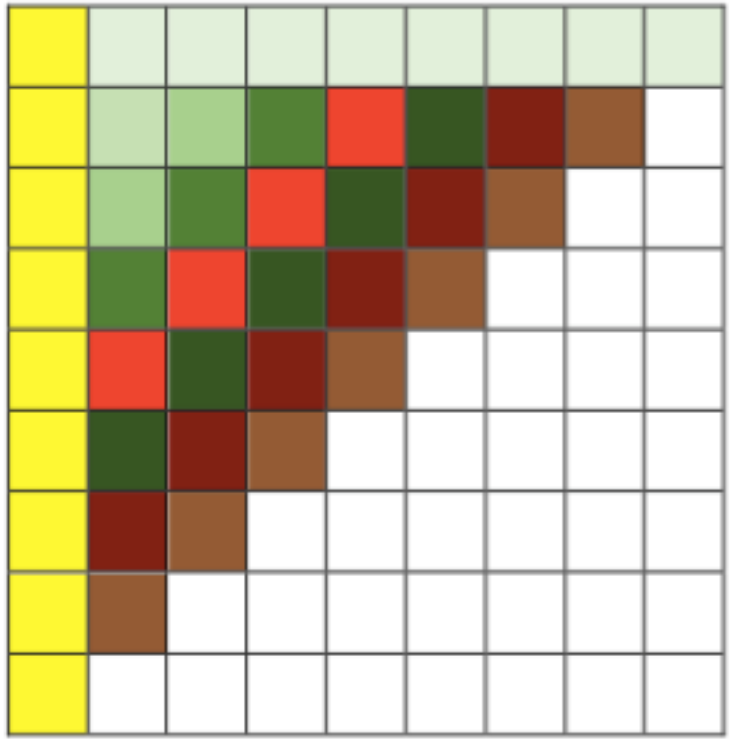
Needleman-Wunsch

Wołk K., Marasek K., „Tuned and GPU-accelerated Parallel Data Mining from Comparable Corpora”, Lecture Notes in Artificial Intelligence, p. 32 – 40, ISBN: 978-3-319-24032-9, Springer, 2015

Poprawa Szybkości Dopasowania



CPU Needleman-Wunsch



GPU Needleman-Wunsch

Wołk K., Marasek K., “Unsupervised comparable corpora preparation and exploration for bi-lingual translation equivalents”, Proceedings of the 12th International Workshop on Spoken Language Translation, Da Nang, Vietnam, December 3-4, 2015, p.118-125

Skrypt Tuningujący

Domena	Poprawa w %
BTEC	13.2
TED	12.5
EMEA	17
EUP	5
OPEN	11.4

Dodatkowe dane z użyciem skryptu tuningującego wagi dopasowania.

Manualne zrównoleglenie – 100 artykułów

Wikipedia - Usprawnienia

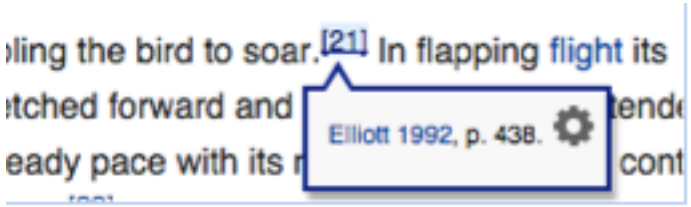
Tytuły.

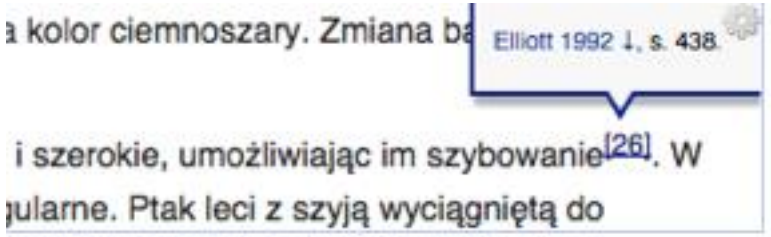
1000 artykułów

5,5%

Podpisy zdjęć, tabel itp.

Analiza przypisów.





