PolEval 2019:
Task 4: Machine Translation

Łukasz Podlodowski

WARSZAWA, 31.05.2019

PolEval Task 4: Machine Translation

Trzy podzadania:

1) EN -> PL
2) PL -> RU
3) RU -> PL

Około 130 tys. par zdań dla EN -> PL;
Około 23 tys. par zdań dla RU <-> PL

PL <-> RU

Mały zbiór uczący oraz wiele odmian lematów

```
>>> butyavka.lexeme
[Parse(word='бутявка', tag=OpencorporaTag('NOUN,inan,femn sing,nomn'), normal_form='бутявка
 Parse(word='бутявки', tag=OpencorporaTag('NOUN,inan,femn sing,gent'), normal_form='бутявка
 Parse(word='бутявке', tag=OpencorporaTag('NOUN,inan,femn sing,datv'), normal_form='бутявка
 Parse(word='бутявку', tag=OpencorporaTag('NOUN,inan,femn sing,accs'), normal_form='бутявка
 Parse(word='бутявкой', tag=OpencorporaTag('NOUN,inan,femn sing,ablt'), normal_form='бутявк
 Parse(word='бутявкою', tag=OpencorporaTag('NOUN,inan,femn sing,ablt,V-oy'), normal_form='б
 Parse(word='бутявке', tag=OpencorporaTag('NOUN,inan,femn sing,loct'), normal_form='бутявка
 Parse(word='бутявки', tag=OpencorporaTag('NOUN,inan,femn plur,nomn'), normal_form='бутявка
 Parse(word='бутявок', tag=OpencorporaTag('NOUN,inan,femn plur,gent'), normal_form='бутявка
 Parse(word='бутявкам', tag=OpencorporaTag('NOUN,inan,femn plur,datv'), normal_form='бутявк
 Parse(word='бутявки', tag=OpencorporaTag('NOUN,inan,femn plur,accs'), normal_form='бутявка
 Parse(word='бутявками', tag=OpencorporaTag('NOUN,inan,femn plur,ablt'), normal_form='бутяв
 Parse(word='бутявках', tag=OpencorporaTag('NOUN,inan,femn plur,loct'), normal_form='бутявк
```

```
funkcja   funkcja subst:sg:nom:f   pospolita
funkcjach funkcja subst:pl:loc:f   pospolita
funkcjami funkcja subst:pl:inst:f  pospolita
funkcje   funkcja subst:pl:acc:f   pospolita
funkcje   funkcja subst:pl:nom:f   pospolita
funkcje   funkcja subst:pl:voc:f   pospolita     rzad.
funkcji   funkcja subst:pl:gen:f   pospolita
funkcji   funkcja subst:sg:gen:f   pospolita
funkcjo   funkcja subst:sg:voc:f   pospolita     rzad.
funkcjom  funkcja subst:pl:dat:f   pospolita
funkcyj   funkcja subst:pl:gen:f   pospolita     arch.|matem.
```

PL <-> RU

Dla DNN potrzebny jest znacznie większy zbiór uczący oraz moc obliczeniowa, a także bardziej skomplikowany model:

1. Sieci neuronowe z wyjściem softmax popularnie stosowane dla tłumaczeń z/do j. angielskiego muszą mierzyć się z problemem kilkanaście razy większej przestrzeni decyzyjnej. Zbiór zdań jest zdecydowanie mniejszy od słownika (bez lematyzacji).
2. W obu językach kolejność występowania części zdania cechuje się wysoką swobodą. Modele uczące się sekwencji (LSTM, Transformer, etc.) mogą wyłapywać szumy związane z nieistniejącymi wzorcami – dodatkowy aspekt utrudniający stosowanie modeli przygotowanych dla j. angielskiego.

