Evaluation and interoperable annotations: the point of view of a parser developer

Gil FRANCOPOULO
TAGMATICA
126 rue de Picpus
75012 PARIS FRANCE
gil.francopoulo@wanadoo.fr

Abstract

The present paper is written within the framework of the French ANR-Passage project that gathers ten parser developers. The main motivations of the project are to evaluate parsers for French, to test their accuracy and robustness on large scale corpora and then to combine the resulting annotations to create a richer and more extensive linguistic resource.

1 Introduction

The present paper is written within the framework of the French ANR-Passage project that gathers ten parser developers[1]. The main motivations of the project are first to evaluate parsers for French, to test their accuracy and robustness on large scale corpora and secondly to combine the resulting annotations to create a richer and more extensive linguistic resource.

The following text does not present any prospective view, nor any critical appraisal because it is obviously too early to judge the results of the project. This is more the point of view of a parser developer whose work is going to be compared with the one of nine others in a few weeks from now.

1 http://atoll.inria.fr/passage

2 Evaluation

2.1 Evaluation campaigns before Technolangue-Easy

The goals of NLP evaluation are to measure one or more qualities of a system, in order to determine to what extend the system answers the needs of its users. Evaluation has received considerable attention, because the definition of a proper evaluation is one way to specify precisely an NLP problem, going thus beyond the vagueness of tasks defined as *language understanding* or *language translation*.

The first evaluation campaign on written texts seems to be a campaign dedicated to message understanding in 1987 (Pallet 1998). Then the Parseval/GEIG project compared phrase-structure grammars (Black 1991). A series of campaigns within Tipster project were realized on tasks like summarization, translation and searching (Hirshman 1998). In 1994, in Germany, the Morpholympics compared German taggers. Then, the Senseval[2] and Romanseval[3] campaigns were conducted with the objectives of semantic disambiguation. In 1996, the Sparkle campaign compared syntactic parsers in four different languages (English, French, German and Italian)[4]. In France, the Grace project compared a set of 21 taggers for French in 1997 (Adda 1999). There were also the large-scale evaluation of dependency parsers performed in the

2 http://www.senseval.org
3 http://aune.lpl.univ-aix.fr/projects/romanseval
4 http://www.ilc.cnr.it/sparkle/wp1-prefinal/wp1-prefinal.html

context of the CoNLL shared tasks in 2006 and 2007.

2.2 Background: Technolangue-Easy

The Easy project was a sub-project of a larger project whose name was Technolangue (Paroubek 2007)[5]. Easy was dedicated to the evaluation of syntactic parsers for French language from 2003 until 2006. The aim was to design and test an evaluation methodology to compare 14 parsers, that means that most parsers for French were gathered and evaluated.

The project produced the following results for the community:

- **An annotation guideline**. This document has been established after a consensual study by a committee of 30 French experts in syntax including the parser developers. The work started from the PEAS specification written in 2003[6]. In this document, two types of elements are accurately specified: i) a decomposition in constituency that defines a flat series of chunks and ii) the relations between these chunks. Six types of chunks are defined: NP, PP, VP, AdjP, AdvP and VprepP (verbal preposition phrase). And fourteen relations are specified: subject, auxiliary, direct-object, verbal complement, verbal modifier, complement, attribute, noun modifier, adjective modifier, adverb modifier, preposition modifier, coordination, apposition and juxtaposition. Let us add that this work is rather specific to the given language, i.e. French, and cannot be easily translated into another language, on the contrary of the rest of the material presented in the current paper.

- **Annotation and distribution of a small corpus** comprising 76K words that served as a gold-standard. These annotations has been inserted manually by five different annotators. The texts are categorized into six different styles:
  - general genre from newspapers, parliamentary minutes and institutional web sites;
  - literary genre from a selection of 19th century novels;
  - email genre after anonymization;
  - medicine genre;
  - transcription of oral street dialogues;
  - questions taken from TREC and AMARYLLIS campaigns.

  Let us note that the email and oral corpora contain a lot of faulty expressions with respect to the French normalized way of expression[7].

- **Quantified results**. First, 150 examples of sentences were distributed to the parser developers in order to let them train and adapt their input and output software modules. Then, the evaluation corpus was given to the competitors who **had one week** to run their parser and to produce their results. In order to avoid any cheat by means of manual annotation during this week, the various sentences of the evaluation corpus were inserted randomly inside a larger corpus comprising 1M words. And obviously, the competitors did not know which sentences were going to be measured. Then, the organizers collected the results and ran a batch of programs to compute the recall and precision, as presented in (Paroubek 2007). Let us note that due to the fact that the texts were rather different and categorized as such, it was possible for each parser to have a separate quantification for each type of text. And indeed, a parser that is good in general genre is not necessarily good on email genre, and vice versa.

- **Campaign know-how**. Both sides (organizer and competitors) learnt a lot.

