Joanna Świątek
Instytut Slawistyki Polskiej Akademii Nauk

POTENCJAŁ I OGRANICZENIA STATYSTYCZNEGO TŁUMACZENIA MASZYNOWEGO[1]

Streszczenie

Przedmiotem niniejszego artykułu są zarówno możliwości jak i ograniczenia statystycznego tłumaczenia maszynowego w świecie współczesnym. Tekst stanowi syntezę wiedzy o MT (*machine translation*) przywołując jego historię, podstawowe pojęcia i możliwości algorytmów i programów do wykonywania tłumaczeń.

Słowa kluczowe: tłumaczenie, tłumaczenie maszynowe, przekład, statystyka

1. O przekładzie maszynowym

Tłumaczenie maszynowe (nazywane również automatycznym, ang. *machine translation, MT*) jest relatywnie młodą dziedziną językoznawstwa, która nadal budzi wiele kontrowersji, przede wszystkim z uwagi na niską jakość przekładów dokonywanych przez darmowe serwisy umożliwiające tłumaczenie każdego rodzaju tekstów. Chociaż narzędzia do wykonywania takich przekładów są stale udoskonalane, podlegają też nieustannej krytyce z uwagi na ich ograniczenia.

Przeciwnicy serwisów i programów wykonujących tłumaczenia mechaniczne zarzucają im, między innymi, że:

„(...) – komputer nie dysponuje wiedzą pozajęzykową, jeśli nie została wprowadzona do programu,

– komputer nie ma emocjonalnego stosunku do wypowiedzi,

– nie ma wiedzy o sytuacji komunikacyjnej, o konotacjach, informacjach implicytnych, odniesieniach intertekstualnych, o ile nie zostały wcześniej wprowadzone do programu,

– nie przeprowadza wyobrażeniowej konceptualizacji tekstu i jego konkretyzacji,

– jego czynność nie jest intencjonalna, wykonuje zadanie zgodnie z programem, nie wiedząc po co, od kogo i dla kogo,

– nie potrafi przejść na pozycję meta wobec własnej wypowiedzi (ironia, dwuznaczność, kłamstwo),

– nie rejestruje relacji składniowych, jeżeli nie zostały wprowadzone do programu,

– ma trudności w przypadku konieczności interpretacji komunikatu, jego fragmentaryczności, tekstów zawierających błędy oraz w przypadku występowania okazjonalizmów." [Lipiński, 2004, 111]

Tymczasem intensywne prace nad doskonaleniem przekładu maszynowego prowadzą do sukcesywnej eliminacji jego technicznych ograniczeń. Powszechnie stosowane statystyczne tłumaczenie maszynowe jest możliwe, dzięki wykorzystywaniu konsekwentnie

[1] Niniejszy tekst został napisany na podstawie pracy magisterskiej dotyczącej statystycznego przekładu maszynowego z języków słoweńskiego i chorwackiego na język polski. Praca została obroniona w 2014 roku w Instytucie Filologii Słowiańskiej Uniwersytetu Śląskiego.

rozbudowywanych baz danych, definiowanych przez Nilsa J. Nilssona [1980, 3] jako komputerowe systemy, przechowujące ogromne ilości informacji o danym temacie, w sposób umożliwiający udzielenie odpowiedzi na pytania użytkownika na ten temat. W ten sposób funkcjonuje najpopularniejszy serwis oferujący możliwość darmowego przekładu maszynowego dowolnego tekstu – Google Translate.

Należy mieć ponadto świadomość, że systemów translatorskich nie tworzy się z myślą o przekładach artystycznych. Praca nad przekładem artystycznym wymaga szczególnego talentu, w jaki nie da się „wyposażyć" maszyny. Jednocześnie wraz z rosnącym zainteresowaniem i zapotrzebowaniem na tłumaczenia techniczne, medyczne lub prawnicze, jakość tego typu tłumaczeń wykonywanych przez komputer stale się poprawia. Tłumaczenie maszynowe zmniejsza również ryzyko kosztownych opóźnień, co ma duże znaczenie w biznesie. Wykorzystywany do takich celów przekład maszynowy jest wygodnym narzędziem przyspieszającym cały proces tłumaczenia i obróbki tekstu (jego edycji, korekty, formatowania) i wymagającym od człowieka tylko sprawdzenia poprawności efektu końcowego. Dla przykładu, już w latach osiemdziesiątych ubiegłego stulecia, zakłady Perkins Engines, dzięki wykorzystaniu MT, były w stanie zaoszczędzić ok. 4000£ i 15 tygodni na każdym przełożonym dokumencie. [Arnold i in., 1994, 11]

