Linguagrid: a network of Linguistic and Semantic Services for the Italian Language.

Alessio Bosca, Luca Dini, Milen Kouylekov, Marco Trevisan
CELI Research
Turin, Italy
E-mail: {bosca, dini, kouylekov, trevisan}@celi.it

Abstract

In order to handle the increasing amount of textual information today available on the web and exploit the knowledge latent in this mass of unstructured data, a wide variety of linguistic knowledge and resources *(Language Identification, Morphological Analysis, Entity Extraction, etc.*). is crucial. In the last decade LRaas (Language Resource as a Service) emerged as a novel paradigm for publishing and sharing these heterogeneous software resources over the Web. In this paper we present an overview of Linguagrid, a recent initiative that implements an open network of linguistic and semantic Web Services for the Italian language, as well as a new approach for enabling customizable corpus-based linguistic services on Linguagrid LRaaS infrastructure. A corpus ingestion service in fact allows users to upload corpora of documents and to generate classification/clustering models tailored to their needs by means of standard machine learning techniques applied to the textual contents and metadata from the corpora. The models so generated can then be accessed through proper Web Services and exploited to process and classify new textual contents.

Keywords: Language Resource as a Service, Corpus Linguistic, Machine Learning

1. Introduction

In order to handle the increasing amount of textual information today available on the web and fully exploit the knowledge latent in this unstructured data a wide variety of linguistic knowledge and resources is crucial for a large set of applicative domains (contents management system, question answering, sentiment analysis, machine translation, etc.). Such linguistic knowledge includes (but is not limited to) morphological analysis, named entities recognition, co-reference resolution (e.g., aliases, synonyms, paraphrases), semantic class labelling, word sense disambiguation and so on.

The need for integrating such a broad spectrum of different linguistic knowledge and resources into a single application inevitably raises technical issues resulting from the complexity of the needed tools: harmonizing different programming languages and software environments, managing different types of licences, tailoring the hardware resources to the data to be processed, etc. However, in the last decade the paradigm of SaaS (Software-as-a-Service) emerged as a general and effective solution to this problem by providing methodologies and instruments for exposing and integrating heterogeneous software resources over the web, (McIlraith et al., 2001; Ferris & Farrel, 2003). In parallel with the development of this technological paradigm, significant efforts in LT research community (Schmidt & Wolff, 2003) have been invested on the topic of service-oriented language infrastructures, thus originating the notion of LRaaS (Language Resource as a Service) as well as concrete implementations of multilingual service platforms that enable registration and sharing of language services over the web. The Language Grid project (Ishida, 2006), developed by the Japanese National Institute of Information and Communications Technology (NICT), represents one of the most important initiative in this field.

In this paper we present an overview of Linguagrid[1], an initiative for the Italian language exploiting the most recent advancements and solutions in the domain of LRaaS. Linguagrid initiative started in 2010 within the context of ICT4Law[2] and GALATEAS[3] research projects and implements an open network of linguistic and semantic Web Services for the Italian language, including: *Morphological Analysis, Dependency Parsing, Keyword Extraction, Entity Recognition and Language Identification*.

Linguagrid operates on the top of the service-oriented infrastructure developed by NICT (under a research agreement) and aims at becoming the first Italian node, federated to the the Language Grid[4] initiative. Linguagrid resources are open to different operators (Universities, Research institutes, Companies) with configurable services access policies: free, restricted to registered users, research or commercial licensing.

In the paper besides presenting an overview of the language services currently available in Linguagrid, we propose a novel approach for enabling customizable corpus-based linguistic services on a standard LRaaS infrastructure. A corpus ingestion service (implemented as a web application) allows users to upload corpora of documents and to generate classification/clustering models tailored to their needs. Standard machine learning techniques applied to the textual contents and metadata from the corpora in order to train customized models that can successively be accessed through proper Web

[1] http://linguagrid.org
[2] http://www.ict4law.org
[3] http://www.galateas.eu
[4] http://langrid.org

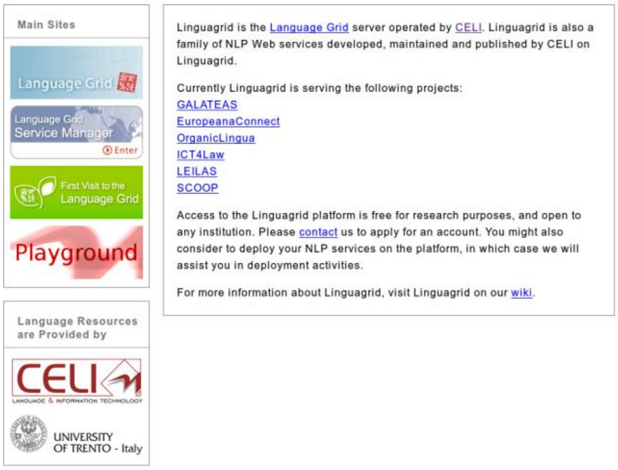
Services and exploited for processing and classifying new textual contents.

