


# Sieci neuronowe

---

## Translator angielsko-polski *Raport*

Sylwia Nowak

polski	angielski		angielski	polski	Przetłumacz

19 stycznia 2017

## 1. Opis problemu badawczego

Problem mechanicznego tłumacza nie został dotychczas w pełni rozwiązany. W kwietniu 2006 roku opublikowano i oddano do użytku najpopularniejszy automatyczny translator - *Google Translate*. Początkowo był on oparty na Rule-Based machine translation, jednakże po 10 latach istnienia ewoluował on w system oparty na tzw. Statistical machine translation wykorzystujący nie tylko tłumaczenia słownikowe ale również zbiór tłumaczeń użytkowników. System ten codziennie uczy się nowych tłumaczeń. Firma *Google* dzięki swojemu zasięgowi wykorzystuje w swoim systemie naukę na dokumentach ONZ dzięki czemu zwiększa to rzetelność tegoż produktu.

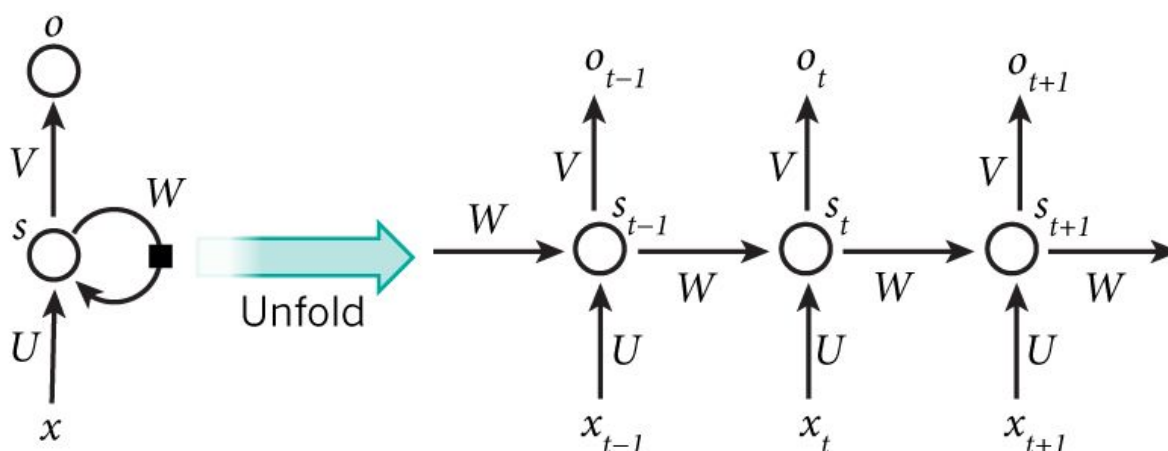
W realizowanym projekcie, w przeciwieństwie do *Google Translate* wspierane jest nie 100 języków a jedynie tłumaczenie z języka angielskiego na polski. Dzięki takiemu ograniczeniu otrzymane wyniki można w prosty sposób zinterpretować ze względu na pełne zrozumienie obu języków i ich gramatyk. System zaprojektowany w celu zrealizowania rozwiązania opisanego problemu w przeciwieństwie do *Google Translate* prezentuje jedną możliwość przetłumaczenia danego słowa, frazy, zdania lub po prostu ciągu słów ograniczonego co do ilości znaków.

