•

Sublanguage Dependent Evaluation: Toward Predicting NLP performances

Gabriel Illouz

LIMSI – CNRS BP 133 91403 ORSAY Cedex France gabrieli@limsi.fr

Abstract

In Natural Language Processing (NLP) Evaluation, such as MUC (Hirshman, 1998), TREC (Harman, 1998), GRACE (Adda et al., 1997), SENSEVAL (Kilgariff, 1998), metrics on the performances, such as precision, recall, or f-measure are used. Nevertheless, performance results are often average measurements computed over the complete test. They do not give any clues about the system's robustness. We conceive evaluations being not only a processs to show how good the systems are on a given dataset, but also as an aid for choosing which system or approach to use to build a NLP application for a specific subset of the language. In this case, knowing which system performs better on average does not help us to find which is the best for a given subset of a language. As a matter of fact, this aspect of the reuse paradigm is rarely investigated in the litterature about workbenches especially designed to adapt quickly to new language resources, such as GATE (Cunningham, 1997), In the present article, the existing approaches which take into account language heterogeneity and offer methods to identify sublanguages are presented. Then we propose a new metric to assess robustness and we study the existence of a correlation between the performance variations observed for POS tagging and the different sublanguages identified in the Penn Tree Bank Corpus. The work we present here is a first step in the development of predictive evaluation methods, intended to propose new tools to help in determining in advance the range of performance that can be expected from a system on a given dataset. **keywords** : (predictive) evaluation, POS tagging, textual typology, sublanguages, performance variations.

1. Introduction

In Natural Language Processing (NLP) Evaluation, such as MUC (Hirshman, 1998), TREC (Harman, 1998), GRACE (Adda et al., 1997), SENSEVAL (Kilgariff, 1998), metrics on the performances, such as precision, recall, or fmeasure are used. Nevertheless, performance results are often average measurements computed over the complete test. They do not give any clues about the system's robustness. We conceive evaluations being not only a processs to show how good the systems are on a given dataset, but also as an aid for choosing which system or approach to use to build a NLP application for a specific subset of the language. In this case, knowing which system performs better on average does not help us to find which is the best for a given subset of a language. As a matter of fact, this aspect of the reuse paradigm is rarely investigated in the litterature about workbenches especially designed to adapt quickly to new language resources, such as GATE (Cunningham, 1997), In the present article, the existing approaches which take into account language heterogeneity and offer methods to identify sublanguages are presented. Then we propose a new metric to assess robustness and we study the existence of a correlation between the performance variations observed for POS tagging and the different sublanguages identified in the Penn Tree Bank Corpus. The work we present here is a first step in the development of predictive evaluation methods, intended to propose new tools to help in determining in advance the range of performance that can be expected from a system on a given dataset.

2. Related works

In the following, *text genre* is employed when the classification is a priori (given by humans), to differentiate from *text type* when it is induced.

In (Slocum, 1986), Slocum shows that different syntactic rules should be used as a function of two sublanguages in German. They are composed of two manuals written by engineers and two brochures written by salesmen. He also proposes a means to automatically characterize the type "manual" (imperative, acronyms, determinants suppression) in comparison with the type "brochure". (long sentences, use of pronouns, richer syntax). This kind of approach seems particularly interesting. Indeed, it allows to define the notion of competence domain of a NLP.

In (DeRose, 1988), the author gives the results according to text genre present in the BROWN, which is interesting to look at the variation from one genre to another, the problem is that his tagger have the same test and train set, making harder to view text genre effect.

In (Biber, 1993), Biber shows that the text genre may have an effect upon the stochastic taggers' results. To perform that, he selects two text genres (expository texts and novels) from the LOB corpus. He gives probability differences of some tags and tags bigram sequences from one text genre to the other. So, language heterogeneity could be lessen using existing or induced classification when it is present for a corpus.

Following this paradigm, Sekine, (Sekine, 1998) trains his syntactic analyzer for each text genre present in the BROWN Corpus. The conclusion is that the performance are always better using a testing set of the same class than the one used for training.

The problem of such an approach is first that corpora we have to deal with are not always classified in any ways (for example Information Retrieval corpora are often raw data), and second that there is no reason why an existing classification should lead to the best way to diminish performance variation.

Partition 1	Nb		Partition 2	Nb		Partition 3	Nb
Informative Prose	374	95.1 (1.5)	Press	88	95.06 (1.10)	A. Press : Reportage	44
						B. Press : Editorial	27
						C. Press : reviews	17
			Misc	176	95.10 (1.46)	D. Religion	17
						E. Skills and Hobbies	36
						F. Popular Lore	48
						G. Belles Lettres, etc	75
			Non-fiction	110	95.18 (1.70)	H. Gov. doc. & misc.	30
						J. Learned	80
Imaginative Prose	126	93.96 (1.1)	Fiction	126	93.96 (1.12)	K. General Fiction	29
						L. Mystery	24
						M. Science Fiction	6
						N. Adv.& Western	29
						P. Romance	29
						R. Humour	9

Table 1: Performance Variations according to text genre

Approaches dealing with text genre classification exist in the literature. In (Biber, 1993), the text genre is induced using a tagger, which is problematic in the present context. In (Karlgren and Cutting, 1994) and (Brett et al., 1997), the authors propose a method using simpler features to extract from raw text. In (Folch et al., 2000), the authors propose a generic architecture to study set of feature to induce the best classification.

