Evaluation of Inter-parser agreement

Anahita Poulad, Melpomeni Chioutakou

Supervisor: Yannick Parmentier

INTRODUCTION

A recurring issue observed in many NLP applications, is the inability to accurately evaluate a parser's performance on un-annotated data. The need for an automated method that provides meaningful insight about the reliability of a parser in real-world unlabeled data, is the motivation behind this project.

Our hypotheses:

- **An annotation is reliable if parsers seem to agree on the dependencies assigned.**

METHODOLOGY

We decided to use LAS, UAS and POS accuracy to compare parses, as well as comparing parser results with the gold standard parse.
To compare scores, we used Pearson Correlation Coefficient (Pearson's r) to see how orrelated the agreement of pairs of parses are. This correlation will then let us explore inter-parser agreement as a way of assessing the reliability of a parser in the absence of gold-parses.

PARSERS

Parser train-set	POS	UAS	LAS
EWT	95.40	86.22	83.59
GUM	95.89	87.06	83.57
LinES	96.88	85.82	81.97
ParTUT	96.15	90.31	87.35

English pretrained parsers

Parser train-set	POS	UAS	LAS
GSD	97.30	91.38	89.05
ParTUT	96.60	90.71	88.37
Sequoia	98.19	90.47	88.34
Spoken	95.49	75.82	70.71

French pretrained parsers

ANNOTATED DATA

Treebank	Domain(s)	Size (Tokens)
ESL	leaner essays	97K
EWT	social networks	254K
GUM	multi	134K
GUMReddit	reddit	16K
LinES	literature, manuals and Europarl	94K
ParTUT	talks, legal texts and Wikipedia	49K
Pronouns	grammar examples	1K
PUD	news and Wikipedia	21K
Total		666K

Used English UD treebanks

Treebank	Domain(s)	Size (Tokens)
FQB	questions	23K
FTB	newspaper	573K
GSD	news, reviews and Wikipedia	400K
ParTUT	talks, legal texts and Wikipedia	28K
PUD	news and Wikipedia	24K
Sequoia	medical, news and Wikipedia	70K
Spoken	spoken language	34K
Total		1152K

Used French UD treebanks

FINDINGS

	combined	ewt	gum	linES	partut
combined		0.268 0.322	0.218 0.263	0.392 0.380	0.279 0.256
ewt	0.258 0.315		0.263 0.311	0.401 0.376	0.303 0.278
gum	0.300 0.503	0.310 0.505		0.553 0.566	0.525 0.494
linES	0.401 0.598	0.392 0.587	0.473 0.568		0.493 0.514
partut	0.503 0.662	0.501 0.658	0.475 0.570	0.519 0.566	

