

Evaluation: a paradigm that produces high quality language resources

Patrick Paroubek

LIMSI-CNRS
BP 133 91403 Orsay cedex, FRANCE
{pap,anne}@limsi.fr

Abstract

We show how the paradigm of evaluation can function as language resource producer for high quality and low cost validated language resources. First the paradigm of evaluation is presented, the main points of its history are recalled, from the first deployment that took place in the USA during the DARPA/NIST evaluation campaigns, up to latest efforts in Europe. Then the principle behind the method used to produce high-quality validated language resources at low cost from the by-products of an evaluation campaign is exposed. It finds its origins in the experiments performed after speech recognition evaluation campaigns in the USA, when the outputs of the participating systems were combined with a simple voting strategy to obtain higher performance results and also in theoretical results from machine learning, which show that one can combine low performance classifiers to obtain an improved system.

1. The Evaluation Paradigm

Apart from the ALPAC event (S. Nirenburg and Wilks, 2003), the evaluation paradigm was initially deployed in the United States in the framework of DARPA projects in 1987, with the organization of a series of evaluation campaigns for speech processing (Pallett, 2003), then for text understanding with the MUC campaigns (Hirschman, 1998) of the TIPSTER program (Harman, 1992), as well as in the scope of other programs run later by the Association for Computational Linguistics (ACL) and by the NIST.

The paradigm is based on a two step process:

- first, create textual or voice data in the form of raw corpora, tagged corpora or lexicons, which are then distributed to main actors in the field of language engineering for the realization of natural applications involving natural language processing, e.g. word sense disambiguation, POS tagging, parsing, natural language database query, message understanding, automatic translation, dictation, oral dialog, character recognition, information retrieval, information extraction, question answering, opinion mining etc.
- second, the systems are tested on similar data and compared. The results of the test sessions and the discussions ensuing from the publication of the results furnish a sound basis to compare pros and cons of the various methods and systems during a workshop.

In addition to the knowledge gained about the algorithms evaluated, the close collaboration that exists between computer scientists and linguists participating in an evaluation campaign and the resulting synergy among the actors are two other benefits from the evaluation paradigm. Collaboration is required to define the common data sets and the gold standard, to propose the evaluation criteria, to define evaluation protocols, and to organize the processing of the data.

In Europe, the first event of the sort happened in 1994 in Germany with the “morpholympics” (Hauser, 1994) on morphological analyzers for German. The same year was started in France the GRACE campaign on Part-Of-Speech

taggers of French (Paroubek, 2000), then there were 7 campaigns of the FRANCIL program (Mariani and Paroubek, 1999) for text and speech, the series of self-supported campaigns Senseval on lexical semantics organized by the ACL-SIGLEX working group (Edmonds and Kilgariff, 2003), its follow-up Semeval (Agirre et al., 2007) or the more recent evaluations campaigns for Portuguese text analysis (Santos and Cardoso, 2006), as well as examples of national programs on evaluation like TECHNOLOGUE (Mapelli et al., 2004) in France with the 8 evaluation campaigns on both speech and text of the EVALDA project or the latest EVALITA (Magnini and Cappelli, 2007) in Italy with its 5 campaigns on text analysis. There were also European project which have addressed the subject of evaluation within the past few years, from EAGLES (King et al., 1996) to the CLEF evaluation series (Agosti et al., 2007).

2. ROVER

The idea to combine the output of systems participating to an evaluation campaign in order to obtain a combination with better performance than the best one is not new. To our knowledge, what now is known as the ROVER (Reduced Output Voting Error Reduction) algorithm was invented by J. Fiscus (Fiscus, 1997) in a DARPA/NIST evaluation campaign about speech recognition. He found out that by aligning the output of the participating speech transcription systems with a dynamic programming algorithm (Allison et al., 1990) and by selecting the hypothesis which was proposed by the majority of the systems, he obtained better performances than with the best system. Since, the idea gained support, first in the speech processing community (Löf et al., 2007), where people now work on refined versions of the algorithm, using the performance of the different speech recognizers as confidence weights in the hypothesis lattice obtained by combining the different outputs and by applying language models to guide the final stage of best hypothesis selection (Schwenk and Gauvain, 2000). In general better results are obtained with retaining only the output of the two or three best performing systems, in which case the relative improvement can go up to 20% with respect to the best performance (Schwenk and Gauvain,

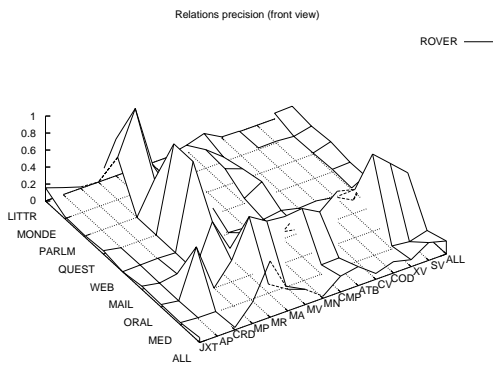


Figure 1: ROVER relative gain of performance in precision for syntactic relation annotation against the best performance

