The LIMA Multilingual Analyzer Made Free: FLOSS Resources Adaptation and Correction

Gaël de Chalendar

CEA, LIST, Vision and Content Engineering Laboratory,
Saclay, F-91191, France;
gael.de-chalendar@cea.fr

Abstract

At CEA LIST, we have decided to release our multilingual analyzer LIMA as Free software. As we were not proprietary of all the language resources used we had to select and adapt free ones in order to attain results good enough and equivalent to those obtained with our previous ones. For English and French, we found and adapted a full-form dictionary and an annotated corpus for learning part-of-speech tagging models.

Keywords: FLOSS; linguistic analyzer; resources adaptation

1. Introduction

At LREC 2010, we presented LIMA, our multilingual analyzer and we concluded with our goal to release it under an open source license (Besançon, Chalendar (de), et al. 2010). This is now done and LIMA is publicly available on GitHub[1]. LIMA is a tool developed in portable C++. It is based on a classical pipeline mechanism, but with a high quality conception and realization, making it usable at industrial scale (it is integrated in several commercial applications). It is also highly configurable making it able to adapt to probably all languages (from French to Arabic and Chinese for example).

To put it under a Free Software licence, we had to replace some non-redistributable proprietary linguistic resources. It was not feasible nor desirable to develop from scratch new resources like dictionaries or corpora. We had to select and adapt resources already available under a compatible free software license. In this article, we present those we chose and the adaptations we had to do to make LIMA as efficient with these free resources than with our commercial resources. We start by reminding the main specifications of the LIMA linguistic analysis framework.

2. Presentation of LIMA

LIMA is available under a dual licensing model. The Free version is available under the Affero General Public License. It is fully functional with modules and resources to analyze English and French texts. Everyone can thus use LIMA for all purposes as soon as the software linked to it or running it through Web services is Free software too. The commercial version is completed with specific modules and resources to analyze other languages (Arabic, Chinese, German, etc.). It is available directly from CEA LIST through R&D partnerships or through an industrial partner with offers including support and adaptation to one's needs.

This platform was developed with the following requirements:

- multilingualism, with a broad spectrum of languages;
- diverse applications (information extraction, information retrieval, summarization, ...);
- extensibility;
- efficiency in an industrial context.

This makes necessary to design a highly modular and flexible architecture with some generic modules and others specific and with resources for each language. All languages are not characterized by the same set of linguistic phenomena and their processing doesn't rely on the combination of the same elementary analyses. Moreover, even if an analysis module can be used for different languages, the linguistic resources it relies on are usually specific to each language.

In our 2010 paper, we described the various possible architectures. Let just say here that we use a configurable processing chain, each element corresponding to levels of linguistic analysis. Treatments use specific linguistic resources and share and modify a common data structure. Modules and resources are loaded as necessary as stated by configuration files. This processing chain is conceived to integrate LIMA in various contexts.

English and French pipelines use the following treatments covering segmentation, morphological analysis and parsing:

- *tokenization*: using a character-based automaton with a window of seven characters. The tokenization dos not consider ambiguity. Later steps are able to group or split some tokens;
- *dictionary check*: each token is checked against a full-form dictionary (a dictionary including all the known forms of single words), and possible lemmas and part-of-speech (PoS) are associated with it;

[1] https://github.com/aymara/lima/wiki

- *hyphenated words*: this unit performs a special treatment to associate lemmas and categories of hyphenated words not present in the dictionary (parts of the split word are looked up in the dictionary);
- *abbreviation split alternatives*: (English only) tokens including a single quote (like "don't", "I'm", etc.) are looked up in the dictionary which indicates the tokens they are made of;
- *idiomatic expressions*: compound expressions recognizing. It reduces the ambiguity before PoS tagging by considering the expression as a whole. It uses a generic pattern recognition unit based on finite state automata.
- *unknown words*: the unknown words are given default PoS using a guesser based on typographical clues;
- *named entities recognition*: the same pattern recognition unit is used with different rules to recognize numbers, dates and named entities;
- *PoS-tagging*: two PoS-taggers are currently implemented in LIMA. The first historical one uses a Viterbi algorithm on PoS trigrams. The second one uses SVMTool++ (Giménez and Màrquez 2004), an LGPL tagger based on SVM. Both use models learned from annotated corpora;
- *parsing*: using a dependency grammar implemented as a set of simple rules executed by the generic pattern recognition unit (Besançon and Chalendar (de) 2005).

