Evaluating the Annotated Corpora using 'CorpEvalSystem'

Mohamed Yoonus. M

Senior Lecturer / Junior Research Officer

yoonussoft@gmail.com

Linguistic Data Consortium for Indian Languages (LDC-IL)

Central Institute of Indian Languages – Mysore

Index

Introduction

Introduction: Evaluation

- Evaluation plays a vital role in every field. It is an indispensable activity to assure the quality of the task done.
- The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system.
- The purpose of evaluation is to provide an assessment of the value of a solution to a given problem.
- One large problem of each annotation project is consistency. The consistency of data development process involves a number of steps including the inter-annotator agreement, finding the confusion labeling of results and validation of final result in the annotated corpora.

Continued…

This Paper:

- Explores different kinds of evaluation with respect to tagged corpus.
- Describes about "CorpEvalSystem" and its algorithm which has been developed for language independent but tagset dependent.
- Provides the different aspects of evaluation results of Tagger along with their discussion.
- Finally, this paper highlights need of several measurement factors including the accuracy calculation, information retrieval metrics, confusion matrix and ambiguous words analysis are required for evaluation of annotated corpora.

Types of Evaluation

Types of Evaluation

- The European project EAGLES* (King and Maegaard, 1998) distinguishes three kinds of evaluation:
  - (1) *progress evaluation*, where the current state of a system is assessed against a desired target state; [undertaken by either researchers/developers or by potential users]
  - (2) *adequacy evaluation*, where the adequacy of a system for some intended use is assessed; [performed by potential users and/or purchasers of systems (individual, companies, or agencies)]
  - (3) *diagnostic evaluation*, where the assessment of the system is used to find where it fails and why. [concern mainly of researcher and developers]

*Expert Advisory Group on Language Engineering Standards

Intrinsic vs. extrinsic evaluation

- Among other general types of evaluation, a number of distinctions are traditionally made in evaluation methodologies on the basis of evaluation procedures:

- An intrinsic evaluation would run the POS tagger on some labeled data, and compare the system output of the POS tagger to the gold standard (correct) output.

- An extrinsic evaluation would run the POS tagger with some other POS tagger, and compare the accuracy of results.

Black-box vs. glass-box evaluation

- Black-box evaluation only sees the final output and its relationship to the original input.
- It measures a number of parameters related to the quality of the process (speed, reliability) and to the quality of the result (e.g. the accuracy of data annotation).
- Glass-box evaluation looks at the design of the system, the algorithms that are implemented, the linguistic resources it uses (e.g. vocabulary size), etc.
- It provides more informative with respect to error analysis or future development of a system.

Automatic vs. manual evaluation

- In many cases, automatic procedures can be defined to evaluate an NLP system by comparing its output with the gold standard one.
- Manual evaluation is performed by human judges, who are instructed to estimate the quality of a system and automatic evaluation is performed by system tools.
- The automatic evaluation is sometimes referred to as *objective* evaluation, while the human being appears to be more *subjective*.

Measurements Terms

Measurements Terminology

- For the measurements, we should consider several factors while evaluating the system - including the accuracy of the tagger, information retrieval metrics, and confusion matrix results.

Accuracy:

- Accuracy is defined as the ratio of the number of word forms correctly tagged over the total number of word forms tagged.

- For example, a common noun classifier that predicts the correct label 75 times in a test set containing 80 common nouns would have an accuracy of 75/80 = 93.6%.

Measurements Terminology

- ***Precision:*** is defined as a measure of proportion of the selected items that the system got right

$$Precision=TP/(TP+FP)$$

- ***Recall:*** is defined as the proportion of the target items that the system selected

$$Recall=TP/(TP+FN)$$

- ***F- Score:*** a measure that combines precision and recall

$$F\text{-}Score=2 * \frac{(Precision * Recall)}{Precision + Recall}$$

NB:TP-True Positive, FP-False Positive, FP-False Negative

Measurements Terminology

Confusion matrix:

- It is a visualization tool typically used in supervised learning.
- Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing between two classes (i.e. commonly mislabeling one as another).

