Propuesta de algoritmo para extender y poblar ontologías

Jorge Cruanes*†, M. Teresa Romá-Ferri‡

†Departamento de Lenguajes y Sistemas Informáticos
‡Departamento de Enfermería
Universidad de Alicante, Apto. Correos 99, Alicante, España
jcruanes@dlsi.ua.es, mtr.ferri@.ua.es

1 Introduction

El objetivo principal de la tesis será el de diseñar y verificar un algoritmo capaz de, a partir de textos escritos en lenguaje natural y una ontología, extraer el conocimiento relevante y usarlo para ampliar dicha ontología. En este planteamiento la ontología usada será base y diana al mismo tiempo. Primero, la ontología filtrará el conocimiento contenido en los textos, etiquetándolo de forma semántica. Posteriormente la ontología se ampliará con nuevo conocimiento contenido en los mismos textos, extendiendo sus conceptos y poblándola con instancias (términos). La propuesta del método se divide en las siguientes etapas:

1. Procesar el documento mediante la herramienta de PLN FreeLing en tres fases. En la primera se usará EuroWordNet (en castellano) como base de conocimiento general para el etiquetado de los tokens del texto. En una segunda fase se usará una ontología de domino, en nuestro caso OntoFIS (dominio farmacoterapéutico). Esta última fase persigue el etiquetado semántico de los tokens conocidos. Por último se mezclan los resultados anteriores. Para ello, tendrá prioridad la etiquetación semántica obtenida usando la ontología de dominio como base de conocimiento.
2. Si un token sólo es identificado con conocimiento general entonces, mediante medidas de distancia semántica, habrá que conocer su equivalencia con los conceptos de la ontología domino. Se establecerán unos umbrales para considerar a un token como rechazado, candidato o equivalente a un concepto de la ontología.
3. Si un token es desconocido, comienza la etapa principal de nuestra propuesta, descubrir conocimiento. Esta etapa se basa en la utilización de patrones tipo tripletas (T1,R,T2), y donde T1 (sujeto) y T2 (predicado) serán conceptos o instancias, y R una relación. Para poder descubrir nuevo conocimiento, al menos uno de los elementos de la tripleta debe ser conocido, para así servir de referencia. En este punto tenemos previsto cuatro aproximaciones:

* Este artículo ha sido cofinanciado por el Ministerio de Ciencia e Innovación (proyecto TIN2009-13391-C04-01), y la Conselleria d'Educació de la Generalitat Valenciana (proyectos PROMETEO/2009/119, ACOMP/2010/286 y ACOMP/2011/001).

- Si conocemos T1 o T2 y R es taxonómica, el token desconocido se comprobará que es un sustantivo común para determinar, mediante el lugar en la tripleta, si es hiperónimo o hipónimo.
- Si conocemos T1 o T2 y R no es taxonómica, el token desconocido será etiquetado semánticamente el dominio y rango de R.
- Si sólo conocemos la relación, pero ésta tiene un único dominio y rango, entonces la clasificación semántica de los tokens desconocidos será acorde a su posición en la relación.
- Si sólo R es desconocida, tendrá como dominio la categoría semántica de T1 y como rango la categoría de T2.

Para refinar los patrones se usarán valores de umbrales para ser aceptados conforme se vaya incrementando los documentos procesados.

4. Finalmente se hará un control de calidad en dos fases. Primero se comprobará que la ontología resultante es válida, es decir, carece de bucles de herencia y no existen instancias pertenecientes a clases disjuntas. En segundo lugar, si existen conflictos serán resueltos de acuerdo a un algoritmo de toma de decisiones, volviendo a realizar el control de calidad.