Applications of NLP Approaches For Georgian Language

Anzor Gozalishvili

22.11.2023

About Me

- Education
- Work Experience
- Research

Education


UNIVERSITAS TPHILISENSIS
1918


VNIVERSITAT POLITÈCNICA VALÈNCIA
UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA


Erasmus+


coursera


MITOPENCOURSEWARE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY


fast.ai


Stanford University
S
NLP
Natural Language Processing


DeepLearning.AI


Bloomberg
ML EDU

Work Experience

MaxinAI

- Machine Learning Engineer
- School of AI Lecturer

Toptal®

Pfizer

- Machine Learning Engineer (NLP)

QUANTORI

Roche

Genentech

A Member of the Roche Group

- Machine Learning Engineer
- Bioinformatics
- MLOps

Research

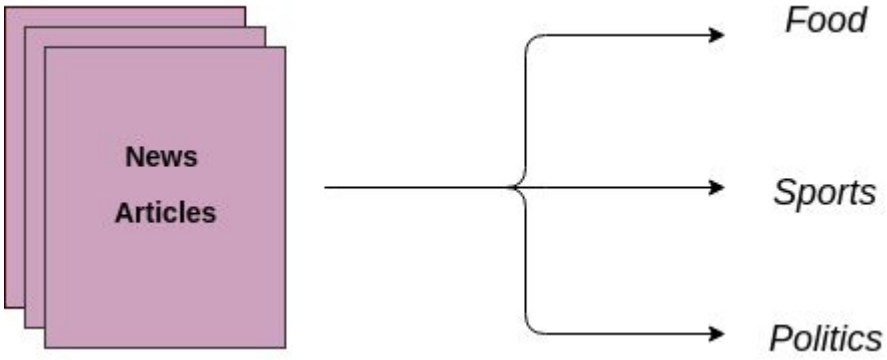
- A new approach to broaden the range of eye colour identifiable by IrisPlex in DNA phenotyping
- Georgian OCR Benchmark Dataset and Evaluation Methods
- Syntactic Annotation of Georgian in the UD Schemes

NLP Tasks

- Supervised Learning
- Unsupervised Learning
- Data vs Model

Supervised Learning

Classification



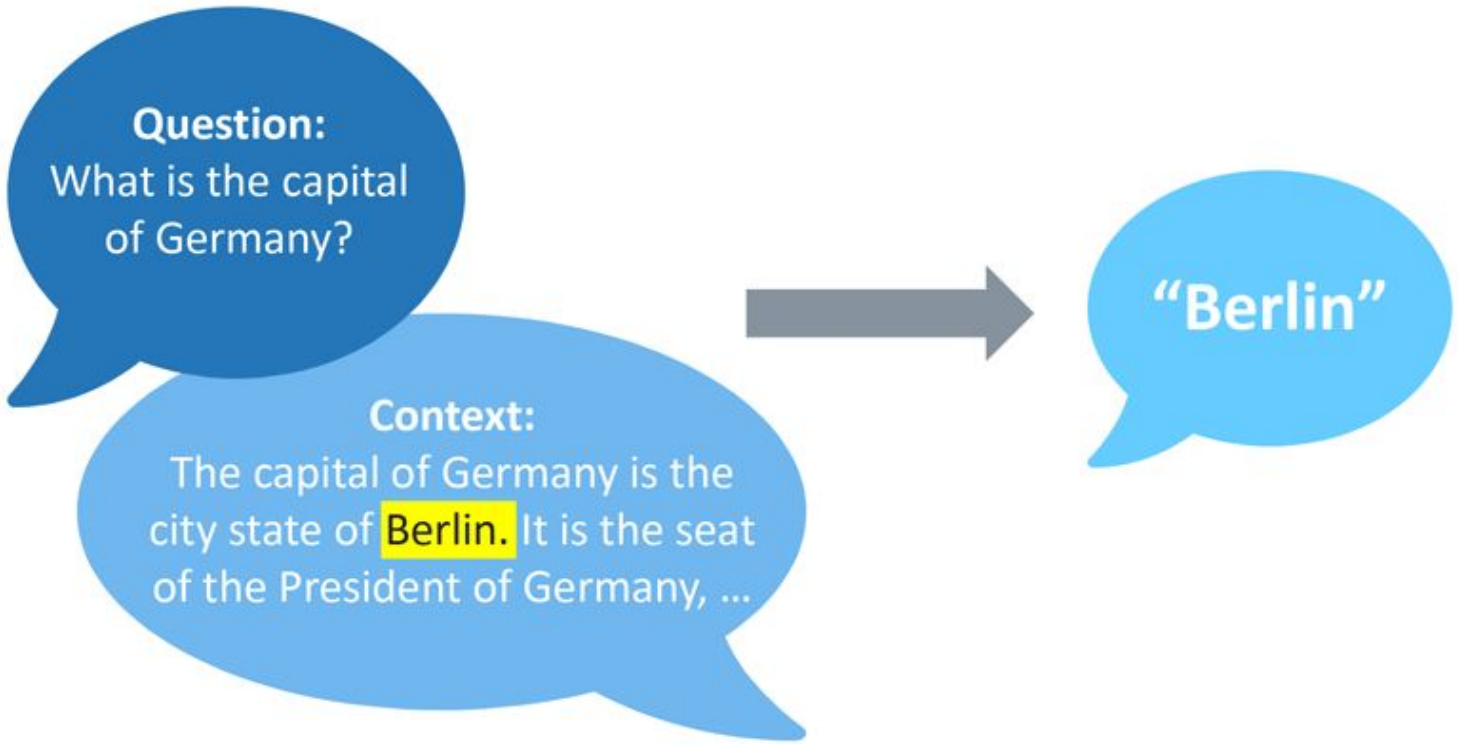
Sentiment Analysis (Aspect based)

"On arrival staff could not off been more helpful , Food was fantastic, the place was spotless. The only let down was the bed was like trying to sleep on a concrete floor it ruined our stay sorry."

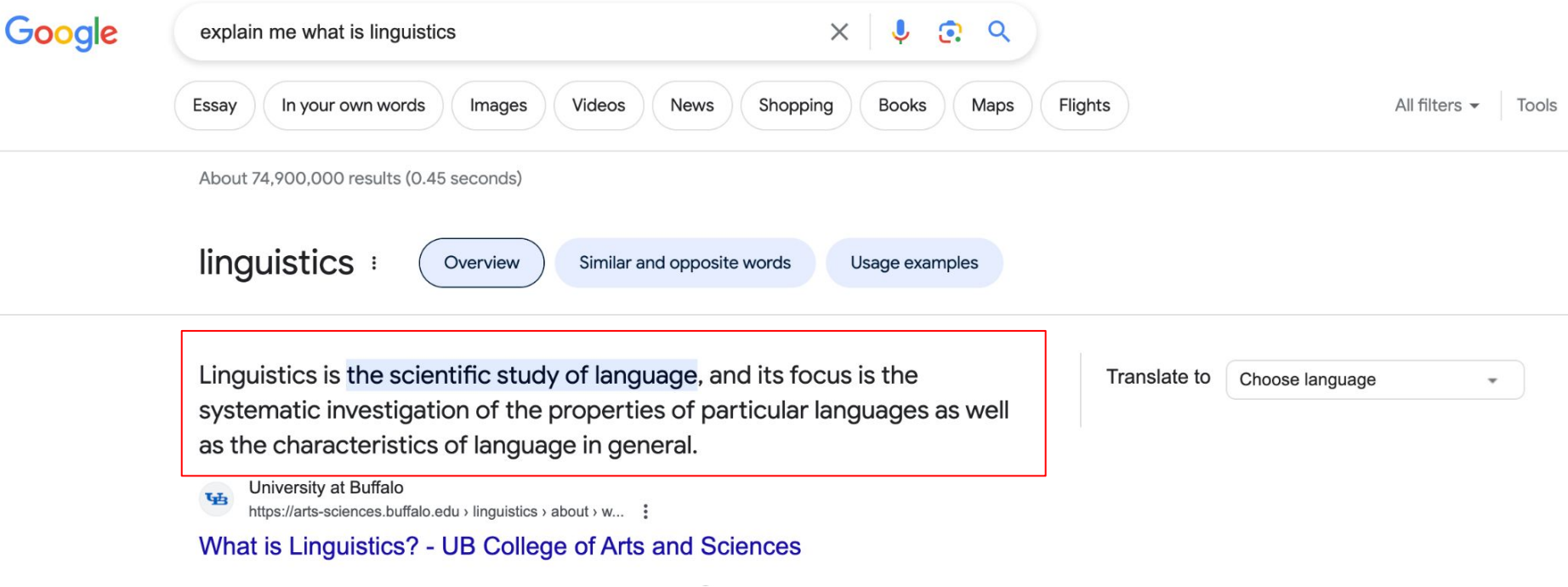
Aspect	Polarity
Staff	Positive
Food	Positive
Cleanliness	Positive
Beds	Negative

Named Entity Recognition (NER)

