Dependency Trees in Automatic Inflection of Multi Word Expressions in Polish

Ryszard Tuora Łukasz Kobyliński

Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland

20 grudnia 2021

Szablonowe NLG

- W NLG dalej jest miejsce dla metod opartych o szablony

  'Anna', 'You have a message from . . . '
- W językach bogatych morfologicznie, implementacja jest trudniejsza.

  'Anna', 'Masz wiadomość od . . . '

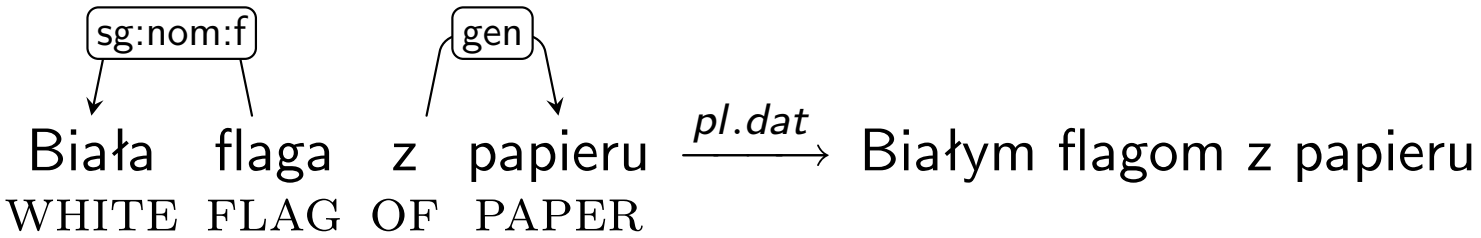
  Konieczne jest wykorzystanie dopełniacza.
- W wielu wypadkach konieczna jest obsługa wyrażeń wielowyrazowych (*Multi-Word Expressions - MWE)*

  **Masz wiadomość od Anna Kowalska.*

  vs.

  Masz wiadomość od Anny Kowalskiej.

Związki zgody i rządu



- *Biała* modyfikuje *flaga*, więc musi zachować zgodną postać w trakcie odmiany.
- *papieru* jest modyfikatorem rzeczownikowym rządzonym przez *z*, więc nie ulegnie odmianie.

NLG 2-etapowe

Szablon:

szukam jakichś #ADJ:gen:pl# #CATEGORY:gen:pl# [PREPSEQ?]

Słownik fraz:

Dodatek:ZŁańcuszkiem:
z łańcuszkiem - PREP, ozdabiane łańcuszkiem - ADJ, ozdobny łańcuszek - SUBST

Wypełniony szablon:

szukam jakichś ozdabianych łańcuszkiem szpilek ze skóry

Zakres interesujących zjawisk językowych

- Frazy rzeczownikowe
  Biała flaga z papieru
- Frazy przymiotnikowe
  Pochodzące z Lazurowego Wybrzeża
- Koordynacje
  Skórzany but i jeansowe spodnie
- Nazwy własne
  Uniwersytet Papieski Jana Pawła II w Krakowie

Szablonowe NLG w czeskim

Generator Seq2Seq tworzy sekwencje lematów i tagów, fleksja odbywa się słownikowo.

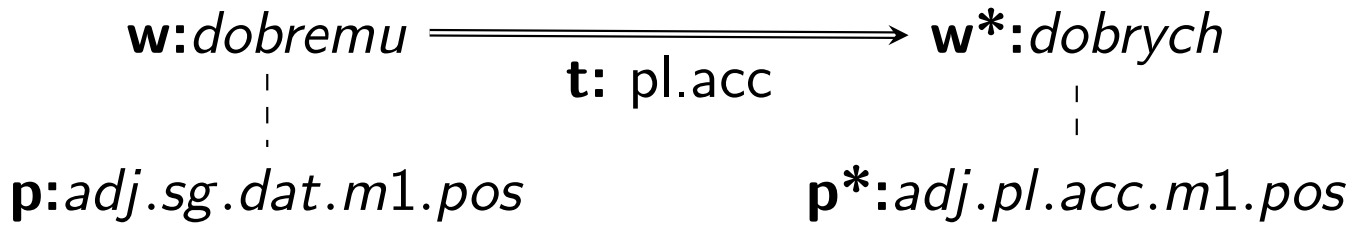
hledat	VB-P---2P-AA---	*vhodný*	AAFS4----1A----	*restaurace*	NNFS4-----A----
search	verb, 2nd person present formal	suitable	adjective, fem sg acc	restaurant	noun, fem sg acc

na	RR--4----------	*X-good_for_meal*	NNFS4-----A----	*?*	Z:----------
for	preposition, acc	slot placeholder	noun, fem sg acc	?	final punctuation

(Dusek i Jurcıcek 2019)

Zadanie

Naszym zadaniem jest reinfleksja, tj. odmiana jednej formy (niekoniecznie lematu) do innej postaci.



