Advanced Information Extraction

1 - Contextualization

Our society is a "document society" (Buckland 2013). "Documents have become the glue that enables societies to cohere. Documents have increasingly become the means for monitoring, influencing, and negotiating relationships with others" (Buckland 2013). With the advent of the web and other technologies the concept of document evolved to include from classical books and reports to complex online multimedia information incorporating hyperlinks.

The number of such documents and rate of increase are overwhelming. Some examples: Governments produce large amounts of documents at the several levels (local, central) and of many types (laws, regulations, minutes of meetings (public), etc); Information on companies' intranets is increasing; more and more exams, reports and other medical documents are stored in servers by health institutions. Our personal documents augment day by day in number and size. As such, health research is one of the most active areas, resulting in a steady flow of documents (e.g. medical journals and Master's and doctoral theses) reporting on new findings and results. There are also many portals and web sites with health information

Much of the information that would be of interest to citizens, researchers, and professionals is found in unstructured documents. Despite the increasing use of tables, images, graphs and movies, a relevant part of these documents adopts at least partially written natural language. The amount of contents available in natural language (English, Portuguese, Chinese, Spanish, etc.) increases every day. This is particularly noticeable in the web.

2 - Justification

Extracting information from natural language unstructured documents is becoming more and more relevant in our "document society". Despite the many useful applications that the information in these documents can potentiate, it is harder and harder to obtain the wanted information. Major problems result from the fact that much of the documents is in a format non usable by humans or machines. There is the need to create ways to extract relevant information from the vast amount of natural language sources.

Problems:

Despite the many useful applications that the information on these documents can potentiate, it is harder and harder to obtain the wanted information. This huge and increasing amount of documents available in the web, companies intranets and accumulated by most of us in our computers and online services potentiate many applications but also pose several challenges to make it really useful.

A major problem results from the fact that much of the documents/data is in a format non usable by humans or machines. Hence, there is the need to create ways to extract relevant information from the vast amount of natural language sources. Natural language is the most comprehensive tool for humans to encode knowledge (Santos 1992), but creating tools to decode this knowledge is far from simple.

The second problem that needs to be solved is how to represent and store the information extracted. One must also make this information usable by machines. Regarding the discovery of information, general search engines do not allow the end-user to obtain a clear and organized presentation of the available information. Instead, it is more or less of a hit or miss, random return of information on any given search. Efficient access to this information implies the development of semantic search systems (Guha et al. 2003) capable of taking in consideration the concepts and not the words.

Semantic search has some advantages over search that directly index text words (Teixeira et al. 2014):

(1) produces smaller sets of results, by being capable of identifying and removing duplicated or irrelevant results;

(2) can integrate related information scattered across documents;

(3) can produce relevant results even when the question and answer do not have common words; and

(4) makes possible complex and more natural queries.

To make possible semantic search and other applications based on semantic information, we need to add semantics to the documents or create semantic descriptions representing or summarizing the original documents. This semantic information must be derived from the documents and this can be done using techniques from Information Extraction (IE) and Natural Language Processing (NLP) fields. In general, to make IE possible, texts are first

pre-processed (ex: to separate into sentences and words) and enriched (ex: to mark words as nouns or verbs) by applying several NLP methods.

This course presents a generic architecture for developing systems that are able to learn how to extract relevant information from natural language documents, and assign semantic meaning to it. It also provides the background and means for students to implement a working system using, in most parts, state-of-the-art and freely available software. Concrete examples of systems/applications are used as case studies to illustrate how applications can deliver information to end. Students will have the opportunity to implement one application.

3 - Scientific Areas

Main areas: Information Extraction and Natural Language Processing.

Other areas: knowledge representation, ontologies, HCI, resources for NLP, tools for NLP, markup languages.

4 - Related courses

Example of courses related to our proposal:

- CS 224N / Ling 284 — **Natural Language Processing**, Stanford Univ., for advanced undergraduate/beginning graduate (http://web.stanford.edu/class/cs224n/syllabus.shtml)

  Including most of the topics selected in NLP for our course but without a clear presentation of the Information pipeline and applications.