Wzrost Ilości Pozyskanych Danych

Metoda pozyskiwania	Ilość par zdań
MONO	510,128
BI	530,480
NW	1,729,061
NWT	2,984,880

Liczba uzyskanych par zdań

Euronews 4,498 par zdań

http://opus.lingfil.uu.se/Wikipedia.php

http://iwslt.org

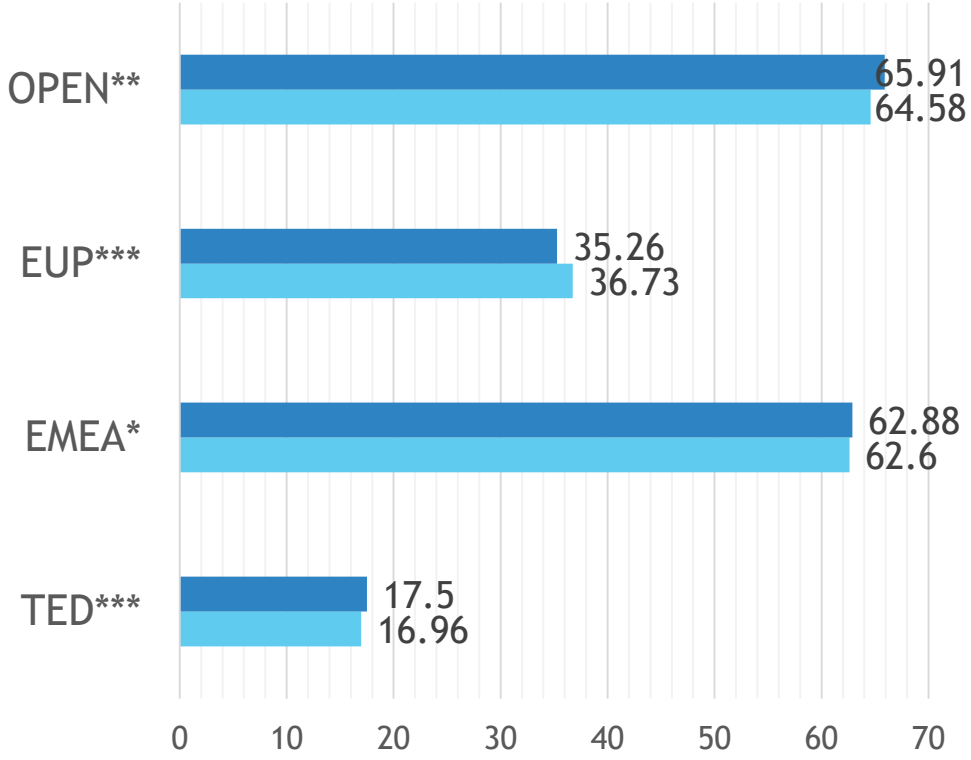
Zmniejszenie Czasu Wydobywania Danych

Metoda wydobycia	Czas obliczeń [s]
YALIGN	89,67
M YALIGN	14,7
NW YALIGN	17,3
GNW YALIGN	15,2

