Evaluación en Procesamiento de Lenguaje Natural

- La evaluación es un componente imprescindible en los sistemas PLN, ya que el objetivo es la construcción de artefactos (ya sea de uso final o intermedio) que deben tener buena performance.
- La evaluación puede ser intrínseca o extrínseca.
- La evaluación puede ser autómatica o humana.

Medidas estándar, tareas compartidas

- La comunidad que trabaja en PLN desarrolla en distintas áreas conferencias que incluyen la resolución de tareas por distintos grupos que comparan sus trabajos mediante medidas comunes, que dependen de la tarea.
- Se habla entonces del estado del arte (SOTA) en traducción, o en tagging, o en parsing. P.ej., en tagging es del 97

Límites inferior (*baseline*) y superior(*upper bound*)

- Se considera como límite inferior la performance de un método simple para resolver el problema. Si es un problema de clasificación, puede ser elegir al azar entre las distintas clases, o dar como salida siempre la clase más frecuente. A veces no es fácil superar este tipo de límites.

Límites inferior (*baseline*) y superior(*upper bound*)

- La cota superior se ha definido para algunas tareas (p.ej., tagging) como la medida asociada al desempeño humano en la tarea. Debemos notar que si bien un humano domina el lenguaje, no necesariamente debe ser competente en tareas que son un tanto artificales. Cuando se habla de un límite superior, se suele pensar en la tasa de concordancia entre varios humanos haciendo la tarea, que involucra en general anotar texto. Por ejemplo, en tagging la cota superior es igual al estado del arte. Se consideraría entonces que esta tarea está resuelta.

Límites inferior (*baseline*) y superior(*upper bound*)

- La cota superior se ha definido para algunas tareas (p.ej., tagging) como la medida asociada al desempeño humano en la tarea. Debemos notar que si bien un humano domina el lenguaje, no necesariamente debe ser competente en tareas que son un tanto artificales. Cuando se habla de un límite superior, se suele pensar en la tasa de concordancia entre varios humanos haciendo la tarea, que involucra en general anotar texto. Por ejemplo, en tagging la cota superior es igual al estado del arte. Se consideraría entonces que esta tarea está resuelta.
- Una medida de interés es la cota del estado del arte (SOTA). P.ej., en tagging es del 97 %.

Particionamiento de los datos

- Cualquiera sea el método que se aplique en un sistema basado en datos, es importante dejar completamente separada una parte de las datos para poder medir el desempeño del sistema
- Lo usual suele ser una partición en 3 conjuntos: entrenamiento (70 %), desarrollo (10 %) y testeo (20 %)
- El conjunto de testeo debe quedar completamente separado en toda la etapa de perfeccionamiento del sistema.

Matriz de confusión

		ALGORITMO		
		pos	neg	neutro
	pos	15	10	100
REAL	neg	10	15	10
	neutro	10	100	1000

- La matriz de confusión resume resultados en un contexto en el que hay datos anotados con clases discretas.
- Por filas tiene la clase según viene en los datos (clases reales o *Gold Standard*). P.ej., sabemos que en los datos hay 125 elementos positivos, porque es la suma de la fila rotulada 'pos'.

Matriz de confusión

		ALGORITMO		
		pos	neg	neutro
	pos	15	10	100
REAL	neg	10	15	10
	neutro	10	100	1000

- La matriz de confusión resume resultados en un contexto en el que hay datos anotados con clases discretas.
- Por columnas tiene la clase que predice el algoritmo. P.ej., sabemos que el algoritmo reconoce 1110 elementos neutros porque es la suma de la columna 'neutro'.

Medidas por clase: *Accuracy*

		ALGORITMO		
		pos	neg	neutro
	pos	15	10	100
REAL	neg	10	15	10
	neutro	10	100	1000

- Cantidad de aciertos/Cantidad de casos
- es una medida global, con mínimo 0 y máximo 1
- Oculta lo que pueda ocurrir con clases pequeñas, las clases con muchos casos dominan.

Medidas por clase: *Precision*

		ALGORITMO		
		pos	neg	neutro
	pos	15	10	100
REAL	neg	10	15	10
	neutro	10	100	1000

- Mide la proporción de elementos recuperados que son correctos.
- Recuperados correctos/Total recuperados
- Tiende a favorecer que se reconozcan pocos elementos
- Para la clase 'pos' la precision es $15/35 = 0{,}42$

Medidas por clase: *Recall*

		ALGORITMO		
		pos	neg	neutro
	pos	15	10	100
REAL	neg	10	15	10
	neutro	10	100	1000

- Mide la proporción de recuperados correctos sobre el total de correctos.
- Mejora cuanto más grande sea el conjunto de recuperados.
- Para la clase 'pos' el recall es $15/125 = 0{,}12$

Medidas por clase: *Medida F*

		ALGORITMO		
		pos	neg	neutro
	pos	15	10	100
REAL	neg	10	15	10
	neutro	10	100	1000

- Las medidas precision y recall tienden a mejorar en sentidos contrarios, la medida F trata de balancear ambas medidas.

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- El factor β regula de la importancia de la precision. Se suele usar con β en 1