## 2. Cel badań

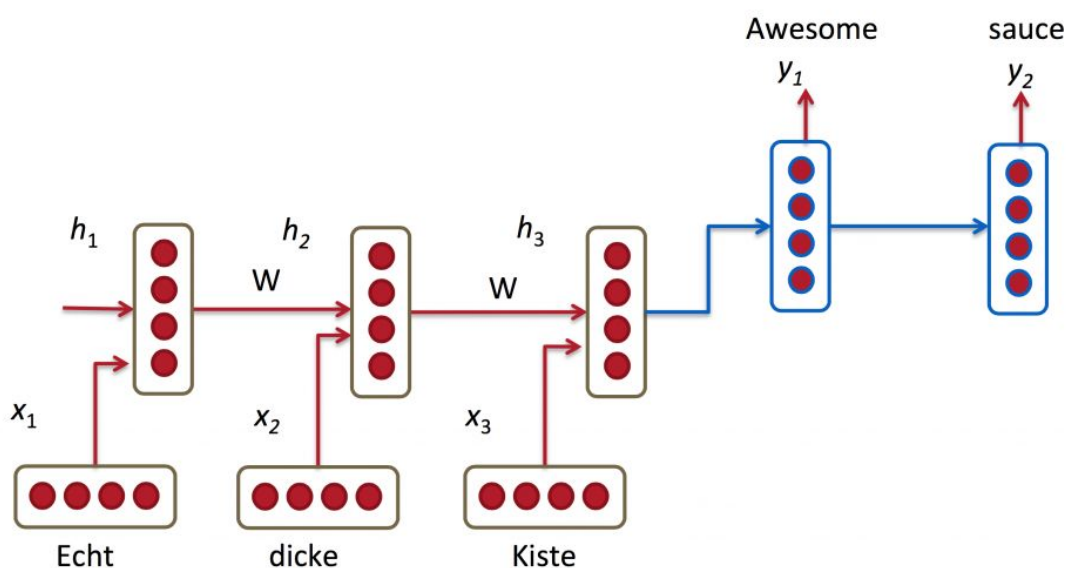
Celem badań jest znalezienie rozwiązania problemu tłumaczenia tekstu z języka angielskiego na polski. Badania pozwolą zbadać dokładność rozwiązania na zaproponowanym zbiorze testowym. Przeprowadzone eksperymenty pomogą wyciągnąć wnioski na temat skuteczności mechanicznego tłumacza podczas tłumaczenia języków o skomplikowanej gramatyce. Przykładowo język polski posiada 3 rodzaje, badania pozwolą sprawdzić jak zaproponowana sieć neuronowa uczy się odmian rodzajów, przypadków itp.

## 3. Wykorzystane techniki

W rozwiązaniu problemu mechanicznego tłumacza zastosowano model sieci neuronowej *Recurrent neural network*. Jest to rodzaj sieci, w której połączenia między neuronami tworzą cykl skierowany. W odróżnieniu od sieci feedforward, RNN są w stanie wykorzystać swoją pamięć wewnętrzną aby przetwarzać arbitralne sekwencje wejściowe. Aby zminimalizować błąd całkowity sieci została użyta metoda stochastycznego spadku wzdłuż gradientu (SGD) będąca algorytmem numerycznym mającym na celu znalezienie minimum lokalnego zbadanej funkcji celu.



A recurrent neural network and the unfolding in time of the computation involved in its forward computation.  
Source: <http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/>



RNN for Machine Translation. Image Source: <http://cs224d.stanford.edu/lectures/CS224d-Lecture8.pdf>

## 4. Opis przykładowych danych

### 4.1. Dane treningowe

Wykorzystana sieć neuronowa rozwiązująca problem mechanicznego tłumacza wykorzysta do swojej nauki dane przygotowane własnoręcznie. W celu prostej weryfikacji przygotowano następujący krótki zbiór treningowy:

```
i ---> ja
i am ---> jestem
i'm ---> jestem
he ---> on
she ---> ona
it ---> to
she's ---> ona jest
is ---> jest
he is ---> on jest
cool ---> mega
super ---> super
she is cool ---> ona jest mega
it's super ---> to jest super
```

## 4.2. Dane testowe

Podczas testów wykorzystano wszystkie dane zbioru treningowego oraz następujące wyrażenia:

```
she is ---> ?
it is ---> ?
he's ---> ?
it's ---> ?
i'm cool ---> ?
he's cool ---> ?
she's cool ---> ?
it's cool ---> ?
he is cool ---> ?
it is cool ---> ?
she's super ---> ?
he's super ---> ?
she is super ---> ?
he is super ---> ?
it is super ---> ?
```

## 5. Weryfikacja rezultatów

Wytrenowana sieć neuronowa na licznych zbiorze danych (> 20000) powinna zostać zweryfikowana w 4 etapach:

- Etap 1 zakłada jedynie weryfikację tłumacza w kontekście prostego słownika (słowo na słowo).
- Etap 2 weryfikuje poprawność fraz np. złożenia 2-3 słów wykorzystywane powszechnie w danym języku.
- Etap 3 sprawdza tłumaczenie pojedynczych zdań. W tym etapie weryfikowane są zdania początkowo względnie krótkie do 20 znaków, następnie do 100 znaków oraz takie na ograniczenie znakowe do 300.

