Unsupervised Lexicon and Punctuation Discovery

Anton Kolonin

akolonin@aigents.com

Facebook: akolonin

Telegram: akolonin

Novosibirsk State University
*THE REAL SCIENCE

SingularityNET

AGI IN RUSSIAN

https://agirussia.org

AIGENTS

https://aigents.com

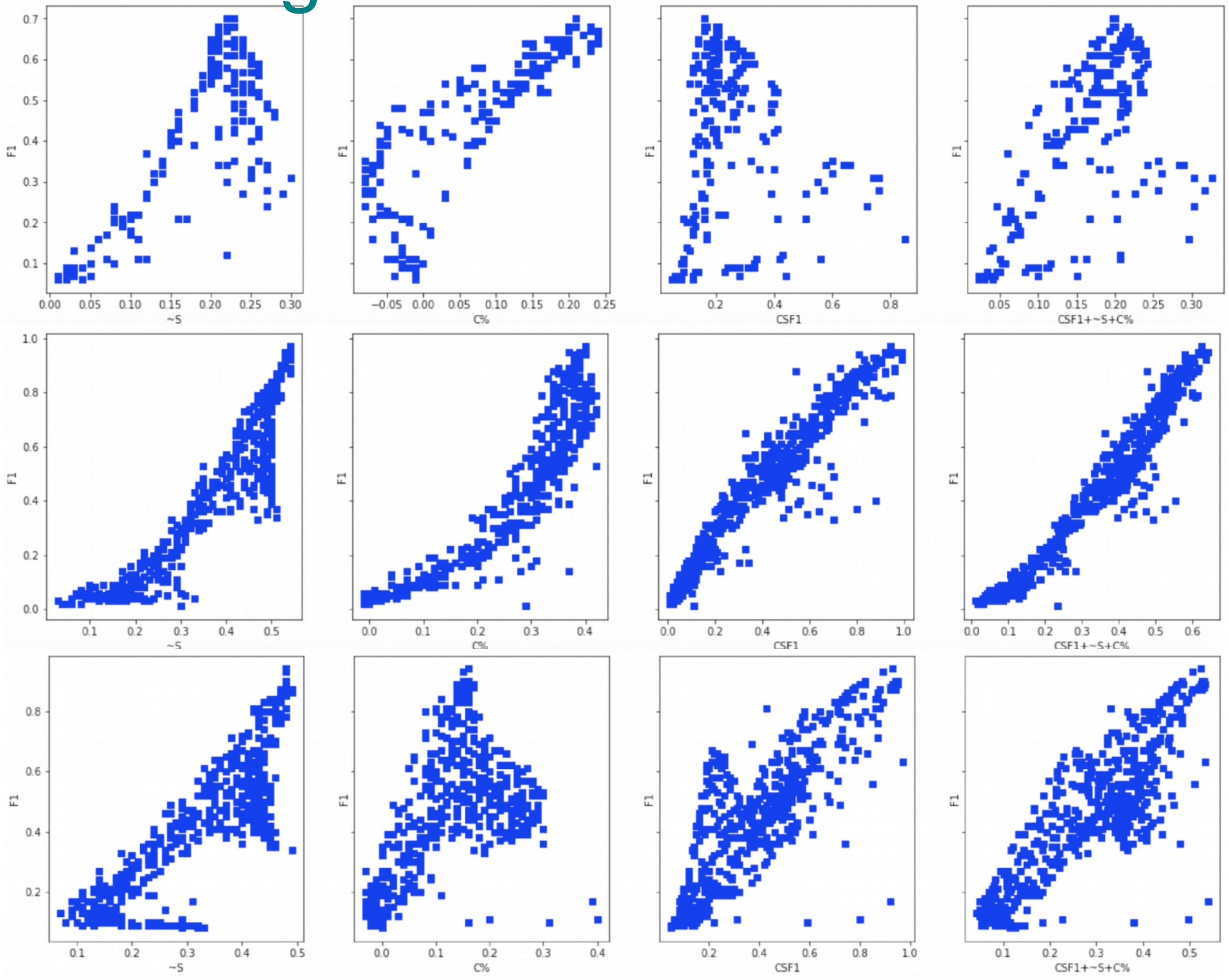
SingularityDAO

Something about Human Intuition?

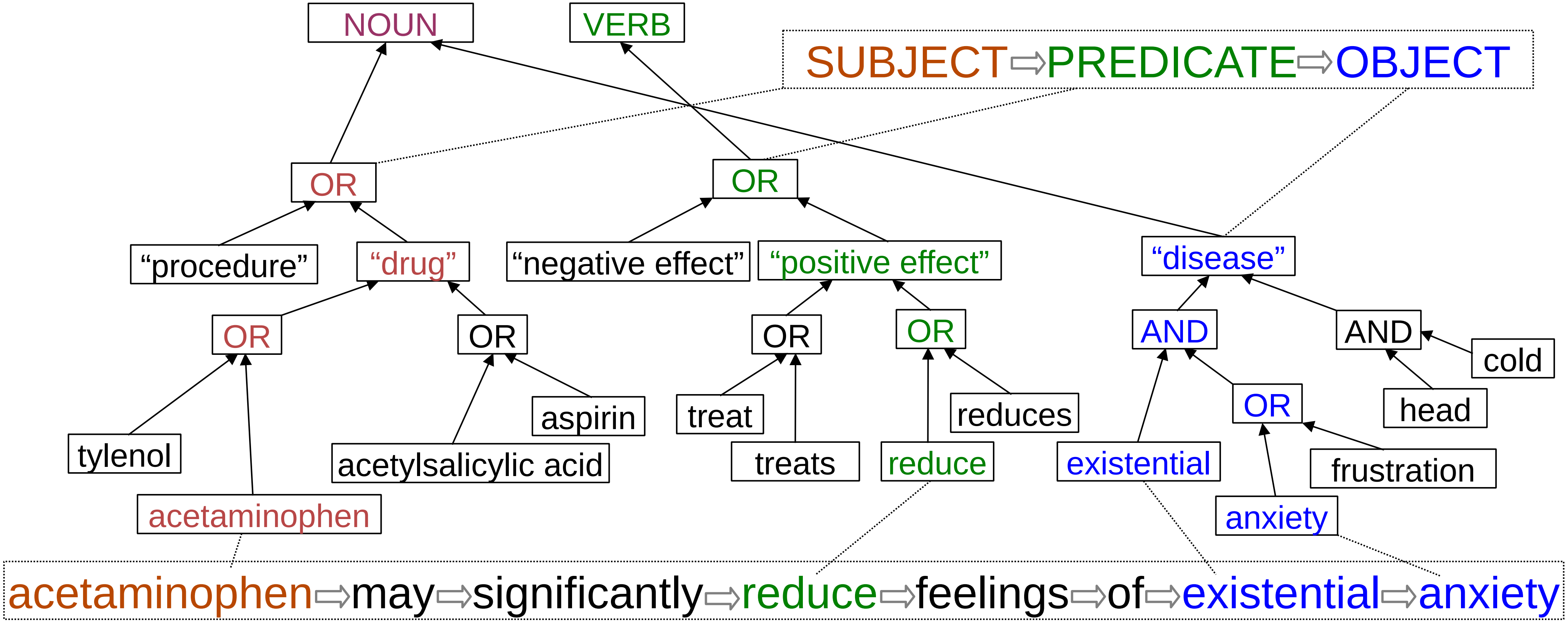
Language 1

Language 2

Language 3



Goal: Discovering NLP patterns (words, punctuation, phrases) for unsupervised language learning (Aigents® “Deep Patterns”)



https://ieeexplore.ieee.org/document/7361868
https://github.com/aigents/aigents-java

https://www.springerprofessional.de/unsupervised-language-learning-in-opencog/15995030
https://www.springerprofessional.de/en/programmatic-link-grammar-induction-for-unsupervised-language-le/17020348
https://github.com/singnet/language-learning/