Olgierd Łukasiewicz zauważył we *Wstępie do teorii tłumaczenia*, że *„niewątpliwie stosunkowo najwyższy procent tekstów całkowicie przekładalnych stwierdzimy wśród tekstów handlowych i technicznych"*. [Wojtasiewicz, 1996, 78] W tekstach tego rodzaju nie mamy do czynienia z problemem wieloznaczności, nie występują również idiomy i środki stylistyczne, wymagające odpowiedniej interpretacji w języku docelowym. Krzysztof Jassem uważa, że komputer nie nadaje się do tłumaczenia poezji ani literatury:

„(...) w tych dziedzinach komputer jeszcze długo nie zastąpi człowieka. Co więc pozostaje dla komputera? Każdy zawodowy tłumacz potwierdzi, że jego praca to w 90% tłumaczenie nudnych, artystycznie bezwartościowych tekstów: dokumentów naukowych i technicznych, transakcji i kontraktów, przepisów administracyjnych, instrukcji obsługi, patentów technicznych czy też aktów prawnych. (...) W tłumaczeniu tych dokumentów niewiele jest miejsca na improwizację, czyli na to co tłumacz-humanista kocha najbardziej. Trzeba tłumaczyć wiernie oraz dokładnie – a do tego właśnie komputer nadaje się wręcz doskonale." [Jassem, 2001, 15]

Większość zamawianych przekładów nie reprezentuje żadnej literackiej wartości, a popyt jest na tłumaczenia dokumentacji, umów, przepisów oraz kodeksów, instrukcji obsługi, podręczników akademickich dla uczelni o profilach technicznych, patentów i raportów [Hutchins i Somers, 1992, 2]; innymi słowy – tekstów wymagających dokładności i terminologicznej konsekwencji.

Co więcej, MT w swoim obecnym kształcie nie jest w stanie wykonywać całej pracy tłumacza, może więc tylko *asystować* człowiekowi redukując ilość jego zadań. Z tego względu, tłumaczenie maszynowe należy postrzegać w kategoriach „przekładu jutra", którego celem będzie dokonywanie szybkich i tanich tłumaczeń, niewymagających czasochłonnej pre- lub postedycji.

2. Opozycja człowiek-tłumacz i komputer-tłumacz

Od tłumacza-człowieka oczekuje się pięciu rodzajów wiedzy, umożliwiających mu dokonanie przekładu. Są to: znajomość języka oryginału, znajomość języka docelowego [co umożliwia wytworzenie tekstu w tym języku], umiejętność przetłumaczenia języka oryginału na język docelowy, znajomość oraz rozumienie tekstu oryginalnego oraz znajomość kontekstu kulturowego. [Arnold i in., 1994, 35]

Tymczasem osoba pracująca jako pre- lub postedytor tłumaczenia maszynowego, wcale nie musi posługiwać się językiem oryginału i jednocześnie językiem docelowym.

Cechą tekstów użytkowych jest bardzo ograniczona liczba dopuszczalnych sformułowań – odejście od terminologii traktowane jest jako brak konsekwencji, dlatego rola tłumacza wysoce specjalistycznego tekstu zostaje zredukowana do znalezienia najlepszego [a często jedynego] ekwiwalentu na poziomie słów, a nie kontekstu. Tego rodzaju schematyczne tłumaczenia, komputer jest w stanie wykonywać w sposób akceptowalny. Trudności mogą się pojawić wraz z pojawieniem się wieloznaczności.

Słowa miewają zazwyczaj więcej niż jedno znaczenie i wybór odpowiedniego jest uzależniony od kontekstu. Nawet jeżeli założymy tylko dwa znaczenia dla każdego wyrazu w danym komputerowi do przełożenia zdaniu, to z każdym kolejnym słowem wybór odpowiednich znaczeń będzie się zwiększał parokrotnie. Na przykład w zdaniu liczącym tylko dwa słowa (i zakładając dwa możliwe znaczenia dla każdego z nich), uzyskamy cztery możliwe, i co ważniejsze: równoważne rozwiązania i tylko 25% szans, że komputer wybierze prawidłowe[2]. Przy trzech słowach, potencjalnych wyników będzie już osiem. Zdanie złożone z dziesięciu takich słów, będzie posiadać już dwa tysiące czterdzieści osiem rozwiązań (2^{11}), a i to przy optymistycznym założeniu, że komputer będzie znał tylko dwie gramatycznie poprawne możliwości uszeregowania tych wyrazów w zdaniu. Co więcej, w ostatecznym rozrachunku tylko jedna z tych dwóch tysięcy czterdziestu ośmiu opcji może okazać się poprawna, w sensie porównywalna pod względem jakościowym z tłumaczeniem dostarczonym przez profesjonalnego tłumacza.