- Dla etapu 1, 2 sprawdzenie odbywa się na zasadzie *jeden do jednego*. Fraza/słowo tłumaczone powinno zwrócić wartość słownikową. Etapy 3 i 4 weryfikowane są przy użyciu człowieka.

## 6. Aplikacja Translator

[illegible]

Podczas trenowania sieci ustawioaa kolejno 2000, 4000 i 10000 iteracji, sieć zbudowana była z 3 warstw (input, hidden, output). Przeprowadzono również test dla sieci z dwiema i trzema warstwami ukrytymi jednakże sieci te znacząco odpowiadały gorzej podczas tłumaczeń niż sieci z 1 warstwą ukrytą. W warstwie ukrytej znajdowało się zawsze 100 neuronów. Podczas jednej iteracji losowane było odpowiednio 8, 10 i 12 danych ze zbioru treningowego i na ich podstawie sieć była uczona. Ze względu na ograniczenie na maksymalną ilość znaków tłumaczonych (30) i ilość znaków w *validCharacters* (oznaczona *N*) każdy znak tłumaczony zamieniany jest na wektor długości  $30 \cdot N$ . Gdzie na jednej pozycji wektora stoi wartość 1 a na pozostałych 0. Trenowana sieć neuronowa ma współczynnik uczenia 0.1, oraz współczynniki backpropagacji odpowiadające za ilość kroków backpropagation równe 10. Podczas zwiększania współczynnika uczenia wyniki otrzymywane podczas testów były zdecydowanie różne od oczekiwanych wyrażeń, przykładowo dla współczynnika 0.5 wynik dla tłumaczenia słowa cool:

```
* = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * =
ENGLISH--->cool
POLISH--->mega
TRANSLATED--->eejj gaessj
* = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * = * =
```

Po zakończeniu trenowania sieci w dostarczonej aplikacji tworzony jest plik *network.txt* zawierający zserializowaną, wytrenowaną sieć neuronową, którą można ponownie wykorzystać podczas testowania tłumaczenia wyrażeń.

## 6.1. Testowanie sieci

```
Your choice:
2
```

```
Enter numbers of tests: 1
Enter text in english to translate (max 30) signs: he is
```

Otrzymana wartość na standardowe wyjście:

```
ENGLISH--->he is
POLISH--->?
TRANSLATED--->on jest
```

Podczas testowania sieci neuronowej, sieć neuronowa wczytywana jest z pliku *network.txt*. Dane testowane zawierają jedynie znaki z *validCharacters*. Na standardowe wyjście wypisywana jest przetłumaczona fraza.

## 7. Otrzymane wyniki

7.1 2000 iteracji podcza treningu, losowanie podczas epoki 8 elementów zbioru treningowego

- Czas trenowania sieci wyniósł ok 5h.
- Testy zbioru treningowego:

ENGLISH--->i  
POLISH--->?  
TRANSLATED--->ja

ENGLISH--->i am  
POLISH--->?  
TRANSLATED--->jaste

ENGLISH--->i'm  
POLISH--->?  
TRANSLATED--->jeste

ENGLISH--->he  
POLISH--->?  
TRANSLATED--->on

ENGLISH--->she  
POLISH--->?  
TRANSLATED--->ona

ENGLISH--->it  
POLISH--->?  
TRANSLATED--->jo

ENGLISH--->she's  
POLISH--->?  
TRANSLATED--->ona jest

ENGLISH--->is  
POLISH--->?  
TRANSLATED--->jest

ENGLISH--->he is  
POLISH--->?  
TRANSLATED--->on jest

ENGLISH--->cool  
POLISH--->?  
TRANSLATED--->mega

ENGLISH--->super  
POLISH--->?