POS

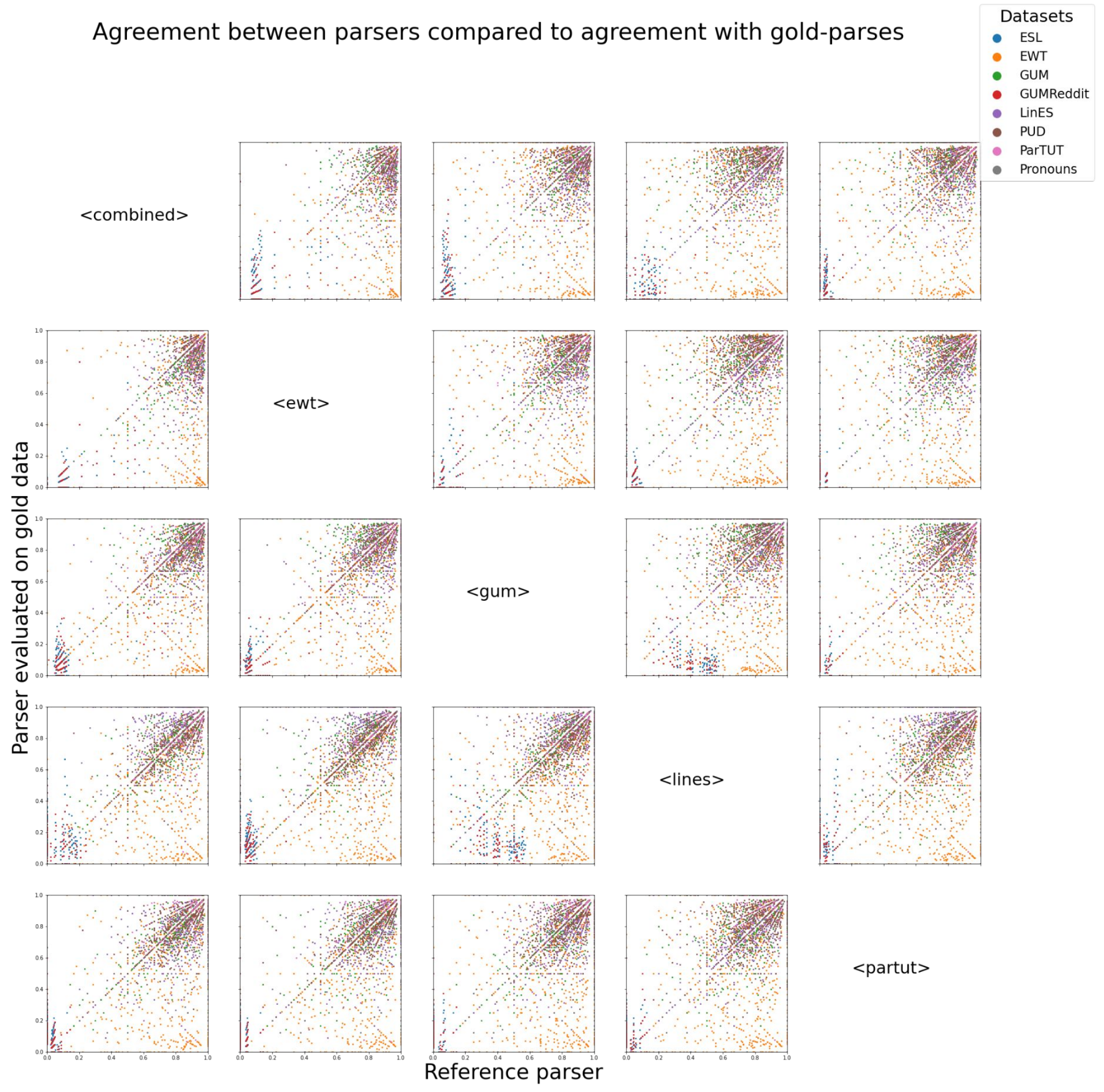
	combined	ewt	gum	linES	partut
combined		0.271 0.274	0.340 0.305	0.302 0.323	0.283 0.284
ewt	0.400 0.411		0.442 0.395	0.374 0.382	0.341 0.335
gum	0.480 0.526	0.465 0.510		0.318 0.398	0.383 0.398
linES	0.494 0.575	0.498 0.559	0.375 0.467		0.465 0.497
partut	0.476 0.555	0.471 0.533	0.452 0.477	0.480 0.497	

