INVESTIGATIONES LINGUISTICAE, VOL. XVI; POZNAŃ, DECEMBER 2008

Wprowadzenie do metod statystycznych w tłumaczeniu automatycznym

An introduction to statistical methods in machine translation

Marcin Junczys-Dowmunt[*]

Institute of Linguistics, Adam Mickiewicz University
ul. Aleja Niepodległości 4, 61-874 Poznań, POLAND

junczys@amu.edu.pl

Abstract

The intention of this article is to provide a concise introduction to the basic mathematical concepts of statistical translation models as they were introduced by Brown et al. (1993) in their groundbreaking work *The Mathematics of Statistical Machine Translation: Parameter Estimation*. We concentrate on a simplified description of the first two translation models known as IBM Model 1 and 2. It is one major aim of this work to serve as tutoring material for students of computational linguistics, mathematics or computer science and therefore a lot of comments, additional examples and step-by-step explanations are given, augmenting the original formula by Brown et al. (1993). For both discussed models the calculations for a small parallel corpus are described in detail.

1 Wstęp

Do końca lat osiemdziesiątych technologie związane z tłumaczeniem automatycznym były oparte na różnych rodzajach reguł: regułach składniowych, regułach transferu leksykalnego, regułach generowania języka, regułach morfologicznych itp. Na rynku polskim większość komercyjnych systemów tłumaczenia automatycznego (prawdopodobnie nawet wszystkie) nadal należy do tego typu systemów. Jednak w roku 1989 dominacja systemów regułowych w nauce[1] została przerwana przez pojawienie się nowych metod, opartych na tzw. korpusach równoległych.

Najbardziej znaczące okazało się odrodzenie metod statystycznych, co pojmowano jako powrót empiryzmu przeciwstawiającego się racjonalizmowi systemów regułowych. Metody stochastyczne odniosły spore sukcesy w dziedzinie rozpoznawania mowy, co skłoniło grupę badaczy w IBM do ponownego przyjrzenia się zastosowaniom statystyki do tłumaczenia automatycznego. Wyniki tego zespołu zostały opisane w pracy *The Mathematics of Statistical Machine Translation: Parameter Estimation* (Brown, Della Pietra, Della Pietra i Mercer, 1993), która stała się jedną z najbardziej wpływowych publikacji w dziedzinie tłumaczenia automatycznego i pokrewnych działów lingwistyki komputerowej. W niniejszej pracy przeglądowej postaramy się przedstawić część wyników Brown et al.

[*]Marcin Junczys-Dowmunt jest stypendystą Fundacji Uniwersytetu im. Adama Mickiewicza w Poznaniu na rok 2009.

[1]W sektorze komercyjnym dopiero w ostatnim czasie zaczęły pojawiać się odpowiednie systemy.

(1993) w sposób nieco bardziej przystępny niż uczyniono to w oryginalnym dziele. Skupimy się głównie na aspektach matematycznych tłumaczenia statystycznego z ograniczeniem do zagadnień estymacji parametrów omawianych modeli. Modele te oparte są na relacjach równoważności tłumaczeniowej zachodzących między wyrazami, inne informacje lingwistyczne nie są brane pod uwagę.

Od czasu pojawienia się omawianej pracy statystyczne metody tłumaczenia stały się najszerzej badanym paradygmatem tłumaczenia maszynowego. Relacje między wyrazami zostały zastąpione relacjami między frazami, wprowadzono dodatkową wiedzę lingwistyczną dotyczącą zjawisk składniowych lub morfologicznych, a równocześnie dzięki internetu dramatycznie poprawiła się dostępność korpusów równoległych.

Mimo tego szybkiego rozwoju, praca Brown et al. nie straciła na znaczeniu. Można nawet stwierdzić, że praktycznie wszystkie metody tłumaczenia statystycznego są jedynie mniej lub bardziej zaawansowanymi modyfikacjami modeli przedstawionych już w roku 1993. W przypadku modeli opartych na frazach, zestawianie fraz (konieczny krok wstępny) odbywa się za pomocą oryginalnych modeli IBM, zaimplementowanych dokładnie w takiej postaci, w jakiej zostały opisane 15 lat temu. Dotychczas jedyną licząca się aplikacją, służącą do dopasowywania tekstów równoległych na poziomie wyrazów jest GIZA++ (Al-Onaizan et al., 1999; Och i Ney, 2003), wierna realizacji modeli IBM 1 do 5. W tej pracy skupimy się jedynie na modelach 1 i 2. Wyniki w częściach przykładowych otrzymano z własnej prostej implementacji tych modeli. W tym momencie zachęcamy do samodzielnej próby implementacji opisanych algorytmów, gdyż wbrew pozorom nie jest to zadanie trudne, przynajmniej w przypadku dwóch pierwszych modeli.

2 Statystyczne modele tłumaczenia

W tym rozdziale omówimy podstawowe pojęcia modelu tłumaczenia, dopasowania na poziomie wyrazów oraz korpusu równoległego. Szczególnie ważna okaże się probabilistyczna definicja procesu tłumaczenia.

2.1 Podstawowe równanie tłumaczenia statystycznego

Metody regułowe skupiają się na procesie tłumaczenia. W tym celu rozwija się lingwistycznie uzasadnione sposoby reprezentacji zjawisk językowych, analizy składniowej oraz transferu. Tłumaczenie statystyczne jest ukierunkowane na wynik, a nie na proces. Oznacza to, że tłumaczenie nie musi przebiegać w sposób podobny jak u człowieka[2], ważne tylko, żeby uzyskane tłumaczenie było możliwie najbardziej prawdopodobnym odpowiednikiem zdania źródłowego. Założenia te można sformalizować w następujący sposób:

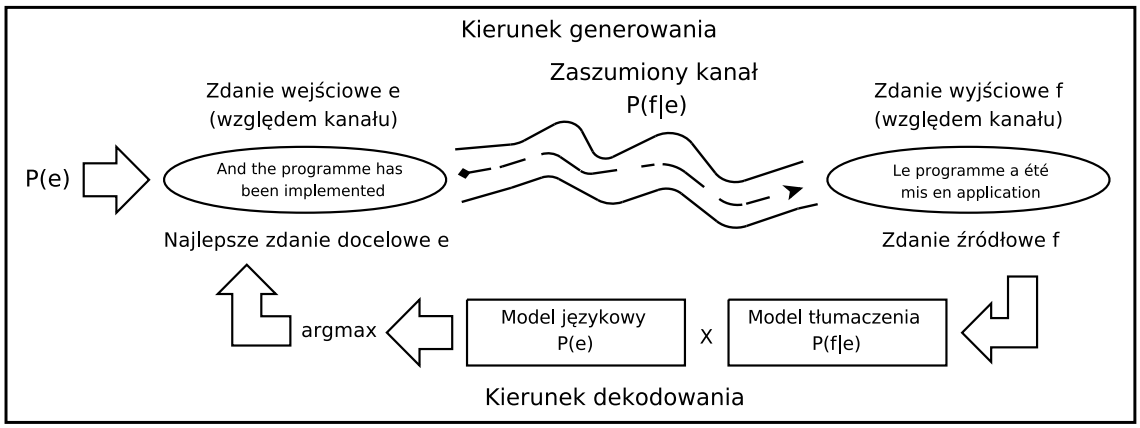
$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\mathbf{E} = \mathbf{e} | \mathbf{F} = \mathbf{f}), \tag{1}$$

gdzie $\mathbf{E}$ oraz $\mathbf{F}$ są zmiennymi losowymi przebiegającymi odpowiednio po wszystkich angielskich zdaniach $\mathbf{e}$ i francuskich zdaniach $\mathbf{f}$.[3] Dla ustalonego zdania francuskiego $\mathbf{f}$ zdanie $\hat{\mathbf{e}}$ maksymalizuje powyższą funkcję prawdopodobieństwa i jest tym samym najbardziej prawdopodobnym tłumaczeniem $\mathbf{f}$. Zastosowanie warunkowego prawdopodobieństwa jest zgodną z intuicją formalizacją faktu, że proces tłumaczenia jest procesem ukierunkowanym, tzn. produkującym zdanie docelowe po zaobserwowaniu (usłyszeniu, przeczytaniu) zdania źródłowego.

Pojawiają się dwa pytania: „Czy na pewno prawdopodobieństwo jest adekwatnym środkiem oceny jakości tłumaczeń?” oraz „W jaki sposób uzyskać prawdopodobieństwo, że zdanie jest tłumaczeniem innego zdania?”. Pytanie pierwsze ma naturę raczej filozoficzną i odpowiedź jest kwestią wysoce sporną. Zdaję się, że znaczące sukcesy w dziedzinie tłumaczenia statystycznego w ciągu

[2] Przy czym wcale nie jest powiedziane, że proces tłumaczenia przez człowieka przebiega tak jak w metodach symbolicznych.

[3] Tutaj oraz w dalszej części pracy będziemy zakładali, że opisujemy system tłumaczący dowolne zdanie języka francuskiego $\mathbf{f}$ na język angielski.



Rysunek 1: Model zaszumionego kanału dla tłumaczenia statystycznego (Jurafsky i Martin, 2000)

ostatnich lat oraz duża popularność tych metod przyczyniły sie do przechylenia szali w stronę odpowiedzi twierdzącej. Pojawiają się jednak głosy, że osiągnięto już szczyt możliwości metod statystycznych.

Według pracy Jurafsky i Martin (2000) tłumaczenie powinno być wierne i płynne, czyli powinno oddawać sens zdania źródłowego możliwie dokładnie, będąc przy tym poprawną albo co najmniej zrozumiałą wypowiedzią w języku docelowym. Takie ujęcie sugeruje pewną modularność problemu tłumaczenia. Powinno być łatwiej, jeśli zamodelowałoby się dwie oddzielne funkcje prawdopodobieństwa dla każdego kryterium, czyli funkcję, która mierzy wierność zdania docelowego względem zdania źródłowego oraz funkcję mierzącą poprawność i płynność zdania docelowego. Metody statystycznego tłumaczenia działają dokładnie w taki sposób. Wystarczy przekształcić równanie (1) za pomocą reguły Bayesa:

$$P(\mathbf{E}=\mathbf{e}|\mathbf{F}=\mathbf{f}) = \frac{P(\mathbf{F}=\mathbf{f}|\mathbf{E}=\mathbf{e})P(\mathbf{E}=\mathbf{e})}{P(\mathbf{F}=\mathbf{f})}, \tag{2}$$

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\mathbf{F}=\mathbf{f}|\mathbf{E}=\mathbf{e})P(\mathbf{E}=\mathbf{e}). \tag{3}$$

Otrzymane równanie (3) jest jednym ze szczególnych przypadków modelu zaszumionego kanału i nosi nazwę *podstawowego równania tłumaczenia statystycznego*. Mianownik $P(\mathbf{F}=\mathbf{f})$ można opuścić, gdyż przy ustalonym $\mathbf{f}$ nie ma żadnego wpływu na wybór najbardziej prawdopodobnego $\mathbf{e}$. Pojawiły się tutaj dwa nowe komponenty, *model tlumaczenia* $P(\mathbf{F}=\mathbf{f}|\mathbf{E}=\mathbf{e})$ oraz *model języka* $P(\mathbf{E}=\mathbf{e})$.

Zastosowanie reguły Bayesa w takim kontekście może być czymś nieoczekiwanym, ponieważ przy tłumaczeniu z języka francuskiego na język angielski pojawia się nagle model, który wydaję się działać w niewłaściwym, odwrotnym kierunku. Jednak taka jest istota modelu zaszumionego kanału. Udaje się, że dane źródłowe zdanie francuskie $\mathbf{f}$ jest zniekształconą wersją jakiegoś angielskiego zdania docelowego $\mathbf{e}$, a zadanie tłumacza polega na „odgadnięciu" tego zdania $\mathbf{e}$, które wygenerowało zaobserwowane zdanie $\mathbf{f}$.

Na rysunku 1 oprócz modelu języka i modelu tłumaczenia umieszczono maszynę dekodującą. Jest to rodzaj automatu, który stara się „pozbyć szumu" powstałego w kanale, czyli właściwy tłumacz. Można to zilustrować za pomocą dziecięcej zabawy w *głuchy telefon*. Jak wiadomo, gra polega na tym, że dziecko na początku kolejki wymyśla jakieś krótkie zdanie szepcząc je do ucha swojego sąsiada. Dzieci są ustawione obok siebie i każde z nich szeptem przekazuje usłyszane zdanie następnemu dziecku. Ostatnie dziecko w kolejce wypowiada zdanie na głos. Gra jest tym zabawniejsza, im większe są zniekształcenia pierwotnych zdań. Wprowadźmy pewne modyfikacje zabawy. Załóżmy, że uczestnicy głuchego telefonu to dzieci polskich emigrantów w Anglii i każde z nich jest bilingwalne.

