STUDIA INFORMATICA 2002
Volume 23 Number 2B (49)

Mirosław GAJER
Akademia Górniczo-Hutnicza w Krakowie, Katedra Automatyki

TRANSLACJA AUTOMATYCZNA W SIECI INTERNET

Streszczenie. Translacja automatyczna jest dziedziną nauki dostarczającą wiedzy o tym, jak programować komputery, aby były w stanie dokonywać automatycznych przekładów pomiędzy językami naturalnymi. Pomimo prawie pół wieku badań nad translacją automatyczną, zadanie to jest jeszcze dalekie od swego ostatecznego rozwiązania. W artykule omówiono stosowane obecnie metody translacji automatycznej oraz rozważono możliwości ich użycia w kontekście sieci Internet.

MACHIN TRANSLATION IN THE INTERNET

Summary. High-quality machine translation between human languages has for a long time been an unattainable dream for many computer scientists involved in this fascinating and interdisciplinary field of the application of computers. After over 50 years of research the problem of automation of translation is still far away from its final solution. The paper is a survey of machine translation techniques with the special attention paid to its applications in the Internet.

1. Wprowadzenie

Translacja automatyczna jest dziedziną zastosowań informatyki, która dostarcza wiedzy o tym, w jaki sposób pisać programy komputerowe, które byłyby zdolne dokonywać w sposób całkowicie automatyczny tłumaczenia pomiędzy językami naturalnymi takimi, jak np. norweski i japoński. Translacja automatyczna nie jest wcale nowym pomysłem, a za jej prekursora uważany jest powszechnie Amerykanin Warran Weaver, który już w 1947 roku wystąpił z ideą zastosowania komputerów do dokonywania przekładów pomiędzy językami naturalnymi. Pierwsza grupa badawcza ukierunkowana na automatyzację translacji została powołana w Austin w Teksasie, już w 1951 roku, a pierwszy publiczny pokaz systemu, który

był w stanie przetłumaczyć z rosyjskiego na angielski 49 prostych zdań, odbył się w roku 1954. Począwszy od tego momentu badania nad translacją automatyczną trwają nieprzerwanie, a sama dziedzina ma już za sobą zarówno poważny kryzys, który wystąpił w latach sześćdziesiątych (tzw. raport ALPAC), jak i renesans, który przyszedł pod koniec lat siedemdziesiątych. Obecnie w wyniku gwałtownej ekspansji sieci Internet translacja automatyczna zaczyna być widziana w zupełnie nowym świetle.

Jak powszechnie wiadomo, Internet to prawdziwa kopalnia informacji. Osoba zainteresowana jakąś dziedziną wiedzy, korzystając ze sprawnej wyszukiwarki internetowej, może tam znaleźć informacje na praktycznie dowolny temat. Jedyna trudność, która występuje na drodze do zrealizowania zamierzonego celu, ma charakter bariery językowej, istniejącej w sposób naturalny pomiędzy poszczególnymi społecznościami świata. Obecnie świat wkracza w epokę, w której ludzkość żyć będzie w jednej wielkiej „globalnej wiosce" – niestety, wielce niepokojący i stanowiący zagrożenie dla dalszego rozwoju jest fakt, że poszczególni mieszkańcy tej „wioski" posługują się różnymi i często całkowicie wzajemnie niezrozumiałymi językami. Aby lepiej zilustrować ten problem załóżmy, że pewna osoba narodowości polskiej chce się czegoś dowiedzieć o wulkanach, ponieważ akurat zainteresowała się tą dziedziną. Posługując się wyszukiwarką internetową może, oczywiście, wpisać hasło „wulkany". Postępując w ten sposób znajdzie być może jakieś polskojęzyczne strony internetowe poświęcone wulkanom. Jeżeli rozważana osoba zna w dobrym stopniu język angielski, może podać wyszukiwarce hasło „volcano" i w ten sposób uzyska dostęp do całej masy angielskojęzycznych stron poświęconych zagadnieniom wulkanizmu. Niestety, osoba taka, nie znając innych języków obcych, nie może w ogóle skorzystać z informacji zawartych na stronach internetowych opracowanych np. w języku niemieckim, francuskim, hiszpańskim czy włoskim – a stron takich jest naprawdę wiele i ich procentowy udział wciąż rośnie. Powstaje zatem pokusa zastosowania w sieci Internet translacji automatycznej. W ten sposób zapytanie użytkownika przed podaniem do wyszukiwarki internetowej byłoby tłumaczone na różne języki narodowe, a następnie obcojęzyczne strony internetowe byłyby na żądanie użytkownika tłumaczone na jego język ojczysty. W ten sposób rozważany polski użytkownik mógłby odczytać w swoim ojczystym języku np. włoską stronę internetową poświęconą wulkanowi Etna.