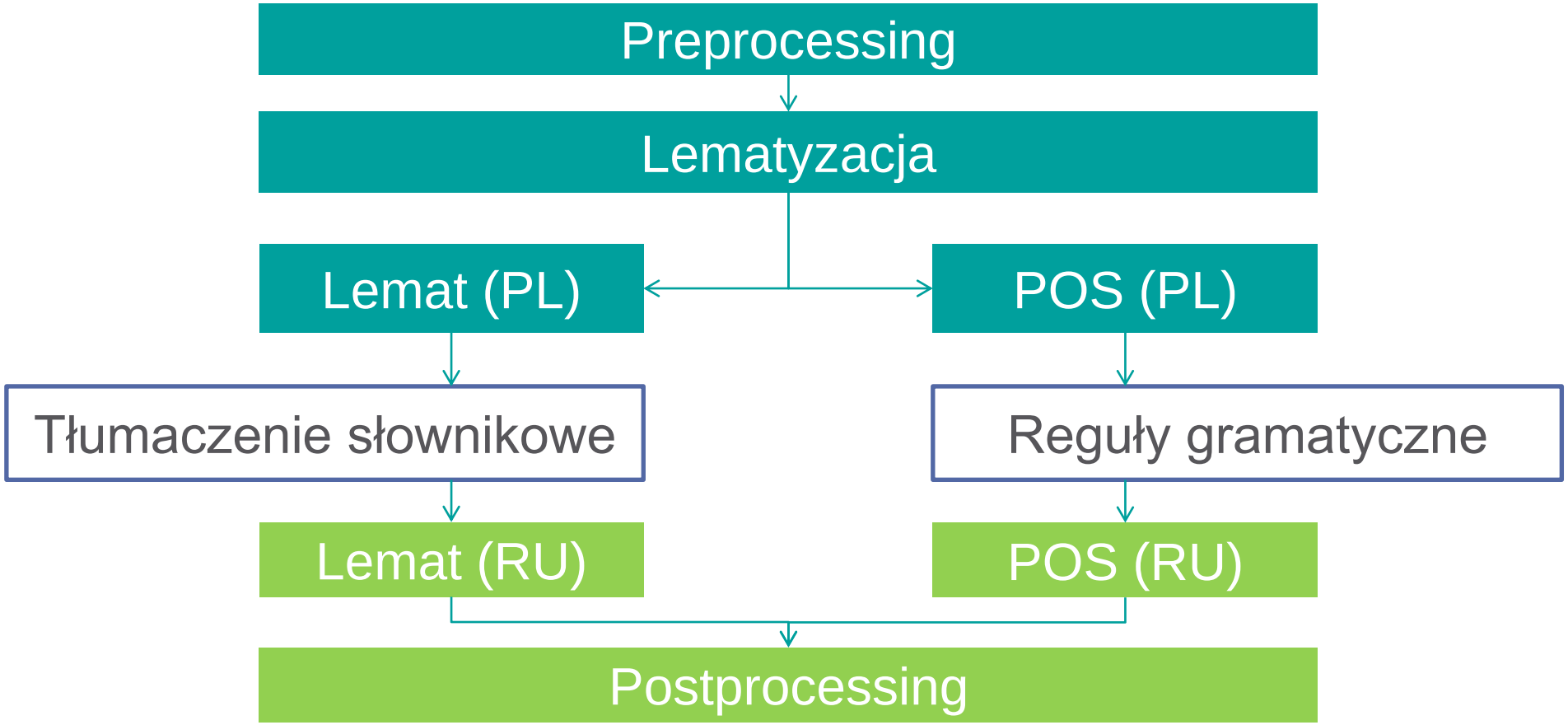
PL <-> RU

J. polski oraz j. rosyjski należą do tej samej grupy języków słowiańskich (m.in. podobieństwo gramatyki):

- Część wyrażeń niewystępujących w j. angielskim ma swoje odpowiedniki w j. rosyjskim/polskim, które mogą być tłumaczone bezpośrednio,
- Wysokie podobieństwo fleksyjne deklinacji oraz podobna struktury gramatyczne związane z wyrażaniem aspektu czasu.

PL <-> RU

Ogólny schemat tłumaczenia PL -> RU (reguły są odwracalne i stosowane także dla tłumaczenia RU -> PL):



PL <-> RU

Tłumaczenie słownikowe:

1. Lematyzuj słowa;
2. Wyznacz indeks odwrotny;
3. Określ przestrzeń kandydatów dla lematu;
4. Oceń każdego kandydata funkcją:

$$\text{score}(w_{pl}) = \frac{|RU_{w_{ru}} \cap PL_{w_{pl}}|^2}{|RU_{w_{ru}}| + |PL_{w_{pl}}|}$$

pyMorphy2	Morfologik
NOUN, NUMR	subst, depr, burk, ger, num
ADJF, ADJS, COMP	adj, adja, adjc, adv
VERB, INFN	verb, ger
PRTF, PRTS, GRND	pant, ppas, pcon
NPRO	ppron12, ppron3, siebie
ADVB	adv,adjp
PRED	pred
PREP	prep
CONJ	conj, comp
PRCL	qub
INTJ	interj

Procedura powtarzana dla ograniczonej przestrzeni z uwzględnieniem POS, bez stoplisty. Oddzielny słownik dla słów ze stoplisty. W rezultacie powstało kilka słowników o różnym stopniu zoptymalizowania precyzji oraz recall.

