Krzysztof Jassem

System tłumaczenia automatycznego opracowany na potrzeby poprawy bezpieczeństwa publicznego

Streszczenie

Celem badań jest opracowanie i wdrożenie systemu tłumaczenia automatycznego wysokiej jakości na potrzeby poprawy bezpieczeństwa publicznego.

Projektowany system tłumaczenia automatycznego zapewni wysoką jakość tłumaczenia dzięki wykorzystaniu korpusów z dziedziny bezpieczeństwa publicznego. Korpusy te zostaną wykorzystane w dwojaki sposób:

1) uzyskany zostanie słownik polsko-angielski fraz związanych z bezpieczeństwem o wielkości przekraczającej 1 milion jednostek,
2) zbudowana zostanie pamięć tłumaczeń, w której przechowywane będzie ponad 8 milionów jednostek tłumaczenia.

System translacji automatycznej może zrealizować postulat niezwykle trudny do osiągnięcia w tłumaczeniu ludzkim: jednorodność tłumaczenia terminologii. Na przykład, z korpusu odpowiadających sobie tekstów z Oficjalnego Dziennika Unii Europejskiej (http://eur-lex.europa.eu/) w języku polskim i angielskim, który składa się z ok. 2 500 000 jednostek tłumaczenia, można automatycznie wyekstrahować kilkaset tysięcy różnych fraz języka angielskiego – terminów, które powinny zostać przełożone na język polski w jeden określony sposób. Tymczasem w analizowanych dokumentach większość terminów posiada dwa lub więcej różnych odpowiedników. Od tłumaczenia dokumentów istotnych dla bezpieczeństwa oczekuje się natomiast, by określone terminy były zawsze przełożone w ten sam sposób.

Prace będą koncentrować się przede wszystkim nad zapewnieniem wysokiej jakości tłumaczenia między językiem angielskim i polskim. Wytworzony system umożliwi jednakże tłumaczenie z i na inne języki: w szczególności niemiecki, rosyjski i francuski, czyli języki najważniejsze z punktu widzenia geopolitycznego położenia Polski. W celu uzyskania poprawnej analizy składniowej tych języków zostaną wykorzystane nowoczesne modele lingwistyczne oparte na istniejących korpusach opisanych składniowo.

Prowadzone badania mają na celu stworzenie prototypu systemu tłumaczenia mowy. We współpracy z zespołem prof. Grażyny Demenko, który w ramach Polskiej

Platformy Bezpieczeństwa Wewnętrznego realizuje projekt rozpoznawania mowy ciągłej języka polskiego, opracowany zostanie prototyp systemu tłumaczenia mowy między językami polskim i angielskim.

Abstract

The aim of the project is to develop and implement a high quality Machine Translation system intended to be used for the needs of public security improvement.
The system will assure high quality of translation thanks to corpus-based methods. The corpora, containing texts from the domain of public security, will be used as the basis for the following resources:

1. A Polish-English dictionary of phrases containing over 1 million units.
2. A translation memory of over 8 million units.

Machine Translation systems may help achieve an aim difficult to obtain in human translation: the homogeneity of terminology translation.
For example, the Official Journal of the European Union (http://eur-lex.europa.eu/) is a multilingual text corpus of over 2 500 000 translation units. It has been possible to extract automatically a few hundred thousand distinct English phrases from the Journal, most of which should be translated into Polish in one, strictly defined way, in order to assure the homogeneity of translation. However, most of the phrases have, in the analyzed corpus, two or more equivalents (this is most probably due to the fact that particular fragments of texts have been translated by various human translators). It is required from the translation of security documents to result in the same translation of the same terms and sentences.
The research will focus on high quality translation between Polish and English. The system will however allow for the translation from and into other languages, in particular: German, Russian, and French, i.e. languages, which are the most important for Poland from the geopolitical point of view.
The research will also aim at the creation of Speech Translation prototype. This objective will be achieved in strict co-operation with the team of Professor Grażyna Demenko, whose project concentrates on the recognition of Polish speech.

1. Geneza projektu

Systemy tłumaczenia automatycznego przeznaczone do tłumaczenia tekstów ogólnych nie dorównują jeszcze jakością tłumaczeniu ludzkiemu. W ostatnich latach

stało się możliwe uzyskanie tłumaczenia automatycznego jakości zbliżonej do jakości ludzkiej (a niekiedy nawet wyższej) przy założeniu ukierunkowania systemu na określoną dziedzinę wiedzy. W tym celu należy wykorzystywać istniejące korpusy tekstów z danej dziedziny przetłumaczonych przez człowieka.

