On Evaluation of Natural Language Processing Tasks
Is Gold Standard Evaluation Methodology a Good Solution?

Vojtěch Kovář, Miloš Jakubíček and Aleš Horák
Natural Language Processing Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic

Keywords: Natural Language Processing, Applications, Evaluation.

Abstract: The paper discusses problems in state of the art evaluation methods used in natural language processing (NLP). Usually, some form of gold standard data is used for evaluation of various NLP tasks, ranging from morphological annotation to semantic analysis. We discuss problems and validity of this type of evaluation, for various tasks, and illustrate the problems on examples. Then we propose using application-driven evaluations, wherever it is possible. Although it is more expensive, more complicated and not so precise, it is the only way to find out if a particular tool is useful at all.

1 INTRODUCTION: GOAL OF NATURAL LANGUAGE PROCESSING

Why do we do computational analysis of natural language? The ultimate goal can be described as "to teach computer understand and use human language". However, computers should have this ability for a purpose: to be able to help us solve everyday tasks that involve understanding of human language. We want computers to correct our writing, to translate our texts, to answer our questions... All of these are *applications* of natural language processing (NLP) technology and research.

It may seem trivial but it is not. In the last years, a significant part of research in NLP was driven by the annotated data available for precise evaluations and comparisons with others, and the need of having as high numbers as possible to be published. At the same time, the papers stop arguing about "what is it good for?" and simply repeat the previous experiments (with better and better results, but often still without practical impacts). The applicability, as the main goal of the NLP research, has partly disappeared from the research.

To a great extent, this was caused by mechanical, almost monopolistic state-of-the-art evaluation methodology using gold standard data, manually annotated on certain levels of linguistic analysis.

In this paper, we argue that this methodology leads to unwanted effects and it should not be used – or at least, it should not be the only serious option – in scientific evaluations of NLP tools. We illustrate that using gold standard evaluation methodology often leads to developments that are irrelevant for real applications, and useless in general. Then we propose an alternative methodology focused on applicability of particular tools.

The structure of the paper is as follows: The next section summarizes the state-of-the-art evaluation methodology that uses gold standards. Section 3 shows the negative effects of the current methodology and illustrates them on examples. Section 4 contains the proposal of the alternative approach to the scientific evaluation of NLP tools, and discusses their advantages and weaknesses.

Examples will be given in English and Czech, as these are the languages that the authors work with.

2 STATE OF THE ART: GOLD STANDARDS

"Gold standard" for an NLP task is a data set of natural language texts annotated by humans for correct solutions of that particular task. Examples include:

- treebanks, for syntactic analysis – natural language corpora where every sentence is annotated by its correct syntactic tree (Marcus et al., 1993; Hajič, 2006)
- parallel corpora, for machine translation where each sentence or segment in the source language

Kovář, V., Jakubíček, M. and Horák, A.
On Evaluation of Natural Language Processing Tasks - Is Gold Standard Evaluation Methodology a Good Solution?.
DOI: 10.5220/0005824805400545
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 2*, pages 540-545
ISBN: 978-989-758-172-4

is annotated by correct translation into the target language (Koehn, 2005)

- corpora with annotated named entities (possibly with relations among them), for named entity recognition task and information extraction (Kim et al., 2003)
- documents with assigned topics or terms from a set of possible options, for topic, keyword or terminology extraction
- and many others

The evaluation then means comparing output of an NLP tool with the data in the gold standard, computing some sort of similarity. Precision, recall and F-measure are most used metrics[1] but there are others – e.g. special tree similarity metrics in case of syntactic analysis (Sampson, 2000) or the BLEU score (Papineni et al., 2002) widely used for evaluation of machine translation.

Development process of majority of NLP tools is then similar to the following:

1. implement a prototype (possibly because a particular application needs it, or because the task seems meaningful)
2. find a suitable gold standard data and evaluate (because proper evaluation and comparison with other tools is needed in order to publish the results)
3. tune the tool until the numbers against the gold standard are publishable, then publish

In the next section, we will try to explain what is wrong with this approach.

3 CRITICISM OF GOLD STANDARDS

There is a number of factors that make the gold standard evaluation methodology problematic.