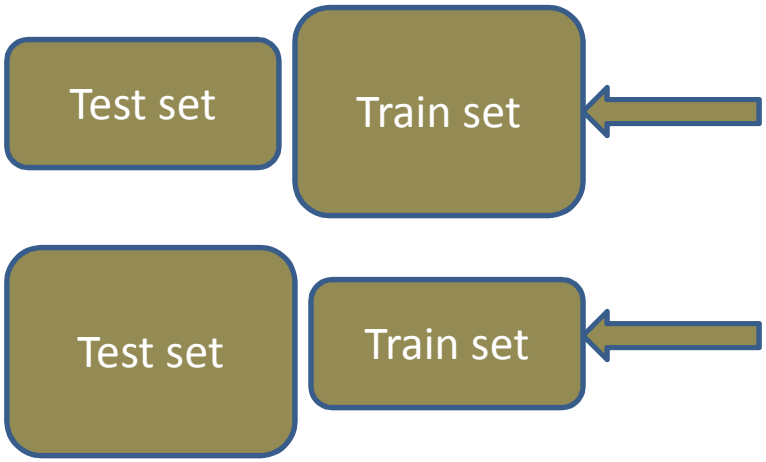
		Predicted class	
		X	Y
Actual class	X	150	25
	Y	15	300

- correct predictions→ 450 (150 + 300).
- incorrect predictions→40 (25 + 15).
- error rate → 40/490 = 0.0816
- overall accuracy rate → 450/490 = 0.9183.

Measurements Terminology

Cross-Validation:

- ***Problem:***



- One **solution** for this is to perform multiple evaluations on different test sets and then, combine the scores from those evaluations, a technique known as **cross validation**.

- In other words, we subdivide the original corpus into *N* subsets called **folds**.

Software Description

Main Algorithm: Confusion Matrix

1. Start the process
2. Read either a file or set of annotated files (gold standard) from the machine.
3. Separate each token into words and tag labels.
4. Store the values in the first list (L1).
5. Read either a file or set of annotated files (system tagged) from the machine.
6. Repeat step 2.
7. Store the values in the second list (L2).
8. If the token size of L1 is equal to the token size of L2 then
    a. Iterate the process to view the confusion result
    b. Go to 'Confusion_View' subroutine, if algorithm satisfies any one of the following conditions:
        i. If user wants to view the error result only, it checks the 'token1' is not equal to the 'token2' then
        call the confusion_view(Token1, Token2)
        ii. If user wants to view the correct result only, it checks the 'token1' is equal to the 'token2' then
        call the confusion_view(Token1, Token2)
        iii. If user wants to view the both result at the same time then,
        Simply, call the confusion_view(Token1, Token2
    c. End of loop
9. Otherwise exit
10. End process

Subroutine: Confusion_View

Confusion_View(Token1, Token2)

Begin

1. Iterate from the first row to the last row of dataset
    a. If row[index] of item[0] value in dataset is equal to Token1
        i. Iterate from the first colum to the last column of dataset
            1. If the column name of column[index] is equal to Token2
                a. CountValue = CountValue + 1
                b. Assign previous value of [row,col]=0
                c. PreviousValue= current value of [row,col] of datset
                d. Update [row,col] of dataset= PreviousValue + CountValue
            2. End if
        ii. ColumnIndex = ColumnIndex + 1
        iii. End loop
    b. End if
2. Reset the CountValue = 0
3. End loop

End

Algorithm

- Apart from the main algorithm, we also use the some other kinds of algorithm to evaluate the POS tagger

- Error Classification: find the right analysis and wrong analysis , calculate the overall accuracy

- Corpus Statistics: number of tokens in test corpus, number of tags in test corpus, etc.,

- Ambiguous words analysis: frequency of ambiguous words and their percentage of coverage in test corpus

Software: CorpEvalSystem

- We have developed a GUI based 'Annotated Corpus Evaluation System' (CorpEvalSystem) for evaluating the different version of same annotated corpora using similar tagset.

