Professional paper

A ReSTful Web Service for Multilingual LRT

Kamel Nebhi
LATL - University of Geneva
2, rue de Candole CH-1211 Genève 4, Switzerland
kamel.nebhi@unige.ch

Summary

Today, almost all language tools and resources are available by download only. Someone who wants to use and combine tools and resources are limited. In this context, the establishment of a Web service dedicated to Natural Language Processing seems inevitable. This paper presents a Web service based on a "RestFul" architecture allowing use of the Fips parser, the ITS-2 translation system and the FipsVox text-to-speech system. Users can access all of these technologies by using just a browser while developers can create Web applications based on these resources.

Key words: Language Resources and Tools, Natural Language Processing, Web services, ReSTful Architecture

Introduction and motivation

For years, researchers have been trying to develop language resources and tools (LRT). Almost all of these technologies are often available by download only and/or restricted to a particular platform or environment.
In fact, the "democratization" of these LRT necessarily requires the creation of Web Services (WS).
WS dedicated to Natural Language Processing (NLP) would allow:

- easy access to LRT
- interoperability between different tools and resources
- development of Web application, for example eLearning application (Goldman et al., 2010)

This paper reports our experience integrating different tools and resources into a ReSTful Web Service. Our services provide access to the multinlingual Parser Fips (Nerima and Wehrli, 2009) (Wehrli, 2007) ,the translation system ITS-2 (Wehrli and Nerima, 2008) and the text-to-speech system FipsVox (Goldman et al., 2001).
From now on, this paper is organized as follows: Section 2 describes tools and resources integrated in the WS. In Section 3, we describe the characteristics of the WS and the available LRT. Finally, Section 4 is devoted to the creation of Web applications based on the WS.

Integrated Tools and Resources

This Web Service aims to use several LRT developed at the *Laboratoire d'Analyse et Technologie du Langage* (LATL) of Geneva. Fips (Nerima and Wehrli, 2009) (Wehrli, 2007), ITS-2 (Wehrli and Nerima, 2008) and FipsVox (Goldman et al., 2001) are the main tools of the infrastructure.

The Fips Parser

Fips is a robust multilingual parser which is based on generative grammar concepts for its linguistic component and object-oriented design for its implementation. It uses a bottom up parsing algorithm with parallel treatment of alternatives, as well as heuristics to rank alternatives.

The syntactic structures built by Fips are all of the same pattern, that is: $[_{XP}$ LX R], where L stands for the possibly empty list of left constituents, X for the (possibly empty) head of the phrase and R for the (possibly empty) list of right constituents. The possible values for X are the usual part of speech Adverb, Adjective, Noun, Determiner, Verb, Tense, Preposition, Complementizer, Interjection.

The parser makes use of 3 fundamental mechanisms: projection, merge and move.

Fips builds the canonical form for a sentence, in which extraposed elements (relative pronouns, clitics, interrogative phrases etc.) are coindexed with empty constituents in canonical positions (i.e., typical argument or adjunct positions). For instance, the sentence in (1) below is assigned by Fips the syntactic structure in (2), in which the canonical position of object for the verb address is taken by the empty constituent **e**. The latter stands for the trace of the noun issue, which has been extraposed through relativization. The trace **e**, the relative pronoun φ (a zero-pronoun), and the noun issue are all linked via the index *i*.

1. This too is an issue the Convention must address
2. $[_{TP}$ $[_{DP}$ This] $[_{VP}$ $[_{AdvP}$ too] is $[_{DP}$ an $[_{NP}$ issue$_i$ $[_{CP}$ $[_{DP}$ φ_i] $[_{TP}$ $[_{DP}$ the $[_{NP}$ Convention]] must $[_{VP}$ address $[_{DP}$ **e** $_i$]]]]]]]

Evaluation

The Fips multilingual parser has been developed for 6 languages (English, French, German, Italian, Spanish and Greek). A comparative evaluation has been conducted to show how the various implementations of Fips compare with respect to a near identical corpus, the European Parliament corpus (Koehn, 2005). We parsed approximately 1 million words in each of the six languages.

