Sistemas híbridos de extracción de información

Pablo Duboue.
Les Laboratoires Foulab, Montreal, Canada.

Resúmen:

En este curso estudiaremos pipelines de extracción de información utilizando métodos basados en reglas y diccionarios en combinación con métodos estadísticos (e.g., entropía máxima, Conditional Random Fields) basados en datos anotados. Se dará énfasis a técnicas con implementaciones libremente disponibles de código abierto (e.g., Apache UIMA y herramientas relacionadas). El curso analizará una pipeline de ejemplo disponible como código abierto para la extracción de información sobre contratos gubernamentales en la ciudad de Montreal, Canadá, escrita por el autor. Este curso requiere familiaridad con el lenguaje de programación Java para apreciar los ejemplos y realizar el trabajo de evaluación final.

Descripción:

Módulo 1
Extracción de información. Generalidades. Competiciones de extracción de información. Pipelines de procesamiento de lenguaje natural. Anotaciones off-stand. Sistemas de tipos. Control de flujo. Serialización de anotaciones.

Módulo 2
Entidades nombradas. Generalidades. Sistemas basados en diccionario. Sistemas basados en desambiguación de sentidos. Bootstrapping usando diccionarios de arranque. Destilado de diccionarios a partir de datos abiertos.

Módulo 3
Extracción de información basados en reglas. Métodos basados en expresiones regulares. JAPE. RefO. Métodos basados en puntos de anclaje. RuTA. Inducción de reglas a partir de ejemplos anotados. Algoritmo Whisk. Algoritmo LP2. Algoritmo KEP. Metodología de programación y evaluación.

Módulo 4
Extracción de información basada en métodos estadísticos. Metodología Begin-Inside-Outside. Sistemas basados en entropía máxima. Sistemas basados en Conditional Random Fields. Criterios de anotación y evaluación.

Módulo 5
Sistemas híbridos de extracción de información. Errores en cascada. Caso de estudio. Puesta en producción y mantenimiento. Cierre.

Evaluación:
El trabajo de evaluación será realizar una modificación sobre el pipeline de ejemplo. Los alumnos tendrán un mes para enviar el trabajo por correo electrónico.

Bibliografía:
Principal

1. Cunningham, Hamish. "Information extraction, automatic." Encyclopedia of language and linguistics, (2005): 665-677.https://gate.ac.uk/sale/ell2/ie/preprint.pdf
2. Documentacion de UIMA: http://uima.apache.org/d/uimaj-current/index.html
3. Documentacion Apache Ruta: https://uima.apache.org/d/ruta-current/tools.ruta.book.html

Opcional

Text Mining, Weiss et al. (2005), Chapter 6
* https://gate.ac.uk/sale/tao/splitch8.html
* http://iepy.readthedocs.org/en/latest/tutorial.html
* http://www.nltk.org/book/ch07.html
* http://mallet.cs.umass.edu/fst.php