Evaluation of TAL Tools

-Report Project TAL-

Vincent DOISNEAU, Florian LE PALLEC, Anais MOKDAD

10 Mars 2019

Polytech Paris Sud
Maison de l'Ingénieur - bât 620 - Centre Scientifique d'Orsay - 91405 Orsay - France
Tel : +33 (0)1 69 33 86 00 - fax : +33 (0)1 69 41 99 58 - www.polytech.u-psud.fr

Preface

This document is the report of results as part of the student project on information retrieval at the Engineer School Polytech Paris-Sud from the Université Paris-XI, France.
The objective of this assignment is the **evaluation of language processing tools**.
It has been carried out by a group of 3 students in the 5-th year of the engineering cycle, ***Vincent DOISNEAU***, ***Florian LE PALLEC*** and ***Anaîs MOKDAD*** under the direction of ***SEMMAR Nasredine***.

We will describe in this document the objectives of the project, the results of the evaluation with the two linguistic analysis platform, the strong points as well as the limitations of each platform.
We will also keep one page to detail what has been the contribution of each member of the group.

Introduction

We have 2 different open-source tools, LIMA and Stanford, though they are quite similar and have the same purpose, these tools differ in their approach. Lima is a platform centered on the use of rules and dictionaries while Stanford focus on statistics.
Our objective will be to experiment with these 2 framework/linguistic analysis platform and evalutate/compare their results.

State of the Art

Linguistic Analysis : Current situation

Many natural language resources are now available and have been an essential part for the release of platforms like LIMA and Stanford. However, not every tool have the same purpose or implement the same method. some will put more emphasis on tokenization, while others may focus on the meaning carried by it. That is why, it is most important to test these tools and understand their differences to get a better grasp of their strengths and weaknesses.

Why did we choose these 2 platforms?

There was 2 points that have driven and defined our choice in the use of which platform.
1 - They need to work with a similar purpose but in a different way. If that is not the case, we do not have any basis on which to compare our results.
A point that is respected as Lima focus on rules/dictionaries and Stanford on statistics.
2 - They also need to be public. Tomorrow, if we want to conceive a tool of technology watch, we will be able to use both Lima and Stanford as they are both open-source platforms.

Description of the platforms

CEA List LIMA

LIMA is a fully functional multilingual analyzer with modules and ressources to analyze texts developed by the CEA LIST.
This platform was developed with the following requirements: diverse applications, extensibility, efficiency in an industrial context, tokenization, dictionary check, hyphenated words, abbreviation split alternatives(English only), idiomatic expressions, unknown words, named entities recognition, PoS-tagging, parsing.

Stanford Core NLP

Stanford CoreNLP is an open-source platform developed by Stanford University providing a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.
Stanford CoreNLP's goal is to make it very easy to apply a bunch of linguistic analysis tools to a piece of text. A tool pipeline can be run on a piece of plain text. CoreNLP is designed to be highly flexible and extensible and integrates many of Stanford's NLP tools that we will describe further in the following.

References

[1] Jenny Rose Finkel, Trond Grenager, and Christopher Manning
Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.
http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf
Consulted on 20/02/2019

[2] de Chalendar, Gaël.
The LIMA Multilingual Analyzer Made Free: FLOSS Resources Adaptation and Correction.
https://www.researchgate.net/publication/269408814_The_LIMA_Multilingual_Analyzer_Made_Free_FLOSS_Resources_Adaptation_and_Correction10.13140
Consulted on 20/02/2019

[3] Universal Dependencies contributors.
CoNLL-U Format.
https://universaldependencies.org/format.html
Consulted on 02/03/2019

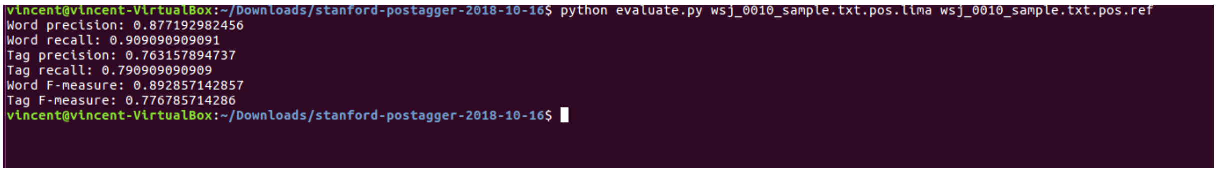
Sommaire

1 Description of the experiments

1.1 Evaluation of the LIMA platform for text analysis

1.1.1 Morpho-Syntaxic part

For a first test, we can try with PTB tags. We launch the following command with the result :

