Text Mining of Transcription Factors to Proteins Interactions

Ashish Baghudana

Supervisor: Prof. Dr. Burkhard Rost

Advisor: Juan Miguel Cejuela

Transcription Factors Interactions

- Binds to specific DNA sequences and controls transcription

Transcription Factors Interactions

- Binds to specific DNA sequences and controls transcription
- Two types of interactions
  - Transcription factor *transcribes* gene

NF-κB regulates BCL3 transcription in T Lymphocytes through an intronic enhancer.

Transcription Factors Interactions

- Binds to specific DNA sequences and controls transcription
- Two types of interactions
  - Transcription factor *transcribes* gene

    NF-κB regulates BCL3 transcription in T Lymphocytes through an intronic enhancer.

  - Protein *modifies* or *interacts with* transcription factor

    Androgen receptor interacts with a novel MYST protein, HBO1.

Transcription Factors Interactions

- Binds to specific DNA sequences and controls transcription
- Two types of interactions
  - Transcription factor *transcribes* gene

    NF-κB regulates BCL3 transcription in T Lymphocytes through an intronic enhancer.

  - Protein *modifies* or *interacts with* transcription factor

    Androgen receptor interacts with a novel MYST protein, HBO1.

  - Estimated 45,000+ such interactions

Related Work

- TRANSFAC
  - Manually curated DB
  - Eukaryotic TF and genomic binding sites
  - Commercial version contains reports for 21,000 transcription factors[1]
  - Public version contains reports for 7,000 transcription factors[1]

[1]: https://portal.biobase-international.com/archive/documents/transfac_comparison.pdf
[2]: http://cbrc.kaust.edu.sa/tcof/
[3]: http://itfp.biosino.org/itfp/

Related Work

- TRANSFAC
  - Manually curated DB
  - Eukaryotic TF and genomic binding sites
  - Commercial version contains reports for 21,000 transcription factors[1]
  - Public version contains reports for 7,000 transcription factors[1]
- TcoF – Transcription co-Factor Database
  - 1365 transcription factors[2]
  - Manually curated from BioGrid, MINT and EBI

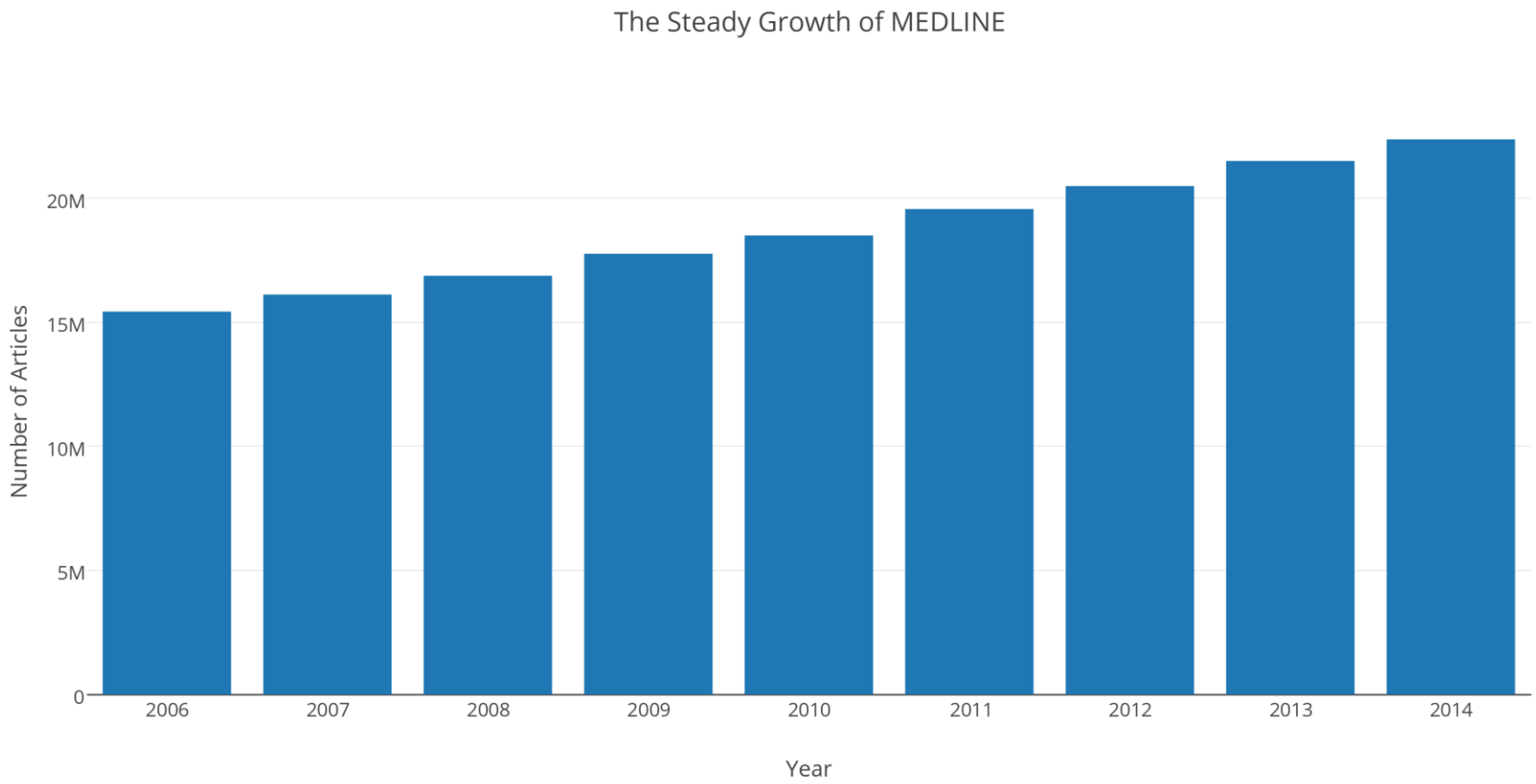
[1]: https://portal.biobase-international.com/archive/documents/transfac_comparison.pdf
[2]: http://cbrc.kaust.edu.sa/tcof/
[3]: http://itfp.biosino.org/itfp/

Related Work

- TRANSFAC
  - Manually curated DB
  - Eukaryotic TF and genomic binding sites
  - Commercial version contains reports for 21,000 transcription factors[1]
  - Public version contains reports for 7,000 transcription factors[1]
- TcoF – Transcription co-Factor Database
  - 1365 transcription factors[2]
  - Manually curated from BioGrid, MINT and EBI
- Integrated Transcription Factor Platform
  - Predicted interactions using sequence data[3]
  - SVMs

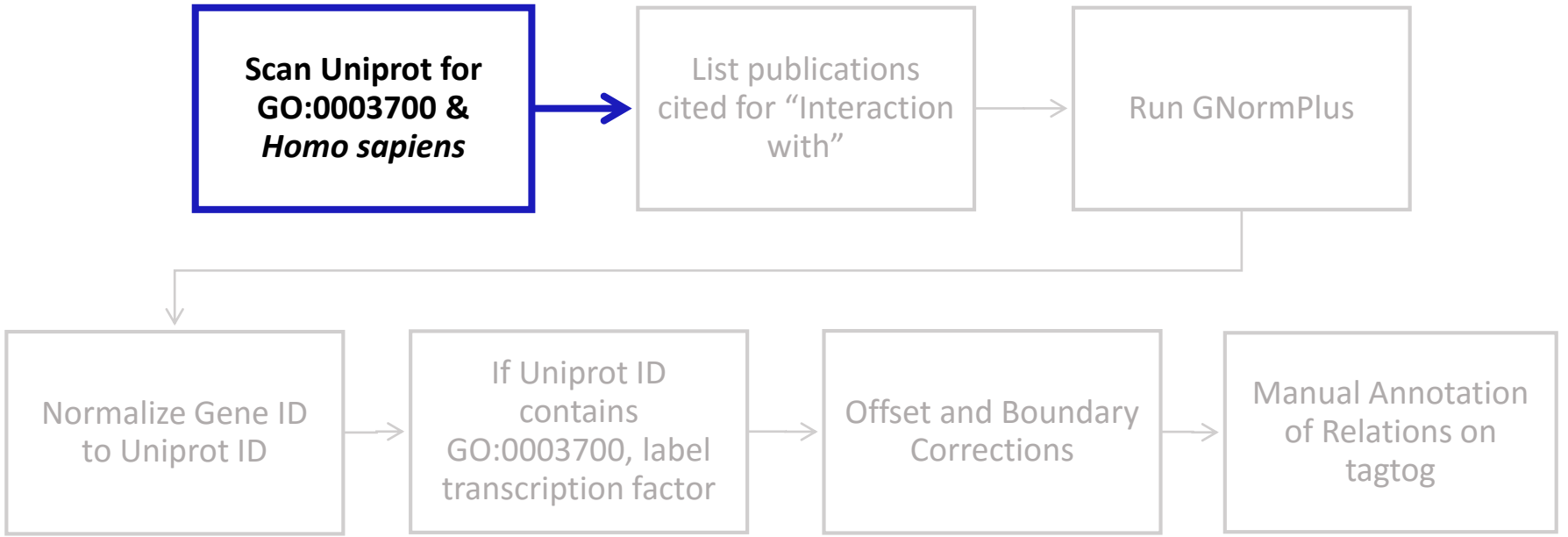
[1]: https://portal.biobase-international.com/archive/documents/transfac_comparison.pdf
[2]: http://cbrc.kaust.edu.sa/tcof/
[3]: http://itfp.biosino.org/itfp/