Na przykład, korpus odpowiadających sobie tekstów Oficjalnego Dziennika Unii Europejskiej w języku polskim i angielskim zajmuje ok. 1 GB powierzchni dyskowej, co kilka lat temu było sporym obciążeniem komputera, a obecnie jest wielkością nie stwarzającą bariery technologicznej. Systemy oparte na statystycznych własnościach korpusów tekstów jeszcze do niedawna były wyłącznie domeną projektów akademickich, a obecnie pojawiają się na rynku masowym, czego przykładem może być tłumaczenie udostępniane przez wyszukiwarkę google (http://www.google.pl/language_tools?hl=pl).

System translacji automatycznej może zrealizować postulat niezwykle trudny do osiągnięcia w tłumaczeniu ludzkim: **jednorodność tłumaczenia terminologii**. Na przykład, w wyżej wymienionym korpusie tekstów Unii Europejskiej udało się automatycznie wyekstrahować 470 000 różnych fraz języka angielskiego (tradycyjne wielkie słowniki zawierają zaledwie między 50 000 a 80 000 fraz). Wstępna analiza wykazuje, że około 60% z tych fraz stanowią terminy, które powinny zostać przełożone na język polski w jeden określony sposób (tymczasem w analizowanych dokumentach większość terminów posiada dwa lub więcej różnych odpowiedników).

Od tłumaczenia dokumentów istotnych dla bezpieczeństwa oczekuje się, by zdania o treści tej samej były zawsze przełożone w ten sam sposób. Ten postulat również dobrze spełniają systemy translacji automatycznej dzięki **pamięciom tłumaczeń**, w których przechowywane są odpowiadające sobie jednostki tłumaczenia (najczęściej zdania), zweryfikowane pod względem jakości przekładu.

2. Cel wdrożeniowy projektu

Celem projektu jest opracowanie i wdrożenie **systemu tłumaczenia automatycznego wysokiej jakości** na potrzeby poprawy bezpieczeństwa międzynarodowego.

W celu zbudowania takiego systemu należy:

- zebrać duże korpusy odpowiadających sobie tekstów,
- automatycznie dopasować do siebie zdania, które są swoimi odpowiednikami i umieścić je w pamięci tłumaczeń,

- automatycznie wyekstrahować z korpusów odpowiadające sobie frazy leksykalne i umieścić je w słowniku systemu tłumaczenia,
- z pozyskanych korpusów automatycznie wyekstrahować reguły składni rządzące tekstami,
- skonstruować algorytmy, które będą korzystać z uzyskanych zasobów, a następnie zoptymalizować ich złożoność czasową i pamięciową.

Algorytm tłumaczenia ma działać następująco: algorytm wyszukuje w pamięci tłumaczeń zdania podobne do zdania wejściowego. W przypadku powodzenia jako wynik tłumaczenia przyjęty zostaje zmodyfikowany odpowiednik z pamięci tłumaczeń. W przypadku niepowodzenia algorytm dokonuje automatycznej translacji zdania w oparciu o:

- reguły translacji uzyskane z analizy zebranych korpusów,
- reguły translacji opracowane przez człowieka,
- słownik wyrazów i fraz pozyskany z zebranych korpusów,
- słownik wyrazów i fraz opracowany przez człowieka.

Projektowany system tłumaczenia automatycznego zapewni wysoką jakość tłumaczenia dzięki wykorzystaniu korpusów z dziedziny bezpieczeństwa publicznego. Korpusy te zostaną wykorzystane w dwojaki sposób:

- uzyskany zostanie słownik polsko-angielski **fraz** o wielkości przekraczającej **1 milion** jednostek;
- zbudowana zostanie pamięć tłumaczeń, w której przechowywane będzie ponad **8 milionów jednostek tłumaczenia**.

Efektem końcowym projektu będzie system tłumaczenia automatycznego wysokiej jakości z zakresu bezpieczeństwa publicznego. Jądro systemu tłumaczenia stanowić będzie istniejący system tłumaczenia Translatica (www.translatica.pl). System zapewnia tłumaczenia między językiem polskim a językami angielskim, niemiecki i rosyjskim.

W projekcie największy nacisk zostanie położony na rozwój pary językowej polski-angielski.