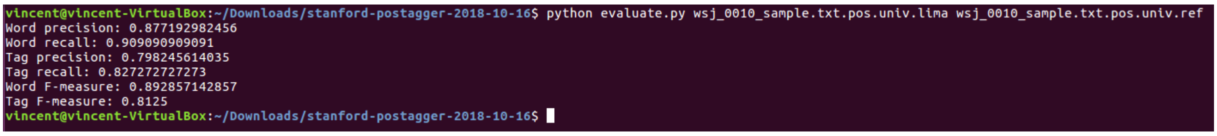


```
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$ python evaluate.py wsj_0010_sample.txt.pos.lima wsj_0010_sample.txt.pos.ref
Word precision: 0.877192982456
Word recall: 0.909090909091
Tag precision: 0.763157894737
Tag recall: 0.790909090909
Word F-measure: 0.892857142857
Tag F-measure: 0.776785714286
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$
```

Only the following lines are interesting for this part : Tag precision, Tag recall, and Tag F-measure.
Tag precision corresponds to the measure of correct tags in the LIMA file, compared to the reference file.
Tag recall is the same as Tag precision : it is just a second try.
Tag F-measure finally, is the average.
So we will focus on the last number, and we have a "precision for tags around 78%", but what does it mean ? Literally, it means that around 1 word for 4 other words has been incorrectly tagged. The word could have been miss-tagged due to many factors, but the most relevant is the one that correspond to the title of this part: the role of the word in the sentence or in the context has been misidentified. For example a proper noun that looks like an adjective or a simple noun leads to an error.

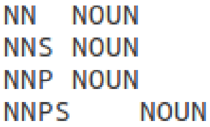
I give you an example to be clearer : If we have a sentence like "*The restaurant Leading room is now open*" : in this sentence Leading room is the name of the restaurant, but it may have been identified by LIMA as Adjective + Simple Noun.

Now that we did a test with PTB tags, we will now do a test on the same LIMA's file, with universal tags. To do that, it required to change PTB tags with the universal equivalent, by using a script. We do the same process with LIMA file : we replace all PTB tags with the Universal tags. Then we can start the test :



```
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$ python evaluate.py wsj_0010_sample.txt.pos.univ.lima wsj_0010_sample.txt.pos.univ.ref
Word precision: 0.877192982456
Word recall: 0.909090909091
Tag precision: 0.798245614035
Tag recall: 0.827272727273
Word F-measure: 0.892857142857
Tag F-measure: 0.8125
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$
```

We now look at lines that are relevant for us, like in the previous test, and the first thing that we notice is : **results are better**.
Previously, we had like 78% of correct words, now we have 81%. So, why do we have improved our results?
If you remember my explanation in the previous point with the PTB tags test, I gave an example about proper nouns, that we assimilate as simple noun with an adjective. Now this error has been half fixed *(adjective error is still here)*, due to the fact that Universal tags unlike PTB tags, are way more simple : there is no difference between the different types of nouns, for example :



```
NN   NOUN
NNS  NOUN
NNP  NOUN
NNPS      NOUN
```

In this context, it is harder to do some mistakes on the morpho-syntaxic part, and it is normal that there is less errors in the second test than the first one.

Now I will give my impression about the result : to be honest, I expected a better improvement on the test with universal tags. I thought that most of the errors made in the first test was due to the fact that LIMA could not differentiate a proper noun and a simple noun. If I take my restaurant example, 50% of the errors are fixed while I hoped for a result similar to 90% of correct tags.

If you ask me why this result, I can't give a short and precise answer, but I have some theories.
First, it could be an error in our scripts : even if we verified that tags have been correctly modified with universal tags, we could have made some mistakes with invisible characters for example.
The second explanation is the source of the error : maybe most of it is not made by the fact that LIMA could not identify precisely the role of the word : for example it analyzed an adjective as verb or a noun as a adverb. In this case, the simplification made by universal tags is useless, an the improvement is not so visible.

c. What does the format of the result of this analysis look like?

