Proceedings of the Language Technologies for All (LT4All) , pages 151–154
Paris, UNESCO Headquarters, 5-6 December, 2019.

MultiTAL
An online Platform to List NLP Tools for Under-Resourced Languages

Damien Nouvel, Mathieu Valette, Driss Sadoun
ERTIM
2, rue de Lille, 75007 PARIS
{damien.nouvel, mvalette}@inalco.fr, driss.sadoun@postlab.fr

Abstract

The diversity and variety of human languages raises indisputable difficulties for processing textual data. Regarding under-resourced languages, many software solutions have been designed, but many are poorly referenced and documented. The ERTIM (INALCO) lab published in 2016 a platform named MultiTAL that addresses this issue. Our platform lists tools available for languages. For each software, the knowledge base provides information concerning : processing tasks, implemented method, OS compatibility, among others. We do not claim to be comprehensive, but people populating the knowledge base are speakers of concerned languages, they downloaded and tested softwares, and provided detailed technical information for installation and use.

Keywords: Multilinguality, NLP Tools

Résumé

La diversité et la variété des langues humaines donne d'incontestables difficultés pour le traitement de données textuelles. Concernant les langages peu dotées, de nombreux logiciels ont été implémentés, mais beaucoup restent peu référencés et mal documentés. L'équipe ERTIM a mis en ligne en 2016 la plateforme MultiTAL qui réalise ce travail. Cette base de connaissances apporte des informations sur des outils, par langue et tâche. Nous ne prétendons pas être exhaustifs, mais les personnes remplissant la base étaient locuteurs des langues concernées, elles ont téléchargé et testé les outils, et ont renseigné des informations détaillées sur leur installation et leur utilisation.

1. Introduction

1.1. A Project for Under-Resourced Languages

It is a known issue that with 150 written languages, a great number of them are considered "under-resourced" from a Natural Language Processing (NLP) point of view. This is indeed even more true for the largest number[1] of oral languages, but we won't consider them in the present work. In the context of globalisation and digitalization, this concern is even more serious as language communities require to access the information-based society and the Internet for various purposes of their everyday life.

Unsurprisingly, languages that are the best equipped with digitalized linguistic resources (e.g. corpora, lexicons, software) are those that have either a long history related to computers or a sufficient economic weight to receive more recent developments. On the other side, those that are not in this case are most of the time very late on such developments. The related communities are often forced to use one of the *lingua franca* already well established on the Internet. As a side-effect, this also raises the risk of language impoverishment. Yet, making those languages exist in the digital world is undoubtedly a necessary step and can't be avoided. When no economic benefits are in sight, those development can only be handled by local or international non-profit organisations such as governments, NGOs, associations or academics.

Our institute, INALCO (National Institute for Oriental Languages and Civilizations) is both a university that teaches around 100 oriental languages from Central Europe, Africa, Asia, America and Oceania and a research center that works on related languages. Within this institute, our research team, ERTIM, frequently collaborates with researchers and teachers on a number of those languages. This position requires that we can quickly establish what is the status of a language in terms of digitalized resources.

To help us with this, we initiated the *MultiTAL*[2] project in 2016, hosted at (`http://multital.inalco.fr`. The *MultiTal* infrastructure aims at providing systemic descriptions of tools (software) for under-resourced languages, so as to document them, promote them, and ease their accessibility. For this purpose, tools are downloaded, installed and tested, we select those that are actually operational, and provide accurate and *critical* documentation, rather than providing lists of tools that have not been tested except by their designers. We also ranked this project as "highly multilingual": the main pages of our website have been localized into 7 languages, so that the largest number of people can understand it, even if they have limited knowledge of English.

