Postulowane kierunki rozwoju systemów komputerowego przekładu

Mirosław Gajer

AGH Akademia Górniczo-Hutnicza, Wydział EAIiE, Katedra Automatyki

Streszczenie: Tematyka artykułu dotyczy zagadnień związanych z automatyzacją przekładu między językami naturalnymi. Jakość pracy współczesnych systemów komputerowego przekładu pozostawia wciąż bardzo wiele do życzenia. W artykule zamieszczono propozycję działań, które powinny przyczynić się do istotnego podniesienia jakości przekładu komputerowego i pozwolić na jego zbliżenie się do rezultatów pracy człowieka. W artykule zaproponowano przeniesienie procesu translacji z poziomu pojedynczych wyrazów na wyższy poziom wielowyrazowych fraz oraz przedstawiono postulat rozwoju systemów komputerowego przekładu zorientowanych na określone obszary tematyczne tłumaczonych tekstów. Wykazano celowość nałożenia swego rodzaju więzów na język źródłowy przekładu, co w znacznym stopniu powinno przyczynić się do eliminacji wieloznaczności tłumaczonych zdań.

Słowa kluczowe: lingwistyka komputerowa, translacja automatyczna, systemy wyspecjalizowane

1. Wprowadzenie

Obecnie badania nad możliwościami automatyzacji przekładu między językami naturalnymi postrzegane są jako jedno z bez wątpienia najważniejszych zagadnień zaliczanych do dziedziny sztucznej inteligencji, którego pomyślne rozwiązanie otworzyłoby drogę dla całej gamy wielu innych zastosowań technik komputerowego przetwarzania i interpretacji języka naturalnego [1]. Historia badań nad przekładem komputerowym jest praktycznie tak samo stara, jak sam wynalazek komputera, a jej początki sięgają co najmniej roku 1951, gdy w Stanach Zjednoczonych na uniwersytecie MIT powołana została specjalna grupa badawcza, mająca za zadanie dokonanie analizy możliwości wykorzystania komputerów w celu automatyzacji przekładu między językami rosyjskim i angielskim [2]. W owym czasie wydawało się, że całkowite zastąpienie człowieka pracującego jako tłumacz przez komputer jest kwestią zaledwie kilku, a być może, w najgorszym wypadku, kilkunastu najbliższych lat prowadzonych w tej dziedzinie zintensyfikowanych badań. Niestety, życie pokazało, jak bardzo się wówczas mylono [3]. Obecnie, gdy historia przekładu komputerowego liczy już ponad 60 lat, całkowite wyeliminowanie tłumacza przysięgłego przez stworzony w tym celu program komputerowy wydaje się być o wiele mniej realne, niż miało to miejsce jeszcze kilkadziesiąt lat temu. Tymczasem badania prowadzone nad automatyzacją przekładu dały początek całkowicie nowej dziedzinie zastosowań informatyki, którą jest lingwistka komputerowa, oraz uświadomiły, jak wiele pracy pozostaje jeszcze do wykonania w obszarze komputerowej analizy składni i semantyki języka naturalnego [4]. W celu realizacji komputerowego przekładu w przeszłości proponowano już wiele różnorodnych podejść, do których zaliczyć można przede wszystkim metody postępowania oparte na regułach, które były inspirowane dokonanymi przez Noama Chomsky'ego odkryciami w zakresie formalnego opisu składni języka naturalnego z wykorzystaniem gramatyk transformacyjno-generatywnych. Ponieważ ostatecznie metody oparte na regułach nie spełniły pokładanych w nich oczekiwań, zaproponowano w tym zakresie liczne podejścia alternatywne, do których można zaliczyć przede wszystkim przekład komputerowy oparty na metodach statystycznych SBMT (ang. Statistical-Based Machine Translation), bazach wiedzy KBMT (ang. Knowledge-Based Machine Translation) oraz przekład komputerowy bazujący na przykładach w postaci korpusów tekstów równoległych EBMT (ang. Example-Based Machine Translation) lub specjalnych przykładach uogólnionych GEBMT (ang. Generalized Example-Based Machine Translation) [5, 6]. Niestety żadne z wymienionych podejść ostatecznie nie spełniło pokładanych w nim nadziei i nie przybliżyło rezultatów prowadzonych badań do pożądanego celu, jakim jest zastąpienie człowieka – tłumacza przez wyspecjalizowany program komputerowy. W związku z istniejącym stanem rzeczy niejako automatycznie nasuwa się niezwykle istotne pytanie, czy dalsze badania nad automatyzacją przekładu między językami naturalnymi mają jeszcze jakikolwiek sens?