Motivation

Absence of explicit start/stop tags in continuous streams of spaces in experiential (reinforcement/self-reinforcement) learning with delayed/sparse feedback

https://www.youtube.com/watch?v=2LPLhJKh95g
https://www.springerprofessional.de/neuro-symbolic-architecture-for-experiential-learning-in-discret/20008336
https://github.com/aigents/aigents-java/tree/master/src/main/java/net/webstructor/agi

Complex, cumbersome, unreliable and expensive language-specific tokenization process for unsupervised language learning in NLP

Low quality of unsupervised parsing and tokenization learning based on mutual information and conditional probabilities

https://www.springerprofessional.de/unsupervised-language-learning-in-opencog/15995030
https://www.springerprofessional.de/en/programmatic-link-grammar-induction-for-unsupervised-language-le/17020348
https://github.com/singnet/language-learning/
https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=6983&context=etd

Tokenization or Text Segmentation as Language Modeling

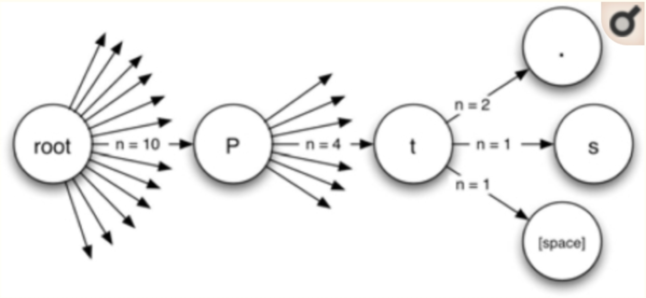


Figure 1

Trie data structure. The probability of observing an ‘s’ given the preceding string “Pt” is ¼, or 25%. The freedom following “pt” is 3.

Metrics/Indicators:

Mutual Information[1]

Conditional Probability[1,2]

Transition Freedom[2,3]

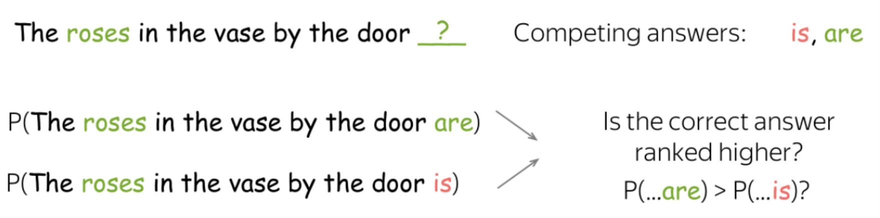
[1] https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=6983&context=etd

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/

[3] **Karl Friston.The free-energy principle: a unified brain theory?** https://www.nature.com/articles/nrn2787

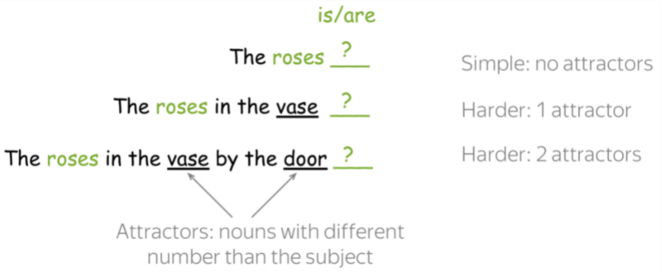
Contrastive Evaluation: Test Specific Phenomena

To test if your LM knows something very specific, you can use contrastive examples. These are the examples where you have several versions of the same text which differ only in the aspect you care about: one correct and at least one incorrect. A model has to assign higher scores (probabilities) to the correct version.



A very popular phenomenon to look at is subject-verb agreement, initially proposed in the Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies paper. In this task, contrastive examples consist of two sentences: one where the verb agrees in number with the subject, and another with the same verb, but incorrect inflection.

Examples can be of different complexity depending on the number of **attractors**: other nouns in a sentence that have different grammatical number and can "distract" a model from the subject.



https://lena-voita.github.io/nlp_course/language_modeling.html

Claims

Transition Freedom (TF) appears to be superior (over **Mutual Information** and **Conditional Probability**) for unsupervised text segmentation (tokenization).

English and Russian require one specific way (variance) of handling the TF while Chinese requires a bit different specific way (derivative-based “peak values”) for the same purpose.

Tokenization quality for Russian and English may be as high as F1=0.96-1.0, depending on training and testing corpora while for Chinese the minimum is F1=0.71-0.92, depending on the assessment assumptions.

Larger training corpora does not necessarily effect in better tokenization quality, while compacting the models eliminating statistically weak evidence typically improve the quality.

TF-based tokenization appear quality same or better than lexicon-based one for Russian and English while for Chinese appears the opposite (as it could be anticipated).

Doing Russian and English tokenization with removed spaces makes the situation similar to Chinese with reasonable quality on lexicon-based tokenization but much worse results on TF-based one.

https://arxiv.org/abs/2205.11443

https://github.com/aigents/pygents

Corpora and Methodology

Train corpora

Chinese
CLUE News 2016 Validation – 270M
CLUE News 2016 Train – 8,500M

English
Brown – 6M
Gutenberg Children – 29M
Gutenberg Adult – 140M
Social Media – 68M
All above – combined

Russian
RusAge Test – 141M
RusAge Previews – 825M

Test corpus

Parallel Chinese/English/Russian – 100 multi-sentence statements on finance

Metrics/Indicators:

Ngram (Character)
- Probability or Conditional Transition Probability (p-/p+)
  - Deviation (dvp-/dvp+)
  - Derivative (dp-/dp+)
- Transition Freedom (f-/f+)
  - Deviation (dvf-/dvf+)
  - Derivative (df-/df+)

Hyper-parameters:

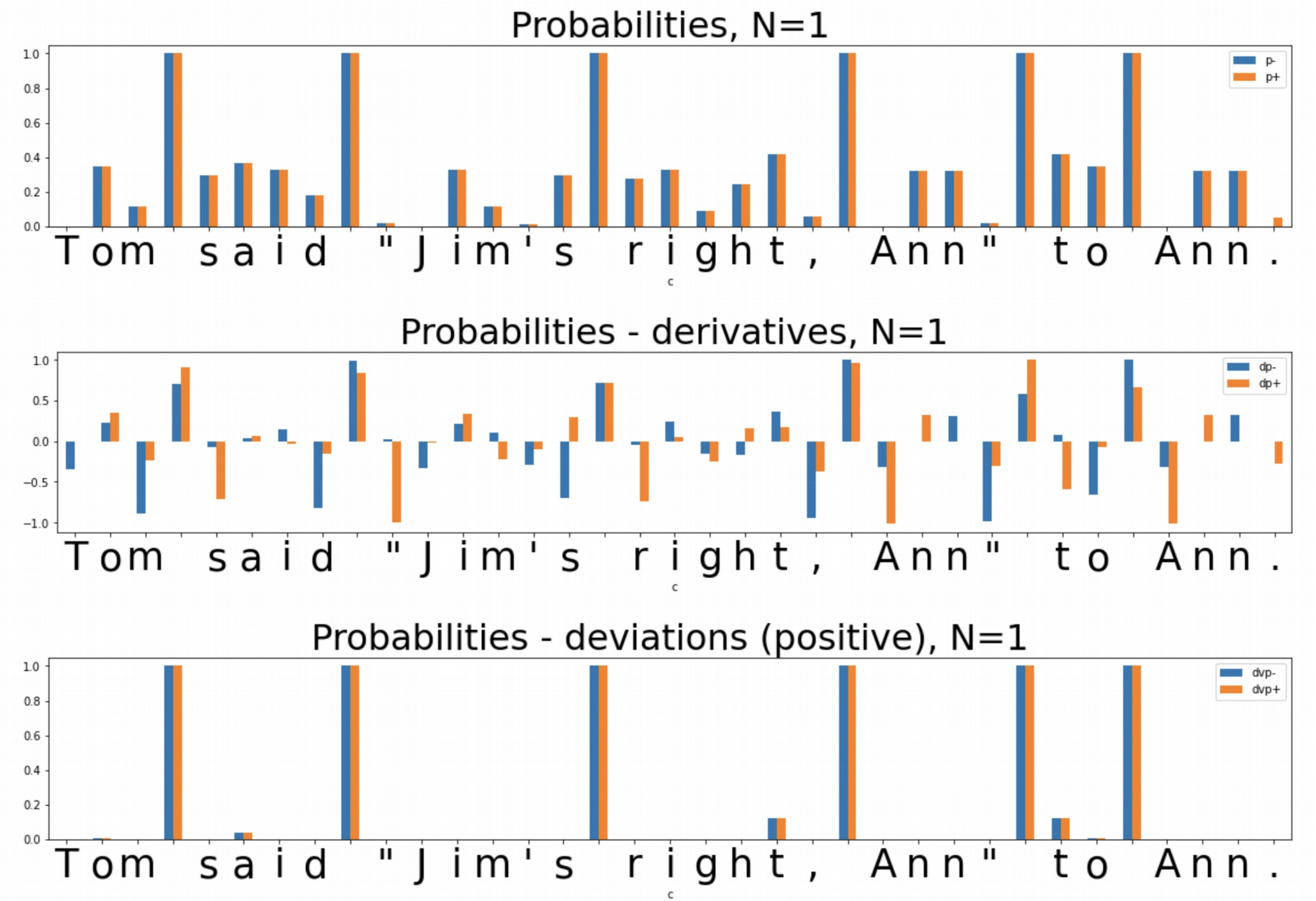
Combination of Ngram ranks N ([1],[2],[3],[1,2],[1,2,3],…)
Threshold for model compression
Threshold for segmentation

Evaluations:

Tokenization F1, on set of tokens found comparing to delimiter-based (English/Russian) or Jieba (Chineze)
Precision on set of tokens found comparing to reference lexicons

Unsupervised Text Segmentation (Tokenization)

Metrics/Indicators:
Ngram (Character)
Probability



Unsupervised Text Segmentation (Tokenization)

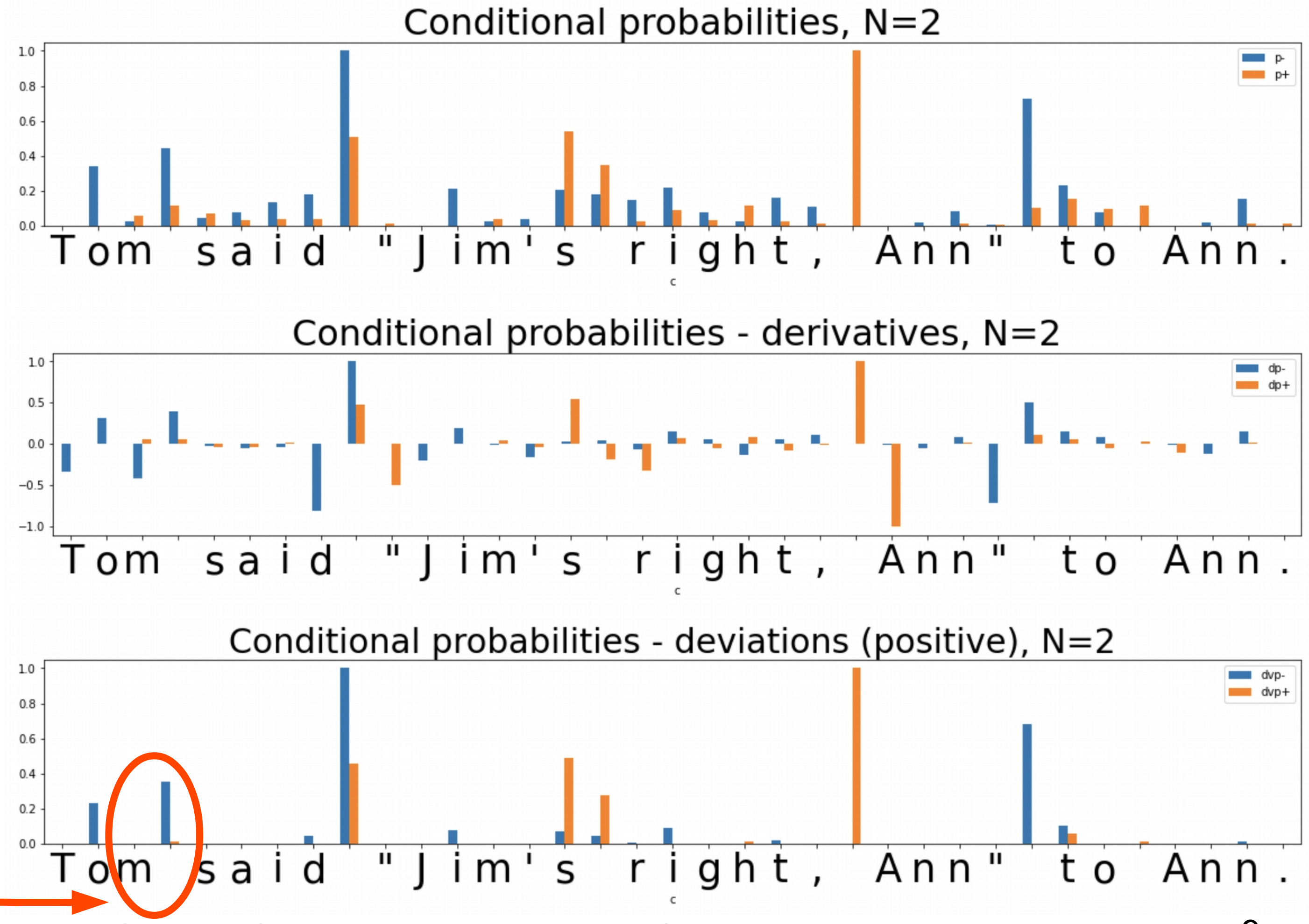
Metrics/Indicators:

Ngram (Character) Conditional Probability

(of Transition)

$P(Ngram_{n+1})/P(Ngram_n)$

P(“m ”)/P(“m”)



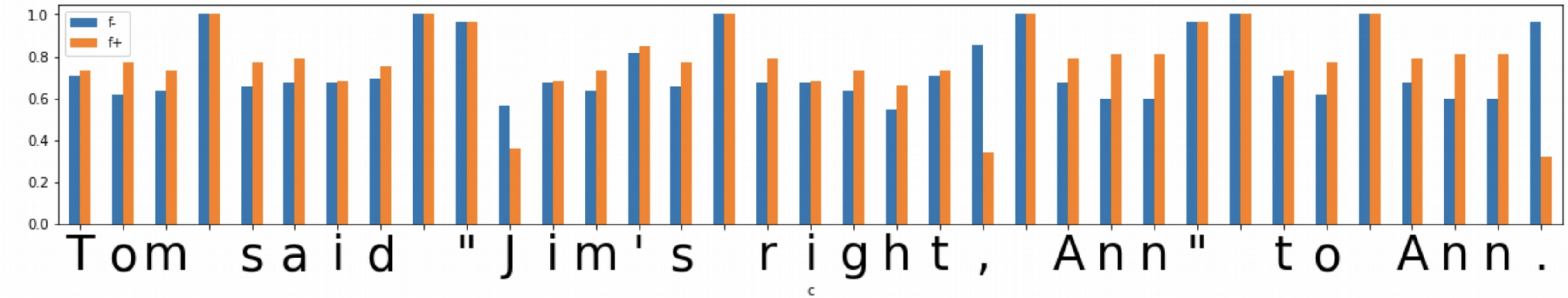
Unsupervised Text Segmentation (Tokenization)