Przedstawiona tutaj idea brzmi niezwykle pociągająco, ale niestety, zagadnienie automatyzacji translacji w taki sposób, aby uzyskiwane za pomocą komputera przekłady nie odbiegały w swej jakości od wyników pracy zawodowego tłumacza, jest jeszcze bardzo dalekie od swego ostatecznego rozwiązania, pomimo prowadzenia przez ponad pół wieku badań w tej dziedzinie. Istnieje wiele przyczyn, z powodu których automatyzacja translacji jest zadaniem niezwykle trudnym. Do najważniejszych można zaliczyć: różnice systemowe

pomiędzy gramatykami różnych języków, różnice w słownictwie i frazeologii, wieloznaczność będącą immanentną cechą wypowiedzi formułowanych w dowolnym języku naturalnym oraz idiomatyczny charakter języków naturalnych. Niestety, nie ma tutaj miejsca na szersze omówienie zasygnalizowanych powyżej zagadnień. Z tego powodu zainteresowany czytelnik odsyłany jest do innych prac autora, np. [1] i [2]. Natomiast w dalszej części artykułu zostaną bliżej omówione stosowane obecnie techniki translacji automatycznej oraz to, co można w wyniku ich zastosowania osiągnąć.

2. Techniki translacji automatycznej

Do chwili obecnej opracowano już wiele różnorodnych podejść do zagadnienia automatyzacji translacji. Systemy translacji automatycznej, które powstały najwcześniej, charakterem swej pracy przypominały elektroniczne słowniki. W systemach takich zdania były tłumaczone wyraz po wyrazie, poprzez proste wyszukiwanie znaczeń wyrazów w dwujęzycznym słowniku i następnie podstawianie ich w miejsce wyrazów języka źródłowego. Oczywiście, jakość uzyskanego w ten sposób przekładu była bardzo niska, ponieważ nie były uwzględniane tutaj żadne reguły gramatyki języka źródłowego, ani języka docelowego.

Kolejnym krokiem naprzód było opracowanie systemów translacji automatycznej opartych na tak zwanym transferze struktur syntaktycznych. Systemy takie oznaczane są skrótem **TBMT** (ang. Transfer-Based Machine Translation). Metoda transferu polega na dokonaniu w pierwszej kolejności analizy gramatycznej tłumaczonego zdania. Celem tej analizy jest dokonanie rozbioru gramatycznego zdania i ustalenie, który z wyrazów tłumaczonego zdania jest jego podmiotem, orzeczeniem, dopełnieniem bliższym, dopełnieniem dalszym, przydawką podmiotową, partykułą, przydawką dopełnieniową, okolicznikiem miejsca, czasu, sposobu, celu itp. Następnie po zakończeniu analizy gramatycznej tłumaczonego zdania dokonywany jest jego transfer do odpowiadającej mu struktury gramatycznej języka docelowego. Proces ten zostanie zilustrowany na następującym przykładzie. Dane jest zdanie w języku angielskim:

If my father had had money my brother would have bought that car.

Po dokonaniu analizy gramatycznej tego zdania w systemie translacji automatycznej opartym na metodzie transferu, okazuje się, że zdanie to jest typowym zdaniem reprezentującym tzw. trzeci tryb warunkowy, które posiada następująca budowę:
<partykuła „if"> <podmiot_1> <orzeczenie_1> <dopełnienie_1> <podmiot_2> <orzeczenie_2> <dopełnienie_2>.

Polski odpowiednik takiego zdania ma bardzo podobna budowę gramatyczną: *<partykuła „gdyby"> <podmiot_1> <orzeczenie_1> <dopełnienie_1> <podmiot_2> <orzeczenie_2> <dopełnienie_2>*.

Pozostaje zatem dokonać tłumaczenia podmiotu, orzeczenia i dopełnienia zdania podrzędnego (<podmiot_1> <orzeczenie_1> <dopełnienie_1>) oraz zdania nadrzędnego (<podmiot_2> <orzeczenie_2> <dopełnienie_2>), aby uzyskać poprawny polski przekład tego zdania, który ma postać następującą:

Gdyby mój ojciec miał pieniądze mój brat kupiłby tamten samochód.

