dr hab. Agnieszka Mykowiecka
Instytut Podstaw Informatyki PAN
Jana Kazimierza 5, Warszawa
Polsko-Japońska Akademia Technik Komputerowych
Koszykowa 86, Warszawa

Warszawa, luty 2016

Recenzja rozprawy doktorskiej

Statistical machine translation of speech enhanced by comparable corpora

Recenzowana rozprawa doktorska mgr. Krzysztofa Wołka dotyczy zagadnień związanych ze statystycznym tłumaczeniem maszynowym z języka polskiego na angielski. Praca powstała pod kierunkiem prof. Krzysztofa Maraska w Polsko-Japońskiej Akademii Technik Komputerowych w Warszawie. W rozprawie postawiono następujące tezy:

1. Odpowiednie metody wykorzystania korpusów porównywalnych pozwolą na wzbogacenie danych treningowych umożliwiające polepszenie jakości systemu tłumaczącego z języka polskiego na angielski.
2. Metody filtrowania tekstu i adaptacji parametrów pozwalają na polepszenie jakości statystycznego systemu tłumaczącego.

Rozprawa napisana jest w języku angielskim i składa się z sześciu rozdziałów poświęconych: informacjom wprowadzającym do tematu (*Introduction*), przedstawieniu ogólnych problemów związanych z tłumaczeniem statystycznym i porównywalnymi korpusami tekstów dwujęzycznych (*Statistical Machine Translation and Comparable Corpora*). Rozdział trzeci to opis aktualnego stanu wiedzy w wybranych dziedzinach (*State of the Art*). Rozdział czwarty stanowi zasadniczą część pracy opisującą wypracowane przez autora pracy metody analizy danych i tworzenia korpusów polsko-angielskich. Osiągnięte rezultaty wynikające z ich analizy oraz rezultaty wytrenowanych na tych danych systemów SMT przedstawione są w rozdziale piątym (*Results and Conclusions*). Krótkie podsumowanie całej pracy przedstawione jest w rozdziale szóstym (*Final Conclusions*). Praca zwiera także spis rysunków (32), tabel (50) oraz obszerną (liczącą 219 pozycji) bibliografię.

W rozdziale pierwszym autor sygnalizuje z jakich narzędzi pozwalających na analizę tekstów w języku naturalnym – polskim czy angielskim – korzystał. Charakteryzuje też problemy jakie stawia przez automatyczną analizą język naturalny. Wspomniane są trudności jakie stwarza swobodny porządek zdania, fleksyjność języka polskiego, niezgodność siatki czasów gramatycznych pomiędzy parami języków, istnienie bądź brak uzgodnień rodzajów podmiotu i czasownika, czy aspektu czasowników. Autor przedstawia też krótko stan prac nad tłumaczeniem tekstów w językach słowiańskich.

W drugiej części pracy autor przedstawia statystyczne podejście do tłumaczenia maszynowego, zbiory danych potrzebne do realizacji tej metody i przyjęte ogólnie zasady ewaluacji wyników. Rozdział ten zawiera opis korpusów równoległych, opis narzędzia pozwalającego na

budowę systemów SMT – Mosses – oraz opis problemów utrudniających analizę zdań w języku naturalnym, a więc i ich tłumaczenie – w szczególności zagadnienia związane z tokenizacją – problem wieloznaczności kropki, wielkości liter, znaków interpunkcyjnych czy słów złożonych. Następnie opisane są modele n-gramowe i metody wykorzystywane do ich wygładzania oraz statystyczny model systemu tłumaczącego opartego o model zaszumionego kanału. Autor opisuje też modele tłumaczeniowe IBM Model 1-6 oraz podejście do tłumaczenia maszynowego oparte o przyporządkowanie fraz. Podrozdział 2.5 poświęcony jest ewaluacji systemów tłumaczących, miarom BLEU, NIST, TER, METEOR i RIBES.

Rozdział 3 („State of the art") rozpoczyna się krótkim przeglądem wyników uzyskiwanych przez systemy tłumaczące dla rożnych par języków. Następnie autor opisuje metody zrównoleglania zdań w korpusach równoległych i porównywalnych.

