Aplicaciones de análisis

Práctica 1: Introducción al PLN

"Garbage-in/Garbage-out" es una expresión en forma de concepto ampliamente usada en el mundo de la informática que significa que, si la información/datos que entran en un sistema son malos, aquello que produzca el sistema, también lo será.

Bajo esta premisa se fundamenta gran parte de la minería de datos. A la hora de recopilar la información necesaria para poder resolver un problema, se deberá de comprobar su calidad, es decir, como de bien representa la información que está presente en un problema. Una vez esa calidad haya sido asegurada, la información deberá de pasar por distintas transformaciones hasta que sea estructurada. Una vez esta información esté preparada para poder ser la entrada de cualquier algoritmo de aprendizaje automático, las probabilidades de que el sistema aprenda serán mucho mayores que si no hubiésemos depurado los datos que tomamos como entrada. Los datos desestructurados componen un total del 80% de los datos que generan a lo largo de internet, por lo que encontrar patrones que ayuden a poder estructurar puede ser a veces una tarea tediosa.

Uno de los ejemplos de datos desestructurados que podemos encontrar son los corpus lingüísticos. Los corpus lingüísticos son una colección de textos orientados a la representación de la lengua o la variedad lingüística en un campo el cual va a ser investigado. Para el análisis de texto y el procesamiento de lenguaje natural, una de las maneras de poder "estructurar" los textos de manera que estos puedan ser mas fáciles de ser procesados, es mediante el etiquetado de distintas partes de un texto. A esto se le conoce como anotación. La anotación variará en función distintos atributos que el corpus tenga. Los atributos mas presentes suelen ser:

1. En lo que se refiere a la **naturaleza de los datos**, es decir, de donde se han obtenido los corpus que vamos a analizar, tenemos los **Corpus Textuales**. Estos corpus son aquellos que se sacan de textos escritos. También se tienen los textos de grabaciones orales o de transcripciones. Estos corpus normalmente son representaciones del lenguaje hablado el cual se quiere analizar. Suele ser el tipo de lenguaje menos formal
2. La **temática** de los corpus puede variar. Desde corpus con propósitos generales a otros que tratan temas más específicos.

3. Los corpus podrán estar escritos en uno o varios idiomas. Los corpus bi,tri o multilingües deberán ser anotados con especial cuidado, llegando incluso a ser necesario el desarrollo y uso de aplicaciones relacionadas con la traducción para una anotación mucho mas ágil.
4. Unos corpus tendrán en mente referirse a uno u otro público. Es aquí donde nos encontramos con el género textual del corpus. Existe artículos que tienen un contexto científico, otro que está mas relacionado con el mundo de las noticias y así sucesivamente.
5. Los corpus pueden estar representando la evolución de un lenguaje (diacrónicos) o del lenguaje en un punto (sincrónicos).

Como se ha explicado en el apartado anterior, la contextualización de los textos es vital para poder saber como anotar los corpus.

Las anotaciones de los corpus lingüísticos es una de las tareas mas importantes dentro del NLP. Es la tarea que le da "forma" al texto desde el punto de vista de la máquina, ya que esta no entiende de contextos o entidades por sí sola. Es aquí donde juega un papel fundamental la metodología MATTER:

M: Model (Modelo de anotación)
A: Annotate (Anotación)
T: Train (Entrenamiento)
T: Test (Evaluación)
E: Evaluation (evaluación del acierto del modelo)
R: Revise (Revisión del modelo)

Como tal, no es un procedimiento unidireccional, si no que se ve mas como un ciclo. De hecho, existen subciclos dentro del ciclo MATTER, entre los dos primeros pasos, se les llama MA-MA (Model Annotate – Model Annotate).

Para la anotación de textos se han de tener un equipo de anotadores. Estos anotadores elaborarán una serie de directrices para el etiquetado de los corpus, adaptándose al contenido de este y lo que quiere representar. Estas directrices de anotación tendrán que ir evolucionando, por lo que se plateará lo que se conoce como **modelo de anotación**.

Para desarrollar correctamente un modelo de anotación, se ha de definir primero la tarea que se ha de resolver. Si no tenemos claro el objetivo de la anotación, nos será complejo anotar los términos que resulten de interés para que los algoritmos de aprendizaje automático. Como anotadores deberemos vernos inmersos en el contexto sobre el que el corpus se basa, por lo que documentarnos usando bibliografía relacionada, será una buena opción para poder destacar las partes del texto y etiquetarlas debidamente. Para verificar que el etiquetado sigue un rumbo correcto, se deberán diseñar heurísticas o

métricas que nos indiquen como de cerca estamos de un correcto funcionamiento del proceso de etiquetado. Esto refleja como de vital importancia es la parte de definir el objetivo que se quiere alcanzar con el etiquetado. Finalmente, con las directrices fijadas, construimos el modelo y comenzamos a probar como de preciso es.

La propia definición del modelo de anotación infiere un esquema en el modelo de anotación. Este esquema se definirá bajo 3 componentes: (T)érminos, (R)elación entre componentes y su (I)nterpretación. Es aquí donde se abre el abanico de las interpretaciones, lo que para un anotador puede representar una cosa, otro puede que lo interprete de otra manera. Una manera de poder paliar este problema es hacer que varios anotadores realicen anotaciones del corpus sin ser influenciados por otros anotadores. Posteriormente se puede comprobar las anotaciones en una matriz de confusión y así ir ajustando poco a poco los criterios de anotación. Estos ajustes entre las distintas interpretaciones de los anotadores se le conoce como “armonización”, ya que se trata de buscar una “armonía” (un sentido contextual) a aquello que se está anotando, siguiendo un criterio definido.

A medida que se van haciendo los ajustes pertinentes, se va llegando a un punto donde las métricas definidas en el modelo y por consecuente el esquema que implemente el modelo sea óptimo. Esto se le conocerá como “gold standard”.

Una vez alcanzado este “gold standard” podríamos decir que hemos elaborado una métrica la cual nos permite etiquetar de manera correcta los textos, identificando correctamente entidades, sentimiento y otros aspectos de un corpus, el cual facilitará en función de la calidad de dicho anotado, unos mejores resultados del algortimo de NPL.