Z perspektywy minionego czasu przyznać należy, że postawiony pierwotnie przed badaczami parającymi się automatyzacją przekładu cel był zdecydowanie zbyt ambitny, a wiara w to, że program komputerowy może dokonywać bezbłędnych przekładów zadanych na jego wejście tekstów o dowolnej tematyce była po prostu zwykłą naiwnością, tak charakterystyczną dla każdego pionierskiego okresu badań. W tej kwestii niebagatelnym czynnikiem są zwykle wygórowane oczekiwania użytkowników systemów komputerowego przekładu, którzy z reguły liczą po prostu na to, że komputer będzie w stanie w sposób bezbłędny przełożyć na wybrany język obcy wprowadzony na jego wejście dowolny tekst. W wyniku tego zdecydowana większość użytkowników programów komputerowych tłumaczy jest najczęściej mocno rozczarowana i niekiedy wręcz zawiedziona oferowanym przez nie poziomem jakości przekładu [7]. Jak uczy historia badań nad automatyzacją przekładu, rozpatrywana dziedzina wiedzy w przeszłości przeżywała już liczne wzloty i upadki, po których na dobrych kilkanaście lat najczęściej przestawano się tą problematyką prawie w ogóle zajmować. Czy zatem obserwowany w ostatnich latach swoisty wysyp różnego rodzaju programów komputerowych tłumaczy, w tym licznych dostępnych za pośrednictwem sieci Internet, stanowi zwiastun rychłego przestoju mającego nastąpić w tej dziedzinie, spowodowanego powszechnym rozczarowaniem użytkowników

i wynikającym z tego zniechęceniem do korzystania z jakichkolwiek automatycznych translatorów?

W opinii autora konieczne jest ponowne trzeźwe i zdroworozsądkowe spojrzenie na całą dziedzinę automatyzacji przekładu i nakreślenie realistycznych celów, które będą możliwe do osiągnięcia w przyszłości. Konieczne jest także uczciwe postawienie sprawy odnośnie tego, czego można na obecnym etapie od komputerowych tłumaczy oczekiwać, a co należy traktować jako aktualnie niemożliwe do spełnienia mrzonki. W związku z powyższym autor postanowił przedstawić kilka postulatów, ukazujących, w jego głębokim przekonaniu, właściwe i tym samym perspektywiczne kierunki rozwoju dziedziny automatyzacji przekładu.

2. Translacja na poziomie fraz

Pierwszym i zarazem najważniejszym, zdaniem autora, postulatem, jaki muszą spełniać wszelkiego typu podejścia do automatyzacji przekładu jest wybór właściwego poziomu analizy językowej, na którym ma przebiegać proces translacji. W opinii autora tego rodzaju poziomem jest poziom wielowyrazowych fraz. Tymczasem w przypadku wielu spotykanych do tej pory systemów komputerowego przekładu proces translacji przebiega w zasadzie na poziomie pojedynczych wyrazów. Podejście takie jest stosunkowo proste do praktycznej realizacji, gdyż w tym celu wystarczy dysponować jedynie elektronicznym dwujęzycznym słownikiem wraz z algorytmicznymi mechanizmami tworzenia różnych pochodnych form fleksyjnych wyrazów i uzgadniania ich końcówek, ale zdaniem autora podejście takie nie jest bynajmniej słuszne.