Rozdział 4 zawiera opis zaproponowanych przez doktoranta metod przetwarzania korpusów równoległych dla języka polskiego. Autor proponuje własny potok narzędzi, które pozwalają uzyskać korpus równoległy z porównywalnych danych zaszumionych, jak na przykład stron Wikipedii w językach polskim i angielskim. Rozdział rozpoczyna się od przedstawienia procedury pozyskania danych w dostępnych zbiorów dwujęzycznych różnego typu. Do wstępnych eksperymentów z metodami zrównoleglania zdań wykorzystano Wikipedię, a do testowania systemów tłumaczących zbiory zawierające teksty wykładów – bazę TED przygotowaną na potrzeby konferencji IWSLT 2014, zbiory z informacjami medycznymi Europejskiej Agencji Leków (EMEA), wystąpienia z parlamentu europejskiego (EUP), zbiór tekstów związanych z turystyka – Basic Travel Expression Corpus (BTEC) i zbiór zawierający ścieżki dialogowe z filmów z witryny OpenSubtitles.org (OPEN). Po charakterystyce tych zbiorów autor przedstawia użyte metody zrównoleglania zebranych danych. Zbiory te zostały zamienione na korpusy równoległe poprzez kolejno zrównoleglenie na poziomie artykułów, a następnie wyszukanie w nich odpowiadających sobie par zdań. Zgromadzone artykuły zostały zapamiętane w bazie danych. Pierwszym filtrem było usunięcie artykułów, dla których nie znaleziono wystarczająco podobnych odpowiedników. Następnie zdania ze sparowanych artykułów zrównoleglone zostały przy wykorzystaniu jednej z dwóch metod. Pierwsza z nich stanowiła zaadaptowaną wersję uniwersalnego schematu z Yalign Tool, która została dostosowana do przetwarzania dużych zbiorów danych poprzez zrównoleglenie i dostosowania do obliczeń z wykorzystaniem procesorów graficznych. Autor zaproponował też metodę doboru wartości parametrów użytej metody pozwalających na uzyskanie lepszych wyników dopasowania zdań. Dobór ten dokonywany jest poprzez porównywanie wyników uzyskanych automatycznie ze specjalnym zbiorem ręcznie opracowanych wylosowanych dokumentów. Wartości parametrów zmieniane są tak, by uzyskać jak największą zgodność z danymi przygotowanymi przez człowieka. Dla artykułów z Wikipedii dodatkowo autor testował metodę zwiększania trafności przypisania zdań poprzez analizę odwołań do pozycji bibliografii. Korzystał też z podpisów pod rysunkami i zdjęciami. Kolejny fragment pracy zawiera opis autorskiej metody zrównoleglania tekstów. Metoda ta dostosowana jest do struktury Wikipedii, której hasła są parowane na podstawie linków i informacji zawartych na stronie z hasłem. Hasła te przetwarzane są następnie programem Hunalign wzbogaconym o opracowane narzędzie pozwalające na dopasowywanie zdań, w których występują uzgodnienia krzyżowe oraz na odfiltrowanie zdań o niskim stopniu odpowiedniości.

W kolejnym podrozdziale Doktorant przedstawia zaproponowane modyfikacje popularnej miary jakości tłumaczenia – BLEU. Modyfikacje polegają na uwzględnieniu pojawienia się w tłumaczeniu synonimów słów, które pojawiły się w tekście oryginalnym (np. przykład użycie 'quiz' zamiast 'exam') oraz zwiększeniu wagi dla dobrego dopasowania słów rzadkich (częściej wnoszących znaczenie). Autor przedstawia dokładną analizę porównawczą różnych miar konkludując, że

zaproponowana miara lepiej oddaje jakość tłumaczenia dla języka polskiego niż oryginalna miara BLEU.