3. Adapting free resources for their use within LIMA

Among the resources evoked above, several were already created by our team: tokenization automatons, idiomatic expressions, named entities and syntactic analysis rules. But two fundamental ones for each language were bought proprietary resources: dictionaries and PoS-annotated corpora. For all of them we had to find and adapt free resources. After the initial adaptation, we improve the resources by iteratively running a ten-fold cross validation, learning disambiguation model on 90% of the corpus and testing on 10%, and correcting errors revealed.

For French, we have chosen the Lefff extensional dictionary (Benoît Sagot 2010) and the Free French Treebank (FFT) (Hernandez and Boudin 2013), both available under LGPL-LR. For English, we use the dictionary of the FreeLing project (Carreras et al. 2004).[2] We initially chose the only free[3] large annotated corpus that we discovered, namely the Open American National Corpus (OANC). Unfortunately, we found the quality of its annotations, automatically obtained with a modified GATE's ANNIE system, too low. The work needed to obtain results of a quality high enough for a public release was not compatible with the time available. We thus decided to use the 10% sub part of the Penn Tree Bank (PTB) which is freely available in NLTK.[4] But this corpus is not free in the sense that it is not redistributable. Thus we are only able to release our byproducts, the taggers compiled resources. We consider LIMA as incomplete without complete source resources that the user or contributor can adapt to its needs. We thus plan to work again on the OANC to correct its problems.[5]

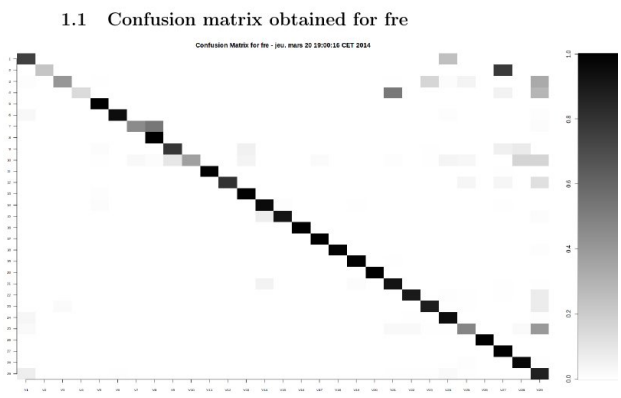

Figure 1: The confusion matrix.

We use several tools to detect and identify the errors. First, a confusion matrix (Figure 1) displays on the y-axis the expected tags and the obtained ones on the x-axis, with a gray level indicating the percentage of cases in the given cell. Tags are sorted by increasing frequency in the corpus.In the example, gray cell at (29, 25) shows that there is a somewhat frequent confusion between ADJ and NC tags.

An error-free analysis would give a matrix with a black diagonal and no other colored cell. It allows to detect immediately systematic errors, due for example to tag misspelling or to annotation errors on rare tokens. Another error-detection tool is the list of most frequent errors at the token granularity. It allows to find some inconsistencies between the dictionary and the corpus, either because a tag is missing in the dictionary for a given token or because a token has systematically the wrong tag in the corpus.

Finally, a very productive tool is the replacement of the tags given by the unknown words tag guesser by a single UNK tag which after analysis directly gives the list of tokens tokens absent from the dictionary. They can be absent for various reasons. For some of them the dictionary must be completed (*e.g.* proper names) while for others the corpus should be corrected: we do not want to learn models including spelling errors. A

[2] http://www.freeling.org

[3] allowing use, reproduction and modification

[4] http://nltk.org/

[5] Another option is to use the Brown Corpus. But its Creative Commons-Attribution-Non Commercial (CC-BY-NC) license make it non-free and thus problematic.

frequent case is the presence in the corpus of foreign words or even complete sentences in foreign language. They are sometimes wrongly annotated with standard tags in the corpus.