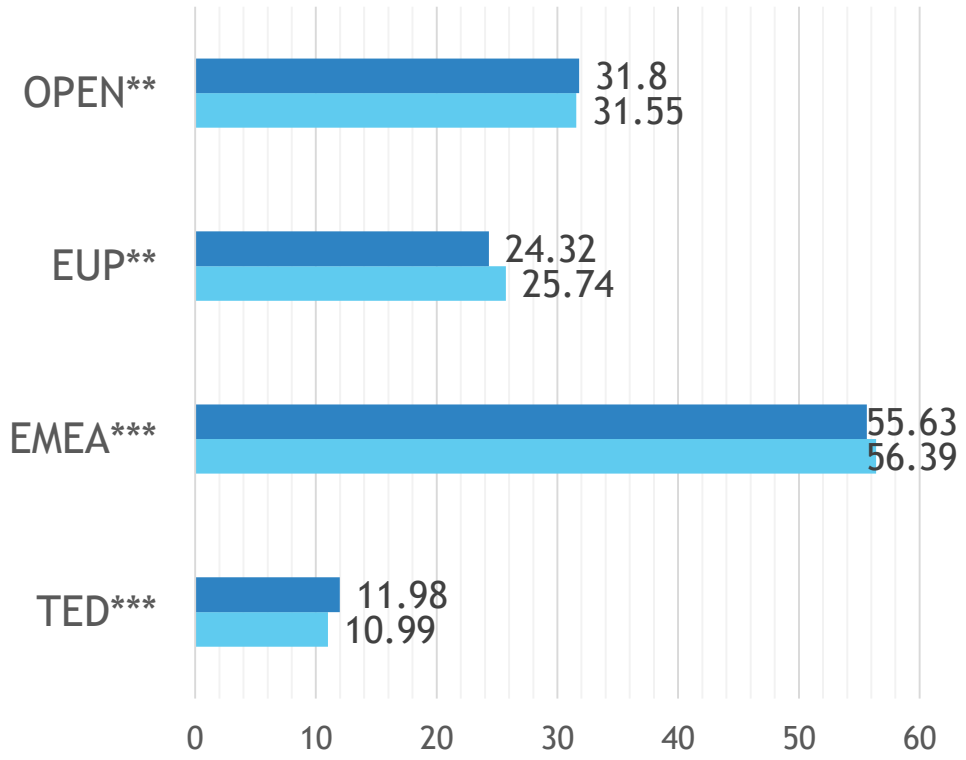
Czas obliczeń uzyskany przez różne wersje YALIGN

Intel Core i7 CPU
GeForce GTX 660 GPU

SMT z Pozyskanymi Korpusami (BLEU)



Polski - Angielski



Angielski - Polski

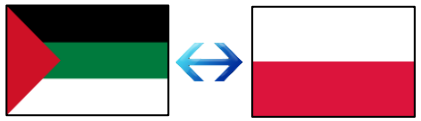
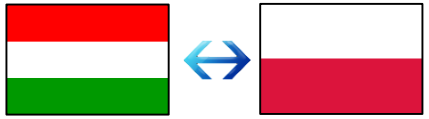
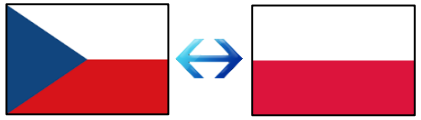
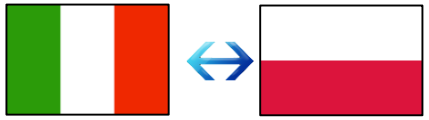
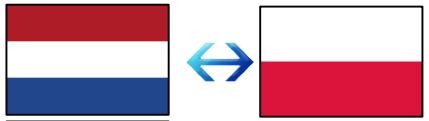
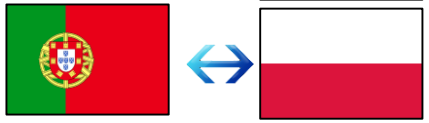
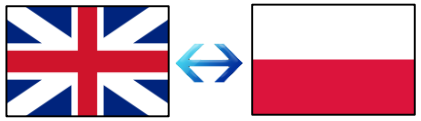
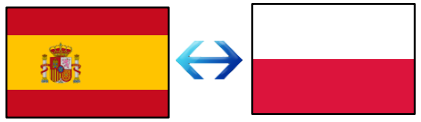
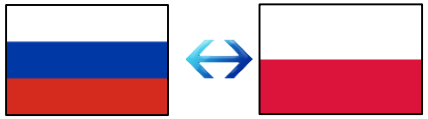
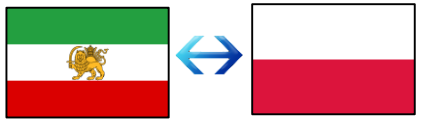
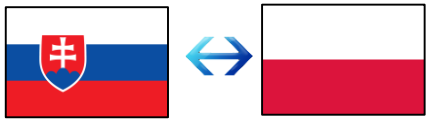
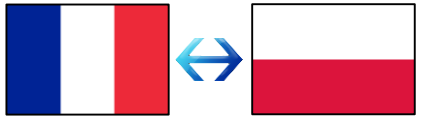
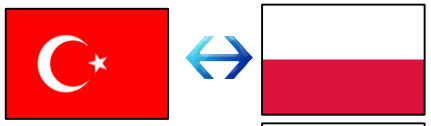
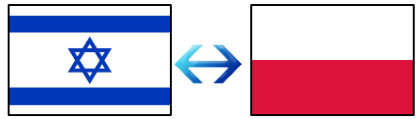
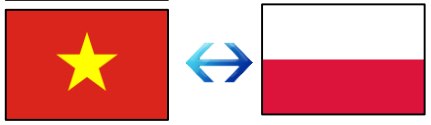
* Niski poziom istotności
** Zmiany istotne
*** Zmiany bardzo istotne