Pierwsze dziecko wymyśla zawsze zdanie angielskie, ale jakiś żartowniś w środku kolejki nagminnie tłumaczy usłyszane zdanie angielskie na język polski i na koniec kolejki dociera za każdym razem zdanie polskie. Mamy więc zaszumiony kanał. Nauczycielka radzi sobie z sytuacją w następujący sposób: każe ostatniemu dziecku zgadnąć, jakie zdanie angielskie zostało wymyślone przez pierwsze dziecko, gdy tylko usłyszy zdanie polskie.

Dziecko wymyślające zdanie angielskie reprezentuje model językowy, żartowniś powodujący całe zamieszanie jest reprezentantem modelu tłumaczenia, a ostatnie dziecko jest odpowiednikiem maszyny dekodującej. Wystarczy założyć, że każde zdanie polskie przeszło kiedyś przez głuchy telefon oraz że każdy tłumacz jedynie odgaduje jego angielskie źródło. Stąd przy omawianiu modeli tłumaczenia język francuski staję się nagle językiem docelowym a język angielski źródłowym.

Model zaszumionego kanału pozwala więc podejść do tłumaczenia automatycznego metodami probabilistycznymi. Pozostaje drugie pytanie, skąd biorą się prawdopodobieństwa $P(\mathbf{F} = \mathbf{f}|\mathbf{E} = \mathbf{f})$ oraz $P(\mathbf{E} = \mathbf{e})$. W tej pracy odpowiemy jedynie na pierwszą połowę pytania, omijając przy tym modele językowe. Trzeba jednak wyraźnie zaznaczyć, że modele językowe są nieodłączną częścią statystycznych metod tłumaczenia automatycznego. Bez nich wynik tłumaczenia zawierałby z grubsza wszystkie istotne tłumaczenia fragmentów zdania źródłowego, ale ze względu na brak jakichkolwiek zasad poprawności językowej przypominałby niezrozumiałe skupowisko angielskich wyrazów. Model tłumaczenia $P(\mathbf{F} = \mathbf{f}|\mathbf{E} = \mathbf{e})$ nie przeznacza większej masy prawdopodobieństwa na poprawne zdania angielskie niż na niepoprawne, dopiero model języka dokonuje wyboru. Istnieją oczywiście różne sposoby tworzenia modeli języka. W erze internetu, który udostępnia gigantyczne zasoby tekstów jednojęzycznych, pozyskanie dobrych modeli języka nie jest zadaniem trudnym.

2.2 Definicja modelu tłumaczenia

Według Brown et al. (1993) modelem tłumaczenia z języka angielskiego na język francuski[4] P z parametrem $\theta \in \Theta$ nazywamy funkcję obliczającą prawdopodobieństwo $P_\theta(\mathbf{F} = \mathbf{f}|\mathbf{E} = \mathbf{e})$ dla dowolnego zdania francuskiego $\mathbf{f}$ oraz dowolnego zdania angielskiego $\mathbf{e}$. Następujące warunki muszą być spełnione

$$\begin{aligned} &P_\theta(\mathbf{F} = \mathbf{f}|\mathbf{E} = \mathbf{e}) \geq 0, \qquad P_\theta(failure|\mathbf{E} = \mathbf{e}) \geq 0, \\ &P_\theta(failure|\mathbf{E} = \mathbf{e}) + \sum_{\mathbf{f}} P_\theta(\mathbf{F} = \mathbf{f}|\mathbf{E} = \mathbf{e}) = 1, \end{aligned} \tag{4}$$

gdzie $failure$ jest specjalnym symbolem niezaliczanym do zdań francuskich. Suma przebiega po wszystkich możliwych zdaniach francuskich $\mathbf{f}$. $P_\theta(\mathbf{F} = \mathbf{f}|\mathbf{E} = \mathbf{e})$ interpretuje się jako prawdopodobieństwo, że tłumacz wyprodukuje zdanie $\mathbf{f}$ po zaobserwowaniu zdania $\mathbf{e}$ oraz $P_\theta(failure|\mathbf{E} = \mathbf{e})$ jako prawdopodobieństwo, że nie powstanie żadne tłumaczenie zdania $\mathbf{e}$. Od tej pory zamiast $P_\theta(\mathbf{F} = \mathbf{f}|\mathbf{E} = \mathbf{e})$ oraz innych podobnych zapisów będziemy korzystali ze skrótów postaci $P_\theta(\mathbf{f}|\mathbf{e})$ z wyjątkiem sytuacji, gdzie jawne użycie zmiennej losowej będzie służyło celom ilustracyjnym. Wtedy możemy $\mathbf{f}|\mathbf{e}$ potraktować jako symbol oznaczający, że $\mathbf{f}$ jest tłumaczeniem $\mathbf{e}$.

Powyższa definicja modelu tłumaczeń nie mówi nam niczego o sposobie obliczania $P_\theta(\mathbf{f}|\mathbf{e})$ dla dowolnego $\mathbf{f}$ i $\mathbf{e}$. Pozornie najprostszym i zupełnie poprawnym rozwiązaniem byłaby zwykła tablica, która zestawia ze sobą wszystkie zdania angielskie, ich francuskie tłumaczenia oraz — np. oszacowane przez doświadczonego tłumacza — prawdopodobieństwa par spełniające warunki (4). Wiadomo jednak z góry, że nie da się tego zrobić dla wszystkich możliwych par zdań. Nawet gdybyśmy opuścili wszystkie pary o zerowym lub pewnym bardzo małym prawdopodobieństwie, byłoby ich wciąż bardzo dużo, może nawet nieskończenie wiele. Wynika z tego, że zdania nie mogą być atomami naszego modelu; trzeba je rozbić na mniejsze jednostki.

W przypadku modeli, które będziemy omawiali w niniejszej pracy, rolę takich jednostek pełnią wyrazy. Ponieważ językoznawcy mają spore problemy z jednoznaczną definicją tego, co potocznie

[4] Tutaj mamy już do czynienia z odwróconym kierunkiem tłumaczenia wymaganym przez model zaszumionego kanału.

nazywamy wyrazem, posłużymy się definicją informatyczną wyrazu. Jest to w pełni uzasadnione, gdyż opisane tutaj abstrakcyjne modele statystyczne są przeznaczone do implementacji na komputerze. Wyrazem jest więc najdłuższy ciąg znaków, który nie zawiera spacji.[5] Francuskie zdanie $\mathbf{f}$ można więc zapisać jako $\mathbf{f}_1^m = f_1, f_2, \ldots, f_m$, gdzie $f_i, i = 1, 2, \ldots, m$ to francuskie wyrazy. Indeks i zaznacza pozycję wyrazu w zdaniu. Podobnie dla dowolnego zdania angielskiego $\mathbf{e}$ mamy $\mathbf{e}_1^l = e_1, e_2, \ldots, e_l$. Zdarzenie $\mathbf{f}|\mathbf{e}$ sprowadza się wtedy do serii zdarzeń typu $f|e$. Sposób powiązania miedzy tymi zdarzeniami zależy od typu modelu. Wyrazów jest znacznie mniej niż zdań i próby tworzenia w miarę możliwości pełnych tablic tłumaczeń pojedynczych wyrazów w postaci dwujęzycznych słowników papierowych podejmuje się często i wytrwale. Istotnie modele statystyczne tłumaczenia automatycznego korzystają właśnie z takich tablic słownikowych, pozostaje jednak problem pozyskania tablic tłumaczeń oraz przyporządkowania odpowiednich prawdopodobieństw.

2.3 Korpus tłumaczeń

W językoznawstwie statystycznym najpopularniejszym sposobem ustalania prawdopodobieństw jest wykorzystanie częstości lub prawdopodobieństwa *a posteriori*. W przypadku prawdopodobieństwa tłumaczenia zdaje się, że wystarczy wykorzystać duży zbiór tłumaczeń dla interesującej nas pary języków. Taki zbiór tekstów nosi nazwę *korpusu tłumaczeń* lub *korpusu równoległego*. Ominiemy tutaj kwestie związane ze zdobywaniem takiego korpusu i przyjmiemy do wiadomości, że nawet dla języka polskiego jako jednego z dwóch języków istnieją równoległe korpusy składające się z ponad miliona zdań. Zasoby tego rodzaju dla bardziej popularnych par języków, np. dla angielskiego i francuskiego, są oczywiście jeszcze większe.

Żeby uzyskać prawdopodobieństwo *a posteriori* zdarzenia $f|e$ dla ustalonego f i e wystarczy zliczyć ile razy w danym korpusie tłumaczeń wyraz e został przetłumaczony na f oraz podzielić przez liczbę tłumaczeń wyrazu e na dowolny wyraz francuski z korpusu. Oznaczając prawdopodobieństwo tłumaczenia $f|e$ jako $t(f|e)$ oraz liczbę wystąpienia $f|e$ w całym korpusie jako $c(f|e)$, mamy

$$t(f|e) = \frac{c(f|e)}{\sum_f c(f|e)}. \tag{5}$$

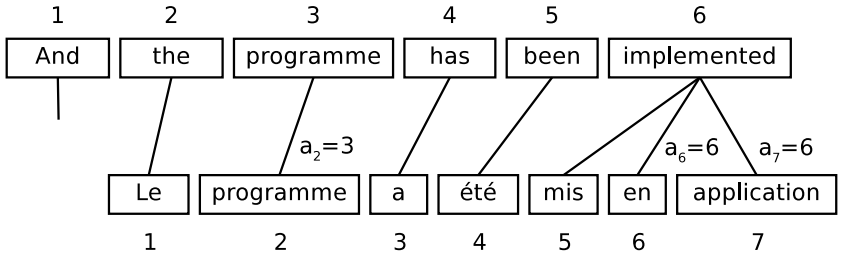
Traktując wszystkie prawdopodobieństwa $t(f|e)$ jako stałe dla ustalonej pary wyrazów $f|e$, uzyskamy wektor parametrów θ dla modelu P. Dokładnie taką postać ma θ w przypadku Modelu 1, który opiszemy nieco później. Przyjmując więc, że stosujemy Model 1 oraz że ustaliliśmy wszystkie $t(f|e)$, możemy obliczyć prawdopodobieństwo $P(\mathbf{f}|\mathbf{e})$ dla dowolnego $\mathbf{f}|\mathbf{e}$.

Ogólna idea opisanego postępowania jest poprawna. Istnieje tylko jeden problem: w korpusie tłumaczeń nie ma nigdzie jawnej informacji, że konkretny wyraz po stronie francuskiej jest tłumaczeniem odpowiadającego mu wyrazu angielskiego. Zakładamy, że mamy do czynienia z korpusem dopasowanym na poziomie zdań. Jedyne informacje dostępne to liczby dotyczące *współwystępowania* (ang. *co-occurrence*) wyrazów francuskich i angielskich w równoległych zdaniach. Wyznaczenie parametrów typu powyższego wymaga jednak równoczesnego wyznaczenia dopasowań na poziomie wyrazów (ang. *word alignment*). Na podstawie dopasowań na poziomie wyrazów jesteśmy wtedy w stanie stwierdzić, które wyrazy są wzajemnymi tłumaczeniami (Model 1), jak zachowuje się szyk wyrazów języka docelowego względem języka źródłowego (Model 2), ile wyrazów docelowych zostało przyporządkowanych jednemu wyrazowi źródłowemu (Model 3) itp.

2.4 Dopasowania tłumaczeń na poziomie wyrazów

Opisane modele tłumaczenia opierają się na koncepcji dopasowań wyrazów, z tego powodu określa się je często jako *modele dopasowań wyrazów* (ang. *word alignment model*). Pamiętamy, że zdanie można przedstawić jako ciąg wyrazów, gdzie indeks przy wyrazie odzwierciedla pozycję wyrazu w

[5] Zgodnie z taką definicją pojedyncza liczba będzie wyrazem, podobnie jak samotny przecinek lub inne znaki interpunkcyjne.



Rysunek 2: Dopasowanie z niezależnymi angielskimi wyrazami

zdaniu. Dopasowaniem pary zdań $\mathbf{f}|\mathbf{e}$ na poziomie wyrazów nazywamy relację $\mathbf{a} \subseteq \{(i,j) : i = 1,\ldots,m \wedge j = 1,\ldots,l\}$, gdzie $\mathbf{f} = \mathbf{f}_1^m$ oraz $\mathbf{e} = \mathbf{e}_1^l$. Dopasowanie można zinterpretować jako zbiór krawędzi w grafie dwudzielnym, w którym dwa rozdzielne zbiory wierzchołków reprezentują odpowiednio pozycje w zdaniu francuskim i angielskim. Krawędź w takim grafie sygnalizuje, że wyrazy powiązane z jej wierzchołkami są wzajemnymi tłumaczeniami. Zbiór $\mathcal{A}(\mathbf{f},\mathbf{e})$ jest zbiorem wszystkich możliwych dopasowań między zdaniami $\mathbf{f}$ i $\mathbf{e}$. Liczba dopasowań w $\mathcal{A}(\mathbf{f},\mathbf{e})$ wynosi 2^{lm}, tyle ile istnieje grafów dwudzielnych z dwupodziałem na m i l różnych wierzchołków. Dla pary zdań z rysunku 2 istnieje więc 2^{42} dopasowań, z których tylko znikoma ilość ma sens. Wprowadza się dodatkowo angielski *wyraz zerowy* e_0, z którym są połączone wszystkie wyrazy francuskie, które nie zostały połączone z którymkolwiek z wyrazów e_1 do e_l. Wyraz zerowy nie zwiększa ogólnej liczby dopasowań w $\mathcal{A}(\mathbf{f},\mathbf{e})$.