Następnym rozwiązywanym przez Doktoranta problemem jest problem filtrowania uzyskanego zbioru zrównoleglonych zdań tak, by budowany korpus jak najbardziej przypominał korpus równoległy, w którym zdania w parach stanowią rzeczywiście swoje odpowiedniki. Zaproponowana metoda polega na wykorzystaniu automatycznych tłumaczeń zdań i porównywania ich ze zdaniami pochodzącymi z filtrowanego zbioru. Dla ewaluacji metody autor uzyskał też zbiór zdań przejrzany przez tłumaczy oraz zbiory uzyskane przy wykorzystaniu kilku innych dostępnych narzędzi. Wszystkie zestawy zdań użyte zostały do wytrenowania modeli tłumaczących, a te modele zewaluowane przy wykorzystaniu standardowych miar SMT. Modele wytrenowane na danych filtrowanych metodą Doktoranta uzyskiwały wyniki lepsze niż modele uzyskane na pozostałych automatycznie wyczyszczonych zbiorach i nieco poniżej modeli uzyskanych dla danych przygotowanych „ręcznie". Autor przeprowadził bardzo wiele eksperymentów dobierając różne metody wstępnej obróbki danych i różne parametry przy tworzenia modeli, wykorzystywał przy tym zarówno wiele narzędzi do przetwarzania tekstów polskich jak i wiele dostępnych programów do zrównoleglania tekstu na poziomie słów.

W piątej części pracy przedstawione jest analiza porównawcza wyników uzyskanych przy użyciu wytrenowanych systemów SMT. Autor rozpoczyna prezentację rozwiązań od porównania (według różnych miar) systemów trenowanych bezpośrednio na danych oryginalnych, na przetworzonej wersji danych wyczyszczonych z błędnych dopasowań, na tekstach zawierających formy podstawowe słów oraz takich, które zawierają wyłącznie zdania w porządku SVO (podmiot-czasownik-dopełnienie). Dla kolejnych zbiorów testowych Autor podaje też wyniki najlepszego systemu (BEST), nie zawsze jest jednak jasno sprecyzowane, co to jest za system. Autor nie wyjaśnia też w jaki sposób uzyskiwał zdania w porządku SVO, a procedura ta mogła wpłynąć na uzyskiwanie gorszych efektów przy użyciu tak przetworzonych danych jako danych treningowych. Kolejno w pracy przedstawione są wyniki wielu eksperymentów z rożnymi wytrenowanymi modelami, różnice między poszczególnymi systemami nie są jednak zbyt duże i jednoznaczne (różne miary wskazują różne uporządkowania wyników. Ta część pracy potwierdza wykonanie przez doktoranta wielu eksperymentów, ich wyniki są jednak dla czytelnika trudne do śledzenia. Autor podsumowuje jednak te wyniki w rozdziale 5.1.5 co nieco ułatwia ich odbiór.

W drugiej części rozdziału piątego autor zajmuje się analizą wykorzystanych metod uzyskiwania równoległych danych. Eksperymenty obejmują jedno- i dwukierunkowe dopasowywanie zdań z wieloma dodatkowymi wariantami. Ponownie, nie zawsze wiadomo, co konkretny wariant oznacza, np. EXT (tabela 46) – to użycie danych dodatkowych, ale nie wiadomo jednoznacznie o jakie dane tu chodzi. Autor przedstawia też analizę wyników poprawionej wersji algorytmu Yalign wykazując zarówno zwiększenie liczby otrzymanych par zdań, jak i zmniejszenie czasu potrzebnego go ich uzyskania. Dołączenie uzyskanych w ten sposób danych wpłynęło pozytywnie na efekty wytrenowanych przy ich wykorzystaniu systemów SMT. Wyniki działania algorytmów zrównoleglających zostały porównane z wynikami pracy tłumaczy. Z tabeli 60 wynika, że prawie wszystkie te zdania zrównoleglone automatycznie, które były zgodne ze zrównolegleniem dokonanym przez tłumaczy, zostały „odfiltrowane" przez zaproponowane metody. Jeżeli ta moja interpretacja tabeli jest słuszna, zasługiwała na szerzy komentarz niż tylko stwierdzenie, że wyniki wykazują „some correlation with human judgements". W szczególności jakiś przykład ze wskazaniem powodów tego odrzucenia mógłby coś wyjaśnić. To jest też ogólna uwaga do tej części pracy, która dotyczy prezentacji wyników, że chociaż praca dotyczy tłumaczenia tekstów w języku naturalnym, nie zawiera żadnych przykładów osiąganych wyników, a jedynie tabele z wartościami wybranych miar.

