Poznań, 4 lutego 2016

prof. UAM dr hab. Krzysztof Jassem,

Uniwersytet im. Adama Mickiewicza w Poznaniu,

Wydział Matematyki i Informatyki

Recenzja pracy doktorskiej Krzysztofa Wolka, pt. „Statistical machine translation of speech enhanced by comparable corpora"

Analiza tytułu pracy

Termin Speech Translation (czy też translation of speech) odnosi się do tłumaczenia mowy i analizuje aspekty takie jak np. wpływ prozodii wypowiedzi na prawidłowe tłumaczenie. Zasobami analizowanymi w tłumaczeniu mowy są korpusy nagrań mowy ludzkiej. Recenzowana praca doktorska analizuje natomiast korpusy tekstów (z których pewną część stanowią stenogramy wypowiedzi), omawia zatem zagadnienia związane z tłumaczeniem tekstu. W mojej opinii bardziej adekwatne do zawartości byłoby zatem pominięcie w tytule wyrazu „speech".

Ocena tez pracy

Autor dysertacji stawia dwie tezy:
Pierwsza z nich to stwierdzenie, że proponowane przez autora metody eksploracji tekstów porównywalnych prowadzą do poszerzenia danych trenujących i w konsekwencji do polepszenia jakości systemu tłumaczenia. Pierwsza część tezy jest oczywista (jeśli do zbioru trenującego dodamy nowe dane, to zbiór trenujący się poszerzy). Druga część tezy wymaga wykazania za pomocą metryki oceny jakości tłumaczenia. Autor wybiera powszechnie stosowane metryki oceny jakości tłumaczenia, jako punkt odniesienia prrzyjmując system tłumaczenia bez zastosowania danych wyekstrahowanych z korpusów porównywalnych. Teza badawcza jest interesująca, gdyż dotąd nie wykonano podobnych eksperymentów dla tłumaczenia automatycznego z lub na język polski.

Druga teza to stwierdzenie, że jakość tłumaczenia automatycznego może być poprawiona poprzez filtrowanie danych trenujących oraz dostrojenie parametrów trenowania. Teza ta jest dość oczywista – filtrowanie danych trenujących ma na celu usunięcie danych o wątpliwej jakości, a parametry trenowania służą właśnie do tego, aby poprzez ich dostosowanie do tłumaczonych języków poprawiać jakość translacji. W mojej opinii celem pracy jest wykazanie, że jakość translacji można poprawić metodami filtrowania i strojenia parametrów **proponowanymi przez Autora**.

Autor wykonał szereg eksperymentów, których celem było zbadanie wpływu na jakość tłumaczenia polsko-angielskiego, różnych operacji pre-processingu (wstępnego przetworzenia) danych trenujących. Eksperymenty te są interesujące, a ich rezultaty mają istotną wartość dla rozwoju systemów tłumaczenia z językiem polskim.

Ocena układu pracy

Układ pracy jest następujący:

W rozdziale pierwszym (wprowadzeniu) autor szkicuje historię rozwoju tłumaczenia automatycznego oraz przedstawia motywację badań w dziedzinie tłumaczenia tekstu z językiem polskim. Można mieć **zastrzeżenie**, że podrozdział 1.1. składa się tylko z jednej sekcji (1.1.1), a w takiej sytuacji wprowadzanie kolejnego poziomu podziału mija się z celem.

Rozdział drugi poświęcony jest teoretycznym podstawom automatycznego tłumaczenia statystycznego. Wprowadzane są zarówno podstawowe pojęcia lingwistyczne, jak i matematyczne podstawy podejścia statystycznego. Omówione są powszechnie stosowane metody ewaluacji jakości tłumaczenia. W części tej znajduje się również charakterystyka systemu Moses, który stanowi platformę umożliwiającą tworzenie translatorów statystycznych. Dziwie mnie omówienie systemu Moses (podrozdział 2.3) przed częścią poświęconą frazowym modelom tłumaczenia (2.4.3), które stanowią bazę teoretyczną systemu.

Rozdział trzeci poświęcony jest omówieniu współczesnego stanu badań w dziedzinie tłumaczenia automatycznego, ze szczególnym naciskiem na badania związane z wykorzystaniem korpusów porównywalnych. Ostatnie zdanie na stronie 81 zaczyna się od słów: „In addition, two other mining methods were investigated". Sądzę, że mowa jest tutaj o rozwiązaniu autorskim, a w tej sytuacji zdanie to powinno znaleźć się w rozdziale czwartym.

