1 Wprowadzenie

Tłumaczenie między językami naturalnymi ma duże znaczenie praktyczne. Od dawna próbowano je zautomatyzować.

Problemy:

- wieloznaczne słowa
- zmiany w szyku zdania
- słowa funkcyjne i konstrukcje gramatyczne zmieniają znaczenie słów

Najprostszą metodą jest użycie słownika, każdemu słowu z języka źródłowego przypisując jego odpowiednik ze słownia. Przykłady:

Boy goes home

może dać

Chłopiec idzie dom

Jest to sensowne, ale brzmi niezbyt dobrze.

Duch jest dobry

może dać

Spirit is good

Co z kolei może dać

Spirytus jest dobry

Powyższy przykład jest sztuczny, ale polskie tłumaczenie książki "Projektowanie obiektowe"(autorzy P. Coad, E. Yourdon) zawiera zdanie:

W niektórych sklepach można wręcz odnieść wrażenie, że maszyna czasu przeniosła nas w rok 1968.

W kontekście książki słowo *sklepach* brzmi dziwnie. Ale jak sobie to z powrotem przetłumaczymy na angielski to sprawa się wyjaśni:

In some shops ...

Angielskie słowo *shop* to zwykle jest *sklep* ale może to być *warsztat* czy też *instytucja*. W tym przypadku ostatenie znaczenie jest najbardziej sensowne, ale tłumacz się pomylił i wziął najbardziej popularne znaczenie.

Pokazuje to że wieloznaczność słów może łatwo prowadzić do przekłamań.

Przez wiele lat panowało przekonanie że by poprawić jakość tłumaczenia należy przeprowadzić analizę gramatyczną (rozbiór) zdania źródłowego. Proponowano użycie przypadków dla ograniczenia wieloznaczności, np. informacje gramatyczne pozwalają stwierdzić że właściwym tłumaczniem w pierwszym przykładzie jest

Chłopiec idzie do domu

Niestety, okazało się to trudne. W praktyce lepsze wyniki osiągano stosując metody ad hoc. Dużym problem była ilość pracy potrzebna do stworzenia systemu tłumaczenia, zarówno metody ad hoc jak i bardziej systematyczne podejścia wymagały starannego ręcznego kodowania dużych ilości informacji.

W początkach tłumaczenia maszynowego próbowano użyć podejście statystyczne. Jednak pierwsze próby nie dały zadawalających wyników i przez wiele lat panowało przekonanie że metody statystyczne są nieprzydatne do tłumacznia. Zmieniło się to po roku 1980. Wtedy zespół z IBM zaproponował statystyczną metodę tłumacznia. Podstawą był tzw. tekst równoległy, to znaczy długi tekst z odpowiadającym mu tłumaczeniem. Zespół z IBM używał Hansard, tzn. protokóły posiedzeń parlamentu Kanadyjskiego. W Kanadzie są dwa języki urzędowe, angielski i francuski. Wystąpienia w jednym języku są tłumaczone na drugi, tak że w efekcie dostajemy tekst równoległy w języku angielskim i francuskim. Nowsze prace w Europie używały protokóły Parlamentu Europejskiego które są tłumaczone na wszystkie jezyki oficjalne Uni Europejskiej.

Istotną trudnością dla metod statystycznych jest rzadkość danych. Konsekwencją prawa Zipfa jest to że większość słów występuje rzadko. Poniżej jako ilustrację pokazuję analizę dziennika ustaw z 2013 roku. Dane wejściowe to były publicznie dostępne pliki PDF. Skonwertowałem je do tekstu i utworzyłem listę słów wraz z częstościami wystąpień. Ze względu na proces konwersji jest sporo śmieci, jednak jakościowo wyniki dla lepszych tekstów (również dla innych języków) są podobne. Języki bez fleksji przy danej długości tekstu będą miały mniej różnych słów, jednak w typowym tekście słowa występujące jednokrotnie stanowią sporą część wszystkich słów.

Kategoria	licznik	suma
do 1	704907	704907
do 2	265986	531972
do 5	205125	756315
do 10	99368	749162
do 18	54597	760006
do 30	33055	783996
do 50	23791	932359
do 100	22108	1567511
do 250	18075	2835379
do 600	9664	3689787
do 1500	5023	4605421
do 10000	3171	10618403
do 100000	455	10634595

```
do 1000000          39              9466664
do 10000000          1              1225113
Razem        1445365                49861590
```

Widać też że jest kompensujący czynnik: cały tekst ma ponad 49 milionów słów, słów występujących jednokrotnie jest 704907 czyli ok. 1.4 procent.

Zwykle o tekście równoległym zakłada się że jest on podzielony na zdania, tak że odpowiadające sobie zdania są swoimi tłumaczeniami. Zwykle surowe teksty źródłowe nie są tak podzielone, ale wyodrębnianie zdań nie jest zbyt trudne, a odpowiedniość między zdaniami można wyznaczyć metodami statystycznymi.

Najnowsze metody tłumaczenia używają sieci neuronowe (uczymy je na tekście równoległym), jednak dalej powiemy o metodach statystycznych.

