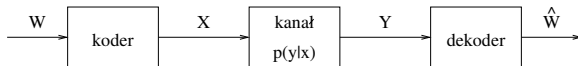


Model zaszumionego kanału



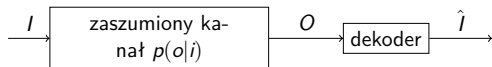
Oryginalna praca Shannona polegała na poszukiwaniu takiego kodowania, które umożliwiało ustalenie nadmiarowości informacji w taki sposób, żeby na wyjściu można było odtworzyć oryginalną wiadomość nawet przy obecności zakłóceń. Dla kanału bez pamięci drugie prawo Shannona mówi, że przepustowość kanału może być określona na podstawie informacji wzajemnej:

$$C = \max_{p(X)} I(X; Y)$$

W zastosowaniach przetwarzania języka naturalnego usiłujemy odtworzyć oryginalne wejście (oryginalny przekaz) na podstawie wyjścia zaszumionego kanału. Ponieważ wyjście jest dane, jest stałe dla wszystkich przypadków i można jego prawdopodobieństwo pominąć:

$$\begin{aligned}\hat{i} &= \arg \max_i p(i|o) = \arg \max_i \frac{p(i)p(o|i)}{p(o)} = \\ &= \arg \max_i p(i)p(o|i)\end{aligned}$$

Model zaszumionego kanału w NLP



| Zastosowanie | Wejście | Wyjście | $p(i)$ | $p(o i)$ |
|-----------------------|-------------------------|-------------------------|-------------------------------|-------------------|
| tłumaczenie maszynowe | ciągi słów języka J_1 | ciągi słów języka J_2 | $p(J_1)$ w modelu j. | model tłumaczenia |
| OCR | tekst | tekst z błędami | $p(\text{text})$ | model błędów OCR |
| oznaczanie POS | ciągi znaczników POS | słowa języka | $p(\text{ciągów znaczników})$ | $p(w t)$ |
| rozpoznawanie mowy | ciągi słów | sygnał mowy | $p(\text{ciągów słów})$ | model akustyczny |

Model zaszumionego kanału w poprawianiu pisowni

W poprawianiu pisowni mamy niepoprawny łańcuch znaków s i słownik D zawierający poprawne słowa. Szukamy takiego $w \in D$, które najprawdopodobniej było słowem, które zostało zmienione w wyniku błędów. Chcemy znaleźć $\arg \max_w P(w|s)$. Na mocy twierdzenia Bayesa i po odrzuceniu mianownika otrzymujemy $\arg \max_w P(s|w)P(w)$.

Model zaszumionego kanału w poprawianiu pisowni

Zakładamy, że osoba pisząca i popełniająca błędy pisowni dzieli słowo na mniejsze części i następnie każdą z tych części pisze – poprawnie lub nie:

| | | | | | | | | |
|---|--|---|--|----|--|---|--|---|
| w | | s | | ch | | ó | | d |
| f | | s | | h | | u | | t |

Prawdopodobieństwo zapisania słowa *wschód* jako *fshut* przy danym podziale słowa wyniesie wówczas $P(f|w) * P(s|s) * P(h|ch) * P(u|ó) * P(t|d)$.

Model zaszumionego kanału w poprawianiu pisowni

Niech $\text{Part}(w)$ będzie zbiorem wszystkich podziałów słowa w . Dla szczególnego podziału $R \in \text{Part}(w)$, gdzie $|R| = j$ oznacza, że R składa się z j przyległych do siebie odcinków, niech R_i będzie i -tym odcinkiem. Wtedy:

$$P(s|w) = \sum_{R \in \text{Part}(w)} P(R|w) \sum_{T \in \text{Part}(s): |T|=|R|} \prod_{i=1}^{|R|} P(T_i|R_i)$$

Rozważając tylko najlepszy podział s i w :

$$P(s|w) = \max_{R \in \text{Part}(w), T \in \text{Part}(s)} P(R|w) \prod_{i=1}^{|R|} P(T_i|R_i)$$

Model zaszumionego kanału w poprawianiu pisowni

Do uczenia modelu potrzebujemy par (s_i, w_i) . Używając odległości edycyjnej, wyrównujemy pary, np.:

| | | | | | |
|---|---|---|---|---|---|
| g | u | c | h | i | s |
| g | ł | u | | s | i |

Otrzymujemy ciąg operacji edycyjnych: $g \rightarrow g$, $\text{ł} \rightarrow \epsilon$, $u \rightarrow u$, $\epsilon \rightarrow c$, $\epsilon \rightarrow h$, $si \rightarrow is$. Aby skorzystać z szerszej informacji kontekstowej, każda zamiana różna od identyczności jest rozszerzana na do N najbliższych operacji. Np. dla $\epsilon \rightarrow h$ i dla $N = 2$, dodajemy jeszcze: $\epsilon \rightarrow ch$, $s \rightarrow hi$ u $\rightarrow uch$, $s \rightarrow chi$, $si \rightarrow his$.

