Language Resources as by-Product of Evaluation: the MULTITAG example

Patrick Paroubek

LIMSI - CNRS

Batiment 508 Universite Paris XI, 91403 Orsay Cedex Email: pap@limsi.fr URL: http://www.limsi.fr/Individus/pap

Abstract

In this paper, 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 (SENSEVAL2/ROMANSEVAL2, CLEF, CLASS etc.). Then the principle behind the method used to produce high-quality validated language at low cost from the by-products of an evaluation campaign is exposed. It was inspired by the experiments (Recognizer Output Voting Error Recognition) performed during speech recognition evaluation campaigns in the USA and consists of combining the outputs of the participating systems with a simple voting strategy to obtain higher performance results. Here we make a link with the existing strategies for system combination studied in machine learning. As an illustration we describe how the MULTITAG project funded by CNRS has built from the by-products of the GRACE evaluation campaign (French Part-Of-Speech tagging system evaluation campaign) a corpus of around 1 million words, annotated with a fine grained tagset derived from the EAGLES and MULTEXT projects. A brief presentation of the state of the art in Part-Of-Speech (POS) tagging and of the problem posed by its evaluation is given at the beginning, then the corpus itself is presented along with the procedure used to produce and validate it. In particular, the cost reduction brought by using this method instead of more classical methods is presented and its generalization to other control task is discussed in the conclusion.

1. The paradigm of Evaluation

Comparative evaluation in language engineering has been used as a basic paradigm in the USA DARPA program on human language technology since 1984. Activities similar in kind, have been pursued in Europe, both at national and at European level, but on a smaller scale and over a limited time (Mariani and Paroubek, 1999). The latest efforts concerning evaluation in Europe are CLEF (Cross Language Text Retrieval System Evaluation in collaboration with NIST and TREC conference), SENSEVAL-2/ROMANSEVAL-2 (Kilgarriff, 1998) and CLASS (evaluation across FP5 project clusters). Comparative evaluation is a paradigm in which a set of participants compare the results of their systems using the same or similar control tasks (Bernsen et al., 1999) and related data with metrics that are agreed upon. More precisely, Comparative evaluation consists in (1) choosing or creating a control task, (2) in gathering system or component developers and integrators who are interested in testing their systems against those of others, (3) in organizing an evaluation campaign which necessarily involves distributing linguistic data for training and testing the systems, and (4) in defining the protocol and the metrics which will be used in the results assessment. A control task is the function that the participating systems have to perform during an evaluation together with the conditions under which this function must be performed (e.g. for parser evaluation, a control task can be the bracketing of the constituents). Every deployment of the paradigm of evaluation in the field of Language Engineering entails the production of linguistic data: the organizers build the reference and test data sets and the participants apply their systems on test data to produce evaluation data. When a quantitative black-box (Sparck-Jones and Galliers, 1995) evaluation methodology is applied the data produced is generally abundant and could easily be re-used as training material if the cost of filtering out the errors was not so high.

2. Combining to Improve

In machine learning, it is well known that ensemble methods or committees of learning machines can often improve the performance of a system in comparison to a single learning machine. A very promising algorithm based on this principle now under investigation is Ada boost (Schwenk, 1999). In the same field, people have applied for a long time "winner take all" strategies to combine, inside the same system, the output of several basic processing units (Simpson, 1990). In the course of its evaluation program on speech recognition (S., 1998), NIST developed the ROVER (Recognizer Output Voting Voting Error Reduction) (Fiscus, 1997), to produce a composite Automatic Speech Recognition system output from the output of several ASR systems. Such composite system has an error rate inferior to the one of any of its components. In the ROVER, the output of several ASR systems is first combined into a single transition network using a modified version of the dynamic programming alignment technique used by NIST to score ASR systems[1]. This network is then explored and a simple voting strategy (highest number of votes) is used to select the best scoring word at each decision point. In (Fiscus, 1997), NIST reports a incremental 5.6% Word Error Rate reduction (12.5% relative) using voting by frequency of occurrence and maximum confidence (the output of the ASR systems was annotated with confidence measures (Chase, 1997)). Concerning, POS tagging, the principle of combination has been used in the past by (Marquez and Padro, 1998) who combines two taggers to annotate a corpus and by (Tufis, 1999) who uses several versions of same tagger but trained on different data.