The table 1 shows the results:

Table 1 – Comparative evaluation of the parsers

Language	German	English	Spanish	French	Greek	Italian
Number of symbols	1106559	1075246	1228240	1350522	343461	1181785
Unknown words	10685	5852	9165	5643	6788	9006
Number of sentences	47058	41488	46216	45694	13328	44124
% of complete analyses	66,54%	75,68%	71,4%	75,97%	51%	70,76%
Speed (word/second)	17	38	132	83	196	112

ITS-2: the interactive translation system

ITS-2 is a large-scale translation system developed at the LATL of Geneva.

The language pairs currently supported are: French-English, English-French, French-German, German-French, Italian-French, Spanish-French.

The system is based on the familiar transfer architecture, with its three main components, parser, transfer and generation. The parser – associates with an input sentence a set of syntactic structures corresponding to GB S-structures, i.e. surface structures enriched with traces of moved elements and other empty categories. The role of the transfer component is to map source structures onto target structures. Transfer, which occurs at the D-structure level, is to a large extent a matter of lexical correspondence. For each lexical head of a SL structure, the lexical transfer component consults the bilingual lexicon to retrieve the most appropriate TL item, which is then projected according to the X-bar specifications of the TL. Applied recursively over the whole SLD-structure, this process determines an equivalent TL D-structure. From these structures, the generation component derives well-formed S-structures, which are finally converted into the target sentence by morphological process.

At the software level, an object-oriented design has been used, similar to the design adopted for the Fips multilingual parser on which it relies (Wehrli, 2007). To a large extent, ITS-2 can be viewed as an extension of the parser. It relies heavily on the detailed linguistic analysis provided by the parser for the supported languages, and exploits the lexical information of its monolingual lexicons. Both systems aim to set up a generic module which can be further refined to suit the specific needs of, respectively, a particular language or a particular language-pair.

To take a simple example, the direct object of the French verb *regarder* in (3) will be transferred to English as a prepositional phrase headed by the preposition at, as illustrated in (5).

This information comes from the lexical database. More specifically, the French-English bilingual lexicon specifies a correspondence between the French lexeme [$_{VP}$ regarder NP] and the English lexeme [$_{VP}$ look [$_{PP}$ at NP]]. For both sentences, we also illustrate the syntactic structures as built by the parser and/or the generator of ITS-2:

3. Paul regardait la voiture.
4. $[_{TP}\,[_{DP}$ Paul] regardait$_i$ $[_{VP}\,\mathbf{e}_i\,[_{DP}$ la $[_{NP}$ voiture]]]]

5. Paul was looking at the car.
6. $[_{TP}\,[_{DP}$ Paul] was $[_{VP}$ looking $[_{PP}$ at $[_{DP}$ the $[_{NP}$ car]]]]]

Evaluation
The last evaluation was for the Fifth and Sixth Workshop on Statistical Machine Translation. LATL participated in the French-English task in both directions.
The table 2 shows the results in terms of BLEU and translation edit rate (TER) using the newstest2010 and newstest2011 corpus as evaluation set.

Table 2 - Translation results from French to English and English to French measured on newstest2010 and newstest2011

Pair of language	BLEU	TER
French-English	16,5	0,785
English-French	20,4	0,690

FipsVox: a French TTS based on a syntactic parser
FIPSVox is a text-to-speech system for French developed at LATL. It is based on FIPS which produces detailed analyses and the MBROLA diphones-concatenation synthesizer. The syntactic information provided by the parser is directly exploited by the grapheme-to-phoneme module to handle heterophone homographs as well as French elision, denasalisation and liaison phenomena. The prosody generation module also uses this information to determine the dependency between phrases, the accentuation of syllables, and to identify particular syntactic structures such as extraposed constructions (cleft, heavy-NP shift, left dislocation structures, etc.), and parentheticals to derive of appropriate prosodic patterns.