Osiągnięte narzędzia będą mogły być wykorzystywane do tłumaczenia w sposób automatyczny dokumentów Unii Europejskiej, np. w **Systemie Informacyjnym Schengen**, tłumaczenia tekstów **Interpolu** oraz przy organizacji imprez masowych, takich jak np. Mistrzostwa Europy w piłce nożnej EURO 2012.

System będzie realizowany równolegle z projektami badawczymi Polskiej Platformy Bezpieczeństwa Wewnętrznego: „Technologie przetwarzania oraz rozpoznawania informacji słownych w systemach bezpieczeństwa wewnętrznego” (kierownik: prof. dr hab. G. Demenko) oraz „Technologie przetwarzania tekstu polskiego zorientowane na potrzeby bezpieczeństwa publicznego” (kierownik: prof. dr hab. Z. Vetulani). W ramach współpracy z zespołem prof. G. Demenko zostanie zrealizowany prototyp systemu tłumaczenia mowy.

Systemy automatycznego tłumaczenia mowy mogą mieć kluczowe znaczenie w wielu sytuacjach związanych z bezpieczeństwem:

- stanowiska przyjmowania zgłoszeń alarmowych (numery 112, 999, 998, 997);
- przyjęcie informacji na miejscu zdarzenia;
- tłumaczenie informacji zebranej przez funkcjonariuszy bezpieczeństwa;
- pościg transgraniczny;
- współdziałanie służb w sytuacjach nadzwyczajnych;
- wsparcie pracy mediatora;
- wsparcie tłumaczenia dla osób posługujących się językiem obcym bez znajomości słownictwa specjalistycznego.

3. Cele naukowe projektu

Wyniki badań przeprowadzonych przy realizacji projektu zostaną wykorzystane w trzech pracach doktorskich (Marcin Junczys-Dowmut, Mikołaj Wypych, Tomasz Kowalski) i jednej pracy habilitacyjnej (Filip Graliński).

Praca doktorska Marcina Junczysa-Dowmuta obejmuje zagadnienia automatycznego ekstrahowania z korpusów i statystycznego tłumaczenia fraz rzeczownikowych.

Praca doktorska Mikołaja Wypycha podejmuje problematykę syntezy mowy.

Praca doktorska Tomasza Kowalskiego obejmuje zagadnienia automatycznej ekstrakcji reguł translacji z korpusów równoległych.

W swojej pracy habilitacyjnej Filip Graliński będzie zajmował się problemem wykorzystania algorytmów sztucznej inteligencji (takich jak algorytm A*) w celu ograniczenia przestrzeni przeszukiwań w analizie składniowej tekstu.

Opracowane algorytmy dopasowywania tekstów (patrz sekcja 4.1.) zostaną udostępnione w formie „*open source*” w celu ich dalszego rozwoju przez społeczność

informatyczną. W obecnej chwili dostępne są opisy niektórych algorytmów dopasowywania (np. algorytmu Moore'a), natomiast nie są udostępnione kody źródłowe programów. Udostępnienie naszych kodów źródłowych z tego zakresu przyczyni się, jak sądzimy, do rozwoju algorytmów dopasowywania, a co za tym idzie, lepszego wykorzystania korpusów językowych w badaniach naukowych (Jassem, Lipski, 2007).

4. Charakterystyka badań

4.1. Opracowanie nowych algorytmów automatycznego dopasowywania tekstów na poziomie dokumentów, elementów i zdań (text alignment)

Jak wspomniano w paragrafie 2, pierwszym etapem tworzenia systemu tłumaczenia dedykowanego na potrzeby poprawy bezpieczeństwa publicznego jest zebranie i przygotowanie korpusów dwujęzycznych. Przez korpus dwujęzyczny rozumiemy tutaj zbiór tekstów w dwóch językach – nazwijmy je umownie językiem źródłowym i językiem docelowym – w którym określono wzajemnie jednoznaczne odwzorowanie pomiędzy tekstami w języku źródłowym i tekstami w języku docelowym. Innymi słowy: dla każdego tekstu w języku źródłowym w korpusie dwujęzycznym musi istnieć jego odpowiednik w języku docelowym i *vice versa*.

Co więcej, od dwujęzycznych korpusów tekstowych w większości zastosowań oczekuje się istnienia odwzorowań na bardziej szczegółowych poziomach: akapitów, zdań lub wyrazów. Proces wyznaczania takiego odwzorowywania nazywa się **dopasowywaniem**. Wynik procesu dopasowywania (czyli dwujęzyczny korpus tekstów, na którym wyznaczono odwzorowanie) nazywa się **dopasowaniem**. W przypadku zastosowania korpusów dwujęzycznych do tłumaczenia automatycznego za pomocą pamięci tłumaczeń oczekuje się wyznaczenia dopasowania na poziomie zdań.