3.1 Overfitting to Gold Standard

Creating the gold standards is expensive – not only a specialist in the field is required who spends typically months annotating a reasonable amount of data; more of them is needed to eliminate errors and random decisions, and they usually need to know extensive annotation instructions in detail (see also below). On the other hand, evaluating a tool against the data is very cheap, once they are created – usually there is a simple script that produces the numbers.

For this reason, there is typically only one (or a few, at most) gold standard data set for each task. This leads to all tools producing one type of output, compatible with this gold standard, because of the need for evaluation.

But this does not correspond to the reality: Each application has slightly different needs. Detecting named entities in Wikipedia (Nothman et al., 2012) is dramatically different from detecting named entities in blogs or Facebook posts (which are probably much more needed). In case of morphological tagging, sometimes it is desirable to distinguish between e.g. passive verb forms and adjectives (which may be tricky); but in many cases it is not, and it would just make the task more complicated for no reason. In case of syntactic analysis there are much more similar cases; typically an application needs to recognize one type of phrases and 80% of the tree structure is useless for it.

However, the gold standard enforces that all tools need to solve all the problems covered by the gold standard and in the same way as the gold standard prescribes (e.g. with the same granularity), otherwise they will lack a sound evaluation according to the state-of-the-art methodology. This way, the NLP tools are designed according to the gold standard "shape" – they need to implement all the details that are implemented in gold standard and need to follow all the arbitrary decisions that the gold standard creators have made – instead of aiming at needs of particular applications.

Figure 1 shows how absurd this mentioned gold standard shape – that the tools are forced to accommodate to – can be.

3.2 Inter-Annotator Agreement and Ambiguity

Inter-annotator agreement (IAA), and even intra-annotator agreement, is a nightmare for creators of manually annotated data, including gold standards. It is rarely published, even for very prominent data sets it is not available or semi-official (Manning, 2011).

Despite of that, high IAA is considered a crucial property of quality data, because – as the argument goes – if people do not agree with each other on the correct solution, how could we expect machines to solve the task well? The tasks where high IAA cannot

[1] Precision is the percentage of correct automatic annotations with respect to all automatic annotations; recall is described as percentage of annotations from the gold standard covered by the automatic annotation; F-measure combines these two into a single number.

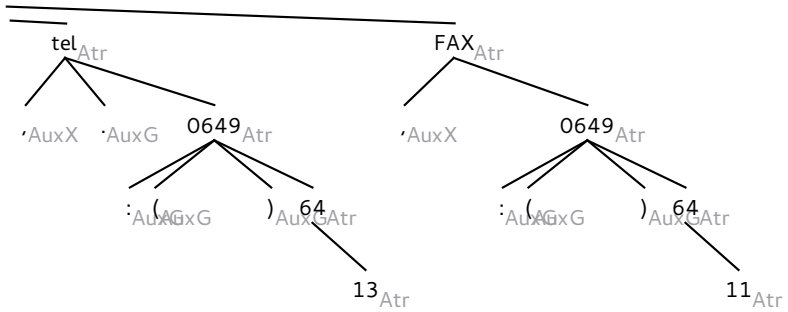

Figure 1: Example of a toxic gold standard – dependency syntactic analysis of Czech, Prague dependency treebank. The segment showed here is: *", tel.: (0649) 64 13, FAX: (0649): 64 11"* (phone and fax number, analyzed arbitrarily according to gold standard rules).

be achieved, are normally perceived as ill-defined.

For well and long studied problems, such as morphological and syntactic analysis, the IAA estimations are known – for morphological tagging of English it is probably around 97% (Manning, 2011),[2] for syntactic analysis around 90% (Mikulová and Štěpánek, 2009).[3]

(Sampson and Babarczy, 2008) provide an interesting study on limits of IAA in case of syntactic analysis of English where they conclude that

- IAA limit on the task is about 95% and cannot be improved by further modifications of annotation instructions, mainly because
- the remaining percentage are structural ambiguities within the language, not a question of technical arrangements

Is 90, or 95 percent pairwise agreement enough for claiming such a data set a universal authority for evaluation? There may be up to 10 percentage points error in the evaluation, half of which are inevitable structural ambiguities – which means that it does not matter whether the tool under evaluation recognized them correctly or not.