Rozdział czwarty poświęcony jest omówieniu rozwiązania autorskiego:

W podrozdziale 4.1. autor szkicuje zastosowane przez siebie techniki pozyskiwania danych równoległych z różnego typu źródeł (korpusów równoległych i korpusów porównywalnych).

W części 4.2 autor opisuje swoje ulepszenie narzędzia Yalign służącego do urównoleglenia korpusów na poziomie zdań. Usprawnienie te dotyczy szybkości działania i zostaje osiągnięte poprzez zastosowanie wielowątkowości.

Podrozdział 4.3. omawia dostrojenie metody Yalign w celu polepszenia jakości dopasowywania. W tym fragmencie znalazło się również – zaburzając moim zdaniem strukturę pracy – omówienie metod pre-processingu danych z korpusów porównywalnych (które nie są częścią dostrajania systemu Yalign).

Podrozdział 4.4. omawia autorskie techniki zastosowane w uzyskiwaniu korpusów równoległych z Wikipedii.

W części 4.5. autor proponuje całkowicie autorską metodę pozyskiwania korpusów równoległych, niezależną od narzędzia Yalign, przyznając jednak, że jest ona dopiero we wstępnej fazie realizacji.

Podrozdział 4.6 poświęcony jest autorskiemu usprawnieniu metody BLEU, służącej do ewaluacji jakości tłumaczenia. W moim odczuciu fragment ten odbiega od głównego nurtu dysertacji.

W części 4.7 autor powraca do głównej tezy pracy, omawiając własne metody pre-processingu danych. Uważam, że fragment ten powinien znaleźć się wcześniej – przed podrozdziałem 4.4 – aby układ pracy odzwierciedlał chronologię przetwarzania danych.

Ponadto w części tej omawia się autorskie technologie filtrowania pozyskanych danych (celem filtrowania jest usunięcie niepoprawnie dopasowanych zdań).

Podrozdziały 4.8 i 4.9 omawiają eksperymenty, których celem było oszacowanie jakości rozwiązania autorskiego. W tym celu wytrenowano system bazowy bez użycia technologii autorskich (opis znajduje się w podrozdziale 4.8), a następnie przeprowadzono eksperymenty z wykorzystaniem metodologii autora (opis w podrozdziale 4.9). Wydaje mi się jednak, że podrozdział 4.9.1. omawia stosowanie metod standardowych – powinien więc być umiejscowiony w podrozdziale 4.8.

W rozdziale piątym omówiono wyniki eksperymentów opisanych w podrozdziale 4.9, i wyciągnięto wnioski z uzyskanych rezultatów.

Pracę kończy podsumowanie, w którym autor stwierdza, że tezy pracy zostały wykazane oraz wskazuje dalsze kierunki badań, które mogą poprawić jakość systemów tłumaczenia automatycznego.

Powyższe omówienie treści poszczególnych części pracy miało na celu wykazanie, że układ treści w recenzowanej dysertacji jest **logiczny**. Kolejność wprowadzanych treści jest w zdecydowanej większości **poprawna** i umożliwia zrozumienie tez autora. Odniesienia do pojęć niewyjaśnionych są **bardzo nieliczne**. Selekcja treści w części teoretycznej jest **zgodna z wymaganiami** dotyczącymi pracy doktorskiej: autor omawia te i tylko te zagadnienia, których poznanie jest niezbędne do merytorycznej oceny rozwiązania autorskiego.

Ocena merytoryczna

Wartość merytoryczną dysertacji ocenię poprzez analizę wartości poszczególnych jej rozdziałów.

Rozdział 1. Wprowadzenie

W pierwszych zdaniach wprowadzenia autor wyjaśnia, że zrealizowany przez niego system tłumaczenia może stać się translatorem mowy, jeśli zostanie połączony z systemami rozpoznawania i syntezy mowy, których implementacja nie leży w zakresie eksperymentu. Potwierdza to więc moje spostrzeżenie (patrz: Analiza tytułu pracy), że dysertacja dotyczy tłumaczenia tekstu a nie mowy.