- **Scientific and social networking.** One of the most important side effects of the campaign is that for the first time, all French parser developers talked to each other. The developers were gathered with common objectives in mind like annotation guidelines specification, quantified results, discussion about parsing algorithms, problems of time

[5] http://www.technolangue.net
[6] www.limsi.fr/Recherche/CORVAL/easy

[7] The standard for French is usually considered as being "Le Bon Usage" from Maurice Grévisse (a belgian linguist). We have a lexicon and a grammar that are published by the "Académie Française" but last editions are dated 1932, so they are considered as obsolete.

and/or space limitations, maintenance of grammars and so on.

A certain number of problems were encountered during this campaign as well on the organizer side as on the competitor side. On the organizer side, different problems were encountered like:

- **Segmentation**. Two types of segmentation were done: for the sentences in the text, and for the words in the sentence. For the latter, a list of grammatical multi-word entries was determined. Alas the segmentation program was not perfect.
- **Manual annotation errors.** When the annotations started, the guidelines were not precise enough, and the annotators took different choices, and thus, made what we consider now as mistakes. Another point to mention is that, during this annotation process, the tools (at the time being) did not have enough strong type checking and a certain number of errors could have been avoided.
- **Physical formats incompatibilities**. A certain number of participants delivered badly XML formated results. The organizers tried to fix these problems automatically, but some problems still remained.

On the parser competitor side, it should be noted that a small number of them were ready, but most of them were not ready to compete. And some of them discovered severe bugs during the evaluation week and produced badly formated results and/or partial outputs. The most difficult task was to follow the annotation guidelines. Two different strategies were adopted. The first strategy was to adapt the result by means of a translation post-processing. On the contrary, some participants [8] estimated that the guidelines were well defined (at least, compared to their current grammar) and changed their core grammar in order to deeply implement the guidelines rules.

On the other hand, even if the quantified results are now difficult to interpret due to the various difficulties encountered on both sides, Easy was a very successful project. And assuredly, Easy was a good technical foundation for ANR-Passage.

3 Annotations

Different types of annotations are encountered and managed in order to build a parser.

3.1 Hand-coded annotated corpora

Hand-coded annotated corpora are generally of small size because human annotation is heavily time consuming. But, if the task is done conscientiously, possibly verified by a cross-checking with different annotators, the data are usually very reliable. A parser obtained automatically from such a small corpus will have some difficulties against large scale corpus.

3.2 Automatic annotation

It is now quite easy to obtain large collections of French raw texts because of the world wide web. It is not very difficult to parse them because of computer speed increase and a better knowledge of parsing optimizations. But as ambiguity is pervasive in natural languages, the border between what is unambiguous (and so reliable) and what is ambiguous (and so unreliable) cannot be determined easily without any external annotation. Nevertheless, a limited number of phenomena may be acquired. For instance, concerning the problem of PP attachment ambiguity, a probability may be computed from unambiguous configurations [9] in order to detect the most likely attachment (Bourigault 2005). Due to the fact that an information is selected in a very precise situation, the size of the corpus must be huge, to be reliable.

Aside from numerical factors, to be used in some limited situations, these results do not give us any broad insight.

3.3 Automatic annotation followed by a manual correction

Supposing that a minimal parser is available, a parse is applied, then possibly after some filtering operations, the annotation is manually checked and

[8] Personally, I took this strategy.

[9] Like a subject configuration NP+PP+VerbalKernel (without any comma between NP and PP) in the beginning of a sentence, where PP is attached to NP in a reliable way. Contrary to a sequence VerbalKernel+NP+PP where PP may be a modifier of NP or a complement of the verb.

fixed, see for instance, the work of Abeillé and al (Abeillé 2003).

The cost is much lower compared to a fully hand-coded process because all regular configurations are automatically resolved. In fact, only problematic configurations are manually fixed. The problem is that this tactic does not function very well when the corpus is large and when the quality of the parsing is in the range of a F-score of 85%-100%, because in this case it is very difficult to detect wrong parsing results within a huge corpus.

Some experiments like (Arun 2005) have been made to induce a parser, but the corpus was limited to the newspaper "Le Monde" and we don't know how such a parser could behave in an evaluation[10] like Easy where the genres are rather diverse.

3.4 Dynamic annotation selection

Another strategy is to adopt a more dynamic behavior. Instead of considering that the parser is just a sort of pre-processing that is not applied anymore, it is possible to improve the parsing process through a series of incremental steps.

A tiny hand-coded corpus is used to serve as a bootstrap to induce an initial parser by means of a machine learning algorithm. This parser is applied to a raw corpus. Problematic sentences are automatically collected and ranked, starting from the simplest ones. And then, manually, a small set of sentences is annotated and added to the hand-coded corpus. A new version of the parser is induced. The system then iterates for another step.