W obecnej chwili znanych jest wiele algorytmów automatycznego zbierania korpusów dwujęzycznych z Internetu. W celach badawczych najczęściej stosuje się algorytmy: STRAND (Resnik, Smith, 2003), BITS (Ma, Liberman, 1999) and PTMiner (Chen, Nie, 2000). Algorytmy te stosowano z dobrym skutkiem (autorzy podają, że ponad 90% proponowanych przez algorytmy par dokumentów faktycznie było swoimi odpowiednikami) dla języków wiodących w Internecie (najczęściej angielskiego i francuskiego). Eksperymenty przeprowadzone przez Monikę Rosińską, pod opieką autora w roku 2007 (Rosińska, 2007) dla korpusów polsko-niemieckich dały znacznie

gorsze wyniki: zastosowanie algorytmu Resnika w czystej postaci dało zaledwie 10% dobrych dopasowań (pozostałe 90% par nie były swoimi odpowiednikami), a kombinacja algorytmów STRAND i BITS poprawiła skuteczność dopasowania do 40%.

Należy więc wnosić, że w celu zgromadzenia korpusów dwujęzycznych z językiem polskim niezbędne jest zmodyfikowanie istniejących algorytmów czy wręcz opracowanie nowych. Co więcej, od opracowanych algorytmów oczekuje się spełniania dodatkowego kryterium – znalezione dokumenty powinny dotyczyć określonej dziedziny – bezpieczeństwa publicznego.

Wyżej wymienione algorytmy przyjmują założenie, że poszczególne wersje językowe dokumentów znajdują się pod różnym adresami URL (których nazwy różnią się nieznacznie – najczęściej oznaczeniem języka). Nasze badania wskazują, że wiele dokumentów dwujęzycznych umieszczonych jest pod tym samym adresem URL (wersje językowe znajdują się w oddzielnych elementach HTML). Do odszukiwania tego typu stron należy opracować nowe algorytmy, które z jednej strony rozpoznają języki stron, a z drugiej analizują współwystępowanie na stronie wyrazów i ich słownikowych odpowiedników w innym języku (w celu odrzucenia stron zawierających słowniki podające najczęściej wyrazy w formie podstawowej algorytm weryfikuje współwystępowanie form fleksyjnych).

Zgromadzone dokumenty należy poddać procesowi dopasowywania na poziomie zdań w celu uzyskania danych do pamięci tłumaczeń.

Pierwsze skuteczne dopasowywania zaczęto rozwijać w latach 90. Obecnie za najbardziej skuteczne (czyli dające największy procent prawidłowego dopasowania) uważa się algorytmy Moore'a (Moore, 2002, 2005) oraz HunAlign (HunAlign, 2007).

Algorytm Moore'a, z pewnymi użytecznymi modyfikacjami, został reimplementowany w języku Java i udostępniony publicznie przez Jarosława Lipskiego (Jassem, Lipski, 2007). Jedną z takich modyfikacji jest możliwość poddania tekstu przetwarzaniu potokowemu. W przetwarzaniu potokowym tekst za pomocą jednego polecenia dzielony jest najpierw na akapity, a następnie na zdania.

Zarówno algorytm Moore'a, jak i HunAlign podają współczynnik dokładności dopasowania na poziomie powyżej 90%. Nasze doświadczenia wskazują, że wyniki te nie sprawdzają się dla tekstów HTML znajdujących się na stronach internetowych (dokładność dopasowania spada drastycznie). Metodą na przezwyciężenie tego problemu jest opracowanie algorytmów dopasowywania na poziomie elementów

HTML. Jest to odpowiednik dopasowywania na poziomie akapitów w plikach tekstowych – pomijany jednak jak dotychczas w literaturze. W ramach projektu opracujemy algorytmy dopasowywania na poziomie elementów.

Uzupełnieniem do algorytmów dopasowywania automatycznego będą narzędzia do ludzkiej korekty dopasowywania. Narzędzia te pozwolą użytkownikowi w sposób ergonomiczny skorygować błędy dopasowania popełnione przez algorytm.