- The software or tool takes two kinds of annotated corpora as input, and compares both. Then, it produces the analyzed result as the output.

It evaluates:

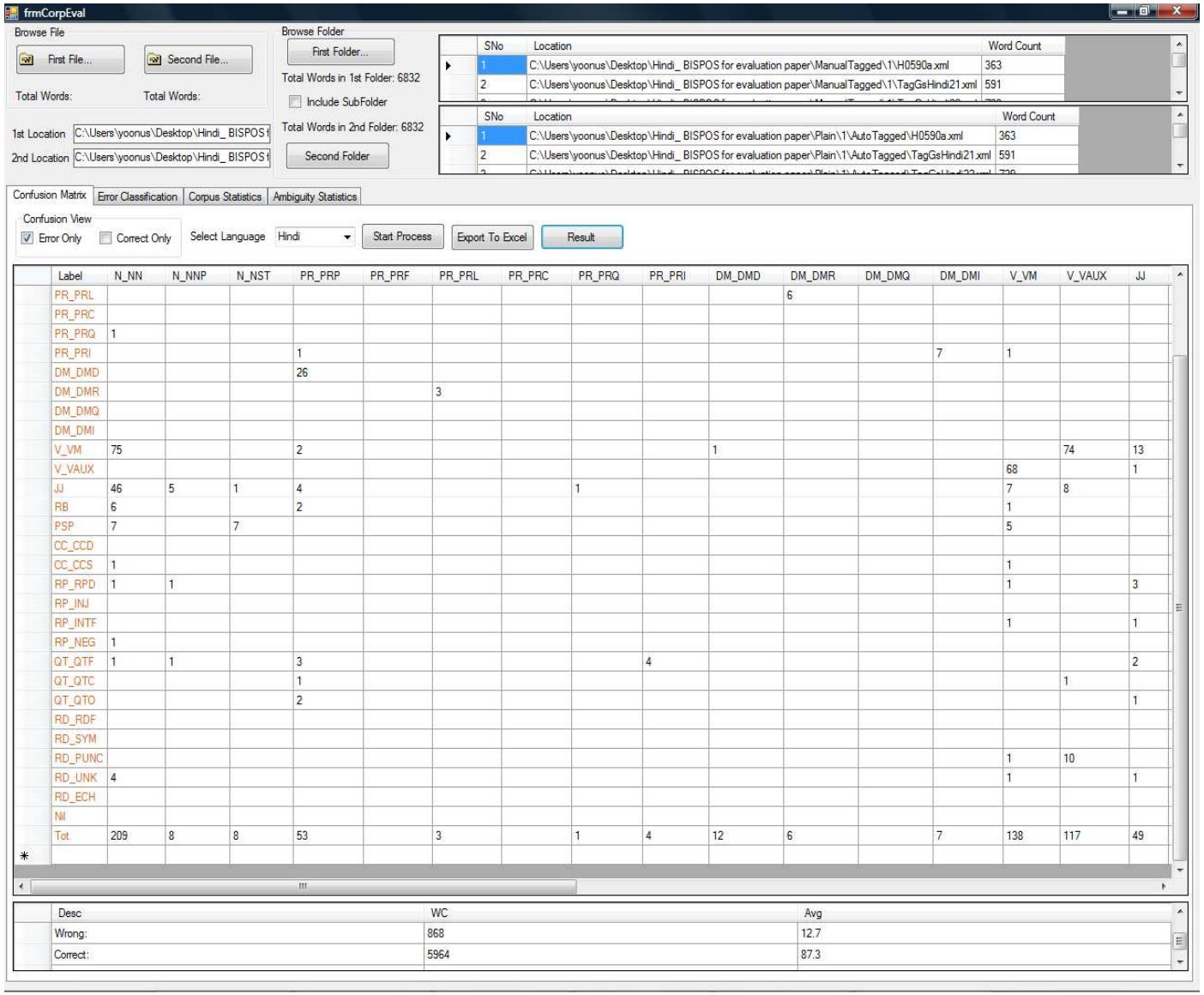
- gold standard corpora with tagger tagged corpora;
- two sets of tagged corpora produced by two different taggers, or
- two sets of manually annotated corpora by different annotators.