2000). For text processing, examples of use of ROVER procedure are more rare, one such instance is MULTITAG (Paroubek, 2000) for POS tagging, where the algorithm was applied to provide POS tags with confidence annotation to yield a validated language resource from data produced in an evaluation campaign. Machine translation evaluation is another area where ROVER algorithms are used (Matusov et al., 2006). The ROVER now begins to be tested as a resource production procedure in the scope of the PASSAGE project where it is used to combine parses to produce linguistic information, see section 3.

3. Improving the Quality of Ressources

MULTITAG (Paroubek, 2000), a French CNRS project, had the goal of producing and making available a 1 Million words corpus annotated with POS tags out of the corpus tagged by the participants of the GRACE evaluation campaign. From the initial aligned corpus tagged by the taggers and the POS annotation mappings provided by the participants were produced the confidence measures by vote counting. Then manual validation was done first only on 38,643 forms (4%) out of the 830000 forms of the test corpus for which the system combination procedure had produced an ambiguous annotation (main morphosyntactic category or subcategory). In a second step, all the forms whose annotations contained number, gender or person information (64,061 forms of the test corpus, roughly 8%) were manually checked. Thus only less than 10% of the corpus needed to be hand checked to obtain a validated annotations.

The syntactic annotations produced by the parsers that participated to the EASY evaluation campaign gave the occasion to test a ROVER algorithm. What we found to work best was by weighting the annotation of a system proportionally to the rank the system obtained at the evaluation, in a way that the annotation of the best system could be changed only if the majority of the other systems voted against it. With this algorithm, we obtained the relative gain of performance in precision for syntactic relation annotation against the best performance shown in figure 1 (Paroubek et al., 2008). In figure 2 we show the difference between recall of union of all participants and best recall performance for syntactic relation annotation at the

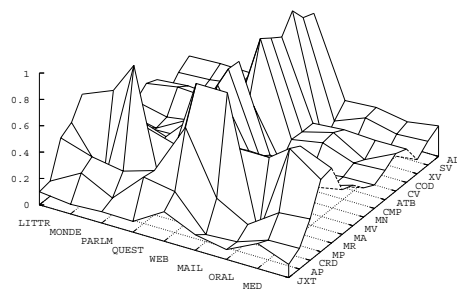


Figure 2: Difference between recall of union of all participants and best recall performance for syntactic relation annotation at EASY campaign

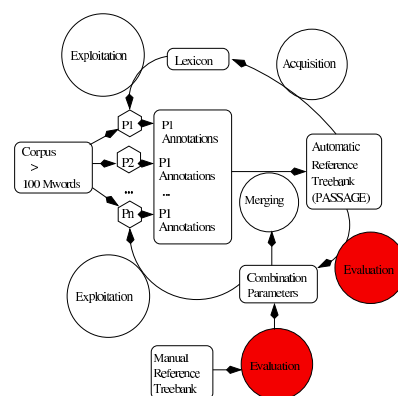


Figure 3: How PASSAGE uses the evaluation paradigm (in grey) to identify ROVER parameters in order to produce automatically a large sized treebank with high quality annotation.

EASY campaign. The interesting fact about this data plot is that it is always positive, it means that the potential gain in recall by combination methods is always possible, for any kind of relation and any kind of corpus genre, provided that one can identify the right weight to give to the output of each parser. That is precisely one of the aim of PASSAGE (de la Clergerie et al., 2008) (Paroubek et al., 2009), whose aim is to use the paradigm of evaluation to identify which parameters to give to a ROVER combination procedure to produce automatically a large sized treebank following the schema given in figure 3. The theoretical grounds behind the ROVER algorithm come from Machine Learning, with Vaillant's PAC model of learning, more precisely with the work of (Javed A. Aslam and Scott E. Decatur, 1993) on boosting the accuracy of weak learning algorithms which fall within the Statistical Query model, a model introduced by Michael Kearns to provide a general framework for efficient PAC learning in the presence of classification noise.

4. Conclusion

We have looked at the recent history of the paradigm of evaluation both in United States and Europe and have shown that be used to produce validated high quality linguistic annotations at a relatively low cost.