Problem Motivation

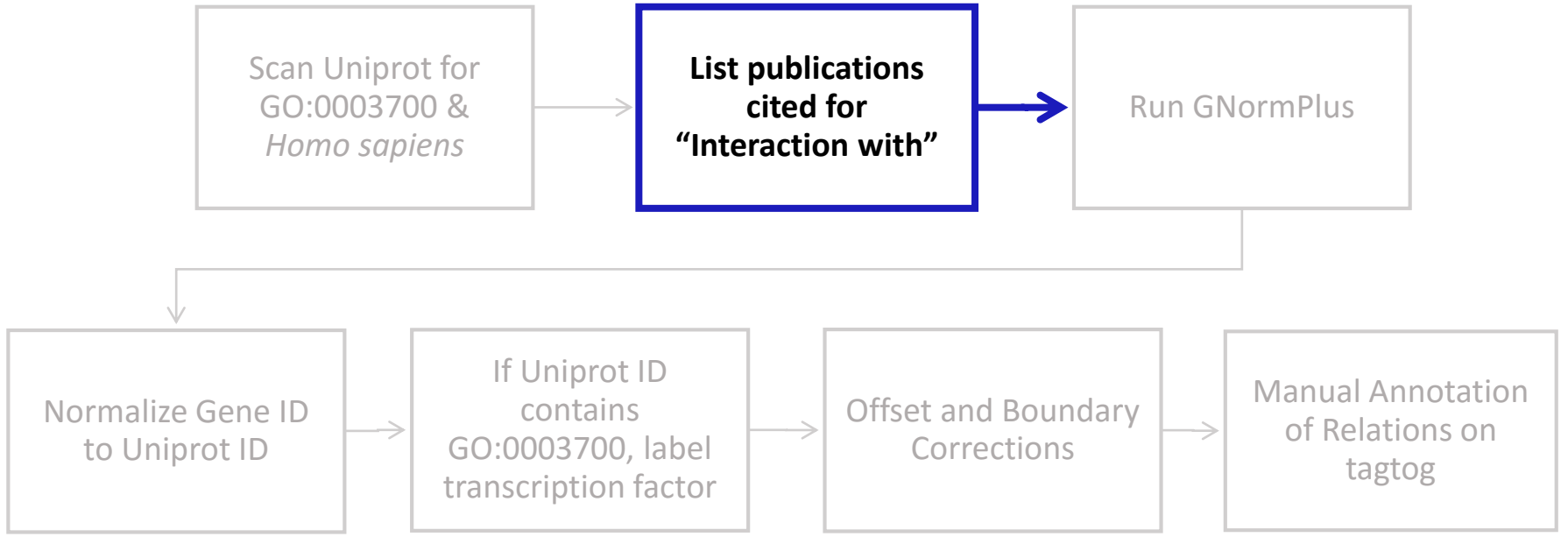


MEDLINE currently has more than 24M articles and adds over 1M articles per year

Corpus

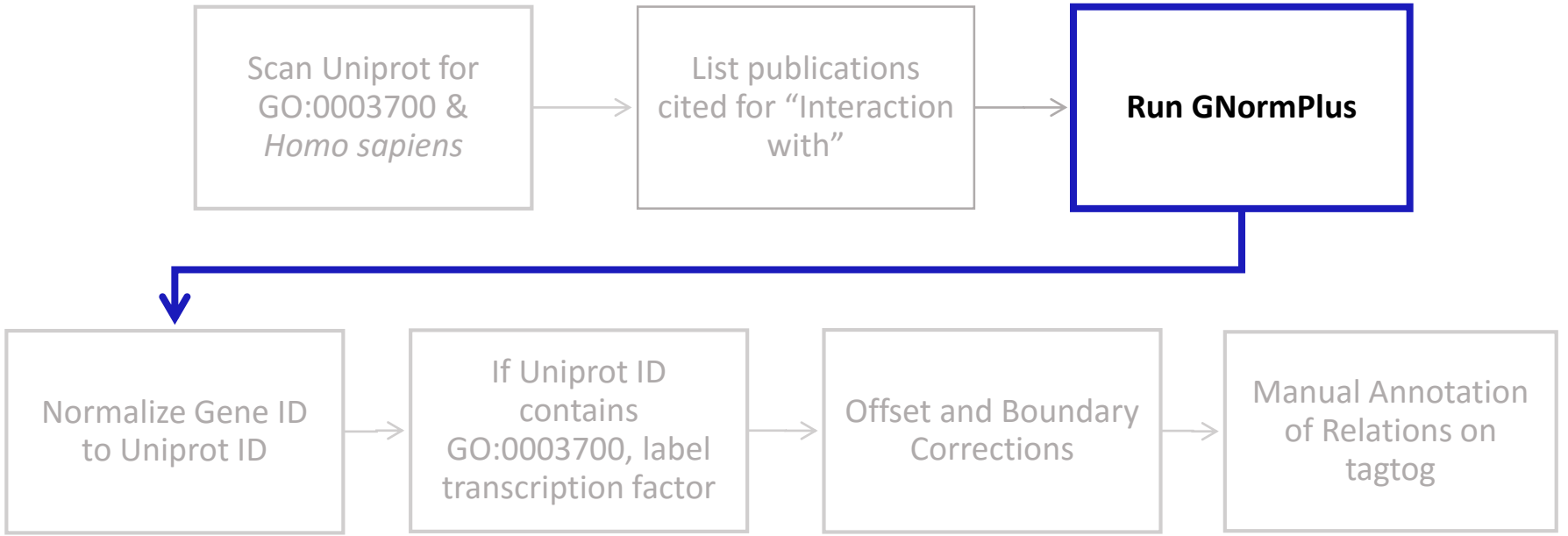


Filter Swissprot for **transcription-factor activity, sequence specific DNA-binding** (GO Term *GO:0003700*) and *Homo sapiens*

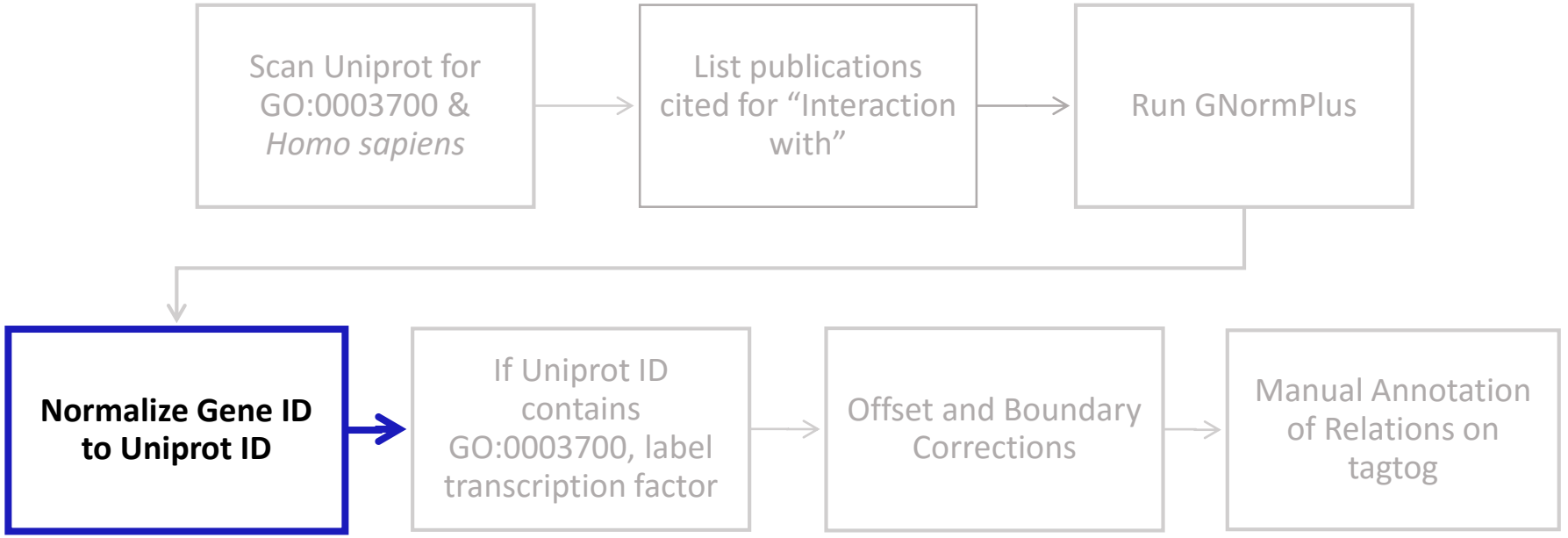


For each protein, obtain list of publications cited for "INTERACTION WITH"

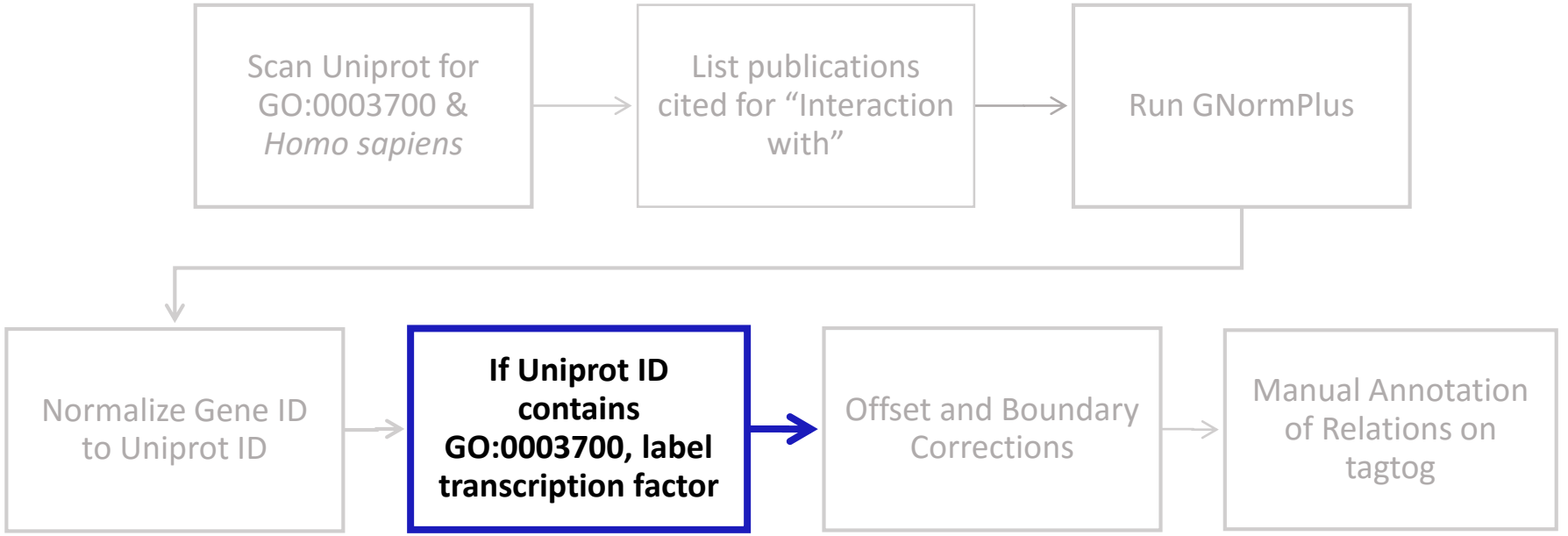
"Cloning and characterization of androgen receptor coactivator, ARA55, in human prostate."
Fujimoto N., Yeh S., Kang H.-Y., Inui S., Chang H.-C., Mizokami A., Chang C.
J. Biol. Chem. 274:8316-8321(1999) [PubMed] [Europe PMC] [Abstract]
Cited for: INTERACTION WITH TGFB1I1.