2 Model zaszumionego kanału

Modele IBM traktują tłumaczenia jako problem odtwarznia sygnału który przesłano przez zaszumiony kanał transmisyjny: tekst w naszym języku N jest po przejściu przez kanał transmisyjny zniekształcony i wychodzi jako tekst w języku obcym F.

Aby uzyskać z powrotem tekst N maksymalizujemy prawdopodobieństwo warunkowe $P(N|F)$, czyli obliczamy

$$\operatorname*{argmax}_N P(N|F)$$

Mamy

$$P(N|F) = \frac{P(N \cap F)}{P(F)} = \frac{P(N \cap F)P(N)}{P(N)P(F)} = \frac{P(F|N)P(N)}{P(F)}$$

czyli

$$\operatorname*{argmax}_N P(N|F) = \operatorname*{argmax}_N \frac{P(F|N)P(N)}{P(F)} = \operatorname*{argmax}_N P(F|N)P(N)$$

gdzie ostatnia równość zachodzi bo $P(F)$ jest niezależne od N i nie wpływa na to gdzie jest osiągane maksimum.

Powyższe przekształcenie jest pomocne, bo przy bezpośrednim użyciu $P(N|F)$ potrzebujemy bardzo dobre oszacownie prawdopodobieństwa. W $P(F|N)P(N)$ mamy rozdzielne różne zadania: $P(F|N)$ dba o to by F i N dobrze sobie odpowiadały, ale nie musi się troszczyć o to czy N i F są dobrze zbudowanie. O jakość N troszczy się $P(N)$, tu używamy danych tylko dla

jednego języka co jest łatwiej dostępne niż teksty równoległe. Zalety tego łatwo widać w przypadku Modelu 1 IBM: $P(F|N)$ jest nieczułe na zmiany kolejności słów, ale $P(N)$ preferuje zdania z właściwą kolejnością.

Model zaszumionego kanału pojawił się w zagadnieniach technicznych. Przy przetwarzaniu języka był wcześniej użyty do rozpoznawania mowy, model tłumacznia zaadoptował i rozwinął techniki stosowane wcześniej dla mowy.

3 Model języka

W statystycznym tłumaczeniu maszynowym typowo używa się modele Markowa na słowach. Klasyczny model Markowa na zbiorze W każdej parze w, v elementów W przypisuje prawdopobieństwo przejścia $p_{w,v}$ tak że

$$P(w_n = w|w_1 w_2 \dots w_{n-1}) = P(w_n = w|w_{n-1}) = p_{w,w_{n-1}}.$$

W zagadnieniach języka taki model często jest za słaby bo modeluje tylko zależności między kolejnymi słowami.

W praktyce używa się modele Markowa wyższego rzędu. Teoretycznie Markowa rzędu k można zdefinowiać jako model Markowa na W^k, taki że

$$p_{w_1 w_2 \dots w_k, v_1 v_2 \dots v_k}$$

jest różne od zera tylko wtedy gdy dla $i = 1, 2, \dots, k-1$ zachodzi $w_i = v_{i+1}$. Innymi słowy,

$$P(w_n = w|w_1 w_2 \dots w_{n-1}) = P(w_n = w|w_{n-k} w_{n-k+1} \dots w_{n-1}) = p_{w,w_{n-k} w_{n-k+1} \dots w_{n-1}}$$

i trzeba tylko podać $p_{w,v}$ gdzie w to słowo zaś v to ciąg słów długości k. W praktyce k często jest większe lub równe 4, a więc chcemy szacować częstości wystąpień ciągów słów długości co najmniej 5. Typowy język zawiera więcej niż $100000 = 10^5$ słów, toteż dla rzędu 4 trzeba $(10^5)^5 = 10^{25}$ prawdopodobieństw. Jasne jest że dla tekstów praktycznych rozmiarów nie da się oszacować wszystkich prawdopodobieństw przez częstości wystąpień, większość ciągów będzie miała częstość zero. Używa się tu estymację Bayesa. Dla większych k często trafiamy na ciągi których nie było w zbiorze treningowym i dlatego mamy tylko szacowanie Bayesowskie. Ale w takim przypadku któtszy ciąg może być w zbiorze treningowym i ma sens użyć pochodzące stąd oszacowanie. Praktycznie realizuje się to przez kombinację wypukłą modeli różnych rzędów:

$$q_{w,v} = \sum_{i=1}^{k} c_i p_{w,v_i v_{i+1} \dots v_k}$$

gdzie c_i to waga modelu rzędu $k + 1 - i$.

Komentarz praktyczny: prawdopodobieństwa mogą być bardzo małe, tak że w arytmetyce komputerowej są nieodróżnialne od zera. Dlatego zwykle pracuje się z logarytmami. Logarytmy upraszczają mnożenie zaś komplikują dodawanie. Na współczesynch komputerach oznacza to że obliczenia na logarytmach będą bardziej kosztowne niż bezpośrednie operacje na prawdopodobieństwach. Ale koszt logarytmów zwykle jest mniejszy niż koszt wielokrotnej precyzji potrzebnej by uzyskać odpowiedni zakres liczb. Bardziej pracochłonną, ale obliczniowo szybszą alternatywą może być skalowanie.