Patrząc na słowo X wyrwane z kontekstu, prawdopodobieństwo odgadnięcia sensu, w jakim zostało użyte jest bardzo niskie i wzrasta proporcjonalnie do N słów, pojawiających się z prawej i lewej strony tego wyrazu. Pytanie brzmi, jaka wartość N, przynajmniej dla większości przypadków, jest wystarczająca do odgadnięcia prawidłowego znaczenia X. Odpowiedź uzależniona jest od rodzaju tekstu z którym mamy do czynienia. Dla bardzo ścisłej terminologii używanej w pracach z zakresu chemii, fizyki lub inżynierii, N może być równe zero, jednak dla poezji będzie już wyższe, choć jego oszacowanie jest praktycznie niemożliwe. Z tego powodu, już w latach czterdziestych profesor Erwin Reifler zasugerował, żeby w procesie pre-edycji, osoba odpowiedzialna za przygotowanie tekstu do przekładu maszynowego, zastąpiła niejednoznaczne wyrazy bardziej jednoznacznymi synonimami. Osoba wykonująca taką pracę powinna biegle posługiwać się językiem oryginału, jednak nie musiałaby znać języka przekładu.

Język docelowy w takim ujęciu staje się przestrzenią probabilistyczną, co znaczy, że w ramach tego języka możemy ocenić, czy dany wyraz jest ekwiwalentny dla wyrazu w języku pierwotnym. Pozwala to określić procentowo, która z dostępnych możliwości jest statystycznie najbardziej prawdopodobna, co z kolei stanowi podstawy działania systemów opartych na statystyce. Eliminuje to ryzyko wystąpienia sytuacji, gdzie wszystkie opcje będą równie prawdopodobne. Dla przykładu, angielskie słowo *play,* można przetłumaczyć na język polski jako *grać* lub *bawić się.* Statystycznie obie możliwości są mało prawdopodobne, jednak jeżeli w bezpośrednim sąsiedztwie *play* system wykryje słowo oznaczające jakiś instrument muzyczny, uzna *grać* za bardziej prawdopodobne. Ponieważ niezwykle rzadko mamy do czynienia z relacją 1:1, w wielu przypadkach niezbędny jest kontekst. Komputer może dokonać selekcji, wybierając statystycznie najczęściej używany odpowiednik danego słowa albo najpierw rozpoznać typ i tematykę tekstu. W tym drugim przypadku, najpierw dokona „skanowania" treści i wybierze wyrazy posiadające tylko jedno znaczenie [np. terminologia medyczna], po czym na tej podstawie „odgadnie" kontekst. Jest w końcu jeszcze jedno kryterium wyboru ekwiwalentu – powiązany z prawdopodobieństwem koszt słowa; komputer będzie chętniej wybierać słowa krótsze i prostsze, niż rzadko spotykane i bardziej wyrafinowane.

[2] Przez „prawidłowe" rozumiem w tym miejscu takie tłumaczenie, które wykonałby tłumacz-człowiek.

Trzeba również pamiętać, że aby „wychwycić sens" wypowiedzi, komputer może wykorzystać składnię i gramatykę danego języka. Zasadniczo przekład odbywa się w trzech etapach – pierwszym jest analiza tekstu w języku oryginału, następnie następuje konwersja na język docelowy i w końcu rekonstrukcja w języku docelowym. Algorytm wykorzystywany przez system tłumaczenia maszynowego jest przeważnie używany przez zwyczajne, pomyślane również o realizowaniu innych zadań komputery. Uzyskany tekst może być później edytowany przez człowieka, którego zadaniem będzie usunięcie błędów i nieścisłości w tłumaczeniu. W większości przypadków, udział korektora jest konieczny do uzyskania przekładu wysokiej jakości.