The text is decomposed into words associated with an identifier according to its position in the sentence, attribute *(subject, DOC ...)*

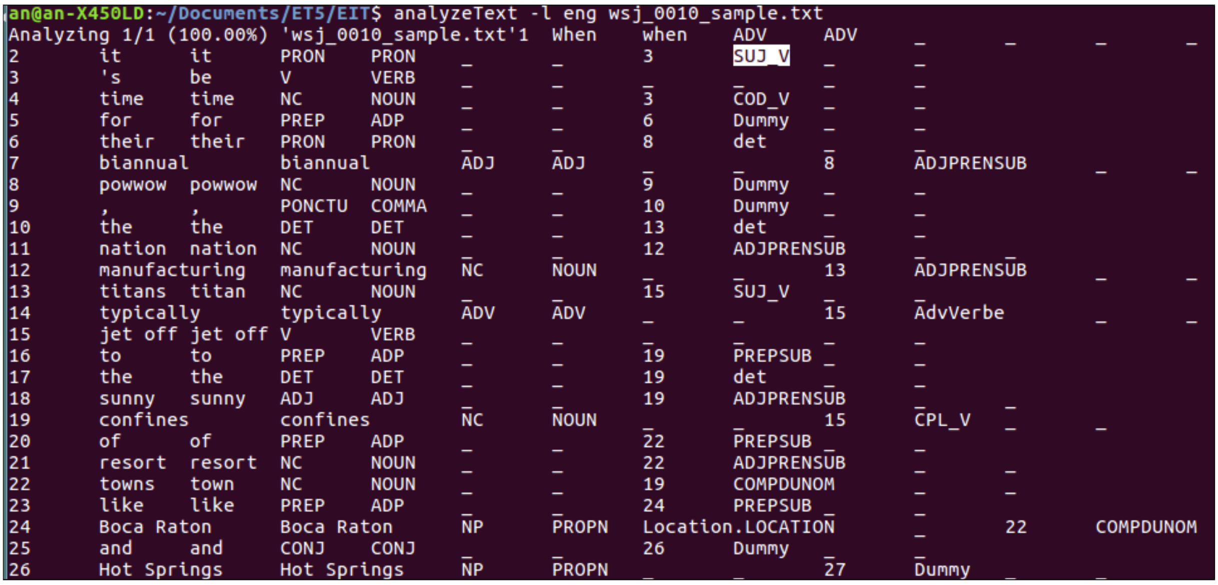
Annotations are encoded in plain text files *(UTF-8, using only the LF character as line break)* with three types of lines:
- Word lines containing the annotation of a word/token in 10 fields separated by single tab characters; see below.
- Blank lines marking sentence boundaries.
- Comment lines starting with hash (#).

Sentences consist of one or more word lines, and word lines contain the following fields:
ID: Word index, integer starting at 1 for each new sentence; may be a range for tokens with multiple words.
FORM: Word form or punctuation symbol.
LEMMA: Lemma or stem of word form.
UPOSTAG: Universal part-of-speech tag drawn from our revised version of the Google universal POS tags.
XPOSTAG: Language-specific part-of-speech tag; underscore if not available.
FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
HEAD: Head of the current token, which is either a value of ID or zero (0).
DEPREL: Universal Stanford dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
DEPS: List of secondary dependencies (head-deprel pairs).
MISC: Any other annotation.

The fields DEPS and MISC replace the obsolete fields PHEAD and PDEPREL of the CoNLL-X format. In addition, we have modified the usage of the ID, FORM, LEMMA, XPOSTAG, FEATS and HEAD fields as explained below.
The fields must additionally meet the following constraints:
Fields must not be empty. Fields must not contain space characters. Underscore is used to denote unspecified values in all fields except ID. Note that no format-level distinction is made for the rare cases where the FORM or LEMMA is the literal underscore processing in such cases is application-dependent. Further, in UD treebanks the UPOSTAG, HEAD, and DEPREL columns are not allowed to be left unspecified.

d. Activate in the file « lima-lp-eng.xml » the loggers "specificEntitiesXmlLogger" and "disambiguatedGraphXmlLogger"
e. start LIMA again on the file and observe the output

After activating the loggers, we can see that 2 files are generated by the system, $wsj_0010_sample.txt.se.xml$ and $wsj_0010_sample.txt.disambiguated.xml$.
Thanks to them, we can apply an XML parsing. In the first one, we can find the entities with their length, position and value and the second one defines the lemma with their macro and micro-categories.