Jakość Pozyskanych Danych

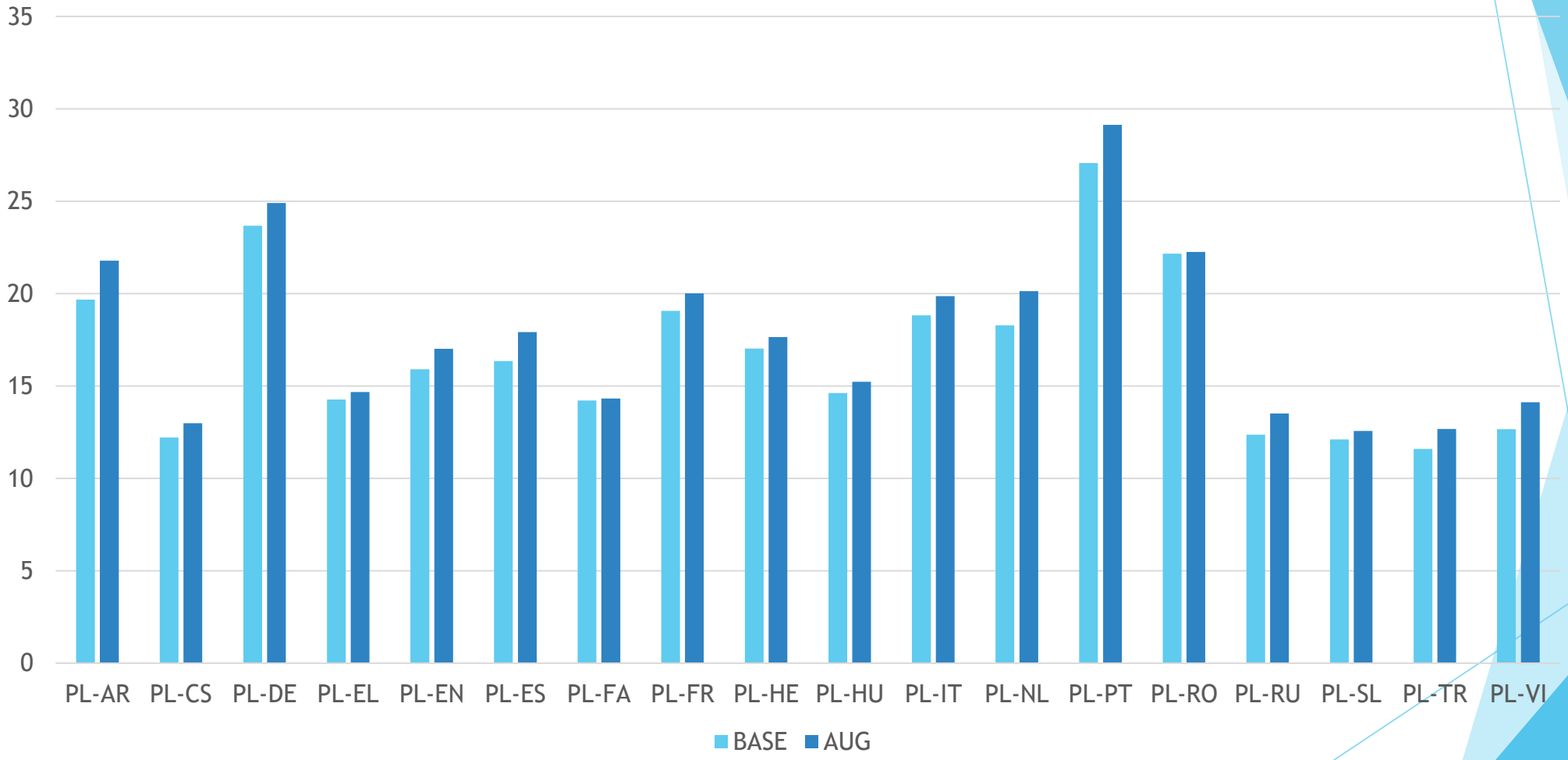
PL	EN
W przeciwnym razie alternatywa zdań jest fałszywa.	In all other cases it is true.
ISBN 9788361182085.	ISBN 0-14-022697-4. .
ASN.1 (skrót od "Abstract Syntax Notation One" - abstrakcyjna notacja składniowa numer jeden) jest to standard służący do opisu struktur przeznaczonych do reprezentacji, kodowania, transmisji i dekodowania danych.	Abstract Syntax Notation One (ASN.1) is a standard and notation that describes rules and structures for representing, encoding, transmitting, and decoding data in telecommunications and computer networking.
Standard ASN.1 określa jedynie składnię abstrakcyjną informacji, nie określa natomiast sposobu jej kodowania w pliku.	ASN.1 defines the abstract syntax of information but does not restrict the way the information is encoded.
Metody kodowania informacji podanych w składni ASN.1 zostały opisane w kolejnych standardach ITU-T/ISO.	X.683 | ISO/IEC 8824-4 (Parameterization of ASN.1 specifications)Standards describing the ASN.1 encoding rules:* ITU-T Rec.
pierwszego).	One criterion .
problemy nierozstrzygalne.	.".
temperaturę) w optymalnym zakresie.	"Algorithmic theories"...
Na początku lat 30.	U.S. Dept.

