Self-tuning hyper-parameters for unsupervised cross-lingual tokenization

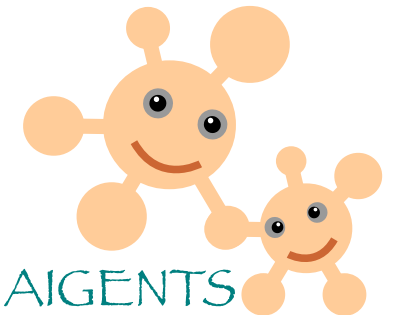
Anton Kolonin

akolonin@aigents.com

Facebook: akolonin

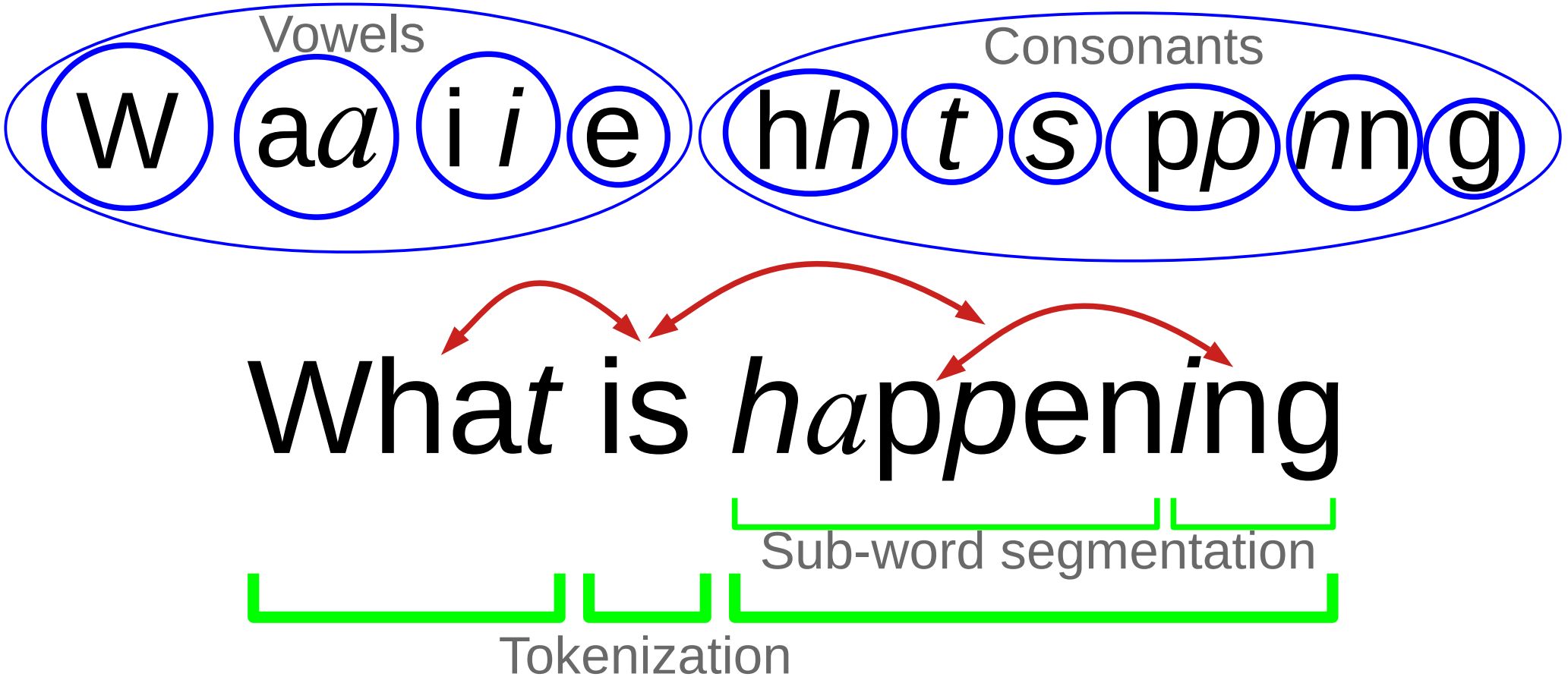
Telegram: akolonin

https://agirussia.org

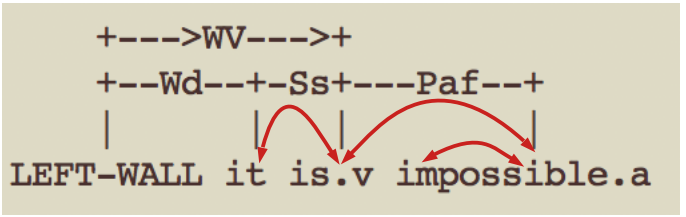


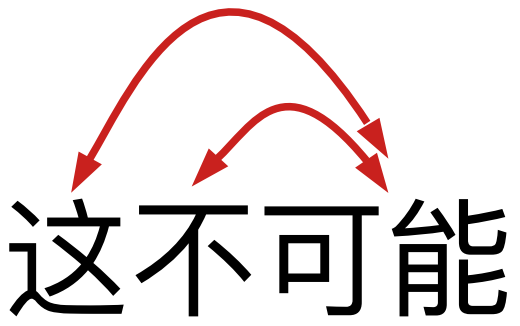
https://aigents.com

Learning Lexicon, Punctuation, Morphology and Grammar
Clustering, Segmentation and Parsing