[1]The SCLITE tool is freely available from http://www.itl.nist.gov/iaui/894.01/software.htm

3. The GRACE POS tagging evaluation campaign and its data.

GRACE(Adda et al., 1999) was the first large scale evaluation campaign for Part-Of-Speech tagging for the French language. It was part of the French program CCIIL (Cognition, Intelligent Communication and Language Engineering), jointly promoted by the Engineering Sciences and Human Sciences departments of the CNRS. The call for tenders was published in November 1995 and the first year has been devoted to bootstrapping the program by defining and installing the different organization committees. From the participants point of view, GRACE was made of 3 phases: training, dry-run and test. The first one was used by the participant to calibrate their systems on untagged data, the two others were complete runs of the evaluation protocol were the participants had to tag a large amount of text and to provide a mapping between their tagset and the reference tagset which had been derived with their collaboration from the EAGLES format (Leech and Wilson, 1995). The training corpus was distributed globally to all the participants in January 1996, while the dry run corpus was distributed individually to each participant in an encrypted form during the fall of 1996. The results were discussed during a workshop restricted to the participants, a satellite event to the Journées Scientifiques et Techniques du Réseau FRANCIL, in April 1997 (Adda. et al., 1997). The test corpus was distributed in the same manner as for the dry run, at the end of December 1997. The preliminary results of the tests were discussed with the participants in a workshop in May 1998. The final results were disclosed on the WEB[2] during fall of 1998 as soon as they had been validated by the organizers (cross validation with two different processing chains based on different algorithms and developed at two different sites) and the participants. At the beginning there were 18 participants from 5 different countries (CA, USA, D, CH, FR), from both public research and industry, and 3 evaluators EPFL, INaLF-CNRS and Limsi-CNRS.The 2 corpus providers were Limsi and INaLF. Out of the 21 initial participants, 17 only took part in the dry run and only 13 completed the tests. The size of the training corpus was around 10 million words of untagged texts, evenly distributed between literary works and newspaper articles. For the dry run, the participants tagged a corpus of roughly 450,000 words with a similar genre distribution and the performance measure was computed over 20,000 words to which a reference description had been manually assigned. For the tests, the participants had to mark a corpus of 650,000 words and the measure was taken over 40,000 words. GRACE used the quantitative black box metrics: Decision and Precision, which were derived especially for GRACE from the metrics used in Information Retrieval (Precision and Recall). Precision measures the ability of a POS tagger to assign a correct tag to a given word form, and Decision measures the capacity of a POS tagger to restrict for a given word form the number of candidate tags with respect to a given tagset. One of the lessons to draw from the GRACE experience, is that ideally, results should be cross-validated with two different processing chains, based on different algorithms (when this is possible) and developed at two different sites in order to ensure their accuracy and quality. The evaluation toolkit of GRACE has been packaged as a demonstration by the ELSE project and is freely available[3]. GRACE proved to be a success; its results are: a better knowledge of the existing systems in each domain and of their state of development; precise evaluation metrics defined in collaboration with the participants; an evaluation toolkit freely available, a new product on the market (one participant decided to add a tagger to his catalogue as a result of his participation); the creation of a community of actors interested in evaluation; and last of all, the initial data to build the new linguistic resource described here.

[2]http://www.limsi.fr/TLP/grace

4. MULTITAG

MULTITAG (of the joint research program in Language Engineering of CNRS departments SHS and SPI)[4] 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. The tags are in the standard format proposed by EAGLES and further refined in MULTEXT (Ide and Véronis, 1994) and GRACE. The corpus and its documentation will represent a very useful material for linguistics studies, an essential resource for POS tagger training but also an interesting material for machine learning in the study of system combination. Work has already started, and a preliminary results of a study of the relationship between the different types of material (genre) composing the corpus and POS tagging performance can be found in (Illouz, 1999). To cut down the cost of proofreading the corpus, it has been semi-automatically corrected by verifying only the forms for which the annotations proposed by the different systems did not converge. The level of convergence in the annotations provided a confidence measure to identify which forms needed to be manually checked.

From the initial aligned corpus tagged by the 15 systems (see Table 1)

occ. #	occ.	system 1 tag	system 2 tag	etc.
000	Sur	PREP	PREP	...
001	la	DTN:sg	DETFS	...
002	couverture	SBC:sg	NFS	...
003	du	DTC:sg	PREPDU	...
004	livre	SBC:sg	NMS	...
005	,	,	YPFAI	...

Table 1: Sample of the corpus tagged by the participants.

and the mapping tables provided by the participants (see Table 2):

we produced the confidence measures through vote counting. An example is given in Table 3, where the number of votes that each tag received is indicated between curly braces { }.

