State of the #NLProc

Vsevolod Domkin

Iforum 2017
2017-05-25

About me

* Lisp programmer
* 5+ years of NLP work at Grammarly
* (m8n)ware NLP consultancy
* volunteer at lang-uk (http://lang.org.ua)
* Lecturer on OS, Algortihms, NLProc (KPI, Projector, UCU)

https://vseloved.github.io

Outline

- Intro to NLProc
- Academic & industrial NLProc: theory & practice
- Problems, algorithms, tools, datasets
- Main directions of current and future development

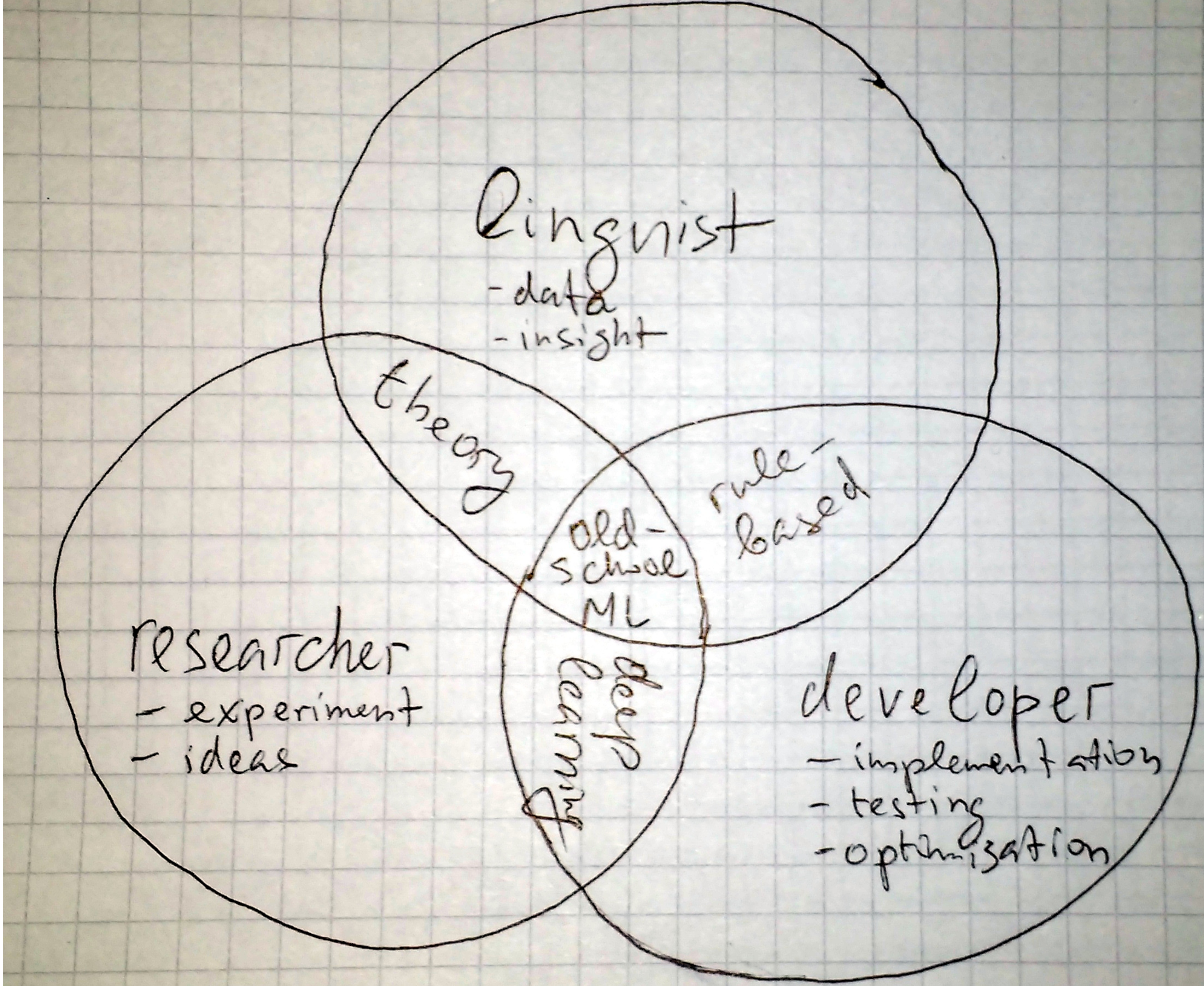
What is Natural Language Processing?

Transforming free-form text into
structured data and back
since 1950

What is Natural Language Processing?