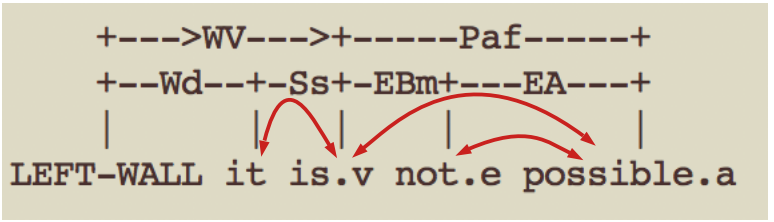


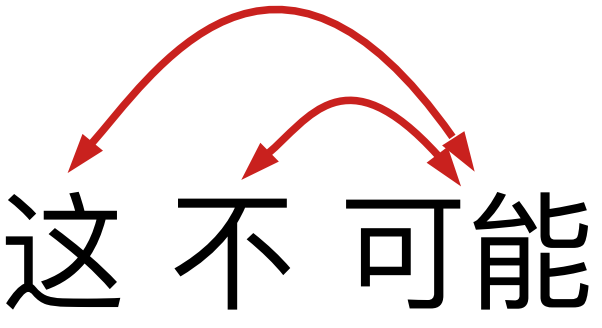
Blurring boundaries between grammatical parsing and morphological parsing causes tokenization ambiguity





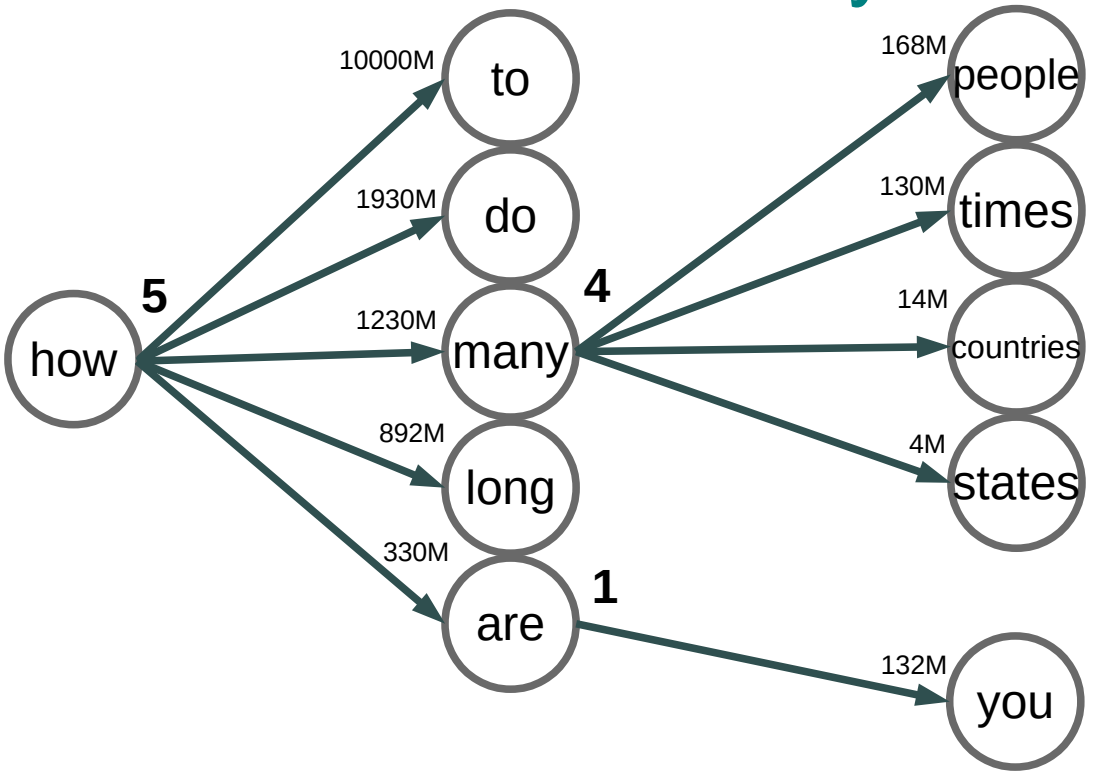
(Google)





(Jieba)

Unsupervised Learning for Text Segmentation based on Probability and Uncertainty Measures



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

Unsupervised Text Segmentation (Tokenization)