Model zaszumionego kanału w poprawianiu pisowni

Do uczenia modelu potrzebujemy par (s_i, w_i) . Używając odległości edycyjnej, wyrównujemy pary, np.:

| | | | | | |
|---|---|---|---|---|---|
| g | u | c | h | i | s |
| g | ł | u | | s | i |

Otrzymujemy ciąg operacji edycyjnych: $g \rightarrow g$, $\text{ł} \rightarrow \epsilon$, $u \rightarrow u$, $\epsilon \rightarrow c$, $\epsilon \rightarrow h$, $si \rightarrow is$. Aby skorzystać z szerszej informacji kontekstowej, każda zamiana różna od identyczności jest rozszerzana na do N najbliższych operacji. Np. dla $\epsilon \rightarrow h$ i dla $N = 2$, dodajemy jeszcze: $\epsilon \rightarrow ch$, $s \rightarrow hi$ u $\rightarrow uch$, $s \rightarrow chi$, $si \rightarrow his$.

Model zaszumionego kanału w poprawianiu pisowni

Do uczenia modelu potrzebujemy par (s_i, w_i) . Używając odległości edycyjnej, wyrównujemy pary, np.:

| | | | | | |
|---|---|---|---|---|---|
| g | u | c | h | i | s |
| g | ł | u | | s | i |

Otrzymujemy ciąg operacji edycyjnych: $g \rightarrow g$, $\text{ł} \rightarrow \epsilon$, $u \rightarrow u$, $\epsilon \rightarrow c$, $\epsilon \rightarrow h$, $si \rightarrow is$. Aby skorzystać z szerszej informacji kontekstowej, każda zamiana różna od identity jest rozszerzana na do N najbliższych operacji. Np. dla $\epsilon \rightarrow h$ i dla $N = 2$, dodajemy jeszcze: $\epsilon \rightarrow ch$, $s \rightarrow hi$ i $u \rightarrow uch$, $s \rightarrow chi$, $si \rightarrow his$.

Model zaszumionego kanału w poprawianiu pisowni

Do uczenia modelu potrzebujemy par (s_i, w_i) . Używając odległości edycyjnej, wyrównujemy pary, np.:

| | | | | | |
|---|---|---|---|---|---|
| g | u | c | h | i | s |
| g | ł | u | | s | i |

Otrzymujemy ciąg operacji edycyjnych: $g \rightarrow g$, $\text{ł} \rightarrow \epsilon$, $u \rightarrow u$, $\epsilon \rightarrow c$, $\epsilon \rightarrow h$, $si \rightarrow is$. Aby skorzystać z szerszej informacji kontekstowej, każda zamiana różna od identyczności jest rozszerzana na do N najbliższych operacji. Np. dla $\epsilon \rightarrow h$ i dla $N = 2$, dodajemy jeszcze: $\epsilon \rightarrow ch$, $s \rightarrow hi$ u $\rightarrow uch$, $s \rightarrow chi$, $si \rightarrow his$.

Model zaszumionego kanału w poprawianiu pisowni

Do uczenia modelu potrzebujemy par (s_i, w_i) . Używając odległości edycyjnej, wyrównujemy pary, np.:

| | | | | | | |
|---|---|---|---|---|---|---|
| g | | u | c | h | i | s |
| g | ł | u | | | s | i |

Otrzymujemy ciąg operacji edycyjnych: $g \rightarrow g$, $\text{ł} \rightarrow \epsilon$, $u \rightarrow u$, $\epsilon \rightarrow c$, $\epsilon \rightarrow h$, $si \rightarrow is$. Aby skorzystać z szerszej informacji kontekstowej, każda zamiana różna od identyczności jest rozszerzana na do N najbliższych operacji. Np. dla $\epsilon \rightarrow h$ i dla $N = 2$, dodajemy jeszcze: $\epsilon \rightarrow ch$, $s \rightarrow hi$ $u \rightarrow uch$, $s \rightarrow chi$, $si \rightarrow his$.

Model zaszumionego kanału w poprawianiu pisowni

Do uczenia modelu potrzebujemy par (s_i, w_i) . Używając odległości edycyjnej, wyrównujemy pary, np.:

| | | | | | |
|---|---|---|---|---|---|
| g | u | c | h | i | s |
| g | ł | u | | s | i |

Otrzymujemy ciąg operacji edycyjnych: $g \rightarrow g$, $\text{ł} \rightarrow \epsilon$, $u \rightarrow u$, $\epsilon \rightarrow c$, $\epsilon \rightarrow h$, $si \rightarrow is$. Aby skorzystać z szerszej informacji kontekstowej, każda zamiana różna od identyczności jest rozszerzana na do N najbliższych operacji. Np. dla $\epsilon \rightarrow h$ i dla $N = 2$, dodajemy jeszcze: $\epsilon \rightarrow ch$, $s \rightarrow hi$ u $\rightarrow uch$, $s \rightarrow chi$, $si \rightarrow his$.