Oczywiście, obiektywna ocena tak wielu systemów SMT mogła polegać wyłącznie na automatycznym wyliczeniu wartości uznanych miar jakości, ale one dobrze oddając wzajemne zależności między podobnie działającymi systemami, mało mówią o wartości samego systemu tłumaczącego dla ludzi czytających te tłumaczenia. Kilka przykładów negatywnych i pozytywnych pozwoliłoby wyrobić sobie wrażenie jakie te wyniki są. W ostatniej części tego rozdziału autor opisuje wyniki eksperymentu z dołączeniem zdań uzyskanych w wyniku znajdowania par słów podobnych i podaje przyczyny, dla których, wbrew oczekiwaniom, bezpośrednie dołączenie takich par pogorszyło wyniki.

Ostatni, szósty rozdział podsumowuje osiągnięte w pracy rezultaty. Najlepsze wyniki dla systemu SMT udało się osiągnąć przy wykorzystaniu danych TED przygotowanych na konferencje IWSLT 2014 i 2015. Jako podstawowy wynik pracy prezentowane jest opracowanie metody uzyskiwania danych treningowych dla systemów SMT z niezrównoleglonych danych dwujęzycznych. Na zakończenie autor wspomina o zyskującej sobie ostatnio coraz większe uznanie metodzie automatycznego tłumaczenia wykorzystującej sieci neuronowe, która może umożliwić uzyskiwanie jeszcze lepszych rezultatów.

Uwagi do tekstu rozprawy podzielić można na kilka grup. Po pierwsze, niektóre zawarte w pracy stwierdzenia wykazują niestety braki w wykształceniu lingwistycznym doktoranta. Jakkolwiek zastosowane metody nie zakładają wykorzystania informacji o głębokiej składni języka, to powoływanie się przy poruszaniu kwestii składniowych na skróconą gramatykę języka polskiego dla obcokrajowców czy na artykuł chiński (przy opisie słów z klas zamkniętych, str. 43) nie było najlepszym wyborem źródeł. Parę innych przykładów:

- Na stronie 28 (rozdział 2.20 autor mylnie używa pojęcia *focus*. W zdaniu "*Jane purchased the house*" rola ta przypisana jest czasownikowi "*purchased*". Bez dodatkowych informacji o kontekście czy sposobie wymówienia zdania, nie możemy jasno stwierdzić, który element zdania jest „głównym obiektem zainteresowania", ale opis, który następuje potem sugeruje, że autorowi chodziło nie o „focus", ale o predykat.
- Zdanie „*Objects needed by a verb are known as arguments*" powinno być sformułowane odwrotne, gdyż to obligatoryjne argumenty nazywane są dopełnieniami.
- Str. 17 – zdanie „kość liże pies" nie powinno być przetłumaczone jako „a bone is licking a dog" gdyż słowo pies jest tu w mianowniku, a nie bierniku – o przyporządkowaniu ról decydują tu bezpośrednio przypadki gramatyczne, nie jest potrzebny kontekst zewnętrzny.
- Na stronie 29 jest użyte sformułowanie „*prepositional representations*", które powinno być zastąpione przez "*prepositional phrases*".
- „helping prepositional phrase" (s.32) nie jest używanym terminem lingwistycznym, lepiej napisać "adjunct prepositional phrases' lub "modifying prepositional phrases"
- Jako tłumaczenia 'the" podano zaskakująco: „który", „która", „któremu" (s.42).
- W opisie rezultatów analizy morfologicznej WCRFT (rozdział 4) zamiast 'lemma' czy 'base form' użyte jest słowo 'stem'.

Drugą grupą uchybień są stwierdzenia ogólne nie poparte żadnym uzasadnieniem lub sformułowane zbyt kategorycznie, na przykład:

- „development of SMT systems for Polish has been substantially slower than for other languages" – to prawda stosunku do niektórych par języków, ale na pewno jest bardzo wiele takich, dla których systemy tłumaczące nie rozwijają się szybciej niż dla polskiego.
- "speakers and authors use ambiguity when it is not clear what the right intension is" (s.32) - celowe wykorzystanie niejednoznaczności w celu zamaskowania niewiedzy lub zmylenia

odbiorcy oczywiście się zdarza, ale znacznie częstsze jest nieświadome formułowanie wypowiedzi niejednoznacznych, gdyż jest to immanentna cecha języka naturalnego.