Evaluation
There is no evaluation available for this tool.

Characteristics of the Web Service

A ReSTful Web Service

The services are created on the architectural model ReST. The notion of "Representational State Transfer" (ReST) was introduced in 2000 by Roy Fielding (Fielding, 2000). ReST is an architectural style simply based on the World Wide Web (WWW).
ReSTful means Web Services built using HTTP, URIs, XML, JSON, etc. It's interesting to deploy a WS with a range of existing infrastructure like web server, client library, proxy server, firewall, etc.

ReST architecture is totaly integrated to the WWW that's why it is the simplest and least expensive to implement.
ReSTful approach is basically composed of four concepts:

- the use of the "Uniform Interface" : all resources can be manipulated using the HTTP protocol and the method : PUT, GET, HEAD, POST, DELETE.
- the identification of resources via URI (Uniform Resource Identifier) : each resource can be uniquely identified and adressed.
- the operation are stateless.
- the use of standards like HTML, XML or JSON.

In short, ReST is not a standard it's an architectural style that makes maximum use of web technologies.

Available Resources

Our services provide access to the multilingual Parser Fips, the machine translation ITS-2 and the speech synthesizer FipsVox.
At the moment, most of these tools are available for German, English, Spanish, Greek and French.

Access to resources – Examples

The Data representation is done using standards (UTF-8, XML, TEI, etc.). Applications can be operated directly by URI. Tables 3, 4 and 5 describes resources and request parameters. Picture 1 gives an example of request and response of the *Analyze* resource.

Table 3: Available Resources

Resources	URI	Method	Description
Analyze	baseURI/Analyze[1]	GET	This resource generates a representation of linguistic analysis using the multilingual Parser Fips.
Translate	baseUri/Translation	GET	This resource generates a translation using the machine translation Its.
Speech	baseUri/Speech	GET	This resource generates a speech synthesis of a sentence using FipsVox.

Table 4: Request parameters for *Analyze*

Parameters	Value	Description
ln	fr, en, de, it, es, el (required)	Language for analyze (French, English, German, Italian, Spanish, Greek)
text	String (required)	Text to analyze

[1] Currently, baseURI is: http://129.194.19.89

Table 5: Request header for *Analyze*

Header	Value	Description
Accept	Parser, Xml, XmlTei, Tagger	Specify types of content which are acceptable for the response. Values correspond to different representation and format.

```
#Request :
GET /Analyze HTTP/1.1
Host : baseURI
Accept : XmlTei
text = le+chat+mange&ln=fr
#Response :
HTTP/1.1 200 OK
Content-Type : application/xml;charser=UTF-8
<TEI xmlns='http://www.tei-c.org/ns/1.0">
<s xml:lang="french">
    <phr type="DP"  function="SUBJ">
        <w type="DETERMINANT-DEFINI SIN MAS" lemma="le">Le</w>
        <w type="NOM-COMMUN SIN MAS" lemma="chat">chat</w>
    </phr>
    <phr type="" function="Predicate">
        <w type="VERBE-IND-PRE 3 SIN" lemma="manger">mange</w>
    </phr>
</s>
</TEI>
```

Picture 1 - Example of request and response for Analyze

Description of the Web Service

We use the WADL[2] specification to describe our application in a simple format. In this file (cf. Picture 2), we are defining:

- methods and parameters
- responses
- protocols and formats
- URIs of service

[2] WADL (Web Application Description Language) is an XML-based file format that provides a machine-readable description of HTTP-based web applications.