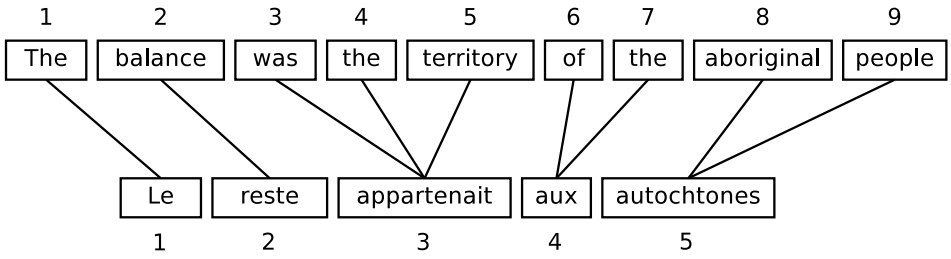
W omawianych modelach ogranicza się możliwe dopasowania między zdaniami $\mathbf{f}$ i $\mathbf{e}$ w taki sposób, że każdy wyraz francuski jest połączony z dokładnie jednym wyrazem angielskim. Relacja $\mathbf{a}$ spełniająca taki warunek jest funkcją i można ją zapisać w postaci $\mathbf{a} \equiv \mathbf{a}_1^m = a_1, a_2, \ldots, a_m$, gdzie $a_i \in \{0,1,\ldots,l\}$. Funkcja $\mathbf{a}$ odwzorowuje pozycję francuskiego wyrazu na pozycję angielskiego wyrazu, którego tłumaczeniem jest wyraz francuski. Takich dopasowań jest $(l+1)^m$, czyli dla przykładu z rysunku 2 mamy 7^7 dopasowań będących funkcjami. Ograniczenie do dopasowań lewostronnie jednoznacznych nie ma uzasadnienia lingwistycznego, autorzy Brown et al. przyjęli je z powodów technicznych. Gigantyczna liczba dopasowań nieograniczonych uniemożliwiłaby przeprowadzenie efektywnych obliczeń potrzebnych do estymacji parametrów modelu tłumaczenia. Nawet mocno okrojony zbiór dopasowań funkcyjnych nie rozwiązuje jeszcze problemu złożoności obliczeniowej, który wynika z następującego faktu:

$$P_\theta(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P_\theta(\mathbf{f},\mathbf{a}|\mathbf{e}). \tag{6}$$

Prawdopodobieństwo $P_\theta(\mathbf{f}|\mathbf{e})$ wyraża się tutaj jako prawdopodobieństwo całkowite $P_\theta(\mathbf{f},\mathbf{a}|\mathbf{e})$ po wszystkich dopasowaniach $\mathbf{a}$. Dla dopasowań, które nie są funkcjami, mamy $P_\theta(\mathbf{f},\mathbf{a}|\mathbf{e}) = 0$. Przykłady z rysunków są najbardziej intuicyjnymi dopasowaniami z pośród wszystkich możliwych dopasowań. Jednak jak pisaliśmy w poprzednim podrozdziale, nie wiemy, które wyrazy są swoimi tłumaczeniami, co jest równoznaczne z faktem, że nie są nam dane najlepsze dopasowania. Stąd musimy sumować po wszystkich możliwych dopasowaniach. Najlepsze dopasowania wyłaniają się stopniowo w procesie estymacji parametrów θ i mają znaczący wpływ na wartości θ.

Bezpośrednią konsekwencją stosowania dopasowań w postaci funkcji jest niezdolność omawianych modeli tłumaczenia do adekwatnej reprezentacji zjawisk językowych, które wychodzą poza narzucone ograniczenia. Wystarczy spojrzeć na przykłady z rysunków 3 i 4, gdzie w pierwszym przypadku jeden wyraz z zdania docelowego może mieć kilka odpowiedników w zdaniu źródłowym. W drugim przykładzie tłumaczenie jest niedosłowne i w ogóle nie można wykonać poprawnego dopasowania na poziomie wyrazów, a na pewno nie takie, które będzie funkcją.

Nowoczesne systemy tłumaczenia statystycznego odrzucają koncepcję, w której proces tłumaczenia odbywa się przez tłumaczenie pojedynczych wyrazów, i opierają się na całych frazach lub



Rysunek 3: Dopasowanie z niezależnymi francuskimi wyrazami

drzewach składniowych. Jednak do wygenerowania dopasowań na poziomie fraz czy drzew nadal wykorzystuje się dopasowania wyrazów uzyskane jako produkt uboczny w trakcie procesu estymacji parametrów modeli opartych na wyrazach. Tworzy się dwa dopasowania korpusu na poziomie wyrazów, odpowiednio zamieniając język źródłowy i język docelowy. Po rekombinacji dwóch dopasowań można uzyskać dopasowania podobne do rysunku 4. Pozostałe parametry modeli tłumaczenia nie są brane pod uwagę. To dzięki opisanym w niej modelom dopasowań praca Brown et al. jest jedną z najczęściej cytowanych prac w lingwistyce komputerowej w ogóle, wykraczając przy tym również poza dziedzinę tłumaczenia statystycznego.

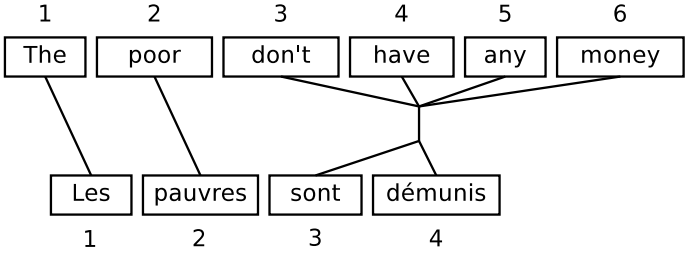
2.5 Model tłumaczenia a model statystyczny

W tej sekcji postaramy się powiązać modele tłumaczenia oraz dotychczas przedstawione informacje z klasycznymi pojęciami rachunku prawdopodobieństwa oraz statystyki. Obiektem matematycznym będącym podstawą omawianych modeli jest rozkład łączny $P(\mathbf{F}=\mathbf{f},\mathbf{A}=\mathbf{a},\mathbf{E}=\mathbf{e})$, gdzie $\mathbf{F}$ i $\mathbf{E}$ są znanymi już zmiennymi losowymi dla francuskich i angielskich zdań, a $\mathbf{A}$ jest zmienną losową opisującą dopasowanie zachodzące między tymi zdaniami. Omawiane modele tłumaczenia powstają przy wykorzystaniu różnych rozkładów brzegowych tego podstawowego rozkładu. Mamy np.

$$\begin{aligned} P(\mathbf{F}=\mathbf{f}|\mathbf{E}=\mathbf{e}) &= \frac{P(\mathbf{F}=\mathbf{f},\mathbf{E}=\mathbf{e})}{P(\mathbf{E}=\mathbf{e})} \\ &= \frac{\sum_{\mathbf{a}} P(\mathbf{F}=\mathbf{f},\mathbf{A}=\mathbf{a},\mathbf{E}=\mathbf{e})}{\sum_{\mathbf{f}}\sum_{\mathbf{a}} P(\mathbf{F}=\mathbf{f},\mathbf{A}=\mathbf{a},\mathbf{E}=\mathbf{e})} \end{aligned} \tag{7}$$

Podobnie za pomocą tego podstawowego rozkładu możemy uzasadnić wzór (6), gdzie

$$\begin{aligned} P(\mathbf{F}=\mathbf{f}|\mathbf{E}=\mathbf{e}) &= \frac{P(\mathbf{F}=\mathbf{f},\mathbf{E}=\mathbf{e})}{P(\mathbf{E}=\mathbf{e})} \\ &= \frac{\sum_{\mathbf{a}} P(\mathbf{F}=\mathbf{f},\mathbf{A}=\mathbf{a},\mathbf{E}=\mathbf{e})}{P(\mathbf{E}=\mathbf{e})} \\ &= \frac{P(\mathbf{E}=\mathbf{e})\sum_{\mathbf{a}} P(\mathbf{F}=\mathbf{f},\mathbf{A}=\mathbf{a}|\mathbf{E}=\mathbf{e})}{P(\mathbf{E}=\mathbf{e})} \end{aligned}$$



Rysunek 4: Ogólne dopasowanie

$$= \sum_{\mathbf{a}} P(\mathbf{F} = \mathbf{f}, \mathbf{A} = \mathbf{a} | \mathbf{E} = \mathbf{e}). \tag{8}$$

Warto rozpatrzyć rozkład brzegowy $P(\mathbf{F} = \mathbf{f}, \mathbf{E} = \mathbf{e})$. Wyeliminowanie zmiennej losowej $\mathbf{A}$ jest tutaj o tyle sensowne, że dopasowania, czyli wartości $\mathbf{A}$, traktujemy jako parametry niejawne, wyłaniające się przez wykonanie kolejnych iteracji algorytmu EM. Możemy wtedy określić łączną zmienną losową (dwuelementowy wektor losowy) $\mathbf{X} = (\mathbf{F}, \mathbf{E})$, w której pary zdań są wartościami. Zmienne losowe $\mathbf{F}$ oraz $\mathbf{E}$ są w oczywisty sposób zależne, ponieważ są ze sobą powiązane jako tłumaczenia — w tym momencie jeszcze bez określenia kierunku tłumaczenia.

Korpus równoległy wykorzystywany do trenowania modelu tłumaczenia można wtedy potraktować jako próbę $\underline{\mathbf{X}} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_S)^T$, niezależnych zmiennych losowych o takim samym łącznym rozkładzie prawdopodobieństwa, gdzie S jest liczbą par zdań w korpusie. Założenie o niezależności poszczególnych zdań jest lingwistycznie wątpliwe. Wystarczy sobie wyobrazić, że korpus zawiera tłumaczenie pewnego stanowiącego logiczną całość tekstu, wtedy treść zdań zależy od innych zdań w tym samym fragmencie tekstu. W praktyce jednak ignoruje się powiązania tego rodzaju.

2.6 Algorytm EM

Algorytm EM (ang. *expectation-maximalization*) (Dempster et al., 1977) jest wykorzystywany w statystyce do wyznaczania wartości największej wiarogodności dla modeli probabilistycznych, gdy model zależy od ukrytych parametrów. Algorytm cyklicznie powtarza dwa kroki: krok obliczania oczekiwanych wartości wiarogodności (E – *expectation step*) w taki sposób, jakby parametry ukryte zostały zaobserwowane, oraz krok maksymalizacji (M – *maximalization step*), który oblicza oszacowania największej wiarogodności parametrów maksymalizując oczekiwane wiarogodności wyznaczone w kroku E. Parametry wyznaczone w kroku M służą wtedy do rozpoczęcia obliczeń w kolejnym kroku E i proces się powtarza aż do zbieżności.

EM jest wykorzystywany do klastrowania danych w uczeniu maszynowym i grafice komputerowej. W przetwarzaniu języków naturalnych ma szczególnie ciekawe zastosowania, np. przy automatycznej indukcji probabilistycznych gramatyk bezkontekstwych oraz modeli Markowa lub właśnie w tłumaczeniu statystycznym. Zilustrujemy jego działania w dalszej części pracy przy estymacji parametrów modelu tłumaczenia.

3 Estymacja parametrów modeli tłumaczenia

Wprowadzamy w tej sekcji ogólną procedurę estymacji parametrów, która jest wspólna dla wszystkich omawianych modeli, również tych bardziej zaawansowanych. W pracy Brown et al. (1993) te informacje zostały uwzględnione jedynie w aneksie wraz ze skąpym komentarzem. Postaramy się dostarczyć brakujące obliczenia i wypełnić ewentualne luki powstałe przez skoki myślowe autorów. Czytelnikom bez matematycznego zacięcia radzimy, nie bać się dużej liczby równań. Nie wykraczają one poza wiadomości nabyte podczas pierwszego roku studiów matematyki lub innych kierunków technicznych. Dla łatwiejszego przyswojenia przedstawionej wiedzy można przeskoczyć od razu do przykładowych obliczeń dla poszczególnych modeli i wracać do rozdziałów matematycznych w razie potrzeby konsultacji.