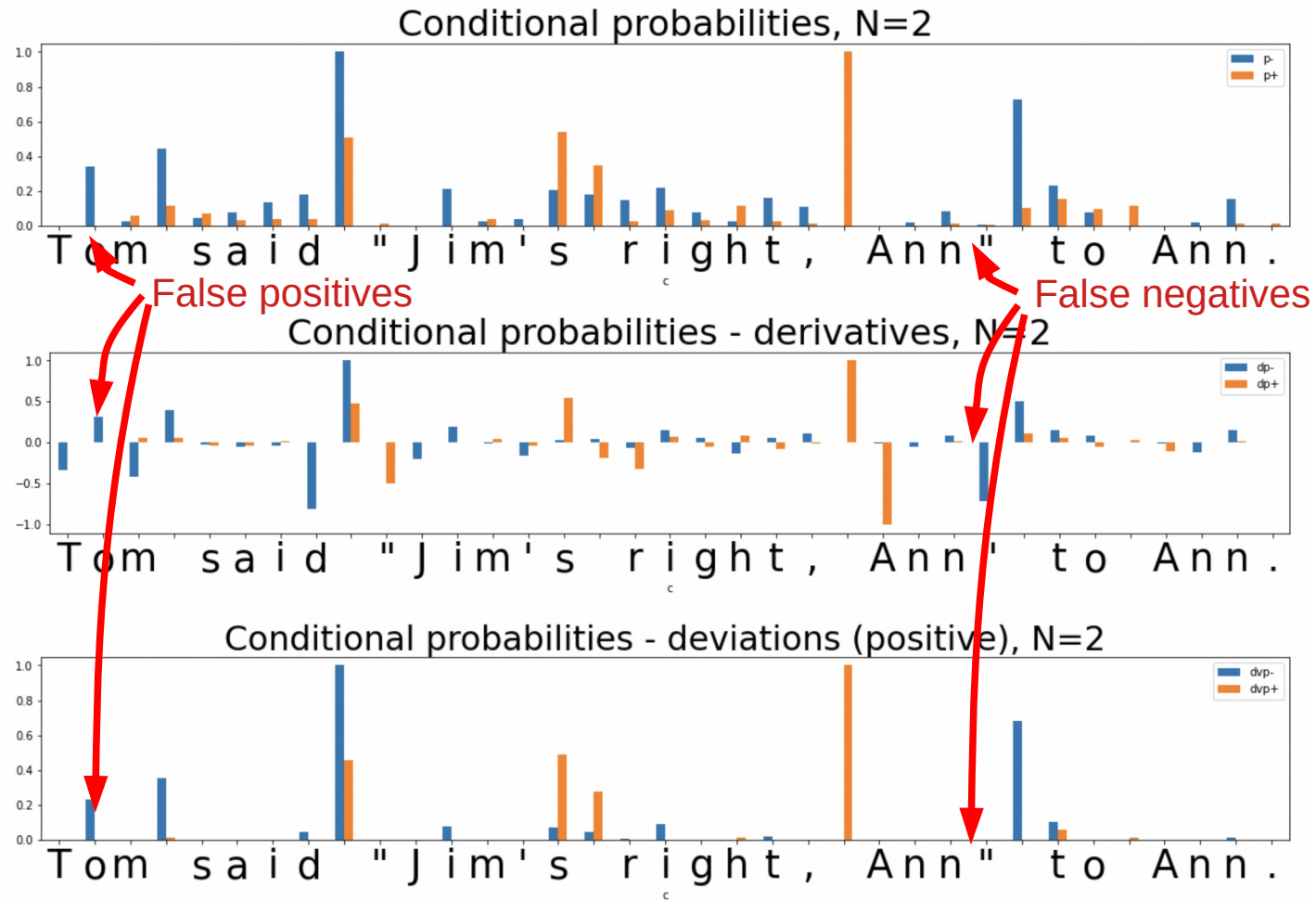
Metrics/Indicators:

Ngram (Character) Conditional Probability

(of Transition)

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

P(“m ”)/P(“m”)