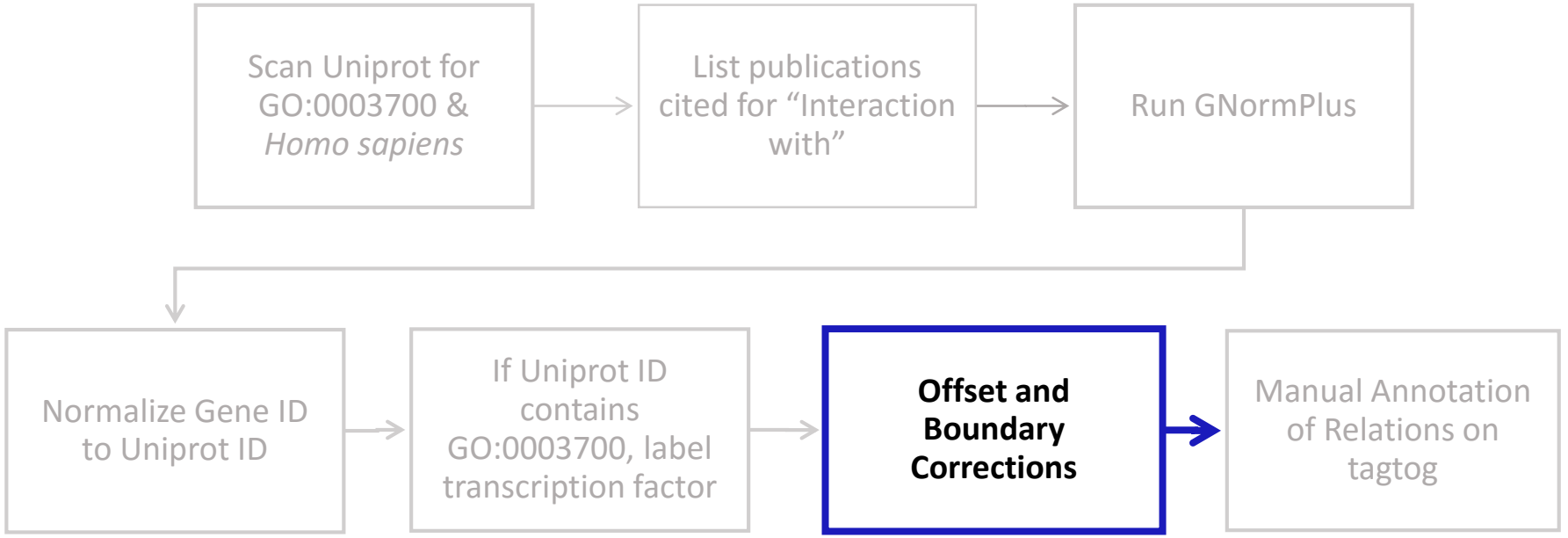
Run **GNormPlus**, a gene tagger, on each of these abstracts – giving **Gene or Gene Product** (GGP)



Entrez Gene IDs **normalized to Uniprot IDs** using **priority selection** (first Swissprot, then TrEMBL)

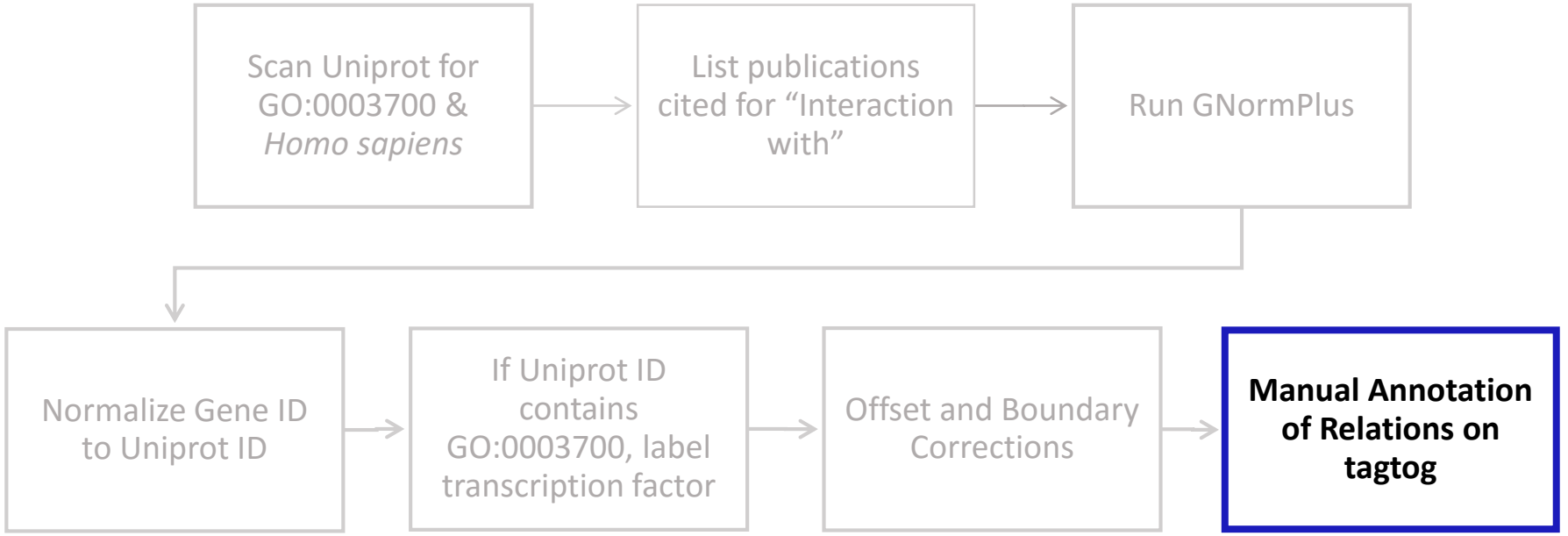


All GGPs **cross-referenced** with GO Term *GO:0003700* and its descendants. If Uniprot ID contains annotation for transcription factor activity, labeled as **transcription factor**



Correct entity boundaries and offsets, and add abbreviations

Androgen Receptor (AR) is involved in → Androgen Receptor (AR) is involved in



Annotation of relations on **tagtog** manually

relna Annotation

TRIM24 mediates ligand-dependent activation of androgen receptor and is repressed by a bromodomain-containing protein, BRD7, in prostate cancer cells.

Abstract

The androgen receptor (AR) is a ligand-dependent transcription factor that belongs to the family of nuclear receptors, and its activity is regulated by numerous AR coregulators. AR plays an important role in prostate development and cancer. In this study, we found that TRIM24/transcriptional intermediary factor 1alpha (TIF1alpha), which is known as a ligand-dependent nuclear receptor co-regulator, interacts with AR and enhances transcriptional activity of AR by dihydrotestosterone in prostate cancer cells. We showed that TRIM24 functionally interacts with TIP60, which acts as a coactivator of AR and synergizes with TIP60 in the transactivation of AR. We also showed that TRIM24 binds to bromodomain containing 7 (BRD7), which can negatively regulate cell proliferation and growth. A luciferase assay indicated that BRD7 represses the AR transactivation activity upregulated by TRIM24. These findings indicate that TRIM24 regulates AR-mediated transcription in collaboration with TIP60 and BRD7.

Relation Extraction

Relation Extraction

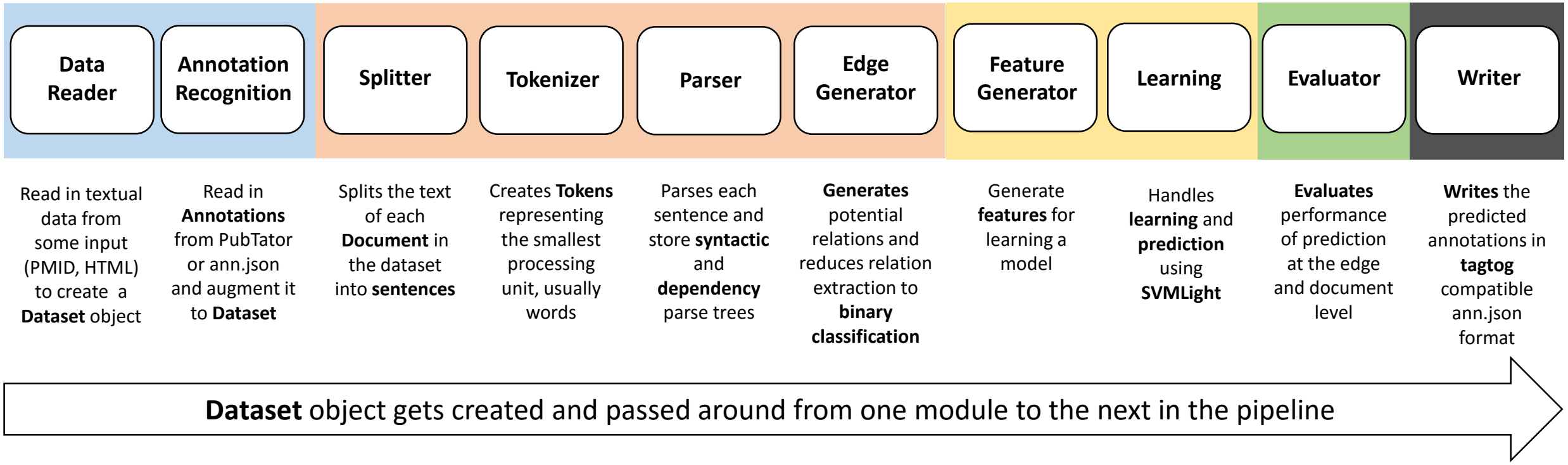
- Binary Classification

effect on AR transactivation. Overexpression of RanBP10 enhanced transcriptional activity of glucocorticoid receptor, but not estrogen receptor alpha. RanBP10 was highly expressed in