TRANSLATED--->ouper

ENGLISH--->she is cool

POLISH--->?

TRANSLATED--->ona jest ma

ENGLISH--->it's super

POLISH--->?

TRANSLATED--->jo jestpsu

- Testy wyuczonej gramatyki:

ENGLISH--->she is

POLISH--->?

TRANSLATED--->ona jest

ENGLISH--->it is

POLISH--->?

TRANSLATED--->jo jest

ENGLISH--->he's

POLISH--->?

TRANSLATED--->on jest

ENGLISH--->it's

POLISH--->?

TRANSLATED--->jo je

ENGLISH--->i'm cool

POLISH--->?

TRANSLATED--->jaste

ENGLISH--->he's cool

POLISH--->?

TRANSLATED--->on jest a

ENGLISH--->she's cool

POLISH--->?

TRANSLATED--->ona jest a

ENGLISH--->it's cool

POLISH--->?

TRANSLATED--->jo jeeema

ENGLISH--->he is cool

POLISH--->?

TRANSLATED--->on jest ma

ENGLISH--->it is cool

POLISH--->?

TRANSLATED--->jo jest ma



ENGLISH--->she's super  
POLISH--->?  
TRANSLATED--->ona jest su

ENGLISH--->he's super  
POLISH--->?  
TRANSLATED--->on jest su

ENGLISH--->she is super  
POLISH--->?  
TRANSLATED--->ona jest mer

ENGLISH--->he is super  
POLISH--->?  
TRANSLATED--->on jest mer

ENGLISH--->it is super  
POLISH--->?  
TRANSLATED--->jo jest mer

7.2 4000 iteracji podcza treningu, losowanie podczas epoki 10 elementów zbioru treningowego

- Czas trenowania sieci wyniósł ok 1,5 dnia.
- Testy zbioru treningowego:

ENGLISH--->i  
POLISH--->?  
TRANSLATED--->je

ENGLISH--->i am  
POLISH--->?  
TRANSLATED--->jestem

ENGLISH--->i'm  
POLISH--->?  
TRANSLATED--->jest m

ENGLISH--->he  
POLISH--->?  
TRANSLATED--->on

ENGLISH--->she  
POLISH--->?  
TRANSLATED--->ona

ENGLISH--->it

POLISH--->?  
TRANSLATED--->jo

ENGLISH--->she's  
POLISH--->?  
TRANSLATED--->ona jest

ENGLISH--->is  
POLISH--->?  
TRANSLATED--->jest

ENGLISH--->he is  
POLISH--->?  
TRANSLATED--->on jest

ENGLISH--->cool  
POLISH--->?  
TRANSLATED--->mega

ENGLISH--->super  
POLISH--->?  
TRANSLATED--->oupea

ENGLISH--->she is cool  
POLISH--->?  
TRANSLATED--->ona jest mega

ENGLISH--->it's super  
POLISH--->?  
TRANSLATED--->jo jst ma

- Testy wyuczonej gramatyki:

ENGLISH--->she is  
POLISH--->?  
TRANSLATED--->ona jest

ENGLISH--->it is  
POLISH--->?  
TRANSLATED--->jo jest

ENGLISH--->he's  
POLISH--->?  
TRANSLATED--->on jest

ENGLISH--->it's  
POLISH--->?  
TRANSLATED--->jo jst

ENGLISH--->i'm cool  
POLISH--->?

TRANSLATED--->jestem ma

ENGLISH--->he's cool

POLISH--->?

TRANSLATED--->on jestgaga

ENGLISH--->she's cool

POLISH--->?

TRANSLATED--->ona jestgega

ENGLISH--->it's cool

POLISH--->?

TRANSLATED--->jo jss mega

ENGLISH--->he is cool

POLISH--->?

TRANSLATED--->on jest mega

ENGLISH--->it is cool

POLISH--->?