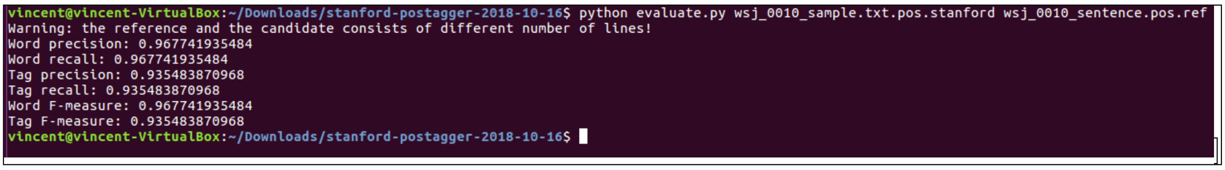


```
an@an-X450LD:~/Documents/ET5/EIT$ analyzeText -l eng wsj_0010_sample.txt
Analyzing 1/1 (100.00%) 'wsj_0010_sample.txt'1  When    when    ADV     ADV     _       _       _       _
2       it      it      PRON    PRON    _       _       3       SUJ_V   _       _
3       's      be      V       VERB    _       _       _       _       _       _
4       time    time    NC      NOUN    _       _       3       COD_V   _       _
5       for     for     PREP    ADP     _       _       6       Dummy   _       _
6       their   their   PRON    PRON    _       _       8       det     _       _
7       biannual        biannual        ADJ     ADJ     _       _       8       ADJPRENSUB      _       _
8       powwow  powwow  NC      NOUN    _       _       9       Dummy   _       _
9       ,       ,       PONCTU  COMMA   _       _       10      Dummy   _       _
10      the     the     DET     DET     _       _       13      det     _       _
11      nation  nation  NC      NOUN    _       _       12      ADJPRENSUB      _       _
12      manufacturing   manufacturing   NC      NOUN    _       _       13      ADJPRENSUB      _       _
13      titans  titan   NC      NOUN    _       _       15      SUJ_V   _       _
14      typically       typically       ADV     ADV     _       _       15      AdvVerbe        _       _
15      jet off jet off V       VERB    _       _       _       _       _       _
16      to      to      PREP    ADP     _       _       19      PREPSUB _       _
17      the     the     DET     DET     _       _       19      det     _       _
18      sunny   sunny   ADJ     ADJ     _       _       19      ADJPRENSUB      _       _
19      confines        confines        NC      NOUN    _       _       15      CPL_V   _       _
20      of      of      PREP    ADP     _       _       22      PREPSUB _       _
21      resort  resort  NC      NOUN    _       _       22      ADJPRENSUB      _       _
22      towns   town    NC      NOUN    _       _       19      COMPDUNOM       _       _
23      like    like    PREP    ADP     _       _       24      PREPSUB _       _
24      Boca Raton      Boca Raton      NP      PROPN   Location.LOCATION       _       22      COMPDUNOM
25      and     and     CONJ    CONJ    _       _       26      Dummy   _       _
26      Hot Springs     Hot Springs     NP      PROPN   _       _       27      Dummy   _       _
```

1.2 Evaluation of the morphosyntactic disambiguation tool from Stanford University

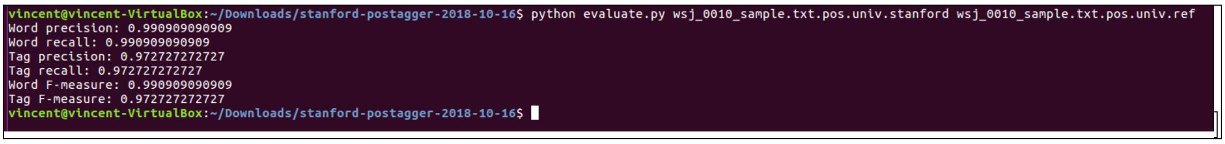
In this part, we'll use a tool provided by Stanford University. This tool works on the morpho-syntaxic part.

After installation, we start with an evaluation on this tool. We have a text file called *wsj_0010_sample.txt* . First, we launch the POS tagger on this file to get a first impression.



```
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$ python evaluate.py wsj_0010_sample.txt.pos.stanford wsj_0010_sentence.pos.ref
Warning: the reference and the candidate consists of different number of lines!
Word precision: 0.967741935484
Word recall: 0.967741935484
Tag precision: 0.935483870968
Tag recall: 0.935483870968
Word F-measure: 0.967741935484
Tag F-measure: 0.935483870968
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$ 
```

According to the results, the precision of this POS Tagger by using PTB tags is actually pretty good : more than 9 out of 10 word's role are recognized. For an other experiment, this time we use Universal tags instead of PTB ones. Before all, we have to create the a file called *wsj_0010_sample.txt.pos.univ.stanford.* To create it, we can use script used to transform the file *wsj_0010_sample.txt.pos.lima* into *wsj_0010_sample.txt.pos.univ.lima.* We just need to adapt an element, due to the fact that Stanford file contains many paragraphs (and LIMA one does not). Then we launch the test :