Linguagrid initiative promotes the aggregation of resources, the collaboration between entities (i.e. service composition) and aims at simplifying/standardizing the use of NLP services by third parties both for research and commercial purposes. The vision behind such initiative, actually goes beyond service provision for specific projects and aims to connect to European initiative such as FlaReNet (Calzolari, 2008) and CLARIN (Tamás, 2008).

The paper is organized as follows: in section 2 we presents an overview of the services currently exposed by Linguagrid and how to obtain access to them. In section 3 we focus on the Corpus Ingestion Service and on the approach used for enabling corpus based services, while in section 4 we conclude the paper by providing a practical road map for the future steps and challenges.

2. Linguagrid Services Overview

The Linguagrid portal includes a Service Management facility (the standard interface provided by Langrid) and a repository with software resources and documentation related to the services exposed. The Service Management tool allows the authorized service providers to register and share their linguistic resources and tools as well as to define the users accounts details and the policies regulating their access to the services. A complementary wiki resource [5] provides additional details (documentation, usage examples, standard software clients, etc.) on the exposed resources.


Main Sites
Language Grid
Language Grid Service Manager
Enter
First Visit to the Language Grid
Playground
Language Resources are Provided by
CELI
LANGUAGE & INFORMATION TECHNOLOGY
UNIVERSITY OF TRENTO - Italy

Linguagrid is the Language Grid server operated by CELI. Linguagrid is also a family of NLP Web services developed, maintained and published by CELI on Linguagrid.

Currently Linguagrid is serving the following projects:
GALATEAS
EuropeanaConnect
OrganicLingua
ICT4Law
LEILAS
SCOOP

Access to the Linguagrid platform is free for research purposes, and open to any institution. Please contact us to apply for an account. You might also consider to deploy your NLP services on the platform, in which case we will assist you in deployment activities.

For more information about Linguagrid, visit Linguagrid on our wiki.

Figure 1: Linguagrid Homepage

The linguistic and semantic services currently available on Linguagrid originate from the software solutions developed by CELI in the progress of different EU research projects (CACAO[6], Organic.Lingua[7] , Galateas) and include:

- **Language Identification**: it offers language identification for *English, French, Italian, Spanish, Portuguese, Dutch, Polish, Hungarian, Swedish and German* languages. The service analyses a chunk of text and produces as output a ranked list of languages and each guess is provided with a confidence value. Different strategies are combined by the service in order to increase the precision and robustness of the service (N-gram character, Corpus frequency, Function Words).

- **Named Entity Recognizer**: is a software module that identifies named entities, such as person names, geographic names, organization names, etc. It is currently available for *Italian* and *French* languages.

- **Morphological Analysis**: is a software module that performs tokenization and lemmatization, but also de-compounding, multi-word detection and part of speech tagging. A morphological analyzer provides as output a lemma plus Part-Of-Speech information (NOUN, ADJ, VERB, ...) and morphological features (tense, number, gender, etc.). It is currently available for *Romanian, Italian German, Danish, Polish and Russian* languages.

- **Dependency Parsing**: it provides an interface to a robust dependency parser for *Italian* and *French* and produce a syntactic analysis of the input text in the well-known CONLL format (in XML format)

- **Dictionary Lookup Service**: consists in a dictionary based translation of terms. The dictionaries contains details about the terms POS as well as an optional domain category (i.e. art, history, etc). If no direct translation is available between the source and target languages the system performs a translation via a default bridge language. The language pairs currently supported include all the combinations from *Italian, French, German, Spanish, English, Portuguese, Polish, Hungarian, Dutch, Swedish*.

- **Sentiment Analysis**: it exposes functionalities for advanced NLP analysis of sentences, with the identification of textual snippets containing opinions and sentiments. It produces a list of the sentiment relations detected in the text along with the target and the polarity of the opinion detected (positive, negative, neutral). It is currently available for the *Italian* and *French* languages.

[5] http://research.celi.it/JSPWiki/Wiki.jsp?page=Linguagrid
[6] http://www.cacaoproject.eu/
[7] http://www.organic-lingua.eu/

- **Word Similarity Service**: it is a corpus-based service that allows to measure the distance between pairs of words. The distance between two words corresponds to the similarity of their contexts in a reference corpus. Words sharing similar contexts (i.e. surrounding terms) are closer than words that appear in dissimilar contexts. The distance is computed using the Random Indexing algorithm, an approximation technique similar to LSA (Karlgren & Sahlgren, 2001). The system is composed by two different modules: a web service for the generation of the semantic models and one for computing the semantic distance between terms given a specific model.