1.2. Related Work

Over the last decade, the number of digitized materials has considerably grown. The willingness to take into account this new digital content has led to the popularization of the use of Language Resources (LR) and NLP technologies. However, LR are still difficult to find because they are drowned in the mass of web content. Moreover, their documentation is often monolingual and written either in the developers' languages (such as Arabic, Chinese, Japanese or

[1] Depending on studies, 6000 or 7000

[2] TAL stands for *Traitement Automatique des langues*, as "Natural Language Processing" in French

Russian) or in a *lingua franca* (such as English or French). This situation makes it difficult for scholars to use or re-use LR that could be useful for their work or research. Hence, storing and distributing LR has become an issue in itself. This has been addressed by many initiatives all around the world (Váradi et al., 2008; Tohyama et al., 2008; Piperidis, 2012; Tonne et al., 2013; Calzolari et al., 2012). These initiatives are essential to promote the research and development of language technologies. They also may provide a real picture of tools and resources that are currently available for several languages (Skadina et al., 2013; TADIĆ, 2012; Del Gratta et al., 2014).

In order to describe and share LR, different meta-data models have been proposed (Gavrilidou et al., 2011; Broeder et al., 2012; McCrae et al., 2015a). The models of each provider depend on their coverage and the kind of LRs they manage. Hence, there are as many meta-data models for describing LRs, which may represent a limit for resource sharing and lead to the re-creation of already existing LR resources (Cieri et al., 2010). To address this issue, different attempts have been made, such as an initiative for harmonising between ELRA and LDC catalogs (Cieri et al., 2010) and more recently ontologies were used to devise interconnections among resources (Chiarcos, 2012) or to make meta-data available from different sources under a common scheme (McCrae et al., 2015a; McCrae et al., 2015b). In the perspective of a possible interoperability between our meta-data model and the existing ones, we chose to use an ontology for storing *MultiTal* infrastructure data.

Most existing LR infrastructures focus on occidental languages and invite developers of resources or tools to describe them themselves. Even if it eases access to LR technologies, when it concerns NLP tools it does not necessarily make their use any easier. Often, their usage instructions remain poorly documented. In our project, we aim to list NLP tools processing written non-occidental languages or more precisely languages taught at INALCO. In this framework, each NLP tool is identified, tested and fully documented by an intern speaking the language the tool processes. Then, if the tool appears to run correctly its information is stored within our meta data model (ontology) and its resulting documentation is made available. Our aim is to ensure that tools described on *MultiTal* infrastructure can be properly installed and executed by end-users. As MultiTal's end-users may not be language technology experts and their mother tongue may vary, we use an ontology verbalisation method (Androutsopoulos et al., 2014; Cojocaru and Trãuşan Matu, 2015; Keet and Khumalo, 2016) detailed in (Sadoun et al., 2016) to automatically produce documentation in multiple languages. So that we provide end-users with simple, structured and organised documents containing NLP tool information and detailing instructions of how to install, configure and run tools fitting their needs.

2. The MultiTAL Platform

2.1. Goals

As already stated in Section 1.1., issues are identified and well-known for many languages. Among those, we focus on three main difficulties:

- Relevant NLP tools are not so easy to find, in particular for under-resourced languages
- Documentation is not always comprehensive (sometimes native)
- Instructions to install, configure and execute NLP tools are not that simple

Facing those issues, our goal is to make technologies more accessible regardless of the expertise or spoken language of end-users by means of providing information with a simple, multilingual, structured and standardised tool documentation.

2.2. Methodology

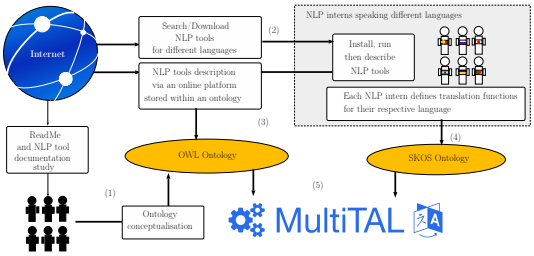

Figure 1: Process of the MultiTAL process

As depicted in 1, we rely mainly on human contributors to enrich our documentation of NLP tools. Their first objective is to verify that the software can easily be downloaded, installed and executed on a standard and minimally configured system (whether Windows, Linux or MacOS) and that the output has minimal accuracy regarding the concerned task so that it can be used out-of-the-box. We believe this to be a major advantage of our platform: we guarantee that we have succeeded in those preliminary steps.