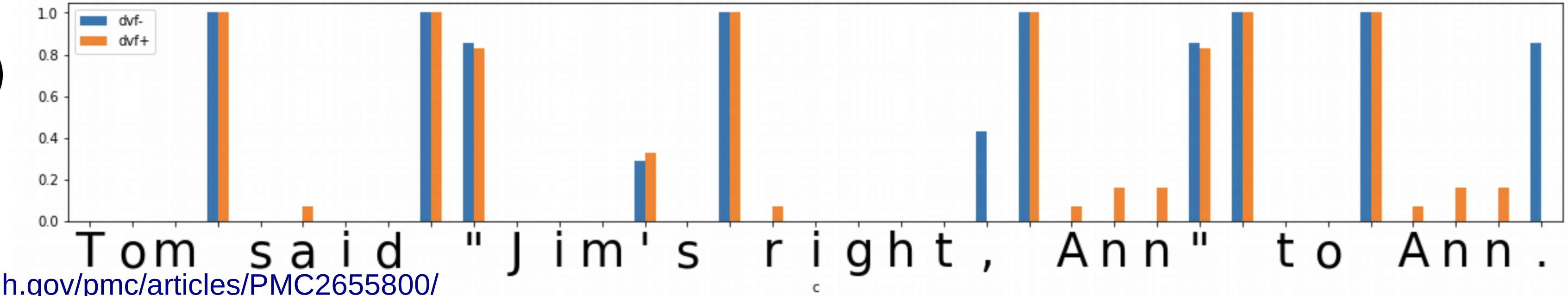
```
Threshold 0.25
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', 'Jim', "'", 's', ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
0.89
Threshold 0.35
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
1.0
```

Metrics/ Indicators:

Transition Freedom Deviation

(Freedom of Transition)





https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/

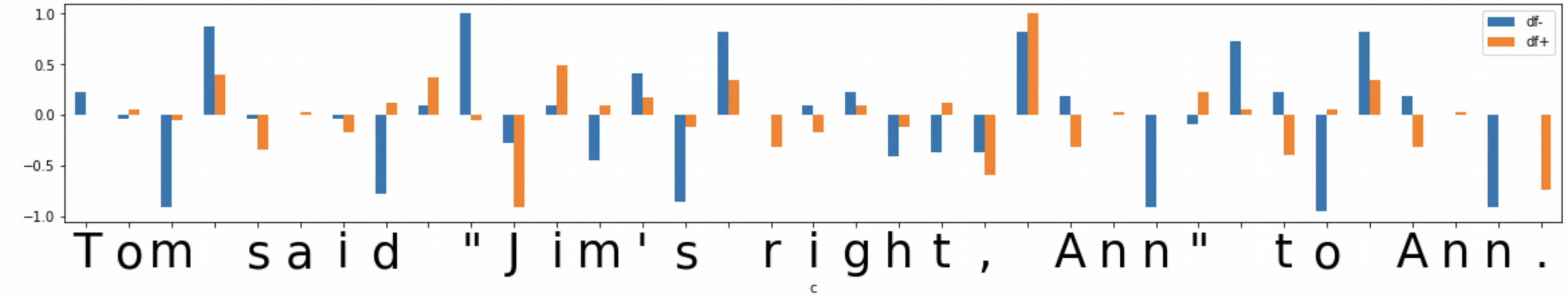
Unsupervised Text Segmentation (Tokenization)

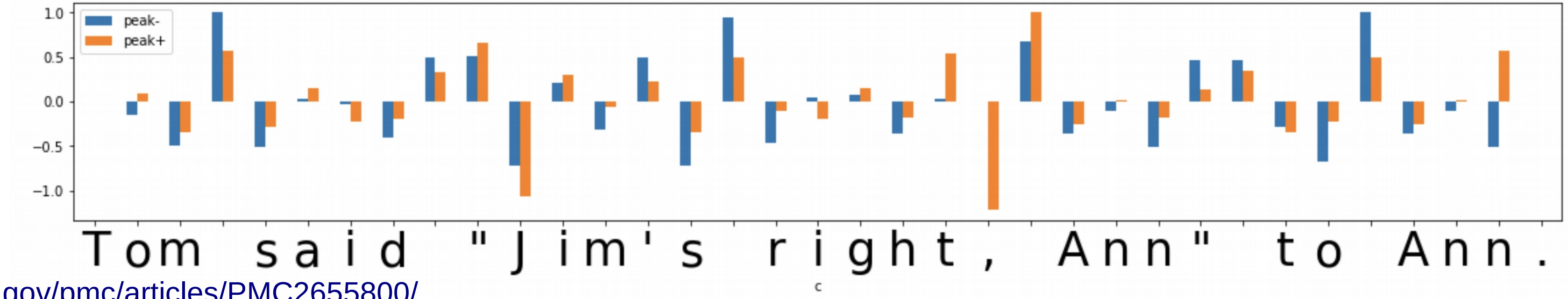
Metrics/ Indicators:

Transition Freedom Derivative and “Peak”

(Freedom of Transition “Peak”)

```
Threshold 0.25
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', 'Ji', 'm', "'s", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
0.89
Threshold 0.35
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', 'Jim', "'s", ' ', 'right', ',', ' ', 'Ann', '"', ' to', ' ', 'Ann', '.']
0.82
```





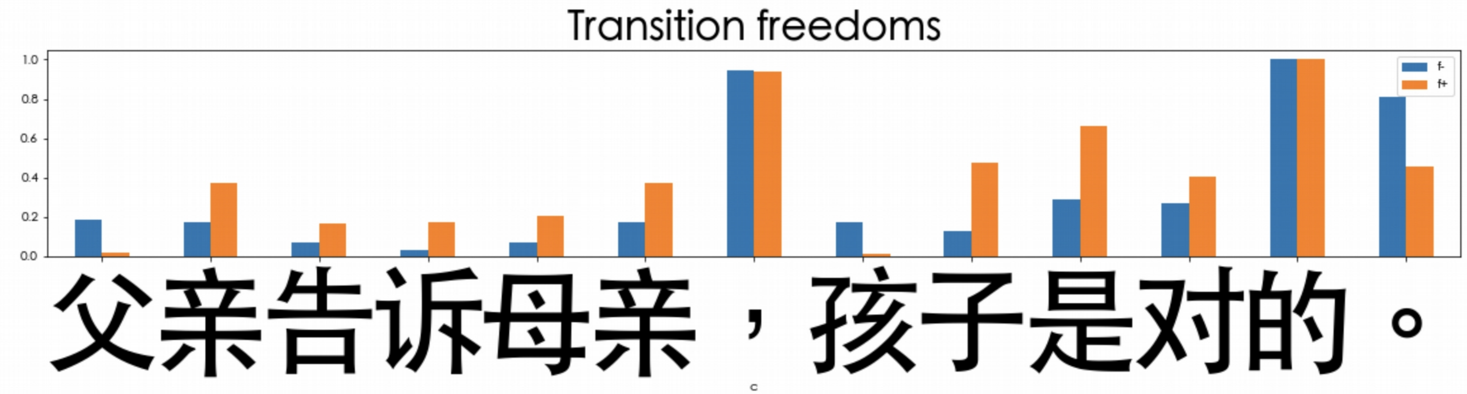
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/