3.1 Funkcja celu

Wspomnieliśmy już, że trenowanie modelu tłumaczenia P_θ polega na znalezieniu parametrów modelu θ takich, aby maksymalizowały prawdopodobieństwo danego zestawu danych trenujących. Pojmując to zadanie jako zagadnienie optymalizacyjne można wyznaczyć intuicyjną *funkcję celu*

$$\psi(P_\theta) \equiv S^{-1} \sum_{s=1}^{S} \log P_\theta(\mathbf{f}^{(s)} | \mathbf{e}^{(s)}), \tag{9}$$

gdzie przenosząc sumę pod logarytm otrzymalibyśmy iloczyn prawdopodobieństw, że każde zdanie francuskie jest tłumaczeniem swojego angielskiego odpowiednika w danym korpusie równoległym. Zastosowanie logarytmu przynosi kilka korzyści. Po pierwsze, jako że logarytm jest funkcją wklęsłą, pozwala na pokazanie kilku przydatnych faktów dotyczących procesu trenowania przedstawionych modeli tłumaczenia, a po drugie znacznie ułatwia implementację. Iloczyn takich prawdopodobieństw przyjąłby bardzo szybko wartości będące bardzo bliskie zera, które wykroczyłyby poza zakres dokładności zwykłych liczb zmiennoprzecinkowych na dowolnym systemie komputerowym. Korzystając z logarytmu przechodzimy do sumowania liczb ujemnych, które zachowują się o wiele bardziej „przyjaźnie" pod względem obliczeniowym.

Za pomocą funkcji $c(\mathbf{f}, \mathbf{e})$, która podaje liczbę wystąpień danej pary zdań w korpusie podzieloną przez S, uogólnia się funkcję celu dla dowolnej pary zdań równoległych do następującej postaci

$$\psi(P_\theta) = \sum_{\mathbf{f},\mathbf{e}} c(\mathbf{f}, \mathbf{e}) \log P_\theta(\mathbf{f}|\mathbf{e}). \tag{10}$$

3.2 Porównywanie modeli

Model tłumaczenia $P_\theta(\mathbf{f}|\mathbf{e})$ można przedstawić jako sumę prawdopodobieństw dopasowań $\mathbf{a}$ między zdaniami $\mathbf{e}$ i $\mathbf{f}$, co przedstawiliśmy dokładniej w poprzednim rozdziale.

$$P_\theta(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) \tag{11}$$

Wykorzystując taki zapis, porównuje się dwa modele tłumaczenia P_θ oraz $\tilde{P}_{\tilde{\theta}}$ za pomocą *relatywnej funkcji celu*

$$R(\tilde{P}_{\tilde{\theta}}, P_\theta) \equiv \sum_{\mathbf{f},\mathbf{e}} c(\mathbf{f}, \mathbf{e}) \sum_{\mathbf{a}} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f}, \mathbf{e}) \log \frac{P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\tilde{P}_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e})}, \tag{12}$$

gdzie $\tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f}, \mathbf{e}) = \tilde{P}_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e}) / \tilde{P}_{\tilde{\theta}}(\mathbf{f}|\mathbf{e})$. Ważnym dla nas faktem jest to, że dla identycznych modeli z równymi parametrami mamy $R(\tilde{P}_{\tilde{\theta}}, \tilde{P}_{\tilde{\theta}}) = 0$. Powiązanie R z ψ wynika z następującej nierówności

$$\psi(P_\theta) \geq \psi(\tilde{P}_{\tilde{\theta}}) + R(\tilde{P}_{\tilde{\theta}}, P_\theta), \tag{13}$$

która jest przeformułowaniem nierówności Jensena dla modeli tłumaczenia. Wynika ona ze skończonej wersji nierówności Jensena dla funkcji wklęsłych, gdzie dla dowolnego $\mathbf{e}$ i $\mathbf{f}$ mamy

$$\begin{aligned}
\sum_{\mathbf{a}} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f}, \mathbf{e}) \log \frac{P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\tilde{P}_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e})} &\leq \log \sum_{\mathbf{a}} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f}, \mathbf{e}) \frac{P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\tilde{P}_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e})} \\
&= \log \sum_{\mathbf{a}} \frac{\tilde{P}_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\tilde{P}_{\tilde{\theta}}(\mathbf{f}|\mathbf{e})} \frac{P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\tilde{P}_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e})} \\
&= \log \frac{P_\theta(\mathbf{f}|\mathbf{e})}{\tilde{P}_{\tilde{\theta}}(\mathbf{f}|\mathbf{e})} = \log P_\theta(\mathbf{f}|\mathbf{e}) - \log \tilde{P}_{\tilde{\theta}}(\mathbf{f}|\mathbf{e}).
\end{aligned} \tag{14}$$

Mnożąc obie strony przez $c(\mathbf{f}, \mathbf{e})$ i sumując po wszystkich francuskich i angielskich zdaniach otrzymamy nierówność (13).

3.3 Iteracyjne poprawianie modeli

Załóżmy, że $\tilde{P}_{\tilde{\theta}}$ jest danym modelem z ustalonym parametrem $\tilde{\theta}$, na podstawie którego chcemy uzykac lepszy model P_θ. Z nierówności Jensena dla modeli tłumaczenia wynika, że $\psi(P_\theta)$ jest większa od $\psi(\tilde{P}_{\tilde{\theta}})$, gdy $R(\tilde{P}_{\tilde{\theta}}, P_\theta)$ przyjmuje wartości dodatnie. Z kolei gdy $\tilde{P} = P$, to z faktu $R(\tilde{P}_{\tilde{\theta}}, \tilde{P}_{\tilde{\theta}}) = 0$ mamy, że dla ustalonego $\tilde{\theta}$ maksimum lokalne funkcji R w θ jest zawsze nieujemne. Stąd przyjęcie

punktu, w którym R osiąga maksimum, za nowy parametr θ gwarantuje nam, że $\psi(P_\theta)$ nie będzie mniejsza niż $\psi(\tilde{P}_{\tilde{\theta}})$.

Powyższe obserwacje skłoniły autorów Brown et al. do wykorzystania algorytmu EM do iteracyjnej optymalizacji modelu tłumaczenia P względem parametru θ według następującej procedury:

1. Ustal początkowe wartosci $\tilde{\theta}$.
2. Powtarzaj kroki 3. i 4. dopóki, dopóty nie nastąpi konwergencja.[6]
3. Przy ustalonych $\tilde{\theta}$ znajdź wartości θ, w których $R(\tilde{P}_{\tilde{\theta}}, P_\theta)$ ma maksimum.[7]
4. Zastąp $\tilde{\theta}$ przez θ.

3.4 Reestymacja parametrów

Żeby móc zastosować powyższy algorytm, trzeba rozwiązać problem identyfikacji maksimum w kroku 3. W przypadku omawianych modeli tłumaczenia odbywa się to jawnie. Dla ilustracji ogólnej metody, autorzy Brown et al. przyjmują następującą uproszczoną postać modelu

$$P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{\omega \in \Omega} \theta(\omega)^{c(\omega;\mathbf{a},\mathbf{f},\mathbf{e})}. \tag{15}$$

Tutaj $\theta(\omega)$, $\omega \in \Omega$, to rzeczywiste parametry spełniające warunki

$$\theta(\omega) \geq 0, \qquad \sum_{\omega \in \Omega_\mu} \theta(\omega) = 1, \quad \mu = 1, 2, \ldots \tag{16}$$

Zbiory Ω_μ, $\mu = 1, 2, \ldots$, tworzą partycję przestrzeni zdarzeń Ω. Grupuje się w ten sposób zdarzenia pokrewne, np. $\Omega_e = \{f|e : f \text{ jest wyrazem francuskim}\}$ jest zbiorem wszystkich zdarzeń takich, że francuski wyraz jest tłumaczeniem danego wyrazu e. $\theta(\omega)$ interpretuje się jako prawdopodobieństwo wystąpienia zdarzania ω, $c(\omega; \mathbf{a}, \mathbf{f}, \mathbf{e})$ jako liczbę wystąpień tego zdarzenia w $(\mathbf{a}, \mathbf{f}, \mathbf{e})$. Logarytmując i różniczkując równanie (15) wyznaczamy $c(\omega; \mathbf{a}, \mathbf{f}, \mathbf{e})$:

$$\begin{aligned}
P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) &= \prod_{\omega \in \Omega_\mu} \theta(\omega)^{c(\omega;\mathbf{a},\mathbf{f},\mathbf{e})} \\
\log P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) &= \sum_{\omega \in \Omega_\mu} c(\omega; \mathbf{a}, \mathbf{f}, \mathbf{e}) \log \theta(\omega) \\
\frac{\partial}{\partial \theta(\omega)} \log P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) &= \frac{\partial}{\partial \theta(\omega)} \sum_{\omega \in \Omega_\mu} c(\omega; \mathbf{a}, \mathbf{f}, \mathbf{e}) \log \theta(\omega) \\
\frac{\partial}{\partial \theta(\omega)} \log P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) &= c(\omega; \mathbf{a}, \mathbf{f}, \mathbf{e}) \frac{\partial}{\partial \theta(\omega)} \log \theta(\omega) \\
c(\omega; \mathbf{a}, \mathbf{f}, \mathbf{e}) &= \theta(\omega) \frac{\partial}{\partial \theta(\omega)} \log P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e})
\end{aligned} \tag{17}$$

W celu znalezienia wartości θ maksymalizującej $R(\tilde{P}_{\tilde{\theta}}, P_\theta)$ wprowadzamy funkcję Lagrange'a

$$L(\theta, \lambda) \equiv R(\tilde{P}_{\tilde{\theta}}, P_\theta) - \sum_\mu \lambda_\mu \Big(\sum_{\omega \in \Omega_\mu} \theta(\omega) - 1 \Big), \tag{18}$$

[6] Czyli $R(\tilde{P}_{\tilde{\theta}}, P_\theta) = 0$ lub $R(\tilde{P}_{\tilde{\theta}}, P_\theta) \leq \epsilon$, gdzie ϵ jest pewną ustaloną małą liczbą. W praktyce jednak ogranicza się po prostu liczbę kolejnych iteracji, zwykle do nie więcej niż 5 dla poszczególnego modelu.

[7] Przy takim sformuowniu kroki E i M algorytmu EM odbywają się w jednym kroku. Rozgraniczenie staje się bardziej jasne w sekcjach z przykładowymi obliczeniami.

gdzie λ_μ to mnożniki Lagrange'a, różne dla każdego podzbioru parametrów Ω_μ. Następnie dla każdego $\omega \in \Omega_\mu$, $\mu = 1, 2, \ldots$, należy rozwiązać układ równań postaci

$$\frac{\partial}{\partial\theta(\omega)} R(\tilde{P}_{\tilde{\theta}}, P_\theta) - \lambda_\mu = 0. \tag{19}$$

Równań jest tyle, ile zdarzeń w całej przestrzeni zdarzeń Ω. Mnożąc powyższe równanie obustronnie przez $\theta(\omega)$ rozwiązujemy je względem $\theta(\omega)$ za pomocą definicji (12) oraz równania (17) i otrzymujemy następujące wyniki

$$\begin{aligned}
\theta(\omega) &= \lambda_\mu^{-1}\theta(\omega)\frac{\partial}{\partial\theta(\omega)} R(\tilde{P}_{\tilde{\theta}}, P_\theta) \\
&= \lambda_\mu^{-1}\theta(\omega)\frac{\partial}{\partial\theta(\omega)} \sum_{\mathbf{f},\mathbf{e}} c(\mathbf{f},\mathbf{e}) \sum_{\mathbf{a}} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e}) \log \frac{P_\theta(\mathbf{f},\mathbf{a}|\mathbf{e})}{\tilde{P}_{\tilde{\theta}}(\mathbf{f},\mathbf{a}|\mathbf{e})} \\
&= \lambda_\mu^{-1} \sum_{\mathbf{f},\mathbf{e}} c(\mathbf{f},\mathbf{e}) \sum_{\mathbf{a}} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e})\theta(\omega)\frac{\partial}{\partial\theta(\omega)} \big(\log P_\theta(\mathbf{f},\mathbf{a}|\mathbf{e}) - \log \tilde{P}_{\tilde{\theta}}(\mathbf{f},\mathbf{a}|\mathbf{e}) \big) \\
&= \lambda_\mu^{-1} \sum_{\mathbf{f},\mathbf{e}} c(\mathbf{f},\mathbf{e}) \sum_{\mathbf{a}} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e})\theta(\omega)\frac{\partial}{\partial\theta(\omega)} \log P_\theta(\mathbf{f},\mathbf{a}|\mathbf{e}) \\
&= \lambda_\mu^{-1} \sum_{\mathbf{f},\mathbf{e}} c(\mathbf{f},\mathbf{e}) \sum_{\mathbf{a}} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e}) c(\omega;\mathbf{a},\mathbf{f},\mathbf{e}).
\end{aligned}$$

Sens tego wyniku stanie się jaśniejszy, gdy uzwięźlimy jego zapis wprowadzając następujące symbole pomocnicze:

$$E_{\tilde{\theta}}[C(\omega;\mathbf{f},\mathbf{e})] = \sum_{\mathbf{a}} c(\omega;\mathbf{a},\mathbf{f},\mathbf{e}) \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e}), \tag{20}$$

$$E_{\tilde{\theta}}[C(\omega)] = \sum_{\mathbf{f},\mathbf{e}} c(\mathbf{f},\mathbf{e}) E_{\tilde{\theta}}[C(\omega;\mathbf{f},\mathbf{e})], \tag{21}$$

wtedy

$$\theta(\omega) = \lambda_\mu^{-1} E_{\tilde{\theta}}[C(\omega)], \qquad \lambda_\mu = \sum_{\omega\in\Omega_\mu} E_{\tilde{\theta}}[C(\omega)]. \tag{22}$$