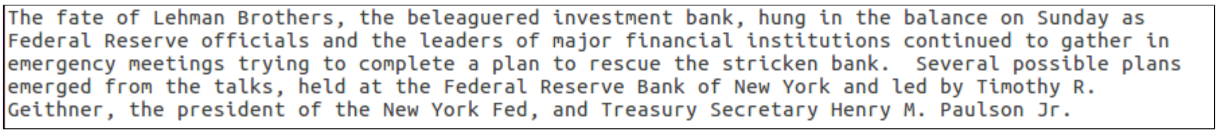
```
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$ python evaluate.py wsj_0010_sample.txt.pos.univ.stanford wsj_0010_sample.txt.pos.univ.ref
Word precision: 0.990909090909
Word recall: 0.990909090909
Tag precision: 0.972727272727
Tag recall: 0.972727272727
Word F-measure: 0.990909090909
Tag F-measure: 0.972727272727
vincent@vincent-VirtualBox:~/Downloads/stanford-postagger-2018-10-16$ 
```

The improvement is very significant and now with universal tags, the Stanford tool does almost no mistakes on tags. So where this improvement come from, and why it exists ?
As we detailed in Lima for the same test, the main reason for this improvement is this one : the stanford tool made some mistake about the role of a word **precisely** : for example, Stanford tool detects that we have a name and put the tag NN, but it was a mistake, and the correct tag was NNP. In an other example, the tool imagines that a word verb a give it the tag VB, but it was VBZ. With the simplification of tags when we come from PTB tags to universal tags, there is no more mistakes like these.

To finish I would conclude on the results for both platforms on this part on analysis: as we can see Stanford has better performance than LIMA. Both of them have better performances when we use universal tags instead of PTB tags. But we can also notice that the performance improvement is better for the Stanford tool (around 6%), than LIMA (around 3%). It means that these two tools have not the same way to process for analysis, and Stanford tool is more adapted to use process with universal tags.

1.3 Evaluation of named entity recognition tools from the CEA List and Stanford University

For this part, we will evaluate both platforms to recognize entities with this corpus :

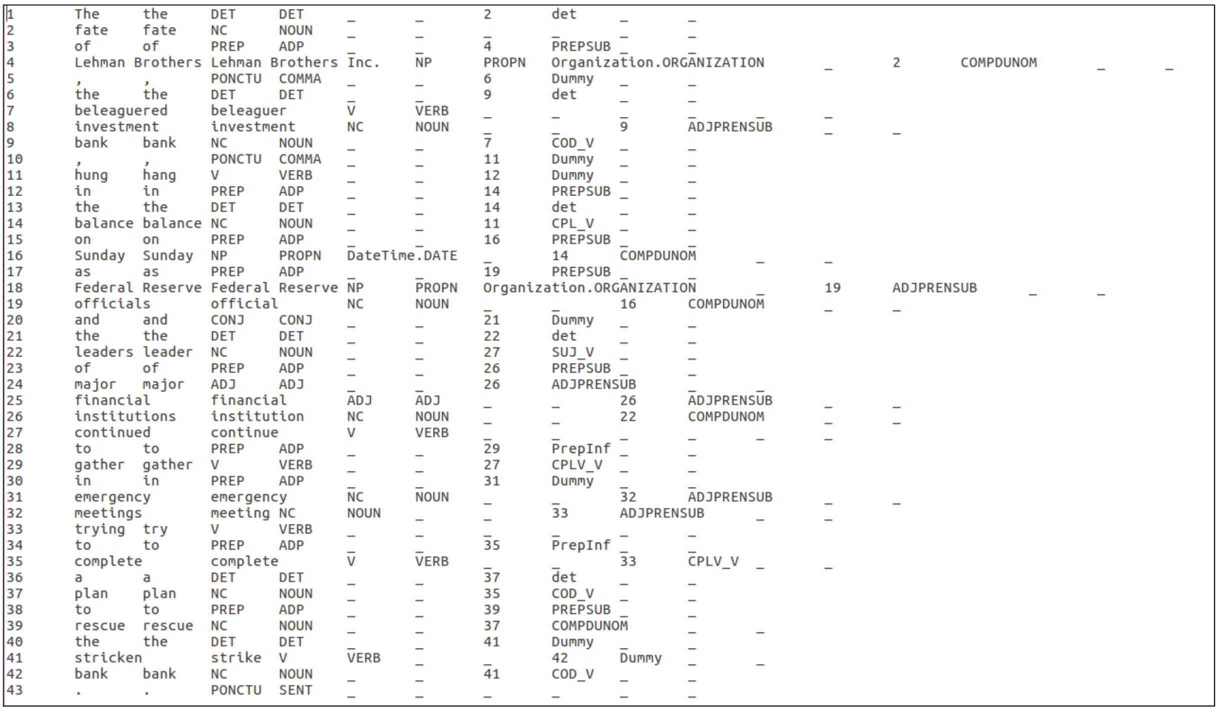


```
The fate of Lehman Brothers, the beleaguered investment bank, hung in the balance on Sunday as
Federal Reserve officials and the leaders of major financial institutions continued to gather in
emergency meetings trying to complete a plan to rescue the stricken bank.  Several possible plans
emerged from the talks, held at the Federal Reserve Bank of New York and led by Timothy R.
Geithner, the president of the New York Fed, and Treasury Secretary Henry M. Paulson Jr.
```