Transforming free-form text into
structured data and back
since 1950

Related, but not quite:
- Compling
- AI
- Data Science
- Regular expressions, Neural Networks...


linguist
- data
- insight
theory
rule-based
old-school ML
researcher
- experiment
- ideas
deep learning
developer
- implementation
- testing
- optimisation

Computational Linguistics

- dictionaries, lexicons, thesauri, ontologies
- labelling (POS, NER, coref, semantic roles, ...)
- parsing (constituency, dependency, semantic, abstract, ...)
- modelling (language, topics, ...)
- embedding (word2vec, doc2vec, ...)

Data

Plenty!

- Dictionaries
- Corpora (Penn Treebank, Brown, Europarl, Gigaword, ...)
- Shared task resources

Algorithms (classic)

- Reliance on rich linguistic data and features
- Ngram-based linear models
- Shift-reduce-based parsing
- PCA-based semantics

Deep Learning

“NLP is kind of like a rabbit in the headlights of the Deep Learning machine, waiting to be flattened.”

Deep Learning

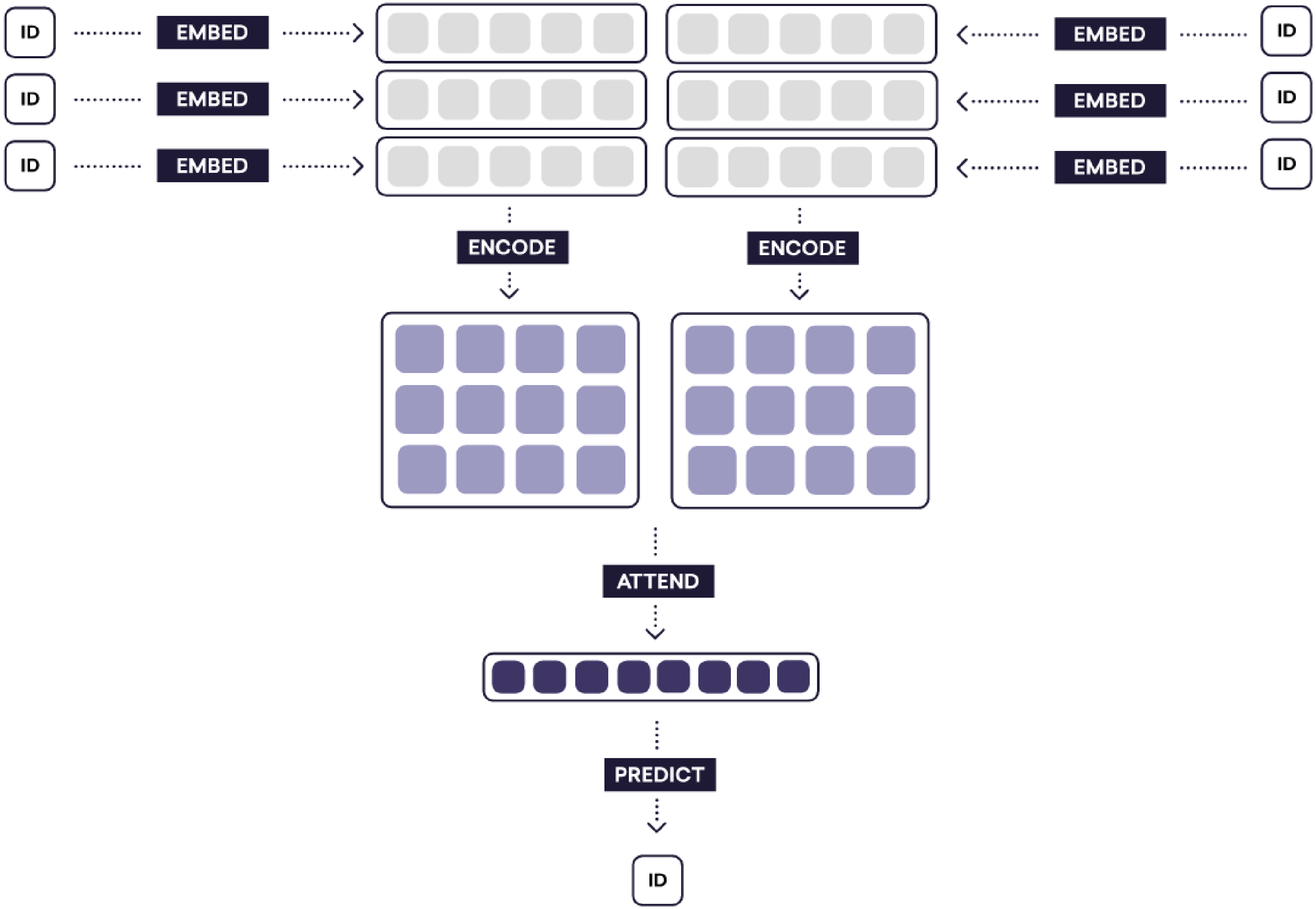
- Vector-space word & document representations
- Recurrent neural networks

Haven't really taken the classic tasks by storm (like in Signal Processing), but seriously expanded what's possible

DL for NLP Formula

“Embed, encode, attend, predict”

https://explosion.ai/blog/deep-learning-formula-nlp



Tools

Everything, basically, happens in 2 major languages: Python & Java

A number of solid NLProc frameworks: Stanford, OpenNLP, Spacy, LingPipe

A plethora of academic-quality targeted tools

word2vec etc.