- "the usage of a language model guarantees that the output is fluent." (str 38) – to jest oczywiście cel, jakiemu służyć ma wykorzystanie tego modelu, jednak użycie modeli n-gramowych przybliża nas jedynie do osiągnięcia płynności wypowiedzi, nie zapewnia jej.
- „rules to prepare words to be put in different places and to change their meaning depending on context" – słowa przed wstawieniem do fazy nie są poddawane jakimś zabiegom zmieniającym ich znaczenie,.

Podsumowując ocenę strony formalnej rozprawy, praca napisana jest dobrym językiem, przedstawia obszerny materiał wprowadzający zwierający opis aktualnego stanu prac nad statystycznym podejściem do maszynowego tłumaczenia i dokładny opis opracowanego rozwiązania, czasami jednak zmiany tematu są zbyt zaskakujące, a tok pracy niezbyt usystematyzowany. Uwagi co do formy dotyczą zwłaszcza pierwszej, wprowadzającej części rozprawy. Drobnym utrudnieniem jest też brak dającego się wykryć porządku w obszernej bibliografii.

Postawionym w pracy celem było „spełnienie określonych wymogów tłumaczeniowych i rozwinięcie dwujęzycznej bazy tekstowej poprzez użycie informacji z korpusów porównywalnych". Autor zgromadził kilka zbiorów danych dwujęzycznych, które w przybliżeniu modelują mowę potoczną. Zbiory te zostały następnie zamienione na korpusy równolegle poprzez kolejno zrównoleglenie na poziomie artykułów, a następnie wyszukanie w nich odpowiadających sobie par zdań. Dane te zostały następnie wykorzystane przy trenowaniu systemów SMT. Uzyskane wyniki potwierdziły dobrą jakość przygotowanych zbiorów.

Pierwszą z głównych wartości pracy jest przetestowanie bardzo wielu kombinacji cech czy metod wpływających na postać systemów tłumaczących między językami polskim i angielskim oraz wskazanie tych, które dają najlepsze rezultaty dla konkretnych danych. Istotnym wynikiem jest też potwierdzenie hipotezy, że dla każdej dziedziny tekstów dobór parametrów trzeba przeprowadzać indywidualnie. Drugim, jeszcze istotniejszym, gdyż bardziej uniwersalnym, wynikiem pracy jest opracowanie skutecznej metody uzyskiwania danych równoległych z danych porównywalnych bądź zaszumionych. Dane takie są niezwykle cenne, gdyż zbiory danych równoległych nie są zbyt obszerne, a od ich wielkości (i zgodności tematycznej) zależy jakość budowanych systemów SMT. Dodanie opracowanych danych podwyższyło jakość budowanych systemów, co dowodzi poprawności zastosowanych kryteriów selekcji.

Podsumowując, Autor wykazał w pracy znajomość zarówno literatury jak i dostępnych narzędzi do analizy tekstów i do budowy statystycznych modeli automatycznego tłumaczenia. Wykonał wiele dobrze zaplanowanych eksperymentów zarówno w fazie przygotowania danych jak i trenowania modeli. Wykazał się zarówno sprawnością w łączeniu wielu różnych źródeł informacji jak i znajomością metod ewaluacji tworzonych narzędzi. Uwagi krytyczne wymienione w niniejszej recenzji dotyczą szczegółów i nie są na tyle poważne by obniżyć moją ogólną pozytywną ocenę rozprawy. Oceniam, że rozprawa ta stanowi rozwiązanie oryginalnego zagadnienia naukowego oraz potwierdza ogólną wiedzę kandydata w wybranej dyscyplinie naukowej i wnosi wkład w rozwój bardzo ważnej obecnie dziedziny jaką jest automatyczne tłumaczenie tekstów z jednego języka naturalnego na inny w kontekście prac nad językiem polskim. Uważam zatem, że recenzowana rozprawa doktorska autorstwa mgr. Krzysztofa Wołka „Statistical machine translation of speech enhanced by comparable corpora" spełnia wymogi stawiane pracom doktorskim i stawiam wniosek o przyjęcie rozprawy i o dopuszczenie Doktoranta do dalszych etapów przewodu doktorskiego.

A. Nykowska