PL <-> RU

Reguły gramatyczne:

1. Tłumacz „1:1” dla czasu nieprzeszłego.
2. Tłumacz „1:1” dla trybu rozkazującego.
3. Forma przeszła <-> zaimek osobowy + pseudoimiesłów (jechałem <-> Я ехал).
4. Imiesłów przymiotnikowy bierny <-> pseudoimiesłów (forma krótka)
5. Tłumacz rzeczowniki „1:1”, ale uwzględnij zmianę płci / liczby (informacja propagowana do pozostałych reguł).
6. Tłumacz przymiotniki „1:1”.

PL <-> RU

Postprocessing:

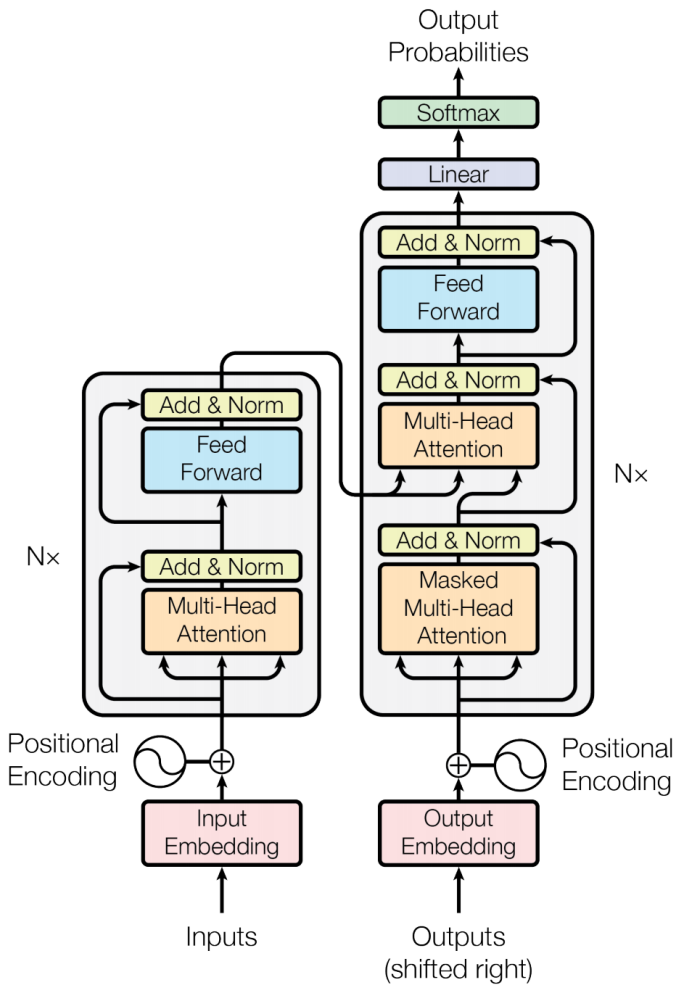
1. Przetłumacz „что” jako polskie „że”, jeżeli następnym słowem jest zaimek osobowy lub jest to początek zdania.
2. Usuń nadmierne zaimki osobowe. (Я бегу -> biegnę).
3. Usuń nadmierne występowanie „давать”. („давать” + czasownik)
4. Przywróć oryginalne formatowanie; znaki interpunkcyjne, wielkość liter.

EN -> PL

Model Transformer oparty o mechanizm atencji:

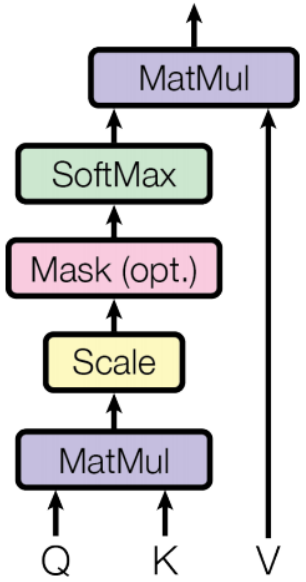
1. Opublikowany w 2017 r.
2. Wyznaczył nowe state-of-the-art dla zadania WMT 2014 English-to-German.
3. Model to głęboka sieć neuronowa zawierający warstwy Multi-Head Attention / Scaled Dot-Product Attention zamiast powszechnie stosowanych modeli rekurencyjnych.
4. Przestrzeń decyzyjna została ograniczona dla słów, które wystąpiły więcej niż 5 razy.
5. Parametryzacja opierała się o podstawową konfigurację oryginalnej pracy, zwiększona została jedynie liczba neuronów ukrytych w warsztwach feed-forward do 4096.