Industrial NLProc

Main end-user applications:

Groupping & labelling

- sentiment analysis
- plagiarism detection
- classification/ clustering
- parts extractions

Transformation

- machine translation
- summarization
- error correction
- generation

Dialog

- information retrieval
- QA
- conversational interfaces

Data

None :(

(b/c verticals)

Need to create your own:

- Extract
- Annotate
- Crowdsource
- Generate as by-product

Algorithms

Hybrid approaches

Rule-based

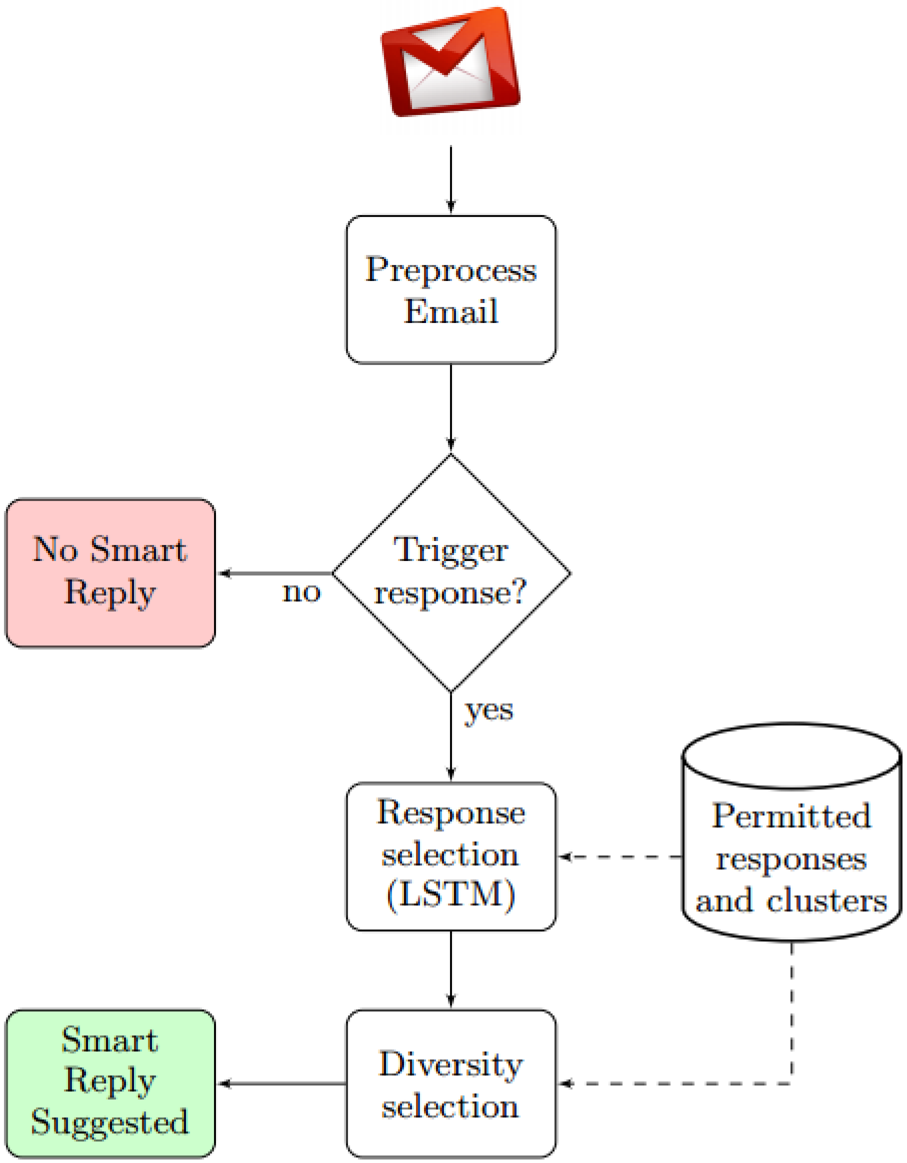
- initial data collection
- pre-/post-processing
- power to domain experts