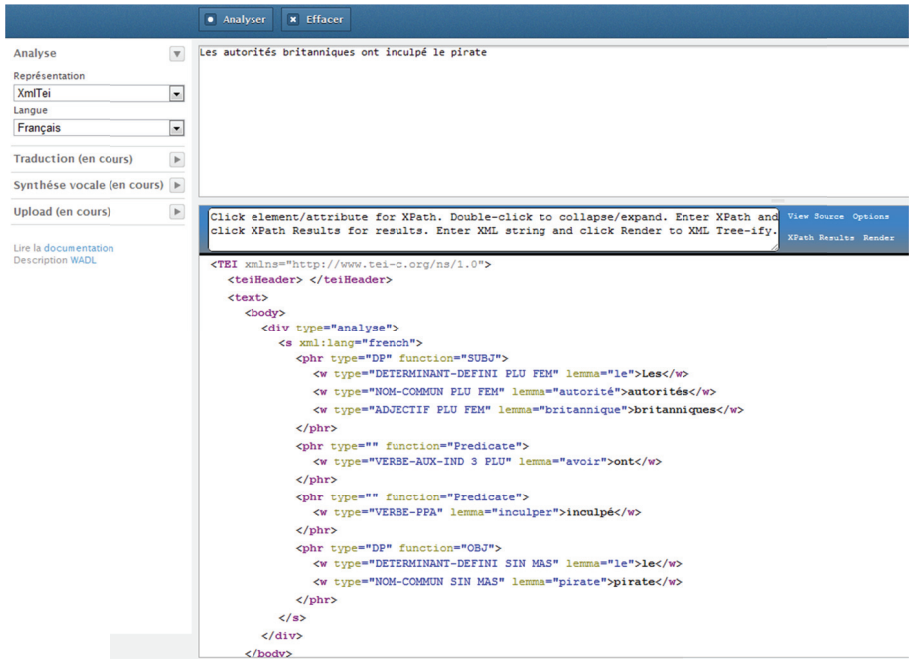
```
<resource path="Analyze">
    <method name="GET">
      <request>
        <param name="ln" type="language" style="query" required="true">
          <doc xml:lang="en" title="ln">[language]</doc>
          <option value="fr"/>
          <option value="en"/>
          <option value="de"/>
          <option value="it"/>
          <option value="es"/>
        </param>
        <param name="text" type="xsd:string" style="query" required="true">
          <doc xml:lang="en" title="text">[xsd:string]</doc>
        </param>
        <param name="Accept" type="xsd:string" style="header" required="true">
          <doc xml:lang="en" title="Accept">[xsd:string]</doc>
          <option value="Parser"/>
          <option value="Tagger"/>
          <option value="Xml"/>
          <option value="XmlTei"/>
        </param>
      </request>
    </method>
  </resource>
```

Picture 2 – WADL Description of Analyse resource

Web Application based on the Web Service

The user interface

It is possible to use the services with an intuitive interface created as a Web 2.0 application. In this interface, the user can analyze, translate or synthesize a text. The user can view them in several formats. (cf. Picture 3)



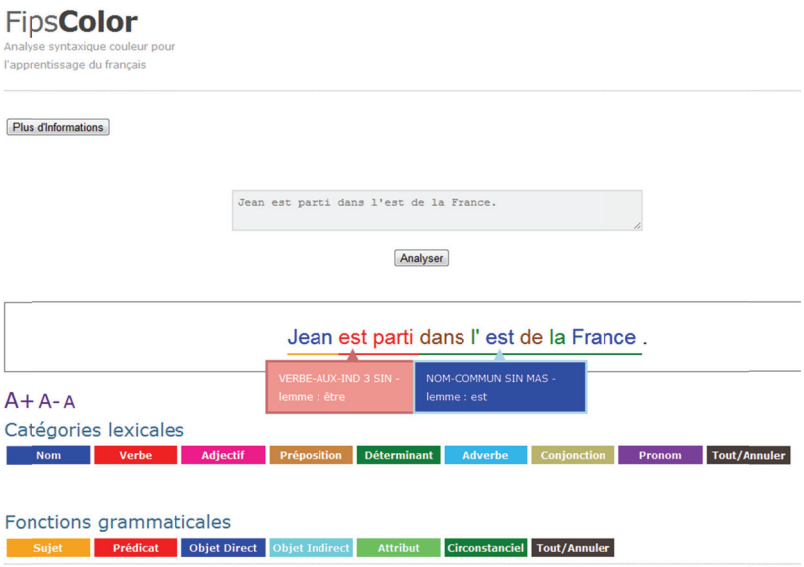
Picture 3- Screenshot of the User Interface