Entity 1	Entity 2	Feature 1	…	Feature n	Class
RanBP10	Glucocorticoid Receptor	1	…	0	True
RanBP10	Estrogen receptor alpha	0	…	1	False

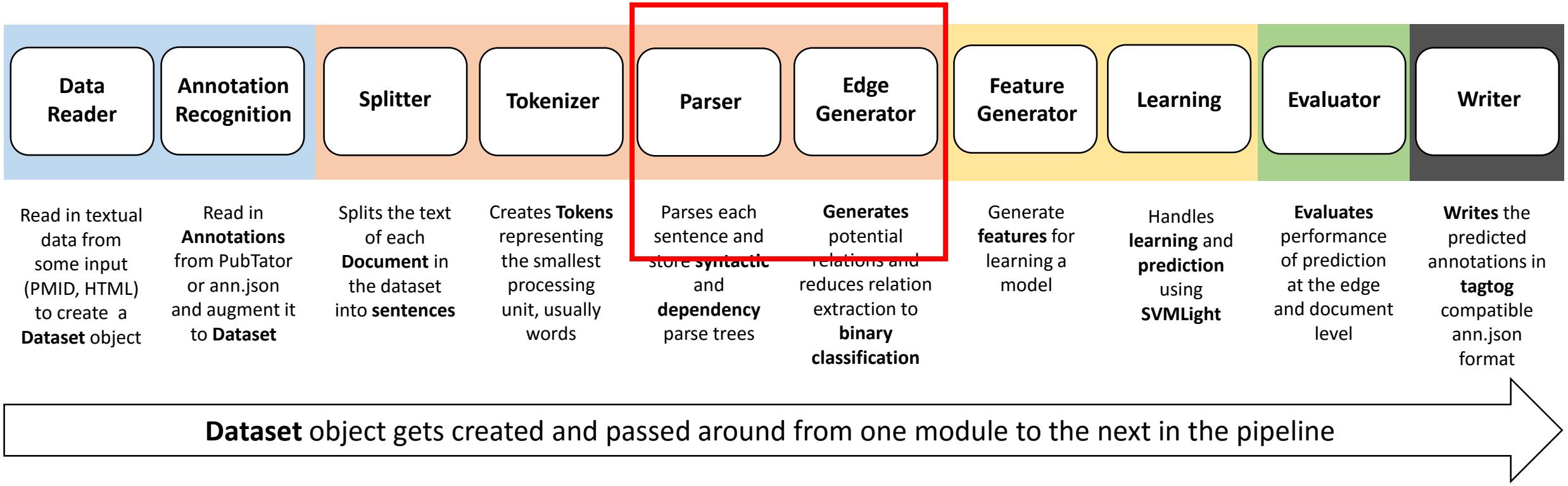
Method Development

- Pipeline (based on *nalaf*)

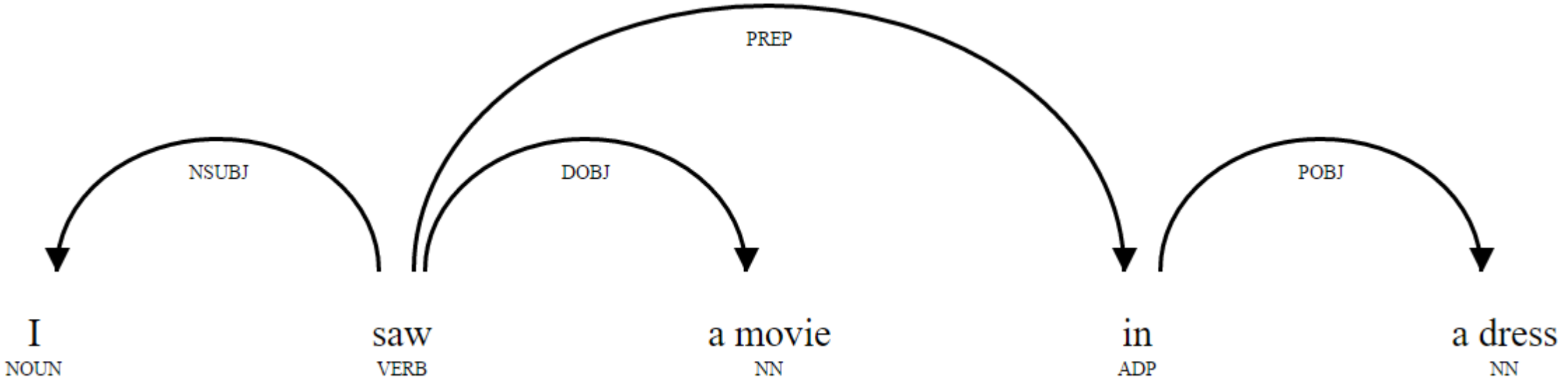


Method Development

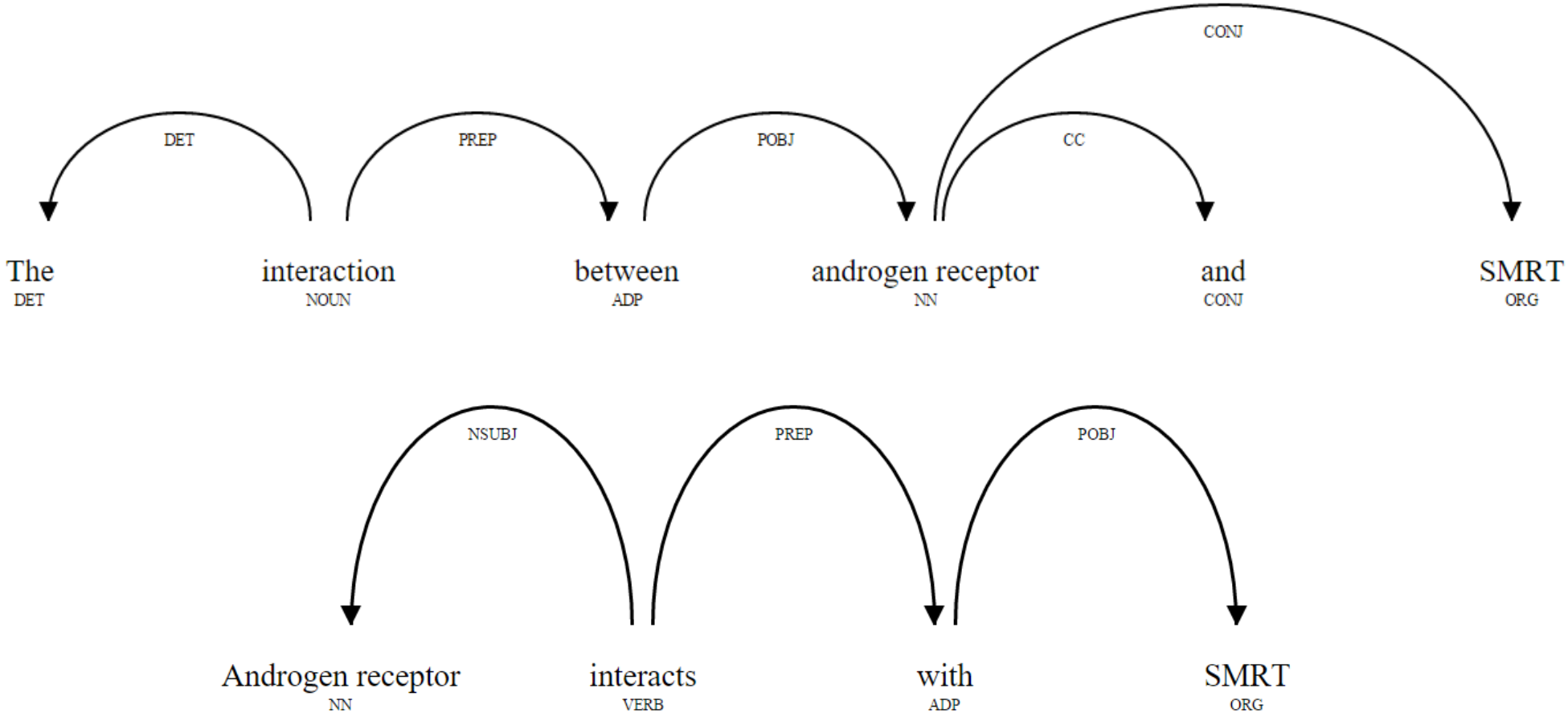
- Pipeline (based on *nalaf*)



Dependency Parsing

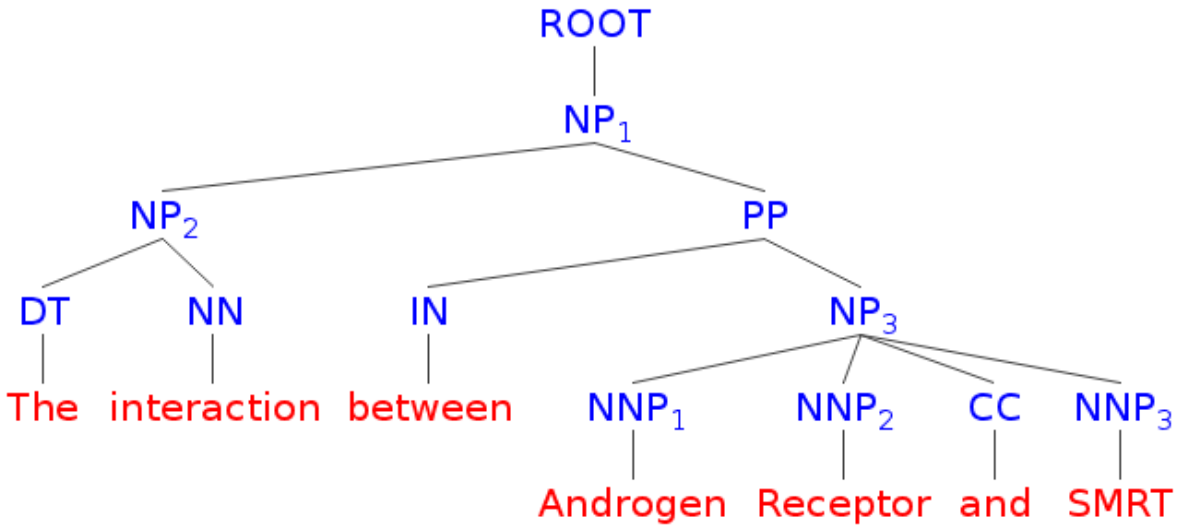


Dependency Parsing



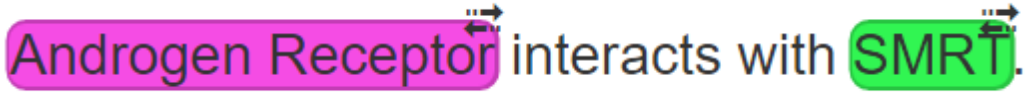
Constituency Parsing

```
(ROOT
  (NP
    (NP (DT The) (NN interaction))
    (PP (IN between)
      (NP (NNP Androgen) (NNP Receptor)
        (CC and)
        (NNP SMRT)))))
```