4.2. Opracowanie nowych algorytmów automatycznego dopasowania tekstów na poziomie wyrazów (word alignment) w celu poprawienia jakości algorytmów tłumaczenia statystycznego

Przy wykorzystaniu modeli dopasowywania IBM 4 i IBM 5 (Och, Ney 2003) oraz metod symetryzacji zostanie zaimplementowane nowe podejście do tłumaczenia statystycznego oparte na dodatkowych informacjach lingwistycznych, tzw. *Factored Phrase Based Translation Models* (Koehn, Hoang 2007; Koehn, Schroeder 2007). Wykorzystując znalezione frazy i ich odpowiedniki jako materiał treningowy można uzyskać tłumaczenia fraz dotąd nienapotkanych w korpusach równoległych, o ile tylko w korpusie znajdują się ich fragmenty. Tłumaczenia uzyskane taką metodą będą cechować się wysoką poprawnością i naturalnością, ponieważ będą odtwarzały słownictwo i styl korpusów źródłowych. Równocześnie połączenie metod statystycznych z podejściem regułowym umożliwi uniknięcie błędów czysto składniowych w generowanych tekstach, co jest najsłabszym ogniwem algorytmów *stricte* statystycznych (Koehn, Knight 2003).

Zbadany zostanie również wpływ na jakość tłumaczenia statystycznego modeli innych niż n-gramowe (np. modeli uwzględniających składnię). Przy ograniczeniu modelu do fraz rzeczownikowych może okazać się, że metody lingwistyczne oparte na wzorcach składniowych sprawdzą się lepiej niż używane dotychczas modele n-gramowe.

Badania z zakresu ekstrakcji fraz z korpusów równoległych są już prowadzone w ramach pracy doktorskiej Marcina Junczysa-Dowmunta, jednego z wykonawców projektu.

4.3. Algorytmy dezambigiuacji znaczenia wyrazów (Word Sense Disambiguation)

Algorytmy dezambigiuacji znaczenia wyrazów stosują metody pozalingwistyczne w celu ustalenia znaczenia niejednoznacznego wyrazu (np. „przypadek", „zamek" w języku polskim). Jednym ze sposobów rozwiązania problemu niejednoznaczności wyrazów jest automatyczne wytrenowanie reguł na bazie korpusów tekstów. Algorytmy dezambigiuacji były dotychczas opracowywane najpowszechniej dla języka angielskiego (Ide, Veronis, 1998; Kilgariff, 2000), nie były natomiast prowadzone badania dla języka polskiego. Wykonawcy chcą rozszerzyć rozwiązania wypracowane dla języka angielskiego poprzez:

- wykorzystanie posiadanych zasobów leksykalnych, głównie słowników zawierających obszerny opis słów (ramy walencyjne, klasy semantyczne, kwalifikatory i dużą liczbę wyrazów złożonych),
- wykorzystanie zebranych dużych korpusów tekstów,
- połączenie metody maszynowego uczenia się (*Machine Learning*) z regułami tworzonymi manualnie.

4.4. Implementacja nowych algorytmów analizy składniowej na podstawie gramatyk ekstrahowanych z korpusów lingwistycznie anotowanych

W systemach tłumaczących opartych na regułach transferu, analizator składniowy systemu jest jednym z kluczowych komponentów. Jakość tłumaczenia jest w sposób bezpośredni uzależniona od jakości wstępnej analizy składniowej języka źródłowego.

W przypadku języków obcych, które zostaną uwzględnione w projekcie, istnieją korpusy anotowane składniowo (tzw. *Treebank Grammars*). Anotacja została stworzona ręcznie przez zespoły specjalistów, jakość tych modeli językowych jest w związku z tym bardzo wysoka. (Telljohann, Hinrichs, Kübler 2004; Marcus, Santorini, Marcinkiewicz, 1993).

Ekstrahując reguły analizy składniowej z tych modeli, można uzyskać zestawy ponad 10 000 reguł (dla porównania zestawy reguł składniowych wytworzonych przez człowieka nie przekraczają na ogół liczby 1000) oraz pełny model probabilistyczny, wykorzystywany przy ujednoznacznianiu wyników analizy.

Jednak wraz ze wzrostem liczby reguł zwiększają się wymagania czasowe i pamięciowe algorytmów analizy składniowej. Aby móc korzystać z podwyższonej

jakości analizy, trzeba równocześnie zagwarantować szybkość i niezawodność działania systemu. Metody, z których skorzystamy w tym celu, pochodzą z dziedziny sztucznej inteligencji, a przykładem może być zastosowanie wyszukiwania A* do analizy składniowej (Klein, Manning 2003).