3. Preliminary Experiments

3.1. Need for a Performance Variation Metrics (PVM)

As preliminary experiments, the performance variations of public domain tagging tools such as the TREETAGGER (Shmid, 1995) on the BROWN CORPUS, part of PENN TREE BANK CORPUS (Marcus Mitchell, 1993) are studied. If we look at the results, the global precision is 94.6%, but important variations are present. Indeed, precision varies from 0.85% to 0.98%. The histogram of the number of text for each range of performance is given in the figure 1.

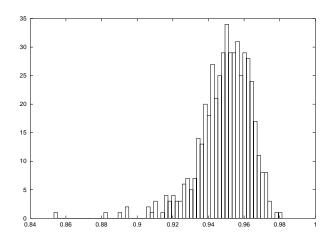


Figure 1: Histogram of performance for tagging

From this experiment, we see the need for a Performance Variation Metric (PVM). We propose a basic one. As the end-user of a system would like to know what kind of performance he/she could expect from a NLP system, we don't use an entropy based method but rather descriptive statistic based method giving the average, and the **standard deviation** which gives a better idea of the most probable cases¹. The amplitude of results (a min-max couple) could be added. In the previous example, the TreeTagger has a standard deviation of 1.5%. This variation needs to be investigated.

3.2. Effect of Existing Classification on PVM

Using existing classification of the BROWN CORPUS: "Imaginative Prose" versus "Informative Prose", we compute our metrics on these two partitions. Surprisingly, even if we find that the TREETAGGER trained on a Newspaper corpus (WALL STREET JOURNAL) does perform better on Informative Prose (95.1%) than Imaginative Prose (94.0%), it is more homogeneous on Imaginative Prose (Min-Max:90.1%-96.2%, PVM: 1.1%) than on Informative Prose (Min-Max:85.4%-98.1%, PVM: 1.5%). Even if we do not lessen drastically the performance variation, we obtain two different average performances, with inferior or equal variations.

At a second level of partitioning, as described in (Karlgren and Cutting, 1994), the finer-grained results are presented in table 1. From these, as intuition would confirm, the two most homogeneous partitions are the Press and fiction one, to be opposed to Miscellaneous and Non-fiction one.

3.3. Effect on Induced Classification on PVM

Then, as an on-going work, we show that we can obtain the same kind of variations using induced classification, based on the model developed for the ELRA/TYPTEX project: a text typology profiler based on Quantitative Linguistics methods (Biber, 1993) to infer text genre from raw data. As a first step, before using sophisticated linguistics features, we use ASCII character set to induce a partition of the data using unsupervised method based on entropy

¹In the case of a normal distribution, 95% of the sample will be in a zone defined by the average plus or minus 1.96 standard deviation

						Total
	Imaginative	Informative				
	Fiction		Press	Misc	Non-Fiction	
Perf	81.29(2.07)	82.27(3.02)	80.78(2.62)	82.59(2.98)	82.95(3.00)	82.02(2.84)
Unknown	4.86(1.72)	5.46(2.83)	6.58(2.59)	4.86(2.51)	5.53(3.21)	5.31(2.61)
Direct	43.35(2.24)	43.65(3.04)	43.20(2.44)	43.86(3.10)	43.69(3.32)	43.58(2.86)
Ambiguous	51.15(2.98)	49.68(3.55)	48.82(3.33)	50.21(3.24)	49.50(4.01)	50.05(3.47)
EasyResolution	37.93(2.78)	38.29(3.16)	37.21(3.02)	38.52(2.85)	38.77(3.52)	38.20(3.07)
Error	12.61(1.69)	10.85(1.40)	11.01(1.09)	11.10(1.46)	10.33(1.39)	11.29(1.67)
Transcat	1.25(0.69)	1.73(1.04)	1.98(0.91)	1.70(1.04)	1.58(1.09)	1.61(0.98)

Table 2: Error Analysis according to a dumb baseline

(Jardino and Beaujard, 1997). We look for 2 classes, which we call IC1, IC2.

	IC1	IC2	Total
Informative prose	53	321	374
Imaginative prose	125	1	126
Total	178	322	500

Table 3: Number of Texts according to Partitions

The two classes obtained are quite similar to the existing ones (less than 11% of texts are classified differently). The variations for the two classes are given in table 4. and could be compared to the one obtained on the existing classification. The variation are a little lessen using this partition.