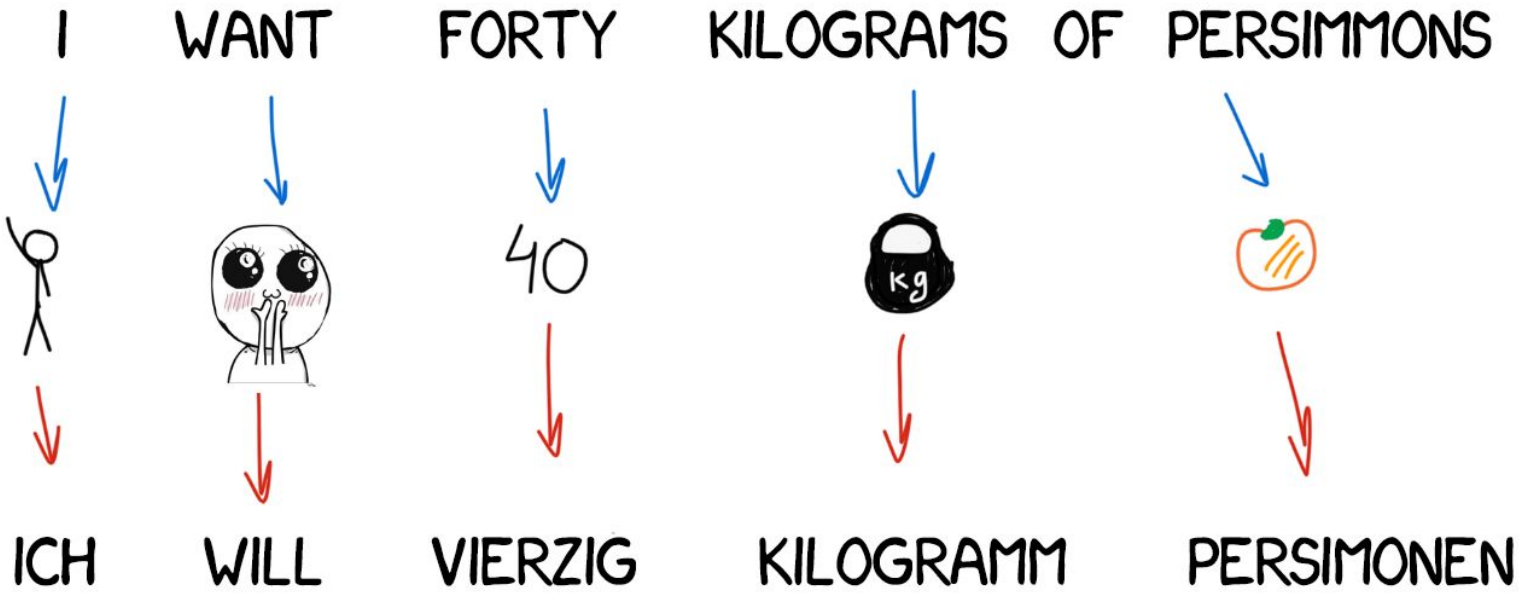
QA - Question Answering



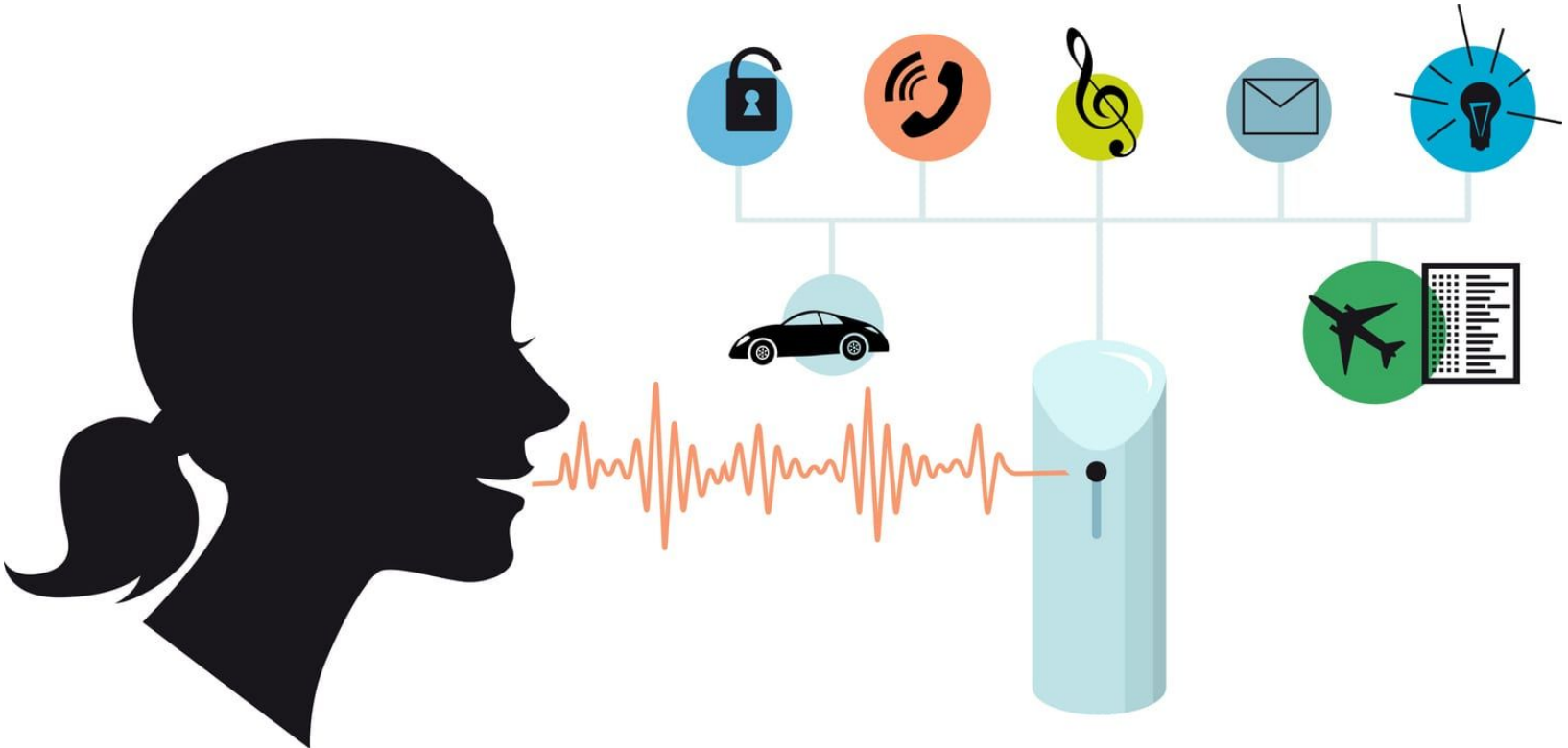
Summarization



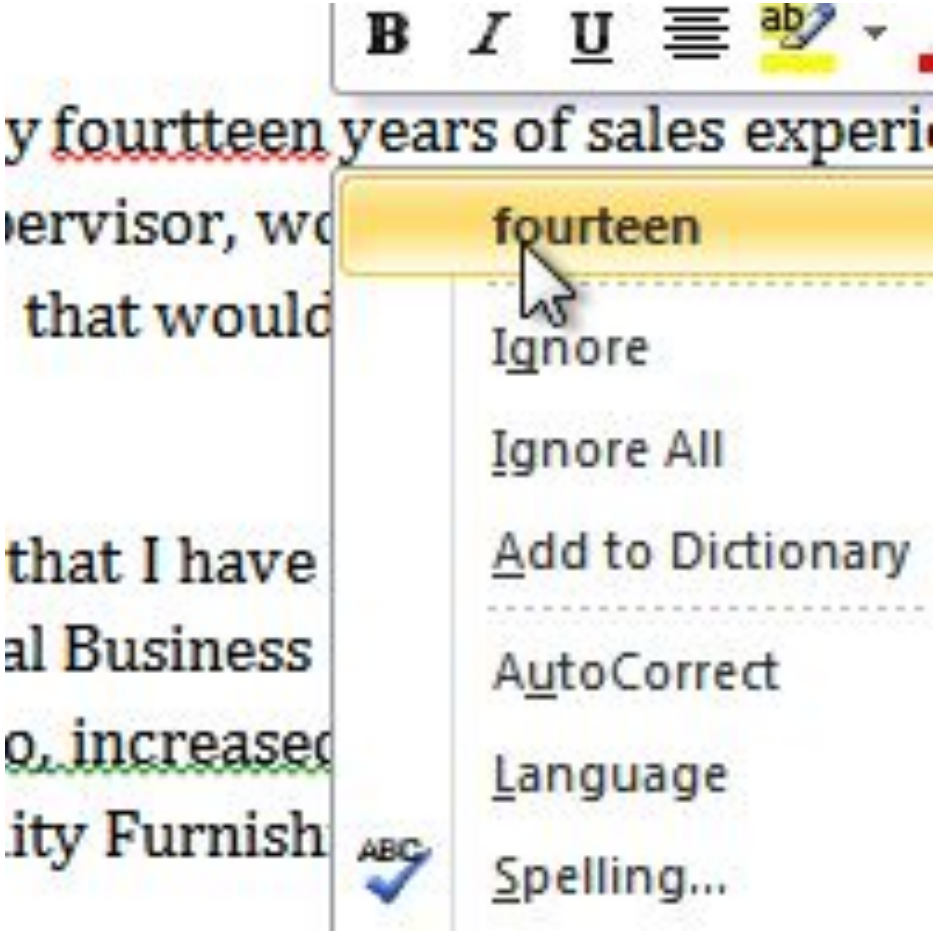
Machine Translation



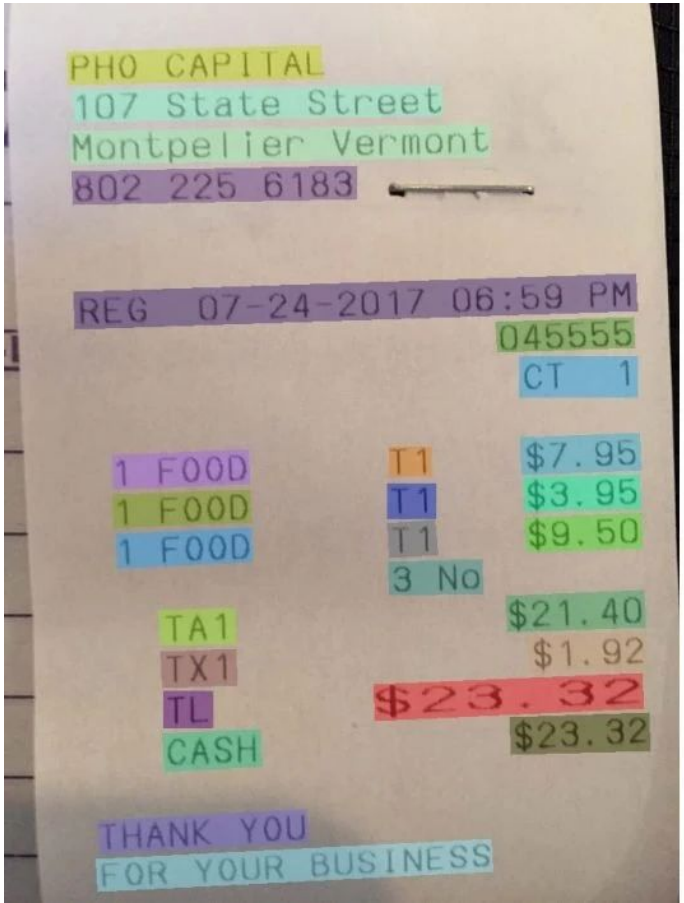
Speech Recognition



Spelling Correction



OCR - Optical Character Recognition (Post OCR Text Correction)



PHO CAPITAL
107 State Street
Montpe l ier Vermont
802 2256183

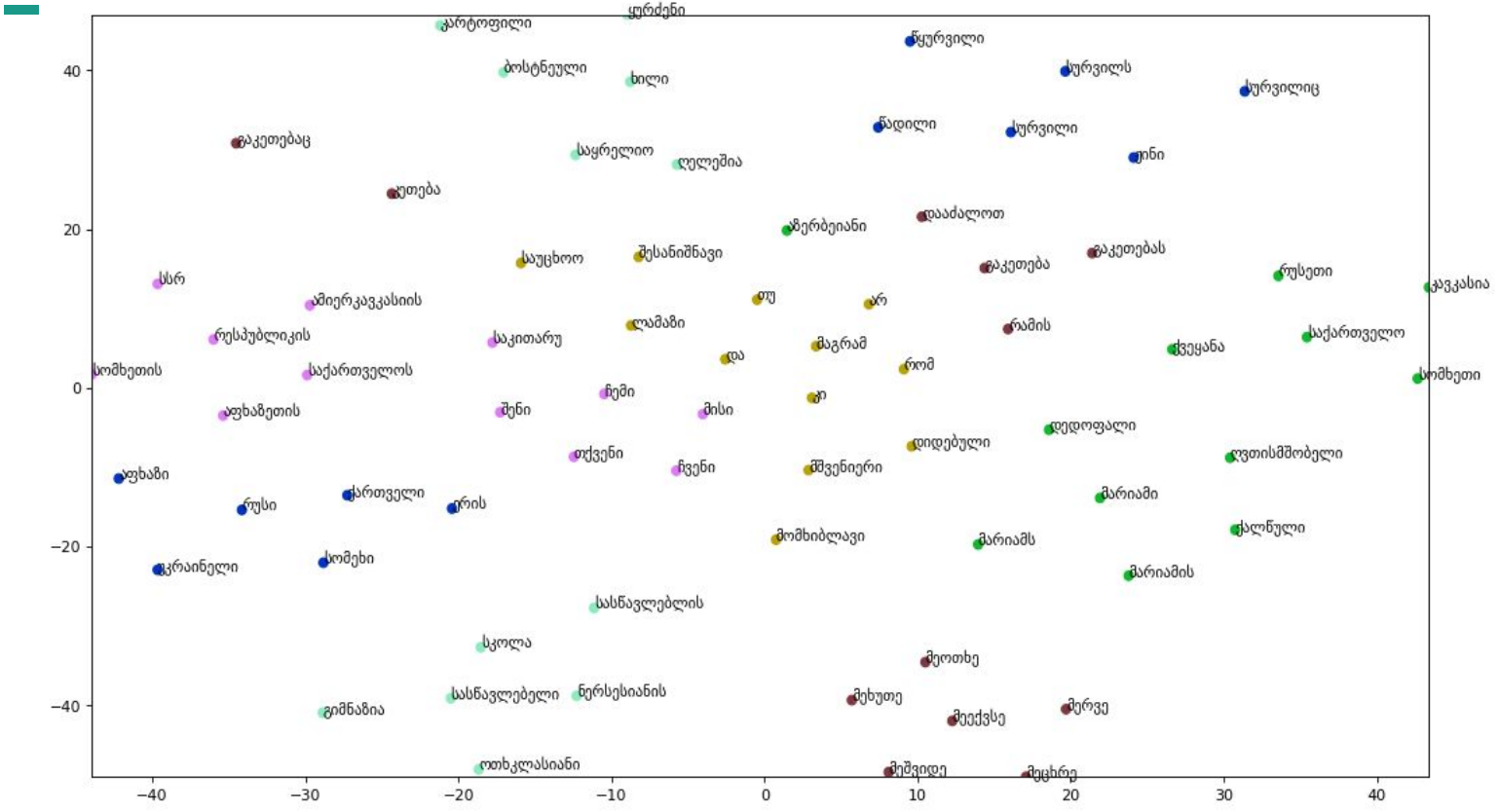
REG 07-24-201706:59 PM
045555
CT1

1 FOOD
1 FOOD
1 FOOD
T1
T1
T1
3 No
$7.95
$3.95
$9.50
$21.40
TA1
TX1
$1.92
TL
$23.32
CASH
$23.32

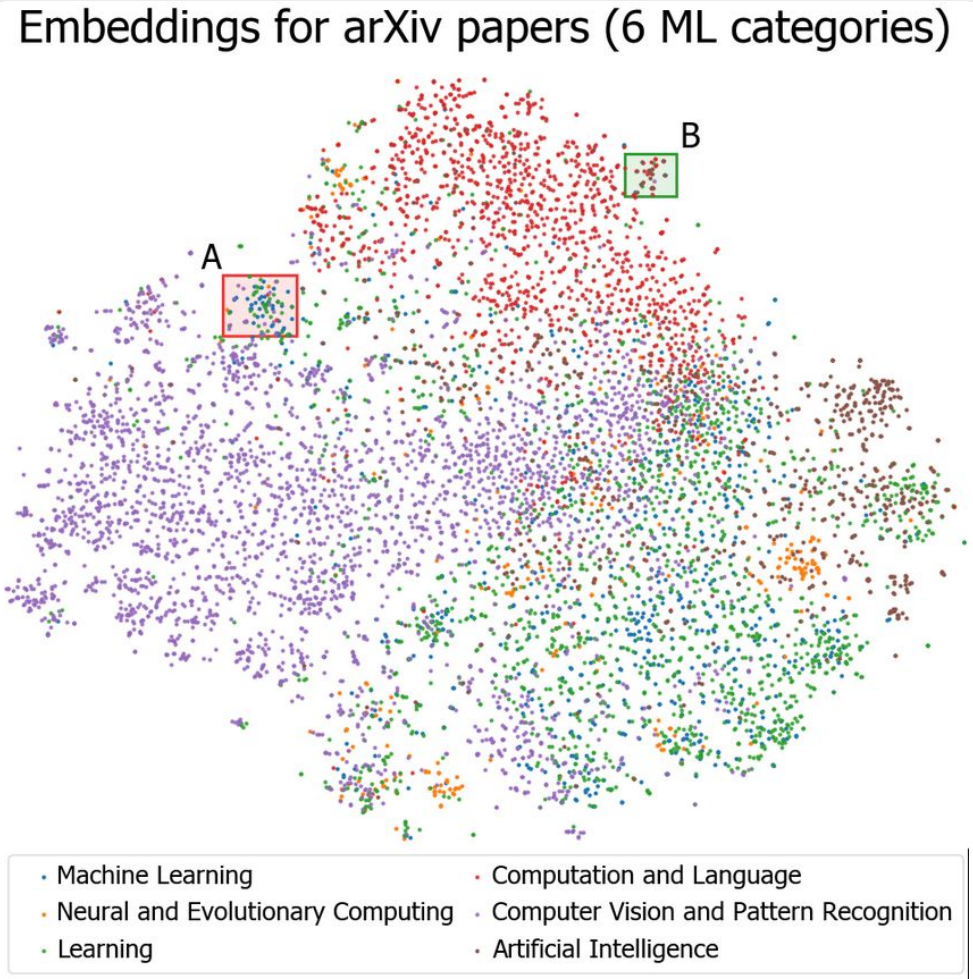
THANK YOU
FOR YOUR BUSINESS

Unsupervised Learning

Text Encoding - Word Embeddings



Clustering


Embeddings for arXiv papers (6 ML categories)
A
B
Machine Learning
Neural and Evolutionary Computing
Learning
Computation and Language
Computer Vision and Pattern Recognition
Artificial Intelligence

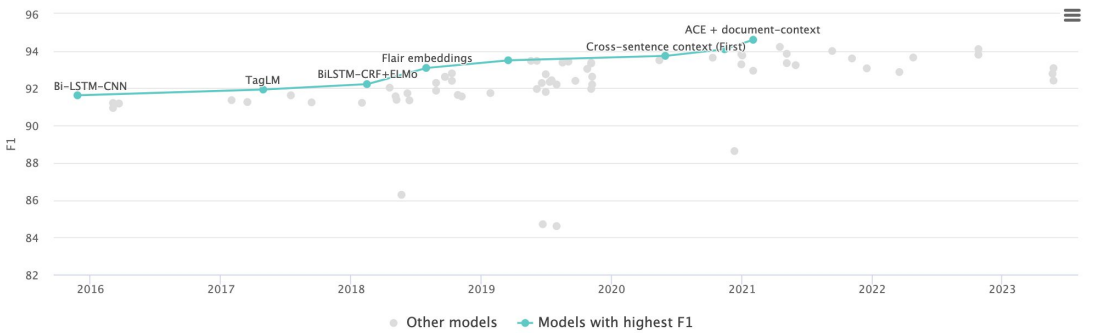
Benchmark vs Model

Benchmark Datasets vs Model Architectures

Named Entity Recognition (NER) on CoNLL 2003 (English)

Leaderboard | Dataset

View F1 by Date for All models



Filter: LSTM | Transformer | MRC-based | knowledge distillation | GCN | untagged

Edit Leaderboard

Rank	Model	F1	Extra Training Data	Paper	Code	Result	Year	Tags
1	**ACE + document-context**	94.6	×	Automated Concatenation of Embeddings for Structured Prediction			2021	LSTM, Transformer
2	**Co-regularized LUKE**	94.22	×	Learning from Noisy Labels for Entity-Centric Information Extraction			2021	knowledge distillation
3	**ASP+T5-3B**	94.1	×	Autoregressive Structured Prediction with Language Models			2022	
4	**FLERT XLM-R**	94.09	✓	FLERT: Document-Level Features for Named Entity Recognition			2020	Transformer

GLUE

The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems. GLUE consists of:

- A benchmark of nine sentence- or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty,
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language, and
- A public leaderboard for tracking performance on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set.

The format of the GLUE benchmark is model-agnostic, so any system capable of processing sentence and sentence pairs and producing corresponding predictions is eligible to participate. The benchmark tasks are selected so as to favor models that share information across tasks using parameter sharing or other transfer learning techniques. The ultimate goal of GLUE is to drive research in the development of general and robust natural language understanding systems.

SuperGLUE

In the last year, new models and methods for pretraining and transfer learning have driven striking performance improvements across a range of language understanding tasks. The GLUE benchmark, introduced one year ago, offered a single-number metric that summarizes progress on a diverse set of such tasks, but performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research.

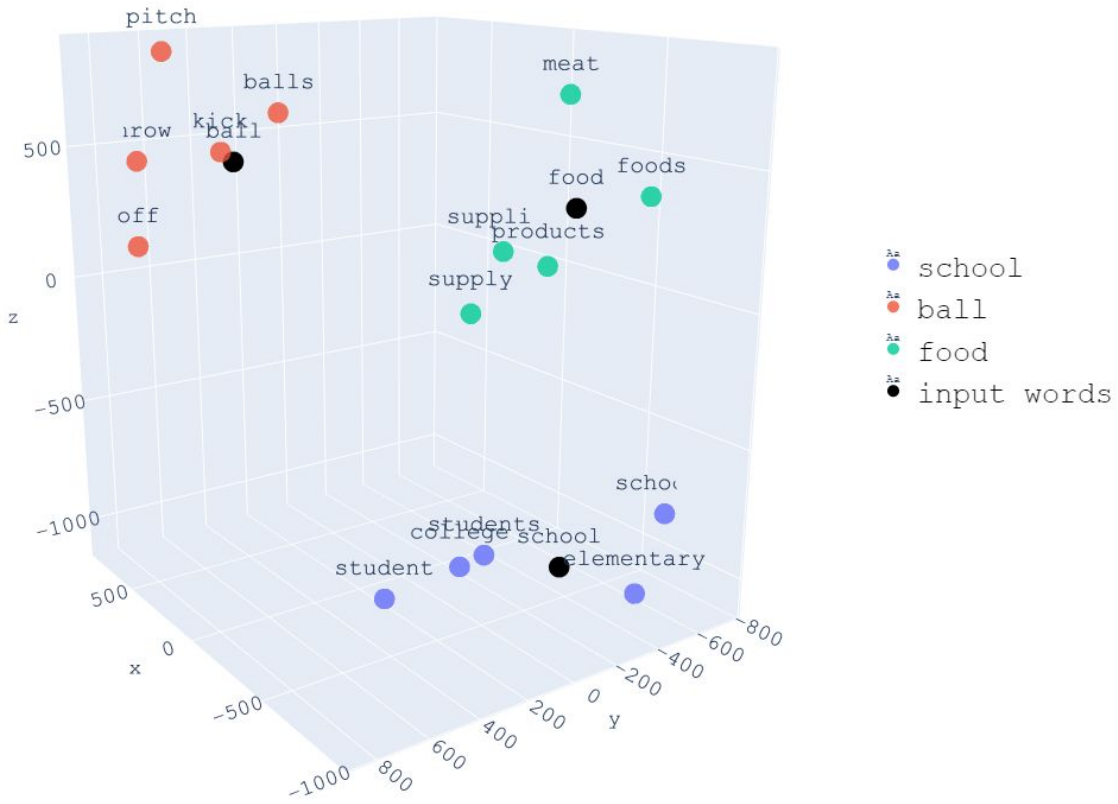
We take into account the lessons learnt from original GLUE benchmark and present SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard.