Z teoretycznego punktu widzenia wszystko wygląda jak najbardziej w porządku. Niestety, w praktyce metoda transferu ma dwie poważne wady. Pierwsza z nich polega na tym, że u podstaw metody transferu leży milczące założenie, zgodnie z którym tłumaczone zdanie ma poprawną i w miarę przejrzystą budowę gramatyczną, co niestety, w praktyce nader często okazuje się nieprawdą. Istniejące obecnie parsery (programy do analizy syntaktycznej) potrafią dokonać poprawnej analizy gramatycznej dla jedynie stosunkowo prostych zdań.

Druga z wad rozważanej metody translacji automatycznej polega na tym, że metoda ta nie posiada żadnych środków, które pozwalałyby się jej zmierzyć z wieloznacznością występującą na poziomie leksykalnym. Bowiem w sytuacji, gdy nawet zostanie dokonana w sposób prawidłowy gramatyczna analiza tłumaczonego zdania i właściwie zostanie wybrana odpowiadająca jej struktura gramatyczna języka docelowego, podczas dokonywania przekładów poszczególnych wyrazów pojawia się nieuchronnie problem, które z ich licznych ekwiwalentów, występujących w języku docelowym, należy wybrać. Na przykład niemieckie słowo „Kiefer" ma dwa polskie ekwiwalenty znaczeniowe: „sosna" i „szczęka". Oczywiście, nie jest rzeczą obojętną, który z nich zostanie podstawiony w miejsce niemieckiego oryginału. Dokonanie błędnego wyboru poskutkuje najprawdopodobniej całkowitym niezrozumieniem przez odbiorcę dokonanego automatycznie przekładu. Jest to naprawdę poważny problem, bowiem wieloznaczność występująca na poziomie leksykalnym jest immanentną cechą każdego języka naturalnego – otwierając dowolny duży dwujęzyczny słownik trudno jest doprawdy znaleźć wyrazy, które posiadałyby w języku docelowym tylko jeden ekwiwalent znaczeniowy.

Wymienione wady systemów translacji automatycznej opartych na metodzie transferu doprowadziły wielu badaczy do wniosku, iż nie można prawidłowo tłumaczyć tekstów bez ich zrozumienia. Sformułowanie takich wniosków doprowadziło bezpośrednio do opracowania nowej klasy systemów translacji automatycznej, które zostały nazwane systemami opartymi na wiedzy. W literaturze systemy takie oznaczane są skrótem KBMT (ang. Knowledge-Based Machine Translation). Celem twórców systemów KBMT było wbudowanie do programów komputerowych dokonujących translacji wiedzy dotyczącej

języka źródłowego i docelowego, jak również wiedzy dotyczącej dziedziny, z którą tłumaczone teksty są związane. Niestety, w przypadku dokonywania przekładów tekstów dotyczących tematyki ogólnej, w komputerowy program automatycznej translacji musiałaby również zostać wbudowana ogólna wiedza o świecie, co zapewne jest sprawą jeszcze bardzo odległej przyszłości. Dlatego też jeszcze do chwili obecnej nie powstał żaden system typu KBMT, który byłby w stanie tłumaczyć teksty należące do języka ogólnego. Natomiast systemy takie są z powodzeniem stosowane na potrzeby translacji tekstów należących do pewnej specjalistycznej i bardzo wąskiej dziedziny wiedzy, której model może zostać z powodzeniem zaimplementowany w programie komputerowym. Na przykład typowym systemem KBMT jest system KANT opracowany na Carnegie-Mellon University, którego zadaniem jest tłumaczenie podręczników obsługi sprzętu ciężkiego, np. koparek, dźwigów, podnośników, spychaczy, wywrotek itp. [5].

Z kolei wielu badaczy zajmujących się translacją automatyczną postanowiło pójść w całkowicie przeciwnym kierunku i zerwać raz na zawsze ze stosowaniem w takich systemach jakiejkolwiek wiedzy lingwistycznej bądź dziedzinowej. Takie postawienie sprawy zaowocowało wypracowaniem kilku niezwykle interesujących koncepcji.