4.5. Opracowanie i zaimplementowanie algorytmów wyszukiwania informacji

W celu wyszukiwania w pamięci tłumaczeń segmentów podobnych do tłumaczonych zastosowane zostaną algorytmy z dziedziny wyszukiwania informacji (*Information Retrieval*). Popularna metoda wyszukiwania informacji – *inverted-file index* – stosowana była we wczesnych implementacjach wyszukiwarek internetowych. Jej złożoność pesymistyczna wynosiła początkowo O(n), gdzie n oznacza liczbę dokumentów, lecz algorytm Browna (Brown, 2004) poprawił złożoność problemu do O(log n).

Wykonawcy projektu planują rozwinięcie metody opartej na modelu przestrzeni wektorowej (*Vector Space Model*). Metoda ta przyjmuje podejście „*bag-of-words*", ignorujące kolejność wyrazów w zdaniu. Jej główną zaletą jest możliwość wyznaczenia stopnia podobieństwa między informacją szukaną i znalezioną (Baldwin, 2004). W zastosowaniu do przeszukiwania pamięci tłumaczeń model przestrzeni wektorowej umożliwi uszeregowanie jednostek tłumaczenia według stopnia podobieństwa do szukanego zdania (fragmentu tekstu).

5. Prace rozwojowe

Projekt ma charakter rozwojowy. Oznacza to, że prace badawcze zostaną wdrożone w działającym systemie informatycznym. W tym celu wykonanych zostanie szereg zadań mających charakter wdrożeniowy.

5.1. Zadania związane z korpusami tekstów

Zadania związane z korpusami tekstów mają na celu uzyskanie wielojęzycznych słowników oraz pamięci tłumaczeń w dziedzinie bezpieczeństwa za pomocą statystycznych algorytmów analizy korpusów tekstowych. Można wyróżnić następujące etapy prac:

5.1.1. Zebranie dwujęzycznych i wielojęzycznych korpusów tekstów równoległych z zakresu bezpieczeństwa publicznego

Korpusy stanowić będą bazę, z której wyodrębniony zostanie automatycznie słownik fraz i ich tłumaczeń (zadania 5.1.5, 5.1.6). Ponadto na podstawie korpusów zostanie utworzona pamięć tłumaczeń (zadanie 5.1.3).

5.1.2. Dopasowanie (zrównoleglenie) dokumentów na poziomie akapitów i zdań

W wyniku tego działania otrzyma się dokumenty, w których oznaczone będą odpowiedniości pomiędzy poszczególnymi akapitami i zdaniami a ich tłumaczeniami.

5.1.3. Stworzenie pamięci tłumaczeń

Dokumenty zrównoleglone w zadaniu 5.1.2. zostaną zaimportowane do pamięci tłumaczeń. Docelowo pamięć tłumaczeń ma zawierać co najmniej 8 milionów jednostek. Jednostką pamięci tłumaczeń jest najczęściej jedno zdanie podane w kilku „wariantach", czyli różnych językach, ale niekiedy jednostkę stanowi część zdania lub kilka zdań.

5.1.4. Dopasowanie (zrównoleglenie) dokumentów na poziomie wyrazów

W wyniku tego działania uzyska się dokumenty, w których oznaczone będą odpowiedniości między poszczególnymi wyrazami tłumaczonych dokumentów. Jest to krok niezbędny dla uzyskania słownika terminologii (zadanie 5.1.5).

5.1.5 Automatyczna ekstrakcja wielojęzycznej terminologii związanej z bezpieczeństwem (na podstawie dokumentów wielojęzycznych zrównoleglonych na poziomie wyrazów)

Otrzymane w tym zadaniu frazy i ich tłumaczenia mogą zawierać błędy i nieścisłości. Dlatego niezbędna będzie ich ludzka weryfikacja – w zadaniu 5.1.6.

5.1.6. Stworzenie specjalistycznego słownika wielojęzycznego z zakresu bezpieczeństwa

Proces weryfikacji terminologii zebranej automatycznie jest nieporównanie bardziej wydajny od pozyskiwania danych słownikowych „od początku".

5.2. Zadania związane z funkcją tłumaczenia automatycznego

5.2.1. Automatyczne pozyskanie reguł analizy języków z anotowanych jednojęzycznych korpusów tekstowych

Reguły uzyskane w sposób automatyczny bardziej obiektywnie odzwierciedlają opisywany język niż reguły przygotowane manualnie.

5.2.2. Automatyczne pozyskanie reguł transferu z dokumentów zrównoleglonych

Posiadanie dużych korpusów tekstów wielojęzycznych, zrównoleglonych na poziomie wyrazów, umożliwi automatyczne pozyskiwanie reguł transferu.