Feature Generation

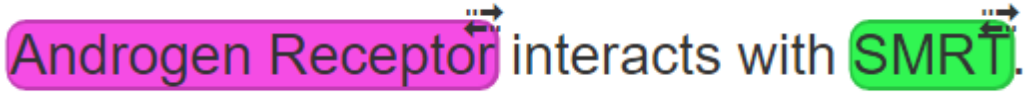
- Sentence Features



Feature example:
“Named Entity Count”: 2

Feature Generation

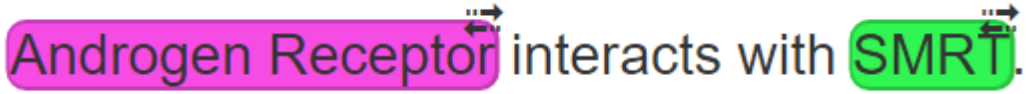
- Sentence Features
- Token Features



Feature example:
“interacts_stem” : “interact”
“interacts_pos” : “VBZ”
“interacts_lem” : “interact”

Feature Generation

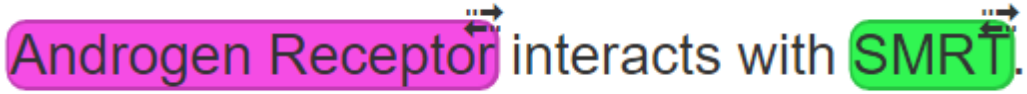
- Sentence Features
- Token Features
- N-gram Features



Feature example:
“androgen receptor”,
“receptor interacts”,
“interacts with”, “with SMRT”

Feature Generation

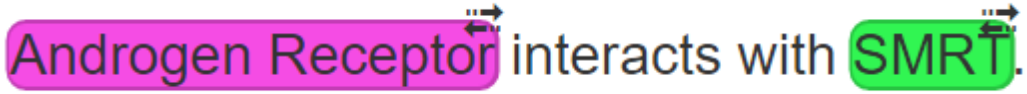
- Sentence Features
- Token Features
- N-gram Features
- Linear Context and Distance between Entities



Feature example:
“linear_distance” : 2

Feature Generation

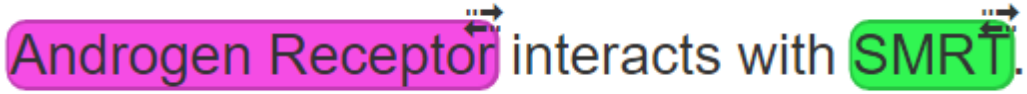
- Sentence Features
- Token Features
- N-gram Features
- Linear Context and Distance between Entities
- Dependency Features
  - Shortest path between the two entities
  - Path constituents
  - Root word
  - Path to root word



Feature example:
receptor -> interacts -> with -> SMRT

Feature Generation

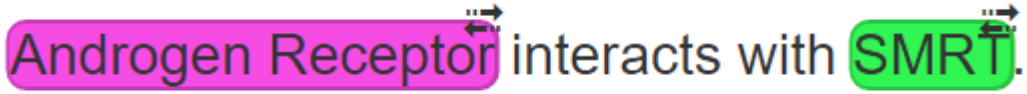
- Sentence Features
- Token Features
- N-gram Features
- Linear Context and Distance between Entities
- Dependency Features
  - Shortest path between the two entities
  - Path constituents
  - Root word
  - Path to root word



Feature example:
[“interacts”, “with”]

Feature Generation

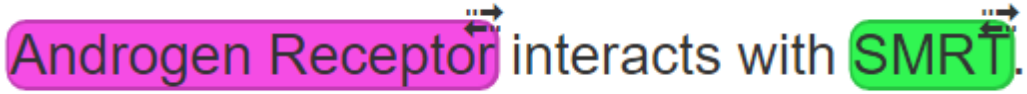
- Sentence Features
- Token Features
- N-gram Features
- Linear Context and Distance between Entities
- Dependency Features
  - Shortest path between the two entities
  - Path constituents
  - Root word
  - Path to root word



Feature example:
“root_word”: “interacts”

Feature Generation

- Sentence Features
- Token Features
- N-gram Features
- Linear Context and Distance between Entities
- Dependency Features
  - Shortest path between the two entities
  - Path constituents
  - Root word
  - Path to root word



Feature example:
receptor -> interacts
SMRT -> with -> interacts

Evaluation

Evaluation Modes

- **Non-Unique**
  - SVM (or Edge) Performance
  - Repeated relations
  - Only if offsets AND text of entities match
- **Unique**
  - Document Performance
  - No repetitions
  - If texts of entities match

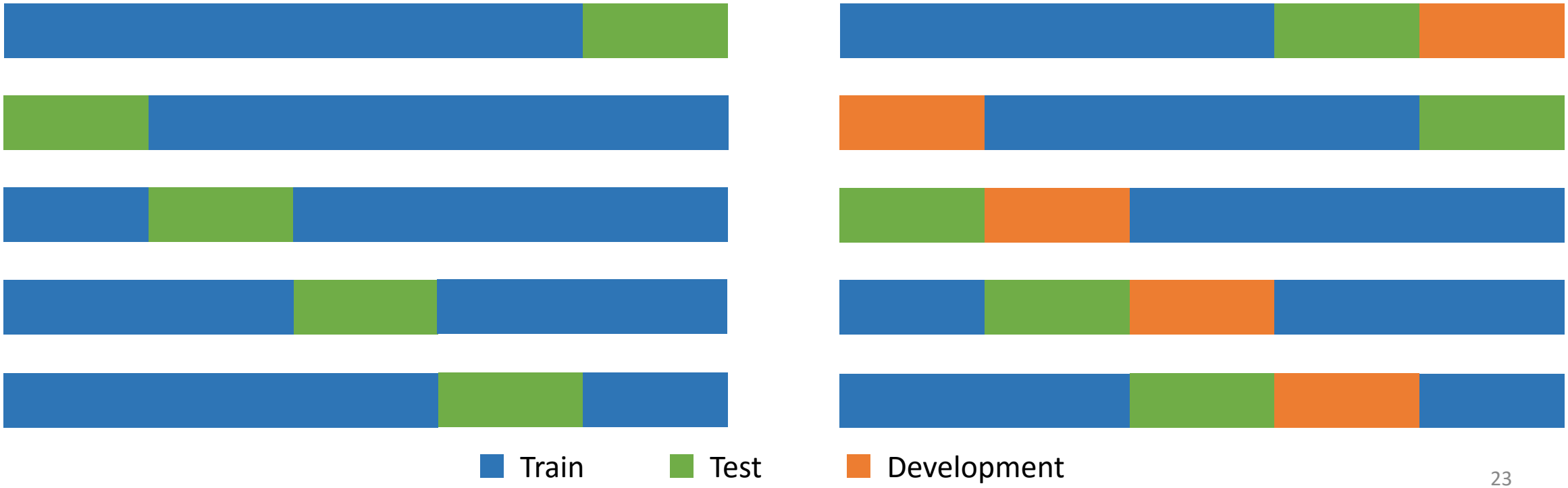
HepG2 hepatocarcinoma cells. IL-6 induces rapid nuclear translocation of Tyr-phosphorylated STAT3 that forms a nuclear complex with CDK9 in nondenaturing co-immunoprecipitation and confocal colocalization assays. To further understand this interaction, we found that CDK9-STAT3 binding is mediated via both STAT NH2-terminal modulatory and COOH-

Non-Unique:
1. CDK9, STAT3 (2nd line)
2. CDK9, STAT3 (4th line)

