Your dream corpus is already there

Stasinos Konstantopoulos
NCSR 'Demokritos', Athens, Greece

The tutorial will provide the attendee with an in-depth look at how to tap the potential of web data as a source of linguistic resources that go beyond frequency counts. Special focus will be given to linguistic resources that can be compiled from web data that was constructed for entirely different purposes. The tutorial will combine the theoretic analysis with the discussion of concrete examples, the problems that had to be overcome, and the solutions provided.

The tutorial is structured in four sessions.

1 Introduction

This session presents and explains approaches where the structure of the corpus is equivalent to that of the data it is constructed from. These include using the web as a text repository paying no attention to web structure. Examples include using the web to calculate frequency counts or to compile collections of definitions.

Another large class of these types of corpus-construction efforts include retrieving collections that result from manual equivalents of a language technology task, such as using manual translations to build machine translation systems or manual summaries to build summarization systems. There are many examples, covering many tasks and domains from creating parallel corpora from multi-lingual web pages or the European Parliament proceedings, to using academic paper abstracts for training summarization systems, to tuning personalized NLG from manually authored natural language renderings of structured weather forecasts or medical records.

2 Data

This session of the tutorial concentrates on identifying data sets for creating linguistic resources from, and especially data sets with radically different purpose or domain than that of the resulting resource. After discussing the kinds of things generally found in large lists or semi-structured text on the web, relevant literature and examples are presented and discussed. These include using title/abstract collections of academic papers to extract important terminology for a scientific domain; using websites about football transfers or national selection games to extract lists of names and nationalities; or using geographic databases listing placenames in multiple languages or baby-names web sites offering variants of the same name in different languages to extract of lists of cognates. This session emphasises the imaginative use of data sets and transformations and, in

particular, approaches that cross domains or build general-purpose tools from specialized data sets.

3 Toolkit

In this session selected machine learning methodologies are briefly presented from the perspective of robustness to dirty or incomplete corpora, such as the ones compiled using the ideas presented in the tutorial. Machine learning background is beneficial but not required, as the tutorial will only superficially explain which features of each methodology make it appropriate for the kinds of tasks discussed here, without delving into technical details.

4 Hands-on

In the closing session participants are urged to identify web resources that can be turned into a new linguistic resource. Participants' ideas are discussed in the group, with the discussion driven around how much effort will the proposed idea involve in scripting and post-editing, the quality and size of the resulting corpus, and the machine learning methodologies that are best suited from creating a given linguistic tool or other resource from such a corpus.