3.1. Tagsets adaptation

The first version of LIMA was designed during the seventies and the eighties at a time the when learning-based statistical taggers were not available. This is why the tagsets used were very large, incorporating a lot of symbolic information about the possible previous and next tags. This level of details allows to use hand-written disambiguation rules or to learn disambiguation rules from a limited annotated corpus. Our initial tagsets were thus absolutely not compatible with those used in the free resources we adopted. We then had to change also our tagsets.

LIMA uses a data structure with a so-called macro-category, roughly equivalent to the tags used in classical PoS annotated corpora, a micro-category, the kind of tag we just described above and various traits like time, genre, number, etc. For the use with the resources we adapt here, the distinction macro vs. micro category does not stand anymore and it will be ignored from now on.

For French, the Lefff contains categories and traits making it straightforward to convert them into a LIMA format. Concerning the PoS tags themselves, we use the 28 tags of the FFT, developed initially by (Crabbé and Candito 2008). The token `hésiterai#V` in the FFT corresponds to the LIMA dictionary entry `hésiterai hésiter Vpifi1-s`. The tag of this entry, `Vpifi1-s`, matches the following traits: `TAG=V, NUMBER=SING, PERSON=1, TIME=FUTURE` and `SYNTAX=TRANS`. All this data are extracted from the following Lefff entry[6]:

```
hésiterai  100 v [pred="hésiter_____1
<Suj:cln|sn,Obj:(à-sinf)>",      @CtrlSujObj,
@pers, cat=v,@F1s ]  hésiter_____1
Default F1s %actif  v-er:std
```

As one can see, the Lefff contains a lot of information that are currently ignored in LIMA. In the future, we could make use of some of them. This includes the subcategorization frame which could be useful for syntactic analysis.

We introduce also little modifications that help the disambiguation with a relatively small corpus. These changes are tag specializations that are reversible after tagging if necessary. There is specialized tags for numerical determiners and adjectives and specific tags for the determiners `de`, `des` and `du`. Appendix 5.1. lists the set of LIMA French tags.

For English, the adaptation was simpler as both the dictionary and the annotated corpus used the same Penn Treebank tagset. We just renamed punctuation tags with symbols for a better readability and easier scripting of some tasks, added a tag for currencies and created some tags for specific tokens (`about`, `not`, `that`, `there`) (see Appendix 5.2.).

3.2. PoS tagger learning corpora

The FFT contains 2,354,146 tokens. The ten-fold cross validation with the Viterbi tagger takes 32'18" on the full corpus. On 100,000 tokens, it takes only 2'10" minutes with a precision higher of around 0.51%. Thus, we do the tests on the 100,000 tokens sub part while applying the corrections on the whole corpus. Similarly, the SVM-based tagger has a around 0.36% better precision on the French 100k tokens corpus but it takes 21'47" for the whole learn-test cycle. Furthermore, the source of the errors are easy to understand with the Viterbi-based tagger while it is largely opaque with the SVM one. Thus we use the Viterbi-based tagger during our daily work on resources but show results obtained with the SVM-based one.

There was an important number of tokenization differences between the reference corpora and the LIMA tokenization (5,041 in French and 6,069 in English). LIMA applies PoS tagging after searching idiomatic alternatives and named entities, replacing the sequences of tokens matched by a single token with usually only one tag. This first adaptation has been to modify the learning corpus using an aligner based on the output of the GNU diff tool. We replaced in the corpus the tokens grouped by LIMA by a single one. In French, this has been done on 116,755 multi-word tokens and 2,463 in English.

After this step, a large number of tokenization differences were remaining where the tags could not be compared. These differences are reported by our errors analysis tool. They proved to be mainly named entities that were wrongly detected, partly due to the changes in the tagsets. For example, there were previously specific tags for numbers. Named entities rules used this fact. We then had to create new rules matching numbers based on their digital characters. At the time of this writing, it remains only 1,361 tokenization differences in French and 1,095 in English.

Finally, we manually corrected 34,807 annotation errors in the FFT and 510 in the PTB.

3.3. Improvements in dictionaries

In both languages dictionaries, we removed a few entries that have no occurrence in corpora but that introduce a lot of ambiguity. In French, there is five such entries, for example `est#ADJ` or `la#NC`. There is slightly more of them in English: 69 including `even#VB` or `less#CC`.