Pierwszą z rozważanych koncepcji jest tzw. tłumaczenie oparte na analizie przypadków. Systemy automatycznej translacji, w których stosowane jest takie podejście, oznaczane są skrótem **CBMT** (ang. Case-Based Machine Translation). Idea leżąca u podstaw systemów CBMT jest bardzo prosta. Załóżmy, że mamy pewien bilingwiczny tekst, ot chociażby taki, jak poniżej:

„*During the whole of a dull, dark, and soundless day in the autumn of the year, when the clouds hung oppressively low in the heavens, I had been passing alone, on horseback, through a singularly dreary tract of country, and at length found myself* ***as the shades of the evening drew on****. within view of the melancholy House of Usher.*"

„*Przez cały mroczny, głuchy I smętny dzień jesienny, pod ciężką posową obłoków, co wlokły się nisko po niebie, nie opuszczałem siodła przebiegając samotnie na koniu dziwnie zamarłą połać kraju i* ***dopiero z nastaniem wieczornych zmierzchów*** *ujrzałem przed sobą posępny Dom Usherów.*"

Górna część zamieszczonego powyżej tekstu to fragment opowiadania Edgara Allana Poe'a zatytułowanego „The Fall of the House of Usher", natomiast jego dolna część to polski przekład tego opowiadania noszący tytuł „Zagłada Domu Usherów", dokonany przez wybitnego tłumacza Stanisława Wyrzykowskiego. Jeżeli teraz z angielskojęzycznej części rozważanego bilingwicznego tekstu wybierze się pewną frazę, na przykład taką, jaka została w nim podkreślona (*as the shades of the evening drew on*), a następnie w polskojęzycznej części odszuka się jej odpowiednik (*dopiero z nastaniem wieczornych zmierzchów*), to

istnieje duże prawdopodobieństwo, ale nie całkowita pewność, że jeżeli taka angielska fraza pojawi się jeszcze raz w tłumaczonym angielskim tekście, to podstawiając w jej miejsce rozważany polski ekwiwalent, otrzyma się ponownie jej prawidłowy przekład. Oczywiście, ktoś może zadać pytanie, dlaczego nie ma stuprocentowej pewności, że tak dokonany przekład będzie prawidłowy? Powód jest bardzo prosty, są nim trudności, jakie napotyka się podczas prób zautomatyzowania translacji – zwłaszcza różnice systemowe w gramatyce rozważanych języków, różnice w ich systemach słownictwa oraz wieloznaczność wypowiedzi występującą na poziomie leksykalnym, syntaktycznym i semantycznym.

Pomimo braku gwarancji na poprawność przekładu dokonanego metodą CBMT, rozważana metoda translacji coraz częściej okazuje się być bardzo poważną alternatywą wobec stosowanych dotychczas technik automatycznej translacji. Dzieje się tak zwłaszcza w sytuacji, gdy języki uwikłane w proces przekładu są blisko ze sobą spokrewnione. Spektakularny przykład w tym zakresie pochodzi z Hiszpanii, gdzie od 1999 roku wydawana jest w języku katalońskim gazeta „El Periodico de Cataluña", która jest tłumaczona z języka hiszpańskiego całkowicie automatycznie, właśnie dzięki zastosowaniu techniki CBMT [4].

Jest rzeczą oczywistą, że metoda CBMT wymaga opracowania bazy danych o pokaźnych rozmiarach, która jest w stanie pomieścić dużą liczbę przekładów fraz należących do języka źródłowego. Oczywiście, teoretycznie liczba takich fraz jest nieskończenie wielka, ale w praktyce okazuje się, że aby dokonywać przekładów o w miarę dobrej jakości, wystarczy, że system automatycznej translacji będzie uwzględniał od kilku do kilkunastu milionów najczęściej występujących w danym języku fraz.

Swoistą odmianę techniki CBMT stanowi technika translacji oparta na przykładach translacyjnych, która w literaturze oznaczana jest skrótem **EBMT** (ang. Example-Based Machine Translation). W metodzie EBMT nie występuje specjalnie tworzona przez lingwistów baza danych zawierająca tłumaczenia fraz, natomiast system automatycznej translacji dysponuje znacznym zbiorem dwujęzycznych tekstów, przy czym zarówno frazy z tekstu w języku źródłowym, jak i ich domniemane przekłady są wybierane przez system w sposób automatyczny.