Next, to achieve the IAA figures over 90 percent in syntactic annotation, extensive annotation manuals are needed and the annotators need to be very familiar with the underlying language theory. Annotation manuals for Penn treebank and for Prague dependency treebank contain about 300 pages (Bies et al., 1995; Hajič et al., 2005). Is this a record of universal language intuition on syntactic level, or rather set of arbitrary decisions which the data set is built on? If the latter is more correct, then the evaluation of a parser means just testing its ability to follow these arbitrary instructions – not the ability to reveal syntactic information.

3.3 Impossible Gold Standards

For many tasks, the inter-annotator agreement above 90%, or even 70-80% is completely unrealistic to achieve. Such tasks include terminology extraction, keywords extraction, text summarization, topic detection, collocation extraction, ...

Gold standards for such tasks either do not exist at all, which makes the task "ill-defined", or they are domain-specific and created according to strict rules that are not general and can be used for specific purposes only. For example, the GENIA corpus (Kim et al., 2003) can be used as a gold standard for terminology extraction, however, comparison with a general terminology extraction system (Kilgarriff et al., 2014a) shows that there are differences (mainly different notions of "term") that skew the resulting numbers significantly.

Despite of that, these tasks need to be solved by automatic systems – the low IAA itself does not mean the task is invalid. Collocation and terminology dictionaries do exist (e.g. (Rundell, 2010)) and are useful (despite the fact that there is probably not a general agreement on what exactly they should contain); topic detection and summarization systems are badly needed in today's world of information overflow... We need a way of evaluating these systems, but gold standards are probably not the right way to go.

3.4 Dependency on Arbitrary Decisions

In Section 3.2 we have mentioned that gold standards depend heavily on arbitrary decisions that are not general and do not reflect language intuition. Back at the syntactic analysis task, comparison of Czech parsers

[2] On a tagset with ca. 50 tags – with a more fine-grained tagsets it would be definitely worse. Tagsets for flective languages such as Czech or Russian contain thousands of tags.

[3] The example is for dependency analysis of Czech, as in Prague Czech-English dependency treebank; the most important "Structure" feature which determines the shape of the syntactic tree, shows pairwise agreements between human annotators between 87 and 89 percent.

made by (Radziszewski and Grác, 2013) and (Kovář, 2014, section 3.4), on two different gold standards, clearly shows *negative correlation* of the two results.

In other words, the better results on one gold standard, the worse on the other one, although both were designed for general analysis of Czech. Also, overfitting of the statistical tools to the gold standard as discussed above, is clearly visible. Either one of the gold standards is plain wrong,[4] or the results are massively inconsistent and the gold standards do not provide reliable evaluations even in traditional tasks like syntactic analysis.

3.5 Application-free Evaluations

Weak and negative correlations are usually found also between application-oriented evaluations and gold standard evaluations. Again, most evidence is in the field of syntactic analysis. (Miyao et al., 2009) report weak correlation of these two evaluations for English parsers, but the following observation is more important: *a 1% absolute improvement in parser accuracy [against a treebank] corresponds roughly to a 0.25% improvement in PPI extraction accuracy [protein-to-protein interaction, the application in focus]*. Parsing accuracy moves around 85 percent and the PPI accuracy around 57 percent, which means that parsing in the current shape actually does not help the application.

Comparison of Czech parsers on collocation extraction application (Kilgarriff et al., 2014b) shows no correlation with gold standard evaluation – and the best result according to the application evaluation was achieved by a specialized shallow parser, the output of which is not comparable with the gold standard.

(Katz-Brown et al., 2011) from Google report negative correlation when using English parsers as part of the machine translation process: the higher the accuracy of the application, the lower the accuracy against a gold standard. Similar reports can be found also in (Mollá and Hutchinson, 2003) and (Galliers and Spärck Jones, 1993).

All of this information indicates that the gold standard methodology does not provide meaningful evaluations of NLP tools with regard to applications, the most important goal of NLP research. Rather than that, it is evaluating their ability to imitate the data present in gold standards which may be very different from needs of applications.