NLP Quick Deep Dive

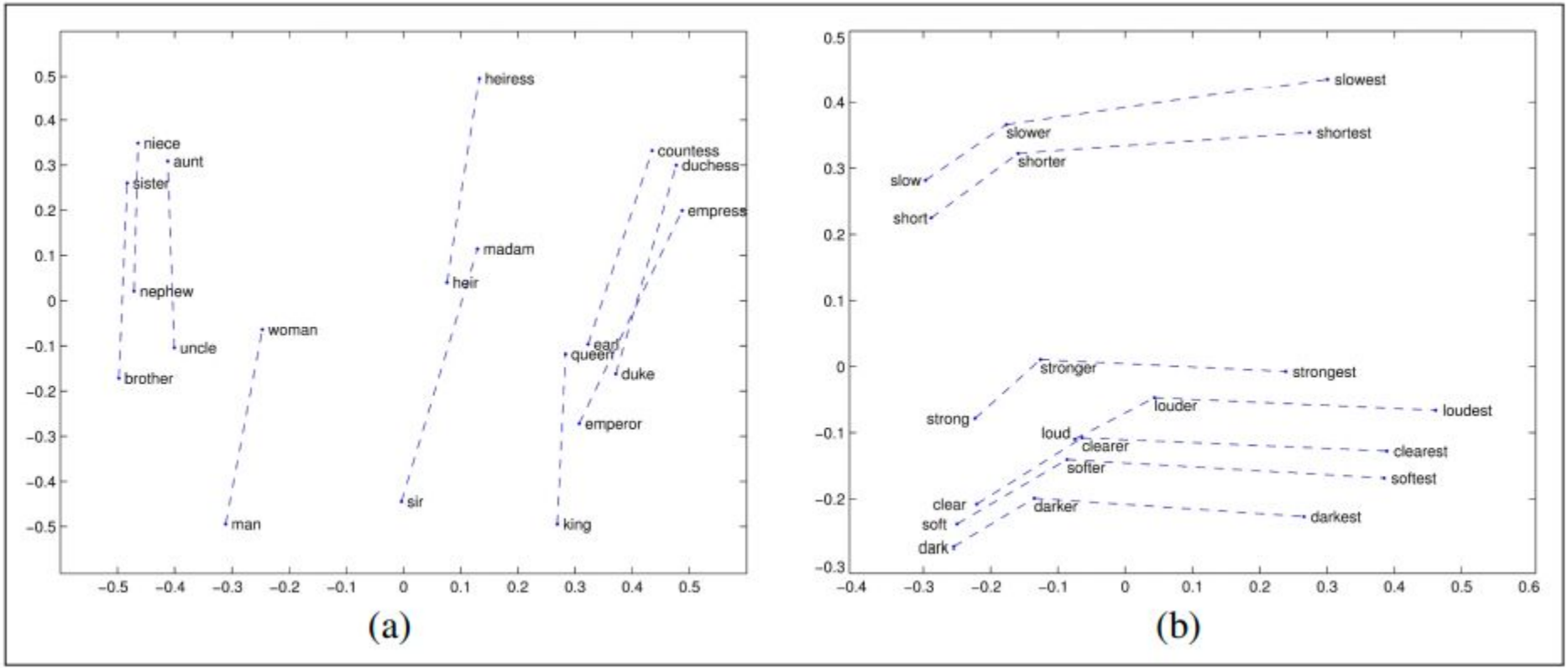
- Embeddings
- Treebanks

Embeddings Intuition

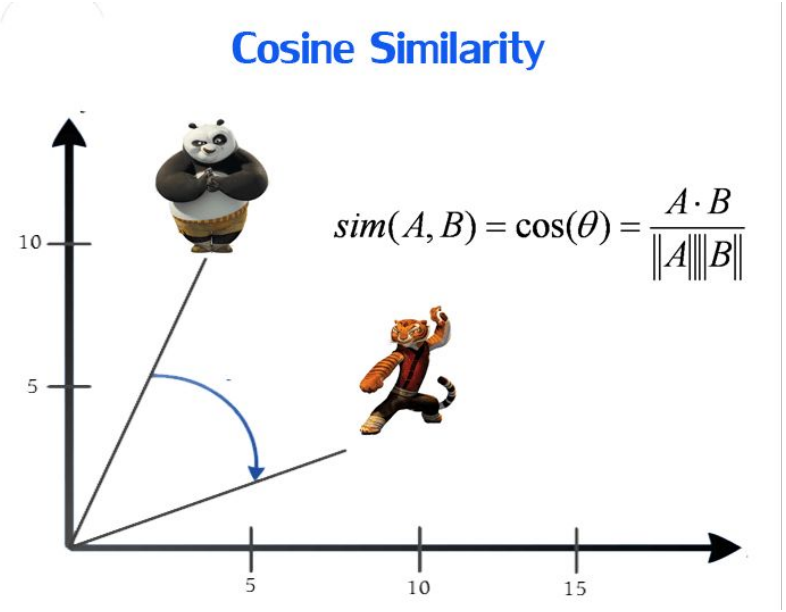
Word Embeddings in 3D Space Encoding Semantic Similarity



Simple Math on Word Vectors



Similarity Measure



Simple Math on Word Vectors - (My Bachelor Thesis 2014)

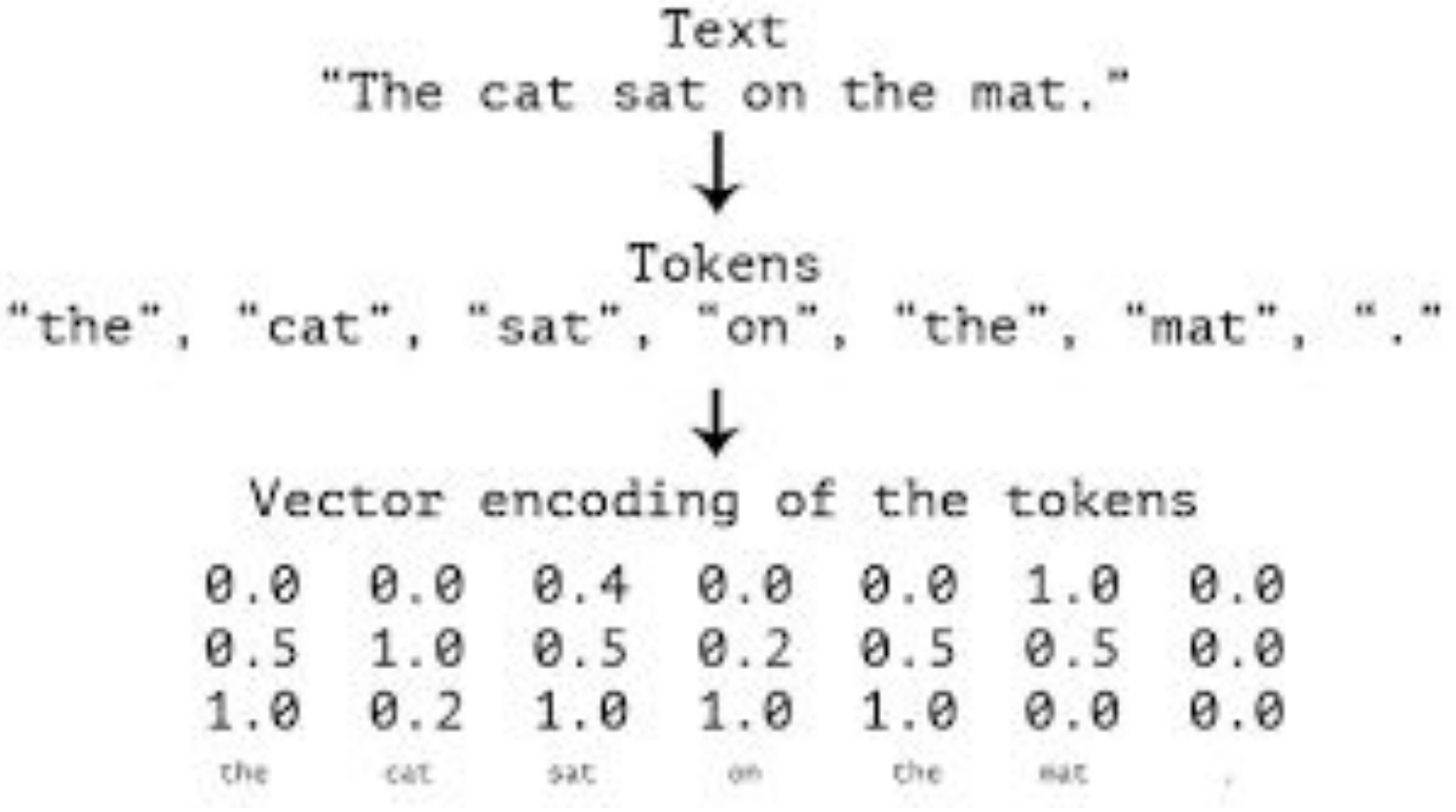
In [3]:
```
import os
import gensim
```

In [7]:
```
# Load word2vec model
model = gensim.models.Word2Vec.load('../models/word2vec_skipgram.w2v', mmap='r')
```

In [8]:
```
print('მეფე - კაცი + ქალი:')
print('')
for word, sim in model.most_similar(positive=['ქალი', 'მეფე'], negative=['კაცი']):
    print('\"%s\"\t- similarity: %g' % (word, sim))
print('')

print('Similarity between კაცი and ქალი:')
print(model.similarity('ქალი', 'კაცი'))
```

```
მეფე - კაცი + ქალი:

/home/anz2/anaconda3/envs/sentinel/lib/python3.6/site-packages/ipykernel_launcher.py:3: De
deprecated `most_similar` (Method will be removed in 4.0.0, use self.wv.most_similar() ins
  This is separate from the ipykernel package so we can avoid doing imports until
"დედოფალი"      - similarity: 0.624178
"უღელაი"        - similarity: 0.51129
"ასურთ" - similarity: 0.509556
"ბეატრისა"      - similarity: 0.505491
"გლიუგსბურგთა"  - similarity: 0.50291
"ქუღუ"  - similarity: 0.502755
"ბეღიბნი"       - similarity: 0.5027
"ფანაკერტელისა" - similarity: 0.502124
"აშურნადინშუმი" - similarity: 0.501499
"სითიხათუნ"     - similarity: 0.499499

Similarity between კაცი and ქალი:
0.6740217978621085
```

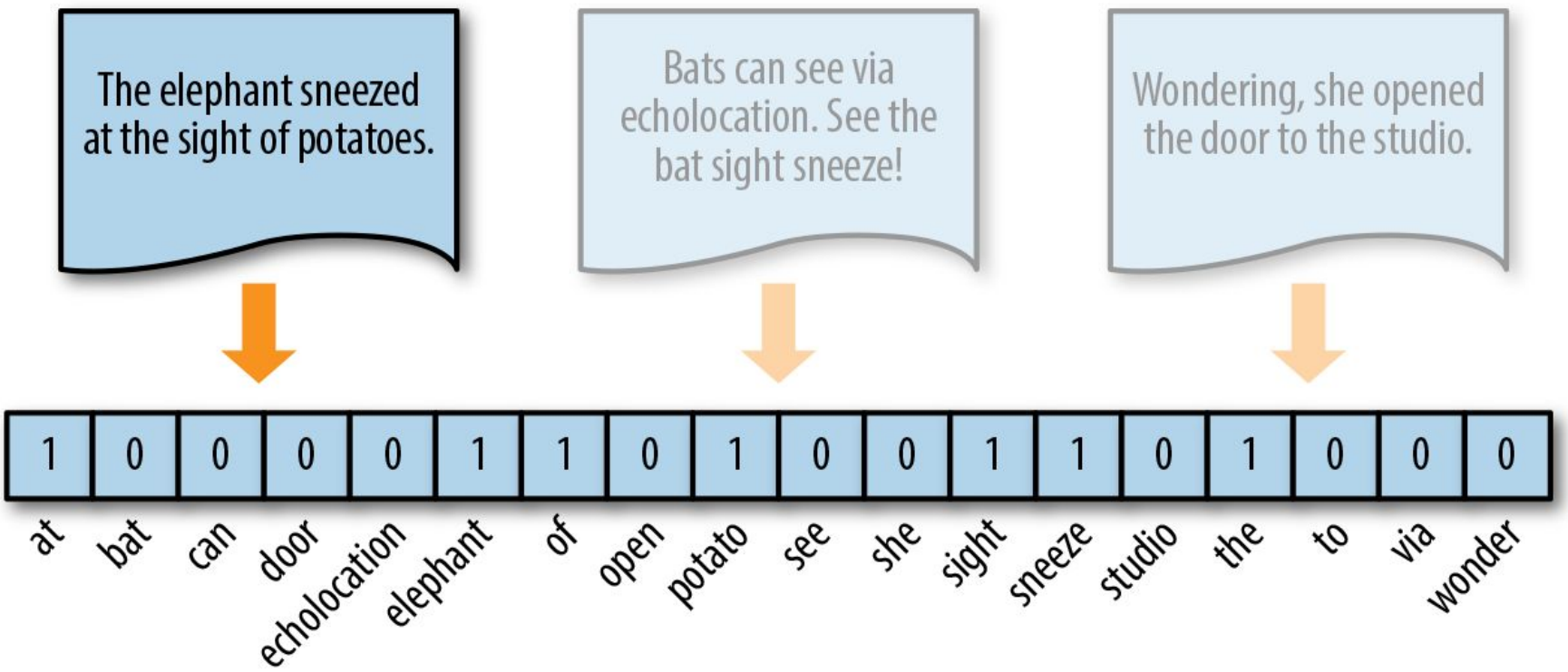
Text -> Number (n dimensional Vector)



https://projector.tensorflow.org/

Embedding Types

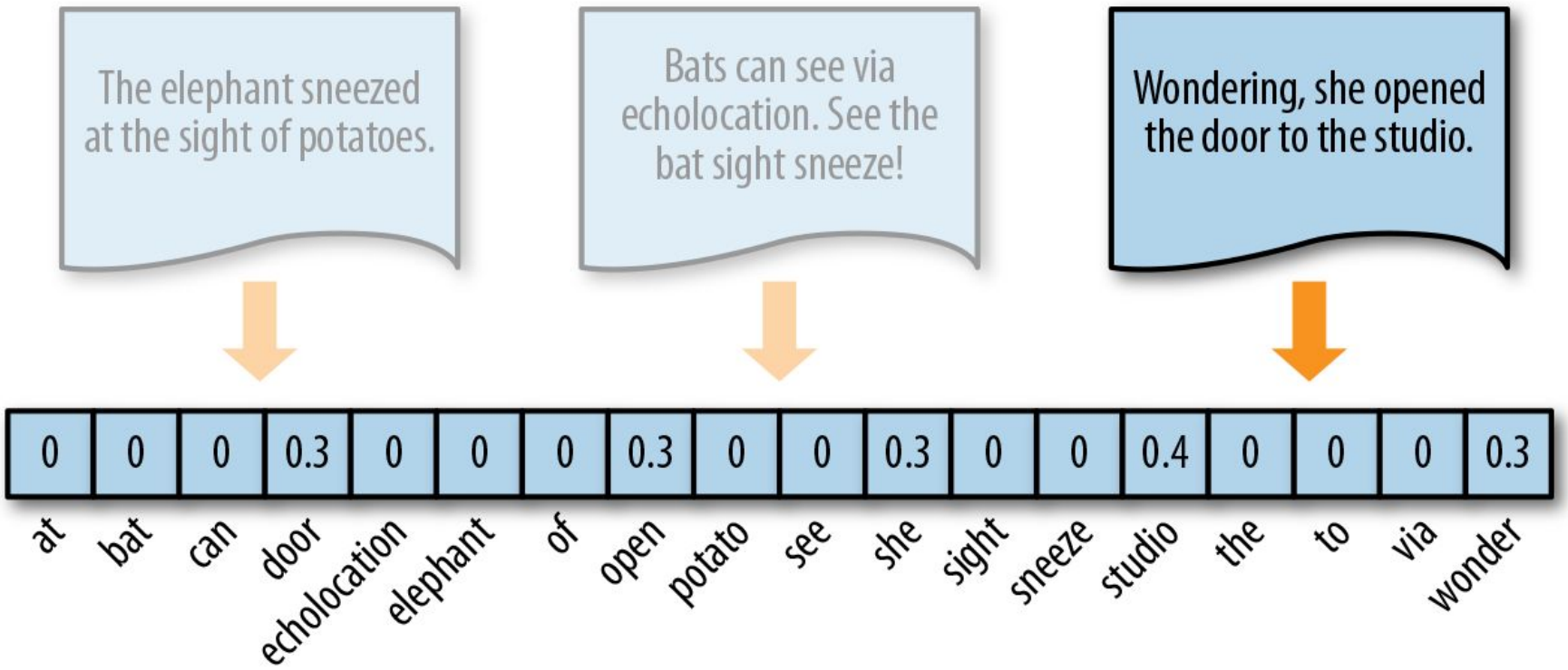
One-Hot Encoding



One-Hot Encoding

- Simple to build
- Very Naive
- Missing word frequency
- Missing word importance
- Categorical

TF-IDF = Term Frequency - Inverse Document Frequency (Count Based)



TF-IDF = Term Frequency - Inverse Document Frequency (Count Based)