- **Classification service**: allows its users to train a classification model using an annotated corpus loaded in the corpora manager. The service allows to classify documents using such models. The system is composed of two separate Web Services: one for the creation and destruction of classification models and another one for the classification of new documents. The algorithm used for training the models is Naive Bayes (Friedman & Kohavi, 2002)

- **LDA Clustering service**: it is a corpus-based service that exposes clustering functionalities over a collection of documents, using the Latent Dirichelet Allocation algorithm (Blei et al., 2003). The system is composed of two different modules, a web service for the creation and destruction of LDA Topic Models and a service for accessing an available model in order to associate a new textual input to one of the detected clusters.

3. Corpus Based Services

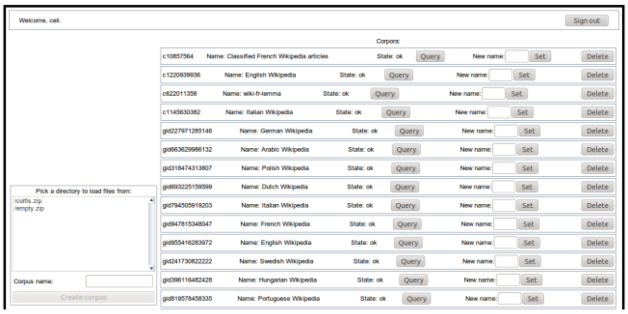
The latter three services (*Word Similarity, Classification, LDA Clustering*) adopt a different interaction model with respect to the previous ones. In fact the linguistic knowledge they expose is not “universal” or language specific, but rather corpus based and dependant on the specific linguistic features (i.e. terms frequencies distribution) of a given domain of interests, represented by the corpus of reference. Corpus linguistics, in fact, is the study of language as expressed in "real world" text, stored in corpora.

In the proposed approach the corpus-based services are composed of two separated modules: one for the generation of the model starting from the reference corpus and one for applying the so generated model to a given textual input; consequently different models can be generated from different corpora and different models will yields different results if applied to the same input. Each model is associated to the account of the user that generated it and cannot be accessed by other user.

These linguistic services are therefore strictly dependant on the corpus used for tailoring the service behaviour to a specific domain of interest; thus a preliminary condition in order to expose them as web services consists of a specific resource for uploading and managing the corpora. In Linguagrid such a specific resource is the Corpus Ingestion service.

3.1 Corpus Ingestion Service

The Corpus Ingestion service [8]can be accessed by a web interface and allows authorized users to upload XML files (properly formatted) and then generate a web accessible corpus from them. The service requires the textual files (used for generating the corpus) to be expressed using the minimal subset of the Text Encoding Initiative (TEI) known as TEI Bare[9] ; TEI is an XML Schema devoted to the mark-up of literary and linguistic texts and is a well known standard for the encoding of electronic texts in the humanities academic community.



c10857564	Name: Classified French Wikipedia articles	State: ok	Query	New name:	Set	Delete
c1220939936	Name: English Wikipedia	State: ok	Query	New name:	Set	Delete
c622011358	Name: wiki-fr-lemma	State: ok	Query	New name:	Set	Delete
c1145630382	Name: Italian Wikipedia	State: ok	Query	New name:	Set	Delete
gid227971285146	Name: German Wikipedia	State: ok	Query	New name:	Set	Delete
gid663629986132	Name: Arabic Wikipedia	State: ok	Query	New name:	Set	Delete
gid318474313807	Name: Polish Wikipedia	State: ok	Query	New name:	Set	Delete
gid693225159599	Name: Dutch Wikipedia	State: ok	Query	New name:	Set	Delete
gid794505919203	Name: Italian Wikipedia	State: ok	Query	New name:	Set	Delete
gid947815348047	Name: French Wikipedia	State: ok	Query	New name:	Set	Delete
gid955416283972	Name: English Wikipedia	State: ok	Query	New name:	Set	Delete
gid241730822222	Name: Swedish Wikipedia	State: ok	Query	New name:	Set	Delete
gid396116482428	Name: Hungarian Wikipedia	State: ok	Query	New name:	Set	Delete
gid819578458335	Name: Portuguese Wikipedia	State: ok	Query	New name:	Set	Delete

Figure 2: Corpus Ingestion Web Application

Each corpus generated by the Corpus Ingestion service is then associated with an univocal ID that is used as a reference for the corpus during the generation of/access to the model. By means of the Corpus Ingestion web application authorized users are allowed to browse their private corpora and delete, rename or explore them.

In fact, the corpora generated by the service are exposed on the web by means of SOLR[10] , the well known open source Search Engine from the Apache project. Exposing the corpora via a standard Search Engine allows for a further, additional benefit: the definition of dynamic corpora.