EN -> PL

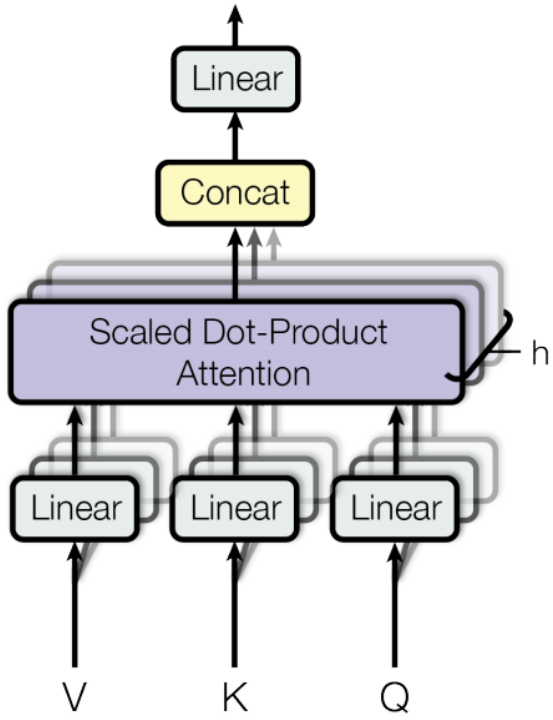

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

EN -> PL

Scaled Dot-Product Attention



Multi-Head Attention



Podsumowanie

- Model regułowy pozwala znacząco ograniczyć wielkość zbioru uczącego kosztem złożoności analizy ludzkiej.
- Stosowanie rozwiązań uczenia głębokiego przygotowanych pod kątem j. angielskiego jest znacząco bardziej skomplikowane dla języka polskiego (bogata fleksja, większa swoboda sekwencji).

Subtask	BLEU	NIST	TER	METEOR
English to Polish	4.92	2.27	86.56	21.74
Polish to Russian	5.38	2.53	83.02	53.54
Russian to Polish	5.51	2.97	85.27	24.08

Bibliografia

Bahdanau D., Cho K. and Bengio Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. „CoRR”, abs/1409.0473.

Banerjee S. and Lavie A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Cho K., Van Merriënboer B., Bahdanau D. and Bengio Y. (2014).On the Properties of NeuralMachine Translation: Encoder-Decoder Approaches. „CoRR”, abs/1409.1259.

Doddington G. (2002).Automatic Evaluation of Machine Translation Quality Using N-gramCo-occurrence Statistics. InProceedings of the 2nd International Conference on Human LanguageTechnology Research (HLT 2002), pp. 138–145, San Francisco, CA, USA. Morgan KaufmannPublishers Inc.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W.,Moran C., Zens R. et al. (2007).Moses: Open Source Toolkit for Statistical Machine Translation.InProceedings of the 45th Annual Meeting of the Association for Computational Linguistics(Companion volume: Proceedings of the Demo and Poster Sessions), pp. 177–180.

Korobov M. (2015).Morphological Analyzer and Generator for Russian and Ukrainian Languages.In Khachay M. Y., Konstantinova N., Panchenko A., Ignatov D. I. and Labunets V. G. (eds.),Analysis of Images, Social Networks and Texts, vol. 542 ofCommunications in Computer and Information Science, pp. 320–332. Springer International Publishing.

Bibliografia

Li Z., Callison-Burch C., Dyer C., Ganitkevitch J., Khudanpur S., Schwartz L., Thornton W. N.,Weese J. and Zaidan O. F. (2009).Joshua: An Open Source Toolkit for Parsing-based MachineTranslation. InProceedings of the 4th Workshop on Statistical Machine Translation, pp. 135–139.Association for Computational Linguistics

Luong M.-T., Pham H. and Manning C. D. (2015).Effective Approaches to Attention-basedNeural Machine Translation. „CoRR”, abs/1508.04025

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2002).BLEU: A Method for Automatic Eval-uation of Machine Translation. InProceedings of the 40th Annual Meeting on Association forComputational Linguistics (ACL 2002), pp. 311–318, Stroudsburg, PA, USA. Association forComputational Linguistics.

Snover M., Dorr B., Schwartz R., Micciulla L. and Makhoul J. (2006).A Study of TranslationEdit Rate with Targeted Human Annotation. InProceedings of Association for Machine Translationin the Americas, pp. 223–231.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. and PolosukhinI. (2017).Attention Is All You Need. „CoRR”, abs/1706.03762.

Bibliografia

Woliński M. (2006).Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In Kłopotek M. A., Wierzcho´n S. T. and Trojanowski K. (eds.),Intelligent Information Processingand Web Mining, pp. 511–520, Berlin, Heidelberg. Springer.

Wu Y., Schuster M., Chen Z., Le Q. V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q.,Macherey K. et al. (2016).Google's Neural Machine Translation System: Bridging the GapBetween Human and Machine Translation. „CoRR", abs/1609.08144.