TRANSLATED--->jo jest mega

ENGLISH--->she's super

POLISH--->?

TRANSLATED--->ona jest me

ENGLISH--->he's super

POLISH--->?

TRANSLATED--->on jest me

ENGLISH--->she is super

POLISH--->?

TRANSLATED--->ona jest na

ENGLISH--->he is super

POLISH--->?

TRANSLATED--->on jest na

ENGLISH--->it is super

POLISH--->?

TRANSLATED--->jo jest nr

7.2 10000 iteracji podcza treningu, losowanie podczas epoki 12 elementów zbioru treningowego

- Czas trenowania sieci wyniósł ok 3 dni.
- Testy zbioru treningowego:

ENGLISH--->i

POLISH--->?  
TRANSLATED--->je

ENGLISH--->i am  
POLISH--->?  
TRANSLATED--->jestem

ENGLISH--->i'm  
POLISH--->?  
TRANSLATED--->jestem

ENGLISH--->he  
POLISH--->?  
TRANSLATED--->on

ENGLISH--->she  
POLISH--->?  
TRANSLATED--->ona

ENGLISH--->it  
POLISH--->?  
TRANSLATED--->jo

ENGLISH--->she's  
POLISH--->?  
TRANSLATED--->ona jest

ENGLISH--->is  
POLISH--->?  
TRANSLATED--->jest

ENGLISH--->he is  
POLISH--->?  
TRANSLATED--->on jest

ENGLISH--->cool  
POLISH--->?  
TRANSLATED--->mena

ENGLISH--->super  
POLISH--->?  
TRANSLATED--->ouper

ENGLISH--->she is cool  
POLISH--->?  
TRANSLATED--->ona jest me e

ENGLISH--->it's super  
POLISH--->?  
TRANSLATED--->jo jest supe

- Testy wyuczonej gramatyki:

ENGLISH--->she is

POLISH--->?

TRANSLATED--->ona jest

ENGLISH--->it is

POLISH--->?

TRANSLATED--->jo jest

ENGLISH--->he's

POLISH--->?

TRANSLATED--->on jest

ENGLISH--->it's

POLISH--->?

TRANSLATED--->jo jest

ENGLISH--->i'm cool

POLISH--->?

TRANSLATED--->jestem je

ENGLISH--->he's cool

POLISH--->?

TRANSLATED--->on jest se

ENGLISH--->she's cool

POLISH--->?

TRANSLATED--->ona jest se

ENGLISH--->it's cool

POLISH--->?

TRANSLATED--->jo jest see

ENGLISH--->he is cool

POLISH--->?

TRANSLATED--->on jest me e

ENGLISH--->it is cool

POLISH--->?

TRANSLATED--->jo jest me e

ENGLISH--->she's super

POLISH--->?

TRANSLATED--->ona jest supe

ENGLISH--->he's super

POLISH--->?

TRANSLATED--->on jest supe

ENGLISH--->she is super  
 POLISH--->?  
 TRANSLATED--->ona jest er

ENGLISH--->he is super  
 POLISH--->?  
 TRANSLATED--->on jest er

ENGLISH--->it is super  
 POLISH--->?  
 TRANSLATED--->jo jest er

## 7.4 Porównanie wyników

Parametry sieci	Pełna poprawność dla zbioru treningowego	*Drobne literówki dla zbioru treningowego	**Znaczące błędy dla zbioru treningowego	Pełna poprawność dla zbioru testowego	Drobne literówki dla zbioru testowego	Znaczące błędy dla zbioru testowego
2000 iteracji, 8 losowanych danych	7	5	1	2	9	4
4000 iteracji, 10 losowanych danych	8	4	1	3	10	2
10000 iteracji, 12 losowanych danych	7	6	0	2	11	2

\* Pełna poprawność oznacza, że przetłumaczony tekst w 100% odpowiada oczekiwaniom.

\*Drobne literówki oznaczają błędy w maksymalnie 3 literach (nie wpływają one na zrozumienie w całości tłumaczonego tekstu).