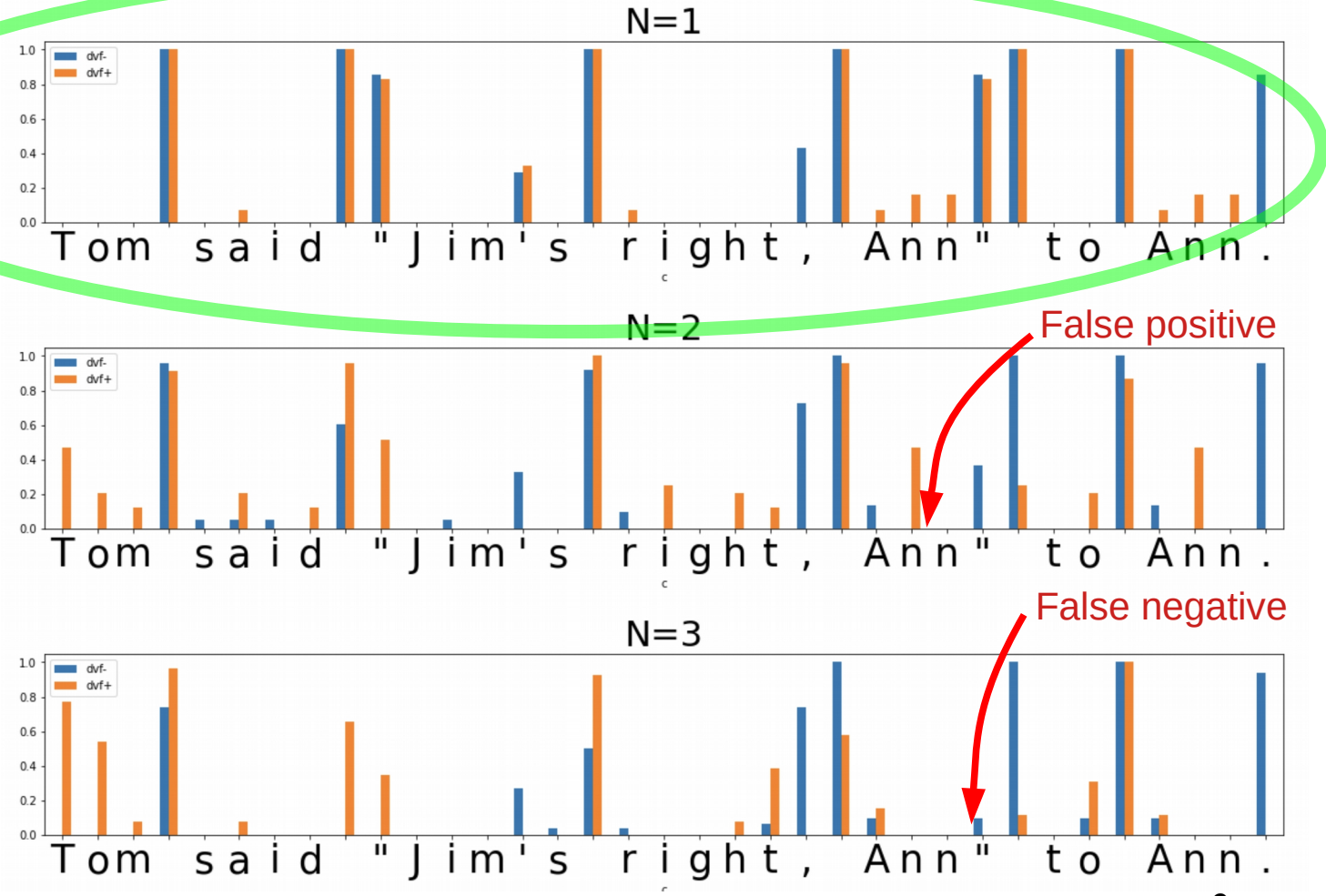


Unsupervised Text Segmentation (Tokenization)

Metrics/ Indicators:

Transition
Freedom
Deviation

(varying “N” of N-gram)



Unsupervised Text Segmentation (Tokenization)

English

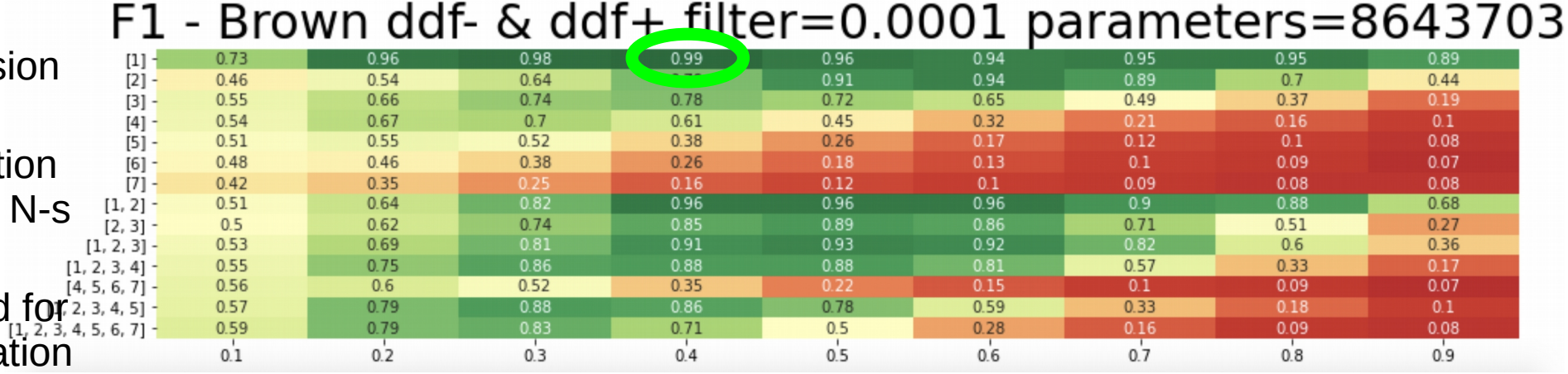
Hyper-Parameters:

Metric: Transition Freedom

Threshold for model compression

Combination of Ngram N-s

Threshold for segmentation



Results – Freedom-based Tokenization against Lexicon-based one (referring to Rule-based)

Language	Tokenizer	Tokenization F_1	Lexicon Discovery Precision
English	Freedom-based	**0.99**	**0.99** (vs. 1.0)
English	Lexicon-based*	0.99	-
Russian	Freedom-based	**1.0**	**1.0** (vs. 1.0)
Russian	Lexicon-based*	0.94	-
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*

