*** Convocatoria de participación tarea BARR - IberEval 2017***

Tarea BARR - Ibereval 2017

Reconocimiento y Resolución de Abreviaturas Biomédicas (Biomedical Abbreviation Recognition and Resolution BARR)

19 de Septiembre de 2017: Workshop SEPLN 2017

http://temu.inab.org/index.html

==========

Descripción General

==========

El reconocimiento y la resolución de abreviaturas, siglas y símbolos es una etapa crítica para muchas tareas, incluyendo el reconocimiento de entidades (Named Entity Recognition, NER), la traducción automática, la recuperación e indexación de la información y la categorización de documentos, entre otras muchas. Por lo tanto, la implementación y la disponibilidad de sistemas de reconocimiento de abreviaturas tiene un gran impacto práctico en la minería de textos y en el procesamiento del lenguaje natural.

En el caso de dominios como la biomedicina y la investigación clínica, las abreviaturas son particularmente frecuentes, refiriéndose a menudo a entidades y conceptos de importancia como genes, enfermedades, síntomas, fármacos, productos químicos o tratamientos. El NER, la extracción de relaciones y los sistemas de codificación de documentos clínicos suelen tener que lidiar con el reconocimiento correcto de las denominadas formas cortas (ShortForm, SF) o abreviaturas.

Las abreviaturas se pueden considerar formas cortas que denotan una palabra o una frase más larga (Forma Larga, LongForm LF), que generalmente se corresponde con su definición. Se han probado diferentes estrategias para detectar formas cortas en textos biomédicos en inglés (Torii et al., 2007), utilizando, por ejemplo, enfoques basados en la alineación, métodos de aprendizaje automático o estrategias basadas en reglas, y algunos corpus

anotados manualmente (por ejemplo, MEDSTRACT, Ab3P O BOADI, véase Islamaj Doğan et al., 2014). Sin embargo, el esfuerzo para detectar pares SF-LF en textos escritos en otros idiomas ha sido mucho menor.

Existe un creciente número de documentos biomédicos y clínicos escritos en español, tales como literatura médica, informes de agencias médicas, patentes y registros de salud particularmente electrónicos. Además, según algunas estimaciones hay más de 500 millones de hispanohablantes en todo el mundo.

Como parte de la iniciativa de IBEREVAL 2017 (http://cabrillo.lsi.uned.es/nlp/IberEval-2017/index.php), hemos propuesto la tarea de Reconocimiento y Resolución de Abreviaturas Biomédicas (Biomedical Abbreviation Recognition and Resolution, BARR) con el objetivo de promover el desarrollo y la evaluación de sistemas de identificación de abreviaturas biomédicas.

Para dicha tarea, los equipos participantes deben detectar menciones de pares de SF y sus LF correspondientes a partir de los resúmenes (abstracts) de artículos médicos escritos en español. Los organizadores de BARR proporcionan un conjunto de entrenamiento etiquetado manualmente con parejas SF-LF (además de otras abreviaturas).

Esta tarea es particularmente interesante ya que algunos resúmenes fueron transcritos manualmente, y se parecen, por tanto, a las características de preprocesamiento que también se encuentran en los documentos clínicos. Además de los corpus BARR de referencia (Gold Standard) anotados manualmente, se publicará la colección de documentos BARR, que consistirá, hasta donde sabemos, en la mayor colección unificada de resúmenes de artículos médicos escritos en español distribuidos mediante un acuerdo especial con la editorial Elsevier para los equipos participantes.

Se pueden consultar detalles adicionales, muestras de ejemplo, preguntas frecuentes, y detalles de inscripción en la URL de la tarea BARR:

http://temu.inab.org/index.html

BARR track URL: http://temu.inab.org/index.html

e-mail de contacto: Martin Krallinger , mkrallinger@cnio.es

Fechas importantes

==========

- ~~**28 de Marzo de 2017:** Publicación de los datos de muestra~~
- ~~**19 de Mayo de 2017:** Publicación del primer subconjunto de los datos de entrenamiento~~
- **30 de Mayo de 2017:** Publicación de los datos de entrenamiento (conjunto completo)
- **19 de Junio de 2017:** Publicación de los datos de prueba (test set)
- **24 de Junio de 2017:** Envío de las participaciones
- **26 de Junio de 2017:** Evaluación de los resultados
- **1 de Julio de 2017:** Artículos de los participantes
- **19 de Septiembre de 2017:** Workshop SEPLN 2017

Organizadores de la tarea BARR

- Martin Krallinger, Biological Text Mining Unit (Bio-TeMUC), CNIO, Spain
- Santiago de la Peña, Biological Text Mining Unit (Bio-TeMUC), CNIO, Spain
- Ander Intxaurrondo, Biological Text Mining Unit (Bio-TeMUC), CNIO, Spain
- Jesús Santamaría, Biological Text Mining Unit (Bio-TeMUC), CNIO, Spain
- Jose A. Lopez-Martin, Medical Oncology, Hospital 12 de Octubre, Spain
- Alfonso Valencia, Life Sciences & Computational Biology, BSC, Spain
- Marta Villegas, Life Sciences & Computational Genomics, BSC, Spain
- Anália Lourenço, Next Generation Computer Systems Group, University of Vigo, Spain
- Gael Pérez, Next Generation Computer Systems Group, University of Vigo, Spain
- Martín Pérez, Next Generation Computer Systems Group, University of Vigo, Spain