\*\*Znaczące błędy oznaczają błędy w zrozumieniu całości tekstu - nie można przykładowo rozpoznać czy osoba jest super czy mega.

## 7.5 Wnioski

Zwiększenie ilości iteracji z 4 tysięcy na 10 tysięcy nie wpłynęło na poprawę otrzymanych wyników. Sieć neuronowa w każdym przypadku potrafiła chociaż częściowo nauczyć się gramatyki języka angielskiego. Ucząc się na frazie *she's* potrafiła dobrze przetłumaczyć *he's* oraz *it's*. Analogicznie z frazą *he is* i frazami *she is*, *it is*. Podczas tłumaczenia krótkich zdań używających czasownika *to be* oraz słów *cool* i *super* sieć rozumiała w jakiej kolejności powinna ustawiać przetłumaczone już słowa. Problem pojawiał

się najczęściej z dokładnym wypełnieniem słów *mega* oraz *super*. Często zamiast na końcu zdania słowa *mega* pojawiała się jedynie *meg*, *me* a itp. Nie wpływało to jednak na zrozumienie całego zdania. Większy problem pojawiał się ze słowem *super*, które mogło być tłumaczone na *er*, *se* itp. co powodowało często brak zrozumienia tłumaczenia. Zauważalny jest problem dla każdego z trzech przypadków z tłumaczeniem pojedynczego słowa *it*. Wszystkie sieci wyuczyły się, że słów *it* zamiast to oznacza *jo*. Powodowało to błędy podczas testowania nowych dla sieci fraz.

## 8. Rozwój projektu

Przygotowano przy wykorzystaniu api *bab.la* 3 zbiory treningowe danych. Przy pomocy przygotowanego zbioru 3000 najpopularniejszych słów języka angielskiego (zbiór S) wygenerowano na podstawie słownika plik *dictionary.txt* zawierający już przetłumaczone słowa angielskie na język polski. Ze względu swoją popularność słowa z S posiadają wiele tłumaczeń co powoduje, że rozmiar zbioru *dictionary* jest równy 25 000. Dodatkowo pobrano wszystkie frazy powiązane z początkowym zbiorem słów S i zapisano je z ich tłumaczeniami do pliku *phrases.txt*. Rozmiar zbioru danych *phrases* to 21 000. Ponadto pobrano przykłady zdań powiązanych ze zbiorem S i zapisano je wraz z tłumaczeniami do pliku *sentences.txt*. Zbiór danych *sentences* jest rozmiaru 114 000.

Aby umożliwić tłumaczenie prostych zdań angielskich (składających się ze słów ze zbioru S) należałoby stworzyć zbiór treningowy rozmiaru

$$114\ 000 + 25\ 000 + 21\ 000 = 160\ 000.$$

Ponadto ograniczenie na maksymalną ilość znaków tłumaczonych nie powinno być 30 tylko 3000 aby móc wykorzystać wszystkie dane zawarte w zbiorze *sentences*. Dodatkowo podczas jednej iteracji nie powinny być losowane 4 dane ze zbioru treningowego tylko minimum 1000 a ilość iteracji 20 000. Taka konfiguracja zbioru treningowego powoduje, że czas nauki sieci neuronowej wyniósłby ok.  $100 * 20\ 000 * 3000/30 * 2$  sekundy = 4 000 000 000 sekund co daje od 4,6 tys dni. Oszacowania te są prawdziwe przy stosowaniu 1 maszyny do uczenia się oraz braku optymalizacji ze względu na równoległe działanie programu.

## 9. Literatura bazowa

Do skonstruowania rozwiązania problemu wykorzystano następujące pozycje:

- a) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation *Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio*  
<https://arxiv.org/abs/1406.1078>
- b) Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation *Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa,*

*Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean*

<https://arxiv.org/abs/1609.08144>

- c) Sequence to Sequence Learning with Neural Networks *Ilya Sutskever, Oriol Vinyals, Quoc V. Le*

<https://arxiv.org/abs/1409.3215>

- d) Przykłady użycia sieci RNN ze strony <https://deeplearning4j.org/>