Unsupervised Text Segmentation (Tokenization)

```
The father told the mother that the child was right.
Threshold  0.15
父亲告诉母亲，孩子是对的。
['父亲', '告诉', '母亲', '，', '孩子', '是', '对', '的', '。']
['父亲', '告诉', '母亲', '，', '孩子', '是', '对', '的', '。']
1.0
```

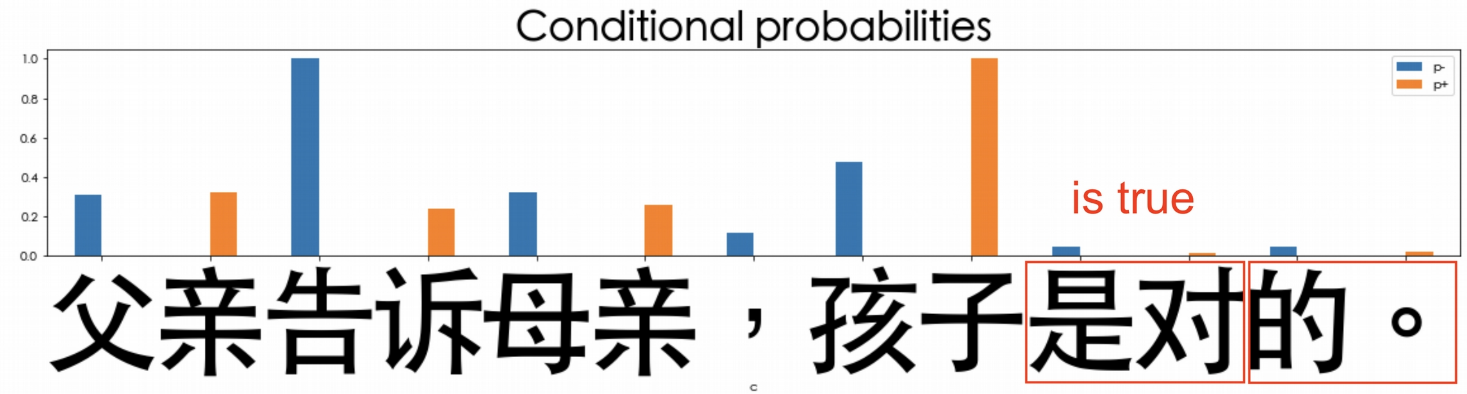
Metrics/Indicators:

Transition Freedom Deviation



Conditional Probability

(of Transition)



Unsupervised Text Segmentation (Tokenization)

Chinese

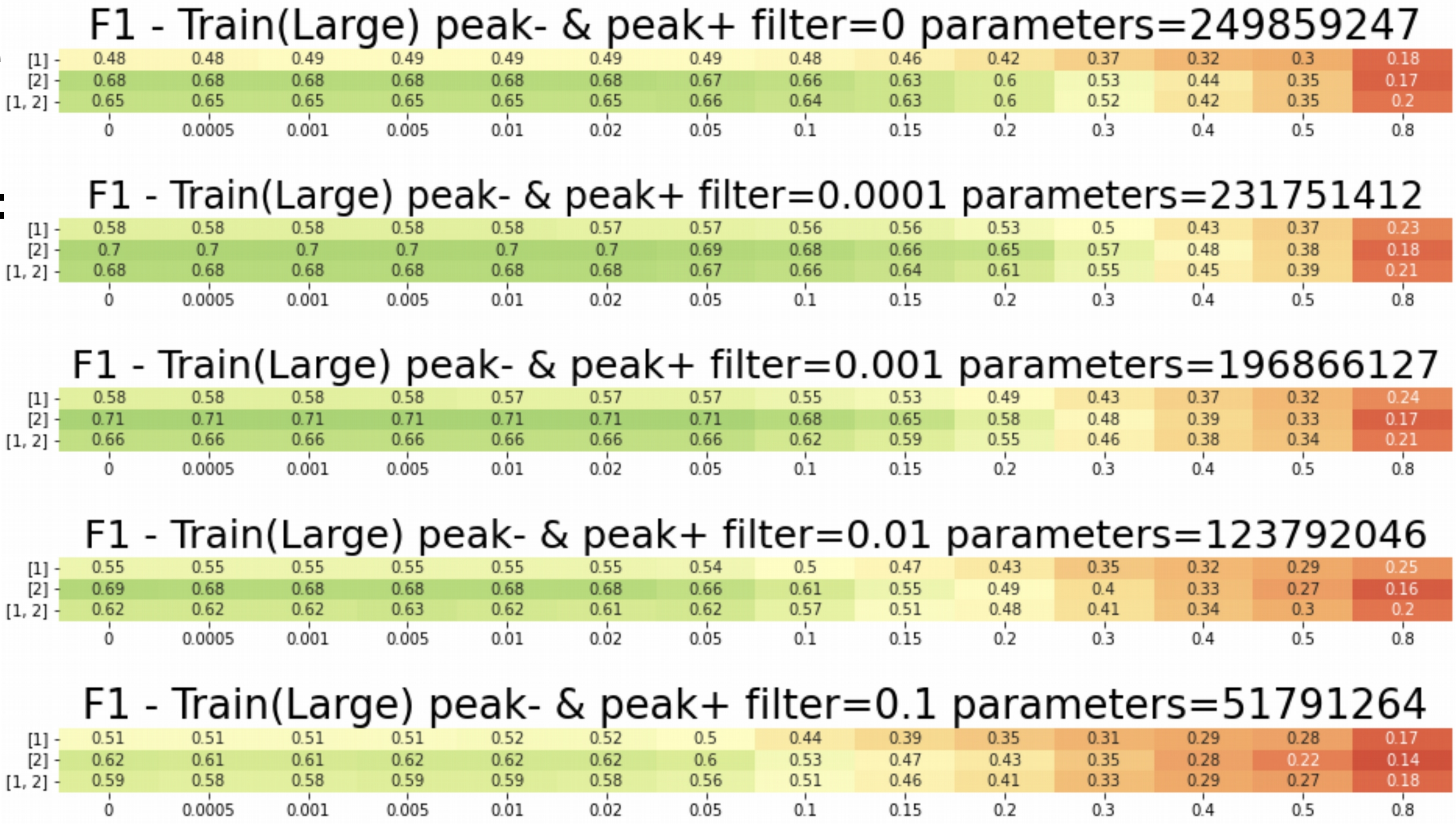
Hyper-Parameters:

TF “Peak”

Threshold for model compression

Combination of Ngram N-s

Threshold for segmentation



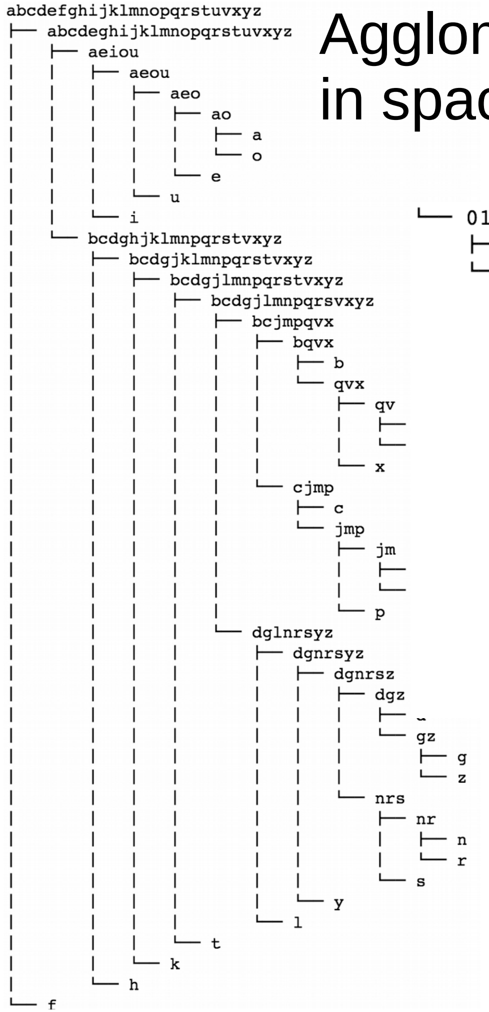
Results – Freedom-based Tokenization against Lexicon