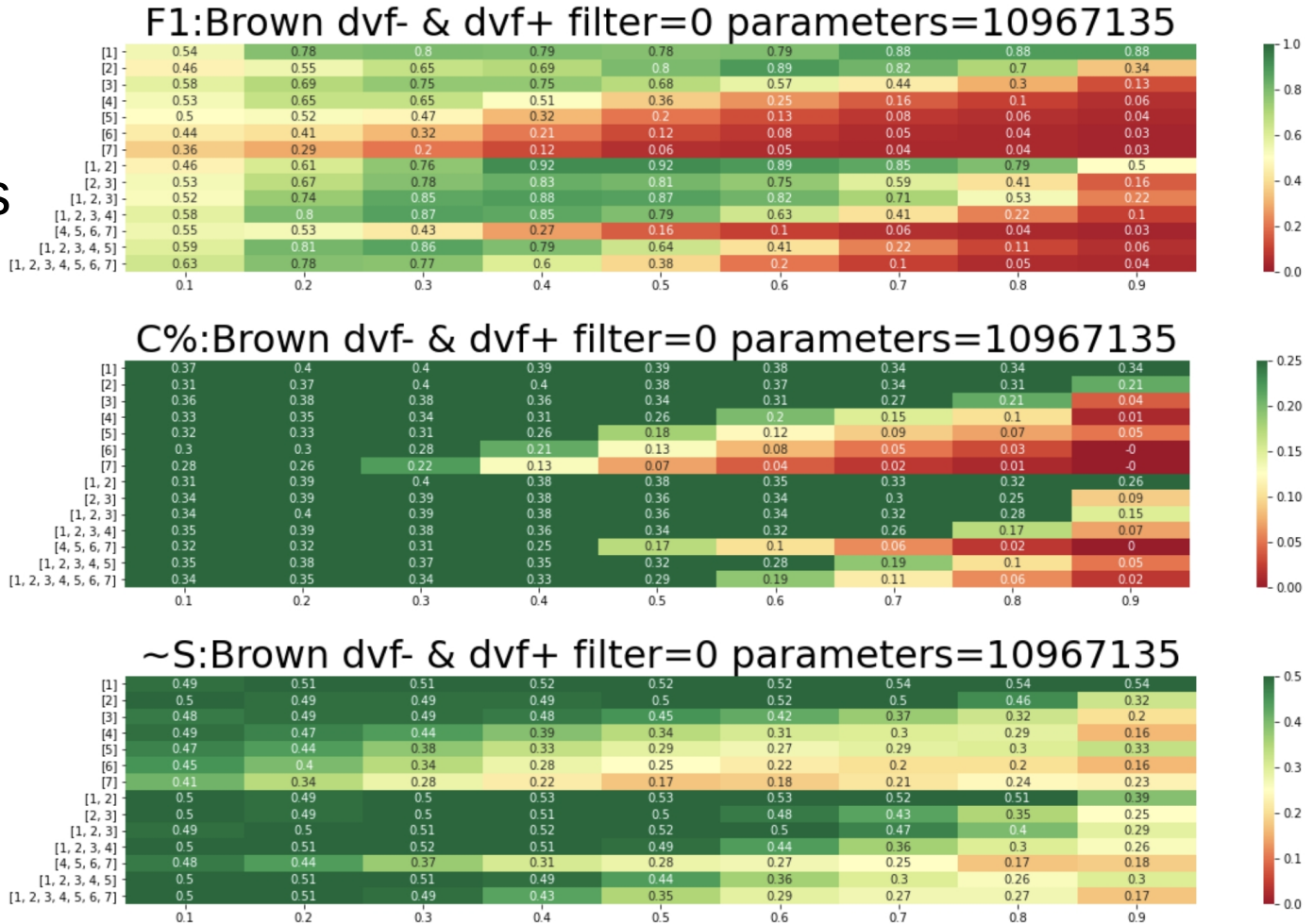
Preprint: https://arxiv.org/abs/2205.11443 **EMNLP-2022**

Hyper-parameters: F1 vs. language-agnostic metrics

English,

Brown corpus

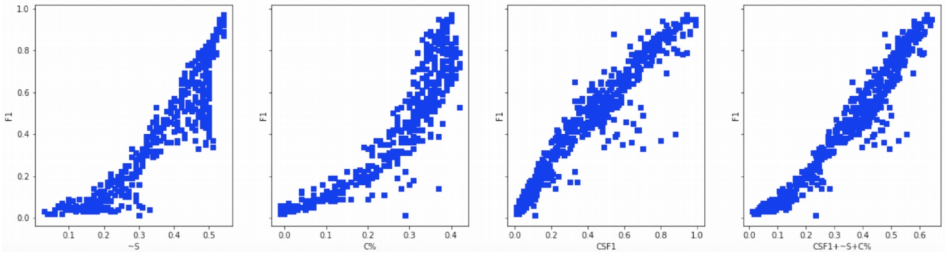
Maximizing F1, compression factor (C%) and normalized anti-entropy (~S) in the space of hyper-parameters



https://arxiv.org/abs/2303.02427

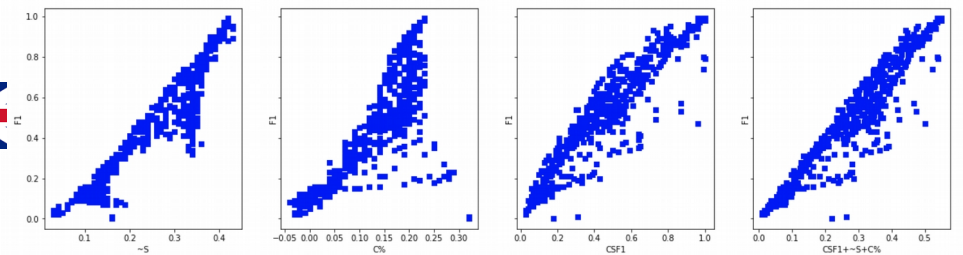
Different corpora, different sizes

English, Train: Brown, Test: Brown 1000

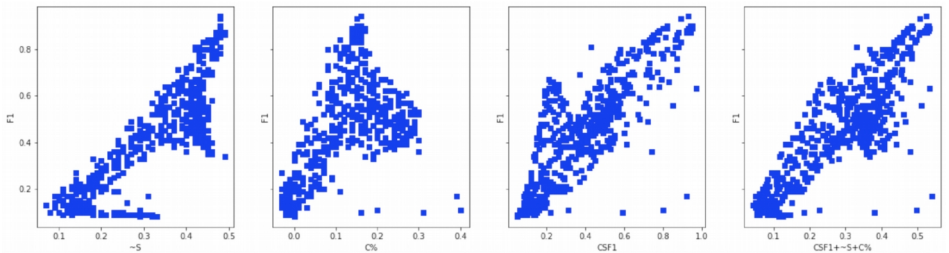


https://github.com/aigents/pygents/blob/main/notebooks/nlp/english/en_tokenizer_auto.ipynb