FipsColor: an eLearning application based on the WS

FipsColor is a Web application based on the *Analyze* resource (Goldman et al., 2010). The Fips parser analyzes a sentence into a syntactic structure reflecting lexical, grammatical and thematic information. The application adapts the structures in terms of constituents as existent in Fips to a grammatical annotation, giving as well a coloured representation.

This tool is used to highlight a particular aspect of French, such as the lexical ambiguity of certain words. Sentence in Picture 4 illustrates this. In this example, “est” is a verb and a noun. The different colors allow to clearly distinguish the correct grammatical category. FipsColor also underline the syntactic function in order to assimilate agreement or a particular structure (for example relative proposition).

This online interactive application can be used freely by teachers and pupils of primary education.



Picture 4 – Screenshot of the Web application FipsColor

Conclusion and Further Work

This paper presented a ReSTful Web Service, which aims at increasing the accessibility and usability of multilingual language resources developed at the LATL of Geneva. We believe that these services can be useful not only for researchers in linguistics, but also for other disciplines which have to analyze corpus. Services also provide creation of Web Application like FipsColor.

In a future version of the Web Service, we would like to include Semantic Web technology in order to provide search and reasoning.

References

Fielding, Roy Thomas. REST : Architectural Styles and the Design of Network-based Software Architectures. Irvine : University of California, 2000

Goldman, Jean-Philippe; Gaudinat, Antoine; Nerima, Luka; Wehrli, Eric. FipsVox : a french tts based on a syntactic parser. // Speech synthesis Workshop / Edinburgh, 2001

Goldman, Jean-Philippe; Laenzlinger, Christopher; Nebhi, Kamel. 2010. Fipscolor : grammaire en couleur interactive pour l'apprentissage du français. // TALN 2010 / Montréal, 2010

Hinrichs, Marie; Zastrow, Thomas; Hinrichs, Erhard. Weblicht: Web-based LRT services in a distributed escience infrastructure. // Seventh conference on International Language Resources and Evaluation (LREC'10) / Valletta, Malta, 2010

Ishida, Toru. Language grid : An infrastructure for intercultural collaboration. // In Proceedings of the IEEE/IPSJ Symposium on Applications and the Internet / Arizona, USA, January. 2006, ,pages 96–100

Koehn, Ph. Europarl: A Parallel Corpus for Statistical Machine Translation. // MT Summit / 2005

Nerima, Luka; Wehrli, Eric. L'analyseur syntaxique Fips. // IWPT 2009 ATALA Workshop : What French parsing systems? / Paris, France, 2009

Schafer, Ulrich. Shallow, deep and hybrid processing with UIMA and heart of gold. // In Proceedings of the LREC-2008 Workshop Towards Enhanced Interoperability for Large HLT Systems : UIMA for NLP / Marrakech, Morocco, 2008, pages 43–50

Villegas, Marta; Bel, Nuria; Bel, Santiago; Rodriguez, Vıctor. A case study on interoperability for language resources and applications // In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) / Valletta, Malta, 2010

Wehrli, Eric; Nerima, Luka. Traduction multilingue: le projet multra. // JEP – TALN – RECITAL Conference / Avignon, France, 2008

Wehrli, Eric. Fips, a deep linguistic multilingual parser // ACL 2007 Workshop on deep Linguistic Processing / Prague, Czech Republic, 2007, pages 120–127