# Statistical structured prediction

Question set (Part 1-B)

Adrián Vázquez Barrera

Polytechnic University of Valencia

**January 2022**

# Pregunta 1

Si asignamos la misma probabilidad a cada una de las producciones de la gramática obtenemos lo siguiente:

1.0 S → Suj Pre
1/3 Suj → Art Nom
1/3 Suj → Art Adj Nom
1/3 Suj → Art Nom Adj
1/2 Pred → Verb Nom
1/2 Pred → Verb
1.0 Art → “la”
1/5 Nom → “vieja”
1/5 Nom → “ayuda”
1/5 Nom → “mujer”
1/5 Nom → “pelea”
1/5 Nom → “demanda”
1/4 Verb → “demanda”
1/4 Verb → “ayuda”
1/4 Verb → “oculta”
1/4 Verb → “pelea”
1/2 Adj → “vieja”
1/2 Adj → “oculta”

Con estas reglas, puede observarse como para la producción de las frases:

„ $\overline{\underset{Art}{la}}\ \overline{\underset{Nom}{vieja}}\ \overline{\underset{Verb}{oculta}}\ \overline{\underset{Nom}{pelea}}$„ y „ $\overline{\underset{Art}{la}}\ \overline{\underset{Nom}{mujer}}\ \overline{\underset{Verb}{demanda}}\ \overline{\underset{Nom}{ayuda}}$„

Al estar compuestas por los mismos elementos gramaticales y en la misma cantidad, Inside-Outside les otorga la misma probabilidad:

$$1 \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot 1 \cdot \frac{1}{5} \cdot \frac{1}{4} \cdot \frac{1}{5} = \frac{1}{600} \approx 0.0016667 \ .$$

Si aplicamos el algoritmo Inside-Outside con una serie reducida de muestras, observamos que las probabilidades no difieren mucho de las establecidas inicialmente. Se intuye que, esta inicialización puede ser útil a la hora de entrenar modelos con pocos datos, lo que ayuda a evitar sesgos al sobre-aprender algunas muestras.

# Pregunta 2

Se pide calcular la estimación de la regla (Suj → Art Nom) con el algoritmo Inside-Outside para la siguiente muestra de entrenamiento que incluye muestras entre parentesis: $\mathcal{D} = \{$(la vieja)(demanda ayuda),la mujer oculta pelea, la vieja ayuda$\}$.

Destacar que, para la primera muestra al estar parentizada, reduce el número de combinaciones aceptadas.

$P_\theta$((la vieja) ( demanda ayuda)) = 0.0009

$P_\theta$(la mujer oculta pelea) = 0.00090 + 0.01176 = 0.01266

$P_\theta$(la vieja ayuda) = 0.007

$$\bar{p}(Suj \rightarrow Art\ Nom) = \frac{\sum_{x \in \mathcal{D}} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj \rightarrow Art\ Nom) P_\theta(x, t_x)}{\sum_{x \in \mathcal{D}} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj) P_\theta(x, t_x)}$$

$$\bar{p}(Suj \rightarrow Art\ Nom) = \frac{\frac{0.0009}{0.0009} + \frac{0.0009}{0.01266} + \frac{0.007}{0.007}}{\frac{0.0009}{0.0009} + \frac{0.0009+0.01176}{0.01266} + \frac{0.007}{0.007}} = \frac{\frac{437}{211}}{3} \approx 0.6904$$

# Pregunta 3

Conociendo los datos del ejercicio anterior es posible estimar la probabilidad con Viterbi, debido a que solo utiliza los arboles con la maxima probabilidad para el calculo. Esto reduce el tiempo de computo a expensas de perder algo de precisión.

$$\bar{p}(Suj \rightarrow Art\ Nom) = \frac{\sum_{t_x} N(Suj \rightarrow Art\ Nom) P_\theta(x, \hat{t_x})}{\sum_{t_x} N(Suj) P_\theta(x, \hat{t_x})}$$

$$\bar{p}(Suj \rightarrow Art\ Nom) = \frac{\frac{0.0009}{0.0009} + \frac{0.007}{0.007}}{\frac{0.0009}{0.0009} + \frac{0.0009+0.01176}{0.01266} + \frac{0.007}{0.007}} = \frac{2}{3} \approx 0.6667$$

# Pregunta 4

Se quiere estimar la probabilidad de la regla (Suj → Art Nom) utilizando el algoritmo Inside-Outside para las muestras: $\mathcal{D} = \{$la vieja demanda ayuda, la mujer oculta pelea, la vieja mujer oculta demanda ayuda$\}$.