[3]http://www.limsi.fr/TLP/ELSE

[4]The teams involved in MULTITAG were: INaLF (CNRS), the limsi (CNRS), the LPL (U. Avignon), and TALANA (U. Paris7)

exception form	participant tag	reference tag(s)
N.A.	PREP	Sp
N.A.	DTN:sg	Ds1mss|Ds2mss|Ds3mss |Ds1fss|Ds2fss|Ds3fss |Ds1msp|Ds2msp|Ds3msp |Ds1fsp|Ds2fsp|Ds3fsp |Ddms|Dims|Dtms |Ddfs|Difs|Dtfs |Damsd|Dafsd|Damsi |Dafsi
N.A.	SBC:sg	Ncms|Ncfs
du	DTC:sg	Sp+Damsd|Dai+Damsd

Table 2: Excerpt of a tagset mapping table

occ	tags & vote #
Sur	Afcms{1}|Afpms{3}|NULL{1} |Sd{4}|Sp{13}|Sp+Dafpd{1} |Sp+Dampd{1}
la	Dafsd{13}|Dafsi{4}|Damsd{1} |Damsi{1}|Ddfs{4}|Ddms{1} |Difs{4}|Dims{1}|Dkfs{1} |Drfs{1}|Ds1fsp{4}|Ds1fss{4} |Ds1msp{1}|Ds1mss{1}|Ds2fsp{4} |Ds2fss{4}|Ds2msp{1}|Ds2mss{1} |Ds3fsp{4}|Ds3fss{4}|Ds3msp{1} |Ds3mss{1}|Dtfs{2}|Dtms{1} |NULL{1}|Ncmp{1}|Ncms{2} |Pp3fsa{2}
couverture	NULL{1}|Ncfs{14}|Ncms{1}
du	Dai+Damsd{4}|Damsd{2} |NULL{1}|Sd{1}|Sp{2} |Sp+Damsd{10}
livre	NULL{1}|Ncfs{4}|Ncms{14} |Vmip1s{2}|Vmip3s{2} |Vmmp2s{2}|Vmsp1s{2} |Vmsp3s{2}
,	Cc{1}|F{14}|NULL{1}

Table 3: Example of POS tagging confidence measure through voting with 15 systems.

The lack of user-friendly interactive tool readily available (and customizable) for correcting the corpus has been a hindrance to the project and regular office spreadsheet software was used in the end. For the linguistics aspects, defining the annotation procedure that the correctors had to follow has been much harder than was anticipated. In addition to the validation of the corpus annotations and of the method of system combination itself, this work yielded a refined annotation manual for the correctors, which will be very handy not only for further annotation work (for what concern decision making and consistency checking), but also for generic linguistic studies and the POS tagging problem itself. This annotation guide is part of the corpus documentation and was refined from what had already been produced in GRACE. The dry run corpus has been normalized and the result of the system combination has been built but no manual validation of the resulting material has been performed yet. Because the test corpus was bigger and more interesting since it had been annotated with the latest version of the GRACE morphosyntactic formalism we decided to work on this one first. It was normalized, then performance test where run to determine the best annotation combination procedure. The results showed that when comparing to the best results obtained by the participants in precision and in decision, it is possible to obtain a system with good performance for both precision and decision, when combining the annotations of the 5 systems with the best result in Precision out of the 15 systems (13 participants and 2 baseline approaches). Figure 1 shows the Precision/Decision performance range triangles of the system with the best precision result, of the system with the best decision result and of the system resulting from the combination of the 5 best systems. These results were measured using the reference data of the GRACE campaign.

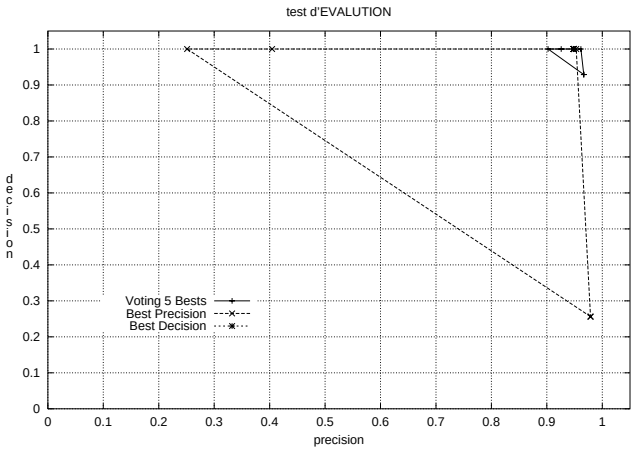

Figure 1: Precision/Decision performance range triangles of the system with best precision (big triangle), of the system with best decision (small star on top right corner) and of the combination of the 5 best precision systems (small triangle on top right corner.

In the first phase, manual validation was done only on 38,643 forms of the test corpus (out of 830000 forms, which represents roughly 4%) for which the system combination procedure had produced an ambiguous annotation for the main morphosyntactic category or the subcategory (independently of other morphosyntactic information like gender or number).

occ.	tags	check ?	Phase 1	Phase 2
Né	Afpms|Vmps-sm	1	Vm	ms-
à	Sp	0	-	- - -
Tarbes	Npfs|Npms	1	-	?s-
,	F	0	-	- - -

Table 4: Examples of POS tagging manual correction, at phase 1 and 2 of MULTITAG.