Count Vectorizer

	blue	bright	sky	sun
Doc1	1	0	1	0
Doc2	0	1	0	1

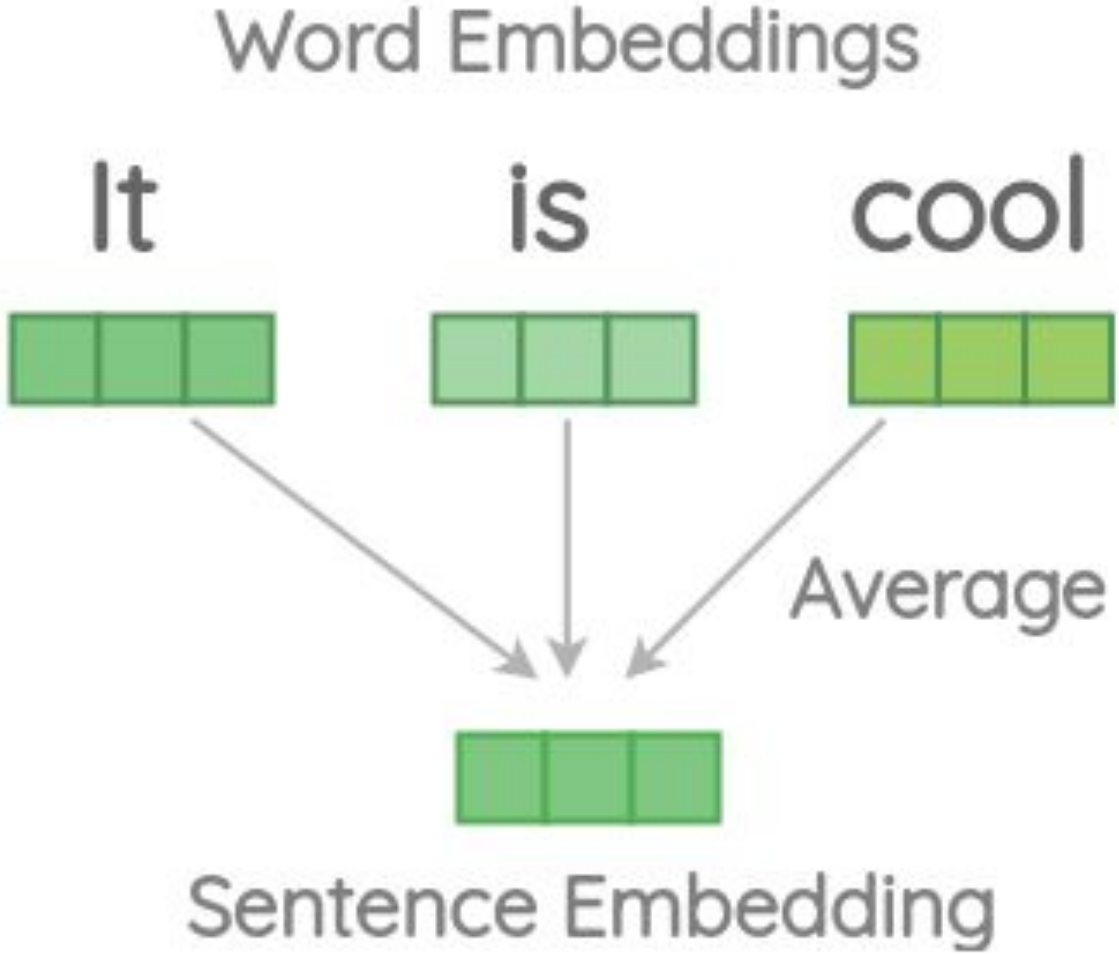
TD-IDF Vectorizer

	blue	bright	sky	sun
Doc1	0.707107	0.000000	0.707107	0.000000
Doc2	0.000000	0.707107	0.000000	0.707107

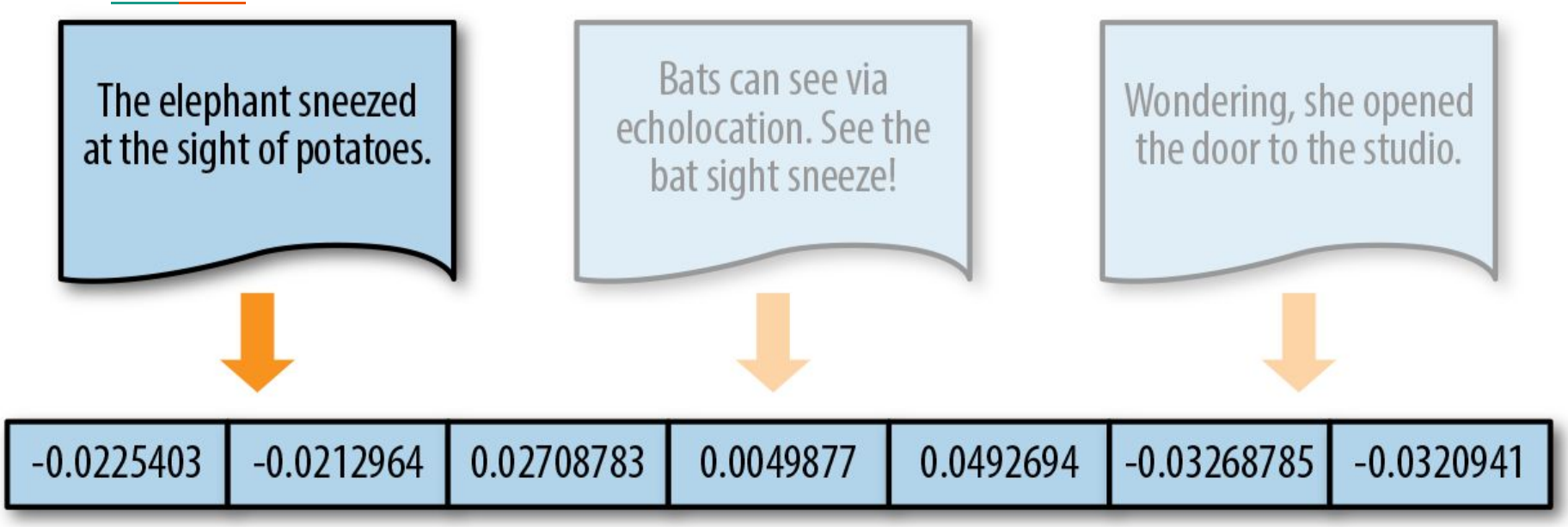
TF-IDF = Term Frequency - Inverse Document Frequency (Count Based)

- Simple to build
- Good Baseline
- Still Very Naive
- No word embeddings (document embedding only)
- Hard to understand word relations beyond n-grams

From Word Embeddings to Sentence Embedding

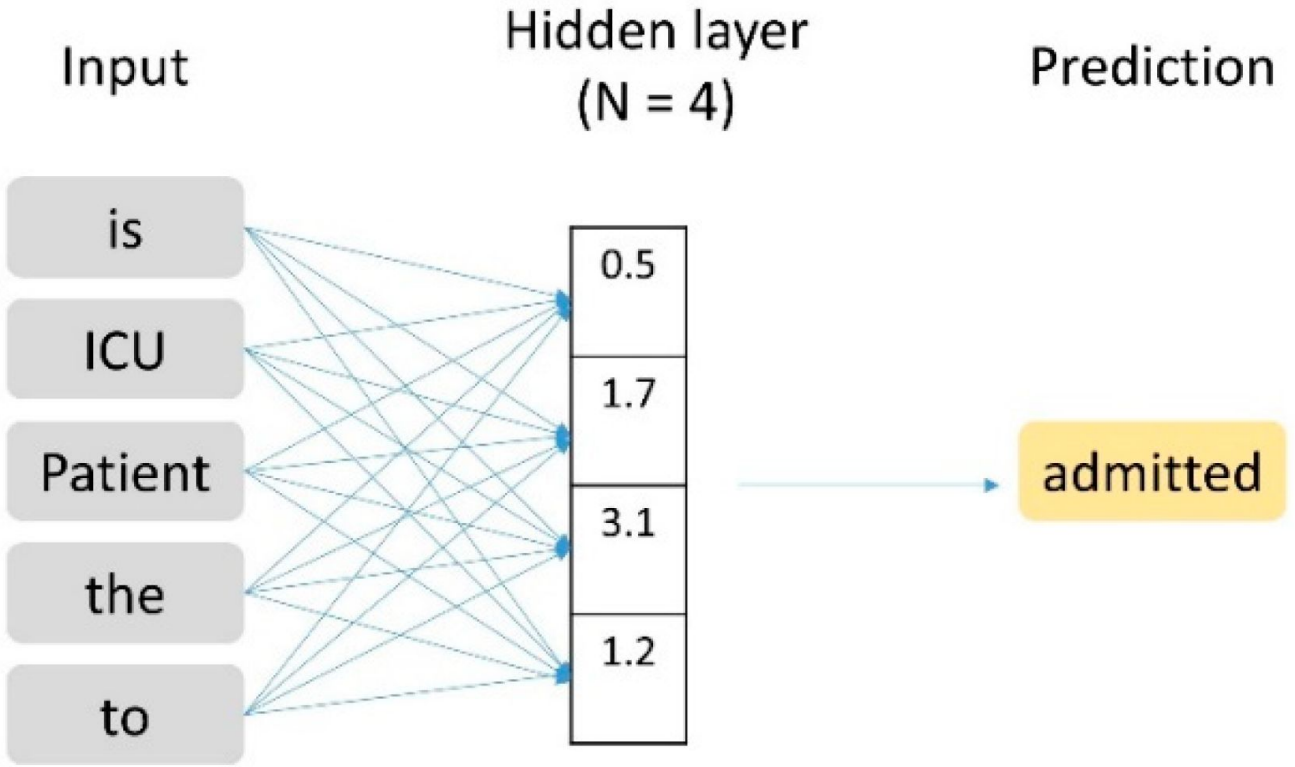


Word2Vec Sentence Embedding

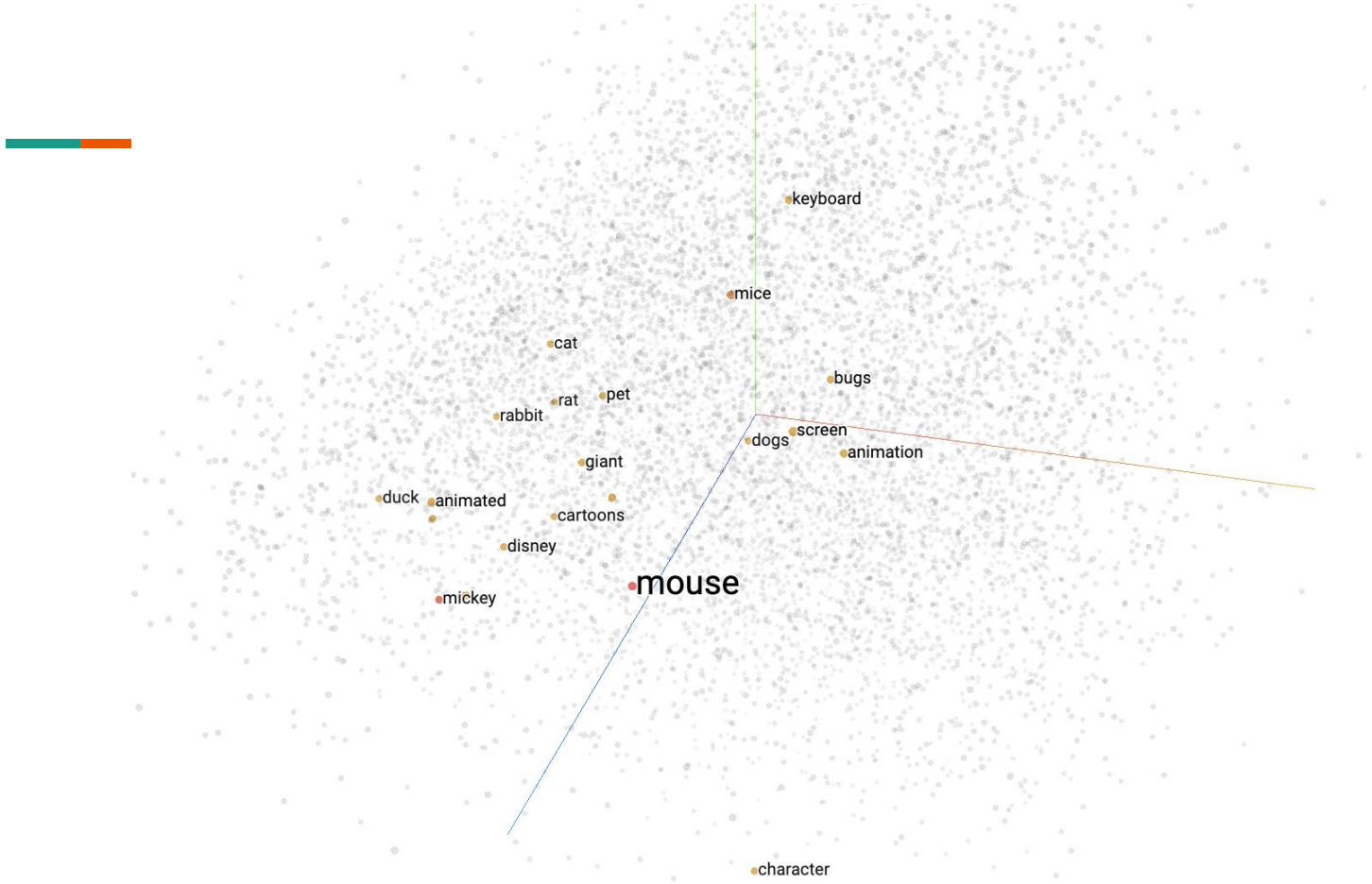


Word2Vec Learning Principle (Similar Words Appear In Same Context)

Sentence: Patient is admitted to the ICU



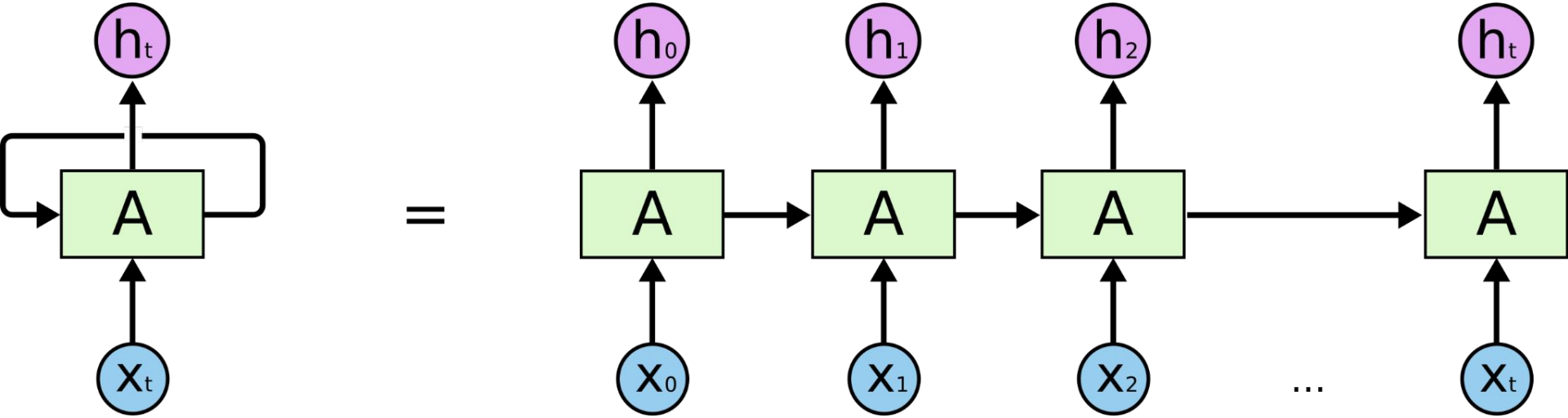
Word2Vec (Non-Contextual Embeddings)



Word2Vec Non-Contextual Embeddings

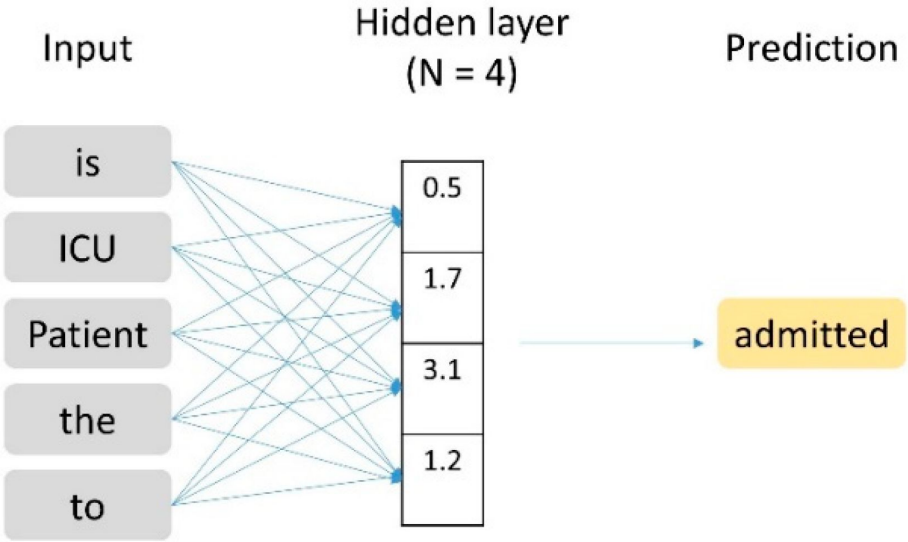
- Learns Relatively Quickly on User Laptop
- Understands the (average) sense of word
- Much Better Baseline than TF-IDF
- Can’t fully understand the sense of word relations
- Can’t understand word meanings in different context