5.2.3. Wykorzystanie systemu tłumaczącego do generowania tłumaczeń nowych jednostek słownikowych nienapotkanych w korpusach równoległych

Może zdarzyć się tak, że dla tekstów zebranych w jednym języku nie będą istniały ich odpowiedniki w innych językach (taka sytuacja ma miejsce na przykład w przypadku większości polskich aktów prawnych). Zadanie to ma na celu zapewnienie poprawnego tłumaczenia terminologii występującej w tego typu dokumentach.

5.2.4. Optymalizacja szybkości działania systemu tłumaczącego poprzez opracowanie nowych algorytmów analizy

Zadanie będzie polegało na wykorzystaniu algorytmów zmniejszających przestrzeń przeszukiwań w analizie składniowej. Szybkość działania stanie się kluczowa, gdyż zestawy reguł pozyskane automatycznie (zadanie 5.2.1) będą znacznie liczniejsze niż zestawy opracowywane przez człowieka.

5.2.5. Opracowanie algorytmów obsługi pamięci tłumaczeń

Niezbędne jest udoskonalenie opracowanych już algorytmów zweryfikowanych dotychczas na bazach zawierających ok. 100 000 jednostek tak, aby można je było wykorzystać w pamięci tłumaczeń o wielkości ponad 8 milionów jednostek.

5.3. Zadania techniczne związane z wymaganiami funkcjonalnymi systemu tłumaczenia

5.3.1. Analiza formatów .txt, .htm, .xml, .doc, .pdf

5.3.2. Zapewnienie integracji z najnowszymi wersjami aplikacji (MS Office, Open Office, przeglądarki internetowe, programy pocztowe)

5.3.3. Zapewnienie działania systemu na najnowszych systemach operacyjnych (np. Windows Vista)

5.4. Zadania związane z prototypem systemu tłumaczenia mowy

5.4.1. Integracja systemu tłumaczenia tekstów z systemem rozpoznawania mowy

Celem tego zadanie będzie zintegrowanie rozwiązań wypracowanych w opisywanym projekcie i projekcie kierowanym przez prof. G. Demenko.

5.4.2 Prace rozwojowe nad syntezą mowy polskiej

Celem jest uzyskanie systemu tłumaczenia *speech-to-speech* z wysokiej jakości syntezą mowy. We wnioskowanym projekcie prace będą koncentrować się nad zagadnieniami oddania poprawnej intonacji czytanego przez maszynę tekstu.

5.4.3. Integracja systemu tłumaczenia tekstów z systemem syntezy mowy

5.5. Testowanie zewnętrzne i wdrażanie systemu

Zadania tej grupy stanowią standardową procedurę wdrażania systemu informatycznego:

5.5.1. Zorganizowanie pilotażowej grupy użytkowników

5.5.2. Szkolenie pilotażowej grupy użytkowników

5.5.3. Analiza danych zebranych od grupy pilotażowej

5.5.4. Wykonanie raportów i przekazanie uwag do zespołu informatycznego

5.5.5. Zainstalowanie programu serwera

5.5.6. Zainstalowanie programów klientów

5.5.7. Szkolenie użytkowników

6. Efekt końcowy

Wynikiem końcowym projektu będzie **system tłumaczenia automatycznego** wysokiej jakości z dziedziny bezpieczeństwa publicznego. System zostanie **wdrożony w Komendzie Wojewódzkiej Policji w Poznaniu**.

System będzie zrealizowany w technologii klient–serwer. Główne funkcjonalności systemu to:

1) tłumaczenie dokumentów różnych formatów: .txt, .htm, .xml, .doc, .pdf
2) integracja z podstawowymi aplikacjami przeglądania i przetwarzania dokumentów:
   - Microsoft Word
   - Microsoft Excel
   - Microsoft PowerPoint
   - Open Office Write

- Open Office Calc
- Microsoft Outlook
- Outlook Express
- Firefox Thunderbird
- Internet Explorer
- Mozilla Firefox

Przed wdrożeniem system zostanie poddany szczegółowej fazie testowania przez pilotażową grupę użytkowników.

Opracowana zostanie dokumentacja techniczna systemu oraz podręcznik użytkownika.

Opracowany zostanie ponadto **prototyp systemu tłumaczenia mowy** z uwzględnieniem wyników uzyskanych w projekcie „Technologie przetwarzania oraz rozpoznawania informacji słownych w systemach bezpieczeństwa wewnętrznego".