Wyniki

@ opus.lingfil.uu.se/Wikipedia.php

	823,715			41,116	
	62,507			210,435	
	169,739			167,081	
	12,222			208,756	
	172,663			6,742	
	151,173			170,227	
	6,092			17,228	
	51,725			15,993	
	10,006			90,428	

Wpływ nowych danych na jakość SMT



Dane wydobyte dla innych par EN-*

151,135

27,723

463,221

104,076

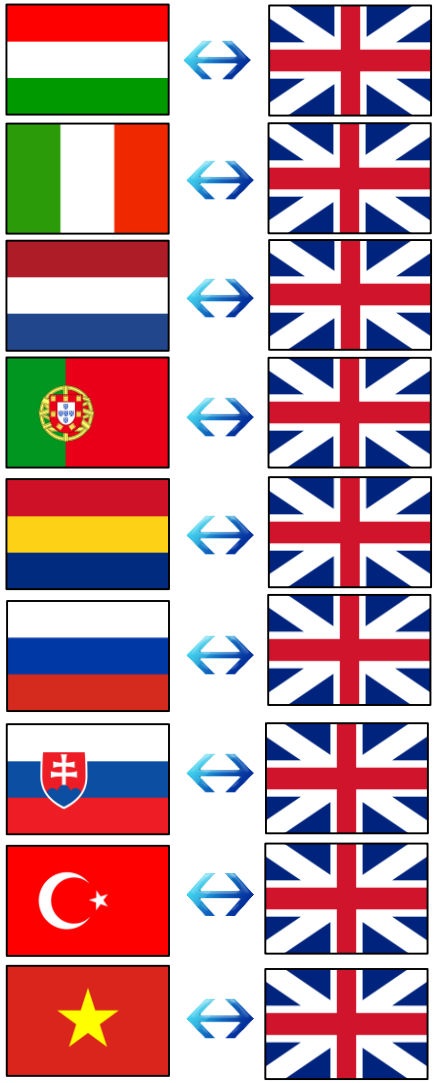
172,663

400,194

97,091

573,608

139,853



61,472

387,829

436,594

435,594

343,982

463,221

479,979

122,591

58,116

Wnioski

- Pomyślnie pozyskano dane równoległe
- Pozyskane dane są dostosowane domenowo
- Korpusy oceniono poprzez ich użycie w tłumaczeniu maszynowym
- Odnotowano wzrost jakości translacji z użyciem danych porównywalnych
- Rozwiązanie jest niezależne od pary języka i automatyczne
- Rozwiązanie udostępniono: https://github.com/krzwolk/yalign