- Info 256. **Applied Natural Language Processing,** Berkeley (http://www.ischool.berkeley.edu/courses/i256)

  Course, at least in the edition having public information covered several topics integrating the proposed course (POS tagging, parsing) but the applications go beyond Information Extraction, including summarization, text classification, clustering. It includes as topics "Finding Semantic Relations" and "Text Mining", with contact points with IE.

- COM6513 Natural Language Processing, Univ. Sheffield (http://www.dcs.shef.ac.uk/intranet/teaching/public/modules/msc/com6513.htm)

Course provides an overview of the field of NLP and its sub-areas, and introduces and explain its key techniques, including their applicability and limitations. In lab classes, students will practice implementing the NLP techniques taught in class, testing their code in application to real language data. Topics covered include:

- Tokenisation, Morphology and Finite State Automata
- N-gram Language Modelling
- Word Classes and Part-of-Speech Tagging
- Lexical Semantics, Word Sense Disambiguation and Lexical Similarity
- Syntactic Theory/The Grammar of English
- Parsing: Chart Parsing, PSG and Feature Representations
- Compositional Semantics

It is essentially a course addressing the NLP part of the proposed course, without the Information Extraction part. NLP is addressed in more detail that we aim for our course.

- Advances in Information Extraction: From Text and Image to Knowledge, Technical University of Kaiserslautern, Computer Science, Course: INF-71-60-V-6 (http://www.dfki.de/~sonntag/courses/SS14/IE.html)

  Major Topics

  - Interactive intelligent systems and multimedia information extraction
  - Overview of several text-based IE tasks including named entity recognition, co-reference resolution, relation extraction as well as image extraction
  - Linguistic (Pre)Processing (NLP)
  - Dependency parsing (NLP)
  - Information Extraction from Biomedical Texts and Images
  - Open Information Extraction at Web Scale
  - Automated Question Answering
  - Machine Learning in IE: integrating clustering and classification, precision/recall, ROC, ANOVA
  - Multimedia Information Extraction
  - Multimedia Knowledge Capture in Ontologies and the Semantic Web
  - Social Multimedia Analysis and Opinion Mining
  - Applications and Projects

This includes additional topics (ex: Multimedia Information Extraction and Automated Question Answering) but, in general, is the more aligned with the proposed course.

There is also a complete PhD program at CMU (Ph.D. in Language and Information Technology) directly related to the topics covered in the proposed course.

5 - Objectives and Learning Outcomes

This course is intended to provide a practical yet state-of-the-art experience in information extraction, with emphasis on the current state of the art tools and technologies.

Course objectives:

- Explanation of the main concepts and technologies of information extraction at large;
- Experiment and analyze existing applications of information extraction
- Demonstrate and encourage students to combine several modules that compose a working system
- Understand the current challenges of concrete designs and solutions
- Knowledge of state of the art, allowing a critical attitude about the possibility of using information extraction technologies in concrete tasks
- Design and implement information extraction applications

6 - Detailed Program

1. Introduction
    a. Motivation
    b. Examples of state-of-the-art systems and tools
    c. Overview of Information Extraction
    d. Demos
    e. Tutorial Example (using, for example, Stanford NLP suite) [sec 5.1 of Rodrigues and Teixeira 2015 book]
    f. Presentation of course modules and teachers
2. Background information
    a. Text and Document Processing
        i. Regular Expressions
        ii. Markup languages and tools (XML, XSLT, schemas, DTDs etc)
    b. Semantics and Knowledge Representation Basics
        i. Semantics
        ii. Taxonomies, Thesauri and Ontologies
        iii. WordNet
        iv. Reasoning basics
        v. OWL, Triple Stores, SPARQL
    c. Natural Language Processing

   i. Processing levels
   ii. Typical NLP pipeline
   iii. NLP common tasks
      1. Segmentation and tokenization
      2. Morphological Analysis and tagging
      3. (Syntactic) Parsing
      4. Named Entities Recognition

3. Information Extraction (IE) - An Overview
   a. Main approaches
   b. Performance metrics
   c. Challenges
   d. General architecture and pipeline,
      i. Process overview