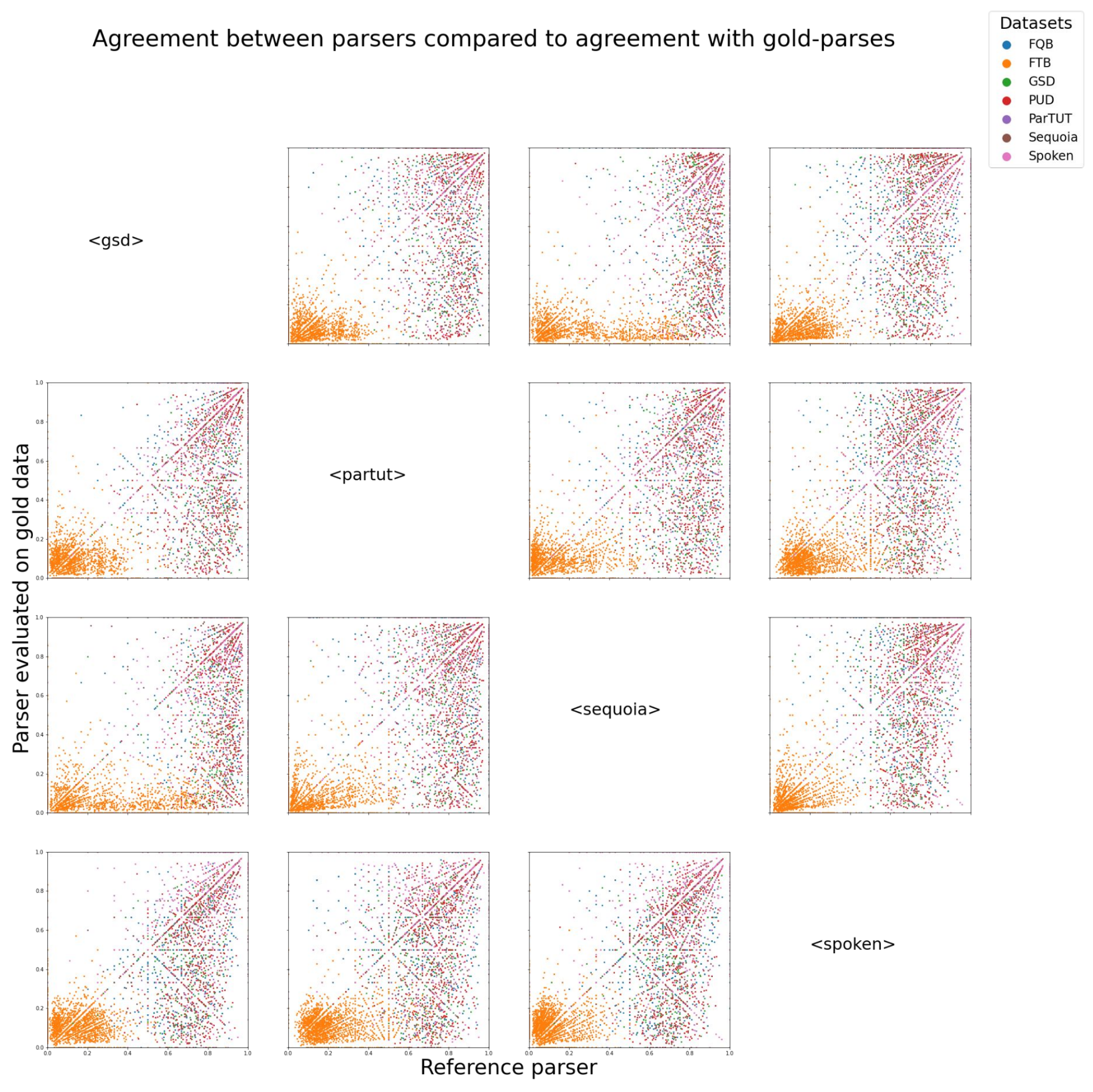
UAS

	combined	ewt	gum	linES	partut
combined		0.356 0.343	0.435 0.386	0.445 0.416	0.365 0.348
ewt	0.430 0.447		0.487 0.450	0.445 0.446	0.403 0.398
gum	0.587 0.608	0.555 0.576		0.495 0.503	0.466 0.465
linES	0.619 0.656	0.588 0.636	0.551 0.570		0.533 0.550
partut	0.608 0.670	0.593 0.652	0.582 0.606	0.562 0.593	

LAS

UAS-IPA vs UAS-Gold-Eval





CONCLUSION

Inter-parser agreement is a promising technique for assessing a parser performance, especially when there is no gold data available for evaluation. To obtain good correlation between IPA and gold-agreement:

- Parsers must be different enough (IPA is too optimistic)
- Data should not be closer to the training data of the assessed parser than the reference parser. (IPA is too pessimistic)