Niezwykle istotną kwestią jest wieloznaczność każdego języka naturalnego, objawiająca się na poziomie jego leksyki, morfologii i składni. Z tego powodu tłumaczenie tekstu z jednego języka na drugi metodami „wyraz po wyrazie", nawet przy użyciu zawansowanych algorytmów analizy kontekstu tłumaczonego wyrazu, daje w praktyce w większości przypadków raczej mierne rezultaty. Co więcej, doświadczenie uczy, że postępując w ten sposób można tłumaczyć skutecznie jedynie bardzo proste zdania, które raczej nie występują w spotykanych w praktyce tekstach, na przekład których istnieje realne zapotrzebowanie. Dopiero przeniesienie procesu przekładu na wyższy poziom analizy językowej, którym jest poziom wielowyrazowych fraz pozwala na uwzględnienie idiomatycznej natury języka i oddawanie tłumaczonych treści w języku docelowym przekładu również w sposób idiomatyczny, dzięki czemu uzyskany na tej drodze przekład komputerowy przypomina w większym stopniu rezultat pracy człowieka niż efekt działania bezdusznej maszyny.

Pierwszą próbę przeniesienia procesu przekładu z poziomu pojedynczych wyrazów do poziomu wielowyrazowych fraz stanowiło wprowadzenie metody automatycznej translacji bazującej na przykładach EBMT. Jednak w celu prawidłowego działania metody EBMT należy dysponować potężnymi korpusami bilingwicznych tekstów, które na dodatek muszą być odpowiednio zsynchronizowane. Ponadto metoda EBMT przewidziana jest przede wszystkim do automatycznego tłumaczenia między językami o budowie analitycznej i raczej nie jest ona odpowiednia w przypadku języków syntetycznych, odznaczających się dużym bogactwem możliwych do utworzenia form pochodnych wyrazów. Z tego powodu autor postanowił wprowadzić własną metodę automatycznego przekładu, określaną mianem metody wzorców translacyjnych PBMT (ang. Pattern-Based Machine Translation), która pozwala na uwzględnienie fleksyjnej natury takich języków, jak na przykład języki słowiańskie, oraz jest w stanie zagwarantować, że między formami fleksyjnymi podmiotu i orzeczenia zdania zawsze spełniony będzie związek zgody, a między orzeczeniem zdania a jego dopełnieniami związek rządu. Z algorytmicznymi mechanizmami funkcjonowania metody PBMT można zapoznać się między innymi na podstawie następujących prac autora [8–10].

3. Systemy wyspecjalizowane

Obecnie zdecydowana większość systemów automatycznej translacji rozwijana jest jako systemy ogólnego przeznaczenia, czyli takie, które w zamierzeniu ich twórców służyć mają do przekładu tekstów o dowolnej i arbitralnie wybranej przez użytkownika tematyce. W opinii autora podejście takie nie jest do końca słuszne i w związku z tym nie może w efekcie zaowocować opracowaniem komputerowych tłumaczy odznaczających się wysoką jakością przekładu. Podobnie, jak w przypadku ludzi parających się na co dzień działalnością translatorską, a zwłaszcza w przypadku tłumaczy przysięgłych, wśród komputerowych translatorów powinna istnieć daleko posunięta specjalizacja. Okazuje się bowiem, że dopiero budowa systemu wyspecjalizowanego do dokonywania automatycznego przekładu określonego typu tekstów o ściśle sprecyzowanym zasięgu tematycznym jest w stanie dostarczyć przekładów, które swoją jakością tylko niewiele, bądź w pewnych wypadkach nawet wcale, nie ustępują rezultatom pracy człowieka.

Dowodem sformułowanej tezy jest opracowany przez autora system automatycznej translacji przewidziany do dokonywania przekładów na język polski tekstów zapisanych w języku angielskim, stanowiących komunikaty opisujące bieżącą sytuację panującą na globalnym rynku wymiany walutowej – *forex* (ang. *foreign exchange*). O szczegółach związanych z zasadami działania wyspecjalizowanego systemu komputerowego tłumacza przeznaczonego do automatycznej translacji tekstów komunikatów dotyczących prognozowanego zachowania się par walutowych, który został zrealizowany z wykorzystaniem zaproponowanej przez autora metody wzorców translacyjnych, można przeczytać między innymi w następujących pracach [11, 12].