Autor przedstawia specyfikę tłumaczenia między językami polskim i angielskim, wskazując potencjalne trudności oraz podobne rozwiązania dla pary czesko-angielskiej oraz omawiając istniejące systemy polsko-angielskie.

Podrozdział 1.2 to omówienie zagadnienia tłumaczenia automatycznego. Nieco mylący jest ogólny tytuł tego podrozdziału (Machine translation), gdyż tytuły jego poszczególnych części dotyczą wyłącznie tłumaczenia statystycznego, który jest tylko jednym z kilku paradygmatów tłumaczenia automatycznego. (Szereg najnowszych eksperymentów wskazuje, że w niedalekiej przyszłości lepszą jakość uzyskać mogą systemy oparte na głębokich sieciach neuronowych.)

Podrozdział 1.3 wskazuje na potencjalne zastosowania translatorów automatycznych.

Ocena merytoryczna części wprowadzającej musi być subiektywna, gdyż autor stanął przez trudnym zadaniem selekcji treści spośród olbrzymiej ilości materiałów. Mój wybór faktów z historii statystycznego tłumaczenia automatycznego byłby nieco inny – skupiłbym się na takich faktach historycznych, jak opracowanie modeli IBM, czy powstanie systemów Pharaoh i Moses.

Oceniam wartość merytoryczną wprowadzenia jako **satysfakcjonującą**.

Rodział 2.

Rozdział drugi poświęcony jest podstawom teoretycznym statystycznego tłumaczenia automatycznego oraz korpusom porównywalnym. Rozdział rozpoczyna **czytelne** wyjaśnienie pojęcia korpusów porównywalnych w kontraście do korpusów równoległych (korpusy równoległe to zestawy tekstów dwujęzycznych, w których zdania lub niewielkie kolekcje zdań są wzajemnym tłumaczeniem; korpusy porównywalne to zestawy tekstów o bardzo zbliżonej treści, które tylko fragmentami spełniają warunek wzajemnej odpowiedniości).

Następnie autor **klarownie** wyjaśnia podstawowy wzór statystycznego tłumaczenia statystycznego oparty o koncepcję kanału zaszumionego.

W kolejnym fragmencie omówione są pojęcia lingwistyczne związane z zagadnieniem statystycznego tłumaczenia automatycznego. **Warto zwrócić uwagę** na fragment 2.2.2. (Sentences), który na wielu przykładach objaśnia potencjalne przyczyny powstawania błędów w zadaniu, które na pozór wydaje się banalne: podziale tekstu na jednostki tłumaczenia. **Pouczający** jest również podrozdział 2.2.3, który wskazuje na trudności w pozyskaniu wysokiej jakości korpusów równoległych, pomimo dynamicznie wzrastającej objętości dostępnych danych wielojęzycznych.

Omówienie systemu MOSES (podrozdział 2.3) trochę **odbiega** od mojej intuicji. Oczekiwałbym informacji o samym systemie, gdy tymczasem autor skupia się na wyjaśnieniu wzorów matematycznych leżących u podstaw tłumaczenia statystycznego opartego na frazach (które implementowane jest również w innych systemach).

W części 2.4 **wyczerpująco** omówiono problematykę tokenizacji, przedstawiając wyzwania, które stoją przed systemem automatycznej translacji. Następnie omawiana jest problematyka wyrazów złożonych. Szczególnie istotne z punktu widzenia tematu dysertacji są podrozdziały 2.4.3 i 2.4.4, w których szczegółowo omówiono zagadnienia związane z generowaniem modelu języka oraz modelu translacji. W tej części pracy warto zwrócić uwagę na formalną **poprawność** zapisów matematycznych (z wyjątkiem wzoru 22, gdzie zapis funkcji *bow* zaczyna się z dużej litery) oraz **skrupulatne** wyjaśnianie wszystkich składowych wzorów matematycznych. Jako **uszczerbek** merytoryczny tej części dysertacji zaliczyłbym brak omówienia algorytmu EM (expectation-maximization), którego zastosowanie jest kluczowe w modelach translacji (o algorytmie EM zdawkowo wspomina się dopiero przy omówieniu Modelu 5, podczas gdy jest on stosowany również we wszystkich wcześniejszych modelach). Moje **zastrzeżenia** budzi również dobór przykładu na rysunku 3, który nie w pełni obrazuje krok dopasowania (w podanym przykładzie krok ten nie zmienia kolejności wyrazów w zdaniu). Również przykład na rysunku 4 jest dla mnie **zaskakujący** – zdanie w czasie przyszłym przetłumaczone jest na zdanie w czasie teraźniejszym.
Hierarchiczny model tłumaczenia automatycznego autor stara się wyjaśnić na przykładzie tłumaczenia zdania *Dzwi otwierajqsię szybko* na *The door opens quickly.* Popełnia przy tym **błąd**. Aby spełnić pożądane tłumaczenie, reguła trzecia na str. 60 powinna mieć postać: *X1 otwierają się → X1 opens*.