4. Data Gathering, Preparation and Enrichment
   a. Objectives
   b. Process overview
   c. Tools
      i. Tokenizers
      ii. Sentence boundary detectors
      iii. Morphological analysers and POS taggers
      iv. Syntactic parsers
   d. Representative software suites

5. Identifying Things and Relations
   a. Identifying Things/Entities (Who, Where and When)
   b. Identifying relations
   c. Information fusion

6. Ontology-based Information Extraction (OBIE)
   a. Basic idea
   b. Types
   c. Architecture
   d. Case study
      i. OBIE system for eGov, developed by Mário Rodrigues
   e. OBIE to the limit: Open Information Extraction

7. Systems and Applications
   a. Case studies
      i. System developed at DETI/IEETA, Univ. Aveiro
         1. IE Applied to Health (MedInx and HealthInX)
         2. IE applied to eGov
      ii. Other systems (selected each year from state-of-the-art)

7- Teaching Methods

The teaching method consists of theoretical-practical classes. A total of 20 hours is planned to present the several parts of the program. Whenever possible, presentations of examples and demonstrations will be performed, to complement the classical exposition, more or less supported in PPTs.
At least 8 hours will be allocated to students presentations of their work and discussion with teachers and colleagues.

8 - Assessment method

The assessment method adopted for the course is based in projects. Students create a concrete application of Information Extraction based in rich NLP processing in two stages. First they will define and develop a complete pipeline to extract information and evaluate it. The pipeline and evaluation results will be presented in class for evaluation, supported by a PPT. Second step will consist in the creation of the complete application, with user inputs and visualization/transmission of the results to the end user. A second public presentation will be made including a demo, for evaluation. The third information for evaluation will be a short written report regarding both application and pipeline.

9 - Bibliography

1. **Mário Rodrigues**, **António Teixeira**. *Advanced Applications of Natural Language Processing for Performing Information Extraction*. Springer, 2015

2. Ingersoll, G. S., Morton, T. S., Farris, A. L., *Taming Text: How to Find, Organize, and Manipulate It*, Manning Publications, 2013

3. Mitkov, R. (2005). *The Oxford handbook of computational linguistics*. Oxford University Press.

4. **Mário Rodrigues**. Model of Access to Natural Language Sources in Electronic Government. PhD Thesis, DETI, Universidade de Aveiro, December 2013 [Supervised by **António Teixeira** and Gonçalo Paiva Dias]

5. Liliana da Silva Ferreira. Medical Information Extraction in European Portuguese / Extracção de Informação de Relatórios Médicos em Português Europeu. PhD Thesis, Universidade de Aveiro, July 2011 [Supervised by **António Teixeira** and João Paulo Cunha].

6. Amy Neustein, S. Sagar Imambi, **António Teixeira**, Liliana da Silva Ferreira, **Mário Rodrigues**. Application of Text Mining to Biomedical Knowledge Extraction: Analyzing Clinical Narratives and Medical Literature. Text Mining of Web-based Medical Content, Amy Neustein (Ed.), De Gruyter, September 2014

7. **António Teixeira**, Liliana da Silva Ferreira, **Mário Rodrigues**. Health Information Semantic Search and Exploration: Reporting on Two Prototypes for Performing Information Extraction on both Intranet and Internet-based Medical Content. Text Mining of Web-based Medical Content, Amy Neustein (Ed.), De Gruyter, September 2014

8. A. Neustein, ed. Text Mining of Web-based Medical Content. De Gruyter, 2014.

References cited

Buckland, M., 2013. The Quality of Information in the Web. *BiD: textos universitaris de biblioteconomia i documentació*, (31).

Guha, R., McCool, R. & Miller, E., 2003. Semantic Search. In *The Twelfth International World Wide Web Conference (WWW)*. Budapest, Hungary, p. 779.

Santos, D., 1992. Natural Language and Knowledge Representation. In *Proceedings of the ERCIM Workshop on Theoretical and Experimental Aspects of Knowledge Representation*. pp. 195–197.

Teixeira, A., Ferreira, L. & Rodrigues, M., 2014. Online health information semantic search and exploration : reporting on two prototypes for performing extraction on both a hospital intranet and the world wide web. In A. Neustein, ed. *Text Mining of Web-based Medical Content*. De Gruyter, pp. 49–73.