Do we still Need Gold Standards for Evaluation?

Thierry Poibeau and Cédric Messiant

Laboratoire d'Informatique de Paris-Nord

28 May 2008

Evaluation Schemes

- Intrinsic evaluation (evaluation against a gold standard).
- Extrinsic evaluation (evaluation turned towards a practical task).
- User-oriented evaluation (experiments with users).

Evaluation Schemes

- Intrinsic evaluation (evaluation against a gold standard).
- Extrinsic evaluation (evaluation turned towards a practical task).
- User-oriented evaluation (experiments with users).

- Why is intrinsic evaluation so popular?
  - Quick and easy, provided that a gold standard is available.
  - Provides scores that makes comparison easy.
- But is it the most relevant scheme?

The Problem with Gold Standards

- Intrinsic evaluation seems to provide a simple and objective scheme.
  - NLP tools provide an output (a resource or an annotated corpus).
  - A manual reference is produced (the gold standard).
  - The evaluation consists in comparing the tool's output with the manual reference.

The Problem with Gold Standards

- Intrinsic evaluation seems to provide a simple and objective scheme.
  - NLP tools provide an output (a resource or an annotated corpus).
  - A manual reference is produced (the gold standard).
  - The evaluation consists in comparing the tool's output with the manual reference.
- However, evaluating against a gold standard is not straightforward.
  - Is the gold standard accurate?
  - Is it comprehensive?
  - Does it contain all the required information?
  - To what extend is it comparable with the tool's output?

NLP and Lexical Information

In this presentation, we take the example of lexical acquisition from corpora.

- A dictionary is a key component for most NLP applications.
  - Comprehensive dictionaries are not available for most languages.
  - Acquisition techniques makes it possible to quickly develop accurate and tunable dictionaries.
  - These dictionaries need to be evaluated.
  - The gold standard scheme is the most popular one.
- We re-investigate this question: we take as a starting point experiments we have done while developping a Subcategorization Frame (SCF) acquisition system for French.

SCF Acquisition as a Typical NLP Task

- SCFs are especially useful for NLP
  - Technical (internal) NLP tasks (e.g. parsing)
  - Practical (user-oriented) applications (e.g. information extraction)

SCF Acquisition as a Typical NLP Task

- SCFs are especially useful for NLP
  - Technical (internal) NLP tasks (e.g. parsing)
  - Practical (user-oriented) applications (e.g. information extraction)
- However, there is no clear definition of what to include into a SCF.
  - The notion of SCF is not completely formalized (what is an argument? What is a adjunct?).
  - It is partially dependent on the domain and the corpus.
  - It is partially dependent on the application
- This is typical of most NLP tasks!

An Example

- A SCF acquisition system has been developed for French.
- A large lexicon of French verbs with SCFs has been produced (see Messiant, Korhonen and Poibeau, LREC 08).
- Below is the example of an entry for the French verb *s'abattre*.

```
:NUM:       05204
:SUBCAT:    s'abattre :  SP[sur+SN]
:VERB:      S'ABATTRE+s'abattre
:SCF:       SP[sur+SN]
:COUNT:     420
:RELFREQ:   0.882
:EXAMPLE:   25458;25459;25460;25461;25462
```

Tentative Gold Standards

- We need a gold standard to evaluate our resource.
- Several electronic dictionaries exist for French
  - Lexicon-grammar (LG) from LADL (Gross, 1994).
  - DicoValence from the University of Leuven (Van Den Eynde and Mertens, 2006).
  - Lefff from University Paris 7 (Sagot et al., 2006)
  - TreeLex from the University of Bordeaux (Kupsc, 2007)
  - TLFI from ATILF (Dendien and Pierrel, 2003)

Tentative Gold Standards

- We need a gold standard to evaluate our resource.
- Several electronic dictionaries exist for French
  - Lexicon-grammar (LG) from LADL (Gross, 1994).
  - DicoValence from the University of Leuven (Van Den Eynde and Mertens, 2006).
  - Lefff from University Paris 7 (Sagot et al., 2006)
  - TreeLex from the University of Bordeaux (Kupsc, 2007)
  - TLFI from ATILF (Dendien and Pierrel, 2003)
- Can we directly use them as a gold standard?

How Gold is the Gold Standard?

All these dictionaries are good starting points for evaluation, but none can be used directly.

- None of the previous dictionaries are comprehensive.
- Some are not fully validated (Lefff).
- Some are not freely available (LG).
- Coverage vary depending on the resource (treeLex vs. TLFI).
- None of them (except TreeLex) include information about productivity.
- When productivity information is include, it is related to a specific corpus, and is hard to be used for another domain (TreeLex based on the Treebank from Paris 7).

