Università degli Studi di Salerno

Dipartimento di Scienze Politiche, Sociali e della Comunicazione

Dottorato in Scienze della Comunicazione, Sociologia,
Teorie e Storia delle Istituzioni, Ricerca Educativa,
Corporeità didattiche, Tecnologie e Inclusione

A Hybrid Framework for Text Analysis

Alessandro Maisto

Supervisor and Tutor:
Prof. Annibale Elia

A.A. 2016/2017

Abstract

In Computational Linguistics there is an essential dichotomy between Linguists and Computer Scientists. The first ones, with a strong knowledge of language structures, have not engineering skills. The second ones, contrariwise, expert in computer and mathematics skills, do not assign values to basic mechanisms and structures of language. Moreover, this discrepancy, especially in the last decades, has increased due to the growth of computational resources and to the gradual computerization of the world; the use of Machine Learning technologies in Artificial Intelligence problems solving, which allows for example the machines to "learn", starting from manually generated examples, has been more and more often used in Computational Linguistics in order to overcome the obstacle represented by language structures and its formal representation.
The dichotomy has resulted in the birth of two main approaches to Computational Linguistics that respectively prefers:

- rule-based methods, that try to imitate the way in which man uses and understands the language, reproducing syntactic structures on which the understanding process is based on, building lexical resources as electronic dictionaries, taxonomies or ontologies;

- statistic-based methods that, conversely, treat language as a group of elements, quantifying words in a mathematical way and trying to extract information without identifying syntactic structures or, in some algorithms, trying to confer to the machine the ability to learn these structures.

One of the main problems is the lack of communication between these two different approaches, due to substantial differences characterizing them: on the one hand there is a strong focus on how language works and on language characteristics, there is a tendency to analytical and manual work. From other hand, engineering perspective finds in language an obstacle, and recognizes in the algorithms the fastest way to overcome this problem.
However, the lack of communication is not only an incompatibility: following Harris, the best way to approach natural language, could result by taking the best of both.

At the moment, there is a large number of open-source tools that perform text analysis and Natural Language Processing. A great part of these tools are based on statistical models and consist on separated modules which could be combined in order to create a pipeline for the processing of the text. Many of

these resources consist in code packages which have not a GUI (Graphical User Interface) and they result impossible to use for users without programming skills. Furthermore, the vast majority of these open-source tools support only English language and, when Italian language is included, the performances of the tools decrease significantly. On the other hand, open source tools for Italian language are very few.

In this work we want to fill this gap by present a new hybrid framework for the analysis of Italian texts. It must not be intended as a commercial tool, but the purpose for which it was built is to help linguists and other scholars to perform rapid text analysis and to produce linguistic data. The framework, that performs both statistical and rule-based analysis, is called *LG-Starship*. The idea is to built a modular software that includes, in the beginning, the basic algorithms to perform different kind of analysis. Modules will perform the following tasks:

- Preprocessing Module: a module with which it is possible to charge a text, normalize it or delete stop-words. As output, the module presents the list of tokens and letters which compose the texts with respective occurrences count and the processed text.

- Mr. Ling Module: a module with which POS tagging and Lemmatization are performed. The module also returns the table of lemmas with the count of occurrences and the table with the quantification of grammatical tags.

- Statistic Module: with which it is possible to calculate Term Frequency and TF-IDF of tokens or lemmas, extract bi-grams and tri-grams units and export results as tables.

- Semantic Module: which use The Hyperspace Analogue to Language algorithm to calculate semantic similarity between words. The module returns similarity matrices of words per word which can be exported and analyzed.

- Syntactic Module: which analyze syntax structures of a selected sentence and tag the verbs and its arguments with semantic labels.

The objective of the Framework is to build an “all-in-one” platform for NLP which allows any kind of users to perform basic and advanced text analysis. With the purpose of make the Framework accessible to users who have not specific computer science and programming language skills, the modules have been provided with an intuitive GUI.

The framework can be considered "hybrid" in a double sense: as explained in the previous lines, it uses both statistical and rule/based methods, by relying on standard statistical algorithms or techniques, and, at the same time, on Lexicon-Grammar syntactic theory. In addition, it has been written in both Java and Python programming languages. LG-Starship Framework has a simple Graphic User Interface but will be also released as separated modules which may be included in any NLP pipelines independently.
There are many resources of this kind, but the large majority works for English. There are very few free resources for Italian language and this work tries to cover this need by proposing a tool which can be used both by linguists or other scientist interested in language and text analysis who have no idea about programming languages, as by computer scientists, who can use free modules in their own code or in combination with different NLP algorithms.

The Framework takes the start from a text or corpus written directly by the user or charged from an external resource. The LG-Starship Framework workflow is described in the flowchart shown in fig. 1.