Step by step, the hand-coded corpus becomes a set of difficulties for the given language. Experiments on a 82M words corpus (made of different genres) showed that a hand-coded corpus smaller than 100K words produces a coverage of 96%[11] for a language like French (Francopoulo 2008).

Another strategy is to combine different parsing results, as presented below.

[10] In fact, there is no parser of this type among the competitors.

[11] The coverage is defined as the ability to produce a result on text that is seen for the first time. Only the most frequent difficulties have been (hand-)coded, this is why the coverage is not 100%.

4 ANR-Passage as a bridge between evaluation and annotation

Passage is a project founded by the new French research agency whose name is "Agence Nationale de la Recherche"[12]. ANR-Passage is built on the results of the Easy project.

The main motivations for ANR-Passage are:

- to evaluate French parsers by means of two campaigns. One campaign is currently launched during the first year of the project, i.e. Fall 2007, as a close rerun of the Easy campaign[13]. And the second campaign will be launched at the end of the project in 2009 in order to measure the improvement.
- to improve the accuracy and robustness of French parsers on large scale corpora of 270M words;
- to exploit the resulting syntactic annotations to create a richer and more extensive linguistic resource: a treebank for French.

The adopted methodology consists of a feedback loop between parsing and resource creation as follows:

- step#1: ten different parsers coming from both the public sector and the private sector are applied to create syntactic annotations;
- step#2: a certain number of annotations are selected and combined on the basis of a quantified evaluation that is similar to a ROVER[14];
- step#3: these annotations are then used to improve linguistic resources like lexicons, grammars and annotated corpora;
- step#4: these resources are integrated into existing parsers;
- return to step#1

[12] http://www.agence-nationale-recherche.fr

[13] In order to avoid the same problems as during the Easy campaign, the badly segmented sentences are excluded and a phase of error correction is undertaken. In order to improve the efficiency of the partners, a collaborative system called "EasyRef" has been developed (by Eric de la Clergerie) and set up for bug reporting. Let us add also that a new set of sentences is added to the gold standard, and the participants do not know this new set.

[14] Recognizer output voting error reduction

As a consequence, it is not obvious that a parser which is the best at the beginning of the project will be the best at the end of the project. Various parameters are to be taken account, like:

- propensity for a parser to smoothly integrate new knowledge;
- psychological motivation of the parser developer to improve and check his system;
- and obviously, the starting point of the parser in the loop.

5 Annotation combination

Annotation combination may be viewed according to two different levels:

- **physical format level.** This is usually a major issue when the aim is to merge annotations coming from different sources. For us, it is not a problem because all parsers must share the same annotation scheme based on the emerging ISO TC37 SC4 standards. Forms should be built upon tokens referring spans of the original documents through standoff notation, following the *Morphosyntactic Annotation Framework* [MAF] proposal (Clément 2005). Besides chunks, constituency should be completed by allowing nested recursive groups as proposed in the *Syntactic Annotation Framework* [SynAF], following the TIGER model (Declerck 2006). And this conformance will be automatically checked.
- **annotation guidelines conformance.** SynAF conformance is not enough. All parsers must respect the annotation guidelines. With this respect, the quality of the specifications is essential. For simple situations, there is no doubt, but certain configurations are not so clear cut. For instance, does the following fragment to be parsed into one or two NPs: "Le député Robert Dupont …"?

It is of course possible to compare the result of ten parsers and select the majority vote. There is a high probability that the majority classification on average outperforms any single system, because it eliminates random errors in individual systems, as shown in the machine learning literature, see for instance (Henderson 1999). But where is the evidence that the majority result is the best one? Is there a rationale behind this? The majority classification may be incorrect and may hide a minority subset which is correct. One should note also that a ballot based system may be biased by the fact that some parsers share the same lexicon.

Let us recall that we know the score of each individual system against the gold standard, and we know this for each genre of text and for each syntactic structure and relation. We will try to use this information in order to be more precise than a naïve majority classification.

Another important aspect to mention is the engineering rationale. We could think that the parsers will make the same errors, most probably based on the same engineering motivations. But this is not sure because the parsers are rather different, so the engineering motivations are different. Most commercial systems are rule based or corpus based. Most public research systems are lexicon driven within a declarative framework.

Most of the parsers being rather mature, there is a high probability that easy and frequent linguistic phenomena will be correctly processed by all competitors. On the contrary, the discrepancy between results will be limited to phenomena that are hard to proceed and rarely encountered.

Taking for granted this hypothesis, the larger discrepancies will deal at a minimum with:

- **multi-word expressions** that are irregular from the point of view of syntax. Most of the parsers do not have the same lexicon. Thus, some parsers will detect the multi-word expression as such, and others will have serious difficulties to build a regular syntactic structure.
- **particularities** concerning the sub-categorization frame of specific words. Here again, the accuracy and coverage of the lexicon will make a difference.
- **long distance attachments.** Some parsers are good at this task while others are much less accurate.
- **coordination** processing capability.
- **error correction.** Some parsers have a good corrector while others do not have any.
- **propensity to parse lengthy sentences.** In French, we have frequently more than 50

words in a sentence[15], particularly in newspapers and institutional texts.