Równanie (20) interpretuje się jako oczekiwaną liczbę wystąpień zdarzenia ω w procesie tłumaczenia zdania $\mathbf{e}$ na $\mathbf{f}$ uzyskaną na podstawie modelu $\tilde{P}_{\tilde{\theta}}$. Stąd $\theta(\omega)$ jest normalizowaną oczekiwaną liczbą wystapień zdarzenia ω we wszystkich tłumaczeniach z zestawu treningowego. Autorzy Brown et al. zamiast z symboli $E_{\tilde{\theta}}[C(\cdot)]$ korzystają z oznaczeń $\tilde{c}_{\tilde{\theta}}(\cdot)$. Uważamy jednak, że nasze oznaczenia bardziej podkreślają istotę algorytmu EM. Nie jest to pusta symbolika, nie do końca oczywistą równość z klasyczną wartością oczekiwaną można zilustrować w następujący sposób:

We wzorze (21) wystarczy wykorzystać własności wartości oczekiwanej. W (20) odwołujemy się do definicji wartości oczekiwanej dla zmiennej losowej dyskretnej. Dla dowolnej pary zdań $\mathbf{f}, \mathbf{e}$ oraz zdarzenia ω zdefiniujemy rodzinę $\{\mathcal{A}_c^\omega(\mathbf{f},\mathbf{e})\}_{c\in\{0,1,2,\ldots\}}$ pomocniczych zbiórów dopasowań między tymi zdaniami, gdzie $\mathcal{A}_c^\omega(\mathbf{f},\mathbf{e}) = \{\mathbf{a} \in \mathcal{A}(\mathbf{f},\mathbf{e}) : c(\omega;\mathbf{a},\mathbf{f},\mathbf{e}) = c\}$ zawierają tylko takie dopasowania, dla których zdarzenie ω zachodzi dokładnie c razy. Zbiory dopasowań o różnych indeksach są wzajemnie rozłączne i tworzą pokrycie zbioru wszystkich dopasowań $\mathcal{A}(\mathbf{f},\mathbf{e})$. Ponadto, ponieważ

istnieje tylko skończona liczba dopasowań, od pewnego k mamy $\mathcal{A}_c^\omega(\mathbf{f},\mathbf{e}) = \emptyset$, gdy $c > k$. Niech $C(\omega;\mathbf{f},\mathbf{e})$ będzie zmienną losową odpowiadającą liczbie wystąpień zdarzenia ω w parze $\mathbf{f},\mathbf{e}$, wtedy

$$\begin{aligned} E_{\tilde{\theta}}[C(\omega;\mathbf{f},\mathbf{e})] &= \sum_{c=0}^{\infty} c\tilde{P}_{\tilde{\theta}}\big(C(\omega;\mathbf{f},\mathbf{e}) = c\big) \\ &= \sum_{c=0}^{\infty} c \sum_{\mathbf{a}\in\mathcal{A}_c^\omega(\mathbf{f},\mathbf{e})} \tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e}) \\ &= \sum_{c=0}^{\infty} \sum_{\mathbf{a}\in\mathcal{A}_c^\omega(\mathbf{f},\mathbf{e})} c(\omega;\mathbf{a},\mathbf{f},\mathbf{e})\tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e}) \\ &= \sum_{\mathbf{a}\in\mathcal{A}(\mathrm{f},\mathrm{e})} c(\omega;\mathbf{a},\mathbf{f},\mathbf{e})\tilde{P}_{\tilde{\theta}}(\mathbf{a}|\mathbf{f},\mathbf{e}). \end{aligned}$$

Podsumowując tę sekcję, widzimy, że wyznaczenie poprawionego modelu P_θ na podstawie modelu $\tilde{P}_{\tilde{\theta}}$ sprowadza się do wyznaczenia wartości $E_{\tilde{\theta}}[C(\omega;\mathbf{f},\mathbf{e})]$ i pośrednio $E_{\tilde{\theta}}[C(\omega)]$ dla każdego $\omega \in \Omega_\mu$, $\mu = 1, 2, \ldots$, co umożliwia nam wyznaczenie wszystkich nowych parametrów jako

$$\theta(\omega) = \frac{E_{\tilde{\theta}}[C(\omega)]}{\sum_{\omega\in\Omega_\mu} E_{\tilde{\theta}}[C(\omega)]}. \tag{23}$$

W zależności od złożoności modelu tłumaczenia wyznaczenie $E_{\tilde{\theta}}[C(\omega;\mathbf{f},\mathbf{e})]$ może być nieskomplikowane — jak w przypadku modeli 1 i 2 — lub obliczeniowo niewykonalne — jak w przypadku pozostałych modeli. Następnie omówimy konkretne modele, w szczególności sposoby wyznaczania $E_{\tilde{\theta}}[C(\omega;\mathbf{f},\mathbf{e})]$ dla różnych rodzajów zdarzeń i parametrów.

4 Model 1

Ogólną postać Modelu 1 (oraz Modelu 2, dla którego Model 1 jest przypadkiem szczególnym) autorzy Brown et al. przedstawiają w następujący sposób:

$$P_\theta(\mathbf{f},\mathbf{a}|\mathbf{e}) = P_\theta(m|\mathbf{e})P_\theta(\mathbf{a}|m,\mathbf{e})P_\theta(\mathbf{f}|\mathbf{a},m,\mathbf{e}) \tag{24}$$

Najpierw dla danego zdania angielskiego $\mathbf{e}$ oblicza się prawdopodobieństwo, że długość jego francuskiego tłumaczenia $\mathbf{f}$ wynosi m. Przy danym $\mathbf{e}$ oraz m oblicza się następnie prawdopodobieństwa danego dopasowania $\mathbf{a}$. Dopiero na podstawie dopasowania można obliczyć prawdopodobieństwo, że dane zdanie francuskie $\mathbf{f}$ jest tłumaczeniem $\mathbf{e}$.

4.1 Założenia

Wspominaliśmy już wcześniej, że zdania nie są atomami w omawianych modelach tłumaczenia. Wszystkie obliczenia sprowadzają się do wyrazów oraz do relacji między nimi, które są opisane za pomocą dopasowań między wyrazami. Dla zdań $\mathbf{f} \equiv \mathbf{f}_1^m$, $\mathbf{e} \equiv \mathbf{e}_1^l$ oraz dla dopasowania $\mathbf{a} \equiv \mathbf{a}_1^m = a_1, \ldots, a_m$, gdzie $a_i \in \{1, \ldots, l\}$, przyjmuje się:

$$P_\theta(m|\mathbf{e}) \equiv \epsilon(m|l) \tag{25}$$

$$P_\theta(\mathbf{a}|m,\mathbf{e}) \equiv \prod_{j=1}^{m} \frac{1}{l+1} = \frac{1}{(l+1)^m} \tag{26}$$

$$P_\theta(\mathbf{f}|\mathbf{a},m,\mathbf{e}) \equiv \prod_{j=1}^{m} t(f_j|e_{a_j}). \tag{27}$$

Stąd $P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e})$ dla Modelu 1 można zapisać jako

$$P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon(m|l)}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}), \tag{28}$$

a $P_\theta(\mathbf{f}|\mathbf{e})$ odpowiednio jako

$$P_\theta(\mathbf{f}|\mathbf{e}) = \frac{\epsilon(m|l)}{(l+1)^m} \sum_{\mathbf{a}} \prod_{j=1}^{m} t(f_j|e_{a_j}) = \frac{\epsilon(m|l)}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}). \tag{29}$$

Tutaj $P_\theta(\mathbf{a}|m, \mathbf{e})$ zależy bezpośrednio od długości m i l zdań $\mathbf{f}$ i $\mathbf{e}$ odpowiednio. Zakłada się, że dla każdego wyrazu francuskiego prawdopodobieństwo dopasowania z każdym z $l+1$ wyrazów angielskich jest równe i wynosi $(l+1)^{-1}$. Widzimy zatem, że w Modelu 1 prawdopodobieństwo, iż wyraz francuski jest tłumaczeniem wyrazu angielskiego, nie zależy od pozycji tych wyrazów w odpowiednich zdaniach, co wydaje się być założeniem nazbyt upraszczającym. Dopiero w Modelu 2 pojawiają się odpowiednie parametry uwzględniające wzajemne pozycje wyrazów.

Parametr $\epsilon(m|l)$ musi zostać wyznaczony w procesie estymacji parametrów modelu. Jednak ponieważ nie zależy on od dopasowań między zdaniami, nie trzeba w tym celu stosować algorytmu EM. Zakładaliśmy wcześniej, że korpus trenujący jest korpusem składającym się z par zdań będących wzajemnymi tłumaczeniami. Długości zdań są więc obserwowalne w sposób bezpośredni i można wykorzystać np. estymatory częstościowe w celu ustalenia $\epsilon(f|m)$.

4.2 Estymacja parametrów

Jedyny rodzaj parametrów podlegający estymacji za pomocą algorytmu EM to prawdopodobieństwa tłumaczenia $t(f|e)$ dla dowolnego f i e. Postać modelu opisanego w równaniu (28) odpowiada ogólnej postaci modeli z równania (15) i możemy wykorzystać wcześniej wyprowadzone równania do estymacji parametrów $t(f|e)$. Istnienie parametru $\epsilon(m|l)$ poza głównym iloczynem nie psuje tej zależności. W relatywnej funkcji celu (12) ulega on skróceniu, a przy wyznaczaniu $c(f|e; \mathbf{a}, \mathbf{f}, \mathbf{e})$ (17) pozbywamy się go przy różniczkowaniu, gdzie jest traktowany jak stała.

Zgodnie z podrozdziałem 3.4 na parametry $t(f|e)$ nakłada się następujący warunek dla każdego wyrazu angielskiego e:

$$t(f|e) \geq 0, \qquad \sum_f t(f|e) = 1. \tag{30}$$

Wtedy pomocnicza funkcja Lagrange'a przyjmuje postać:

$$L(t, \lambda) \equiv R(\tilde{P}_{\tilde{\theta}}, P_\theta) - \sum_e \lambda_e \Big(\sum_f t(f|e) - 1 \Big). \tag{31}$$

Rozwiązaliśmy już ogólną postać tego równania (18), otrzymując w końcu wynik (23). Wystarczy więc określić rodzaj parametru oraz typ zdarzenia, żeby otrzymać

$$t(f|e) = \frac{E_{\tilde{\theta}}[C(f|e)]}{\sum_f E_{\tilde{\theta}}[C(f|e)]} = \frac{\sum_{s=1}^{S} E_{\tilde{\theta}}[C(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)})]}{\sum_f \sum_{s=1}^{S} E_{\tilde{\theta}}[C(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)})]}, \tag{32}$$

gdzie

$$\begin{aligned} E_{\tilde{\theta}}[C(f|e; \mathbf{f}, \mathbf{e})] &= \sum_{\mathbf{a}} P_{\tilde{\theta}}(\mathbf{a}|\mathbf{e}, \mathbf{f}) c(f|e; \mathbf{a}, \mathbf{f}, \mathbf{e}) \\ &= \sum_{\mathbf{a}} P_{\tilde{\theta}}(\mathbf{a}|\mathbf{e}, \mathbf{f}) \sum_{j=1}^{m} \delta(f, f_j) \delta(e, e_{a_j}). \end{aligned} \tag{33}$$

Tutaj $\delta(x, y)$ jest funkcją delta Kroneckera, która przyjmuje wartość 1, gdy jej argumenty są równe, oraz 0 w przypadku przeciwnym. Wyrażenie $\delta(f, f_j)\delta(e, e_{a_j})$ jest równe 1 wtedy i tylko wtedy, gdy f jest równe f_j oraz e jest równe wyrazowi, z którym f_j jest połączony w danym dopasowaniu. Połączenie f z e w danej parze zdań przy danym dopasowaniu jest równoznaczne z fakten, że f jest tłumaczeniem e. Zatem suma $\sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})$ poprawnie zlicza liczbę razy, ile dany wyraz f jest tłumaczeniem e przy danym dopasowaniu w danej parze zdań. Przypominamy, że $P(\mathbf{a}|\mathbf{f}, \mathbf{e}) = P_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e})/P_{\tilde{\theta}}(\mathbf{f}|\mathbf{e})$.

Zakładając, że $\tilde{\theta}$ jest znane, mamy wszystkie środki potrzebne do wykonania kolejnej iteracji algorytmu EM. Przy inicjalizacji wystarczy założyć, że dla każdego f oraz e wszystkie prawdopodobieństwa $t(f|e)$ są równe, o ile zachowane są warunki (30).