The removal of valid entries from the dictionary is obviously not an adequate solution, but a temporary one. Depending on the tagger used, other more complicated but more sound solutions are envisageable. With the Viterbi-based tagger, the only correct solution would be to augment the training corpus size in order to include more occurrences of rare words.

[6]There is in fact four entries in the Lefff for these traits as the Lefff encodes syntactic alternations that we currently don't use in LIMA.

With the SVM-based tagger, we could inject the LIMA dictionary instead of using only the dictionary built at training time as we did up to now. This would allow to obtain classification data at least based on the possible tags of the word.

In both languages we added entries in the dictionaries for tokens of the corpus that were absent from the dictionaries (2,001 in French and 3,267 in English). In English, they were 1,549 proper nouns but also 343 adverbs and 1,323 various forms among which 962 were present but without the tag used in the dictionary. 361 were completely absent from the dictionary. They will have to be checked individually to avoid introducing spelling errors. The categorization of these entries is still to be done but a first look suggests that they include a lot of spelling error.

3.4. Other improvements

The last improvements we made were on a one hand on idiomatic expressions extraction rules and on named entities expressions extraction rules. Concerning idiomatic expressions, 4 wrong rules were removed and 12 added in French and 6 were corrected and 80 added in English. For example, the French expression "lors de" is not a subordinating conjuction. It has been removed. Also, several expressions were tagged a preposition while they are concatenated prepositions and determiners, like "quant au" or "à cause des".

The named entities changes were made to reduce the number of tokenization differences between LIMA and the corpora. This includes improved dates detection rules, missing event detection rules and the addition of the rules recognizing people names with particles. In English, we did the same kind of improvements but also added generic rules for organizations detection, like:

```
Association:
  (National)?:
  of (t_capital_1st|t_capital|\&){1-5}:
  ORGANIZATION:
```

This rule detects organization names like "National Association of Home Builders".

3.5. Results

	English	French
Raw	80.09	86.22
Corpus	81.06	90.70
Tagset		90.94
Dictionary	95.00	94.05

Table 1: Evolution of precision scores in percentage with various corrections in resources .

Precision scores before corrections were, respectively in French and in English, of 86.22 and 80.09. Table 1 shows the progression of scores with the various improvements to resources. At time of writing, the English SVM-based tagger is already usable even if we are still lower than state of the art (95% vs. 97.5% for SCCN (Søgaard 2011)). On the contrary, French performance at 94.05% is too low compared to MElt (97.8%) (Denis and Benoît Sagot 2009) but the work on its resources is still ongoing. One can note the gain of more than three points when introducing the few specific tags described in section 3.1. in French. The main paths of improvement will be to work on alignment errors and to track down annotation errors in the corpus. We are confident being able to reach higher scores, at least a 95% threshold as in English.

4. Conclusion and future work

In this article, we have described the adaptation of freely available natural language resources that were necessary for the release of the LIMA platform as free software. It was relatively simple to adopt new tagsets, dictionaries and annotated corpora. It was also quite easy to adapt our other resources (syntactic analysis rules, idiomatic expressions and named entities recognition rules) to these new tagsets. The result is a fully free, powerful, extendable and adjustable multilingual natural language processing system. Its performance must still be improved to join those of its proprietary version or of state of the art systems but the path towards such a goal is clear. We have used and improved in this work several free resources. We will soon be able to provide these improvements upstream.

It is not too much difficult to add support to other languages as soon as one can find freely available resources, starting with a dictionary and a tagged corpus. Other resources, like named entities rules, can often be copied from one of the existing languages as a first approximation. If a language needs special treatment, like tokenization of Chinese ideograms or splitting of compound words in German, then new pipeline units must be developed. They currently must be implemented in C++ but we are developing APIs allowing to program them in any other language like python. Documentation for resource and code development is still insufficient on the LIMA site but we are working on it and everybody's contribution is welcome.

The free version of LIMA is offered with support for French and English but LIMA already has commercial support for Spanish, German, Arabic and Chinese, inter alia. We hope to be able to release some of them under a Free Software license in the future.