W części 2.5 autor omawia miary ewaluacji tłumaczenia. Najwięcej miejsca autor poświęca mierze BLEU, co jest zrozumiałe, zważywszy, że w rozwiązaniu autorskim autor proponuje własną

modyfikację tej miary. Omówienie poparte jest kilkoma przykładami, **jasno** wyjaśniającymi genezę wzoru stosowanego w tej mierze.

Reasumując, rozdział drugi w sposób **wyczerpujący** przedstawia podstawy lingwistyczne i matematyczne statystycznego tłumaczenia automatycznego.

Rozdział 3

Rozdział 3 omawia aktualny stan wiedzy z dziedziny statystycznego tłumaczenia automatycznego, koncentrując się na technologiach przydatnych do tłumaczenia z i na język polski.

Podrozdział 3.1. omawia konkursy tłumaczenia automatycznego. Bardziej szczegółowo omówiony jest konkurs IWSLT. Prezentowane są dwa wykresy (Figure 16 i Figure 17), obrazujące progres systemów translacji, odpowiednio pomiędzy latami 2012 a 2013, i pomiędzy 2013 i 2014. **Zaskakujące** dla mnie jest to, że wartości poszczególnych miar dla roku 2013 są różne na wykresie 16. i na wykresie 17.

Podrozdział 3.2. omawia badania związane z eksploracją korpusów porównywalnych w celu pozyskania danych trenujących. Przedstawiony jest algorytm stosowany w narzędziu Yalign oraz stosowane tam metody ograniczenia przestrzeni przeszukiwań: algorytm A* oraz algorytm Needlemana-Wunscha. Obie metody zobrazowane są przykładami oraz grafikami pomocnymi do ich zrozumienia.

Uważam, że rozdział 3. **dobrze** przygotowuje czytelnika do lektury rozdziału 4., w którym na tle współczesnego stanu badań prezentowane jest rozwiązanie autorskie.

Rozdział 4

Rozdział 4. poświęcony jest rozwiązaniu autorskiemu.

W podrozdziale 4.1. omawia się czynności wstępne, których celem jest pozyskanie i przygotowanie korpusów równoległych, tak aby można było uruchomić na nich algorytm Yalign dopasowywania na poziomie zdań. Z mojego punktu widzenia **najbardziej wartościowym** elementem tej części prac jest opracowanie korpusów porównywalnych (np. Wikipedii). Temu działaniu poświęcony jest osobny podrozdział – 4.4.

W podrozdziale 4.2. omawia się autorskie metody usprawnienia metody Yalign. Warto w tym miejscu nadmienić, że nawet w dobie wzrastających mocy obliczeniowych eksperymenty, których celem jest przyspieszenie procesu dopasowywania zdań, są niezwykle **wartościowe**. Wydaje się bowiem, że objętość potencjalnych źródeł translacji, czyli wielojęzycznych korpusów tekstowych, zwiększa się w jeszcze szybszym tempie niż wydajność mocy obliczeniowych.

Podrozdział 4.3. poświęcony jest eksperymentom mającym na celu poprawę jakości dopasowywania poprzez dostrojenie narzędzia Yalign. Autor wykazuje, poprzez ewaluację wyników na kilku korpusach tekstów, że proponowane przez niego rozwiązania **istotnie** poprawiają jakość algorytmu.

W podrozdziale 4.5 autor wprowadza nową metodologię dopasowywania na poziomie zdań, niezależną od narzędzia Yalign. Wyniki są **obiecujące** i mogą w przyszłości, po dopracowaniu, konkurować z obecnie stosowanymi metodami.