	IC1	IC2	Total
Informative prose	94.8 (1.3)	95.1 (1.5)	95.1 (1.5)
Imaginative prose	93.4 (1.1)	93.9 (0.0)	93.9 (1.1)
Total	94.2 (1.3)	95.1 (1.5)	94.8 (1.5)

Table 4: Variations of performance according to Partitions

4. Qualitative analysis of the results

What these changes are due to? To study them, we define a baseline methods on these data using CELEX lexicon (CLR), we could then distinguish 4 main different cases, with two subcases):

- The token has no entry in the lexicon (Unknown),
- The token entry has only one POS and it matches (Direct)
- The token entry has more than one POS (Ambiguous)
 - it is the first (Easy Resolution)
 - it is another (Error)
- The POS that should be find is not in the choices (TransCategorisation)

This analysis frame according to this dumb baseline let us make a more precise analysis. We give in table 2 the proportion and standard deviation for each case. This results confirm the previous one in term of heterogeneity of the data used.

We see the difficulties to find some stable ratios explaining the performance variations. From a raw text, only Unknown, multiple-tag and single-tag token could be obtained.

5. Toward a new task: predicting performance

5.1. method

To give a more precise idea on how this results could be used. A methodology is defined. We look for a predictive function such as :

$$predict(\overrightarrow{V_{Text}}) = s \pm \varepsilon$$

where s is a performance score

 ε is a number representing an accepted variation

 $\overrightarrow{V_{Text}}$ is a vector representing a count made on various features on a given Text.

To evaluate this subtask of evaluation, we could on a test set, compute in how many case the prediction was correct and the average ε used, the smaller the better.

5.2. results

The 500 samples collection is here divided in a train set of 444 texts, and 56 text randomly selected. We then compute the result of predicting method for the existing classification. The result for an ε of 1.65 * σ (corresponding to 10% error), and 1.96 * σ (corresponding to 5% error) is given. We see that according to the small number of text the results are quite close of what could be expected.

	Train	nb	Test	(1.65σ)	(1.96σ)
Total	94.8(1.5)	444	56	94.6	98.2
Info.	95.1(1.5)	332	42	92.8	95.2
Imag.	93.9(1.1)	112	14	85.7	92.8
Press	95.1(1.1)	79	9	88.8	100
Misc	95.1(1.4)	154	22	90.9	95.4
NonF.	95.1(1.7)	99	11	90.9	100
Fiction	93.9(1.1)	112	14	85.7	92.8

6. Conclusion and future works

To conclude, we show that it is possible to partition the data in way where the evaluation results are more homogeneous (reduced standard mean deviation) inside each subset than on the global set. Furthermore, this partition can be based either from an existing classification or from an induced one, using easy to extract features. This opens new perspectives in the reuse of evaluation results to help in choosing wich approach is the best suited for handling a specific language subset.

From these preliminary experiments, it seems possible to have, at no extra cost, evaluation campaigns which will provide an idea of the robustness of the participating systems with respect to data heterogeneity.

Future works will concern extensive tests, on larger data, using different feature sets to induce the most appropriate classification and the application of our methodology to other control tasks than POS tagging.

7. References

- Gilles Adda, Josette Lecomte, Joseph Mariani, Patrick Paroubek, and Martin Rajman. 1997. Les procédures de mesures automatique de l'action grace pour l'évaluation des assignateurs de parties de discours pour le fran cais. In *JST-FRANCIL*, Avignon.
- Douglas Biber. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):243–258.
- Kessler Brett, Nunberg Geoffrey, and Hinrich Schuetze. 1997. Automatic detection of text genre. In *Proceedings ACL/EACL*, pages 32–38, Madrid.
- R. Cunningham. 1997. Software infrastructure for natural language processing. In *ANLP*'97.
- Stephen J. DeRose. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.
- Helka Folch, Serge Heiden, Benoit Habert, Serge Fleury, Gabriel Illouz, Pierre Lafon, Julien Nioche, and Sophie Prévost. 2000. Typtex : Inductive typological text classification by multivariate statistical analysis for nlp systems tuning/evaluation. *Computational Linguistics*.
- D. Harman. 1998. The text retrieval conference (trecs) and the cross-language track. In *LREC*, Grenade, Espagne.
- L. Hirshman. 1998. Language understanding evaluations: Lessons learned from muc and atis. In *LREC*, Grenade, Espagne.
- Michele Jardino and Christelle Beaujard. 1997. Role du contexte dans les modèles de langage 'n-classes' application et évaluation sur mask et railtel. In *Actes des 1 ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'Aupelf-Uref.*
- Jussi Karlgren and Douglas Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING*, pages 1071–1075, Kyoto, Japan.
- Adam Kilgariff. 1998. Senseval: An excercise in evaluating word sense disambiguation programs. In *LREC*, Grenade, Espagne.
- Mary Ann Marcinkiewicz Marcus Mitchell, Beatrice Santorini. 1993. Building a large annoted corpus of en-

glish: the penn treebank. *Computational Linguistics*, 19(1):313–330.

- Satoshi Sekine. 1998. The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington, march-april.
- Shmid. 1995. Improvements in part of speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*.
- Johnathan Slocum. 1986. How one might automatically identify and adapt to a sublanguage. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains*, chapter 11, pages 195–210. Lawrence Erlbaum Ass., Hillsdale, NJ.