Para empezar, la ultima muestra no puede generarse a partir de la gramática proporcionada, lo que nos obligará a crear una nueva regla (con su ajuste de probabilidades correspondiente) para poder realizar el ejercicio:

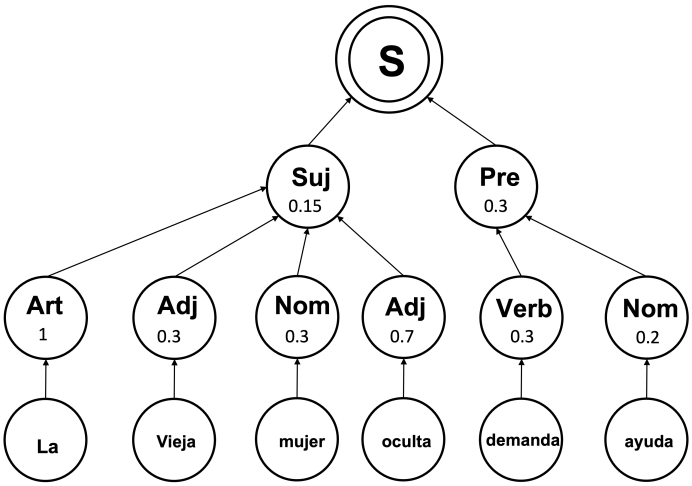
0.4 Suj → Art Nom
0.25 Suj → Art Adj Nom
0.2 Suj → Art Nom Adj
0.15 Suj → Art Adj Nom Adj

Con las reglas actualizadas, la probabilidad para cada una de las muestras es:

$P_\theta(la\ vieja\ demanda\ ayuda) = (1{\cdot}0.4{\cdot}1{\cdot}0.1{\cdot}0.3{\cdot}0.3{\cdot}0.2){+}(1{\cdot}0.25{\cdot}1{\cdot}0.3{\cdot}0.2{\cdot}0.7{\cdot}0.2) = (0.00072) + (0.0021) = 0.00282$

$P_\theta(la\ mujer\ oculta\ pelea) = (1{\cdot}0.4{\cdot}1{\cdot}0.3{\cdot}0.3{\cdot}0.1{\cdot}0.2){+}(1{\cdot}0.2{\cdot}1{\cdot}0.3{\cdot}0.7{\cdot}0.7{\cdot}0.4) = (0.00072) + (0.01176) = 0.01248$

$P_\theta(la\ vieja\ mujer\ oculta\ demanda\ ayuda) = 1 \cdot 0.15 \cdot 1 \cdot 0.3 \cdot 0.3 \cdot 0.7 \cdot 0.3 \cdot 0.3 \cdot 0.2 = 0.0001701$



$$\bar{p}(Suj \rightarrow Art\ Nom) = \frac{\sum_{x \in \mathcal{D}} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj \rightarrow Art\ Nom) P_\theta(x, t_x)}{\sum_{x \in \mathcal{D}} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj) P_\theta(x, t_x)}$$

$$\bar{p}(Suj \rightarrow Art\ Nom) = \frac{\frac{0.00072}{0.00282} + \frac{0.00072}{0.01248}}{\frac{(0.00072)+(0.0021)}{0.00282} + \frac{(0.00072)+(0.01176)}{0.01248} + \frac{0.0001701}{0.0001701}} \approx 0.104$$

# Pregunta 5

Para poder realizar el ejercicio es necesario obtener los dos mejores arboles de derivación y repetir el proceso de la pregunta 4. No obstante, vemos como no existe ninguna muestra que tenga más de dos arboles de derivación, por lo que los resultados serán los mismos.