4. Języki kontrolowane

Obecnie w wielu międzynarodowych korporacjach istnieje potrzeba dokonywania przekładu tekstów różnego typu dokumentów, do których zalicza się przede wszystkim dokumentację produkowanych urządzeń i wyrobów, ich książki serwisowe, instrukcje obsługi, materiały reklamowe, promocyjne itp. Wydaje się, że byłoby rzeczą wysoce pożądaną zlecenie dokonywania przekładu tego rodzaju stereotypowych dokumentów wyspecjalizowanym w tym celu programom komputerowych translatorów. Jednak w celu ułatwienia

procesu automatycznego przekładu i wydatnego podniesienia jakości uzyskiwanych za pomocą komputerów tekstów w językach docelowych należy rozważyć możliwość nałożenia swego rodzaju więzów na tekst tworzony pierwotnie w języku źródłowym przekładu.

W takim wypadku teksty wyjściowe, które następnie mają zostać przełożone przez komputer na wybrane języki docelowe, tworzone są w tzw. języku kontrolowanym, który w swym założeniu stanowi pewien podzbiór języka naturalnego, z którego się pierwotnie wywodzi. Dodatkowo, z założenia każde zdanie utworzone w języku kontrolowanym powinno być poprawnym zdaniem języka naturalnego, w oparciu o który dany język kontrolowany został utworzony. Natomiast bynajmniej nie każde zdanie języka naturalnego musi być równocześnie poprawnym zdaniem języka kontrolowanego utworzonego w oparciu o dany język naturalny [13].

W przypadku języków kontrolowanych, w porównaniu z właściwymi dla nich językami naturalnymi, ograniczeniu ulegają przede wszystkim konstrukcje składniowe. Zwykle język kontrolowany dopuszcza jedynie najprostsze i bezwzględnie w danym języku konieczne konstrukcje składniowe, dzięki czemu tworzone w danym języku kontrolowanym zdania są jednoznaczne pod względem ich analizy składniowej, tzn. ich rozbioru gramatycznego można dokonać tylko na jeden poprawny sposób. W wielu językach kontrolowanych ograniczeniu podlega także dostępny zasób słownictwa, który zawiera jedynie takie jednostki leksykalne, które są niezbędne z punktu widzenia zastosowań danego języka kontrolowanego. Ponadto wszystkie jednostki leksykalne języka kontrolowanego powinny być semantycznie jednoznaczne w kontekście dziedziny jego zastosowań.

Tworzenie tekstów wyjściowych w językach kontrolowanych z reguły pozwala na znaczne podniesienie jakości uzyskiwanych za pomocą komputera przekładów, ponieważ w tym wypadku praktycznie całkowitej eliminacji ulega najważniejsza z barier, stojąca na drodze do pełnej automatyzacji przekładu, w postaci wieloznaczności wypowiedzi formułowanych w języku naturalnym.

Jako przykład praktycznego wykorzystania języka kontrolowanego opartego na języku angielskim można podać projekt ACEMA Simplified English, który jest wykorzystywany przez korporacje lotnicze na potrzeby łatwiejszego korzystania z różnego typu dokumentacji technicznej. W pewnym sensie na kontrolowanym podzbiorze języka angielskiego bazuje również rozwijany w Carnegie Mellon University projekt KANT, gdzie stworzony został system automatycznej translacji przeznaczony do dokonywania komputerowych przekładów tekstów dokumentacji technicznej dotyczącej sprzętu ciężkiego na kilkanaście języków docelowych, w tym również kilka języków orientalnych [13].