Model zaszumionego kanału w poprawianiu pisowni

Do uczenia modelu potrzebujemy par (s_i, w_i) . Używając odległości edycyjnej, wyrównujemy pary, np.:

| | | | | | |
|---|---|---|---|---|---|
| g | u | c | h | i | s |
| g | ł | u | | s | i |

Otrzymujemy ciąg operacji edycyjnych: $g \rightarrow g$, $\text{ł} \rightarrow \epsilon$, $u \rightarrow u$, $\epsilon \rightarrow c$, $\epsilon \rightarrow h$, $si \rightarrow is$. Aby skorzystać z szerszej informacji kontekstowej, każda zamiana różna od identyczności jest rozszerzana na do N najbliższych operacji. Np. dla $\epsilon \rightarrow h$ i dla $N = 2$, dodajemy jeszcze: $\epsilon \rightarrow ch$, $s \rightarrow hi$ u $\rightarrow uch$, $s \rightarrow chi$, **$si \rightarrow his$** .

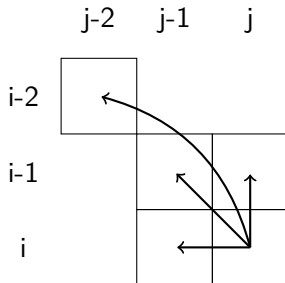
Model zaszumionego kanału w poprawianiu pisowni

Prawdopodobieństwo zamiany $\alpha \rightarrow \beta$ można policzyć licząc liczby wystąpień: $\text{liczba}(\alpha \rightarrow \beta) / \text{liczba}(\alpha)$. Jeśli dysponujemy tylko parami (s_i, w_i) , to policzenie $\text{liczba}(\alpha)$ jest trudne. Można ją przybliżyć licząc liczbę wystąpień α w reprezentatywnym zbiorze tekstów, a następnie dostosowując ją do naszego oszacowania częstości, z jaką ludzie popełniają błędy w pisowni.

Stosowanie modelu

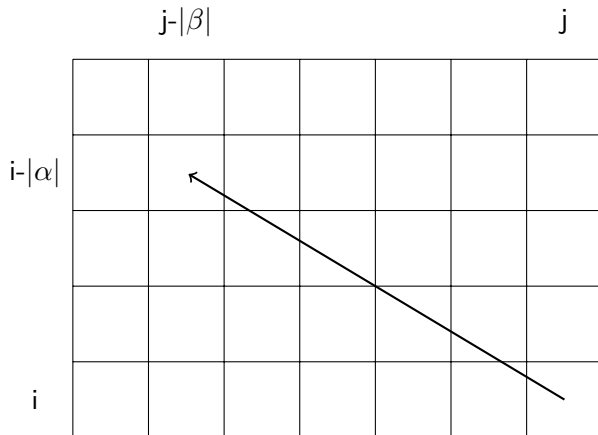
Mając słownik D w postaci automatu i prawdopodobieństwa $P(\alpha \Rightarrow \beta)$ możemy znaleźć najlepszego kandydata na zastąpienie niepoprawnej formy w podobny sposób do podanego na wcześniejszej godzinie wykładu.

Poprzednio dla elementu macierzy o indeksach (i, j) mogliśmy rozpatrywać tylko elementy o indeksach $(i, j-1)$, $(i-1, j-1)$ i $(i-1, j)$ oraz $(i-2, j-2)$.



Stosowanie modelu

Teraz musimy sięgać często znacznie dalej: o $(|\beta|, |\alpha|)$. Zastępujemy nieprawidłowy fragment β prawidłowym α , po czym usuwamy oba z końców słów. Dla każdej pary (α, β) ...



Prawdopodobieństwa $p(\alpha \rightarrow \beta)$ można przechowywać w drzewie w taki sposób, że ścieżki od korzenia do liści odpowiadają łańcuchom α , natomiast w liściach przechowywane są korzenie drzew zawierających łańcuchy β . W liściach tych drugich drzew przechowywane są prawdopodobieństwa zamiany.

- Błędy OCR, rozpoznawania mowy, rozpoznawania pisma odręcznego na bieżąco
- Dodanie uzależnienia prawdopodobieństwa zastąpienia od miejsca w wyrazie (początek, środek, koniec)
- Uzależnienie prawdopodobieństwa zastąpienia od poprzedzających wyrazów
- Przyspieszanie wyszukiwania (automaty)

- 1 Karen Kukich, *Techniques for Automatically Correcting Words in Texts*, Machine Learning 24(4), str. 377–439, 1992.
- 2 Kemal Oflazer, Cemalettine Guzey, *Spelling Correction in Agglutinative Languages*, materiały konferencji Applied Natural Language Processing, Stuttgart, Niemcy, 1994. Dostępne pod: <http://acl.ldc.upenn.edu/A/A94/A94-1037.pdf>
- 3 Stoyan Mihov, Klaus U. Schulz, *Fast Approximate Search in Large Dictionaries*, Computational Linguistics 30(4), str. 451–471, grudzień 2004. Dostępne pod: <http://acl.ldc.upenn.edu/J/J04/J04-4003.pdf>
- 4 Eric Brill, Robert C. Moore, *An Improved Error Model for Noisy Channel Spelling Correction*, Proceedings of the 38th Meeting of the Association for Computational Linguistics, Hong Kong, październik 2000. Dostępne pod: <http://acl.ldc.upenn.edu/P/P00/P00-1037.pdf>