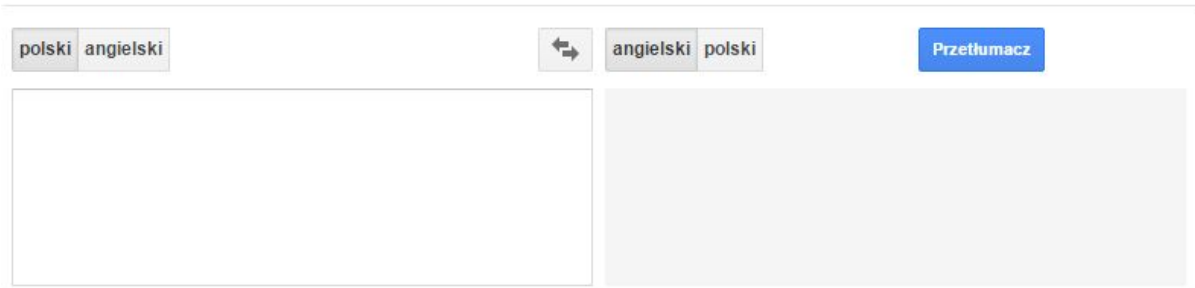
Sieci neuronowe

Translator angielsko-polski
Konspekt

Sylwia Nowak
Radosław Kutkowski



1 listopada 2016

1. Opis problemu badawczego

Problem mechanicznego tłumacza nie został dotychczas w pełni rozwiązany. W kwietniu 2006 roku opublikowano i oddany 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 będzie nie 100 języków a jedynie tłumaczenie z języka angielskiego na polski i odwrotnie. 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 będzie w analogiczny sposób jak *Google Translate* prezentować kilka możliwości przetłumaczenia danego słowa, frazy, zdania lub zdań. Dzięki takiemu rozwiązaniu można prezentować w kolejności od najbardziej prawdopodobnego wyniku do najmniej z zastosowaniem pewnego progu ograniczającego prawdopodobieństwo poprawnego tłumaczenia. W przypadku tłumaczenia zdania bądź zdań (niekoniecznie mających sens logiczny) prezentowany będzie zawsze tylko jeden najlepsz wynik.

2. Cel badań

Celem badań jest znalezienie rozwiązania problemu tłumaczenia tekstu z języka polskiego na angielski i odwrotnie. Badania pozwolą zbadać dokładność rozwiązania na zaproponowanym zbiorze testowym. Dzięki statystykom udostępnionym przez firmę *Google* można będzie porównać otrzymane wyniki z dotychczas istniejącym rozwiązaniem. Przeprowadzone eksperymenty można 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. Literatura bazowa

Do skonstruowania rozwiązania problemu zostaną wykorzystane następujące pozycje:

a) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation *Kyunghyun Cho, Bart van Merrienboer, 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

4. Wykorzystane techniki

W rozwiązaniu problemu mechanicznego tłumacza zastosowany będzie model sieci neuronowej *Recurrent neural network*. Jest to rodzaj sieci, w której połączenia pomiędzy 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 zostanie użyta metoda gradientu prostego będąca algorytmem numerycznym mającym na celu znalezienie minimum lokalnego zbadanej funkcji celu.

5. Opis danych

5.1. Dane treningowe

Wykorzystana sieć neuronowa rozwiązująca problem mechanicznego tłumacza wykorzysta do swojej nauki dane słownikowe. Dzięki udostępnionemu api *Google Translate* zostaną wykorzystane tłumaczenia istniejące już w bazie tego translatora. Biblioteka ta pozwala na tłumaczenie dowolnego tekstu ze wspieranego języka na inny. Do nauki słów i fraz zostanie pobrany słownik podstawowych słów i wyrażeń języka angielskiego i polskiego. Dzięki takiemu podejściu zostanie zbudowany pierwszy zbiór treningowy dla mechanicznego tłumacza. Dodatkowo przeprowadzony będzi eksperyment tłumaczenia zdań z wybranej książki angielskiej i polskiej jedynie przy pomocy sieci nauczonej na słowach i frazach. Weryfikacja poprawności takiego tłumaczenia wykorzysta omówione api *Google*. Eksperyment ten pozwoli pokazać, że sieć ucząca się jedynie na słowach i frazach nie nauczy się nigdy gramatyki danego języka. Następne dane treningowe będą posiadały zdania z wybranej książki przetłumaczona na 2 sposoby:

1. zgodnie z translatorem Google
2. zgodnie z wiedzą autorów rozwiązania.

Sieć nie będzie uczyła się na fragmentach tekstu zawierających więcej niż jedno zdanie.

5.2. Dane testowe

Do danych testowych zostanie wykorzystany słownik, na którym sieć została wyuczona. Dodatkowo wybrane zostaną fragmenty tekstu z tej samej książki. Fragmenty te

będą mogły zawierać pojedyncze słowa, frazy, zbiory słów (niepełne zdania), zdania oraz skończoną liczbę zdań.

6. Weryfikacja rezultatów

Skonstruowane rozwiązanie omawianego problemu będzie weryfikowane w 4 etapach. Dla każdego etapu weryfikacji będzie sprawdzane początkowo tłumaczenie z języka angielskiego na polski, następnie odwrotnie.

- Etap 1 zakłada jedynie weryfikację tłumacza w kontekście prostego słownika.
- 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 będą zdania początkowo względnie krótkie do 5 słów, następnie do 10 słów oraz takie, bez ograniczeń na ilość słów.
- Etap 4 jest weryfikacją tłumaczenia całego tekstu.

Dla etapu 1 i 2 sprawdzenie odbywa się na zasadzie *jeden do jednego.* Zarówno słowo tłumaczone jak i fraza powinny zwrócić wartości słownikowe. Istnieje możliwość tłumaczenia danego słowa bądź frazy na kilka sposób jeżeli słownik, na którym jest nauczona sieć taką możliwość posiada. Etapy 3 i 4 będą weryfikowane na 2 sposoby. Pierwszym krokiem sprawdzającym tłumaczenie będzie porównanie z wynikiem otrzymanym przez *Google Translate*. Kolejnym krokiem będzie weryfikacja przy użyciu człowieka. Tłumaczenie dostanie 2 oceny. Końcowy wynik będzie prezentowany zarówno dla kroku 1 jak i 2.