Jeszcze większą odpornością na pojawienie się wieloznaczności w formułowanych wypowiedziach, w porównaniu z językami kontrolowanymi, charakteryzują się języki sztuczne, odznaczające się całkowitą regularnością gramatyki i brakiem w tym względzie jakichkolwiek wyjątków. Ogólnie rzecz biorąc wszystkie języki sztuczne można podzielić na języki aposteriorycznie, czyli takie, które są z zasady wzorowane na wybranych językach naturalnych, i aprioryczne, które z założenia mają być tworzone niejako całkowicie od nowa, czyli bez jakichkolwiek związków z istniejącymi obecnie bądź w przeszłości językami naturalnymi. W przypadku sztucznych języków aposteriorycznych jesteśmy w stanie uwolnić się całkowicie od wieloznaczności na poziomie analizy morfologicznej języka, ponieważ z założenia w tego typu językach poszczególne części mowy przybierają odmienne, jednoznacznie je definiujące końcówki. Na przykład w przypadku najbardziej znanego i rozpowszechnionego ze sztucznych języków aposteriorycznych, czyli języka *esperanto*, wszystkie rzeczowniki otrzymują końcówkę „-o”, przymiotniki „-a”, przysłówki „-e”, a wyznacznikiem liczby mnogiej jest zawsze końcówka „-j”. Języki aposterioryczne także w znacznym stopniu są w stanie zagwarantować jednoznaczność na poziomie analizy syntaktycznej języka, jednak w tym zakresie mogą pojawić się pewne wyjątki. Do chwili obecnej zgłoszonych zostało już bardzo wiele propozycji sztucznych języków aposteriorycznych, przy czym do najbardziej znanych należą, oprócz wspomnianego języka *esperanto*, języki, takie jak *ido*, *interlingua*, *novial* i *occidental*, jednak, żaden z nich jak dotąd nie zdobył większej popularności [14]. Z kolei jako przykład sztucznego języka apriorycznego można podać język *lojban*, w przypadku którego budowa zdania została oparta na rachunku predykatów, w związku z czym każde zdanie zapisane tym w języku jest zawsze jednoznaczne pod względem swej budowy składniowej.

5. Zakończenie

W artykule dokonano podsumowania aktualnego stanu badań w dziedzinie automatyzacji przekładu pomiędzy językami naturalnymi. Obecnie istniejące systemy komputerowego przekładu, przeznaczone do automatycznego tłumaczenia tekstów o arbitralnie wybranej przez użytkownika tematyce, nie dają gwarancji uzyskania odpowiednio wysokiej jakości tekstów docelowych. W opinii autora zmiana istniejącego stanu rzeczy wymaga spełnienia kilku postulatów. Przede wszystkim najważniejszą kwestią wydaje się trwałe zerwanie z podejściami, w przypadku których proces przekładu zachodzi na poziomie pojedynczych wyrazów. Przeprowadzone podczas licznych badań w wielu wiodących światowych ośrodkach eksperymenty z tłumaczeniem komputerowym opartym na przykładach, z wykorzystaniem metod EBMT i GEBMT, oraz dotychczasowe doświadczenia translatorskie związane w wykorzystaniem narzędzi informatycznych stanowiących pamięci translacyjne stanowią dostatecznie przekonujący argument za przeniesieniem procesów automatycznego przekładu na wyższy poziom wielowyrazowych fraz i związków frazeologicznych języka źródłowego. Kolejną istotną sprawą jest zdefiniowanie obszaru tematycznego tekstów tłumaczonych za pomocą opracowywanych programów komputerowych. Obecnie stworzenie automatycznego translatora będącego w stanie zapewnić wysoką jakość tłumaczonych tekstów w przypadku arbitralnego wyboru przez użytkownika ich tematyki wydaje się być mało realne. Zamiast tego rodzaju działań wysiłki należy raczej skupić na rozwijaniu systemów wyspecjalizowanych do przekładu tekstów należących do ściśle określonego kręgu tematycznego, ponieważ jest to warunek konieczny, aby jakość uzyskanego za pomocą programu komputerowego przekładu była odpowiednio wysoka. Dodatkowo w celu podniesienia jakości przekładu komputerowego należy rozważyć możliwość,

w przypadku której teksty przewidziane do przetłumaczenia przez komputer są specjalnie do tego celu dostosowywane już na etapie ich tworzenia. W tym kontekście rozsądną propozycją wydaje się nałożenie pewnego rodzaju więzów na język źródłowy przekładu, sprowadzających się najczęściej do ograniczenia dopuszczalnych konstrukcji składniowych do jedynie podstawowych i bezwzględnie koniecznych w systemie danego języka.