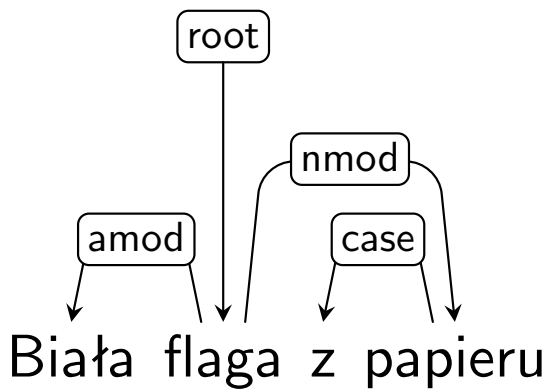
Rysunek. Odmiana *dobremu* do postaci dobrych, tylko liczba ($sg \rightarrow pl$) i przypadek ($dat \rightarrow gen$) ulegają zmianie

Flexer

Bierzemy pod uwagę dwie metody odmiany pojedynczych słów:

- Metoda słownikowa (oparta o Morfeusza 2)
- Metoda neuronowa - sieć seq2seq oparta na znakach

Następnie rozszerzamy to na MWE, korzystając z drzew zależnościowych.
Korzystamy z lematów, tagów POS, rozpoznanych cech morfologicznych, oraz analiz zależnościowych systemu COMBO (Klimaszewski i Wróblewska 2021)



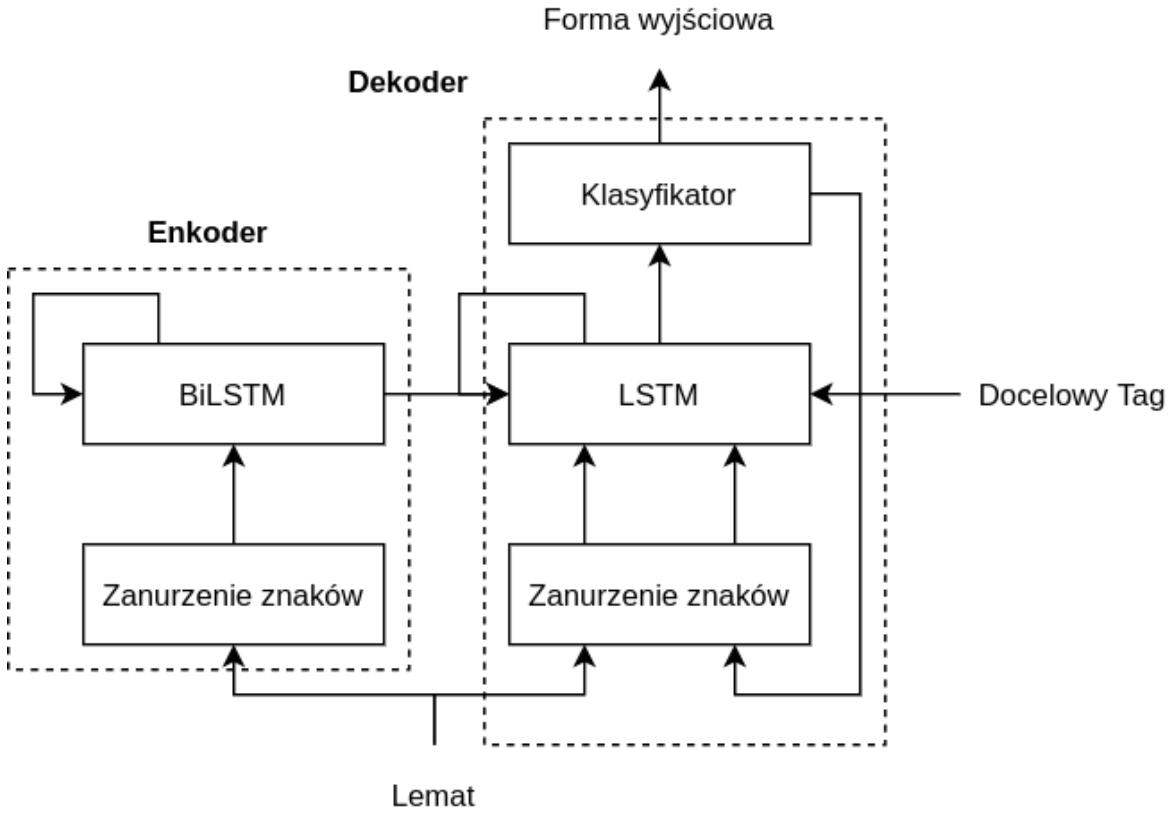
Metoda słownikowa

$$dobremu + pl.acc \rightarrow dobrych$$

	m1	pozostałe rodzaje
pos	**dobrych**:4	dobre:6
com	lepszych:6	lepsze:8
sup	najlepszych:6	najlepsze:8

Tablica. Zbiór form należących do leksemu od słowa "dobremu", zawężony do form spełniających pożądane cechy (*pl.acc*). Spośród dopuszczalnych, wybieramy formę o najmniejszej symetrycznej różnicy cech morfologicznych (po dwukropku).

Metoda neuronowa



Rysunek. Architektura sieci

Metoda neuronowa

- Sieć jest trenowana bez kontekstu zdaniowego, na danych ze słownika.
- Żeby zapobiec 'odwrotnemu efektowi frekwencyjnemu', przykłady są losowane zgodnie z ich frekwencją.