Rozdział 4.6 poświęcony jest autorskiemu usprawnieniu metryki BLEU. Nie jestem przekony co do celowości tej części pracy. Jak wspomniałem przy omawianiu układu pracy, odbiega ona od głównego nurtu rozprawy i nie wprowadza rozwiązań nowatorskich. Autor proponuje m.in., aby miara BLEU przyznawała dodatkowe punkty za synonimy. Pomysł ten stosowany jest już w mierze METEOR, również dla języka polskiego.

Autor stara się wykazać przydatność nowej metody za pomocą szeregu eksperymentów, które w mojej opinii są nieprzekonujące. Tabele 9. I 10. oraz przeprowadzony test Wilcoxona wskazują, że różnice między nową miarą, a istniejącymi nie są statystycznie istotne. Dalsze obliczenia wskazują ponadto, że nowa miara jest dobrze skorelowana z już istniejącymi. (Przy omawianiu wyników autor popełnia drobne błędy – na str. 115 pojawiają się w omówieniu inne dane niż w tabeli 17.). W tej sytuacji powstaje pytanie o sensowność jej wprowadzania nowej miary. **Nie zgadzam** się z tezą autora, że eksperymenty wykazują, iż nowa miara jest bardziej wiarygodna niż standardowa miara BLEU. **Nie widzę ponadto podstaw** do stwierdzenia, że miara autorska jest bardziej zbliżona do ewaluacji ludzkiej.

Moim zdaniem prawidłowo przeprowadzone eksperymenty powinny porównać korelację pomiędzy istniejącymi miarami (a w szczególności miarą BLEU) z oceną ludzką a korelacją nowej miary z oceną ludzką. Dopiero istotna różnica na korzyść tej drugiej wartości udowodniłaby tezę autora o przydatności wprowadzenia usprawnień do miary BLEU.

Podrozdział 4.7 omawia autorską metodę filtrowania danych trenujących, której celem jest odrzucenie zdań, które nie są poprawnie dopasowane. Rezultaty wykazują pewną poprawę jakości, trudno jednak zgodzić się z autorem, że poprawa rzędu 0.003 jest „wysoce znacząca", szczególnie że osiągnięto ją kosztem odrzucenia prawie 20% zdań (Tabela 20.).

Podrozdział 4.8 omawia skrótowo wersję bazową systemu, która stanowić będzie odniesienie do oceny rozwiązania autorskiego.

W podrozdziale 4.9 omawia się eksperyment autorski. Część ta rozpoczyna się dość niefortunnie: omówiona zostaje metoda symetryzacji dopasowań. W obrazującym metodę przykładzie (Figure 31 oraz Figure 32) przedstawione są rezultaty przecięcia i sumy dopasowań, ale brakuje danych wejściowych – czyli dopasowań jednostronnych. W omówieniu przykładu znajdują się odniesienia do wyrazów (*wznowienie, adjourned*), które w przykładzie w ogóle nie występują.

Dalsza część rozdziału 4.9 opisuje wartościowe eksperymenty z wykorzystaniem istniejących narzędzi przetwarzania tekstu: Wrcolaw Natural Language Processing Tools oraz lematyzatora z zestawu PSI-Toolkit. W tej części opisano ciekawy eksperyment polegający na wytrenowaniu translatora – pośrednika, między językiem naturalnym (polskim, angielskim) a językiem form bazowych. Pomimo, że ewaluacja wyników nie wykazuje progresu, sam pomysł wykorzystania translatora-pośrednika oceniam wysoko.

Następnie opisuje się eksperymenty polegające na trenowaniu translatorów na istniejących korpusach równoległych. Interesujące wydają mi się eksperymenty oparte na danych z korpusu medycznego EMEA, gdyż porównuje się tam skuteczność stosowania rozmaitych metod przygotowania danych trenujących.

Omawia się również eksperymenty mające na celu przyspieszenie procesu tłumaczenia poprzez zmniejszenie tablic translacji. Z reguły takie eksperymenty, przyspieszając proces translacji, pogarszają jakość. W recenzowanej dysertacji autor nie podaje, w jakim stopniu eksperymenty wpłynęły na jakość tłumaczenia.

Podsumowując, rozdział czwarty **poprawnie** prezentuje eksperymenty autorskie na tle obecnego stanu wiedzy przedstawionego w rozdziale trzecim.