W opinii autora wprowadzenie w życie wymienionych w artykule postulatów powinno doprowadzić do wydatnego podniesienia jakości uzyskiwanych za pomocą komputera przekładów oraz w przypadku wybranych zastosowań powinno umożliwić nawet całkowite zastąpienie człowieka – tłumacza przez wyspecjalizowany program komputerowej translacji.

Bibliografia

1. Hutchins W.: *Machine translation – past, present, future*, Ellis Horwood Series in Computers and their Applications, London, 1986.
2. Arnold D., Balkan L., Meijer S., Humphreys R.L., Sadler L.: *Machine translation: an introductory guide*, NCC Blackwell, London, 1994.
3. Allen J.F.: *Natural language understanding*, The Benjamin/Cummings Publishing Company, New York, 1995.
4. Whitelock P., Kilby K.: *Linguistic and computational techniques in machine translation system design*, UCL Press, London, 1995.
5. Gajer M.: *Wprowadzenie do systemów komputerowego przekładu opartych na metodzie Example-Based Machine Translation*, „Informatyka Teoretyczna i Stosowana", Rocznik 6, Nr 10, 2006, 215–224.
6. Gajer M.: *Wprowadzenie do systemów komputerowego przekładu opartych na metodzie Statistical-Based Machine Translation*, „Informatyka Teoretyczna i Stosowana", Rocznik 6, Nr 10, 2006, 225–232.
7. Melby A.: *Machine translation and philosophy of language*, "Machine Translation Review", no. 9, 1999, 6–17.
8. Gajer M.: *Systemy translacji automatycznej bazujące na metodzie wzorców translacyjnych*, IV Krajowa Konferencja Metody i Systemy Komputerowe, 2003, 713–718.
9. Gajer M.: *The pattern-based French-to-Polish machine translation system*, "Machine Translation Review", no. 13, 2002, 7–41.
10. Gajer M.: *Wielojęzyczne systemy automatycznego przekładu oparte na metodzie wzorców translacyjnych*, AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków, 2008.
11. Gajer M.: *Wyspecjalizowany system automatycznego przekładu zrealizowany metodą wzorców translacyjnych*, AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków, 2008.
12. Gajer M.: *Specialized fully automatic machine translation system delivering high quality of translated texts*, "Task Quarterly", vol. 13, no. 4, 2009, 347–354.
13. Szczepaniak L., Królikowski Z." *Kontrolowane języki naturalne – przegląd rozwiązań i zastosowań*, „Pro Dialog", vol. 11, 2000, 47–67.
14. Gajer M.: *Analiza języków sztucznych i kontrolowanych w kontekście systemów translacji automatycznej*, „Informatyka Teoretyczna i Stosowana", Rocznik 6, Nr 10, 2006, 181–192.

Postulated directions of development of machine translation systems

Abstract: The paper discusses issues of machine translation between natural languages. The quality of contemporary machine translation systems is still relatively low and below users' expectations. In the paper we present some suggestions of activities that could result in raising the quality of machine translation and make it similar to the results of work of human translators. We propose the transfer of the translation process from the level of single words to a higher level of multiword phrases. Further, we postulate to develop machine translation systems that are oriented towards specific thematic areas of translated texts. We also demonstrate that some limitations must be placed upon the source language in order to eliminate the ambiguity of sentences to be translated.

Keywords: computational linguistics, machine translation, specialized systems

dr inż. Mirosław Gajer

Zatrudniony na stanowisku adiunkta w Katedrze Automatyki Akademii Górniczo-Hutniczej w Krakowie. Swoje zainteresowania naukowe łączy z obszarem badawczym sztucznej inteligencji i lingwistyki komputerowej, koncentrując się w szczególności na zagadnieniach automatyzacji przekładu i symulacji procesu ewolucji języków naturalnych.

e-mail: mgajer@ia.agh.edu.pl