Główną zaletą wszelkich podejść opartych na wzorcach translacyjnych jest fakt, że w systemach takich w naturalny sposób minimalizowane jest ryzyko pojawienia się wieloznaczności na poziomie leksykalnym, ponieważ podczas translacji poszczególne wyrazy nie są tłumaczone oddzielnie, ale jednocześnie tłumaczona jest cała ich grupa, dzięki czemu system korzysta z dodatkowych informacji kontekstowych. Ponadto przekłady dokonane metodą opartą na wzorcach translacyjnych brzmią dla człowieka w sposób bardziej naturalny, podobny do efektu pracy zawodowego tłumacza. Na przykład system translacji automatycznej oparty na transferze przetłumaczyłby angielską frazę *natural gas* zapewne jako *naturalny gaz*,

podczas gdy prawidłowe tłumaczenie to *gaz ziemny* – zatem, jak widać, tłumaczenie całych fraz może wydatnie zwiększyć jakość automatycznych przekładów.

3. Zakończenie

Z zamieszczonych powyżej rozważań widać, że pomysł zautomatyzowania translacji pomiędzy językami naturalnymi wciąż pozostaje niedoścignionym ideałem, można by wręcz powiedzieć – św. Graalem poszukiwań badawczych. Trzeba być jednakże optymistą. Bowiem nawet jeśli wymarzony ideał nigdy nie zostanie osiągnięty, to i tak jest rzeczą wysoce prawdopodobną, że dzięki nieustannie wzrastającej mocy obliczeniowej komputerów i dzięki intensywnym badaniom prowadzonym przez interdyscyplinarne zespoły złożone z informatyków i lingwistów, jakość pracy systemów automatycznej translacji zbliży się w przynajmniej stopniu zadowalającym do pożądanego ideału, jakim jest system całkowicie automatycznej translacji dostarczający przekładów o wysokiej jakości dla tekstów należących do języka ogólnego. Należy ponadto zauważyć, że w przypadku języków o bliskim stopniu wzajemnego pokrewieństwa ideał ten już w zasadzie osiągnięto, czego przykładem może być automatyczne tłumaczenie gazety „Periodico de Cataluña" z języka hiszpańskiego na kataloński [4]. Zatem w przyszłości komputery będą zapewne tłumaczyć teksty pomiędzy różnymi językami równie biegle jak już dzisiaj komputery potrafią np. grać w szachy. Niestety, nie może natomiast być żadnej nadziei na to, że komputery będą cokolwiek z tłumaczonych przez siebie tekstów rozumieć.

LITERATURA

1. Gajer M.: System translacji automatycznej oparty na dialogu z użytkownikiem i przykładach translacyjnych, Pro Dialog 14(2002), Wydawnictwo NAKOM, Poznań, s. 77-115.
2. Gajer M.: Fakty i mity na temat systemów translacji automatycznej, Elektronizacja 1-2/2002, Wydawnictwo SIGMA-NOT, Warszawa, s. 33-36.
3. Waibel A., Geutner P., Tomokiyo L. M., Schultz T., Woszczyna M.: Multilinguality in Speech and Spoken Language Systems, Proceedings of the IEEE, vol. 88, no. 8, August 2000, s. 1297-1313.
4. Rico C.: From Novelty to Ubiquity: Computers and Translation at the Close of Industrial Age, http://www.accurapid.com/journal/15mt2.htm.

5. Nyberg E., Mitamura T., Carbonell J.: The KANT Machine Translation System: From R&D to Initial Deployment, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, USA, 1997.

Recenzent: Dr inż. Nina Suszczańska

Wpłynęło do Redakcji 29 marca 2002 r.

Abstract

Machine translation is a science that delivers the knowledge how to program the computers, so as they were able to translate between human languages, for example, between Danish and Bulgarian. It may be amazing, but the field of machine translation is almost as old as the invention of computer itself. In 1949 an American scientist Warren Weaver sent the memorandum to The Rockefeller Foundation (American institution supporting the scientific research), in which he demanded starting the research on the automation of translation between natural languages. Warren Weaver was inspired by cryptographic techniques, which were developed very strongly during the years of The Second World War, and he thought that there existed some fundamental similarities between these cryptographic techniques and the process of translation between human languages. After the first successes it soon appeared that the problem of machine translation is far more complicated and far more harder than Warren Weaver had ever imagined. After over 50 years of research the problem of automation of translation is still far away from its final solution. The paper is a survey of machine translation techniques with the special attention paid to its applications in the Internet.