English, Train: Brown, Test: MagicData 100

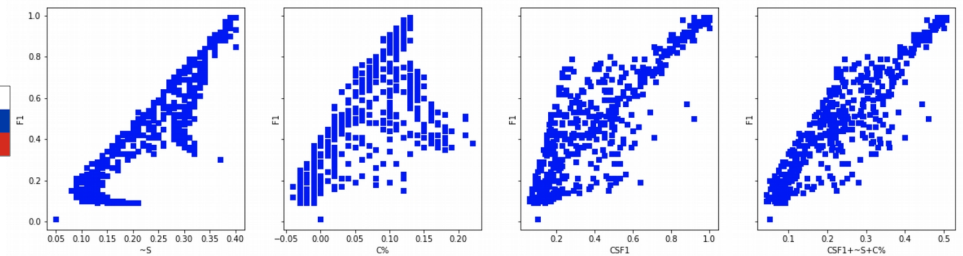


Russian, Train: RusAge, Test: RusAge 1000

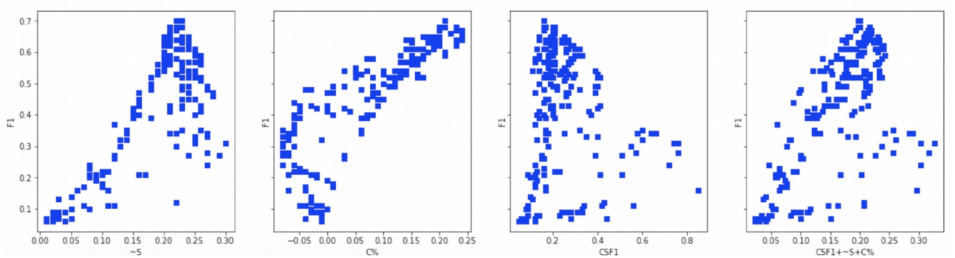


https://github.com/aigents/pygents/blob/main/notebooks/nlp/russian/ru_tokenizer_auto.ipynb

Russian, Train: Brown, Test: MagicData 100

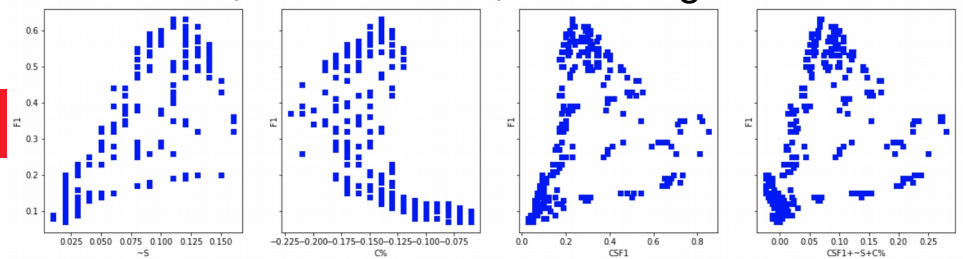


Chinese, Train: CLUE News, Test: CLUE News 1000



https://github.com/aigents/pygents/blob/main/notebooks/nlp/chinese/zh_tokenizer_auto.ipynb

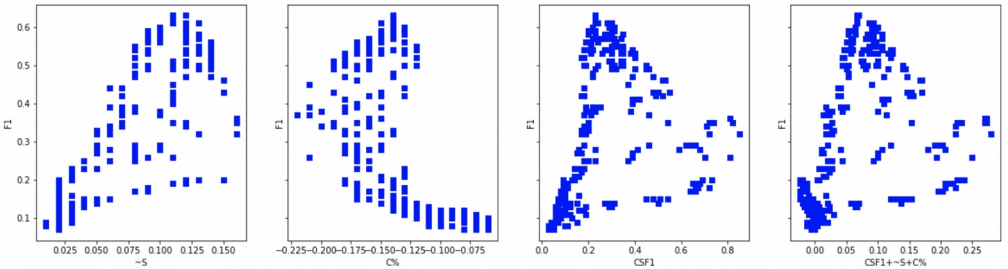
Chinese, Train: Brown, Test: MagicData 100



https://github.com/aigents/pygents/blob/main/notebooks/nlp/tokenization/brown/tokenization_brown_en_ru_zh.ipynb