OPUS

http://opus.nlpl.eu/Wikipedia.php

Korpusy quasi-porównywalne WYNIKI

Corpus	Number of bi-sentences	Number of unique EN tokens	Number of unique foreign tokens
TED	2,459,662	2,576,938	2,864,554
EXT	818,300	1,290,000	1,120,166

Na 1,000 słów kluczowych po 5 stron.

https://github.com/krzwolk/Comparable-Corpora-Builder

Wydobywanie danych quasi-porównywalnych

- GIZA++, słownik 44K słów
- Google Search – po 5 stron wyników
- 3,5 GB danych tekstowych
- 45M segmentów EN
- 16M segmentów PL
- Ujednolicenie kodowań i oczyszczenie danych

Język	System	Kierunek	BLEU
PL-EN	BASE	->EN	30.21
	EXT	->EN	31.37
	BASE	EN<-	21.07
	EXT	EN<-	22.47

Korpus	Ilość zdań równoległych
TED	2.459.662
EXT	818.300

Kierunek	P-VALUE
PL-EN	0.0203**
EN-PL	0.0021***

Wołk, K., & Marasek, K. (2017). Unsupervised Construction of Quasi-comparable Corpora and Probing for Parallel Textual Data. In *Multimedia and Network Information Systems* (pp. 307-320). Springer International Publishing.

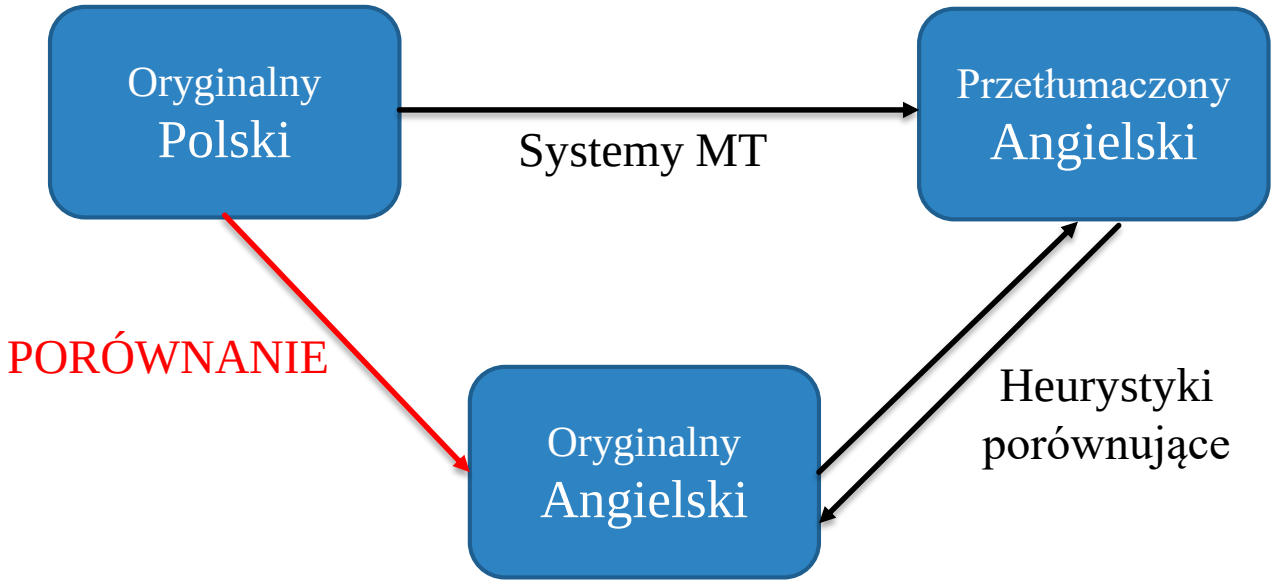
Usługa dostępna jest pod adresem:

http://ld.clarin-pl.eu:9000

http://ld.clarin-pl.eu:9000/#/aligner/9fxqqzwnd0irtrkt94epgf18cryby9xg

Filtrowanie Korpusów - Statystyczne

https://github.com/krzwolk/Statistical-Parallel-Corpora-Filtration



Ilość zdań wejściowych	1000
Ilość niepoprawnych zdań	182
Ilość odfiltrowanych niepoprawnych zdań	154
Ilość odfiltrowanych poprawnych zdań	12