Literatura

1. Baldwin T., 2004, *Translation Memory Engines: A Look under the Hood and Road Test*, w: Proceedings of the 15th International Japanese/English Translation Conference (IJET-15) Yokohama, Japan.
2. Banerjee S., Lavie A., 2005. METEOR: *An automatic metric for Mt evaluation with improved correlation with human judgments*, w: Proceedings of ACLWorkshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.
3. Brown R.D., 2004, *A Modified Burrows-Wheeler Transform for Highly-Scalable Example-Based Translation*, w: *Machine Translation: From Real Users to Research*, Proceedings of the 6th Conference of the Association for Machine Translation (AMTA-2004).
4. Chen Jiang, JianYun Nie, 2000, *Web parallel text mining for Chinese English*. *Crosslanguage information retrieval*, w: International Conference on Chinese Language Computing, Chicago, Illinois.
5. Doddington G., 2002, *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*, w: HLT-01.
6. HunAlign, 2007 (http://mokk.bme.hu/resources/hunalign).
7. Ide N., Veronis J., 1998, *Word sense disambiguation: the state of art*, w: *Computational Lingustics*, 24(1), 1–40.

8. Jassem K., Lipski J., 2007, *A new tool for the bilingual text aligning at the sentence level*, w: Proceedings of Intelligent Information Systems, Zakopane 2008, http://iis.ipipan.waw.pl/2008/proceedings.html.
9. Kilgariff A., Palmer M., 2000, *Introduction to the Special Issue on SENSEVAL*, "Computers and the Humanities".
10. Klein D., Manning C.D., 2003, *A* Parsing: Fast Exact Viterbi Parse Selection*, w: *HLT-NAACL 2003*.
11. Koehn P., Schroeder J., 2007, *Experiments in Domain Adaptation for Statistical Machine Translation*, w: ACL Workshop on Statistical Machine Translation.
12. Koehn P., Hoang H., 2007, *Factored Translation Models*, w: Conference on Empirical Methods in Natural Language Processing 2007.
13. Koehn P., Knight K., 2003, *Feature-Rich Statistical Translation of Noun Phrases* (w:) Proceedings of ACL-2003.
14. Ma Xiaoyi, M. Liberman, 1999, Bits: *A method for bilingual text search over the web* (w:) Machine Translation Summit VII, September. Również: http://www.ldc.upenn.edu/Papers/MTSVII1999/BITS.ps.
15. Marcus M., Santorini B., Marcinkiewicz M., 1993, *Building a large annotated corpus of English: the Penn Treebank*, w: Computational Linguistics, 19.
16. Moore R.C., 2002, *Fast and Accurate Sentence Alignment of Bilingual Corpora*, w: *Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany.
17. Moore R.C., 2005, *Association-Based Bilingual Word Alignment*, w: Proceedings, Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, Ann Arbor, Michigan.
18. Och F.J., Ney H., 2003, *A Systematic Comparison of Various Statistical Alignment Models*, w: *Computational Linguistics*, 29(1).
19. Papineni K., Roukos S., Ward T., Zhu W.J., 2002, *Bleu: a method for automatic evaluation of machine translation*, w: ACL-02, Philadelphia, PA.
20. Resnik P., 1998., *Parallel strands: A preliminary investigation into mining the Web for bilingual text*, w: Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA98.

21. Rosińska M., 2007, *Collecting Polish-German Parallel Corpora*, w: Proceedings of the International MultiConference on Computer Science and Information Technology, Volume 2.
22. Telljohann H., Hinrichs E.W., Kübler S., 2004, *The TüBa-D/Z Treebank – Annotating German with a Context-Free Backbone*. w: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)

Nota o autorze

Krzysztof Jassem (jassem@amu.edu.pl) jest doktorem habilitowanym nauk technicznych w specjalności informatyka. Jest zatrudniony na Uniwersytecie im. Adama Mickiewicza w Poznaniu na Wydziale Matematyki i Informatyki. Obszarem działalności naukowej jest lingwistyka komputerowa, a w szczególności tłumaczenie automatyczne oraz słowniki wielojęzyczne. Pełni funkcję Prezesa firmy informatycznej Poleng Sp. z o.o., której głównym produktem jest system tłumaczący Translatica. Był koordynatorem prac nad Wielkim Słownikiem Polsko-Niemieckim oraz Wielkim Słownikiem Niemiecko-Polskim, wydanym przez Wydawnictwo Naukowe PWN.