4.3 Aspekty obliczeniowe

Aspektem, którego nie można ignorować, jest złożoność obliczeniowa wyrażana liczbą operacji, które trzeba wykonać, żeby uzyskać powyższe wyniki. Gdy rozpiszemy równanie (33) w następujący sposób

$$\begin{aligned} E_{\tilde{\theta}}[C(f|e; \mathbf{f}, \mathbf{e})] &= \sum_{\mathbf{a}} P_{\tilde{\theta}}(\mathbf{a}|\mathbf{e}, \mathbf{f}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \\ &= P_{\tilde{\theta}}(\mathbf{f}|\mathbf{e})^{-1} \sum_{\mathbf{a}} P_{\tilde{\theta}}(\mathbf{f}, \mathbf{a}|\mathbf{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \\ &= \frac{\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} \tilde{t}(f_j|e_{a_j}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})}{\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} \tilde{t}(f_j|e_{a_j})}, \end{aligned} \tag{34}$$

widzimy, że liczba operacji konieczna do uzyskania $E_{\tilde{\theta}}[C(f|e; \mathbf{f}, \mathbf{e})]$ jest proporcjonalna do $(l+1)^m$, gdzie l jest liczbą wyrazów w zdaniu $\mathbf{e}$ a m liczbą wyrazów w $\mathbf{f}$. W przypadku długich zdań obliczenia stają się szybko niewykonalne nawet dla jednego zdania, np. w przypadku pary zdań z 20 wyrazami w każdym zdaniu, liczba operacji ma 26 zer. Dodatkowo zestaw trenujący może składać się kilku milionów równoległych zdań.

Model 1 ma jednak własność, która umożliwia uzyskanie tych samych wyników za pomocą o wiele mniejszej liczby operacji. Wystarczy zauważyć, że mianownik równania (34) można zapisać jako

$$\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) = \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i),$$

co staje się bardzie jasne na przykładzie, gdzie dla $l = 3$ oraz $m = 1$ mamy $t_{10}t_{20}t_{30} + t_{10}t_{20}t_{31} + \ldots + t_{11}t_{21}t_{31} = (t_{10} + t_{11})(t_{20} + t_{21})(t_{30} + t_{31})$. Podobnie licznik można przekształcić do następującej postaci

$$\begin{aligned} &\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \\ &\qquad = \frac{t(f|e)}{\sum_{i=0}^{l} t(f|e_i)} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i) \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=0}^{l} \delta(e, e_i), \end{aligned} \tag{35}$$

co nie jest już takie oczywiste jak w przypadku mianownika.

Wystarczy jednak zauważyć, że $t(f_1|e_{a_1}) \ldots t(f_m|e_{a_m}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})$ jest różne od zera tylko wtedy, gdy przynajmniej jeden z czynników jest równy $t(f|e)$. Stąd $t(f|e)$ możemy wyłączyć przed sumę. Dla przykładu przyjmijmy $t(f|e) = t(f_1|e_0)$, wtedy mamy $t_{10}t_{20}t_{30} + t_{10}t_{20}t_{31} + \ldots +$

fr	en
maison bleue	blue house
chien rouge	red dog
chien vert	green dog

$V_e = \{\text{blue, dog, green, house, red}\}$
$V_f = \{\text{bleue, chien, maison, rouge, vert}\}$

Tablica 1: Bardzo uproszczony korpus równoległy

$t_{10}t_{21}t_{31} = t_{10}(t_{20}+t_{21})(t_{30}+t_{31}) = t_{10}(t_{10}+t_{11})^{-1}(t_{10}+t_{11})(t_{20}+t_{21})(t_{30}+t_{31})$. Podstawiając powyższe przekształcenia do rownania (34) po skróceniu otrzymujemy

$$E_{\tilde{\theta}}[C(f|e;\mathbf{f},\mathbf{e})] = \frac{\tilde{t}(f|e)}{\sum_{i=0}^{l}\tilde{t}(f|e_i)}\sum_{j=1}^{m}\delta(f,f_j)\sum_{i=0}^{l}\delta(e,e_i). \tag{36}$$

Przez $\tilde{t}$ oznaczamy prawdopodobieństwa tłumaczenia z $\tilde{\theta}$. W równaniu (36) liczba operacji jest proporcjonalna do $m+l$, czyli dla pary zdań z 20 wyrazami w każdym zdaniu trzeba wykonać tylko kilkadziesiąt operacji, otrzymując ten sam wynik co w (34).

4.4 Przykładowe obliczenia

W tym rozdziale przedstawimy estymację parametrów modelu 1 dla bardzo uproszczonego korpusu równoległego przedstawionego w tabelce 1. Korpus składa się z trzech par fraz równoległych w języku francuskim i angielskim. Nie pokażemy tutaj działania algorytmu dla całych zdań, ponieważ duża liczba parametrów nie dałaby się wtedy przedstawić w sposób przyjazny dla czytelnika. Różne wyrazy występujące w każdej połowie korpusu tworzą słownictwo danego języka, oznaczane dla francuskiego i angielsko odpowiednio przez V_f oraz V_e. Liczba parametrów w przypadku modelu 1, gdzie jedynym rodzajem parametrów jest prawdopodobieństwo tłumaczenia dla wyrazów, wynosi $|V_f| \cdot (|V_e|+1)$. Jedynka dodana do V_e jest skutkiem założenia angielskiego wyrazu zerowego e_0, o którym wspomniano już w rozdziale dotyczącym dopasowań (2.4).

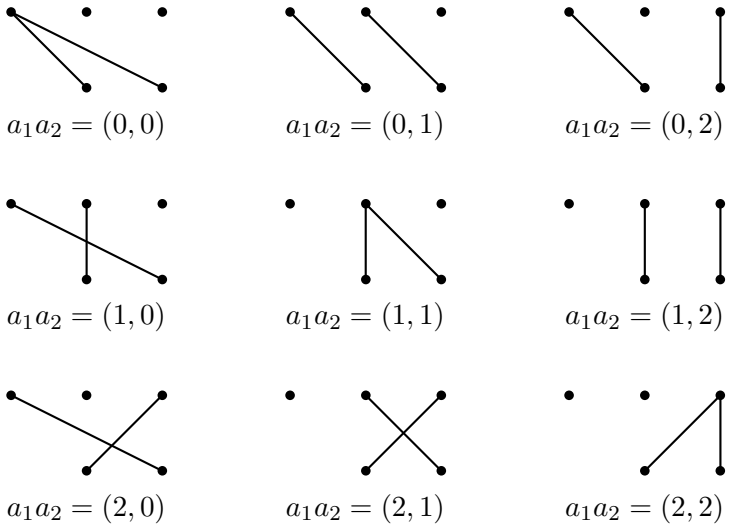
Pierwszym krokiem jest inicjalizacja parametrów modelu. Wystarczy w tym celu przypisać każdej parze wyrazów prawdopodobieństwo tłumaczenia $1/|V_f|$. Wynikiem takiej operacji jest tabela 2. Widać, że warunek $\sum_f t(f|e) = 1$ jest spełniony dla każdego angielskiego wyrazu e.

Możemy zatem przejść do estymacji oczekiwanych liczebności $E[C(f|e)]$, co uczynimy nieco dokładniej dla pary $chien|dog$. Najpierw należy wyznaczyć liczebności dla konkretnych zdań w korpusie $E[C(chien|dog);\mathbf{e}^{(i)},\mathbf{f}^{(i)}]$, gdzie $\mathbf{e}^{(i)},\mathbf{f}^{(i)}$ to i-ta para zdań z korpusu.

Na rysunku 5 mamy wszystkie możliwe — przy wcześniej omówionych ograniczeniach — dopasowania dla par zdań, gdzie $m = 2$ i $l = 2$, czyli dla wszystkich par w przykładowym korpusie. Górny

t(bleue|e_0) = 0.20	t(bleue|blue) = 0.20	t(bleue|dog) = 0.20
t(chien|e_0) = 0.20	t(chien|blue) = 0.20	t(chien|dog) = 0.20
t(maison|e_0) = 0.20	t(maison|blue) = 0.20	t(maison|dog) = 0.20
t(rouge|e_0) = 0.20	t(rouge|blue) = 0.20	t(rouge|dog) = 0.20
t(vert|e_0) = 0.20	t(vert|blue) = 0.20	t(vert|dog) = 0.20
t(bleue|green) = 0.20	t(bleue|house) = 0.20	t(bleue|red) = 0.20
t(chien|green) = 0.20	t(chien|house) = 0.20	t(chien|red) = 0.20
t(maison|green) = 0.20	t(maison|house) = 0.20	t(maison|red) = 0.20
t(rouge|green) = 0.20	t(rouge|house) = 0.20	t(rouge|red) = 0.20
t(vert|green) = 0.20	t(vert|house) = 0.20	t(vert|red) = 0.20

Tablica 2: Prawdopodobieństwa tłumaczeń po inicjalizacji



Rysunek 5: Możliwe dopasowanie dla par zdań, gdzie $m = 2$ i $l = 2$

rząd węzłów reprezentuje angielskie zdanie z dodatkowym wyrazem zerowym e_0 na pierwszym miejscu, dolny rząd odpowiada francuskiemu zdaniu. W przypadku pierwszego zdania liczebność oczekiwana $E_1[C(chien|dog); \mathbf{e}^{(1)}, \mathbf{f}^{(1)}]$ wynosi oczywiście 0, ponieważ w tej parze zdań nie występuje para wyrazów $chien|dog$. Indeksy przy symbolach odzwierciedlają numer iteracji. W przypadku, gdy wartości pochodzą z inicjalizacji, stosujemy indeks 0. Wykonując obliczenia według wzoru (34) musimy uwzględnić wszystkie dopasowania. Niech $f = chien$ oraz $e = dog$, mamy zatem dla zdania drugiego

$$\begin{aligned}
P_0(\mathbf{f}^{(2)}|\mathbf{e}^{(2)}) &= \sum_{a_1=0}^{2}\sum_{a_2=0}^{2}\prod_{j=1}^{2} t_0(f_j|e_{a_j}) = \\
&= t_0(f_1|e_0)t_0(f_2|e_0) + t_0(f_1|e_0)t_0(f_2|e_1) + t_0(f_1|e_0)t_0(f_2|e_2) \\
&+ t_0(f_1|e_1)t_0(f_2|e_0) + t_0(f_1|e_1)t_0(f_2|e_1) + t_0(f_1|e_1)t_0(f_2|e_2) \\
&+ t_0(f_1|e_2)t_0(f_2|e_0) + t_0(f_1|e_2)t_0(f_2|e_1) + t_0(f_1|e_2)t_0(f_2|e_2) \\
&= 9 \cdot 0.2 \cdot 0.2 \\
&= 0.36
\end{aligned}$$

$$\begin{aligned}
E_1[C(f|e); \mathbf{e}^{(2)}, \mathbf{f}^{(2)}] &= P_0(\mathbf{f}^{(2)}|\mathbf{e}^{(2)})^{-1} P_0(\mathbf{f}^{(2)}, \mathbf{a}|\mathbf{e}^{(2)}) \sum_{k=1}^{m} \delta(f, f_k)\delta(e, e_{a_k}) \\
&= (0.36)^{-1} \sum_{a_1=0}^{2}\sum_{a_2=0}^{2}\prod_{j=1}^{2} t_0(f_j|e_{a_j}) \sum_{k=1}^{m} \delta(f, f_k)\delta(e, e_{a_k}) \\
&= (0.36)^{-1}\big(t_0(f_1|e_0)t_0(f_2|e_0) \cdot 0 + t_0(f_1|e_0)t_0(f_2|e_1) \cdot 0 \\
&+ t_0(f_1|e_0)t_0(f_2|e_2) \cdot 0 + t_0(f_1|e_1)t_0(f_2|e_0) \cdot 0 \\
&+ t_0(f_1|e_1)t_0(f_2|e_1) \cdot 0 + t_0(f_1|e_1)t_0(f_2|e_2) \cdot 0 \\
&+ t_0(f_1|e_2)t_0(f_2|e_0) \cdot 1 + t_0(f_1|e_2)t_0(f_2|e_1) \cdot 1 \\
&+ t_0(f_1|e_2)t_0(f_2|e_2) \cdot 1\big) \\
&= (0.36)^{-1} \cdot 3 \cdot 0.2 \cdot 0.2 \\
&= (0.36)^{-1} \cdot 0.12 \\
&\approx 0.33.
\end{aligned}$$

Te same wartości uzyskamy korzystając ze skróconego sposobu obliczania oczekiwanej liczebności $E_1[C(chien|dog); \mathbf{e}^{(2)}, \mathbf{f}^{(2)}]$ na podstawie wzoru (36). Otrzymujemy wtedy w o wiele bardziej zwięzły sposób

$$\begin{aligned} E_1[C(f|e); \mathbf{e}^{(2)}, \mathbf{f}^{(2)}] &= \frac{t_0(f|e)}{\sum_{i=0}^{2} t_0(f|e_i)} \sum_{j=1}^{2} \delta(f, f_j) \sum_{i=0}^{2} \delta(e, e_i) \\ &= \frac{t_0(f|e)}{t_0(f|e_0) + t_0(f|e_1) + t_0(f|e_2)} \cdot 1 \cdot 1 \\ &= \frac{0.2}{0.6} \approx 0.33. \end{aligned}$$

Dla trzeciego zdania $E_1[C(chien|dog); \mathbf{e}^{(3)}, \mathbf{f}^{(3)}]$ również wynosi w przybliżeniu 0.33. Mamy zatem całkowitą liczebność pary $chien|dog$ z

$$\begin{aligned} E_1[C(chien|dog)] &= \sum_{i=1}^{3} E_1[C(chien|dog); \mathbf{e}^{(i)}, \mathbf{f}^{(i)}] \\ &\approx 0 + 0.33 + 0.33 \\ &\approx 0.67 \end{aligned}$$

Podobnie obliczamy wartości oczekiwane dla pozostałych par wyrazów. Tym samym kończymy krok E algorytmu EM. Prawdopodobieństwo tłumaczenia $t(chien|dog)$ wyznaczamy ze wzoru (32) w następujący sposób

$$\begin{aligned} t_1(chien|dog) &= \frac{E_1[C(chien|dog)]}{\sum_f E_1[C(f|dog)]} \\ &\approx \frac{0,67}{0 + 0.67 + 0 + 0.33 + 0.33} \\ &\approx 0.5. \end{aligned}$$

Tabela 3 przedstawia wszystkie wartości oczekiwanych liczebności oraz parametrów modelu 1 po pierwszej iteracji. Widzimy, że prawdopodobieństwa tłumaczenia uległy sporej zmianie w porównaniu z wartościami z tabeli 2. Wyrazy, które nie współwystępują w ramach par zdań równoległych otrzymały zerowe prawdopodobieństwo, inne parametry zmniejszyły lub zwiększyły się.