Pozostaje jeszcze do rozważenia czysto techniczna kwestia, skąd komputer *wie*, gdzie w dłuższym tekście kończy się jedno słowo a zaczyna następne. W językach naturalnych, koniec podstawowego segmentu sygnalizuje się zazwyczaj spacją lub znakiem interpunkcyjnym (przecinek, kropka, wykrzyknik, pytajnik, średnik), jednak komputer powinien być w stanie rozpoznać więcej typów elementów tekstów (czyli segmentów lub tokenów), do których, powtarzając za Agnieszką Mykowiecką należą:

„– ciąg małych liter poprzedzonych wielką literą, np. Kraków,
– ciąg składający się tylko z wielkich liter, np. PZU,
– ciąg małych liter, np. dom,
– ciąg liter małych i wielkich, np. PeKaO,
– ciąg cyfr,
– ciąg cyfr z wewnętrzną kropką lub przecinkiem,
– znak interpunkcyjny.
– data, godzina, adres, numer telefonu,
– adres e-mail, adresy stron www, tagi języka HTML,
– wzory cząsteczek związków chemicznych." [Mykowiecka, 2007, 65]

Ponadto, elektroniczny system tłumaczeniowy powinien być w stanie wychwycić nie tylko nazwiska ale również tytuły znanych na całym świecie arcydzieł, np. *Mona Lisa* lub *Narodziny Wenus,* których nie można tłumaczyć w sposób dowolny. Systemy oparte na statystyce są w stanie wychwycić takie segmenty, podczas gdy systemy oparte na regułach mogą je przełożyć inaczej, o ile wszystkie tego typu wyjątki nie zostały umieszczone w ich bazie danych.

3. Wybrane problemy z zakresu tłumaczenia maszynowego

Jedną z najbardziej rozpowszechnionych metod tłumaczenia automatycznego, jest statystyczne tłumaczenie maszynowe, działające w oparciu o określone modele statystyczne (jedna para językowa w ramach jednego korpusu]. Z metody tej korzysta chociażby Google Translate, wybierając najbardziej prawdopodobny przekład na podstawie olbrzymiej ilości tekstów dających pewne wyobrażenie na temat kształtu konkretnego języka.

U podstaw tłumaczenia statystycznego leży teoria informacji, a także pojęcia informacji opisujących i identyfikujących, natomiast:

„Sam proces przekładu polega na wybraniu statycznie najlepszej (najbardziej prawdopodobnej) wersji docelowej informacji podanej w języku źródłowym, a także najbardziej prawdopodobnej kombinacji słów w języku docelowym. Statystyka zastępuje tutaj reguły językowe, nie ma zatem konieczności opracowywania skomplikowanych gramatyk czy leksykonów." [Bogucki, 2009, 34]

Teoria informacji to dyscyplina nauki zajmująca się przetwarzaniem oraz transmisją komunikatów, a z jej osiągnięć czerpią przede wszystkim informatyka i telekomunikacja. Tłumaczenie maszynowe, polegając na wygenerowaniu tego samego komunikatu w innym języku, również korzysta z założeń tej teorii.

Przede wszystkim, dla każdej asocjacji istnieją dwie pary komunikatów (z języka A na język B, oraz z języka B na język A), co oznacza konieczność wyróżnienia dwóch różnych transformacji oznaczanych jako T_{ab} oraz $T_{ba.}$ Komunikat poddany transformacji nazywamy komunikatem pierwotnym, a komunikat otrzymany w wyniku transformacji komunikatu pierwotnego to komunikat wtórny. Transformacją z kolei nazywamy proces, któremu poddajemy jeden komunikat asocjacji, by uzyskać drugi komunikat tej samej asocjacji. [Mazur, 1970, 42] Z formalnego punktu widzenia nie ma znaczenia, czy po wykonaniu transformacji komunikat pierwotny przestaje istnieć zastąpiony przez komunikat wtórny, aczkolwiek utrata komunikatu pierwotnego może spowodować utratę informacji do powtórnego przetworzenia, jeżeli komunikat wtórny okaże się nieczytelny. [Brillouin, 1969, 54-55]