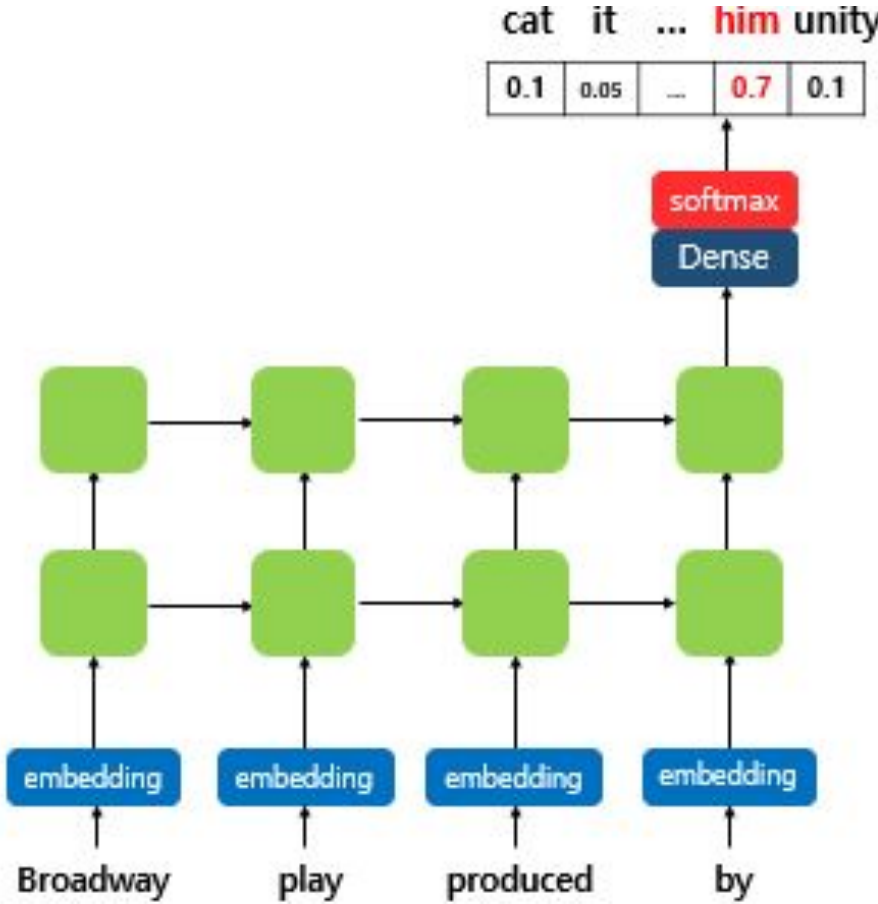
RNNs (better encodings by modeling sequence relations)



RNN VS W2V (Predict word by neighbours VS Predict next word)

Sentence: Patient is admitted to the ICU



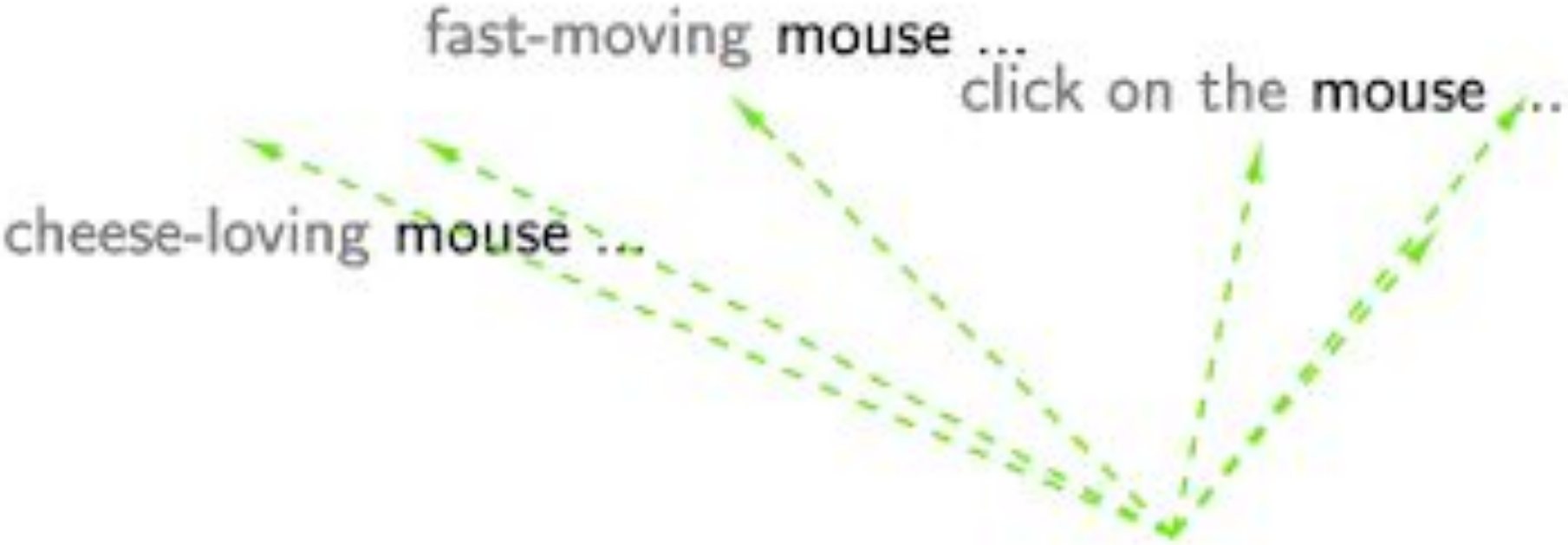


RNNs (better encodings by modeling sequence relations)

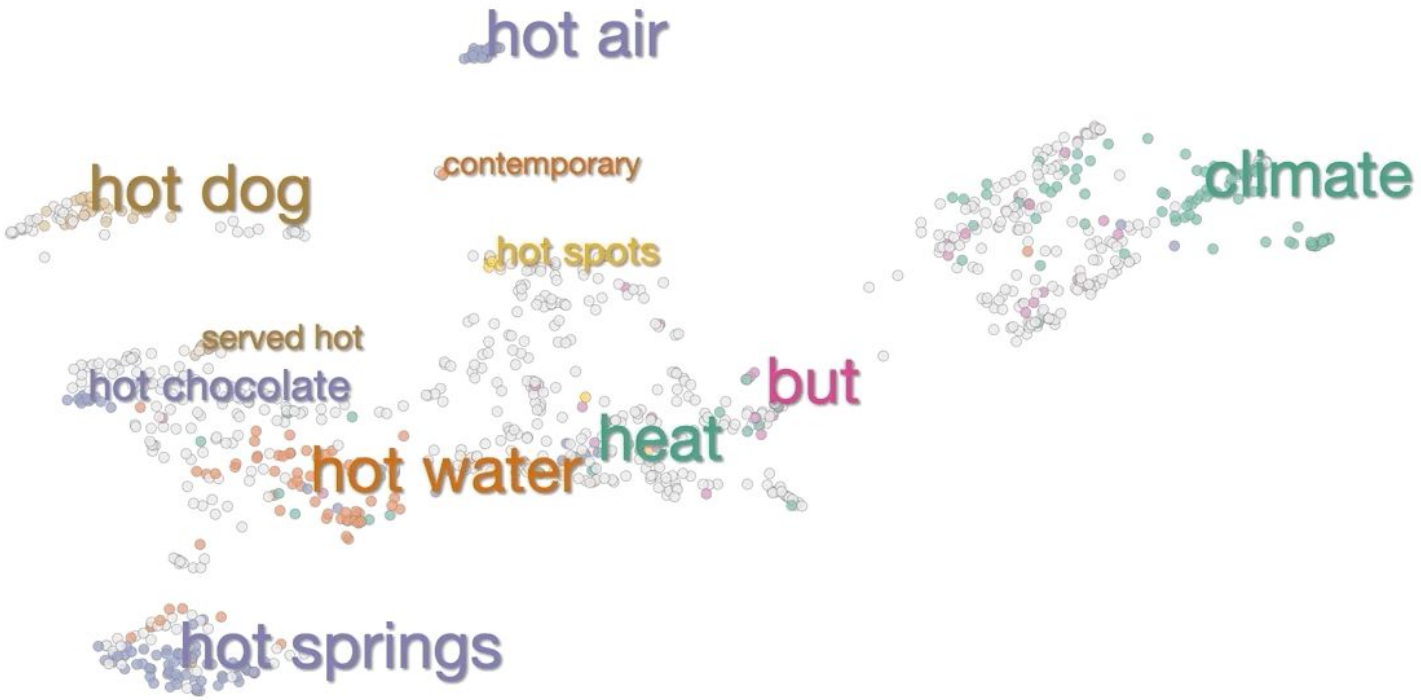
- Understands the sense of word by learning new word embeddings (not only using word2vec)
- Understands the meaning of sentence by analyzing the sequence of words and some relation between them
- More than a Baseline (still might be used in production in some applications)
- Requires GPU for training
- Trains Slow
- Not very optimal in terms of compute requirements

Deep Contextualized Word Representations (ELMo) by AllenAI **https://arxiv.org/abs/1802.05365**

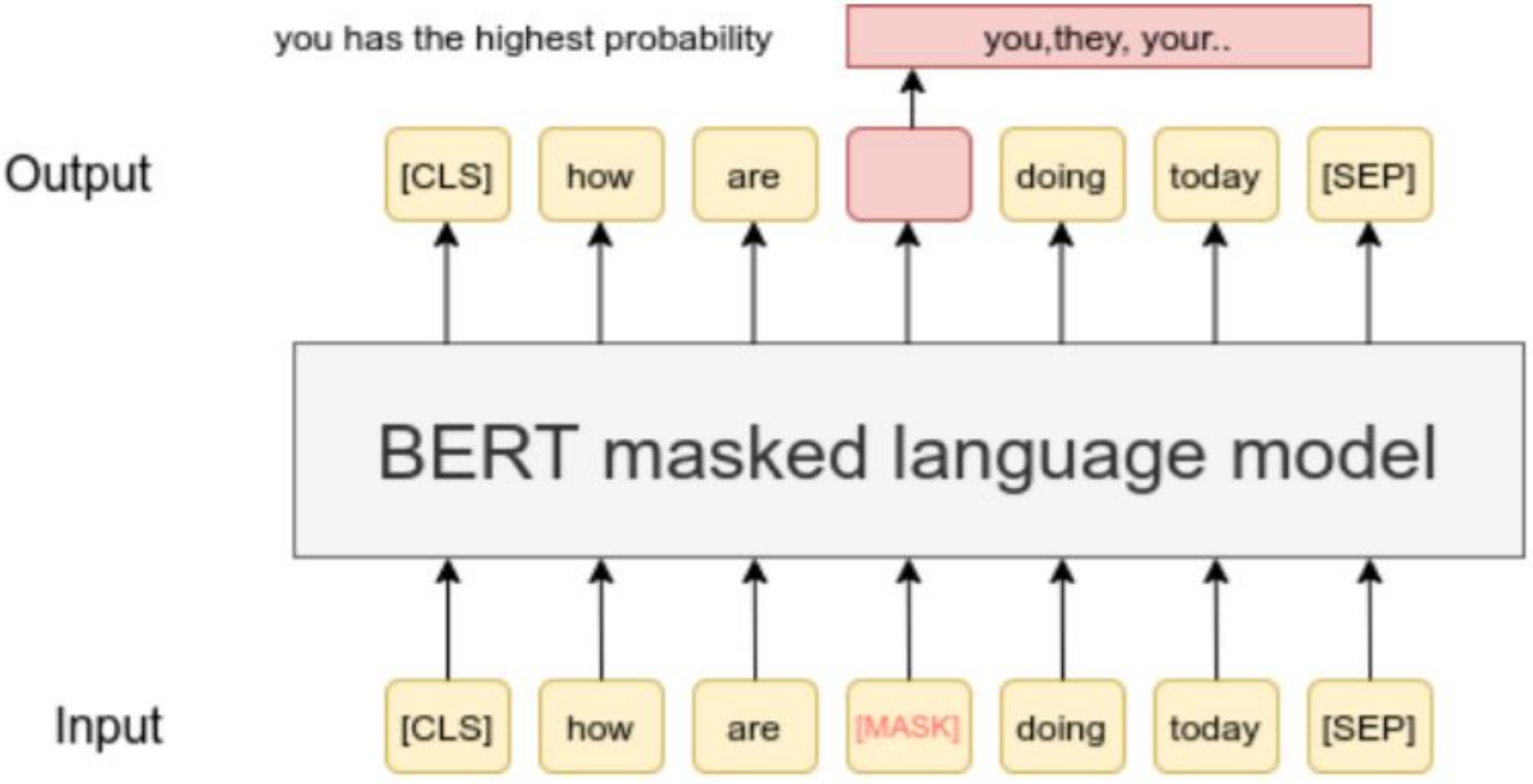
Transformers (contextual embeddings)



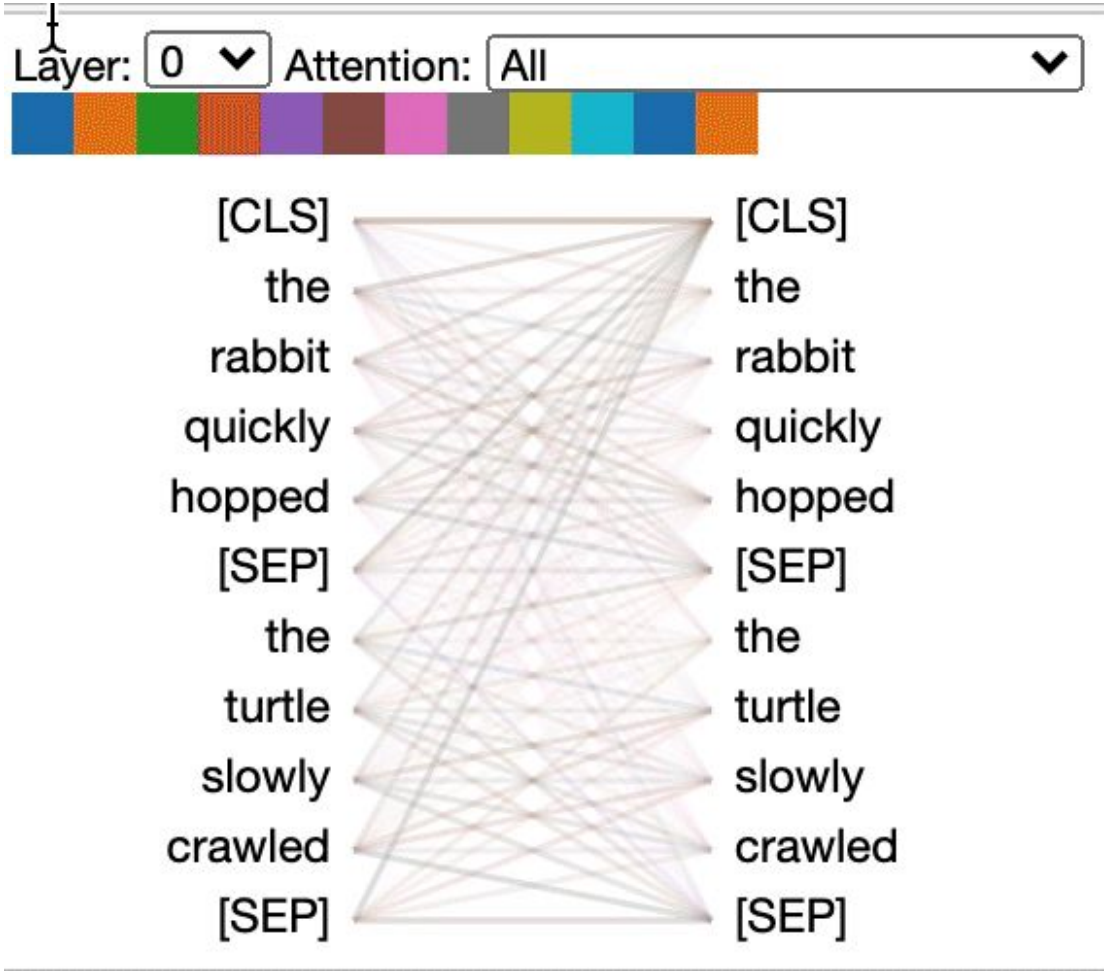
Transformers (contextual embeddings)



Transformers (masked language modeling)



Transformers (Attention Mechanism)