[4] And in that case – how to find out that a gold standard is plain wrong?

4 APPLICATION-DRIVEN EVALUATION

Our proposal builds on the premise that we should aim at the final goal of the development in NLP: the applications useful for people. It is irrelevant how well each component of a complex application works, only the overall result is important for a final evaluation and for comparisons.

We propose to abandon the intrinsic gold standard evaluations and use purely extrinsic application-driven evaluation methods. That is, to design a real-world application (or build on one that is already available) that can be useful for a group of people, and ask these people to use it with a particular data and to quantify how useful it actually is.

The quantification could be done by various methods, and it depends on the particular application which one is the most suitable. The most coarse-grained method for cases where there is no better option, would probably be just to ask the users about their feeling from the application, or better, about improvement/deterioration from a previous version. The most precise way, for applications that allow it, would be to build a gold standard data set *for the final application* and perform automatic and precise evaluations.

On the first sight, the latter may seem as the same methodology that is already used, and was criticized above. The crucial difference from the criticised approach is that the gold standard data will be prepared *only for the final application* and not for any sub-task that is not directly usable. Such evaluations would do the same service as human-oriented ones, just in a faster and cheaper way.

The usual case would be probably somewhere in the middle – for most of real-word applications the users would evaluate their behaviour in small parts (e.g. sentence by sentence), and the result could be interpreted in a reasonably precise quantitative way.

4.1 Examples

4.1.1 Parsing

Are you developing a parser? What is it for – are you claiming that parsing is a corner-stone of any advanced NLP application? You need to prove it.[5]

Pick one of the possible such real-world applications, implement it (a fairly basic version may be enough) and show how it needs your parser. Or, if it

[5] Most of current successful NLP applications are statistical and operate on the word level not exploiting *any* structured information, let alone syntactic trees.

is technically feasible, build your parser into an existing advanced application and show that the parser has improved the results. Then, measure any future development of the parser by the results of the real world application, not by tree similarity metrics – they are useless in context of the application.

Examples of such applications are quite obvious and include:

- (partial) grammar checking
- extracting structured knowledge from text
- extracting short answers to questions
- measuring fluency of text (e.g. for student writing evaluation, translation evaluation, ...)

In case of grammar checking, the evaluation metric can be number of errors fixed (in terms of precision and recall); for extracting knowledge, number of correct extractions; question answering – number of correct answers; text fluency – correlation with native speaker judgments.

4.1.2 Terminology Extraction

Are you working on terminology extraction and no suitable gold standard is available? Build an application – e.g. checking consistency of term translations, or whatever else you think it is good for – and let people evaluate your application. For checking translation consistency, the measure can be the number of false alerts and number of errors that were missed by the tool (which is a bit complicated to find out, but still doable).

4.2 Discussion

Obviously, the proposed methodology has a lot of disadvantages. Here we discuss the most important of them.

Price. Evaluation involving human annotations or rating will always be more expensive than gold standard evaluation. But evaluation for publication is not very frequent, and everyday evaluations for development purposes can be covered by automated tests for regressions.

Replicability. Human ranking will not be objective, and will not be perfectly reproducible. However, the reproducibility of results is a known weakness of NLP research anyway. Besides, the human evaluations *will* be replicable *to a significant extent*, like experiments in humanities with a similar group of people – if not, the evaluation is not valid. Again, it will be probably more expensive, but much more meaningful. Even a non-replicable evaluation by real users of an application would be more valuable than evaluation against gold standard that has nothing in common with any application.

Sensitivity. We will not be able to produce precise numbers, there will be deviations between measurements, probably several percentage points (but it depends on exact circumstances). Yes, this is inevitable – but what does +1 percentage point mean on a gold standard when it can mean -10 percentage points on another gold standard or on application? Gold standard evaluations are very precise but the numbers are problematic; and due to the IAA problems mentioned earlier, the precision of the numbers is debatable as well.

Specificity. The proposed methodology cannot measure the general accuracy of a tool, only the bits important for particular applications. But there is nothing like general accuracy, a purpose of the tool is to be used in applications. If you want more general results, test on more different applications.