Poprawa jakości SMT po filtracji: 1-4%

“ Wołk K., Marasek K., “A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation.”, Advances in Intelligent Systems and Computing volume 275, p.107-114, Publisher: Springer, ISSN 2194-5357, ISBN 978-3-319-05950-1

Filtrowanie Korpusów - Heurystyczne

https://github.com/krzwolk/Differential-Parallel-Corpora-Cleaning-Tool

Filtering can be done using absolute and relative difference:
abs_diff = n rel_diff = n/m
Sentence will be preserved if absolute or relative difference is more for it then provided in script options
Example command:
$ python sentence_cleaner.py --abs_diff 3 --dictionary dictionary.txt --stop_words stop.txt text1 text2

Wykrywanie języka

https://github.com/krzwolk/langdedect

“ Wołk K., Marasek K., “Unsupervised comparable corpora preparation and exploration for bi-lingual translation equivalents”, Proceedings of the 12th International Workshop on Spoken Language Translation, Da Nang, Vietnam, December 3-4, 2015, p.118-125

Filtrowanie w oparciu o reguły

https://github.com/krzwolk/Parallel-Corpora-Cleaning-Tool

Wyłącznie PL-EN
Szybkie
Korekta pisowni
Możliwa półautomatyczna korekta

“ Wołk K., Marasek K., “Polish – English Speech Statistical Machine Translation Systems for the IWSLT 2013.”, Proceedings of the 10th International Workshop on Spoken Language Translation, Heidelberg, Germany, p. 113-119, 2013

Common Crawl

Najbliższe plany

Big Data Language Model of Contemporary Polish

Model: http://goo.gl/hO1hTz

Wstępne przetwarzanie danych

- Common Crawl – 7 lat - 145 TB @ 2015, 280 teraz
- Problemy z kodowaniem
- Normalizacja danych
- Podział zdań
- Deduplikacja
- Wykrywanie języka
- Niska jakość filtrowania danych
- Tokenizacja i Truecasing

MOSES
statistical
machine translation
system

Sentence 1:ABCD Sentence 2: WXYZ

ABABCD.	ABCBCD.
ABCD.ABCD.	ABWXYZCD.

Deduplikacja i normalizacja

TABLE I.

DEDUPLICATION RESULTS

	Size in GB	**Number of sentences**	**Number of unique words**
Before	296,1	1,962,047,863	87,543,726
After	94,8	920,517,413	87,543,726

Ewaluacja

TABLE II.

TEST CORPORA SPECIFICATION

	Number of sentences	Number of PL words	Number of EN words
TED	210,549	218,426	104,177
EMEA	1,046,764	148,230	109,361
OPEN	33,570,553	1,519,948	758,238

TED Ideas worth spreading

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

opensubtitles.org

TABLE III.