# Pregunta 7

Podemos observar como al aumentar el numero de símbolos no terminales se incrementa proporcionalmente el número de trángulos rectángulos. Al aumentar la flexibilidad del modelo, lo dotamos de más reglas, por lo que es capaz de adaptarse mejor a la configuración de los triángulos.

| **# non-terminal symbols** | **# rectangle triangles** |
|---|---|
| 5 | 29 |
| 10 | 63 |
| 15 | 61 |
| 20 | 84 |

# Pregunta 8

Observamos como el algoritmo Inside-Outside es capaz de reconocer mejor los triángulos rectángulos y equiláteros que Viterbi. Por otro lado, este último es capaz de clasificar mejor los triángulos isosceles.

En cualquiera de los casos, se observa como al utilizar muestras parentizadas el error cometido a rasgos generales aumenta. Esto es debido a que, al entrenar con estas se corre el riesgo de no reconocer determinadas estructuras que si podríamos llegar a ver en el conjunto de test sin parentizar. A pesar de ello, se corrige el número de muestras mal clasificadas por Viterbi para los triángulos rectángulos y en el caso de Inside-Outside con los Isosceles.

El algoritmo Inside-Outside (en ambos casos, parentizado y sin parentizar) ofrece el menor error. Esto es debido a que utiliza en sus calculos las probabilidades de todos los aroboles de derivación por cada muestra, al contrario que Viterbi. Por contra, el coste computacional es considerablente superior.

| Inside-Outside (Bracketed) | | | | | |
|---|---|---|---|---|---|
| | Equilateral | Isoscalen | Right | Error | Error % |
| Equilateral | 794 | 206 | 0 | 206 | 20.6 |
| Isoscalen | 531 | 225 | 244 | 775 | 77.5 |
| Right | 108 | 145 | 747 | 253 | 25.3 |
| Error | 41.13% | | | | |

| Inside-Outside (Non-Bracketed) | | | | | |
|---|---|---|---|---|---|
| | Equilateral | Isoscalen | Right | Error | Error % |
| Equilateral | 783 | 217 | 0 | 217 | 21.7 |
| Isoscalen | 483 | 366 | 151 | 634 | 63.4 |
| Right | 48 | 187 | 765 | 235 | 23.5 |
| Error | 36.20% | | | | |

| Viterbi (Bracketed) | | | | | |
|---|---|---|---|---|---|
| | Equilateral | Isoscalen | Right | Error | Error % |
| Equilateral | 77 | 843 | 80 | 923 | 92.3 |
| Isoscalen | 70 | 850 | 80 | 150 | 15.0 |
| Right | 12 | 676 | 312 | 688 | 68.8 |
| Error | 58.70% | | | | |

| Viterbi (Non-Bracketed) | | | | | |
|---|---|---|---|---|---|
| | Equilateral | Isoscalen | Right | Error | Error % |
| Equilateral | 67 | 933 | 0 | 933 | 93.3 |
| Isoscalen | 171 | 612 | 217 | 388 | 38.8 |
| Right | 55 | 372 | 573 | 427 | 42.7 |
| Error | 58.27% | | | | |

# Pregunta 9

Observamos como a media que se aumenta el número de muestras, el error obtenido disminuye, esto tiene sentido, pues a más muestras diferentes vistas en el entrenamiento, más preparado estará el modelo. De nuevo y tal y como se comentó en el ejercicio anterior, esto se debe a que Viterbi utiliza para el calculo, los arboles con mayor probabilidad. De esta manera, el coste computacional es mucho más contenido que el ofrecido por Inside-Outside, que ofrece mejores resultados a cambio de un tiempo de computo superior.

| Error % | | |
|---|---|---|
| Size | Viterbi | Inside-Outside |
| 10 | 57.57 | 53.43 |
| 100 | 58.97 | 43.97 |
| 500 | 49.23 | 42.73 |
| 1000 | 43.90 | 38.17 |