Ustalanie nowych parametrów zawiera w sobie krok M, ponieważ znajdujemy w ten sposób parametry maksymalizujące naszą funkcję celu. Powtarzając powyższe obliczenia wielokrotnie, otrzymujemy po piątej iteracji wartości przedstawione w tabeli 4. Parametr $t_5(chien|dog)$ przyjął już wartość 0.77. Z kolejnymi iteracjami zbliża się coraz bardziej do 1, np. $t_7(chien|dog) = 0.85$, $t_{10}(chien|dog) = 0.91$, $t_{15}(chien|dog) = 0.95$ itd.

Porównując wartości po pierwszej i po piątej iteracji, widać jednak, że prawdopodobieństwa tłumaczenia dla par złożonych z francuskich wyrazów $bleue$, $maison$ oraz angielskich wyrazów $blue$, $house$ (pogrubione) ustabilizowały się wszystkie na poziomie 0.5. Są to wyrazy, z których składa się pierwsza para zdań z korpusu. Model 1 nie jest w stanie wychwycić zależności miedzy $bleue$ i $blue$ oraz $maison$ i $house$, ponieważ wyrazy te nie pojawiają się w innych zdaniach. Model 1 nie uwzględnia też informacji o pozycji wyrazów w zdaniu, więc wszystkie kombinacje są równie prawdopodobne. W przypadku $chien$ i dog informacje są bardziej jednoznaczne. W łatwy sposób można stwierdzić, że $chien$ pojawia się po stronie francuskiej tylko wtedy, gdy pojawia się dog w części angielskiej. Co więcej, dzieję się tak w dwóch zdaniach, w których pozostałe wyrazy się nie powtarzają. Człowiek nie znając żadnego z danych języków szybko znajdzie poprawne odpowiedniki na podstawie sposobu współwystępowania wyrazów. W przypadku pierwszej pary zdań musiałby odwołać się do innych informacji. Mógłby np. stwierdzić, że w drugim i trzecim zdaniu tłumaczenia układają się na krzyż[8]. Model 1 nie potrafi czynić tego rodzaju obserwacji, model 2 już tak.

[8]Na rysunku 5 odpowiada to pozycji $a_1 a_2 = (2, 1)$.

E[C(bleue|e_0)] = 0.33	E[C(bleue|blue)] = 0.33	E[C(bleue|dog)] = 0.00
E[C(chien|e_0)] = 0.67	E[C(chien|blue)] = 0.00	E[C(chien|dog)] = 0.67
E[C(maison|e_0)] = 0.33	E[C(maison|blue)] = 0.33	E[C(maison|dog)] = 0.00
E[C(rouge|e_0)] = 0.33	E[C(rouge|blue)] = 0.00	E[C(rouge|dog)] = 0.33
E[C(vert|e_0)] = 0.33	E[C(vert|blue)] = 0.00	E[C(vert|dog)] = 0.33
E[C(bleue|green)] = 0.00	E[C(bleue|house)] = 0.33	E[C(bleue|red)] = 0.00
E[C(chien|green)] = 0.33	E[C(chien|house)] = 0.00	E[C(chien|red)] = 0.33
E[C(maison|green)] = 0.00	E[C(maison|house)] = 0.33	E[C(maison|red)] = 0.00
E[C(rouge|green)] = 0.00	E[C(rouge|house)] = 0.00	E[C(rouge|red)] = 0.33
E[C(vert|green)] = 0.33	E[C(vert|house)] = 0.00	E[C(vert|red)] = 0.00

t(bleue|e_0) = 0.17	t(bleue|blue) = 0.50	t(bleue|dog) = 0.00
t(chien|e_0) = 0.33	t(chien|blue) = 0.00	t(chien|dog) = 0.50
t(maison|e_0) = 0.17	t(maison|blue) = 0.50	t(maison|dog) = 0.00
t(rouge|e_0) = 0.17	t(rouge|blue) = 0.00	t(rouge|dog) = 0.25
t(vert|e_0) = 0.17	t(vert|blue) = 0.00	t(vert|dog) = 0.25
t(bleue|green) = 0.00	t(bleue|house) = 0.50	t(bleue|red) = 0.00
t(chien|green) = 0.50	t(chien|house) = 0.00	t(chien|red) = 0.50
t(maison|green) = 0.00	t(maison|house) = 0.50	t(maison|red) = 0.00
t(rouge|green) = 0.00	t(rouge|house) = 0.00	t(rouge|red) = 0.50
t(vert|green) = 0.50	t(vert|house) = 0.00	t(vert|red) = 0.00

Tablica 3: Oczekiwane liczebności oraz prawodopobieństwa tłumaczeń po 1. iteracji

5 Model 2

Tak samo jak dla modelu 1, prawdopodobieństwo zdania francuskiego oraz wybranego dopasowania przy danym zdaniu angielskim jest dla modelu 2 opisane jako iloczyn prawdopodobieństw w następujący sposób:

$$P_\theta(\mathbf{f}, \mathbf{a}|\mathbf{e}) = P_\theta(m|\mathbf{e})P_\theta(\mathbf{a}|m, \mathbf{e})P_\theta(\mathbf{f}|\mathbf{a}, m, \mathbf{e}). \tag{37}$$

Prawodpodobieństwa $P_\theta(m|\mathbf{e})$ oraz $P_\theta(\mathbf{f}|\mathbf{a}, m, \mathbf{e})$ nie ulegają zmianie, pojawia się natomiast nowy sposób opisu $P_\theta(\mathbf{a}|m, \mathbf{e})$. Mamy

$$P_\theta(m|\mathbf{e}) \equiv \epsilon(m|l) \tag{38}$$

$$P_\theta(\mathbf{a}|m, \mathbf{e}) \equiv \prod_{j=1}^{m} a(a_j|j, m, l) \tag{39}$$

$$P_\theta(\mathbf{f}|\mathbf{a}, m, \mathbf{e}) \equiv \prod_{j=1}^{m} t(f_j|e_{a_j}), \tag{40}$$

gdzie $a(i|j, m, l)$ jest dodatkowym parametrem interpretowanym jako prawdopodobieństwo, że dla zdania francuskiego o długości m i dla zdania angielskiego o długości l wyraz angielski na i-tej pozycji jest dopasowany z wyrazem francuskim na j-tej pozycji. Parametry tego typu będziemy nazywać *prawdopodobieństwem dopasowania*.

Prawdopodobieństwa dopasowania modelują inne zjawisko niż wprowadzone przy modelu 1 prawdopodobieństwa tłumaczeń. Prawdopodobieństwa tłumaczenia nie zależą od pozycji danych wyrazów tylko od jego ortograficznej postaci. Stąd w przypadku modelu 1 zawsze mamy $P(\mathbf{f_1}|\mathbf{e}) = P(\mathbf{f_2}|\mathbf{e})$, gdy zdania francuskie $\mathbf{f_1}$ oraz $\mathbf{f_2}$ różnią się pozycjami wyrazów. Natomiast prawdopodobieństwa dopasowań nie zależą w ogóle od postaci wyrazów, istotne są tylko ich pozycje w parach zdań o określonych długościach.

E[C(bleue|e_0)] = 0.07	E[C(bleue|blue)] = 0.47	E[C(bleue|dog)] = 0.00
E[C(chien|e_0)] = 0.78	E[C(chien|blue)] = 0.00	E[C(chien|dog)] = 0.92
E[C(maison|e_0)] = 0.07	E[C(maison|blue)] = 0.47	E[C(maison|dog)] = 0.00
E[C(rouge|e_0)] = 0.12	E[C(rouge|blue)] = 0.00	E[C(rouge|dog)] = 0.14
E[C(vert|e_0)] = 0.12	E[C(vert|blue)] = 0.00	E[C(vert|dog)] = 0.14
E[C(bleue|green)] = 0.00	E[C(bleue|house)] = 0.47	E[C(bleue|red)] = 0.00
E[C(chien|green)] = 0.15	E[C(chien|house)] = 0.00	E[C(chien|red)] = 0.15
E[C(maison|green)] = 0.00	E[C(maison|house)] = 0.47	E[C(maison|red)] = 0.00
E[C(rouge|green)] = 0.00	E[C(rouge|house)] = 0.00	E[C(rouge|red)] = 0.74
E[C(vert|green)] = 0.74	E[C(vert|house)] = 0.00	E[C(vert|red)] = 0.00

t(bleue|e_0) = 0.06	**t(bleue|blue) = 0.50**	t(bleue|dog) = 0.00
t(chien|e_0) = 0.67	t(chien|blue) = 0.00	t(chien|dog) = 0.77
t(maison|e_0) = 0.06	**t(maison|blue) = 0.50**	t(maison|dog) = 0.00
t(rouge|e_0) = 0.10	t(rouge|blue) = 0.00	t(rouge|dog) = 0.12
t(vert|e_0) = 0.10	t(vert|blue) = 0.00	t(vert|dog) = 0.12
t(bleue|green) = 0.00	**t(bleue|house) = 0.50**	t(bleue|red) = 0.00
t(chien|green) = 0.17	t(chien|house) = 0.00	t(chien|red) = 0.17
t(maison|green) = 0.00	**t(maison|house) = 0.50**	t(maison|red) = 0.00
t(rouge|green) = 0.00	t(rouge|house) = 0.00	t(rouge|red) = 0.83
t(vert|green) = 0.83	t(vert|house) = 0.00	t(vert|red) = 0.00

Tablica 4: Oczekiwane liczebności oraz prawodopobieństwa tłumaczeń po 5. iteracji

5.1 Estymacja parametrów

Dodatkowy parametr zmienia postać podstawowych rozkładów modelu do następujących postaci

$$P(\mathbf{f},\mathbf{a}|\mathbf{e}) = \epsilon(m|l)\prod_{j=1}^{m}\big(t(f_j|e_{a_j})a(a_j|j,m,l)\big) \tag{41}$$

$$P(\mathbf{f}|\mathbf{e}) = \epsilon(m|l)\sum_{a_1=0}^{l}\ldots\sum_{a_m=0}^{l}\prod_{j=1}^{m}\big(t(f_j|e_{a_j})a(a_j|j,m,l)\big) \tag{42}$$

Przy dodatkowych warunkach postaci

$$\sum_{i=0}^{l} a(i|j,m,l) = 1 \tag{43}$$

dla dowolnych m, l oraz $1 \leq j \leq m$, otrzymujemy nową funkcję wiarogodności

$$L(t,a,\lambda,\mu) \equiv R(\tilde{P}_{\tilde{\theta}},P_\theta) - \sum_e \lambda_e\big(\sum_f t(f|e)-1\big) - \sum_j \mu_{jml}\big(\sum_i a(i|j,m,l)-1\big), \tag{44}$$

gdzie λ_e i μ_{jml} to czynniki Lagrange'a. Wiemy już jak wygląda ogólna postać rozwiązania tego równania i możemy określić

$$t(f|e) = \frac{E_{\tilde{\theta}}[C(f|e)]}{\sum_f E_{\tilde{\theta}}[C(f|e)]} = \frac{\sum_{s=1}^{S} E_{\tilde{\theta}}[C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})]}{\sum_f \sum_{s=1}^{S} E_{\tilde{\theta}}[C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})]} \tag{45}$$

$$a(i|j,m,l) = \frac{E_{\tilde{\theta}}[C(i|j,m,l)]}{\sum_f E_{\tilde{\theta}}[C(i|j,m,l)]} = \frac{\sum_{s=1}^{S} E_{\tilde{\theta}}[C(i|j,m,l;\mathbf{f}^{(s)},\mathbf{e}^{(s)})]}{\sum_f \sum_{s=1}^{S} E_{\tilde{\theta}}[C(i|j,m,l;\mathbf{f}^{(s)},\mathbf{e}^{(s)})]}. \tag{46}$$