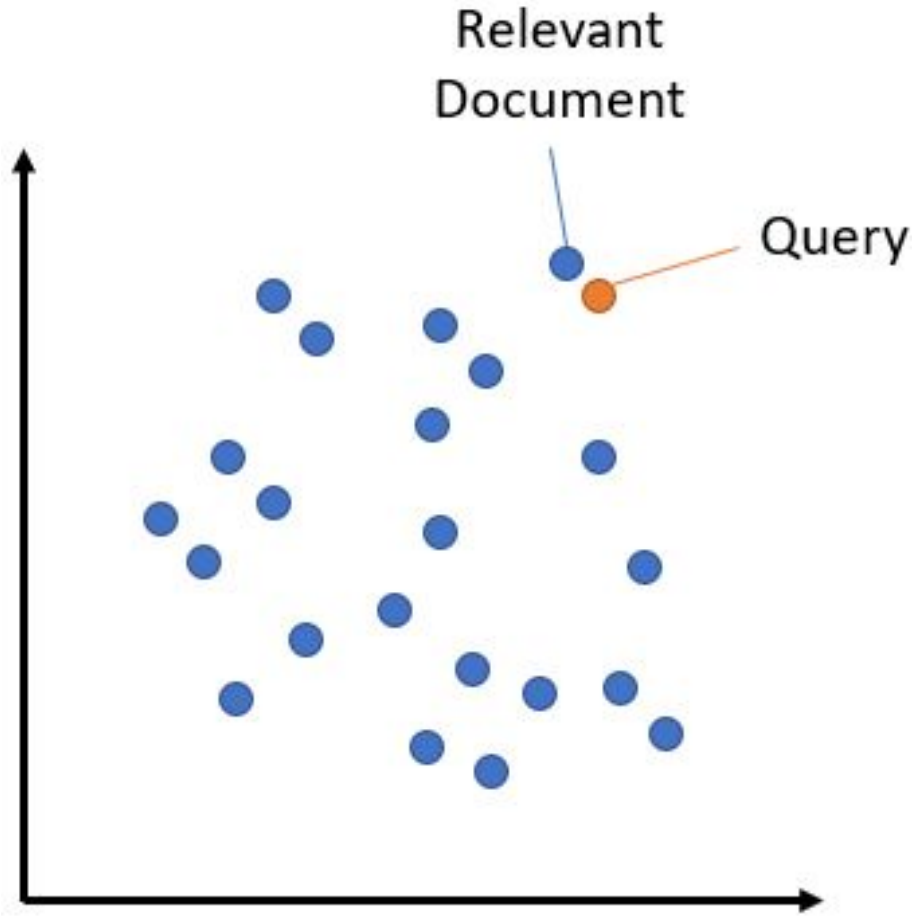


Transformers

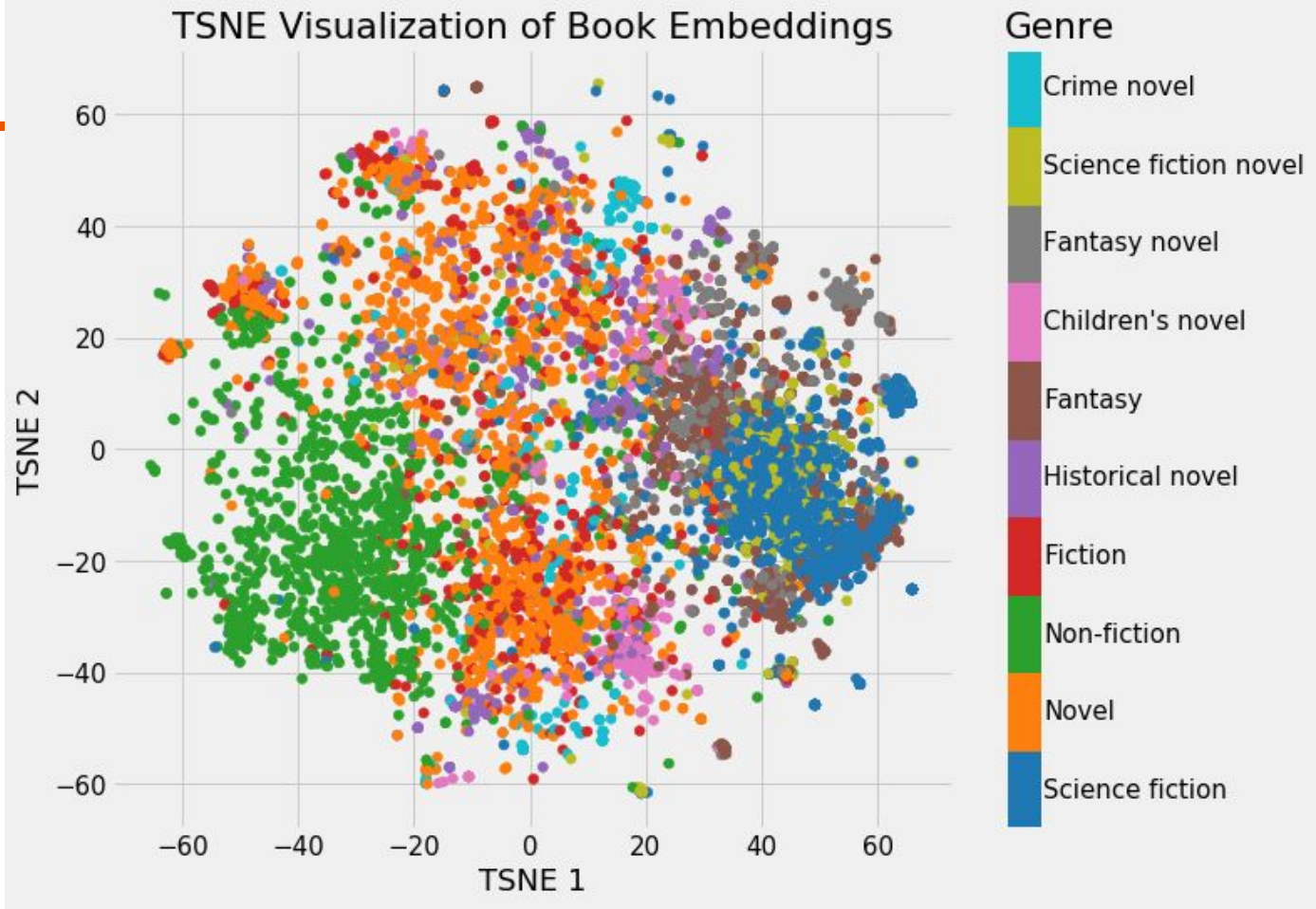
- Understands the sense of word by learning new word embeddings
- Understands the meaning of sentence by analyzing the sequence of words and some relation between them
- Currently the Best Model
- Very good at general language understanding
- Zero Shot Learning Abilities
- Requires huge Amount of GPU Resources
- Trains Efficiently Than RNN
- Data Hungry

How We Use Embeddings?

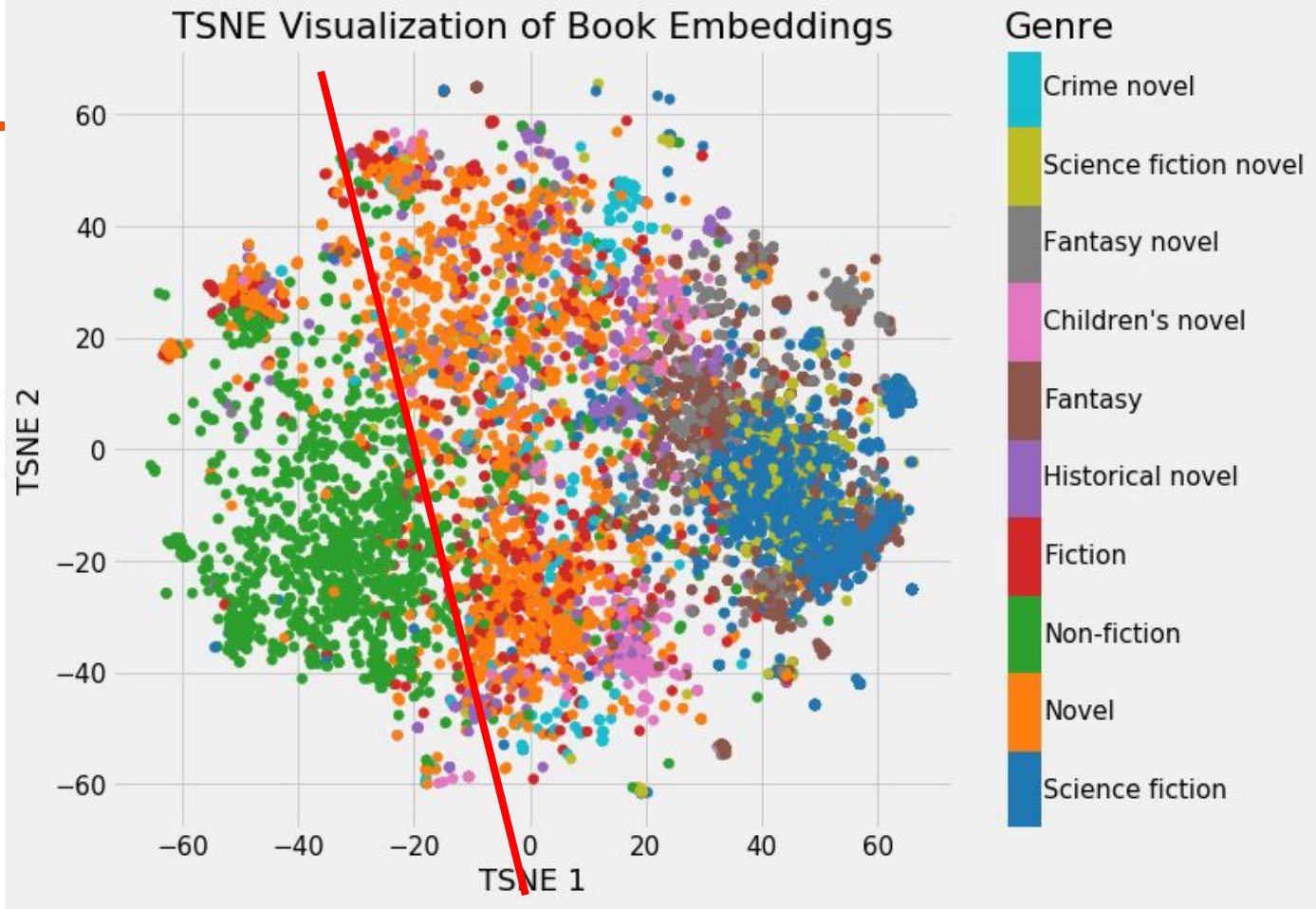
Semantic Search



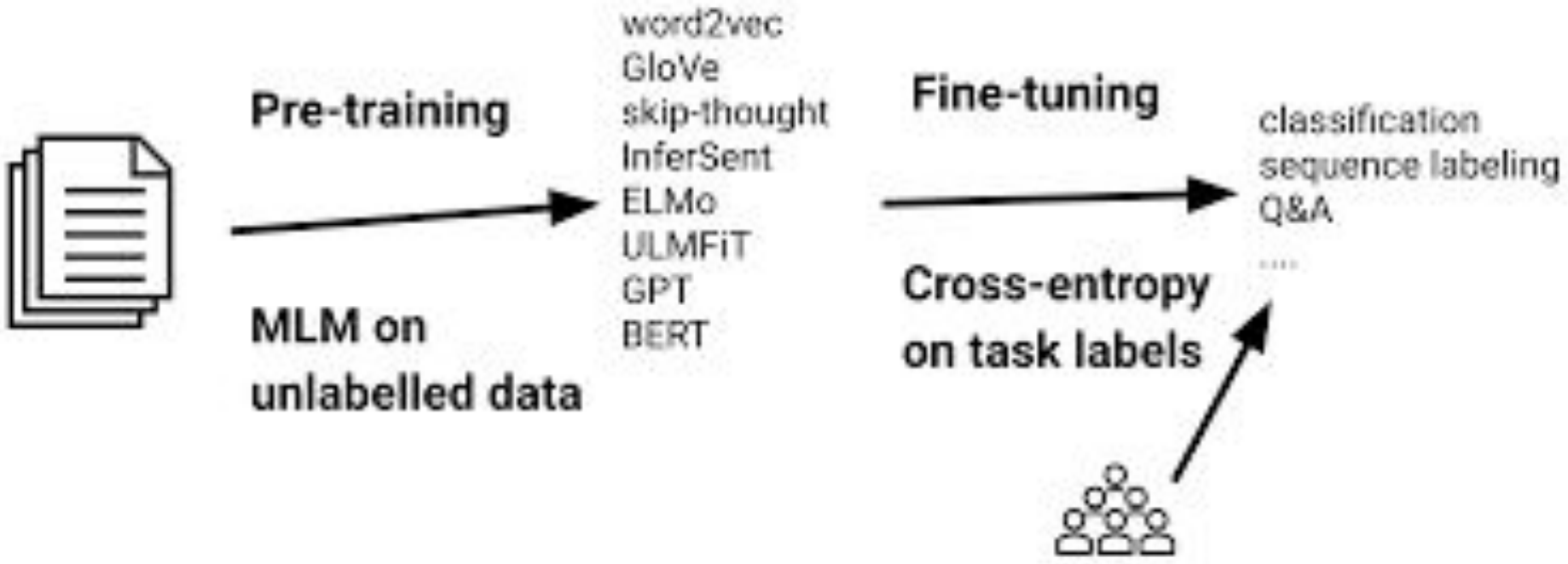
Clustering / Filtering / Outlier Removal, etc



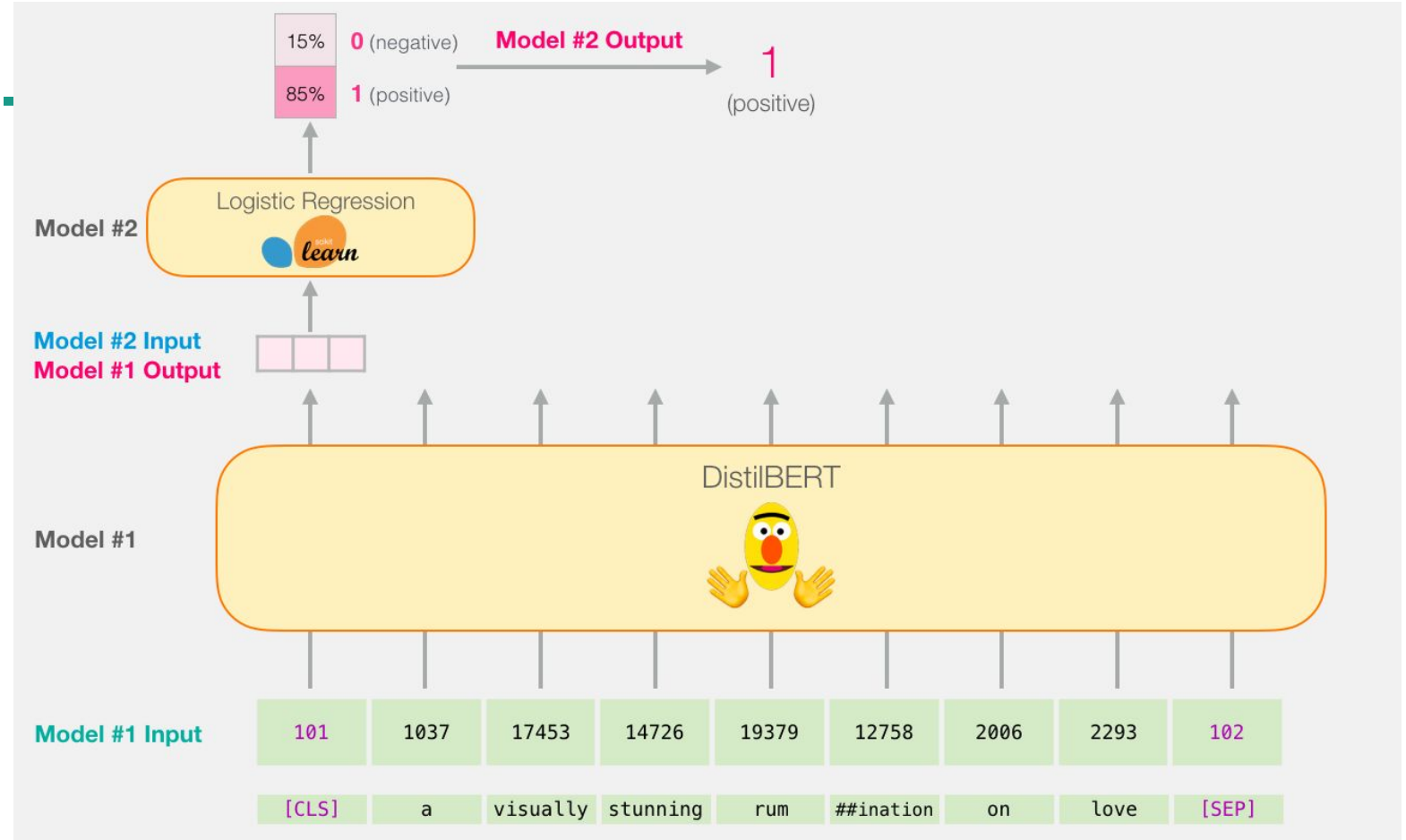
Fiction VS Non-Fiction Classifier (80+ % accuracy)



