KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
INTERKADRA
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt „Interdyscyplinarna kadra akademicka na rzecz rozwoju gospodarki opartej na wiedzy" współfinansowany przez Unię Europejską ze środków Europejskiego Funduszu Społecznego w ramach Poddziałania 4.1.1 „Wzmocnienie potencjału dydaktycznego uczelni" Programu Operacyjnego Kapitał Ludzki.

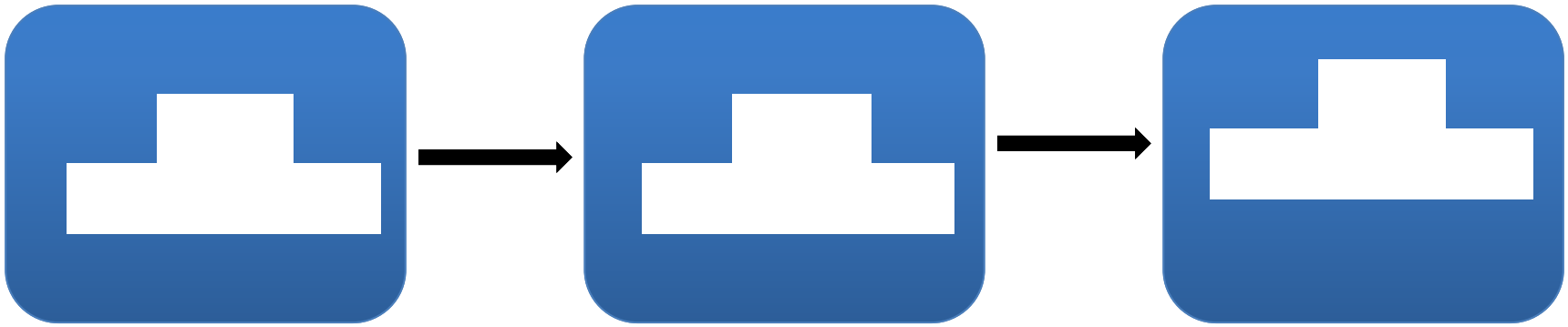
Telemedicine as a specia translation

Author: Krzysztof Wołk

Importance

- Narrow text domain
- Important and promising field of research
- Medical records translation
- Foreign medical history access
- Communication between patient and doctor abroad
- Telemedicine
- Lower costs that human translators

The idea …



Domain	WER	Vocabulary size
Polish Senate	19,6	87k
TV news	15,72	42k
Lectures	27,75	210k

EU-BRIDGE Project

Requirements

- Bilingual Parallel Data
- Alignment
- Monolingual Data for Language Model
- Data in good quality
- Data that is natural
- Really Lots of data

Problems with SMT for PL – EN Pair

- Not Enough Polish Data
- Poor Quality of Accessible Polish Data
- Bad, poor or too indirect translations

Sentence 1:ABCD	Sentence 2: WXYZ
ABABCD.	ABCBCD.
ABCD.ABCD.	ABWXYZCD.

- E.g. about 10% spelling 17% insertions 6% translations in TED

Differences in laguages

- Different syntactic order (SVO)

I bought myself a new car.	**Kupiłem sobie nowy samochód.** **Nowy samochód sobie kupiłem.** **Sobie kupiłem nowy samochód.** **Samochód nowy sobie kupiłem.**

- Bigger Vocabulary (E.g. TED, PL: 92135, EN: 41684)
- More complex grammar

  Cases:
  PL: 7 vs EN: 3

  In Polish 15 gender forms for nouns and adjectives, with additional dimensions for other word classes.

The EMEA Corpora

European Medical Agency (EMEA)
Leaflets from the drugs + Documents

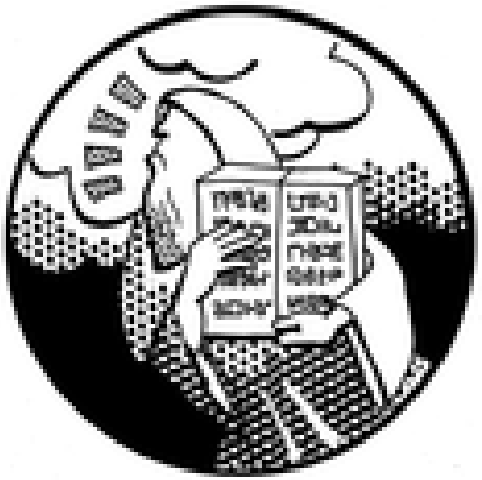
Corpora size: about 80 MB

1,044,764 bi-sentences

11.67M words that were not tokenized
148,170 unique PL words
109,326 unique EN words