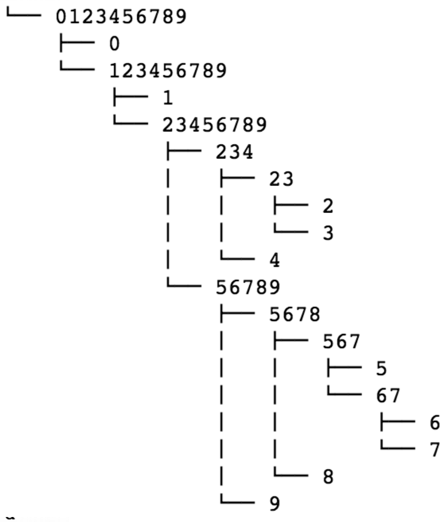
Language	Tokenizer	Tokenization F1	Lexicon Discovery Precision
English	Freedom-based	**0.99**	**0.99** (vs 1.0)
English	Lexicon-based	0.99	-
English no spaces	Freedom-based	0.42	-
English no spaces	Lexicon-based	0.79	-
Russian	Freedom-based	**1.0**	**1.0** (vs 1.0)
Russian	Lexicon-based	0.94	-
Russian no spaces	Freedom-based	0.26	-
Russian no spaces	Lexicon-based	0.72	-
Chinese	Freedom-based	**0.71**	**0.92** (vs 0.94)
Chinese	Lexicon-based	0.83	-

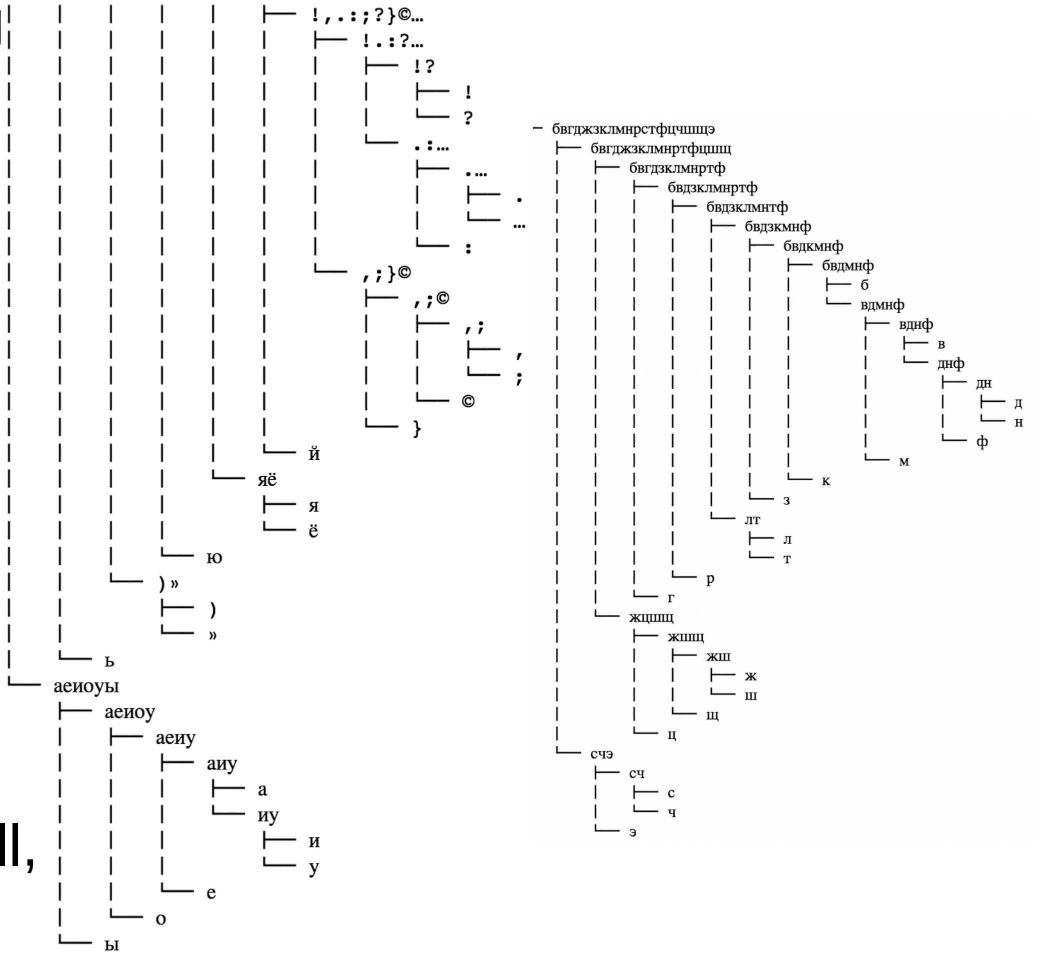
Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency

Unsupervised Character Category Learning

Agglomerative Clustering
in space of Transitions



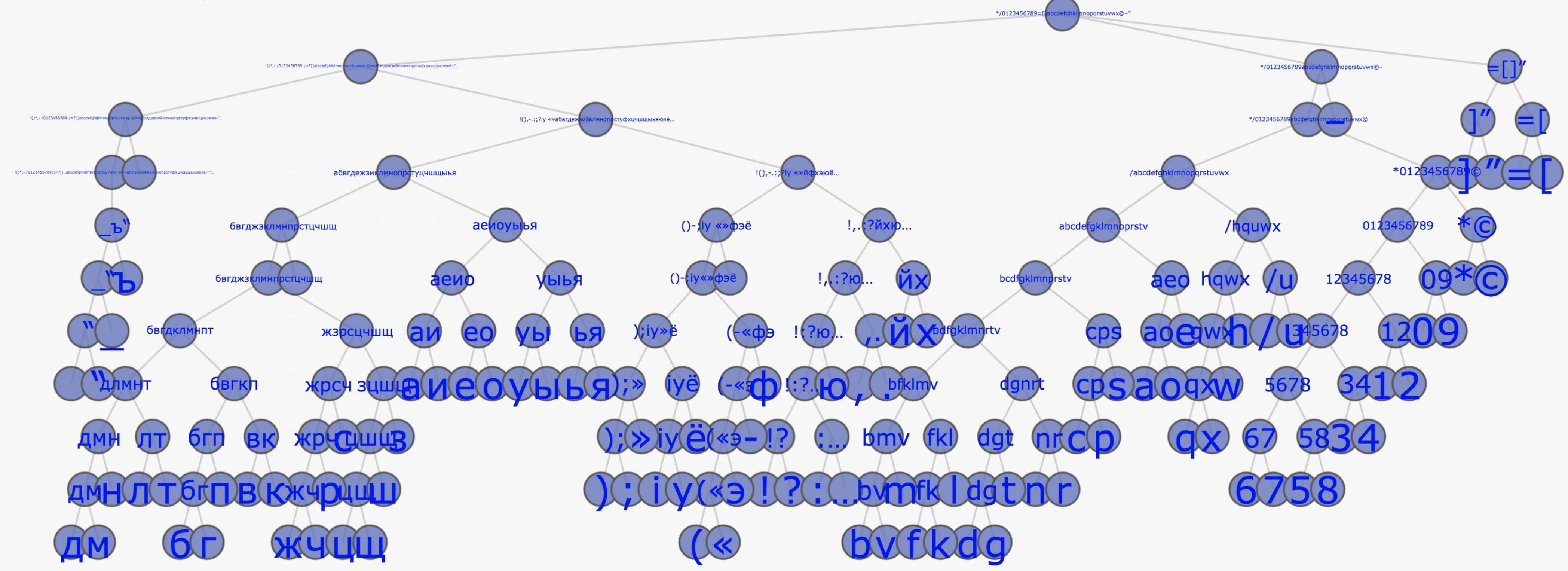




RusAge Test/Small,
Cosine Similarity

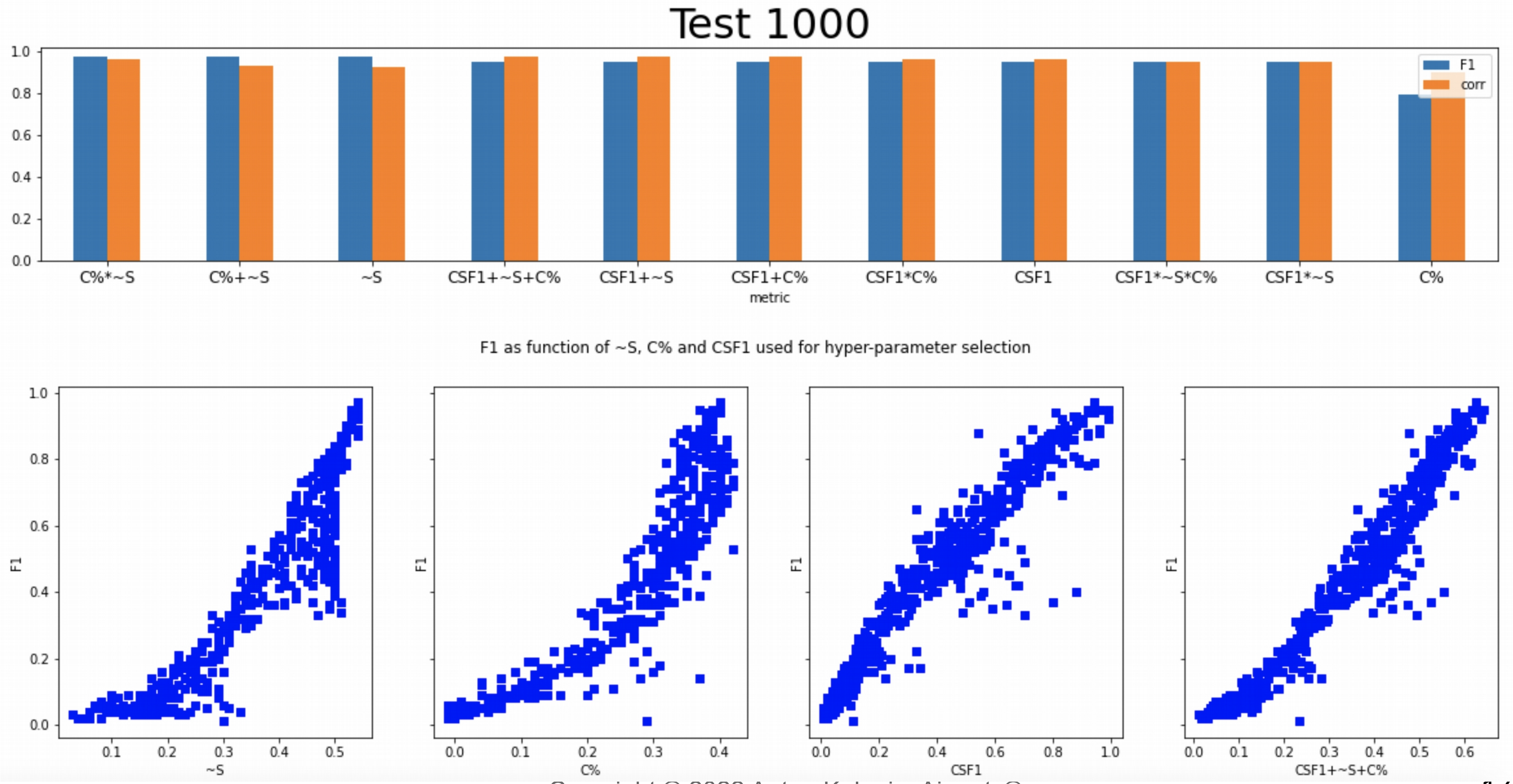
Unsupervised Character Category Learning

