Streszczenie

W pracy została poruszona kwestia możliwości ulepszenia tłumaczenia maszynowego między językami polskim i angielskim. Podczas, gdy dla wielu popularnych języków istnieją doskonałe systemy tłumaczeniowe, w wypadku zestawienia polsko-angielskiego jest jeszcze wiele do zrobienia. Niniejsze badanie opiera się głównie na technikach tłumaczenia maszynowego. Wykorzystano metody słownikowe [145], regułowe [146] oraz składniowe [147]. Najpopularniejsze aplikacje i metodologie nie są odpowiednio przystosowane do struktur języka polskiego, wymagają więc adaptacji. Ponadto, w dotychczasowych rozwiązaniach brakuje zasobów łączących język polski i angielski – w szczególności danych równoległych. Dla języka polskiego problemem jest również brak dostępności danych jednojęzycznych. Dlatego też głównym przedmiotem rozprawy była budowa automatycznego i niezawodnego systemu tłumaczeniowego pomiędzy językami polskim a angielskim. Głównym celem było zaś spełnienie określonych wymogów tłumaczeniowych i rozwinięcie dwujęzycznej bazy tekstowej poprzez użycie informacji z korpusów porównywalnych.

Eksperymenty zostały oparte głównie na mowie codziennej. Na bazę użyć poszczególnych zwrotów złożyły się przykłady zaczerpnięte z wykładów [15], napisów filmowych [14] oraz zapisu posiedzeń Parlamentu Europejskiego [36] i Europejskiej Agencji Leków [35]. Celem badań było przeprowadzenie wnikliwej analizy istniejących problemów i poprawienie jakości podstawowych systemów. Przykładem może być próba przystosowania technik i parametrów szkoleniowych w celu ulepszenia wyników miary BLEU (Bilingual Evaluation Understudy) [27] w celu uzyskania maksymalnej jakości tłumaczeń. Kolejnym celem było stworzenie dodatkowych źródeł danych dwujęzycznych i jednojęzycznych przy użyciu dostępnych źródeł internetowych oraz poprzez wydobywanie informacji z korpusów porównywalnych. Do tego celu wykorzystano maszynę wektorów nośnych i algorytm Needleman'a-Wunsch'a [19] oraz narzędzia specjalistyczne działające w określonej procedurze. Dane wyjściowe zostały wykorzystane do wzbogacenia informacji systemów tłumaczeniowych. Przystosowano je do konkretnych tekstów metodą interpolacji liniowej [82] i Filtrowania Moore'a-Lewisa [60]. Kolejnym celem pośrednim była analiza dostępnych danych i poprawa ich jakości, a co za tym idzie podniesienie jakości danych wyjściowych systemów tłumaczeń maszynowych. Stworzono specjalistyczne narzędzie do zrównoleglania oraz filtrowania bilingwalnych korpusów.

Praca ta jest nowatorska dla badań prowadzonych nad językiem polskim, szczególnie biorąc pod uwagę niedostatek dobrej jakości danych i niedoskonałość dostępnych systemów tłumaczeń maszynowych. Ważne jest to, że opracowanie w dużej mierze spełnia naglące wymogi dotyczące tłumaczeń. Badanie to dowodzi, iż systemy w nim opisane mają duży potencjał wspomagania wiernych tłumaczeń trudnych tekstów przy minimalnej ilości błędów. Choć nastąpił wyraźny postęp, należy mimo wszystko podkreślić konieczność przeprowadzenia dalszych badań w celu udoskonalenia systemów tłumaczenia maszynowego. Mimo to ulepszenia w tłumaczeniu zostały pozytywnie ocenione podczas kampanii ewaluacyjnych IWSLT[1]. Portale Wikipedia i Euronews w zadowalający sposób posłużyły za źródła zdań równoległych oraz porównywalnych. Zaprezentowano też udoskonalenia systemów tłumaczeń maszynowych przy użyciu porównywalnych korpusów. Dla potrzeb języka polskiego wprowadzono dodatkowe ulepszenia w mierze BLEU.