TEST CORPORA SPECIFICATION

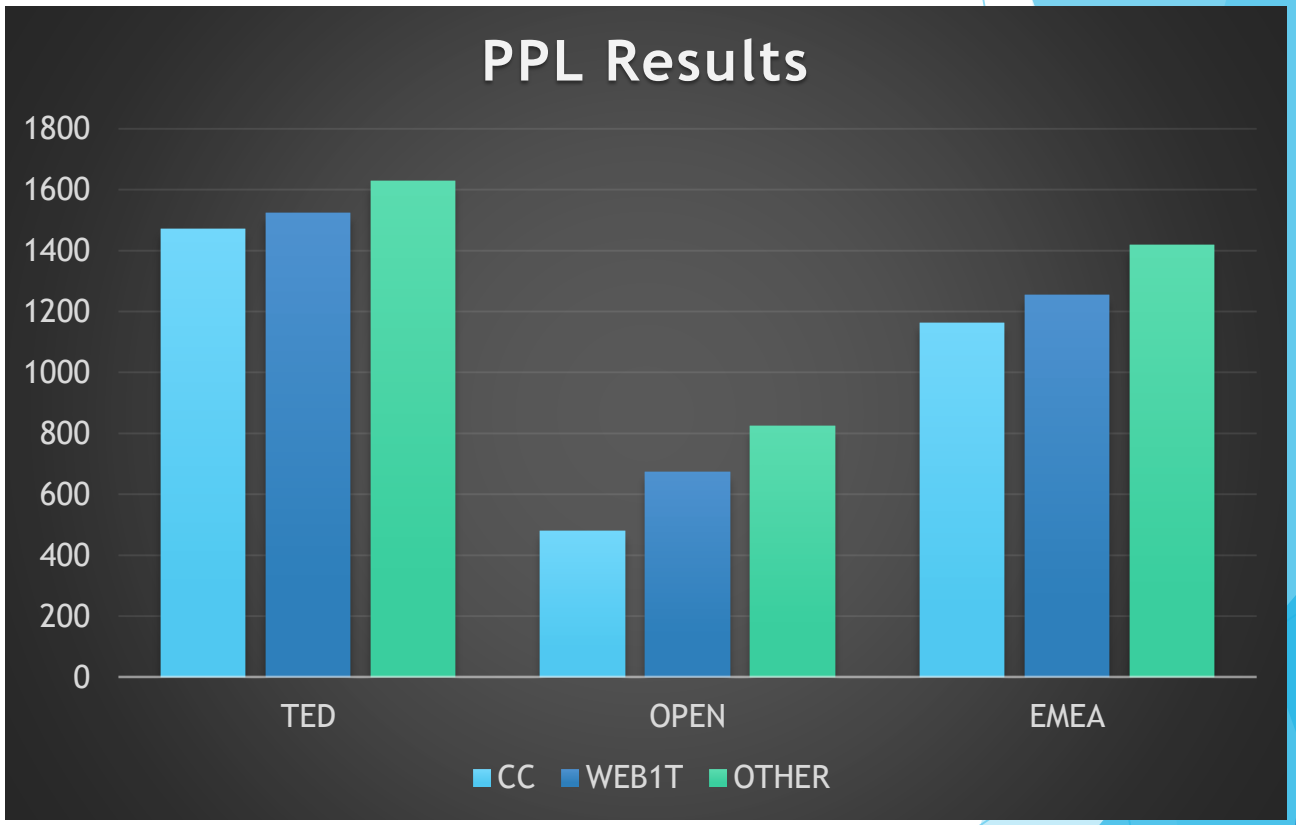
Corpus Name	Baseline system score (BLEU)
TED	17,42
EMEA	36,74
OPEN	58,52

Wyniki PPL

TABLE VI.

PERPLEXITY-BASED LANGUAGE MODEL EVALUATION

CORPUS	MODEL	PERPLEXITY (PPL)
TED	Common Crawl	1471
TED	WEB1T	1523
TED	OTHER	1628
OPEN	Common Crawl	480
OPEN	WEB1T	671
OPEN	OTHER	823
EMEA	Common Crawl	1163
EMEA	WEB1T	1253
EMEA	OTHER	1417



OPUS http://opus.lingfil.uu.se/

Wyniki MT

TABLE VII.

SMT-BASED LANGUAGE MODEL EVALUATION

CORPUS	LANGUAGE MODEL	Baseline BLEU	Augumented BLEU	Delta
TED	Common Crawl	17.42	18.33	0.91
TED	WEB1T	17.42	17.97	0.55
TED	OTHER	17.42	17.76	0.34
OPEN	Common Crawl	58.52	59.23	0.71
OPEN	WEB1T	58.52	59.01	0.49
OPEN	OTHER	58.52	58.79	0.27
EMEA	Common Crawl	36.74	38.34	1.6
EMEA	WEB1T	36.74	37.93	1.19
EMEA	OTHER	36.74	37.26	0.52

Korpus równoległy z Commoncrawl?

- Deduplikacja
- Odfiltrowanie danych PL i EN
- Analiza na podstawie adresów URL
- Analiza na podstawie dystansu Levenstein’a
- Klastrowanie pozostałych danych
- Zrównoleglenie na poziomie klastrów
- Między-klastrowe dopasowanie na poziomie artykułów (Levenstein, lista rankingowa)
- Zrównoleglenie naszym narzędziem
- Dodatkowa filtracja danych

Dziękuję za uwagę

kwolk@pja.edu.pl