Agglomerative Clustering in space of Transitions

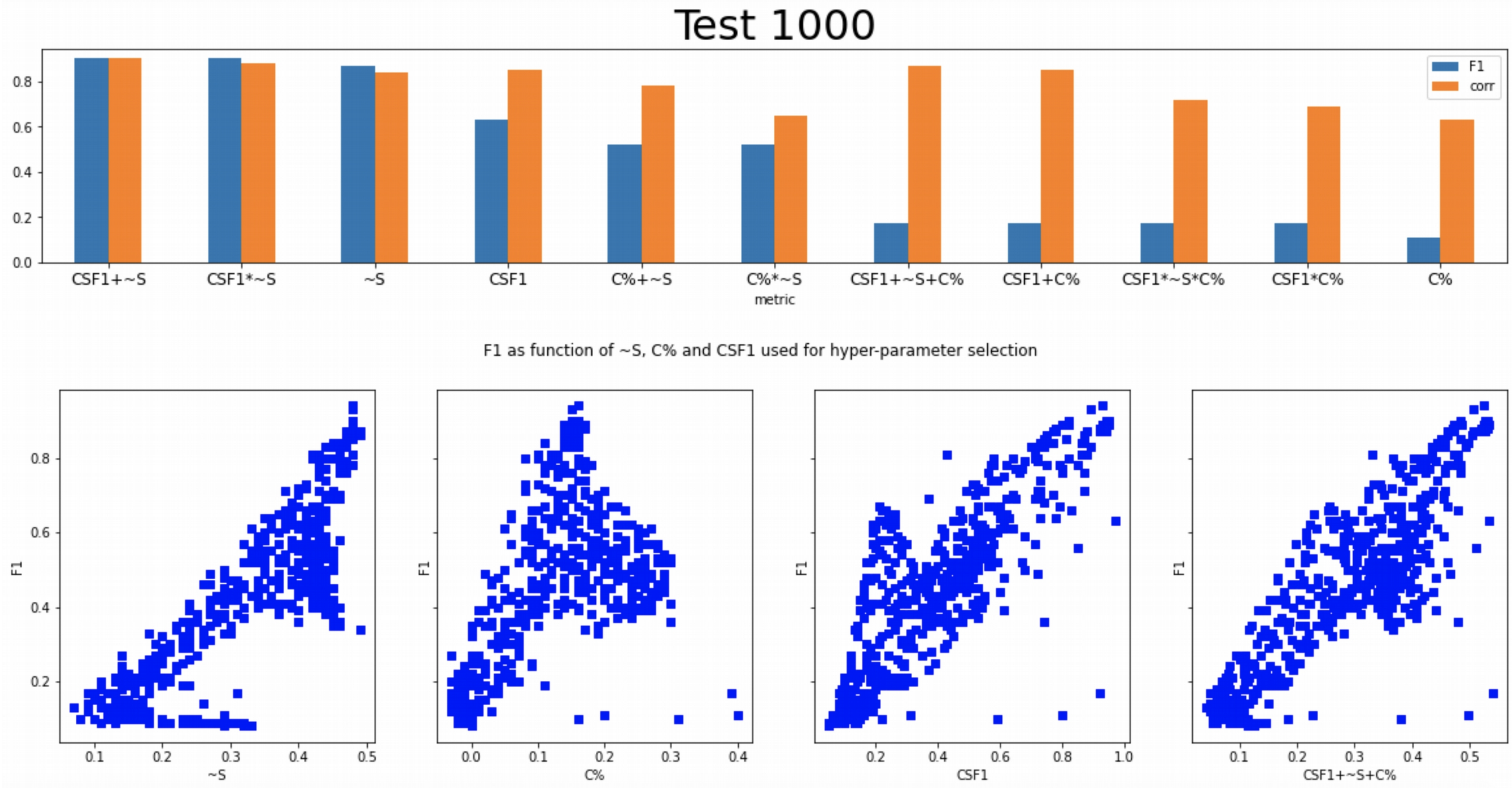


RusAge Previews/Full, Jackard Similarity

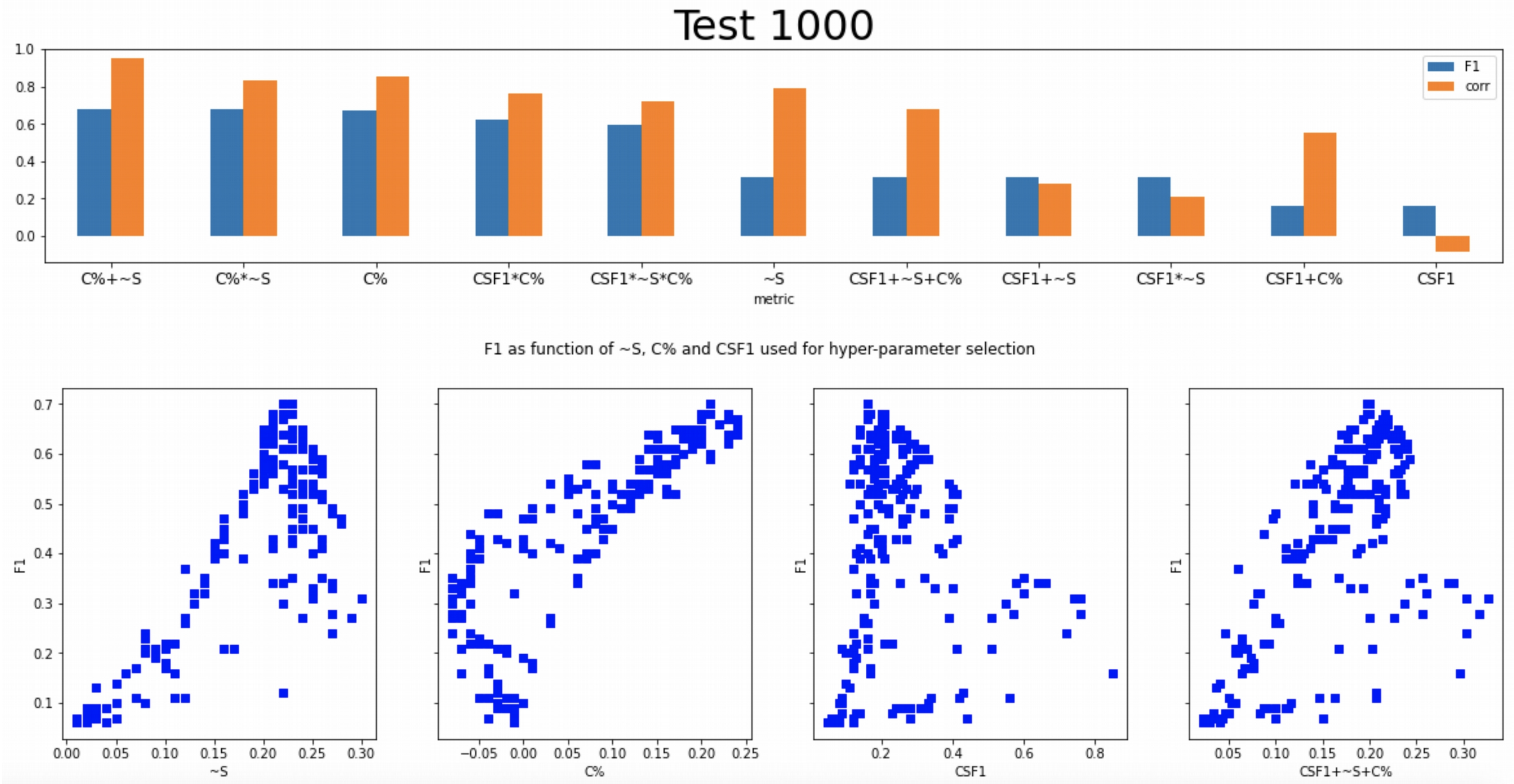
Self-tuning Hyperparameters – English (TF variance)



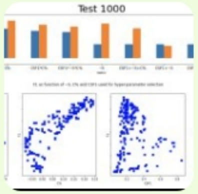
Self-tuning Hyperparameters – Russian (TF variance)



Self-tuning Hyperparameters – Chinese (TF “peak”)



Something about Human Intuition!



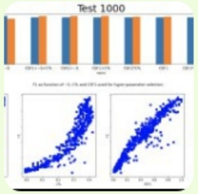
Screen Shot 2022-06-16 at 11.08.54.png

247.8 KB

OPEN WITH

Language 1

11:22



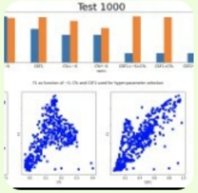
Screen Shot 2022-06-16 at 11.09.45.png

256.8 KB

OPEN WITH

Language 2

11:23



Screen Shot 2022-06-16 at 11.09.59.png

276.4 KB

OPEN WITH

Language 3

11:23

Poll

11% 1 - Russian

23% 2 - Russian

70% 3 - Russian

5% 1 - English

70% 2 - English

17% 3 - English

82% 1 - Chinese

0% 2 - Chinese

5% 3 - Chinese

VIEW RESULTS

3 11:25

Conclusion and Further Work

Unsupervised Tokenization based on Transition Freedom (TF) recall and precision appears good enough as initial approximation for further applications of self-reinforcement learning as part of interpretable unsupervised learning of natural language.

Optimal thresholds and specific TF-based metrics are specific to language. The process and policy of their discovery and adjustment should be further explored.

Clustering or parts of speech on space of transition graphs may provide some insights on morphology and punctuation structure of low-resource and domain-specific languages.

Hybridization of TF-based tokenization approach with lexicon-based one might be efficient for low-resource and domain-specific languages.

Further unsupervised grammar learning experiments can be run on the basis of suggested unsupervised tokenization approach.

Applications for other Experiential Learning environments, including the ones with delayed/sparse feedback.

Using Reinforcement Learning techniques with self-reinforcement on historical data under Unsupervised Learning setup.

https://arxiv.org/abs/2205.11443
https://github.com/aigents/pygents

Thank You and Welcome!

Anton Kolonin

akolonin@aigents.com

Facebook: akolonin

Telegram: akolonin

Novosibirsk State University

*THE REAL SCIENCE

SingularityNET

AGI IN RUSSIAN

SingularityDAO

https://agirussia.org

AIGENTS

https://aigents.com