Subjectivity, and more Space for Cheating. Yes, you can use your students for evaluation, tell them to be generous and then publish that the evaluation was done by independent experts. But such cheating is possible in gold standard world, too – selection of suitable data, tuning the tool for the testing data, ... It is a general question of ethics in science. On the other hand, human evaluations may be even easier to disprove: hiring a group of evaluators is technically very easy whereas running the computer evaluation is often not. Therefore, disproving some results may be interpreted as not understanding an evaluation program – this is not possible in case of human evaluations.

In general, despite the disadvantages, we consider the application-based evaluations the only way how to really prove the usefulness of a particular tool.

5 CONCLUSIONS

In the paper we have formulated some serious problems of gold standard evaluation methodology, currently massively used in all areas of NLP research. We have illustrated the problems on examples and showed that gold standard evaluations can be very misleading. Then we have proposed an alternative, based purely on particular applications, in contrast to seemingly general gold standards.

Although the formulations in the paper are sometimes very strict, it should not be read as complete denial of the gold standard methodology, we believe it can be useful in certain cases, namely when directly reflecting the needs of an application. Rather than

that, we want to discourage from mechanical usage of gold standard evaluation methodology, start a discussion on evaluation methodology in NLP, as well as a shift towards evaluations driven by particular applications. There is no such discussion going on now and the gold standard methodology is usually taken as a dogma.

ACKNOWLEDGEMENTS

This work has been partly supported by the Grant Agency of CR within the project 15-13277S. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

REFERENCES

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M. A., and Schasberger, B. (1995). Bracketing guidelines for treebank II style Penn treebank project.

Galliers, J. and Spärck Jones, K. (1993). Evaluating natural language processing systems. Technical Report UCAM-CL-TR-291, University of Cambridge, Computer Laboratory.

Hajič, J. (2006). Complex corpus annotation: The Prague dependency treebank. *Insight into the Slovak and Czech Corpus Linguistics*, page 54.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A., Štěpánek, J., Pajas, P., and Kárník, J. (2005). Annotations at analytical level: Instructions for annotators.

Katz-Brown, J., Petrov, S., McDonald, R., Och, F., Talbot, D., Ichikawa, H., Seno, M., and Kazawa, H. (2011). Training a parser for machine translation reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 183–192. Association for Computational Linguistics.

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., and Suchomel, V. (2014a). Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conferencethe European Chapter of the Association for Computational Linguistics*, pages 53–56, Gothenburg, Sweden. The Association for Computational Linguistics.

Kilgarriff, A., Rychlý, P., Jakubíček, M., Kovář, V., Baisa, V., and Kocincová, L. (2014b). Extrinsic corpus evaluation with a collocation dictionary task. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1–8, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Kovář, V. (2014). *Automatic Syntactic Analysis for Real-World Applications*. Phd thesis, Masaryk University, Faculty of Informatics.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011*, pages 171–189. Springer, Berlin.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Mikulová, M. and Štěpánek, J. (2009). Annotation procedure in building the Prague Czech-English dependency treebank. In *Slovko 2009, NLP, Corpus Linguistics, Corpus Based Grammar Research*, pages 241–248, Bratislava, Slovakia. Slovenská akadémia vied.

Miyao, Y., Sagae, K., Sætre, R., Matsuzaki, T., and Tsujii, J. (2009). Evaluating contributions of natural language parsers to protein–protein interaction extraction. *Bioinformatics*, 25(3):394–400.

Mollá, D. and Hutchinson, B. (2003). Intrinsic versus extrinsic evaluations of parsing systems. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable?*, Evalinitiatives '03, pages 43–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2012). Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Radziszewski, A. and Grác, M. (2013). Using low-cost annotation to train a reliable Czech shallow parser. In *Proceedings of Text, Speech and Dialogue, 16th International Conference*, volume 8082 of *Lecture Notes in Computer Science*, pages 575–1156, Berlin. Springer.

Rundell, M. (2010). *Macmillan Collocations Dictionary*. Macmillan.

Sampson, G. (2000). A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, 5(01):53–68.

Sampson, G. and Babarczy, A. (2008). Definitional and human constraints on structural annotation of English. *Natural Language Engineering*, 14(4):471–494.