W teorii informacji wyróżniamy kilka rodzajów i podziałów transformacji, z czego za najbardziej podstawowe uchodzi rozróżnienie transformacji banalnej od niebanalnej. W wyniku transformacji banalnej uzyskujemy komunikat wtórny identyczny z komunikatem pierwotnym, podczas kiedy transformacja niebanalna zakłada różnice między komunikatami jednej asocjacji. W tym rozumieniu, przekład stanowi transformację niebanalną, ponieważ nie tłumaczymy niepowiązanych ze sobą słów, ale dłuższe fragmenty tekstu [zazwyczaj zdania], a więc chociażby odmieniamy odmienne części mowy i układamy je w odpowiedniej dla języka docelowego kolejności. [Mazur, 1970, 43] Transformacja jest odwrotna do transformacji komunikatu pierwotnego w komunikat wtórny wtedy i tylko wtedy, jeżeli wynikiem transformacji komunikatu wtórnego jest komunikat pierwotny. [Ibidem, 44] W tłumaczeniu maszynowym oznaczałoby to, że najpierw przetłumaczylibyśmy jakiś tekst z języka A na B, a następnie w drugą stronę z B na A uzyskując dokładnie tę samą treść, jaką wprowadziliśmy oryginalnie do komputera.

W przypadku tłumaczenia maszynowego posługujemy się pojęciem transformacji asocjacyjnej, której *„zastosowanie do komunikatu pierwotnego asocjacji daje w wyniku komunikat wtórny tej asocjacji”*. [Ibidem, 49] Oznacza to wyznaczenie pary dwóch komunikatów, a taki warunek spełnia np. dowolne hasło słownika dwujęzycznego, więc w tym rozumieniu, słownik dwujęzyczny staje się kodem asocjacyjnym, którego wadą jest bezsilność wobec przypadków spoza zbioru – nawet najbardziej obszerny słownik nie umożliwi nam przetłumaczenia wyrazu, niefigurującego w tym słowniku. [Ibidem, 64] Choć przyporządkowanie na poziomie liter jest możliwe, przyporządkowanie na poziomie słów [ale również idiomów lub przeważnie występujących razem innych grup wyrazów] jest bardziej efektywne, ponieważ transliteracja nie jest tym samym co przekład. Ewentualnym wyjątkiem od reguły może być transliteracja niektórych nazw własnych z języka polskiego na przykład na język rosyjski, a więc cyrylicę. Słowo w takim układzie jest nośnikiem informacji, a w przypadku tekstu pisanego, informacja ta jest kodowana przy użyciu liter. [Brillouin 1969, 74]

Alternatywą dla kodu asocjacyjnego (dwóch odpowiadających sobie komunikatów w jednym zbiorze) jest kod zasadniczy, inaczej reguła. Przykładem kodu zasadniczego może być np. wzór regularnego stopniowania przymiotników, odmiana czasownika przez osoby lub rzeczownika przez przypadki. Zakłada się, że *„im większy jest zbiór kodów asocjacyjnych, tym korzystniej jest zastąpić go kodem zasadniczym, oraz że im bardziej skomplikowany jest kod zasadniczy, tym korzystniej jest zastąpić go zbiorem kodów asocjacyjnych”*. [Mazur, 1970, 66] Innymi słowy, kod zasadniczy powinien objąć swoim zasięgiem wszystkie regularności, natomiast wyjątki bezpieczniej jest wyróżnić w formie kodów asocjacyjnych. Kod zasadniczy częściej niźli w tłumaczeniu statystycznym, stosuje się w tworzeniu systemów wykorzystujących regułę.

W przestrzeni obsługiwanej przez kod zasadniczy, operacje na płaszczyźnie liter są o wiele istotniejsze aniżeli w przypadku kodu asocjacyjnego. Litery biorą bowiem udział w procesach słowotwórczych, a te nie zawsze są regularne. Jeżeli na kod zasadniczy nie zostaną nałożone żadne ograniczenia dotyczące ciągu dopuszczalnych symboli, możemy uzy-

skać zbitki liter niemożliwe do wymówienia lub całkowicie pozbawione sensu. [Brillouin, 1969, 57] Pomocna na tym etapie jest statystyka, która pozwala wyeliminować niektóre ciągi liter jako nierealne, ponieważ *„w przypadku wystąpienia określonej litery prawdopodobieństwo tego, że po niej nastąpi inna, określona litera, nie jest równe prawdopodobieństwu a priori pojawienia się tej litery"*. [Ibidem, 47] Eliminuje to szanse wystąpienia nielogicznych ciągów samogłosek lub spółgłosek. Tak skonstruowany system zareaguje również na błędy ortograficzne wynikające ze zignorowania reguły, np. umiejscowienia polskiej litery *ó* przed końcówką *-je*.