Unique:
1. CDK9, STAT3

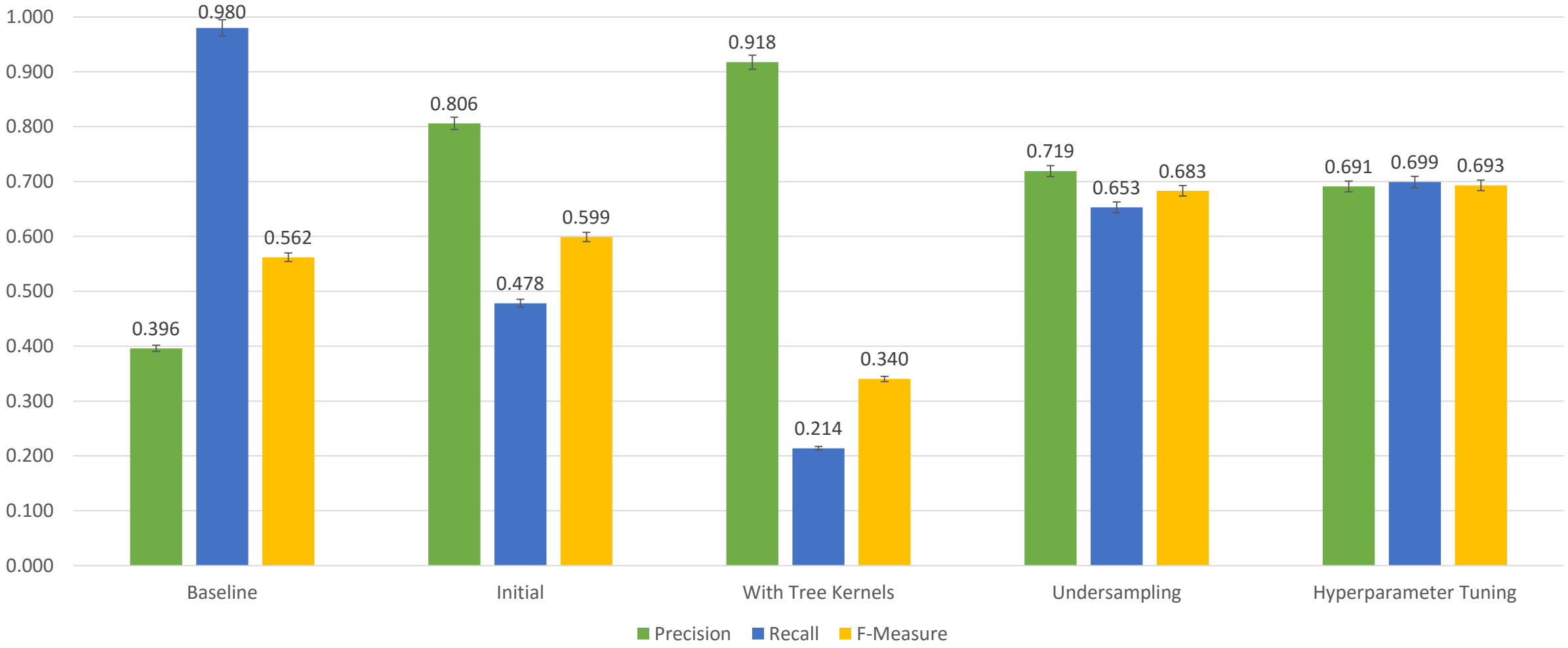
Training and Cross Validation

- Two methods
  - 80:20
  - 60:20:20

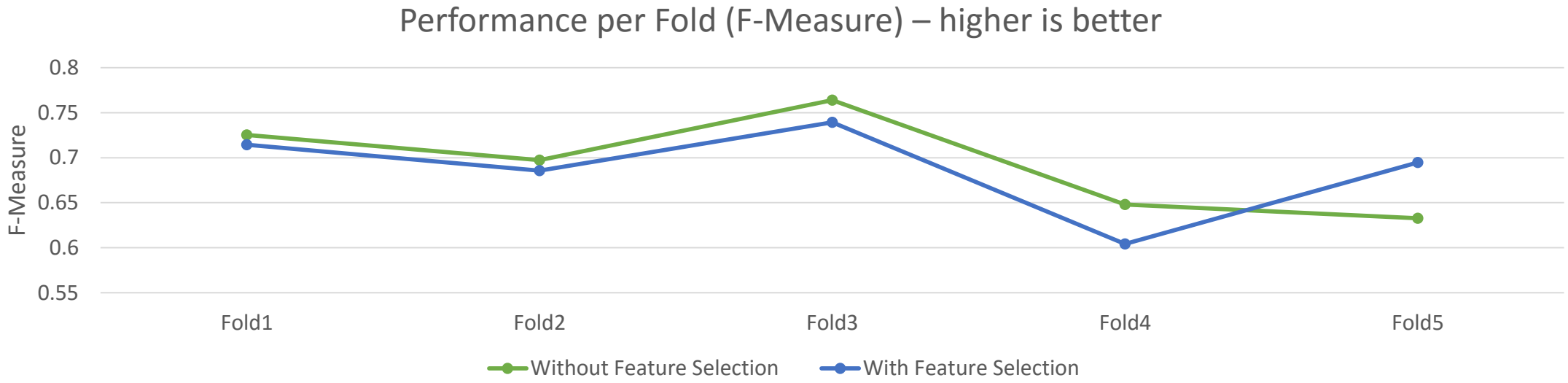


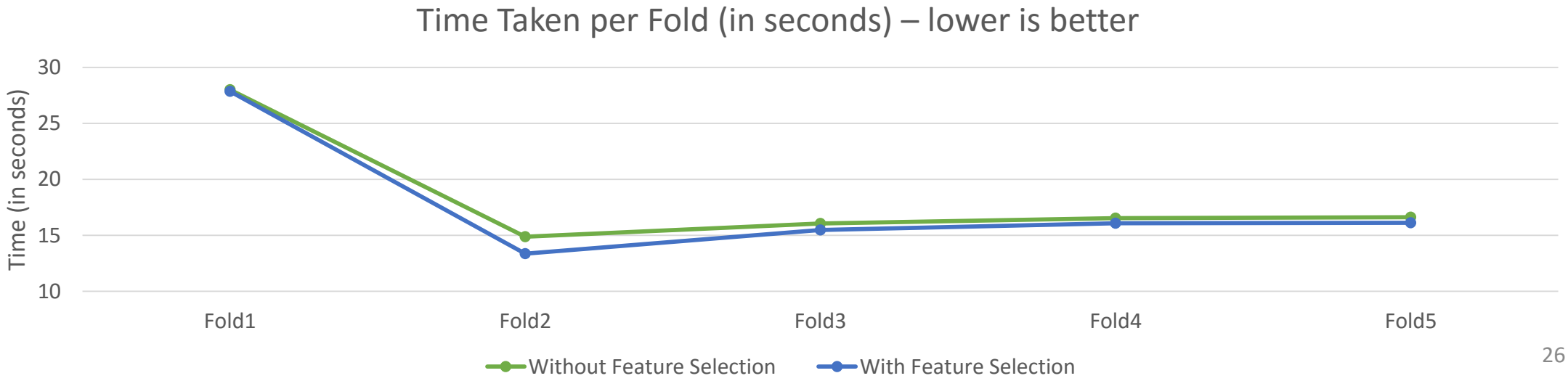
Results

Results – Comparison of Methods



Results

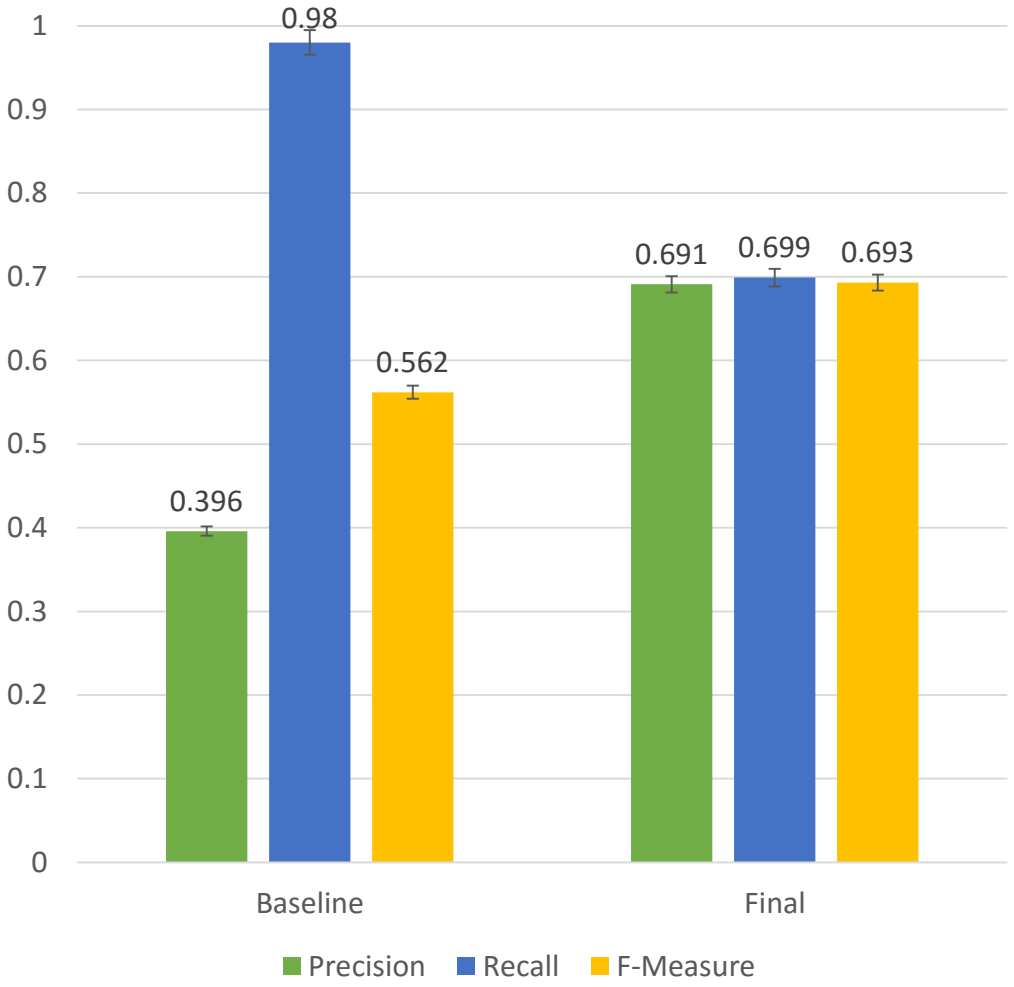

Performance per Fold (F-Measure) – higher is better
F-Measure
0.8
0.75
0.7
0.65
0.6
0.55
Fold1
Fold2
Fold3
Fold4
Fold5
Without Feature Selection
With Feature Selection


Time Taken per Fold (in seconds) – lower is better
Time (in seconds)
30
25
20
15
10
Fold1
Fold2
Fold3
Fold4
Fold5
Without Feature Selection
With Feature Selection