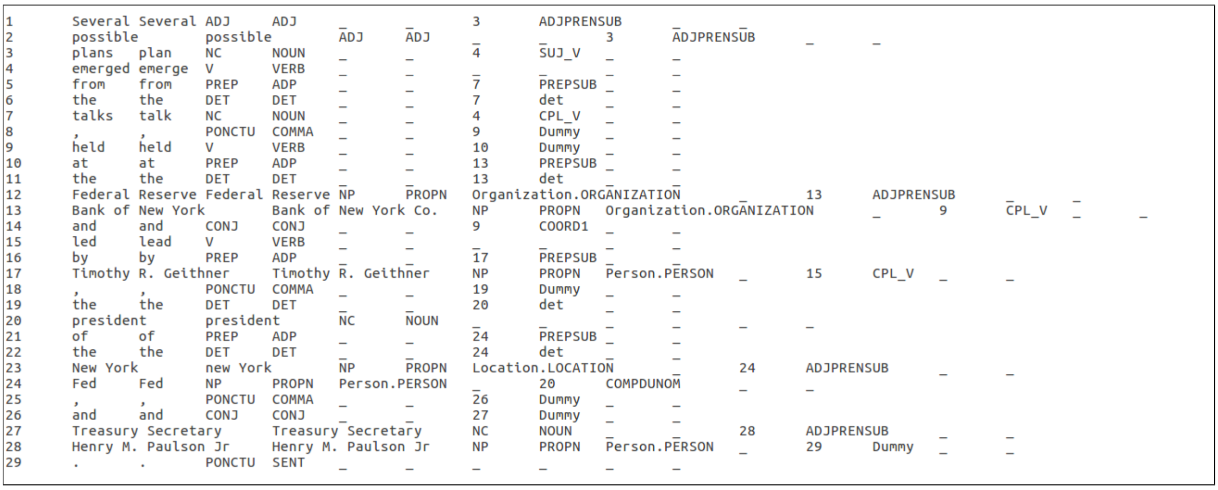
We chose this text because it contains many proper nouns : organizations, locations and persons. Making it a good choice to test our tools performance.

1.4 CEA List

Here is the result for CEA List tool:



```
1       The     the     DET     DET     _       _       2       det     _       _
2       fate    fate    NC      NOUN    _       _       _       _       _       _
3       of      of      PREP    ADP     _       _       4       PREPSUB _       _
4       Lehman Brothers Lehman Brothers Inc.    NP      PROPN   Organization.ORGANIZATION       _       2       COMPDUNOM       _       _
5       ,       ,       PONCTU  COMMA   _       _       6       Dummy   _       _
6       the     the     DET     DET     _       _       9       det     _       _
7       beleaguered     beleaguer       V       VERB    _       _       _       _       _       _
8       investment      investment      NC      NOUN    _       _       9       ADJPRENSUB      _       _
9       bank    bank    NC      NOUN    _       _       7       COD_V   _       _
10      ,       ,       PONCTU  COMMA   _       _       11      Dummy   _       _
11      hung    hang    V       VERB    _       _       12      Dummy   _       _
12      in      in      PREP    ADP     _       _       14      PREPSUB _       _
13      the     the     DET     DET     _       _       14      det     _       _
14      balance balance NC      NOUN    _       _       11      CPL_V   _       _
15      on      on      PREP    ADP     _       _       16      PREPSUB _       _
16      Sunday  Sunday  NP      PROPN   DateTime.DATE   _       14      COMPDUNOM       _       _
17      as      as      PREP    ADP     _       _       19      PREPSUB _       _
18      Federal Reserve Federal Reserve NP      PROPN   Organization.ORGANIZATION       _       19      ADJPRENSUB      _       _
19      officials       official        NC      NOUN    _       _       16      COMPDUNOM       _       _
20      and     and     CONJ    CONJ    _       _       21      Dummy   _       _
21      the     the     DET     DET     _       _       22      det     _       _
22      leaders leader  NC      NOUN    _       _       27      SUJ_V   _       _
23      of      of      PREP    ADP     _       _       26      PREPSUB _       _
24      major   major   ADJ     ADJ     _       _       26      ADJPRENSUB      _       _
25      financial       financial       ADJ     ADJ     _       _       26      ADJPRENSUB      _       _
26      institutions    institution     NC      NOUN    _       _       22      COMPDUNOM       _       _
27      continued       continue        V       VERB    _       _       _       _       _       _
28      to      to      PREP    ADP     _       _       29      PrepInf _       _
29      gather  gather  V       VERB    _       _       27      CPLV_V  _       _
30      in      in      PREP    ADP     _       _       31      Dummy   _       _
31      emergency       emergency       NC      NOUN    _       _       32      ADJPRENSUB      _       _
32      meetings        meeting NC      NOUN    _       _       33      ADJPRENSUB      _       _
33      trying  try     V       VERB    _       _       _       _       _       _
34      to      to      PREP    ADP     _       _       35      PrepInf _       _
35      complete        complete        V       VERB    _       _       33      CPLV_V  _       _
36      a       a       DET     DET     _       _       37      det     _       _
37      plan    plan    NC      NOUN    _       _       35      COD_V   _       _
38      to      to      PREP    ADP     _       _       39      PREPSUB _       _
39      rescue  rescue  NC      NOUN    _       _       37      COMPDUNOM       _       _
40      the     the     DET     DET     _       _       41      Dummy   _       _
41      stricken        strike  V       VERB    _       _       42      Dummy   _       _
42      bank    bank    NC      NOUN    _       _       41      COD_V   _       _
43      .       .       PONCTU  SENT    _       _       _       _       _       _
```