Wartość $E_{\tilde{\theta}}[C(f|e;\mathbf{f},\mathbf{e})]$ tak jak w przypadku modelu 1 opisuje oczekiwaną liczbę razy, że w parze zdań $\mathbf{f}$ i $\mathbf{e}$ wyraz f jest tłumaczeniem wyrazu e. Sposób obliczenia $E_{\tilde{\theta}}[C(f|e;\mathbf{f},\mathbf{e})]$ pozostaje taki sam jak dla modelu 1, mamy więc

$$E_{\tilde{\theta}}[C(f|e;\mathbf{f},\mathbf{e})] = \sum_{\mathbf{a}} P_{\tilde{\theta}}(\mathbf{a}|\mathbf{e},\mathbf{f}) \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j}). \tag{47}$$

Odpowiednio symbol $E_{\tilde{\theta}}[C(i|j,m,l;\mathbf{f},\mathbf{e})]$ oznacza oczekiwaną liczbę razy, że w parze zdań o długościach m i l angielski wyraz na i-tej pozycji jest połączony z francuskim wyrazem na j-tej pozycji. Mamy

$$E_{\tilde{\theta}}[C(i|j,m,l;\mathbf{f},\mathbf{e})] = \sum_{\mathbf{a}} P_{\tilde{\theta}}(\mathbf{a}|\mathbf{e},\mathbf{f})\delta(i,a_j). \tag{48}$$

5.2 Aspekty obliczeniowe

Widać po równaniach (47) oraz (48), że również w przypadku modelu 2 trzeba sumować po wszystkich dopasowaniach, o ile nie uprości się powyższych wzorów. Podobnie jak dla modelu 1 możemy uprościć równanie (42) do postaci

$$P(\mathbf{f}|\mathbf{e}) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_{a_j})a(i|j,m,l). \tag{49}$$

Wykorzystując powyższą postać $P(\mathbf{f}|\mathbf{e})$, można pokazać, że

$$E_{\tilde{\theta}}[C(f|e;\mathbf{f},\mathbf{e})] = \sum_{j=1}^{m}\sum_{i=0}^{l} \frac{t(f|e)a(i|j,m,l)\delta(f,f_j)\delta(e,e_i)}{t(f_j|e_0)a(0|j,m,l) + \ldots + t(f_j|e_l)a(l|j,m,l)} \tag{50}$$

oraz

$$E_{\tilde{\theta}}[C(i|j,m,l;\mathbf{f},\mathbf{e})] = \frac{t(f_j|e_i)a(i|j,m,l)}{t(f_j|e_0)a(0|j,m,l) + \ldots + t(f_j|e_l)a(l|j,m,l)}. \tag{51}$$

Liczba operacji konieczna do obliczenia równania (50) jest proporcjonalna do $m \cdot l$, a nie $m+l$ jak to miało miejsce w modelu 1. Niemniej jest to o wiele bardziej korzystne niż wykładnicza złożoność przy obliczaniu pierwotnych równań. W przypadku równania (51) liczba operacji jest proporcjonalna do l.

5.3 Przykładowe obliczenia

Wykorzystamy ten sam przykład co w sekcji 4.4. Dla prawdopodobieństw tłumaczenia po inicjalizacji mamy takie same wartości jak dla modelu 1. Dodatkowo przypisujemy parametrom typu a odpowiednio równe wartości początkowe, zgodnie z warunkami nałożonymi na tego rodzaju parametry.

Pamiętamy, że model 1 nie był w stanie wyznaczyć jednoznacznie powiązania między wyrazami $maison$ i $house$ oraz między $bleue$ i $blue$. Informacje zawarte w przykładowym korpusie równoległym nie były wystarczające przy założeniu, że pozycja wyrazów nie ma znaczenia. Jak widać po równaniach 50 i 51, to w modelu 2 prawdopodobieństwa tłumaczenia oraz dopasowania wpływają na siebie wzajemnie. Model 1 był w stanie wyznaczyć, że prawdopodobieństwo tłumaczenia dla wyrazów $chien$ i dog jest wysokie, podobnie dla $vert$ i $green$. Nie miało to jednak żadnego wpływu na parę $maison$ i $house$. W modelu 2 wysokie prawdopodobieństwo tłumaczenia $t(chien|dog)$ powoduje zwiększenie się prawdopodobieństwa dopasowania $a(2|1,2,2)$, ponieważ dog znajduję sie w zdaniu angielskim na drugiej pozycji, a $chien$ na pierwszej. Z drugiej strony ten sam parametr $a(2|1,2,2)$ jest wykorzystywany przy obliczaniu $t(maison|house)$, ponieważ $house$ znajduję się na drugim miejscu w zdaniu angielskim, a $maison$ na pierwszym w zdaniu francuskim. W naszym korpusie składa się tak, że wszystkie zdania mają taką samą długość. Stąd zwiększanie się wartości

t(bleue|e_0) = 0.07	**t(bleue|blue) = 0.66**	t(bleue|dog) = 0.00
t(chien|e_0) = 0.52	t(chien|blue) = 0.00	t(chien|dog) = 0.75
t(maison|e_0) = 0.26	**t(maison|blue) = 0.34**	t(maison|dog) = 0.00
t(rouge|e_0) = 0.07	t(rouge|blue) = 0.00	t(rouge|dog) = 0.13
t(vert|e_0) = 0.07	t(vert|blue) = 0.00	t(vert|dog) = 0.13
t(bleue|green) = 0.00	**t(bleue|house) = 0.25**	t(bleue|red) = 0.00
t(chien|green) = 0.34	t(chien|house) = 0.00	t(chien|red) = 0.34
t(maison|green) = 0.00	**t(maison|house) = 0.75**	t(maison|red) = 0.00
t(rouge|green) = 0.00	t(rouge|house) = 0.00	t(rouge|red) = 0.66
t(vert|green) = 0.66	t(vert|house) = 0.00	t(vert|red) = 0.00

a(0|1,2,2) = 0.06	a(0|2,2,2) = 0.01
a(1|1,2,2) = 0.32	**a(1|2,2,2) = 0.91**
a(2|1,2,2) = 0.62	a(2|2,2,2) = 0.08

Tablica 5: Prawdopodobieństwa tłumaczenia oraz dopasowania po 5. iteracji

$a(2|1,2,2)$ powoduję zwiększanie się wartości $t(maison|house)$. Interakcja wszystkich parametrów prowadzi do wytworzenia się bardziej wyraźnego obrazu dla danego korpusu równoległego. Tabela 5 ilustruje wpływ dodatkowych parametrów na wyznaczany model tłumaczenia po piątej iteracji. Porównując pogrubione wartości tej tabeli z odpowiednimi wartościami z tabeli 4 widzimy, że model 2 poradził sobie lepiej z interpretacją danych. Decydujący jest wpływ pogrubionych parametrów $a(2|1,2,2)$ oraz $a(1|2,2,2)$, które odzwierciedlają krzyżujące się dopasowania przedstawione w rysunku 5 jako typ $a_1a_2 = (2,1)$.

5.4 Najlepsze dopasowanie

Wspominaliśmy w sekcji 2.4, że opisywane modele tłumaczenia określa się często jako modele dopasowań wyrazów. Do tej pory sumowaliśmy po wszystkich możliwych dopasowaniach będących funkcjami zawartych w zbiorze dopasowań $\mathcal{A}(\mathbf{f},\mathbf{e})$ miedzy zdaniami $\mathbf{f}$ i $\mathbf{e}$. Większość tych dopasowań nie ma sensu lingwistycznego. Wystarczy spojrzeć na rysunek 5, gdzie dla danych zdań w przykładowym korpusie tylko jedno dopasowanie z dziewięciu jest poprawne, czyli najlepsze. Najlepsze dopasowanie między zdaniami $\mathbf{f}$ i $\mathbf{e}$ oznaczamy przez $V(\mathbf{f},\mathbf{e})$[9].

W przypadku modelu 1 oraz modelu 2 istnieje prosty algorytm wyznaczania najlepszego dopasowania $V(\mathbf{f},\mathbf{e})$ dla danej pary zdań. Pamiętamy, że każde dopasowanie $\mathbf{a}$ jest ciągiem $a_1a_2\ldots a_m$, gdzie m jest długością zdania francuskiego. Elementy ciągu $a_j = i$ to pozycje i-tego wyrazu angielskiego w zdaniu o długości l dopasowanego z francuskim wyrazem na j-tej pozycji. Stąd przez $V(\mathbf{f},\mathbf{e})_j$ oznaczamy j-ty wyraz ciągu $V(\mathbf{f},\mathbf{e})$. Wtedy

$$V(\mathbf{f},\mathbf{e})_j = \arg\max_i t(f_j|e_i)a(i|j,m,l). \tag{52}$$

Dla pary zdań *maison bleue* i *blue house* mamy stąd $V(\mathbf{f},\mathbf{e}) = (2,1)$. Czyli możemy odczytać, że *maison* jest odpowiednikiem *house* oraz *bleue* odpowiednikiem *blue*. Badania nad modelami dopasowań i jakością dopasowań stanowią dział językoznawstwa komputerowego, który zdążył już uniezależnić się od metod tłumaczenia automatycznego.

6 Podsumowanie

W pracy opisaliśmy wyniki opublikowane w artykule *The Mathematics of Statistical Machine Translation* autorów Brown, Della Pietra, Della Pietra i Mercer z roku 1993, przy czym skupiliśmy się

[9] Oznaczenie pochodzi od określenia *Viterbi alignment*.

na metodach estymacji parametrów dwóch pierwszych statystycznych modeli tłumaczenia, tzw. *IBM Model 1* oraz *IBM Model 2*. Przedstawiono koncepcje tłumaczenia statystycznego opartego na wyrazach oraz pojęcie dopasowania na poziomie wyrazów. Zastosowanie algorytmu EM do estymacji parametrów modeli tłumaczenia statystycznego zostało zilustrowane zarówno ze strony teoretycznej jak i za pomocą konkretnych przykładów uzyskanych na podstawie implementacji na komputerze. Przedstawiliśmy różnice w działaniu modeli 1 oraz 2 dla tych samych korpusów równoległych.

Wspominaliśmy na początku pracy, że podlegające modelom tłumaczenia modele dopasowań nadal znajdują szerokie zastosowanie wśród najróżniejszych dziedzin lingwistyki komputerowej oraz w językoznawstwie ogólnym. Większość nowoczesnych systemów tłumaczenia statystycznego ma swoją genezę w opisanej pracy. Nie są to jednak jedyne podejścia, alternatywne modele zostały przedstawione np. w pracy Melamed (2000) oraz innych wcześniejszych pracach tego autora. Te modele nie są skierowane, zamiast tego autor wychodzi z założenia, że tłumaczenia są zawsze swoimi wzajemnymi tłumaczeniami. Również kolejność wyrazów gra podrzędną rolę, mówi się wtedy o tzw. *Bag of words models*, czyli „modelach z workami słów".

Niemniej prace te nie wywarły takiego wpływu na społeczność badaczy, jak publikacja Brown et al., którą można swobodnie zaliczyć do jednej z najważniejszych prac lingwistyki komputerowej. Widać tutaj przy okazji, jak bardzo korzystny jest wpływ metod statystycznych oraz matematycznych na dziedziny, które wydają się być w pierwszym momencie tak oddalone od matematyki jak językoznawstwo. O ile powiązania językoznawstwa z matematyką są oczywiste dla wąskiego kręgu lingwistów komputerów i matematycznych, to mogą one zaskoczyć zarówno matematyków jak i klasycznych językoznawców. Spadkobiercy Brown et al. udostępniają owoce tego połączenia już dzisiaj darmowo w internecie. Kilka miesięcy temu język polski pojawił się jako jeden z możliwych języków źródłowych i docelowych w systemie statystycznego systemu tłumaczenia Google Translate.

Bibliografia

Al-Onaizan, Y., J. Curin, M. Jahr, K. Knight, J. Lafferty, I. Melamed, F. Och, D. Purdy, N. Smith i D. Yarowsky (1999). Statistical machine translation. Rap. tech., JHU workshop.

Brown, P. F., V. J. Della Pietra, S. A. Della Pietra i R. L. Mercer (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Dempster, A. P., N. M. Laird i D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Jurafsky, D. i J. H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (International Edition)*. Prentice Hall.

Kay, M. i M. Röscheisen (1993). Text-translation alignment. *Comput. Linguist.*, 19(1):121–142.

Knight, K. (1999). A Statistical MT Tutorial Workbook. Niepublikowane.

Manning, C. D. i H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Melamed, I. D. (2000). Models of translational equivalence among words. *Comput. Linguist.*, 26(2):221–249.

Och, F. J. i H. Ney (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Somers, H. (2001). Bilingual parallel corpora and Language Engineering.