ML/DL

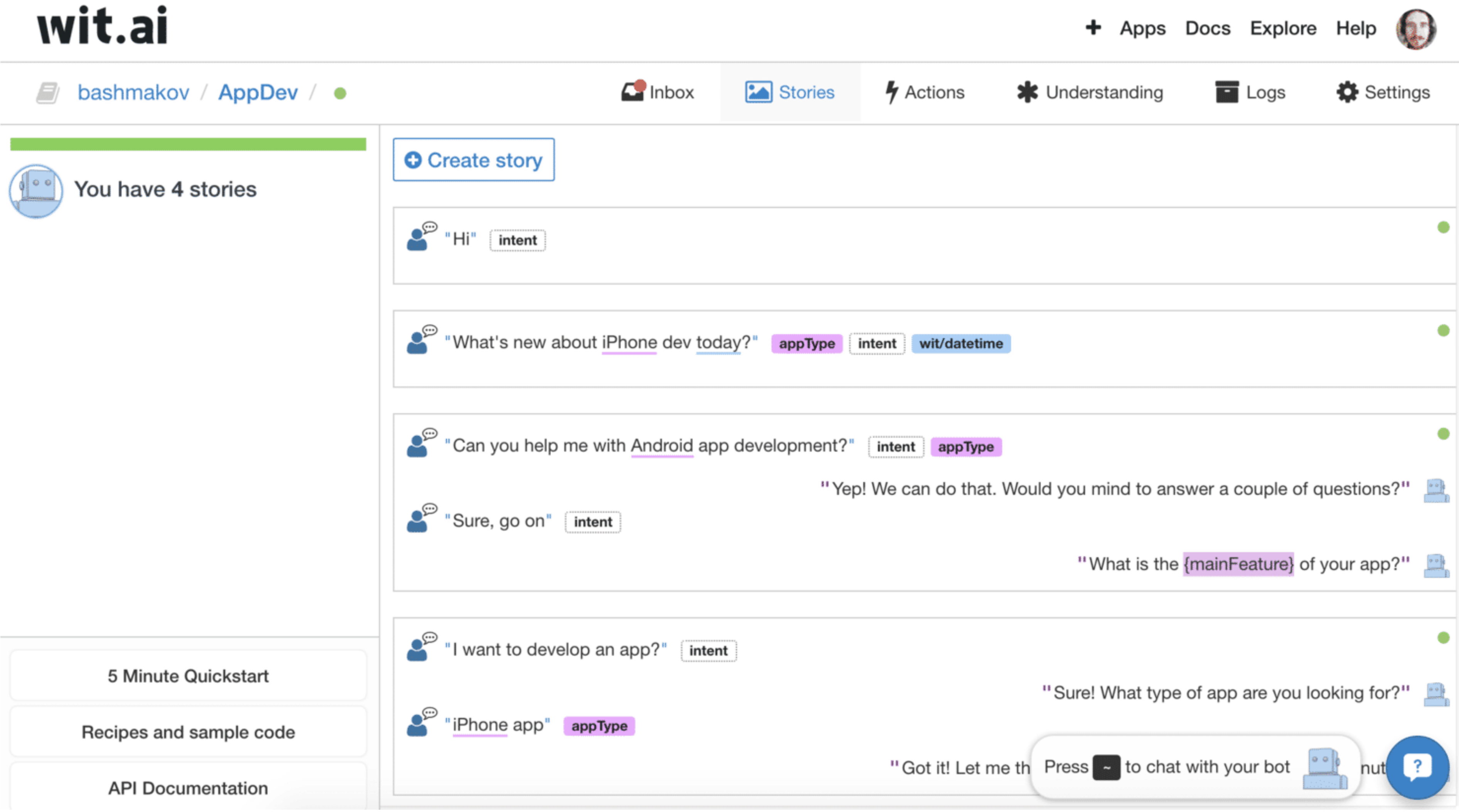
- often needs lots of data
- linguistic features + word vectors
- simple ngram-based models
- seq2seq neural network models

Ex: Gmail smart reply

- Data analysis
- Traditional NLP processing
- FNN
- LSTM
- Semi-supervised graph learning
- Rule-based post-processing
- Engineering



Ex: wit.ai



You have 4 stories

"Hi" intent

"What's new about iPhone dev today?" appType intent wit/datetime

"Can you help me with Android app development?" intent appType

"Yep! We can do that. Would you mind to answer a couple of questions?"

"Sure, go on" intent

"What is the {mainFeature} of your app?"

"I want to develop an app?" intent

"Sure! What type of app are you looking for?"

"iPhone app" appType

"Got it! Let me th Press ~ to chat with your bot nut

What's next

Some predictions:

- Bots go bust
- Deep learning goes commodity
- AI is cleantech 2.0 for VCs
- MLaaS dies a second death
- Full stack vertical AI startups actually work

http://www.bradfordcross.com/blog/2017/3/3/five-ai-startup-predictions-for-2017

What should work:

- NLP automation
- NLP + IR
- ???