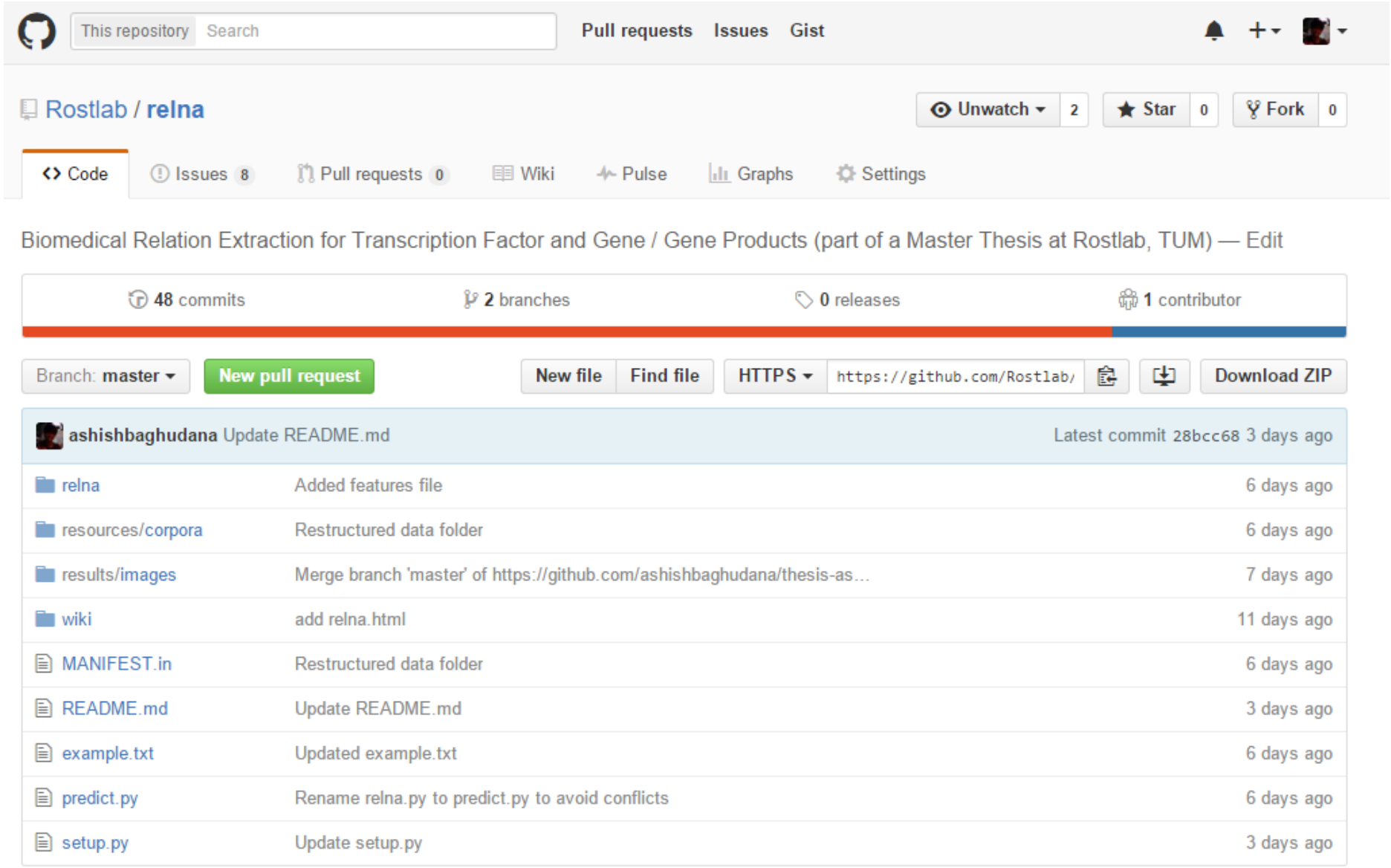
Demo

Conclusion

Conclusion



Performance	Percentage
Precision (P)	69.1 ± 0.61%
Recall (R)	69.9 ± 0.56%
F-Measure (F)	69.3 ± 0.59%


This repository Search
Pull requests
Issues
Gist
Rostlab / relna
Unwatch 2
Star 0
Fork 0
Code
Issues 8
Pull requests 0
Wiki
Pulse
Graphs
Settings
Biomedical Relation Extraction for Transcription Factor and Gene / Gene Products (part of a Master Thesis at Rostlab, TUM) — Edit
48 commits
2 branches
0 releases
1 contributor
Branch: master
New pull request
New file
Find file
HTTPS
https://github.com/Rostlab/
Download ZIP
ashishbaghudana Update README.md
Latest commit 28bcc68 3 days ago
relna Added features file 6 days ago
resources/corpora Restructured data folder 6 days ago
results/images Merge branch 'master' of https://github.com/ashishbaghudana/thesis-as... 7 days ago
wiki add relna.html 11 days ago
MANIFEST.in Restructured data folder 6 days ago
README.md Update README.md 3 days ago
example.txt Updated example.txt 6 days ago
predict.py Rename relna.py to predict.py to avoid conflicts 6 days ago
setup.py Update setup.py 3 days ago

Conclusion

- Development of new corpus with semi-automatic annotation
  - Published at PubAnnotation: http://pubannotation.org/projects/relna

Conclusion

- Development of new corpus with semi-automatic annotation
  - Published at PubAnnotation: http://pubannotation.org/projects/relna
- Development of new method for extracting relations of transcription factors
  - Registered with Elixir Tools: https://bio.tools/tool/RostLab/relna/0.1.0
  - Available on GitHub: https://github.com/Rostlab/relna

Conclusion

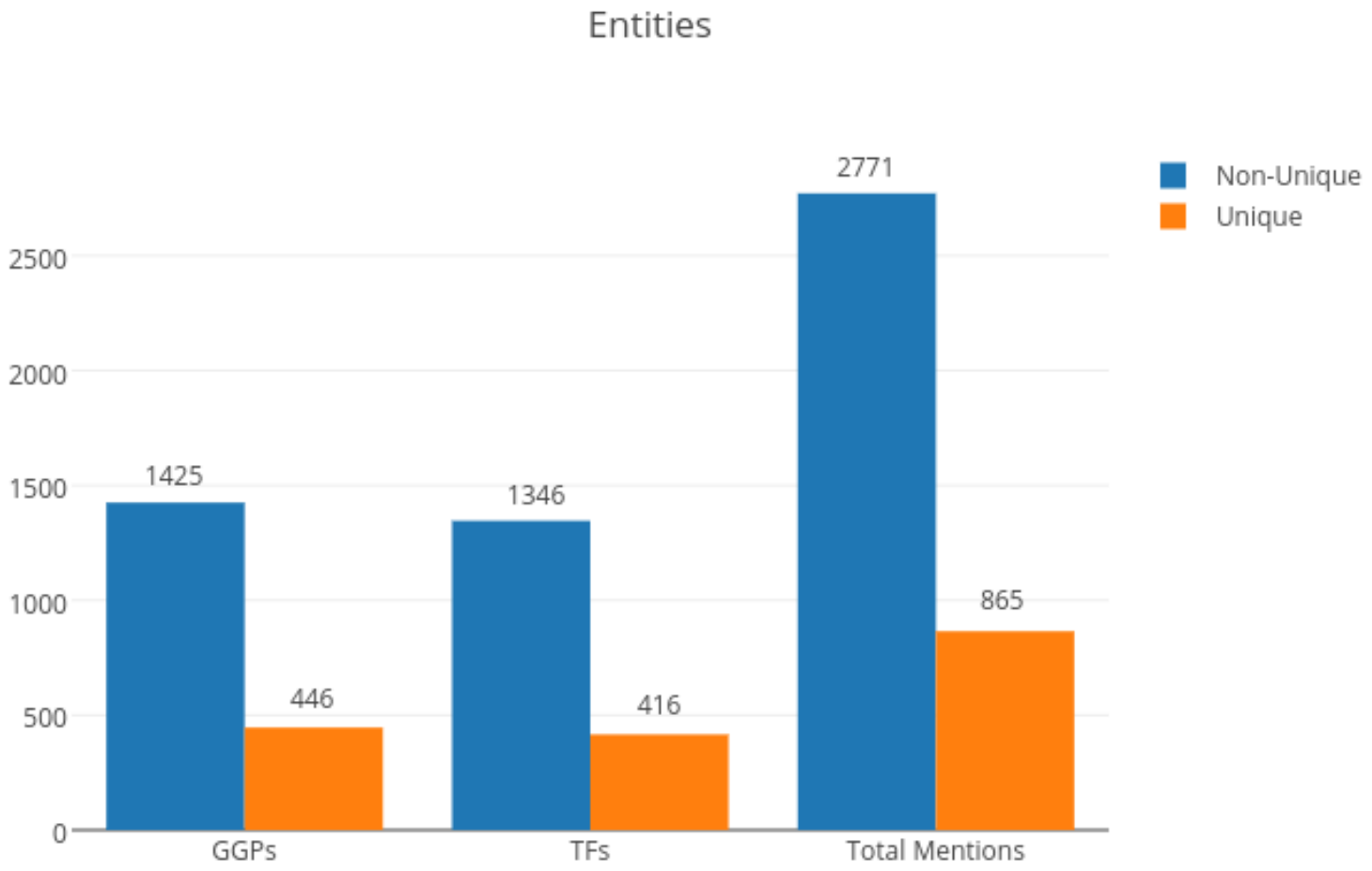
- Development of new corpus with semi-automatic annotation
  - Published at PubAnnotation: http://pubannotation.org/projects/relna
- Development of new method for extracting relations of transcription factors
  - Registered with Elixir Tools: https://bio.tools/tool/RostLab/relna/0.1.0
  - Available on GitHub: https://github.com/Rostlab/relna
- Integration into *nalaf* and building a generalized relation extraction tool

Future Work

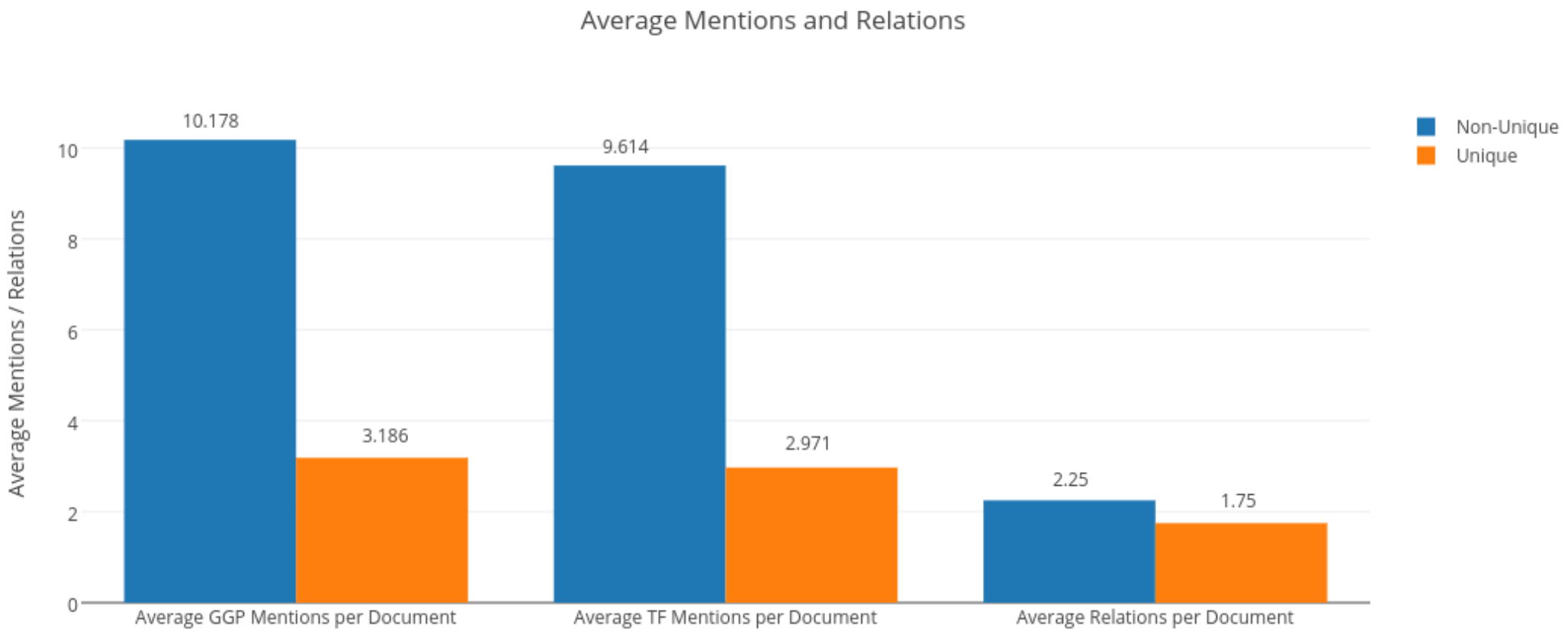
- Coreference resolution techniques
- Generalizing method for spanning multiple sentences
- Further testing with neural networks