Some more Difficult Issues

Some more theoretical issues also need to be examined further.

- All the dictionaries are based on specific theories
  - They do not have the same format
  - They do not contain the same information.
  - A translation process has to be defined in order to be able to use their content.

Some more Difficult Issues

Some more theoretical issues also need to be examined further.

- All the dictionaries are based on specific theories
  - They do not have the same format
  - They do not contain the same information.
  - A translation process has to be defined in order to be able to use their content.
- Examples
  - DicoValence is based on "the pronominal approach" (Van en Eynde and Benveniste, 1978)
  - LG is based on Gross' theory (a translation process has been defined (Gardent *et al.*, 2005))

Some more Difficult Issues

Some more theoretical issues also need to be examined further.

- All the dictionaries are based on specific theories
  - They do not have the same format
  - They do not contain the same information.
  - A translation process has to be defined in order to be able to use their content.
- Examples
  - DicoValence is based on "the pronominal approach" (Van en Eynde and Benveniste, 1978)
  - LG is based on Gross' theory (a translation process has been defined (Gardent *et al.*, 2005))
- There is thus a need to develop an accurate gold standard from these resources.

What do we Learn from the Evaluation?

- Imagine we now have a gold standard that is as accurate and comprehensive as possible. It is then possible to compute scores for precision and recall
- However, when there is a mismatch between the system and the gold standard, it does not say if:
  - The system is wrong,
  - The gold standard is wrong,
  - Both of them are right/wrong (e.g. if the SCF is specific to a given corpus).
- Only a manual analysis of the results can explore the reasons of the mismatches.

We must be Cautious when Comparing Results against a Gold Standard

- Scores needs to be analyzed manually.
- This analysis is far from obvious for the reasons given before:
  - Performance is always relative to a domain, a corpus and a theory.
  - Human (post-)validation is time-consuming and error-prone.
- Therefore, scores are not as objective as they may appear!
- However, *we should not throw the baby out with the bath water!*
  - Intrinsic evaluation remains a quick and valuable way of evaluating NLP systems.
  - It is relevant provided the fact that the gold standard is accurate enough.

Intrinsic vs Extrinsic Evaluation

- Gold standard based evaluation tends to favour systems that produce results similar to manual ones.
- They are not always appropriate (e.g. to evaluate productivity information – corpus "representativeness" is then a key factor).
- Moreover, the significance of an error largely depends on the task.
  - e.g. for IE, the distinction between arguments and adjuncts may not be so fundamental,
  - whereas, it is for parsing (productivity information is then fundamental!)
- Therefore, other kinds of evaluation may be relevant, in addition to intrinsic evaluation.

Evaluating in an Applicative Context

- Extrinsic evaluation allows one to check the usefulness of a result for a certain task.
- e.g. Evaluating the usefulness of a resource for an Information Extraction task.
  - It offers a better view of the utility of a resource.
  - It shows the interest of the automatic acquisition approach.
- Information extraction is especially relevant in our case
  - It requires specific resources in order to be efficient.
  - It requires efficient techniques to quickly acquire these resources.

Extrinsic Evaluation.

- When integrating the SCF information in an IE system, one can see that:
  - The system performs better when incorporating lexical acquisition technique than when simply using an existing dictionary.
  - The acquired data need to be completed with existing dictionaries in order to make the system efficient.
- Practical applications show:
  - How data can be integrated in order to give satisfactory results.
  - How relevant an approach/a result is for a given task (this result can be quite different from the one obtained from an intrinsic evaluation).
- Therefore, extrinsic evaluation naturally complements intrinsic evaluation.

What for Other Kinds of Tasks?

- Is SCF acquisition a special case for evaluation?
  - Cf. R. Bod (ACL07, about parsing): "It is well known that any evaluation on hand-annotated corpora unreasonably favours supervised parsers. There is thus a quest for designing an evaluation scheme that is independent of annotations".
  - Then Bod proposes to evaluate how machine translation could benefit from his parsing algorithm .

Extrinsic evaluation

- Extrinsic evaluation is an invaluable source of knowledge to assess the usefulness of a resource or of a tool.
- However, it remains heavy to organize.
- It is generally difficult to understand where errors come from.

Conclusion

- Finally we have re-investigated two classical evaluation schemes:
  - Intrinsic evaluation,
  - Extrinsic evaluation.
- Intrinsic evaluation is by far the most popular evaluation scheme.
- Most often, it is not as “objective” as it may seems.
- It can be pertinently complemented by extrinsic evaluation.

Thank you!

thierry.poibeau@lipn.univ-paris13.fr
cedric.messiant@lipn.univ-paris13.fr