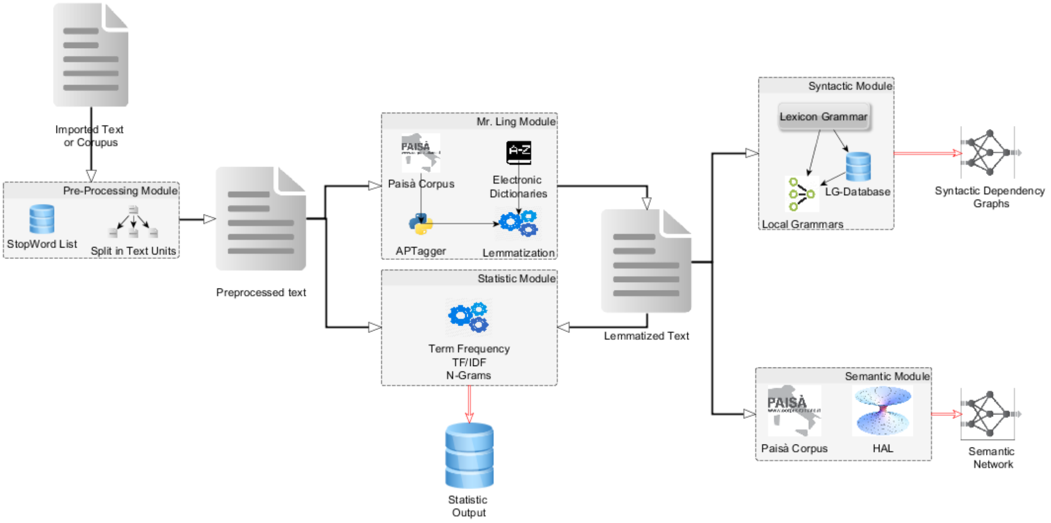

Figure 1: Workflow of the LG-Starship Framework.

The pipeline shows that the Pre-Processing Module is applied on original imported or generated text in order to produce a clean and normalized preprocessed text. This module includes a function for text splitting, a stop-word list and a tokenization method. On the text preprocessed the Statistic Module or the Mr. Ling Module can be applied. The first one, which includes basic statistics algorithm as Term Frequency, tf-idf and n-grams extraction, produces as output databases of lexical and numerical data which can be used to produce charts or perform more external analysis; the second one, is divided in two main task: a Pos tagger, based on the Averaged Perceptron Tagger [?] and trained on the "Paisà Corpus" [Lyding et al., 2014], perform the Part-Of-Speech Tagging and produce an annotated text. A lemmatization method, which relies on a set of electronic dictionaries developed at the University of Salerno [Elia, 1995, Elia et al., 2010], take as input the Postagged text and produces a new lemmatized version of original text with information about syntactic and semantic properties.

This lemmatized text, which can also be processed with the Statistic Module, serves as input for two deeper level of text analysis carried out by both the Syntactic Module and the Semantic Module.
The first one lays on the Lexicon Grammar Theory [Gross, 1971, 1975] and use a database of Predicate structures in development at the Department of Political, Social and Communication Science. Its objective is to produce a Dependency Graph of the sentences that compose the text.
The Semantic Module uses the Hyperspace Analogue to Language distributional semantics algorithm [Lund and Burgess, 1996] trained on the "Paisà Corpus" to produce a semantic network of the words of the text.

These workflow has been included in two different experiments in which two User Generated Corpora have been involved.

The first experiment represent a statistical study of the language of Rap Music in Italy through the analysis of a great corpus of Rap Song lyrics downloaded from on line databases of user generated lyrics.
The second experiment is a Feature-Based Sentiment Analysis project performed on user product reviews. For this project we integrated a large domain database of linguistic resources for Sentiment Analysis, developed in the past years by the Department of Political, Social and Communication Science of the University of Salerno, which consists of polarized dictionaries of Verbs, Adjectives, Adverbs and Nouns.
These two experiment underline how the linguistic framework can be applied to different level of analysis and to produce both Qualitative data and Quantitative data.

For what concern the obtained results, the Framework, which is only at a Beta Version, obtain discrete results both in terms of processing time that in terms of precision. Nevertheless, the work is far from being considered complete. More algorithms will be added to the Statistic Module and the Syntactic Module will be completed. The GUI will be improved and made more attractive and modern and, in addiction, an open-source on-line version of the modules will be published.

Bibliography

A. Elia. Dizionari elettronici e applicazioni informatiche. In *JADT*, 1995. (Cited on page 5.)

A. Elia, F. Marano, M. Monteleone, S. Sabatino, and D. Vellutino. Strutture lessicali delle informazioni comunitarie all'interno di domini specialistici. In *Statistical Analysis of Textual Data, Proceedings of 10th International Conferences" Journées D'Analyse Statistique des Données Textuelles" Roma, Università" La Sapienza*, pages 9–11, 2010. (Cited on page 5.)

M. Gross. *Transformational Analysis of French Verbal Constructions.* University of Pennsylvania, 1971. (Cited on page 5.)

M. Gross. *Méthodes en syntaxe.* Hermann, 1975. (Cited on page 5.)

K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996. (Cited on page 5.)

V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, and V. Pirrelli. The paisa corpus of italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, 2014. (Cited on page 5.)