In a second phase of validation, all the forms whose annotations contained number, gender or person information (64,061 forms of the test corpus, roughly 8%) were manually checked.

The first release of the corpus has been delivered to ELRA for assessment and is now undergoing final quality validation while the latest administrative details are being cleared up for its distribution. Future work, will concern residual error analysis and improving future releases of the corpus, for instance by adding lemma information as at least four participants have provided lemma information and early measurement seem to indicate that lemma annotations differ only for approximately 10% of the word occurrences of the corpus.

occ. #	occ.	tag	Manu. checked?
00007	mes	Ds1fps	0
00008	fonctions	Ncfp	0
00009	m	Pp1msa/1.2	1
00010	’	Pp1msa/2.2	1
00011	ont	Vaip3p	0
00012	successivement	Rgp	0
00013	appelé	Vmpssm	0
00014	à	Sp	0

Table 5: An example of MULTITAG final data format.

5. Conclusion

The GRACE and MULTITAG experiments have proved that the paradigm of evaluation, when it uses black-box quantitative methods, can also function as a producing activity for low cost and very high quality linguistic resources. Such an approach could easily be generalized to other control task and provides the means to alleviate the cost of deploying evaluation on a large scale through the valorization of its by-products.

6. References

G. Adda., J. Lecomte, J. Mariani, P. Paroubek, and M. Rajman. 1997. Les procédures de mesure automatique de laction grace pour lévaluation des assignateurs de parties du discours pour le français. In *1 ères Journées Scientifiques et Techniques du Réseau Francophone de lIngenierie de la Langue de lAupelf-Uref*, Avignon, Avril.

G. Adda, J. Mariani, P. Paroubek, M. Rajman, and J. Lecomte. 1999. L’action grace d’évaluation de l’assignation des parties du discours pour le français. *Langues*, 2(2):119–129, June.

N.-O. Bernsen, N. Calzolari, J.-P. Chanod, K. Choukri, L. Dybkjær, R. Gaizauskas, S. Krauwer, I.de Lamberterie, J. Mariani, K. Netter, P. Paroubek, A. Popescu-Belis, M. Rajman, and A. Zampolli. 1999. A blueprint for a general infrastructure for natural language processing systems evaluation using semi-automatic quantitative black box approach in a multilingual environment. Deliverable 1.1, EU project LE4-8340, Evaluation in Language and Speech Engineering. http://www.limsi.fr/TLP/ELSE/ELSED11EN.HTML.

L. Chase. 1997. Word and acoustic confidence annotation for large vocabulary speech recognition. In *European Conference On Speech Communication And Technology (Eurospeech)*, pages 815–818, Rohdes, Greece, September.

J. G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *IEEE Workshop on Automatic Speech Recongition and Understanding*.

N. Ide and J. Véronis. 1994. Multext: Multilingual text tools and corpora. In *15th International conference on computational linguistics (COLING)*, pages 588–592, Kyoto, Japan.

G. Illouz. 1999. Méta-Étiqueteur adaptatif, vers une utilisation pragmatique des ressources linguistiques. In *Conférence Traitement Automatique du Langage Naturel*, pages 185–194, Cargèse, France., July.

A. Kilgarriff. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *1st International Conference on Language Resources and Evaluation (LREC98)*, volume 1, pages 581–585, Granada, Spain, May.

G. Leech and A. Wilson. 1995. Eagles morphosyntactic annotation - eag-csg/ir-t3.1. Technical report, Consiglio Nazionale delle Ricerche, Instituto di Linguistica Computazionale, Pisa.

J. Mariani and P. Paroubek. 1999. Human language technologies evaluation in the european framework. In *DARPA Broadcast News Workshop, ISBN-1-55860-638-6*, pages 237–242, Whashingthon, February. Morgan Kaufman Publishers.

Marquez and Padro. 1998. On the evaluation and comparison of taggers: the effect of noise in test corpora. In *COLING/ACL*, Montreal.

Pallett D. S. 1998. The nist role in automatic speech recognition benchmark tests. In *1st International Conference on Language Resources and Evaluation (LREC98)*, volume 1, pages 433–441, Granada, Spain, May.

H. Schwenk. 1999. Using boosting to improve a hybrid hmm/neural network speech recognizer. In *IEEE International Conference On Acoustics, Speech, and Signal Processing.*, Phoenix, USA, March.

P. K. Simpson. 1990. *Artificial Neural Systems - Foundations, Paradigms, Applications and Implementations.* Pergamon Press, 1 edition.

K. Sparck-Jones and J.R. Galliers. 1995. *Evaluating Natural Language Processing Systems*. Springer-Verlag.

D. Tufis, 1999. *Tiered Tagging and combined classifier*, chapter Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence 1692, Jelineck F. and Nörth E. eds. Springer.