Jednym z największych problemów z jakim można się spotkać adaptując teorię informacji na potrzeby tłumaczenia maszynowego jest zagadnienie pseudoinformowania, czyli po prostu synonimów, które wykorzystuje się aby uniknąć powtórzeń i poprawić styl wypowiedzi. Pseudoinformowaniem nazywamy sytuację, gdy *„autor jakiegoś artykułu posługuje się w różnych zdaniach wyrazami: fabryka, wytwórnia, zakład wytwórczy, zakład produkcyjny, zakład przemysłowy, wszystkie one bowiem mają to samo znaczenie"*. [Mazur, 1970, 120] Z jednej strony wzbogaca to przekaz, z drugiej może być mylące, szczególnie, jeżeli któreś z użytych w ten sposób synonimów nie ma swojego odpowiednika w języku docelowym.

Kolejnym problemem jest niemożliwa do przewidzenia treść potencjalnie przepuszczonego przez system tłumaczeniowy komunikatu, dlatego zakłada się, że jest to jakaś wielkość losowa. Potencjalnie nadana informacja jest wielkością nieznaną, ponieważ nawet jeśli określony został słownik i zestaw reguł, nie można przewidzieć, jaka kombinacja zostanie poddana transformacji. [Sobczak, 1984, 9-10]

W praktyce, początkowo tłumaczenie statystyczne odbywało się na poziomie wyrazów. Obecnie możliwe jest tłumaczenie dłuższych fraz, aczkolwiek zazwyczaj jest to przekład daleki od doskonałego. Należy jednak podkreślić, że systemy statystyczne nie tłumaczą zdań w takim sensie, w jaki „widzą" je ludzie, ale rozczłonkowują je na krótsze frazy, rozpoznane i wydobyte z korpusu – to znaczy wcześniej użyte w jak największej ilości tekstów. Ponadto, do tak skonstruowanych systemów przeważnie nie wprowadza się manualnie reguł językowych [wyjątek stanowią systemy hybrydowe], więc modelem nie kieruje żadna logika za wyjątkiem reguł statystycznych, gdzie wszystko rozbija się o dokonanie wyboru o najwyższym wskaźniku prawdopodobieństwa. Wprowadza to monumentalne ograniczenia na poziomie idiomów, metafor czy chociażby homonimów. Co więcej, w konsekwencji braku jakiegokolwiek paradygmatu, baza danych „nie dostrzega" związku między odmianami jednego wyrazu, co niesie ze sobą ryzyko nie przetłumaczenia nie znalezionego w systemie ciągu znaków będącego zaledwie wariantem. Te niedopatrzenia tworzą pole do popisu dla *rule-based machine translations* (RBMT).

W przeciwieństwie do systemu opartego na statystyce, czyli korpusie *(corpus-based translation*), RBMT funkcjonuje w oparciu o szereg paradygmatów i, *nomen-omen*, reguł, przeważnie gramatycznych. O ile więc CBMT czerpie z żywego języka i dokonuje wyboru w oparciu o częstotliwość występowania konkretnego słowa, tak RBMT przypomina w działaniu sztuczną inteligencję i kieruje się wyłącznie logiką. W tym miejscu należy zadać sobie pytanie, czy ludzka zdolność do dokonywania tłumaczenia może być porównywalna do umiejętności gry w szachy, ponieważ najlepsze, wygrywające z mistrzami systemy szachowe, nie postępują według wgranych do ich pamięci strategii ale wykorzystują historię zarejestrowanych gier. Jeżeli więc systemy oparte na statystyce okażą się wystarczająco efektywne, dalsze, czaso- i pracochłonne badania języka przez pryzmat MT będą zbędne, ponieważ wdrożenie złożonych, bez porównania bardziej skomplikowanych RBMT, wymaga większych nakładów niż CBMT.

RBMT wykorzystują znane z teorii informacji kody zasadnicze albo reguły gramatyki generatywnej. Gramatyka generatywna *„musi przyporządkować każdemu z nieskończonego szeregu zdań opis strukturalny, wskazujący, jak owo zdanie rozumie idealny użytkownik języ-*

ka". [Chomsky, 1982, 17] W ramach gramatyki generatywnej można ponadto wyróżnić zdania „bardziej dopuszczalne", tj. *„bardziej składne, łatwiej zrozumiałe i w pewnym sensie bardziej naturalne oraz te, których prawdopodobieństwo utworzenia jest większe"*. [Ibidem, 24]

Celem gramatyki generatywnej jest stworzenie systemu powtarzalnych reguł, umożliwiających wygenerowanie nieskończenie wielkiej ilości komunikatów. Podstawą do stworzenia takiego systemu jest odpowiednio obszerny słownik.