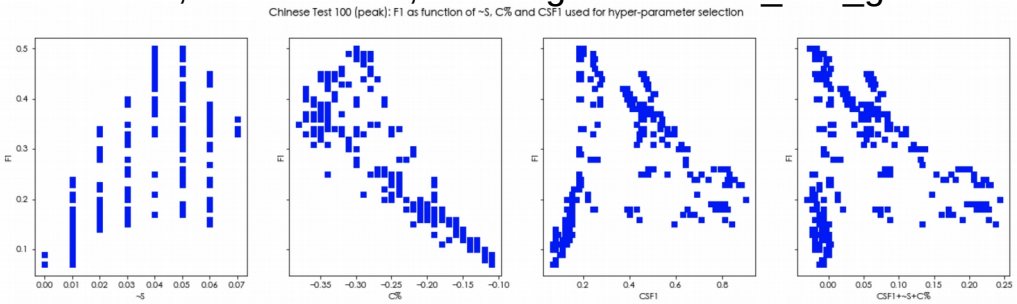
Chinese corpora, different sizes

Chinese, Train: Brown, Test: MagicData 100

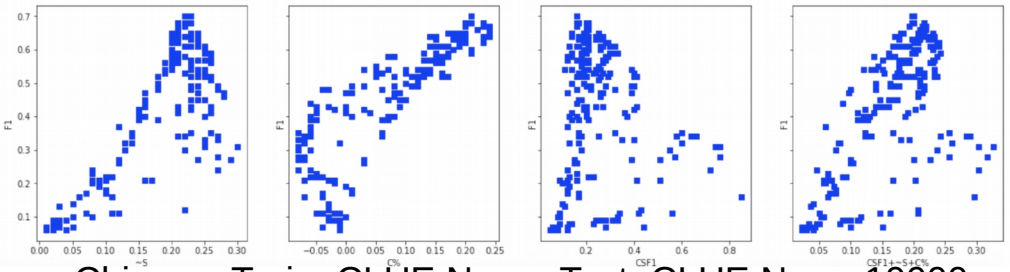


https://github.com/aigents/pygents/blob/main/notebooks/nlp/tokenization/brown/tokenization_brown_en_ru_zh.ipynb

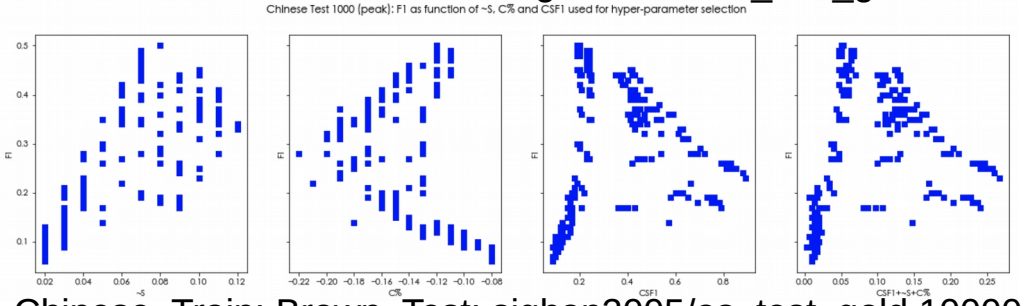
Chinese, Train: Brown, Test: sighan2005/as_test_gold 100



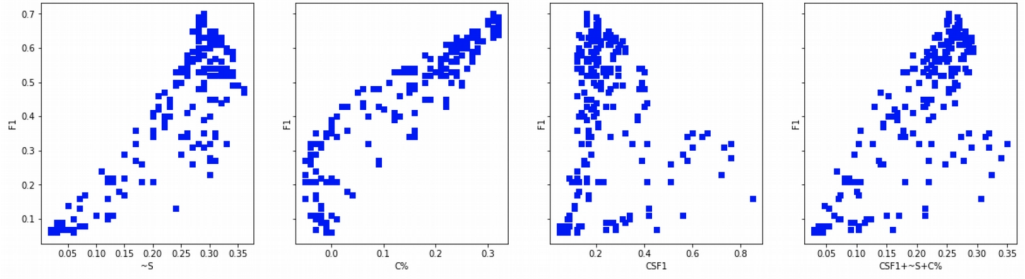
Chinese, Train: CLUE News, Test: CLUE News 1000



Chinese, Train: Brown, Test: sighan2005/as_test_gold 1000

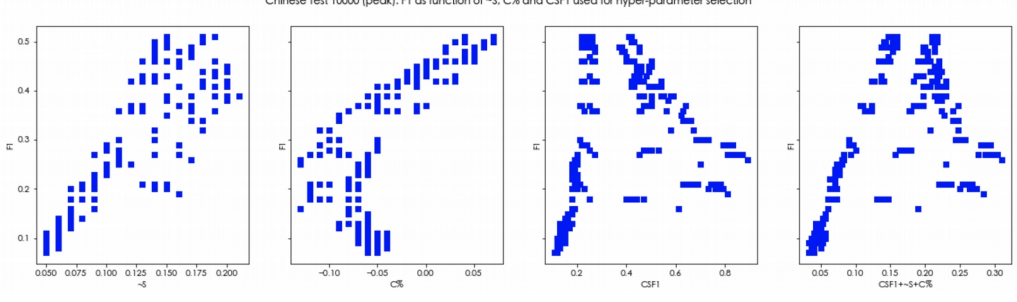


Chinese, Train: CLUE News, Test: CLUE News 10000



https://github.com/aigents/pygents/blob/main/notebooks/nlp/chinese/zh_tokenizer_auto.ipynb

Chinese, Train: Brown, Test: sighan2005/as_test_gold 10000



https://github.com/aigents/pygents/blob/main/notebooks/nlp/chinese/zh_tokenizer_multi-criteria-cws.ipynb