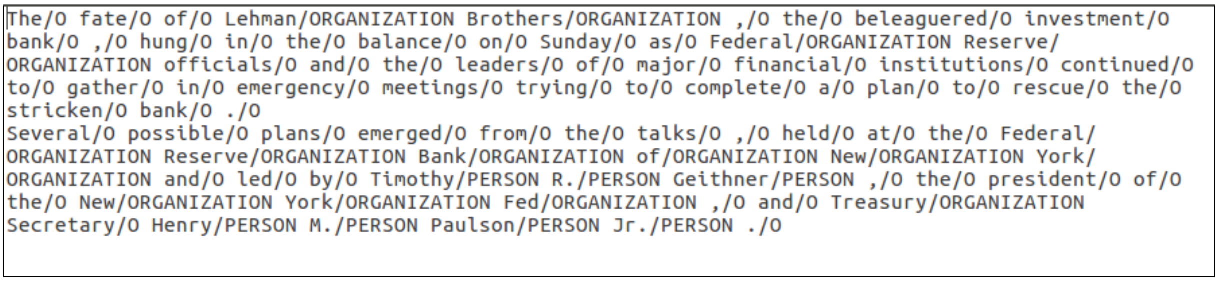


```
1       Several Several ADJ     ADJ     _       _       3       ADJPRENSUB      _       _
2       possible        possible        ADJ     ADJ     _       _       3       ADJPRENSUB      _       _
3       plans   plan    NC      NOUN    _       _       4       SUJ_V   _       _
4       emerged emerge  V       VERB    _       _       _       _       _       _
5       from    from    PREP    ADP     _       _       7       PREPSUB _       _
6       the     the     DET     DET     _       _       7       det     _       _
7       talks   talk    NC      NOUN    _       _       4       CPL_V   _       _
8       ,       ,       PONCTU  COMMA   _       _       9       Dummy   _       _
9       held    held    V       VERB    _       _       10      Dummy   _       _
10      at      at      PREP    ADP     _       _       13      PREPSUB _       _
11      the     the     DET     DET     _       _       13      det     _       _
12      Federal Reserve Federal Reserve NP      PROPN   Organization.ORGANIZATION       _       13      ADJPRENSUB      _       _
13      Bank of New York        Bank of New York Co.    NP      PROPN   Organization.ORGANIZATION       _       9       CPL_V   _       _
14      and     and     CONJ    CONJ    _       _       9       COORD1  _       _
15      led     lead    V       VERB    _       _       _       _       _       _
16      by      by      PREP    ADP     _       _       17      PREPSUB _       _
17      Timothy R. Geithner     Timothy R. Geithner     NP      PROPN   Person.PERSON   _       15      CPL_V   _       _
18      ,       ,       PONCTU  COMMA   _       _       19      Dummy   _       _
19      the     the     DET     DET     _       _       20      det     _       _
20      president       president       NC      NOUN    _       _       _       _       _       _
21      of      of      PREP    ADP     _       _       24      PREPSUB _       _
22      the     the     DET     DET     _       _       24      det     _       _
23      New York        new York        NP      PROPN   Location.LOCATION       _       24      ADJPRENSUB      _       _
24      Fed     Fed     NP      PROPN   Person.PERSON   _       20      COMPDUNOM       _       _
25      ,       ,       PONCTU  COMMA   _       _       26      Dummy   _       _
26      and     and     CONJ    CONJ    _       _       27      Dummy   _       _
27      Treasury Secretary      Treasury Secretary      NC      NOUN    _       _       28      ADJPRENSUB      _       _
28      Henry M. Paulson Jr     Henry M. Paulson Jr     NP      PROPN   Person.PERSON   _       29      Dummy   _       _
29      .       .       PONCTU  SENT    _       _       _       _       _       _
```