Thank you

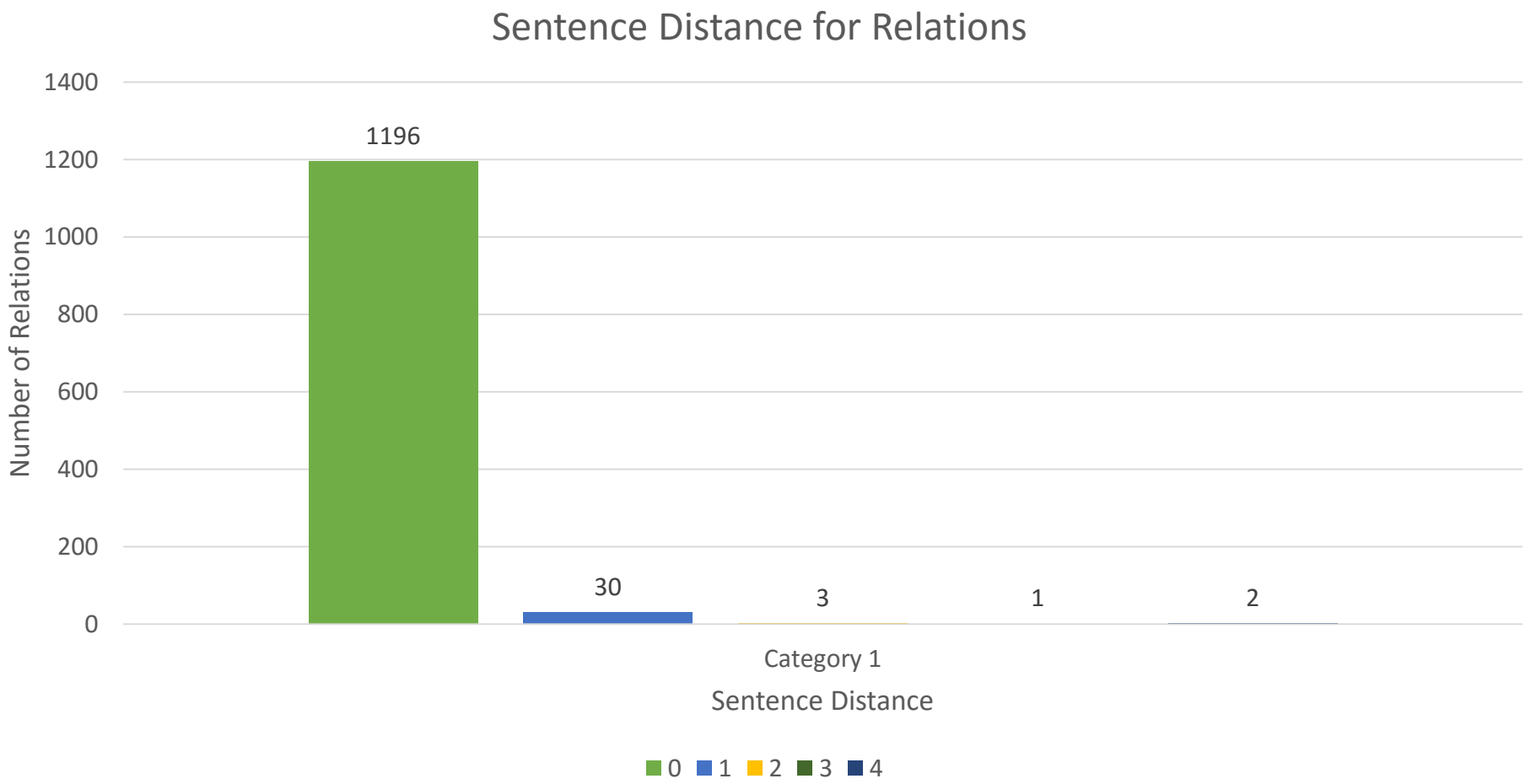
Corpus Statistics



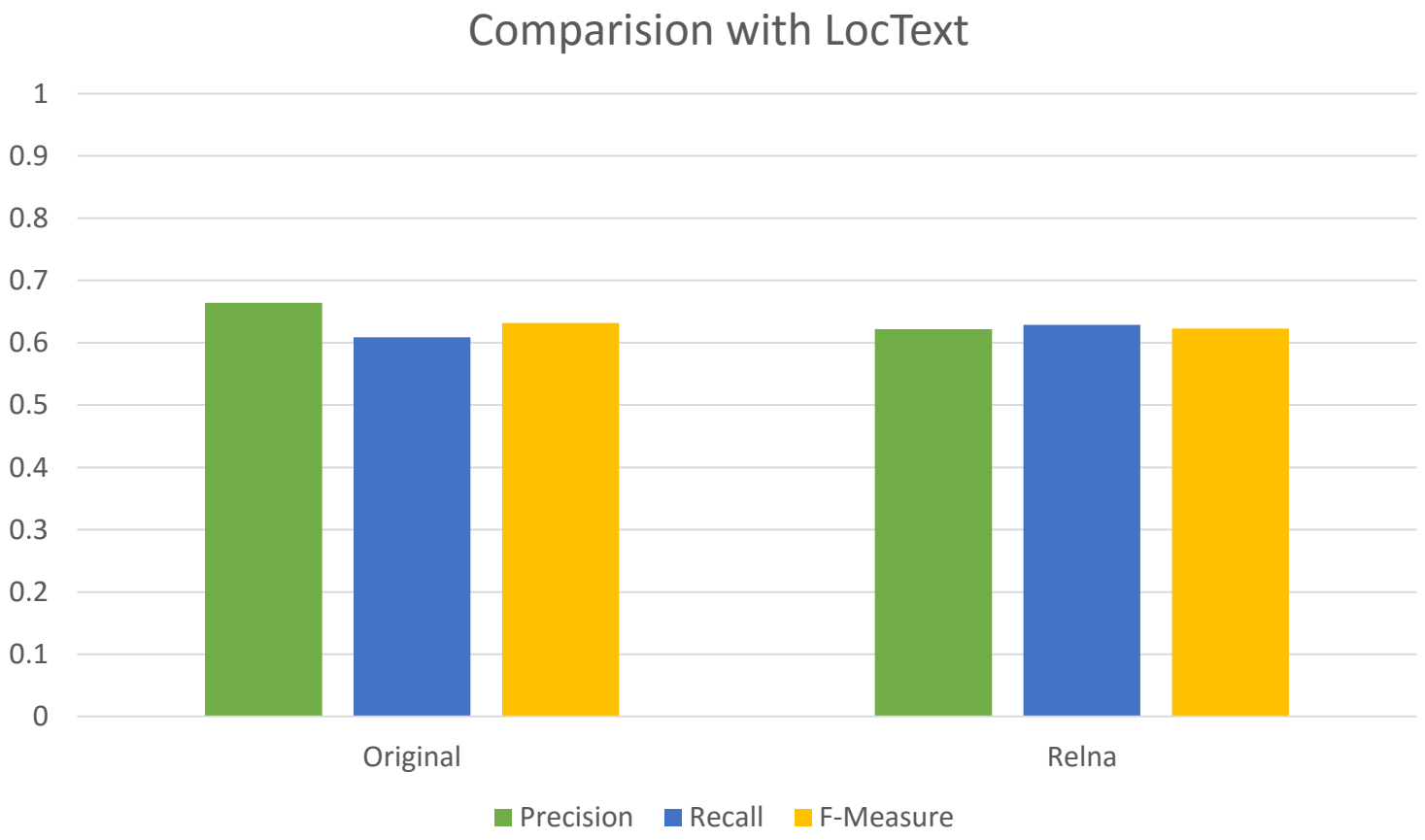
Corpus Statistics



Corpus Statistics



Performance on LocText



Parsing

- Dependency Parsing
  - Identify the relations between words
  - $O(n^3)$ algorithm, with n as the number of words in the sentence
- Constituency Parsing
  - Identify phrases (noun chunks, verb chunks etc.) and their relative structure and hierarchy in the sentence
  - $O(n^5)$ algorithm, with n as the number of words in the sentence

Exhaustive List of Features

- Sentence Features
  - BOW, Stem
  - #entities, #BOW count
- Token Features
  - Token Text, Masked Text
  - Stem, POS
  - Capitalization, Digits, Hyphens and other Punctuations
  - Char bigrams and trigrams
- Dependency Feature for Shortest Paths
  - Path direction (eg. FFRFR)
  - Dependency types in path
  - Path length
  - Intermediate Tokens
  - Path Constituents (eg. “interact”, “bind” etc.)
  - Root word of the sentence
- Linear Context

Other Features

- Linear distance between entities
- Presence of specific words in the sentence
- Prior tokens
- Intermediate tokens
- Post tokens
- N-gram features
  - Bigram
  - Trigram
- Relative Entity Order
- Conjoint Entity Text

N-gram Features

- The cow jumped over the moon
  - “the”, “cow”, “jumped”, “over”, “moon”
  - “the cow”, “cow jumped”, “jumped over”, “over the”, “the moon”
  - “the cow jumped”, “cow jumped over”, “jumped over the”, “over the moon”
  - …