„Każde hasło słownika winno wyszczególniać:

a) aspekty struktury fonetycznej nie dające się przewidzieć na mocy ogólnej reguły (...);

b) własności istotne z punktu widzenia funkcjonowania reguł transformacyjnych (...);

c) własności istotne z punktu widzenia interpretacji semantycznej formatywu (...);

d) cechy leksykalne wskazujące w rządku przedterminalnym miejsca, gdzie można wprowadzić (...) formatyw leksykalny.

Słowem, każde hasło słownika winno zawierać zarówno informacje wymagane przez dział fonologiczny i semantyczny gramatyki oraz przez składnik transformacyjny, działu syntaktycznego, jak również informacje określające właściwe umieszczanie haseł słownika w zdaniach." [Ibidem, 33]

Tak skonstruowany, niezwykle obszerny i jednocześnie precyzyjny słownik, powinien być podstawą dla funkcjonowania systemu opartego na regułach. Oczywiście stworzenie takiego słownika jest o wiele bardziej pracochłonne niż wykorzystanie samouczących się systemów opartych na statystyce.

4. Ocena jakości przekładu maszynowego

Wyróżniamy kilka różnych metod oceny tłumaczenia automatycznego, a ich podstawowy podział obejmuje ocenę dokonywaną przez człowieka i ocenę automatyczną. Ocena wykonywana przez człowieka przeważnie polega na przepuszczeniu przez ten sam system najpierw tekstu źródłowego na język docelowy, a później z docelowego na źródłowy (tzw. *round-trip translation*). Wadą tego procesu jest testowanie nie jednego, ale dwóch systemów – para języków A – B nie jest tym samym co B – A. Metodę „tam i z powrotem" rzadko traktuje się jako poważną metodę sprawdzania jakości, ponieważ uzyskane wyniki zazwyczaj są śmieszne i pozbawione sensu.

Jedną z najczęściej wykorzystywanych metod ewaluacji przekładu maszynowego drogą automatyczną jest algorytm BLEU (*Bilingual Evaluation Understudy*). Punktem odniesienia dla BLEU jest tłumaczenie wykonane przez ludzkiego tłumacza w tym sensie, że im przekład maszynowy jest mu bliższy, tym ostateczna ocena wyższa, przy czym pod uwagę nie są brane zrozumiałość oraz poprawność gramatyczna. Tekst oryginału jest również całkowicie bez znaczenia.

Działanie algorytmu ilustruje poniższy przykład:

Zdanie w języku oryginału: *Ivan never had a dog.*
Zdanie referencyjne 1 wykonane przez człowieka: *Iwan nigdy nie miał psa*

Zdanie w języku polskim jest naszym pierwszym zdaniem referencyjnym. Aby uzyskać drugie zdanie referencyjne, potrzebujemy drugiego przekładu tego samego zdania na język polski, również wykonane przez człowieka (zdania referencyjne mogą ale nie muszą być identyczne).

Zdanie referencyjne 2 wykonane przez człowieka: *Iwan nigdy nie posiadał psa*

Kandydatem do oceny przez algorytm BLEU jest przekład zdania „Ivan never had a dog" na język polski przez Google Translate, wykonany dnia 12 marca 2015 roku. Wygenerowany przez system przekład brzmi:

Zdanie kandydat: *Iwan nigdy nie miał psa.*

Kandydat	**Iwan**	**nigdy**	**nie**	miał	**Psa**
Ref. 1	**Iwan**	**nigdy**	**nie**	**miał**	**Psa**
Ref. 2	**Iwan**	**nigdy**	**nie**	posiadał	**Psa**

Precyzję tłumaczenia obliczamy z następującego wzoru:

$$P = m / w_t$$

gdzie *m* jest liczbą słów tekstu kandydata znalezionych w referencji, a w_t liczbą wszystkich słów w referencji. Im bliższy 1 jest wynik, tym wyższa ocena tłumaczenia. Dla każdego wyrazu w tłumaczeniu kandydującym, algorytm bierze maksymalną całkowitą liczbę wystąpień w tłumaczeniach referencyjnych. Pozycja słowa w zdaniu nie ma znaczenia. Za maksymalną referencyjną liczbę danego słowa uznajemy największą ilość występowania tego słowa w którejkolwiek referencji. Dla tego przykładu uzyskany rezultat wynosi:

$$P = 1/5 + 1/5 + 1/5 + 1/5 + 1/5 = 5/5 = 1$$

Tak wysoki wynik oznacza, że zdanie zostało przełożone perfekcyjnie. Szybko zauważono jednak zasadniczą wadę BLEU – krótkie fragmenty kandydujące mogły uzyskać absurdalnie wysokie wyniki. Z tego powodu wprowadzona została kara zwięzłości (ang. *brevity penalty*). Jeżeli liczbę słów w korpusie referencyjnym oznaczymy przez *r,* a liczbę słów tłumaczenia kandydującego przez *c*, to jeśli *c* będzie większe lub równe[3] *r,* stosuje się karę zwięzłości wyrażoną przez $e^{(1-r/c)}$. Jeżeli zdania referencyjne są różnej długości, za *r* przyjmuje się zdanie o długości najbardziej zbliżonej do długości zdania kandydującego.

Chociaż komputer nie jest w stanie zastąpić człowieka w tłumaczeniu każdego rodzaju tekstów, tłumaczenie maszynowe jest ciągle rozwijającą się dziedziną nauki, z której dobrodziejstw korzystamy coraz częściej. W sytuacji, w której znajdujemy napisany w niezrozumiałym dla nas języku artykuł, co do którego istnieje podejrzenie, że może zawierać interesujące nas informacje, wgranie go do najpopularniejszego darmowego tłumacza Google jest na ten moment najtańszą i najszybszą metodą umożliwiającą nam stwierdzenie, czy ewentualnie nie bylibyśmy zainteresowani lepszym jakościowo przekładem tego tekstu. Rozumując w ten sposób, komputer jest narzędziem w rękach człowieka, a nie jego „konkurentem" na rynku pracy. Istnieje jednak przypuszczenie, że w kolejnych latach systemy wykonujące tłumaczenia automatyczne będą coraz bardziej zaawansowane i zdolne przyspieszać realizacje zamówień na określony rodzaj przekładów. Przyszłość tłumaczenia maszynowego maluje się w jasnych barwach.

LITERATURA

Arnold Douglas i in. 1994. *Machine Translation: An Introductory Guide*. London: NCC Blackwell

Bogucki Łukasz. 2009. *Tłumaczenie wspomagane komputerowo*. Warszawa: Wydawnictwo Naukowe PWN.

[3] Zakłada się, że tłumaczenie automatyczne zazwyczaj generuje więcej słów niż w tłumaczeniu referencyjnym.

Brillouin Leon. 1969. *Nauka a teoria informacji*. Warszawa: Wydawnictwo Naukowe PWN.
Chomsky Noam. 1982. *Zagadnienia teorii składni.* Wrocław, Warszawa, Kraków, Gdańsk, Łódź: Ossolineum.
Hutchins W. John. Harold L. Somers. 1992. *An introduction to machine translation.* London: Academic Press.
Jassem Krzysztof. 2001. POLENG: system tłumaczenia automatycznego z języka polskiego na angielski. W *Wybrane zastosowania współczesnej informatyki,* red. Gawiejnowicz Stanisław. Poznań: Polskie Towarzystwo Przyjaciół Nauk.
Lipiński Krzysztof. 2004. *Mity przekładoznawstwa.* Kraków: Egis.
Mazur Marian. 1970. *Jakościowa teoria informacji*. Warszawa: Wydawnictwo Naukowo-Techniczne.
Mykowiecka Agnieszka. 2007. *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym.* Warszawa: Polsko-Japońska Wyższa Szkoła Technik Komputerowych.
Nilsson Nils. 1980. *Principles of Artificial Intelligence.* California: Palo Alto.
Sobczak Wojciech. 1984. *Statystyczna teoria systemów przesyłania informacji.* Warszawa: Wydawnictwo Komunikacji i Łączności.
Wojtasiewicz Olgierd. 1996. *Wstęp do teorii tłumaczenia*. Warszawa: Translegis.

THE POTENTIAL AND LIMITS OF STATISTICAL MACHINE TRANSLATION

Abstract

The presented paper focuses on the problem of machine translation, especially statistical machine translation. The idea behind MT was the concept of the computers' ability to translate some basic and schematical texts by using databases of grammar and vocabulary. Currently MT is not only a fast-developing field of study, but also a popular and (in many cases) free of charge method of translating texts for personal use, for example via Google Translate.

Keywords: translation, statistical machine translation, machine translation