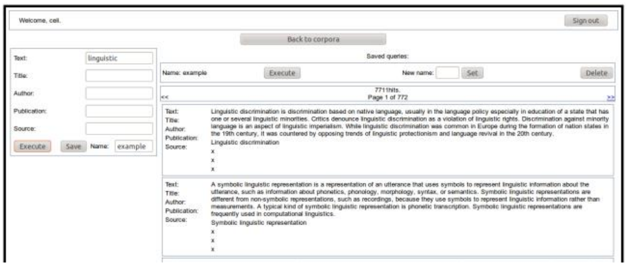
3.1 Dynamic Corpora Generation

The Dynamic Corpora are generated using filter queries specified by users; the Corpus Ingestion service in fact expose a simple form for querying an uploaded corpus and allows to save queries associated to a corpora in order to reuse them.

A given search query can then be used for filtering the documents from the original corpus and the results set can be used as a “dynamic” corpus in the phase of model generation by the corpus-based services, thus allowing to

[8] http://research.celi.it:8095/linguagrid.corpora.wa
[9] http://www.tei-c.org/Vault/Bare/
[10] http://lucene.apache.org/solr/

reuse an existing corpus by focusing it on a specific sub domain.



Text: linguistic
Title:
Author:
Publication:
Source:
Name: example

Name: example

7711hits.
Page 1 of 772

Text: Linguistic discrimination is discrimination based on native language, usually in the language policy especially in education of a state that has one or several linguistic minorities. Critics denounce linguistic discrimination as a violation of linguistic rights. Discrimination against minority language is an aspect of linguistic imperialism. While linguistic discrimination was common in Europe during the formation of nation states in the 19th century, it was countered by opposing trends of linguistic protectionism and language revival in the 20th century.
Title: Linguistic discrimination
Author: x
Publication: x
Source: x

Text: A symbolic linguistic representation is a representation of an utterance that uses symbols to represent linguistic information about the utterance, such as information about phonetics, phonology, morphology, syntax, or semantics. Symbolic linguistic representations are different from non-symbolic representations, such as recordings, because they use symbols to represent linguistic information rather than measurements. A typical kind of symbolic linguistic representation is phonetic transcription. Symbolic linguistic representations are frequently used in computational linguistics.
Title: Symbolic linguistic representation
Author: x
Publication: x
Source: x

Figure 3: Dynamic Corpora Generation

The proposed approach allows to fully exploit the customization capabilities of corpus-based services and the generation of customized models (for classification or clustering purposes) tailored to the specific needs of the users.

4. Conclusions

In this short paper we provided an overview of the linguistic and semantic services publicly available on Linguagrid with the focus on a novel approach for generating custom, corpus-based, classification models.

We intend to direct our future efforts towards two different goals. On one hand we plan to increase the functionalities offered by the corpus ingestion service, by increasing the set of the classification/clustering algorithms available for the generation of models as well as the textual formats supported for the corpora upload.

On the other hand, we plan to facilitate the integration of our linguistic resources into specific system for text processing/enrichment by developing specific client for those system. In particular we are currently evaluating the adoption of Stanbol[11], a novel project from the Apache Foundation for semantic content management.

5. Acknowledgments

This work has been supported and founded by Organic.Lingua and Galateas EU projects.

6. References

Calzolari Nicoletta (2008) *New European Infrastructural and Networking Initiatives.* In C. Delogu (ed.), LangTech 2008, Rome, p.19;

Váradi Tamás, Steven Krauwer, Peter Wittenburg, Martin Wynne and Kimmo Koskenniemi, (2008*) CLARIN: Common Language Resources and Technology Infrastructure*. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)

Ferris, Ch.; Farrell, J. (2003). *What are Web Services?* In: Communications of ACM 46(6) (2003), p. 31.

McIlraith, S.A.; Son, T.C.; Honglei Zeng; Knowledge Syst. Lab., Stanford Univ. (2001) *Semantic Web services*. IEEE Intelligent Systems (vol 16)

Schmidt, F.; Wolff, Ch. (2003) *Linguistic Knowledge Services – Developing Web Services in Language Technology*. Innovative Internet Community Systems. Lecture Notes in Computer Science, 2003, Volume 2877

Toru Ishida. (2006) *Language Grid: An Infrastructure for Intercultural Collaboration*. In IEEE/IPSJ Symposium on Applications and the Internet (2006)

Karlgren, J., & Sahlgren, M. (2001). *From words to understanding*. In Foundations of Real-World Intelligence, pp. 294–308.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003) *Latent dirichlet allocation*. Journal of Machine Learning Resources, pp. 993-1022.

Nir Friedman and Ron Kohavi (2002) *Bayesian classification*, in Handbook of Data Mining and Knowledge Discovery, pp 282-288.

[11] http://incubator.apache.org/stanbol/