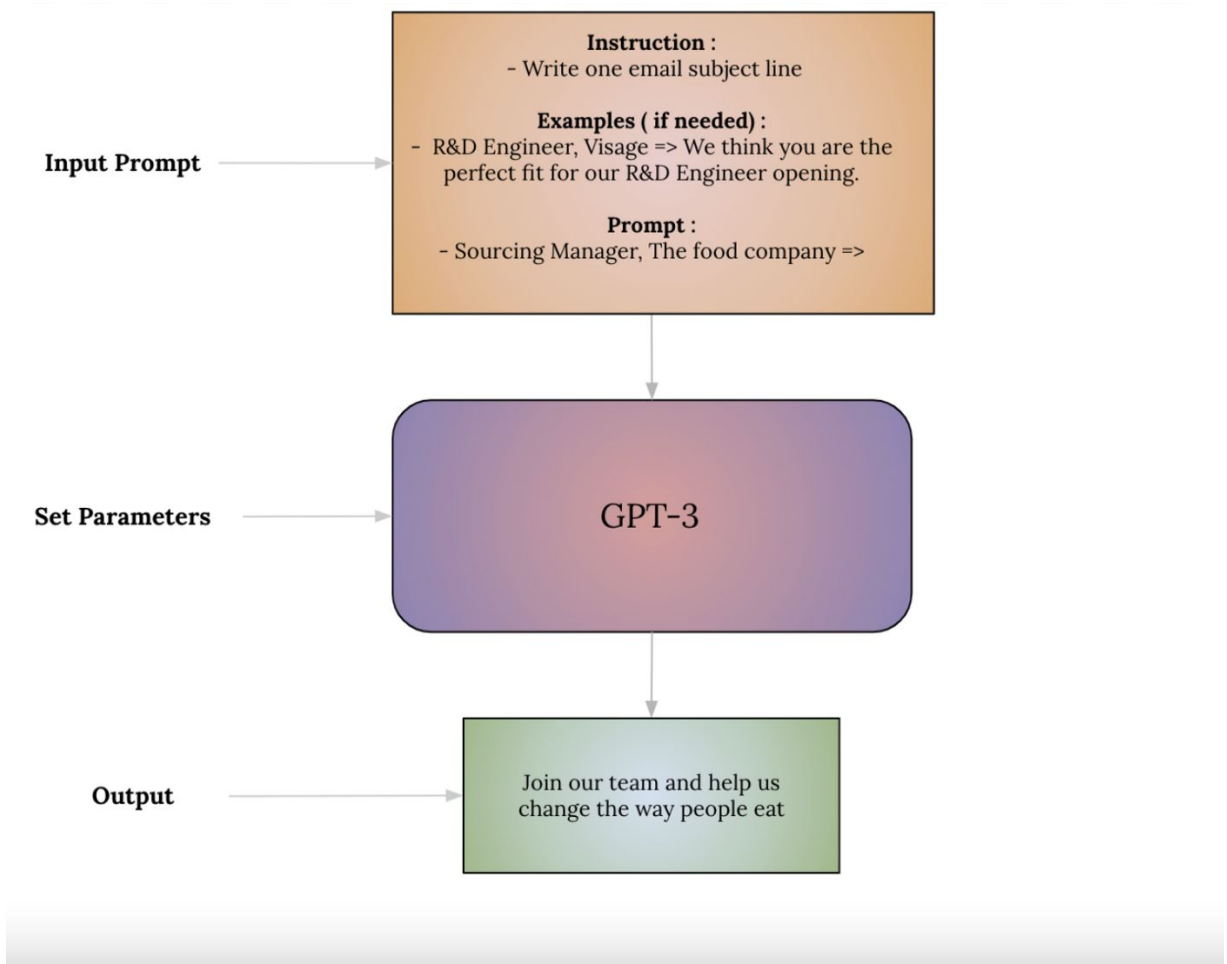
Pre-Training + Downstream Task Fine-Tuning



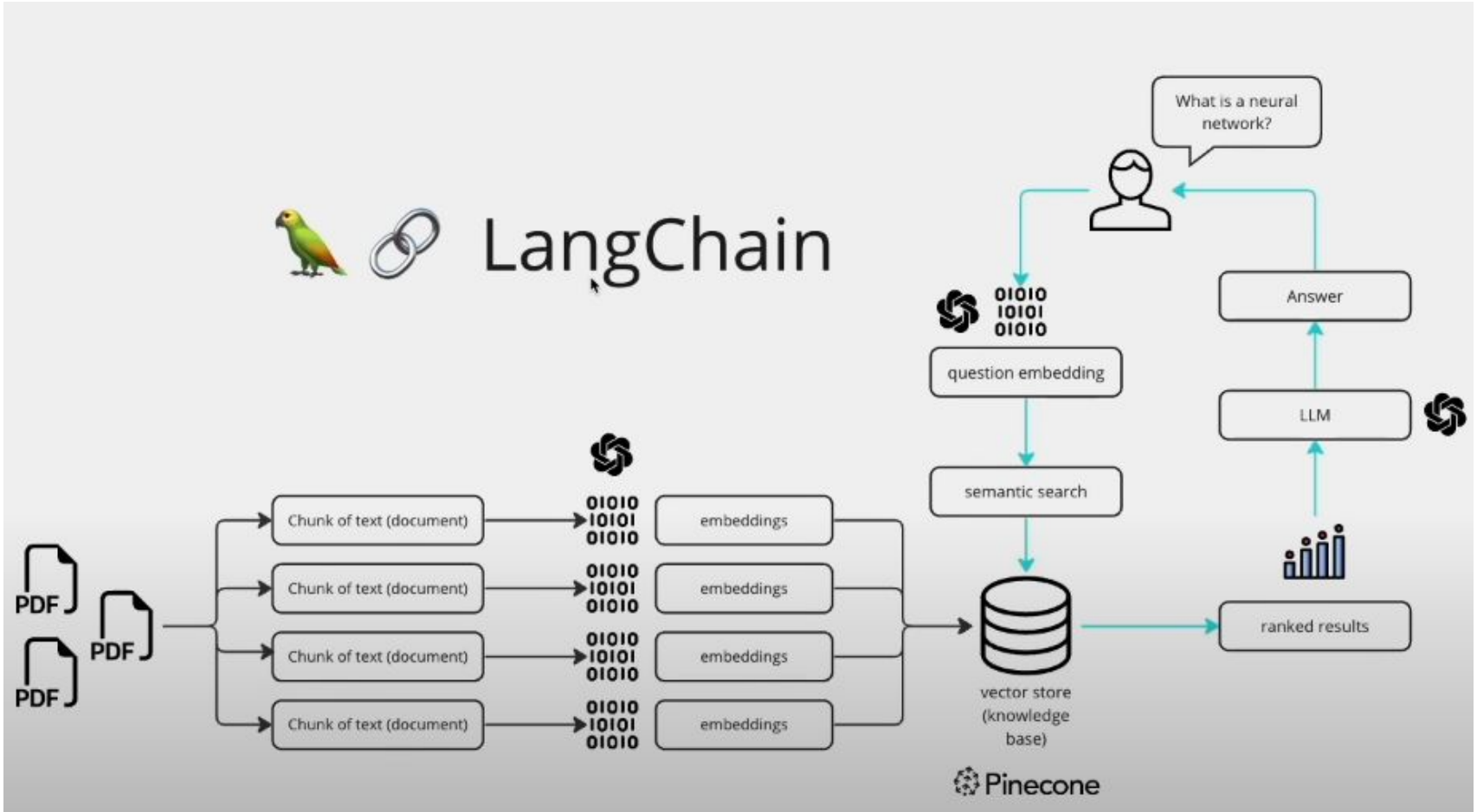
Pre-Training + Downstream Task Fine-Tuning



LLMs and Prompting (GPT-3, 175Billion Params)



LLMs and LangChain



GPT-4 Turbo (OpenAI Live - Nov 6 2023)

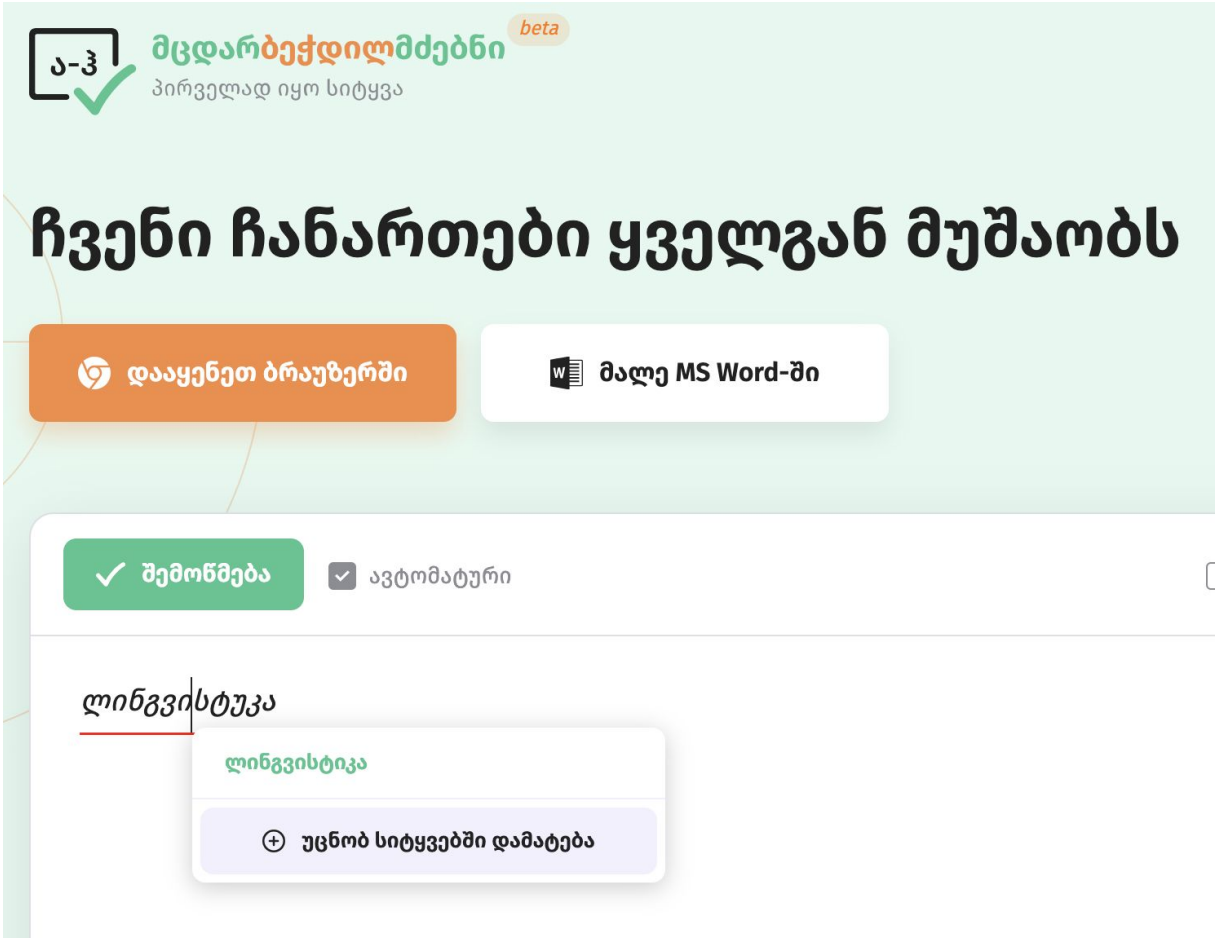
- Larger context length: 128K tokens (GPT4 had 32K)
- More Control:
  - JSON Mode: ensures valid JSON responses (Useful for large amount of API calls)
  - Multiple Function Calls: can call multiple API tools in single request
  - Reproducible Output: random seed guarantee to get consistent outputs
  - LogProbs: allows you to get logprobs of output (generated tokens)
- Better Knowledge:
  - Updated World Knowledge up to April 2023 (GPT4 had up to September 2021)
  - Retrieval System: you can add extra knowledge into DB and it can use it
- New Modalities:
  - DALL-E 3: amazing text to image generation model
  - GPT4-Turbo with Vision: Be able to understand Image content
  - TTS (Whisper V3): you can talk to GPT, even call it
- Customization: Be able to fine-tune model (experimental)
- Higher Rate Limits: 2k tokens per tokens per minute (possibility to request more)

https://www.youtube.com/watch?v=U9mJuUkhUzk&t=2385s&ab_channel=OpenAI

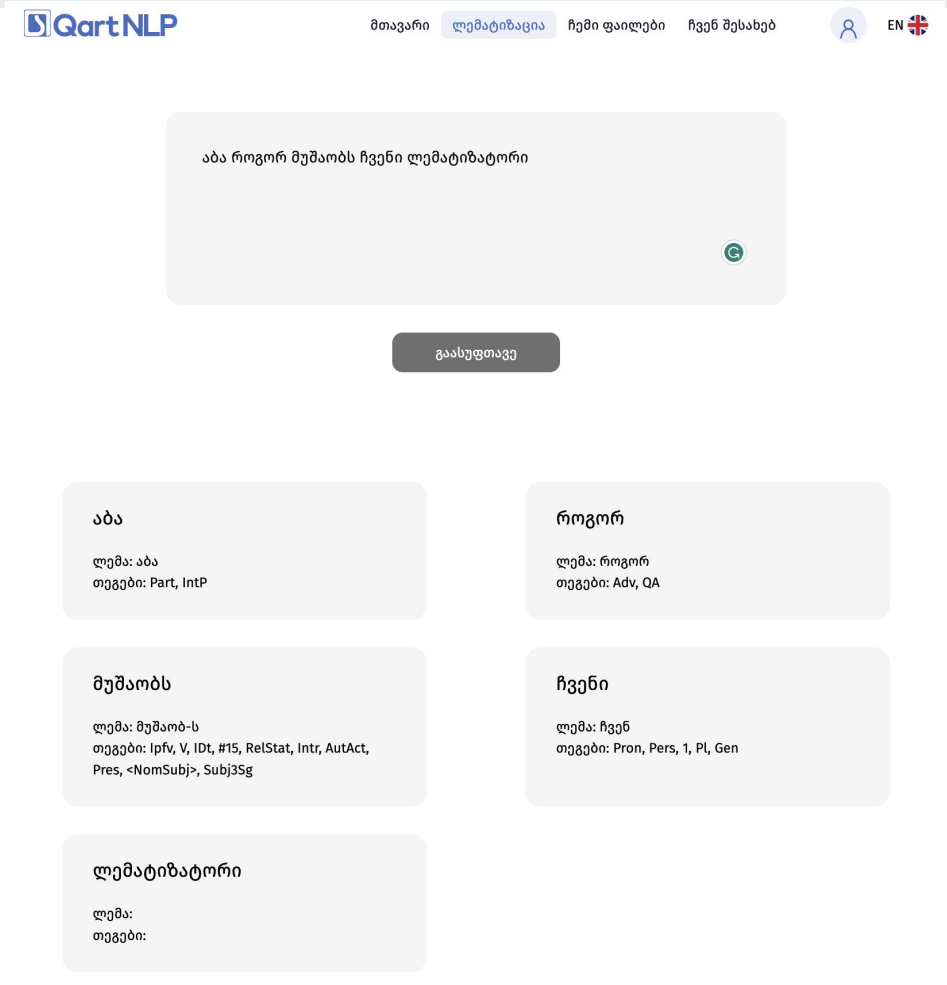
What We have?

- Spell Checker
- Lemmatizer
- Tagger
- Open Source

Spellchecker მცდარმბეჭდელი (word level)



QartNLP Lemmatizer and Tagger (iliauni)



GNC Lemmatizer and Tagger (TSU) - (Paul Meurer)

Parse a text:

- http://clarino.uib.no/gnc/parse-api?command=parse&session-id=2420972048580 72&text=გამარჯობა.

returns:


```
{"tokens":[{"word":"გამარჯობა", "msa":[{"lemma":"გამარჯობა","features":"Interj"}]}, {"word":".","msa":
[{"lemma":".","features":"Punct Period"}]}]}
```


You should URLencode the text string, or (preferably) use a POST request when parsing text.

The most important features you can set are:

- **lemma** (default: true): whether to return the lemma
- **features** (default: true) : whether to return morphosyntactic features
- **disambiguate** (default: true) : whether to run the CG
- **dependencies** (default: false) : whether to include dependency relations in the output (this is experimental)

Georgian Automatic Speech Recognition (Common Voice ~10hr)

m3hrdadfi / **wav2vec2-large-xlsr-georgian** like 3

Automatic Speech Recognition | Transformers | PyTorch | JAX | common_voice | Georgian | wav2vec2 | audio | speech | xlsr-fine-tuning-week | Eval Results | Inference Endpoints

License: apache-2.0

Model card | Files and versions | Community 1 | Train | Deploy | Use in Transformers

Edit model card

Wav2Vec2-Large-XLSR-53-Georgian

Fine-tuned facebook/wav2vec2-large-xlsr-53 in Georgian using Common Voice. When using this model, make sure that your speech input is sampled at 16kHz.