Fleksja MWE

- Stosujemy analizy zależnościowe zgodne z UD jako substytut dla formalizmu opisującego związki zgody i rządu.
- Zmiany morfologiczne w nadrzędniku, są propagowane do jego podrzędników.
- Propagacja opiera się o reguły akomodacji o postaci:

  $$dep \rightarrow attrs$$

  Gdzie `dep` oznacza etykietę relacji, i `attr` reprezentuje zbiór cech mających podlegać uzgodnieniu. Na przykład:

  $$amod \rightarrow number.case.gender$$

  Reprezentuje uzgodnienie zachodzące pomiędzy modyfikatorem przymiotnikowym i jego nadrzędnikiem.

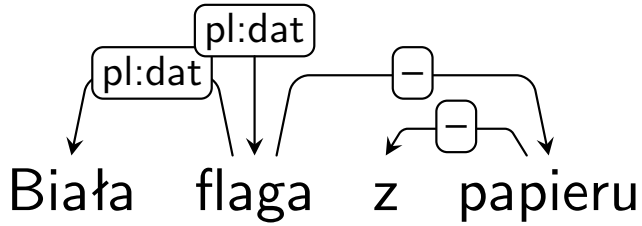
Indukcja reguł

Deprel	number	gender	person	case
amod	**99.78**	**99.81**		**99.48**
appos	94.33	82.57		88.30
conj	83.62	60.57	88.02	**95.97**
det	**99.36**	**98.59**		**98.31**
flat	**98.00**	92.92		92.77
nummod	**99.59**	**95.44**		**97.18**
obj	62.85	20.61	36.59	40.59
obl	65.39	25.34	59.79	11.27

Tablica. Procentowa zgodność, co do danego atrybutu, pomiędzy podrzędnikiem i nadrzędnikiem, o ile łączy je relacja o danej etykiecie. Wyselekcjonowano z PDB.

Propagacja cech

Cechy są propagowane po drzewie, o ile zezwalają na to reguły



$\rightarrow$ Białym flagom z papieru

Lematyzacja MWE

Podobny algorytm może zostać wykorzystany do lematyzacji MWE.

1. Lematyzujemy korzeń frazy
2. Zbieramy cechy morfosyntaktyczne które zostały zmodyfikowane
3. Propagujemy je po drzewie zgodnie z regułami akomodacji

Ewaluacja

Do ewaluacji wykorzystujemy słownik SEJF (Czerepowicka, Monika and Savary, Agata 2018).

	metoda słownikowa	metoda neuronowa
accuracy fleksji	90.54	85.21
accuracy lematyzacji	79.97	78.96

Tablica. Accuracy względem form wzorcowych.

Przykłady

base	target	gold	flexed	correct
giełda towarowa	pl:nom:f	giełdy towarowe	giełdy towarowe	True
mrówka faraona	sg:inst:f	mrówką faraona	mrówką faraoną	False
latający talerz	sg:acc:m3	latający talerz	latający talerz	True
ciąża spożywcza	pl:acc:f	ciąże spożywcze	ciąże spożywcze	True
urwanie głowy	sg:dat:n	urwaniu głowy	urwaniu głowy	True
skała magmowa	pl:voc:f	magmowe skały	skały magmowe	True
niewart złamanego szeląga	sg:voc:m3	niewart złamanego szeląga	niewrry złamanego szeląga	False
dufny w sobie	pl:dat:m2	dufnym w sobie	dufnym w sobie	True
wąskie gardło	pl:dat:n	wąskim gardłom	wąskim gardłom	True
potyczka słowna	pl:nom:f	potyczki słowne	potyczki słowne	True

Ograniczenia

- Zakładamy, że proces odmiany nie wprowadza, ani nie eliminuje nowych słów.
- Konieczne jest dostarczenie kompatybilnych danych: słownika fleksyjnego, oraz treebanku.
- Operujemy w grupie języków fuzyjnych, o przewidywalnych wzorcach afiksacji.
- Dbanie o związki rządu, i zależności leksykalne przerzucamy na użytkownika.

Dziękujemy za uwagę!

Bibliografia I


Czerepowicka, Monika and Savary, Agata (2018). "SEJF - A Grammatical Lexicon of Polish Multiword Expressions". W: *Human Language Technology. Challenges for Computer Science and Linguistics*. Red. Vetulani, Zygmunt and Mariani, Joseph and Kubis, Marek. Cham: Springer International Publishing, s. 59–73. ISBN: 978-3-319-93782-3.


Dusek, Ondrej i Filip Jurcıcek (2019). "Neural Generation for Czech: Data and Baselines". W: *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*. Red. Kees van Deemter, Chenghua Lin i Hiroya Takamura. Association for Computational Linguistics, s. 563–574. DOI: 10.18653/v1/W19-8670. URL: https://aclweb.org/anthology/papers/W/W19/W19-8670/.

Bibliografia II

Klimaszewski, Mateusz i Alina Wróblewska (2021). *COMBO: State-of-the-Art Morphosyntactic Analysis*. arXiv: 2109.05361 [cs.CL].