Software: CorpEvalSystem

Advantages:

- File(s) Read: It can be read either a single file or multiple files of two kinds of tagged corpora.
- Well designed layout: The GUI layout has designed for different size of screen.
- Customizable tag set: Tag set of languages can be customized by user.
- Tab navigation: User can easily navigate from one window to another.
- Export option: Export the results into excel format.
- Using this system, we can measure the accuracy of data, calculate the precision, recall and F-score values, and accomplish the cross folder validation and get the confusion matrix view. In addition to these, we can get the ambiguous words list with their frequency from the system.

CorpEvalSystem: GUI



Evaluation Results & Discussion

Evaluation Results and Discussion

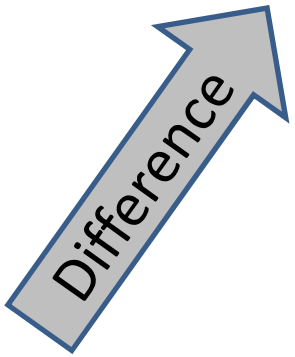
- Conducted experiment for LDC-IL POS Tagger
- Hindi Corpora (69,723 tokens =test set + training set)
- In these evaluation results, initially we compared two sets of tagged corpora and calculated the accuracy of the tagger.

Evaluation Result-1		
Description	**Word**	**Percentage**
Total Words	6832	100
Wrong	868	12.7
Correct	5964	87.3

Evaluation Results and Discussion

- Subsequently we performed a cross-validation experiment. Whole data is divided into ten data sets.
- It was noted that evaluation of each %age of result varies among different fold of data sets.
- Hence, Cross- validation evaluation technique can be quite suitable for obtain the reliable result.

- 87.3 - 86.9 = 0.4



Evaluation Result-2		
S. No	**Folds**	**Percentage**
1	Fold-1	87.3
2	Fold-2	85.6
3	Fold-3	88.2
4	Fold-4	87.2
5	Fold-5	89.1
6	Fold-6	84.7
7	Fold-7	87.5
8	Fold-8	86.3
9	Fold-9	86.4
10	Fold-10	87.1
Total		**86.9**

Evaluation Results and Discussion

- To evaluate the system, we have used the standard Information Retrieval (IR) metrics of Precision, Recall and F-Score.
- The Precision, Recall and F-score evaluation results as shown in Table . This is the average result of the ten folders.

Evaluation Result-3	
Description	**Percentage**
Precision	86.9
Recall	99.6
F-Score	93.0

Evaluation Results and Discussion

- The error analysis report provides information about the nature of error that the system makes.
- In the present experiment, to ascertain the nature of error with respect to the POS tags assigned by LDC-IL hybrid POS Tagger, we have used confusion matrix method.
- Table shows some part of error analysis result for the first folder that contains 6832 tokens. The vertical labels denote the gold standard corpora and the horizontal labels denote the auto tagged corpora.

Part of error analysis result(Hindi)

GS/Auto	N_NN	N_NNP	N_NST	PR_PRP	PR_PRF	PR_PRQ	PR_PRI	V_VM	V_VAUX	JJ	RB	PSP
N_NN		1		7				41	17	16		46
N_NNP	62			5				10	6	11		24
N_NST	1											17
PR_PRP	3								1			
PR_PRF												
PR_PRQ	1											
PR_PRI				1				1				
V_VM	75			2					74	13		25
V_VAUX								68		1		1
JJ	46	5	1	4		1		7	8		6	14
RB	6			2				1				4
PSP	7		7					5			1	

Evaluation Results and Discussion

- On the basis of the confusion matrix, it was found that the most of the errors occur with respect to Noun, Verb, Adjective and Postposition categories in the tested language (Hindi).

- It is often the case that, in a language, Common Noun and Proper Noun are often tagged reverse. The similar misappropriation of tags is witnessed between Main Verb and Auxiliary Verb, and Adjective and Noun. The Noun, Verb and Adjective categories are confused with Postposition.