Unsupervised learning for subword segmentation

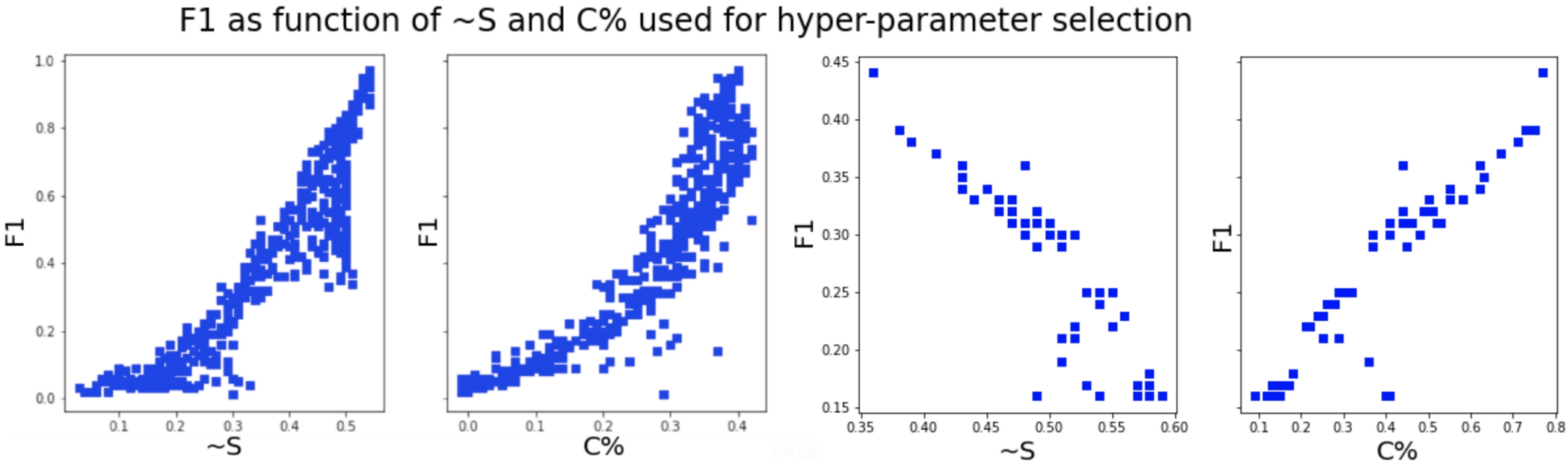
Reference	Morphology-based	BPE	DPE	Transtion-freedom-based
['euro', 'zone']	['eu', 'rozone']	['eurozone']	['euro', 'zone']	['euro', 'z', 'one']
['entrepreneur', 'ship']	['ent', 're', 'pre', 'neur', 'ship']	['entrepreneurship']	['entrepreneur', 'ship']	['entre', 'preneur', 'sh', 'ip']
['pre', 'sent', 'ly']	['pre', 's', 'ent', 'ly']	['pres', 'ently']	['present', 'ly']	['pre', 's', 'ently']
['bloc']	['bloc']	['blo', 'c']	['bl', 'oc']	['b', 'lo', 'c']
['tree', 's']	['tr', 'ee', 's']	['tre', 'es']	['tr', 'ees']	['tre', 'es']
['multi', 'lateral', 'ism']	['multi', 'lat', 'er', 'al', 'ism']	['multilater', 'alism']	['multilateral', 'ism']	['multi', 'later', 'al', 'ism']
['motive', 's']	['mot', 'ive', 's']	['mo', 'tives']	['motiv', 'es']	['mo', 'tiv', 'es']
['progress', 'ive', 'ly']	['pro', 'gr', 'ess', 'ive', 'ly']	['pro', 'gressively']	['progressive', 'ly']	['pro', 'gressiv', 'ely']
['de', 'cent', 'ral', 'isation']	['dec', 'ent', 'r', 'al', 'isation']	['decent', 'ralisation']	['decent', 'ral', 'isation']	['de', 'centralis', 'ation']
['margin', 'al', 'is', 'ed']	['marginali', 's', 'ed']	['margin', 'alised']	['marginal', 'ised']	['mar', 'ginal', 'is', 'ed']
['re', 'cast']	['re', 'cast']	['rec', 'ast']	['re', 'cast']	['re', 'c', 'ast']
['out', 'line', 's']	['out', 'l', 'ine', 's']	['out', 'lines']	['outline', 's']	['out', 'l', 'ines']
['pre', 'vent', 'at', 'ive']	['pre', 'v', 'ent', 'ative']	['preven', 'tative']	['prevent', 'ative']	['pre', 'vent', 'ative']
['en', 'danger', 'ed']	['end', 'an', 'g', 'er', 'ed']	['endang', 'ered']	['endanger', 'ed']	['en', 'dang', 'ered']
['vulner', 'abil', 'ity']	['vulnerabil', 'ity']	['vul', 'n', 'era', 'bility']	['vul', 'ner', 'ability']	['vul', 'ner', 'ability']
F1	**0.46**	**0.05**	**0.55**	**0.25**

Ref:
https://arxiv.org/pdf/2005.06606.pdf

https://github.com/aigents/pygents/blob/main/data/corpora/English/morphology/prefixes.txt
https://github.com/aigents/pygents/blob/main/data/corpora/English/morphology/suffixes.txt

Tokenization vs. sub-word segmentation

F1 connection to anti-entropy and compression factor (English)



Takeaways

Grammar and syntactic/semantic word categories **can** be learned, given we can learn tokenization and parses unsupervisedly.

Tokenization and character categories **can** be learned unsupervisedly based on transition freedom metric.

Hyper-parameters for unsupervised text segmentation learning **can** be found based on culture-agnostic metrics such as compression factor (and anti-entropy in case of tokenization).

Unsupervised learning for morphological units and morpho-parsing **may** be possible but remains non-trivial and needs more study.

Welcome to the Interpretable Natural Language Processing Community and Series of Workshops

AGI-in-Russian on Telegram
https://t.me/agirussia

INLP on Telegram
https://t.me/internlp

INLP Workshops
https://aigents.github.io/inlp/