- **coverage.** Let us recall that in French, for certain phrases like noun phrases, the number of combinations of determiners, adjectives, adverbs and nouns is rather important.
- **propensity** to deal with mechanisms which cross sentence boundaries like inserted discourses.
- **semantic information.** Some lexicons have semantic markers while others do not have any. In certain situations, this makes a difference.

Which result will be endorsed?

6 Interoperable annotations

As said before, one of the objectives of the project is to create a rich and extensive linguistic resource: a treebank for French. We don't know yet if we will deliver the full 270M words corpus or only a reliable subset.

By the way, the delivered corpus will respect MAF (ISO 24611) and SynAF (ISO 24615) for the structure of the annotations. As specified in these two ISO documents, all these structures will be adorned by data category values to be taken from the ISO data category registry, following the work of the two ISO thematic domain groups: the first for morphosyntactic values (like /feminine/) and the second for syntactic values (like /subject/) (Ide 2004).

Following explicit and internationally certified standards seems to be the only way to have interoperable annotations. We just don't know any other way to proceed.

7 Conclusion

It is important to gather in a same project both i) a treebank building and ii) a parsing evaluation. We could have considered a parsing evaluation without treebank building. But we could not have considered treebank building without parsing evaluation.

In fact, it is not important to know if such and such developer obtains a good score. The most important point is to determine if a given technology gives good results, possibly taking into account the style of text.

To be reliable, such a comparison could consider another parameter: this the number of man-years spent in the parser building. But this information is almost impossible to obtain because, aside from problems of confidentiality for private partners, the time spent in lexicons and software modules, is very difficult to estimate.

Knowing whether a technology is better that another one is beyond the reach today. We all hope to have more accurate information at the end of the project.

To conclude, I just wanted to say that from my personal point of view as a developer of a parser, namely TagParser (Francopoulo 2005, 2008), the Easy campaign has been very fruitful. And, at the level of French scientific community, we must not forget the psychological aspect: it is unquestionable that these evaluations gave a serious boost to syntactic parsing in France, thanks to the spirit of a fair competition.

Acknowledgements

This work was supported in part by the EU (eContent project 22236 LIRICS[16]) and in part by the French ANR-Passage project (Action ANR-06 MDCA-013).

References

Abeillé A., Clément L., Toussenel F. 2003 Building a treebank for French, in Treebanks: building and using parsed corpora, Abeillé ed, Kluwer Dordrecht

Adda G., Mariani J., Paroubek P., Rajman M. 1999 L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. Langues vol-2

Arun A., Keller F. 2005 Lexicalization in Crosslinguistic Probabilistic Parsing: the case of French. ACL Ann Arbor MI

Black E., Abney S., Flickinger D., Gdaniec C., Grishman R., Harrison P., Hindle D., Ingria R., Jelinek F., Klavans J., Liberman M., Marcus M., Reukos S.,

[15] Concerning this point, it is very surprising as a French native speaker to see that research on English parsers and machine learning applications for English are frequently limited to sentences shorter than 40 words. This raises the question of the transfer of certain methods from English to French (and languages with lengthy sentences).

16 http://lirics.loria.fr

Santoni B., Strzalkowski T. 1991 A procedure for quantitatively comparing the syntactic coverage of English grammars. DARPA Speech and Natural Language Workshop

Bourigault D., Frérot C. 2005 Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. TALN Dourdan

Clément L., Villemonte de la Clergerie E. 2005 MAF: a morphosyntactic annotation framework. 2nd Language & Technology Conference. Poznan. Poland

Declerck T. 2006 SynAF: towards a standard for syntactic annotation. LREC Genoa

Francopoulo G. 2005 TagParser et Technolangue-Easy TALN Dourdan

Francopoulo G. 2008 TagParser: well on the way to ISO-TC37 conformance. ICGL Hong Kong

Henderson J., Brill E. 1999 Exploiting diversity in Natural Language Processing: combining parsers. Fourth Conference on Empirical Methods in NLP. College Park MD

Hirshman L. 1998 Language understanding evaluation: lessons learned from MUC and ATIS. LREC Granada

Ide N., Romary L. 2004 A registry of standard data categories for linguistic annotation. LREC Lisbon

Pallet D.S. 1998 The NIST role in automatic speech recognition benchmark tests. LREC Granada

Paroubek P., Vinat A., Robba I., Ayache C. 2007 Les résultats de la campagne Easy d'évaluation des analyseurs syntaxiques du français. TALN Toulouse