- Through this evaluation, we can study and analysis the tagger for further improvements of the system.

Evaluation Results and Discussion

- Apart from the accuracy calculation, cross-validation and confusion evaluation tests, it will become very clear that the evaluation also depends on the frequency analysis of the data and size of ambiguous words.

- Moreover, not only the size of the corpus, but also its type can have an influence on the accuracy measure.

- Furthermore, in this evaluation, we found that 25539 (Percentage 36.63) words are ambiguous out of total 69723 words.
- The following table shows that the top 20 ambiguous words out of 764 distinct ambiguous words which were extracted from the annotated corpus using evaluation tool.

Ambiguous words: top 20 frequency

S.No	Word	Tag	Count	Percentage
		Total Tokens	69723	100.000
1	है	[V_VAUX-1048] [V_VM-655]	1703	2.443
2	की	[PSP-1317] [V_VM-77]	1394	1.999
3	से	[PSP-1081] [RP_RPD-2]	1083	1.553
4	और	[CC_CCD-976] [JJ-9] [QT_QTF-27] [RP_INTF-2]	1014	1.454
5	हैं	[V_VAUX-565] [V_VM-144]	709	1.017
6	तो	[CC_CCD-8] [CC_CCS-318] [RP_RPD-269]	595	0.853
7	पर	[CC_CCD-66] [PSP-468]	534	0.766
8	था	[V_AUX-1] [V_VAUX-322] [V_VM-129]	452	0.648
9	हो	[V_VAUX-70] [V_VM-357]	427	0.612
10	वह	[DM_DMD-25] [PR_PRP-383]	408	0.585
11	यह	[DM_DMD-141] [PR_PRP-202]	343	0.492
12	थे	[V_VAUX-254] [V_VM-80]	334	0.479
13	कर	[V_VAUX-85] [V_VM-238]	323	0.463
14	इस	[DM_DMD-240] [PR_PRP-48]	288	0.413
15	थी	[V_VAUX-156] [V_VM-88]	244	0.350
16	लिए	[PSP-228] [V_VAUX-2] [V_VM-3]	233	0.334
17	वे	[DM_DMD-13] [PR_PRP-213]	226	0.324
18	कुछ	[PR_PRI-96] [QT_QTF-120]	216	0.310
19	जो	[CC_CCS-1] [DM_DMR-66] [PR_PRL-141]	208	0.298
20	गया	[V_VAUX-195] [V_VM-9]	204	0.293

Evaluation Results and Discussion

- Therefore, the accuracy of tagger depends upon a number of factors
  - ❑ corpus type and its size
  - ❑ size of ambiguous words
  - ❑ Tagset
  - ❑ methodology used.

- Consequently, the test results may also vary depending on the different types and sizes of the test corpus.

Conclusion

Conclusion

- Algorithms were implemented in C# using Visual Studio 2008.
- Provided a perspective on the overview of evaluation and its types.
- The core part of evaluation system algorithm has been explained. The measurement techniques and the software are also briefly elucidated.
- Corpus based automatic evaluation procedures provided most of the useful information regarding accuracy of data, confusion and correctness of system tagged data, information retrieval metrics and ambiguous words analysis.

Continued…

In future, it is intended

- to compare the LDC-IL tagger with other taggers.
- to find the accuracy of known and unknown words of the tagger.
- to add more functionality in the ‘CorpEvalSystem’ for evaluating tagger.

Acknowledgements

Acknowledgements

- My special thanks to Dr. L. Ramamoorthy and Er. M. Venkatesan for valuable advice and encouragement.

- I would also like to thank Dr. Richa for her valuable guidance and suggestions.

- Last but not least, thanks to LDC-IL Team for support

Thank You
&
?, if any

Mohamed Yoonus. M

Senior Lecturer/ Junior Research Officer

yoonussoft@gmail.com

Linguistic Data Consortium for Indian Languages (LDC-IL)
Central Institute of Indian Languages – Mysore