Rozdział piąty

W rozdziale piątym autor dokonuje ewaluacji przeprowadzonych eksperymentów. Autor konkluduje, że zarówno zwiększenie zbioru danych trenujących poprzez zastosowanie korpusów porównywalnych, jak i zastosowanie modyfikacji metody Yalign poprawia jakość translacji. Fakty te **uzasadniają prawdziwość założonych tez**.

Wstępne eksperymenty z użyciem autorskiej metody niezależnej od technologii Yalign na chwilę obecną nie przewyższają obecnego stanu wiedzy, ale dają nadzieję na postęp w przyszłych badaniach.

Język pracy

Należy podkreślić, że praca napisana jest w języku angielskim, który nie jest dla autora językiem rodzimym, a mimo to stosowany jest **bardzo poprawnie**. Znalazłem zaledwie kilka błędów językowych, które wyszczególniam w Dodatku.

Moim zdaniem stosowanie języka angielskiego w dysertacji doktorskiej z informatyki jest bardzo pożądane. Z jednej strony umożliwia to lepsze rozpowszechnienie wyników poza granicę naszego kraju, a z drugiej nie utrudnia lektury polskim informatykom, którzy zazwyczaj korzystają z lektur angielsko-języcznych.

Podsumowanie recenzji

Autor dysertacji postawił tezę, że możliwe jest poprawienie jakości statystycznego tłumaczenia automatycznego poprzez działania w dwóch kierunkach: poszerzenie zbioru trenującego danymi wyekstrahowanymi z korpusów porównywalnych (niekoniecznie równoległych) oraz udoskonalenie istniejących narzędzi dopasowywania zdań i poprawienie metod filtrowania korpusów. Tezy pracy zostały wykazane poprzez przeprowadzenie odpowiednich eksperymentów i pozytywną weryfikację ich rezultatów. Dysertacja prawidłowo opisuje przebieg i wartość eksperymentów na tle obecnego stanu wiedzy.

Stwierdzam zatem, że **recenzowana praca spełnia wymogi stawiane dysertacjom doktorskim.**

Krzysztof Jassem

Dodatek. Wybrane błędy, których korekta może podwyższyć wartość pracy

Przykładowy błąd interpunkcyjny
Jeśli dobrze rozumiem pierwszą tezę pracy, to pomiędzy wyrazami „corpora" i „along" powinien wystąpić przecinek.

Przykładowy błęda składniowy
There is no ... that exists → *There exist no...*(str. 62)

Przykładowy błąd typograficzny
experimtns → experiments (str 10)

Stosowanie nieformalnego stylu
Domain adaptation is also very important. (str. 66). Zdanie nie niesie naukowej wartości.

Besides being an excellent classifier, an SVM ... (str. 83). Uważam, że stosowanie epitetów bez odpowiedniego uzasadnienia nie powinno mieć miejsca w rozprawie naukowej.

Say you are aligning subtitles...(str. 83)

*The same goes for any figures...(*str. 100)

Przykładowe błędy sementyczne
More precision and less recall → Higher precision and lower recall.

Przykładowe błędy logiczne
A lower value (of a threshold of accepting an alignment; KJ) means more (higher; KJ) precision and less (lower; KJ) recall. Zgodnie z moją wiedzą obniżenie progu akceptacji podwyższa pokrycie (recall), obniżając precyzję (precision).

Niespójność stylu
Wszystkie rozdziały rozpoczyna krótkie streszczenie zawartości podane w czasie teraźniejszym. Tylko w rozdziale trzeci streszczenie podano w czasie przyszłym.

Błędy redakcyjne
Brakuje wyróżniania oznaczeń specjalną czcionką, co jest szczególnie kłopotliwe w lekturze, gdy oznaczenia są jednoliterowe (np. na stronie 50 niewyróżnione litery *e, f* oznaczają zdania w języku angielskim i obcym, niewyróżniona litera *a* na stronie 52 pod wzorem 24 oznacza funkcję).

Na stronie 61 występuje znak greckiego epsilona zamiast znaku należenia do zbioru.

Niepoprawne referencje

Referencja do tabeli 23. zamiast do tabeli 21na str. 124

Referencja do sekcji 4.1.2 (strona 128), która nie istnieje.

James