References

Besançon, Romaric and Gaël Chalendar (de) (June 2005). "L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY". In: *actes de la 12e conférence annuelle sur le Traitement Automatique des Langues Naturelles, TALN 2005*. Dourdan, France.

Besançon, Romaric, Gaël Chalendar (de), Olivier Ferret, Faïza Gara, and Nasredine Semmar (May 2010). "LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation". In: *Proceedings of Language Resources and Evaluation Conference, 2010*. Malta.

Carreras, Xavier, Isaac Chao, Lluí Padró, and Muntsa Padró (2004). "FreeLing: An Open-Source Suite of Language Analyzers". In: *4th International Conference on Language Resources and Evaluation (LREC'04)*.

Crabbé, Benoît and Marie Candito (June 2008). "Expériences d'analyse syntaxique statistique du français". French. In: *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'08*. Avignon, France, pp. 44–54.

Denis, Pascal and Benoît Sagot (2009). "Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort". In: *Pacific Asia Conference on Language, Information and Computation*. Hong Kong, Chine.

Giménez, Jesús and Lluís Màrquez (2004). "SVMTool: A general POS tagger generator based on Support Vector Machines". In: *Proceedings of the 4th LREC*. Lisbon, Portugal.

Hernandez, Nicolas and Florian Boudin (June 2013). "Construction automatique d'un large corpus libre annoté morpho-syntaxiquement en français". Français. In: *Actes de la conférence TALN-RECITAL 2013*. Sables d'Olonne, France.

Sagot, Benoît (2010). "The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French." In: *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*. Istanbul, Turkey.

Søgaard, Anders (2011). "Semisupervised condensed nearest neighbor for part-of-speech tagging". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*. HLT '11. Portland, Oregon: Association for Computational Linguistics, pp. 48–52. ISBN: 978-1-932432-88-6.

5. Appendix

In this appendix, we list the tags used in French and English. Their declaration can be found in the sources of the project in files code-fre.xml and code-eng.xml alongside the other morphologic traits used.

5.1. French tagset

LIMA French tag	Signification
ADJ	Adjective
ADJWH	Interogative adjective
ADJNUM	Numeral adjective
ADV	Adverb
ADVWH	Interogative adverb
CC	Coordinating conjunction
CS	Subordinating conjunction
DET	Determiner
DETDE	"de" token as determiner
DETDES	"des" token as determiner
DETDU	"du" token as determiner
DETWH	Interogative determiner
DETNUM	Numeral determiner
U	Unknown word
ET	Foreign word
PREF	Prefix
I	Interjection
NC	Common noun
NPP	Proper noun
PONCT	in-sentence punctuation
PONCTU_FORTE	sentence separator
P	Preposition
P+D	Preposition and determiner
PRO	Pronoun
PROREL	Relative pronoun
PROWH	Interogative pronoun
CL	Clitic
CLO	Object clitic
CLR	Relative clitic
CLS	Subject Clitic
V	Verb
VIMP	Imperative verb
VINF	Infinitive verb
VPP	Past participle
VPR	Present participle
VS	Subjonctive verb

5.2. English tagset

LIMA English tag	Signification
ABOUTIN	“About” as prep.
ABOUTRB	“About” as adverb
CC	Coordinating conjunction
CD	Cardinal number
COLON	Colon
COMMA	Comma
CPAR	Closing parenthesis
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Prep. or subord. conj.
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
NOT	“Not” token
OPAR	Openinfg parenthesis
OQU	Opening quote
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
QUOT	Closing quote
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SENT	Sentence delimiter
SYM	Symbol
THATDT	“That” as determiner
THATIN	“That” as prep.
THATPRP	“That” as perso. pron.
THATRB	“That” as adverb
THATWDT	“That” as interrog. det.
THERERB	“There” as adverb
TO	“To” token
UH	Interjection
UNK	Unknown word
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present part.
VBN	Verb, past participle
VBP	Verb, non-3rd pers. sing. pres.
VBZ	Verb, 3rd person sing. present
WDT	Interogative determiner
WP	Interogative pronoun
WP$	Poss. interrog. pronoun
WRB	Interogative adverb