As we can see, there is some mistakes made by the tool.
For example, for the part "*the New York fed*" , it considers New York as a location, instead of an institution. It also classifies "*The federal reserve bank of New York*" as two institutions : '*federal reserve*" and "*bank of New York*".

1.5 Stanford

Here is the result for stanford tool :



```
The/O fate/O of/O Lehman/ORGANIZATION Brothers/ORGANIZATION ,/O the/O beleaguered/O investment/O
bank/O ,/O hung/O in/O the/O balance/O on/O Sunday/O as/O Federal/ORGANIZATION Reserve/
ORGANIZATION officials/O and/O the/O leaders/O of/O major/O financial/O institutions/O continued/O
to/O gather/O in/O emergency/O meetings/O trying/O to/O complete/O a/O plan/O to/O rescue/O the/O
stricken/O bank/O ./O
Several/O possible/O plans/O emerged/O from/O the/O talks/O ,/O held/O at/O the/O Federal/
ORGANIZATION Reserve/ORGANIZATION Bank/ORGANIZATION of/ORGANIZATION New/ORGANIZATION York/
ORGANIZATION and/O led/O by/O Timothy/PERSON R./PERSON Geithner/PERSON ,/O the/O president/O of/O
the/O New/ORGANIZATION York/ORGANIZATION Fed/ORGANIZATION ,/O and/O Treasury/ORGANIZATION
Secretary/O Henry/PERSON M./PERSON Paulson/PERSON Jr./PERSON ./O
```

As we can see, all word that need to receive a tag got it. So the tool is, in that aspect, very effective. We can also notice some choices made by Stanford : For the entity "*Federal Reserve Bank of New York*", New York is considered as an Organization and not a Location, due to the fact that before New York, we have Federal Reserve Bank, with an article to link both terms.

2 Conclusion and Perspective

2.1 Limitations

The first limitation that came to mind is the data we used for this project, especially the text files. Even if the text was relatively complex and implement many aspects of the English language, only one text is not enough to make a good opinion on these tools. Training on a few complex parts of the language like morpho syntax is not enough, and many aspects of syntax have not been covered.

2.2 Proposals to enhance platforms performances

Performances for both platforms, are currently quite good, even if Stanford tool is better than LIMA. But the two platforms that we tested are not perfect, and there is still some work that can be done to get better performances :

- **Upgrade the scope**. I think that a way to improve performances is to take more elements into account to determine the role of a word.
  For example you can introduce an element at the beginning of a text, at the middle of the text or at the end, you can also have some references to this element and you can't guess the role of the world if you did not take into account the introduction of this word at the beginning. That is why these tools should use something similar to a memory to register words. Then tools would be able to avoid "*traps*" like some type of references and proper nouns. It could also be useful to get a better recognition for organizations, Locations and persons (especially for CEA List tool).

3 Division of the work

3.1 Main organization

For the realization of this project, we divided the project in independent tasks for each member as the subject was quite long and also because the tools were not working on every computer.

As such, Anais and Vincent handled the testing of the TAL tools while Florian was focusing on the report.

3.2 Details on the contribution

Vincent

After installing the different tools, I started to work on the stanford tool. I evaluated the performance of this tool. To do that, I needed to create scripts used to change PTB tags into universal tags. After that, I did the same process with Lima platform. Then, I worked on the part 4 : I installed tools, and I created the script used to represent data in the given format. Finally, I compared CEA List and Stanford tool for Entity recognition.

Anais

During this project, I was focused on the LIMA platform for text analysis. So I needed to install the tools to be able to work on the morpho-syntaxic part. After that, I worked on the extraction of named entities as well as the analysis of the results.

Florian

I started this project by trying to install the tools but some libraries were not working on my computer and as we were losing time and someone had to handle the report, I volunteered to take care of it.

Therefore, my primary focus during the project has been to handle the results provided by my colleagues to present the data in our report and make some thoughts on these results.

I also helped my teammates for the analysis part when possible, for example, by creating the script in python parsing the xml output from lima.