Usage

The model can be used directly (without a language model) as follows:

Requirements

```
# requirement packages
!pip install git+https://github.com/huggingface/datasets.git
!pip install git+https://github.com/huggingface/transformers.git
!pip install torchaudio
!pip install librosa
!pip install jiwer
```

Normalizer

Downloads last month
10

Hosted inference API

Automatic Speech Recognition | Examples

Browse for file or Record from browser or

Realtime speech recognition

Audio recorded from browser [7:52:24 PM]

0:01

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 1.634 s

ავა მითხარი როკორარის სიტყუა თამაში ინგლისურადაცი

JSON Output | Maximize

Georgian Common Voice Dataset: https://commonvoice.mozilla.org/ka

კრებულის ჩამოტვირთვა

გარკვეული ცვლილებებია. ნაწილობრივი „დელტა“ გამოშვებები მოიცავს წინა სრული გამოშვების შემდგომ დამატებულ ჩანაწერებს მხოლოდ. გაეცანით ვრცლად ამ სიახლეს.

მიუთითეთ სასურველი ენის კრებული და აირჩიეთ ვერსია ჩამოსატვირთად.

ენა
ქართული

ვერსია	თარიღი	ზომა	ჩაწერილი საათი	დამოწმებული საათი	ლიცენზია	ხმების რაოდენობა	ხმის ფორმატი
Common Voice Corpus 15.0	14/09/2023	3.01 GB	154	111	CC-0	1,254	MP3
Common Voice Delta Segment 15.0	14/09/2023	520.71 MB	26	24	CC-0	89	MP3
Common Voice Delta Segment 14.0	28/06/2023	1.61 GB	83	59	CC-0	642	MP3
Common Voice Corpus 14.0	28/06/2023	2.5 GB	128	87	CC-0	1,165	MP3
Common Voice Delta Segment 13.0	24/04/2023	297.29 MB	15	9	CC-0	116	MP3
Common Voice Corpus 13.0	15/03/2023	914.82 MB	45	29	CC-0	523	MP3
Common Voice Delta Segment 12.0	22/12/2022	253.42 MB	13	5	CC-0	83	MP3

Open Source Datasets and Models

- Common Crawl (cc100) - web crawl data on 100 languages (1.1GB raw text for Georgian)
- Common Voice - voice and text pairs for multiple languages (164hr records for KA)
- WikiANN - cross lingual name tagging based on wikipedia (282 languages)
- WikiMatrix - 136M parallel sentences for 1620 language pairs from wikipedia (LASER - facebook research)
- CCAligned - A Massive Collection of Cross-lingual Web-Document Pairs
- Oscar - cc100 filtered (used for training BART) 138GB of text

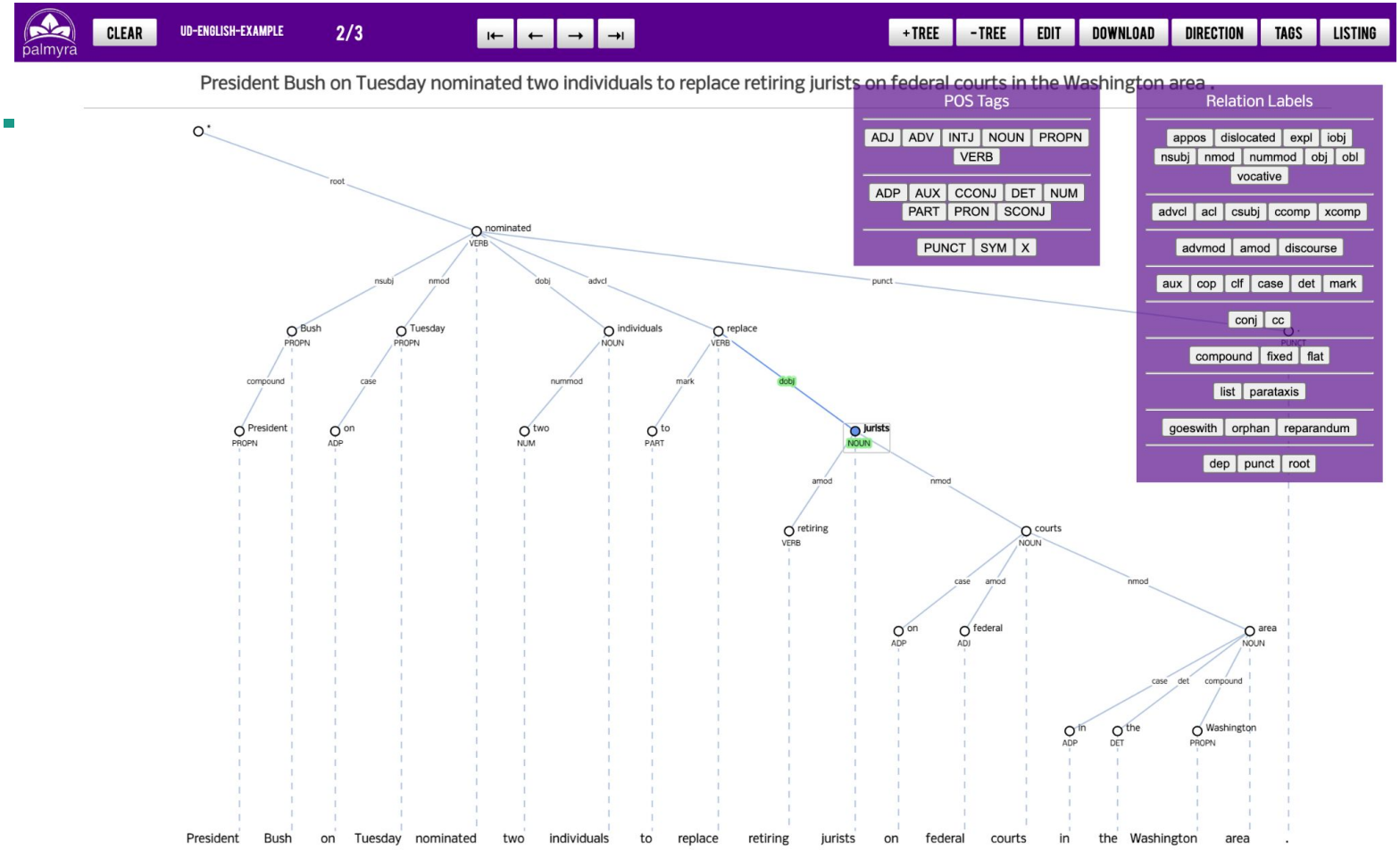
- Tesseract OCR (google)
- Tesseract OCR (iliauni)
- FastText word embeddings (300 dim, 4GB)
- Wav2Vec2-Large-XLSR-53-Georgian

Any More Open-Source Resources?

What We Need to Have?

- Treebank
- Benchmark Dataset for nlp tasks
- Large Text Corpora
- Benchmarks for LLMs

English Treebank



CoNLL Shared Tasks for LMs for English

- Chunks
- NER
- POS Tagging
- Dependency Parsing
- Coreference Resolution
- etc

GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

Large Text Corpora for English

- BookCorpus
- CommonCrawl (petabytes of web crawl data)
- ArXiv (1000s of papers)
- Wikipedia

LLM benchmarks

- Big-Bench
- ElutherAI’s evaluation harness
- Super GLUE

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

Planned / Ongoing Works

- Georgian Large Corpora / LM / LLM / Benchmarks
- Common Voice
- English to Georgian Translation
- Georgian Treebank
- Georgian OCR Benchmark

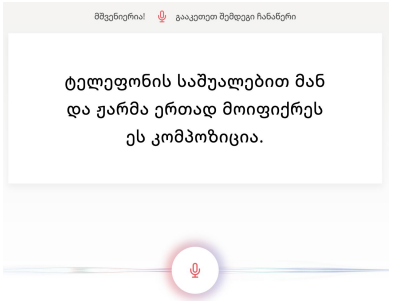
KALMO - Open Source Research Team

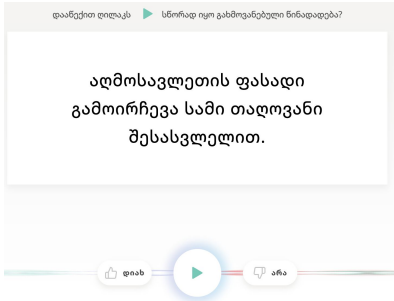
- Resources To Be Created And Published
  - Current Working On:
    - Georgian Large Corpora
    - Georgian LM Benchmarks
    - Georgian LM
  - Future Plans:
    - Georgian LLM
    - Georgian LLM Benchmarks
- Ultimate Goals:
  - All Open Source
  - Publish Scientific Paper
  - Popularize ML Research (motivated by improving Georgian NLP)

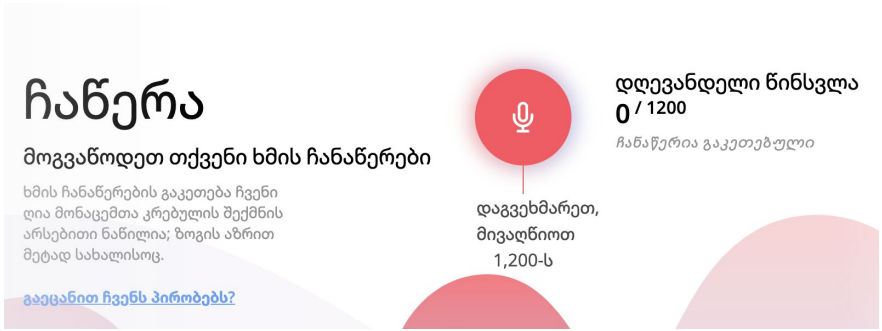
Georgian Common Voice Data Labeling

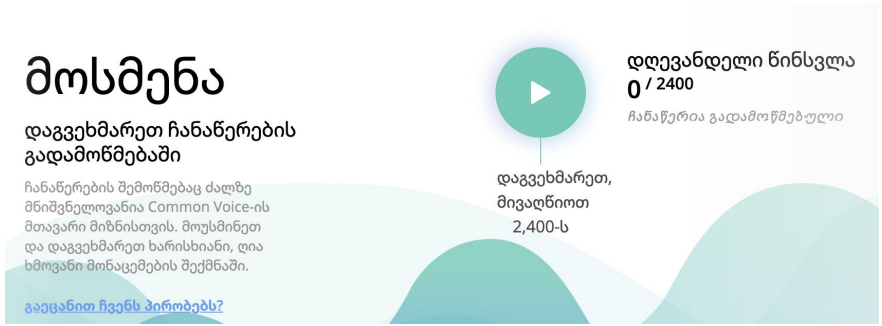
რაზმიკ ბადალიანი

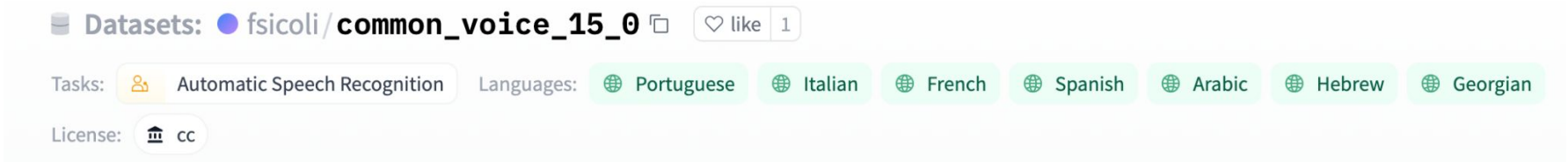
ვასწავლოთ ხელოვნურ ინტელექტს ქართული ენა











https://huggingface.co/datasets/fsicoli/common_voice_15_0

https://commonvoice.mozilla.org/ka

AI Lab (Zaal Gachechiladze): Machine Translation

- Current Work:
  - Machine Translation Dataset (Manually Labeled)
  - Machine Translation Model
  - Narrow Domain: A fairy Tale
  - Better than Google Translate in this domain
- Future Goals
  - Add more domains
  - Publish models for free (Open Source)
  - Publish Scientific Paper
  - Participate in WMT conference
  - Organize Machine Translation Hackathons in Georgia

Georgian Universal Dependencies

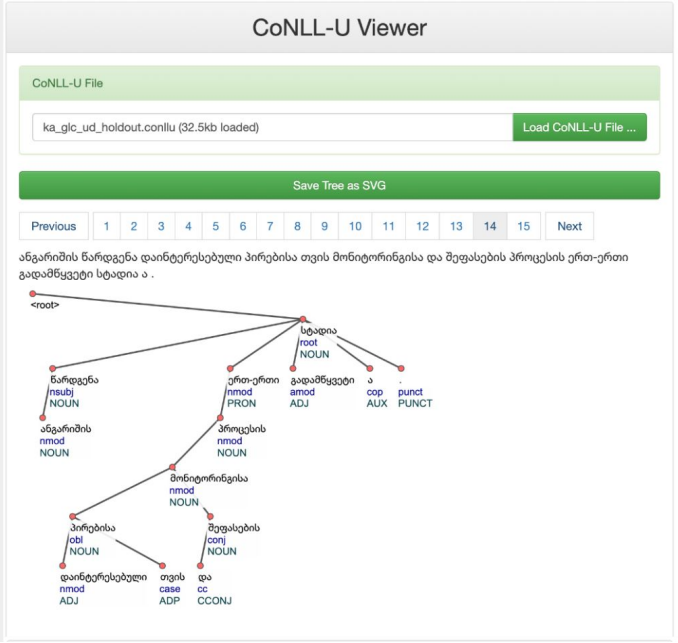
▸	Georgian	1	2K		Kartvelian

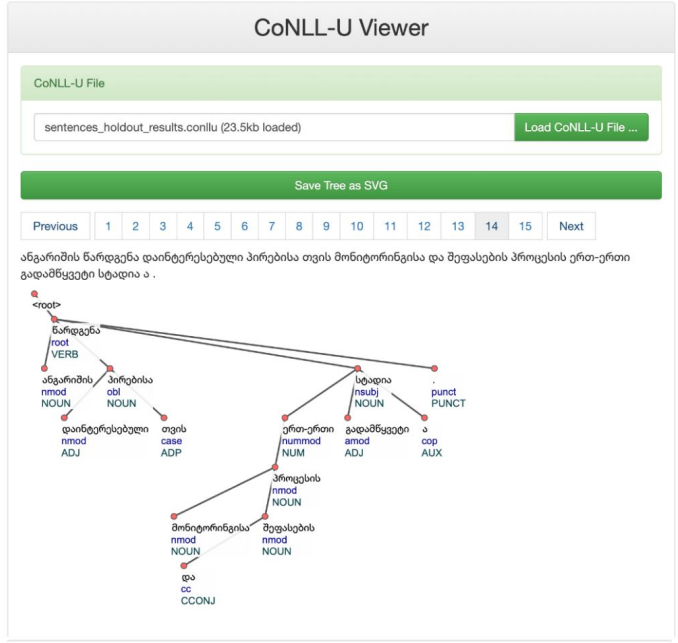
UD_Georgian-GLC Public

https://github.com/UniversalDependencies/UD_Georgian-GLC/tree/master

https://universaldependencies.org/ka/index.html

https://www.anz2.website/nlp/unlocking-georgian-language-understanding-the-ud-treebank-project-at-ilia-state-university/





Component	Steps/Iterations	Metrics
Tokenizer	Epoch 32, logprob: -9.6536e+02, training accuracy: 99.93%	Heldout tokens: 100.00% Precision/100.00% Recall, Sentences: 100.00% Precision/100.00% Recall
Tagger	Iteration 20: Accuracy 100.00%	Heldout accuracy: 72.26% Tokens / 83.22% Lemmas / 64.73% UPOS
Parser	Iteration 10: Training logprob -2.7755e+03	Heldout UAS (Unlabeled Attachment Score): 55.82%, Heldout LAS (Labeled Attachment Score): 49.66%

Georgian OCR Benchmark & Evaluation Package

- Total Documents (Single Page Image): 77
- Annotations:
  - Block
  - Paragraph
  - Word
- Document Categories:
  - Journal
  - Newspaper
  - Book
  - Flyer / Brochure
  - Personal Document
  - Cheque
  - Road Sign (3D text)
  - Etc

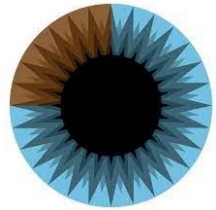
- New Evaluation Method Features:
  - Word level matching
  - Word Group (in single line) level matching
  - Calculated Metrics:
    - Accuracy
    - Precision
    - Recall
    - F1
  - Uses edit distance algorithm to:
    - Find missing symbol
    - Find extra symbol
    - Find incorrectly detected symbol
  - Summary of edit operations to have global picture on what symbols are mistaken and in which form

Good NLP Resources

- ML Platforms
- Personal Blogs
- Papers
- Conferences

NLP Resources

- Huggingface
- NLP progress (sebastian ruder)
- Paperswithcode
- Spacy, stanza, nltk
- Conll
- ICML, NLP Summit, ACL, ICLR,
- Coursera
- Colah's blog
- 3blue1brown (youtube channel)



https://course.spacy.io/en/

მადლობა

გასაკეთებელი

ეგრეე

გმირობის

იქნება

მალადეც

მადლობაა

მადლობა ყურადღებისთვის

ყოჩაღ