Introducción al Procesamiento de Lenguaje Natural

Grupo de PLN - InCo

Detección y corrección de errores ortográficos

El modelo del Canal Ruidoso

Modelos probabilísticos

- Similar al reconocimiento de voz
- Método de inferencia

$$p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow p_4$$

Modelado: *canal ruidoso*

Se intenta reconstruir la señal... pero...
pueden haber varios candidatos...
entonces...
tratamos de determinar la ***probabilidad máxima***

Detección de errores

- Detección de palabras inexistentes (*tmate*)
- Corrección aislada (*tmate* -> *tomate*)
- Detección y corrección dependiente del contexto

  (*calor* -> *color*)

Detección de errores

Pueden deberse a:

- Inserción (*toomate*)
- Borrado (*tmate*)
- Sustitución (*tpmate*)
- Trasposición (*tmoate*)

Estudio (Kernighan, 1990)

1% a 3% palabras con errores

de estos, el 80% eran por borrado o inserción

Detección de errores

Detección de palabras inexistentes

- Diccionario
- Tener en cuenta la morfología (morfotácticas)

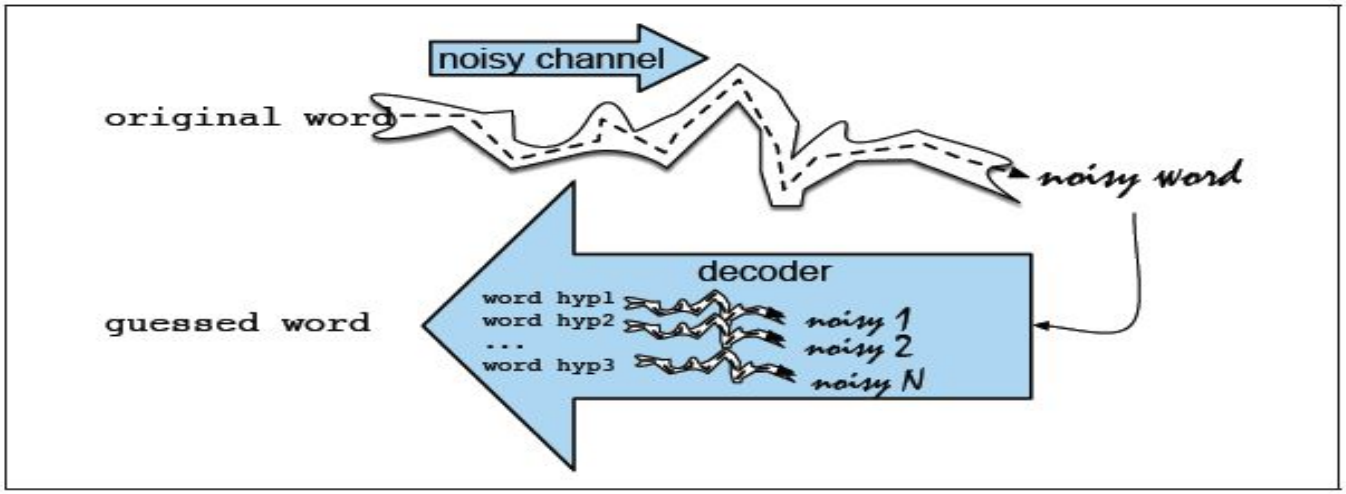
  Recordar que los morfemas pueden combinarse de acuerdo a ciertas reglas

  inevitable
  *inelefante
  rápidamente
  inelefantemente ?

Modelos probabilísticos

Modelos probabilísticos para corrección

Canal Ruidoso



Caso particular de inferencia Bayesiana

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w|O)$$

Modelos probabilísticos

Regla de Bayes

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} \frac{P(O|w)P(w)}{P(O)}$$

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} \frac{P(O|w)P(w)}{P(O)} = \underset{w \in V}{\operatorname{argmax}} P(O|w)\,P(w)$$

Corrección de errores

Algoritmo bayesiano (Kernighan - 1990)

- Hipótesis: errores son por inserción, borrado, sustitución y transposición
- Aplico todas las transformaciones posibles a la palabra observada y busco lista de candidatos válidos considerando la DME

		Transformation			
Error	Correction	Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	—	2	deletion
acress	cress	—	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	—	s	5	insertion
acress	acres	—	s	4	insertion

Distancia de Damerau-Levenshtein

Corrección de errores

$$\hat{w} = \underset{w \in C}{\operatorname{argmax}} \overbrace{P(x|w)}^{\text{channel model}} \overbrace{P(w)}^{\text{prior}}$$

w	count(w)	p(w)
actress	9,321	.0000231
cress	220	.000000544
caress	686	.00000170
access	37,038	.0000916
across	120,844	.000299
acres	12,874	.0000318

$$P(w) = \frac{Count(w) + 0{,}5}{N + 0{,}5V}$$

probabilidad “a priori” que aparezca la palabra en el corpus

Modelos probabilísticos

Tenemos P(w), que es la probabilidad a priori

Pero... cómo calculamos P(x|w)?

- En un corpus de errores, ¿cuántas veces se sustituye?
- Matriz de confusión que contiene las cantidades de ocurrencias en que aparece una letra delante de otra

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)
actress	t	-	c|ct	.000117
cress	-	a	a|#	.00000144
caress	ca	ac	ac|ca	.00000164
access	c	r	r|c	.000000209
across	o	e	e|o	.0000093
acres	-	s	es|e	.0000321
acres	-	s	ss|s	.0000342

probabilidad de que se borre una "**t**" después de una "**c**"

Modelos probabilísticos

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)	P(w)	10^9*P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	0.00078
caress	ca	ac	ac|ca	.00000164	.00000170	0.0028
access	c	r	r|c	.000000209	.0000916	0.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Modelos probabilísticos

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)	P(w)	10^9*P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	0.00078
caress	ca	ac	ac|ca	.00000164	.00000170	0.0028
access	c	r	r|c	.000000209	.0000916	0.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Corrección de errores

...was called a "stellar and versatile **acress** *whose combination of sass and glamour has defined her ....*

Corrección de errores

...was called a "stellar and versatile **acress** *whose combination of sass and glamour has defined her ....*

daría

...was called a "stellar and versatile **across** *whose combination of sass and glamour has defined her ....*

Corrección de errores

...was called a "stellar and versatile **acress** *whose combination of sass and glamour has defined her ....*

daría

...was called a "stellar and versatile **across** *whose combination of sass and glamour has defined her ....*

debiera dar

...was called a "stellar and versatile **actress** *whose combination of sass and glamour has defined her ....*

Referencias

J.Martin & D.Jurafsky. Speech and Language Processing. Tercera Edición. Capítulo 5

M. Kernighan, K. Church, W. Gale. A spelling correction program based on a noisy channel model