Here is the list of information we do provide about tools:

- The name
- Last update (at the moment we documented it)
- Main programming language
- Accessibility (download, online or web service)
- Licenses
- Authors
- References (mostly academics)
- Description
- Implemented "tasks" with for each:
  - Type(s) of task(s)
  - NLP methods
  - Languages covered
  - Input and output (types and encodings)
- Installation and execution procedures (see below)

We did extensive documentation work for the last item regarding installation and execution. The documentation also describes and gives concrete examples of commands that may be issued,as depicted in Figure 2, for the MeCAB[3] tool. Again, we are driven by end-users consideration, especially those people working with NLP tools that are not necessarily developers but may have a minimum knowledge of using computers to execute this kind of tools by using the command line. In this documentation process, validation is important: tools are entered into the knowledge base, but are not published until the contributor has actually tested the tool.

Installation procedure [Linux]

1 - download mecab-0.996.tar.gz from http://taku910.github.io/mecab/#download

```
download mecab-0.996.tar.gz http://taku910.github.io/mecab/#download
```

2 - uncompress -xzvf mecab-0.996.tar.gz

```
tar -xzvf mecab-0.996.tar.gz
```

3 - go to directory mecab-0.996

```
cd mecab-0.996
```

4 - type the command: ./configure

```
./configure
```

5 - type the command: make

```
make
```

6 - type the command: sudo make install

```
sudo make install
```

Execution procedure [Linux]

1 - type the command: mecab input_file.txt -O output_format -o output_file.txt

```
mecab  input_file.txt -O output_format -o output_file.txt
```

Figure 2: Installation and Execution Documentation

2.3. Architecture

Behind the scenes, our platform is a simple LAMP[4] webserver that is used both for public access and for backend administration of the knowledge base. The latter is an OWL ontology where each tool is described using the properties as reported in Figure 3. Our conceptualisation of NLP processing raised a number of questions: we have been led to distinguish a tool from the tasks it accomplishes (each task is thus validated separately upon testing). In practice, we observed that indeed a tool may have very different accuracy for a given task depending on the language, but we didn't enter into this level of detail.

The platform is hosted at `http://multital.inalco.fr`. By default, the search engine displays a view on tasks, for the sake of users looking for a tool that implements a specific type of processing on a particular language. User interface has been more or less extensively translated into 7 languages : Arabic, Chinese, English, French, Hindi, Japanese, Russian, Tibetan. Both the ontology and the SKOS can be queried using SPARQL or downloaded as RDF/XML files.

[3] `http://taku910.github.io/mecab/`

[4] Linux Apache MySQL PhP

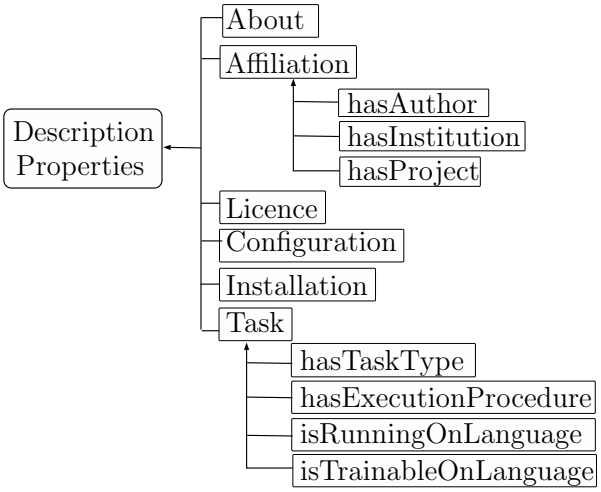

Figure 3: Ontology Properties

2.4. Current status

Currently, the platform documents:

- 45 NLP task types
- 47 languages
- 91 of them are published (among 167 tested)
- 33 NLP methods
- 112 tool authors

Tables 1 and 2 give a rough overview of the documentation per task and language. Be aware that displayed numbers contain redundancy, since a tool may actually implement multiple processings (as tasks) for each language.