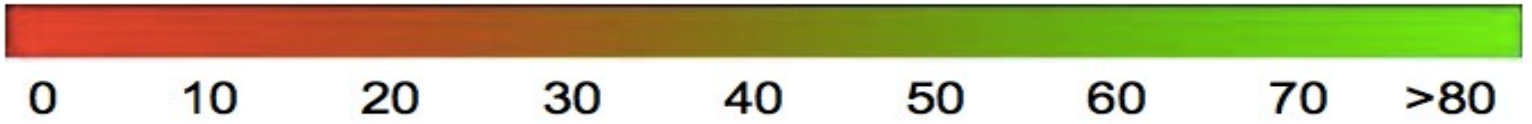
TOOLS

- Moses
  - Pipeline
  - Decoder
- Wroclaw NLP
- JMX Tagger
- EMS



Automatic Evaluation Methods for MT

- BLEU – Bilingual Evaluation Understudy



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

RESULTS

Polish-to-English translation

System	BLEU	NIST	METEOR	TER
00	70.15	10.53	82.19	29.38
01	64.58	9.77	76.04	35.62
02	71.04	10.61	82.54	28.33
03	71.22	10.58	82.39	28.51
04	76.34	10.99	85.17	24.77
05	70.33	10.55	82.28	29.27
06	71.43	10.60	82.89	28.73
07	71.91	10.76	83.63	26.60
08	71.12	10.37	84.55	29.95
09	71.32	10.70	83.31	27.68
10	71.35	10.40	81.52	29.74
11	70.34	10.64	82.65	28.22
12	72.51	10.70	82.81	28.19

RESULTS

English-to-Polish translation

System	BLEU	NIST	METEOR	TER
00	69.18	10.14	79.21	30.39
01	61.15	9.19	71.91	39.45
02	69.41	10.14	78.98	30.90
03	68.45	10.06	78.63	31.62
04	73.32	10.48	81.72	27.05
05	69.21	10.15	79.26	30.88
06	69.27	10.16	79.30	31.27
07	68.43	10.07	78.95	33.05
08	67.61	9.87	77.82	29.95
09	68.98	10.11	78.90	31.13
10	68.67	10.02	78.55	31.92
11	69.01	10.14	79.13	30.84
12	67.47	9.89	77.65	33.32

Conclusions

- Hierarchical Models – biggest improvement
- Truecasing & Punctuation Normalization
- PL -> EN better than EN -> PL
- High BLEU score (over 70)
- Beneficial to the patients
- Aid for human interpreters

TRANSLATION EXAMPLES O-ORIGINAL T-TRANSLATED

O: I think we should reflect carefully before we do that again.

T: I believe that we should reflect carefully before we do that again.

O: International cooperation is an essential tool for an effective fight against this scourge, at both national and multilateral levels.

T: International cooperation is an essential tool for an effective fight against this scourge, at both national and international levels.

O: We have completely overhauled our policy style and the way in which we prepare and review legislation.

T: We have completely overhauled our policy style and the way in which we prepare and review legislation.

O: Abilify is a medicine containing the active substance– aripiprazole.

T: Abilify is a medicine containing the active substance aripiprazole.

TRANSLATION EXAMPLES O-ORIGINAL T-TRANSLATED

O: As far as I am aware, a rather successful long term social dialogue has been conducted on these issues and, in my opinion, it is thus necessary to reconcile seemingly contradictory procedures.

T: As far as I know, in this regard have been carried out rather successful long-term social dialogue and, in my view, there is a need to reconcile seemingly contradictory procedures.

O: They are concerned with the creation of a milk fund that will support reforming activities in this sector, such as the promotion of milk consumption, including consumption in schools, or support for milk production in mountainous areas.

T: They are concerned with the creation of a milk fund that will support reforming activities in this sector, such as the promotion of milk consumption, including consumption in schools, or support for milk production in mountain areas.

TRANSLATION EXAMPLES O-ORIGINAL T-TRANSLATED

O: Split Influenza virus, inactivated, containing antigens equivalent to the following strains:

T: Split influenza virus, inactivated, containing antigens equivalent to the following strains:

O: The immunological response may be diminished if the patient is undergoing immunosuppressant treatment.

T: The immunological response may be wakened if the patient is undergoing immunosuppressant treatment.

O: Optisulin was originally studied in 10 studies, in both type 1 and type 2 diabetes.

T: Optisulin was originally studied in 10 studies in diabetes type 1 and in type 2.

Future Work

- Combining all experiments
- Extending training data and language models
- Stems instead of surface forms
- Merging with TTS and ASR systems
- Filtering and cleaning parallel data
- Changing word order to meet SVO order

THANK YOU FOR ATTENTION !

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt „Interdyscyplinarna kadra akademicka na rzecz rozwoju gospodarki opartej na wiedzy" współfinansowany przez Unię Europejską ze środków Europejskiego Funduszu Społecznego w ramach Poddziałania 4.1.1 „Wzmocnienie potencjału dydaktycznego uczelni" Programu Operacyjnego Kapitał Ludzki.