5. References

- Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June.
- Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro, Donna Harman, and Carol Peters, 2007. *Proceedings of the 11th Conference on Research and Advanced Technology for Digital Libraries*, chapter The Future of Large-Scale Evaluation Campaigns for Information Retrieval in Europe, pages 509–512. Springer Verlag. ISBN 3540748504, 9783540748502.
- L. Allison, C. S. Wallace, and C. N. Yee. 1990. When is a string like a string? In *Proceedings of International Symposium on Artificial Intelligence in Mathematics (AIM)*, Ft. Lauderdale, Florida, January.
- Phil Edmonds and Adam Kilgarriff. 2003. Special issue based on senseval-2. *Journal of Natural Language Engineering*, 9(1), January.
- Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (rover). In *In proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–357, Santa Barbara, CA.
- Donna Harman. 1992. The darpa tipster project. *ACM SIGIR Forum*, 26(2):26–28. ISSN:0163-5840.
- Roland Hauser. 1994. Results of the 1. morpholympics. *LDV-FORUM*, 11(1), June. ISSN 0172-9926.
- L. Hirschman. 1998. Language understanding evaluations: lessons learned from muc and atis. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, pages 117–122, Grenade, Espagne.
- Maghi King, Bente Maegaard, Jörg Schütz, Louis des Tombes, Annelise Bech, Anne Neville, Antti Arppe, Loran Balkan, Colin Brace, Harry Bunt, Lauri Carlson, Shona Douglas, Monika Höge, Steven Krauwer, Sandra Manzi, Cristina Mazzi, Ane June Sieleman, and Ragna Steenbakkere. 1996. *EAGLES Evaluation of Natural Language Processing Systems*. Center for Sprogteknologi, Copenhagen, october. ISBN 87-90708-00-8.
- J. Löff, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, , and H. Ney. 2007. The rwth 2007 tc-star evaluation system for european english and spanish. In *In proceedings of the Interspeech Conference*, pages 2145–2148.
- Bernardo Magnini and Amadeo Cappelli, editors. 2007. *Evalita 2007: Evaluating Natural Language Tools for Italian*, volume IV n°2, Roma, June. Associazione Italiana Intelligenza Artificiale (AI*IA). ISSN 1724-8035.
- Valérie Mapelli, Maria Nava, Sylvain Surcin, Djamel Mostefa, and Khalid Choukri. 2004. Technolangu: A permanent evaluation and information infrastructure. In *In proceedings of the 4th international Conference on Language Resources and Evaluation (LREC)*, volume 2, pages 381–384, Lisboa, Portugal, May. ELDA.
- Joseph Mariani and Patrick Paroubek. 1999. Human language technologies evaluation in the european framework. In *Proc. of the DARPA Broadcast News Workshop*, pages 237–242, Herndon, VA, February. Morgan Kaufmann.
- Evgeny Matusov, N. Ueffing, and Herman Ney. 2006. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 158–165, Trento, Italy.
- E. de la Clergerie, O. Hamon, D. Mostefa, C. Ayache, P. Paroubek, and A. Vilnat. 2008. Passage: from French parser evaluation to large sized treebank. In ELRA, editor, *In proceedings of the sixth international conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May.
- Javed A. Aslam and Scott E. Decatur. 1993. General bounds on statistical query learning and pac learning with noise via hypothesis boosting. In *Proceedings of the 34th Symposium on Foundations of Computer Science*, Foundations of Computer Science, pages 282–291. IEEE, November.
- David Pallett. 2003. A look at nist’s benchmark asr tests: past, present, and future. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 483–488, Virgin Islands, USA, November. IEE. ISBN:0-7803-7980-2 / DOI:10.1109/ASRU.2003.1318488.
- P. Paroubek, I. Robba, A. Vilnat, and C. Ayache. 2008. EASY, evaluation of parsers of French: what are the results? In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.
- Patrick Paroubek, Eric de la Clergerie, Sylvain Loiseau, Anne Vilnat, and Gil Francopoulo. 2009. The passage syntactic representation. In *In proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, Groningen, January. Netherlands Graduate School of Linguistics.
- Patrick Paroubek. 2000. Language resources as by-product of evaluation: the multitag example. In *In proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, volume 1, pages 151–154.
- H. Sommers S. Nirenburg and Y. Wilks, editors, 2003. *Readings in Machine Translation*, pages 131–135. MIT Press, Cambridge, Massachusset. ISBN-10: 0-262-14074-8, ISBN-13: 978-0-262-14074-4, <http://www.hutchinsweb.me.uk/ALPAC-1996.pdf>.
- Diana Santos and Nuno Cardoso, 2006. *A Golden Resource for Named Entity Recognition in Portuguese*, pages 69–79. Springer, Berlin / Heidelberg. ISBN 978-3-540-34045-4, DOI 10.1007/11751984_8.
- Holger Schwenk and Jean-Luc Gauvain. 2000. Improved rover using language model information. In *In proceedings of the ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, pages 47–52, Paris, September.