Task Type	# Tasks
Part-of-speech tagging	36
Segmentation	32
Tokenization	18
Morphological analysis	15
Lemmatisation	13
Morphological tagging	13
Parsing	12
Dependency parsing	9
Named Entity Recognition	8
Stemming	8
Transcription	6
Transliteration	6
Concordances	5
Diacritization/vocalization	5

Table 1: Tasks by Tasks types

3. Conclusion

In this paper, we described our our MultiTAL platform, which provides documentation for NLP tools restricted to some of the languages taught in our institute INALCO

Language	# Tasks
Arabic	47
Chinese - Mandarin (simplified)	35
Chinese - Mandarin (traditional)	33
Japanese	22
Russian	16
Hindi	15
English	12
French	9
Bulgarian	5
Hungarian	5
Tibetan	5
Ukrainian	5

Table 2: Tasks by Language

(most of the time, oriental and under-resourced languages). Our platform not only provides a structured description of the tool, but also includes additional features. First, it guarantees that we were able to use the tool (by downloading and executing it or online). Second, we detail documentation on the aspect of the installation and execution procedures as commented scripts. The user interface has been translated in 7 languages. We believe our project may be a starting point to establish guidelines and best practices for improving NLP tool documentation.

4. Acknowledgements

Many thanks to all our contributors! This work was founded by the USPC (université Sorbonne-Paris-Cité).

5. Bibliographical References

Androutsopoulos, I., Lampouras, G., and Galanis, D. (2014). Generating natural language descriptions from OWL ontologies: the naturalowl system. *CoRR*, abs/1405.6164.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., and Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In *the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1387–1390. European Language Resources Association (ELRA).

Calzolari, N., Gratta, R. D., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE map. Harmonising Community Descriptions of Resources. In *LREC*, pages 1084–1089.

Chiarcos, C. (2012). Ontologies of linguistic annotation: Survey and perspectives. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Cieri, C., Choukri, K., Calzolari, N., Langendoen, D. T., Leveling, J., Palmer, M., Ide, N., and Pustejovsky, J. (2010). A road map for interoperable language resource metadata.

Cojocaru, D. s. A. and Trãuşan Matu, c. (2015). Text generation starting from an ontology. In *Proceedings of the Romanian National Human-Computer Interaction Conference - RoCHI*, pages 55–60.

Del Gratta, R., Frontini, F., Khan, A. F., Mariani, J., and Soria, C. (2014). The lremap for under-resourced languages. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, page 78.

Gavrilidou, M., Labropoulou, P., Piperidis, S., Francopoulo, G., Monachini, M., Frontini, F., Arranz, V., and Mapelli, V. (2011). A metadata schema for the description of language resources (lrs). *Language Resources, Technology and Services in the Sharing Paradigm*, page 84.

Keet, C. M. and Khumalo, L. (2016). Toward a knowledge-to-text controlled natural language of isizulu. *Language Resources and Evaluation*, pages 1–27.

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., and Cimiano, P., (2015a). *The Semantic Web: ESWC 2015 Satellite Events*, chapter One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web, pages 271–282.

McCrae, J., Cimiano, P., Doncel, V., Vila-Suero, D., Gracia, J., Matteis, L., Navigli, R., Abele, A., Vulcu, G., and Buitelaar, P. (2015b). Reconciling Heterogeneous Descriptions of Language Resources. *ACL-IJCNLP 2015*, page 39.

Piperidis, S. (2012). The meta-share language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Sadoun, D., Mkhitaryan, S., Nouvel, D., and Valette, M. (2016). Readme generation from an owl ontology describing nlp tools. In *2nd International Workshop on Natural Language Generation and the Semantic Web*.

Skadina, I., Vasiljevs, A., Borin, L., LindÃ©n, K., Losnegaard, G., Pedersen, B. S., Rozis, R., and De Smedt, K. (2013). Baltic and nordic parts of the european linguistic infrastructure. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 195–211.

TADIĆ, T. V. M. (2012). Central and south-east european resources in meta-share. In *24th International Conference on Computational Linguistics*, page 431.

Tohyama, H., Kozawa, S., Uchimoto, K., Matsubara, S., and Isahara, H. (2008). Construction of a metadata database for efficient development and use of language resources.

Tonne, D., Rybicki, J., Funk, S., and Gietz, P. (2013). Access to the dariah bit preservation service for humanities research data. In *Parallel, Distributed and Network-Based Processing (PDP), 21st Euromicro International Conference*, pages 9–15.

Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., and Koskenniemi, K. (2008). Clarin: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.