Machine Learning

Natural Language Processing

Karol Przystalski

May 15, 2024

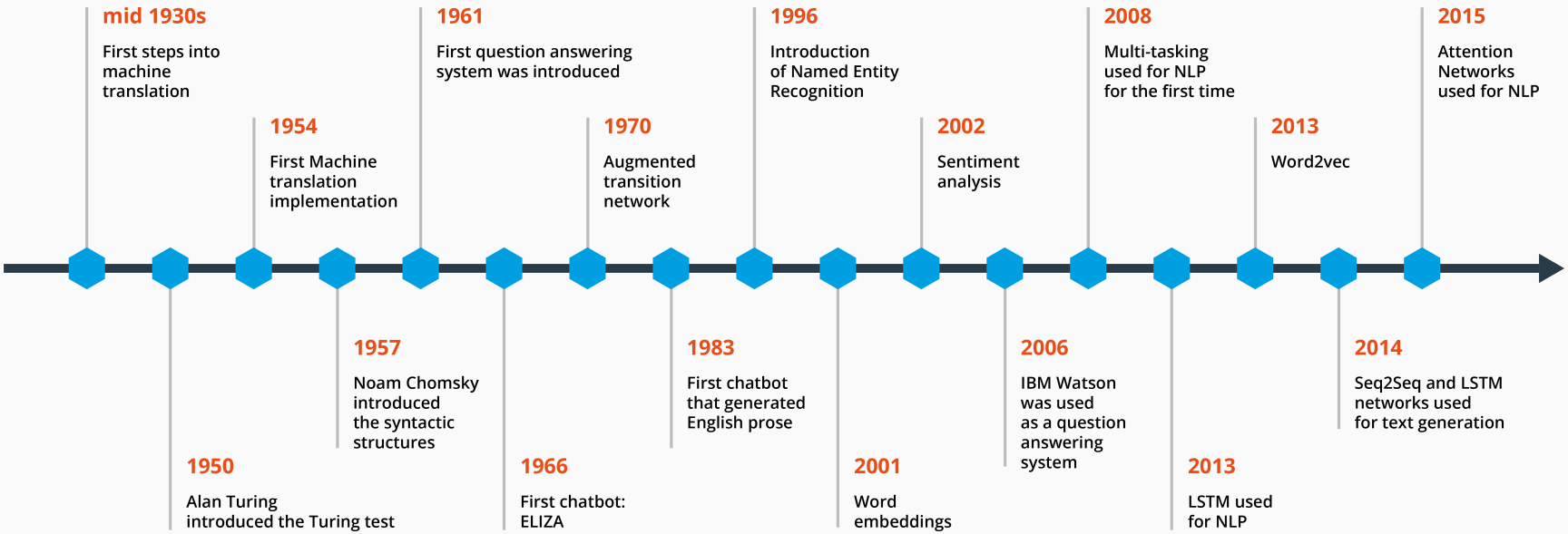
Department of Information Technologies, Jagiellonian University

Agenda

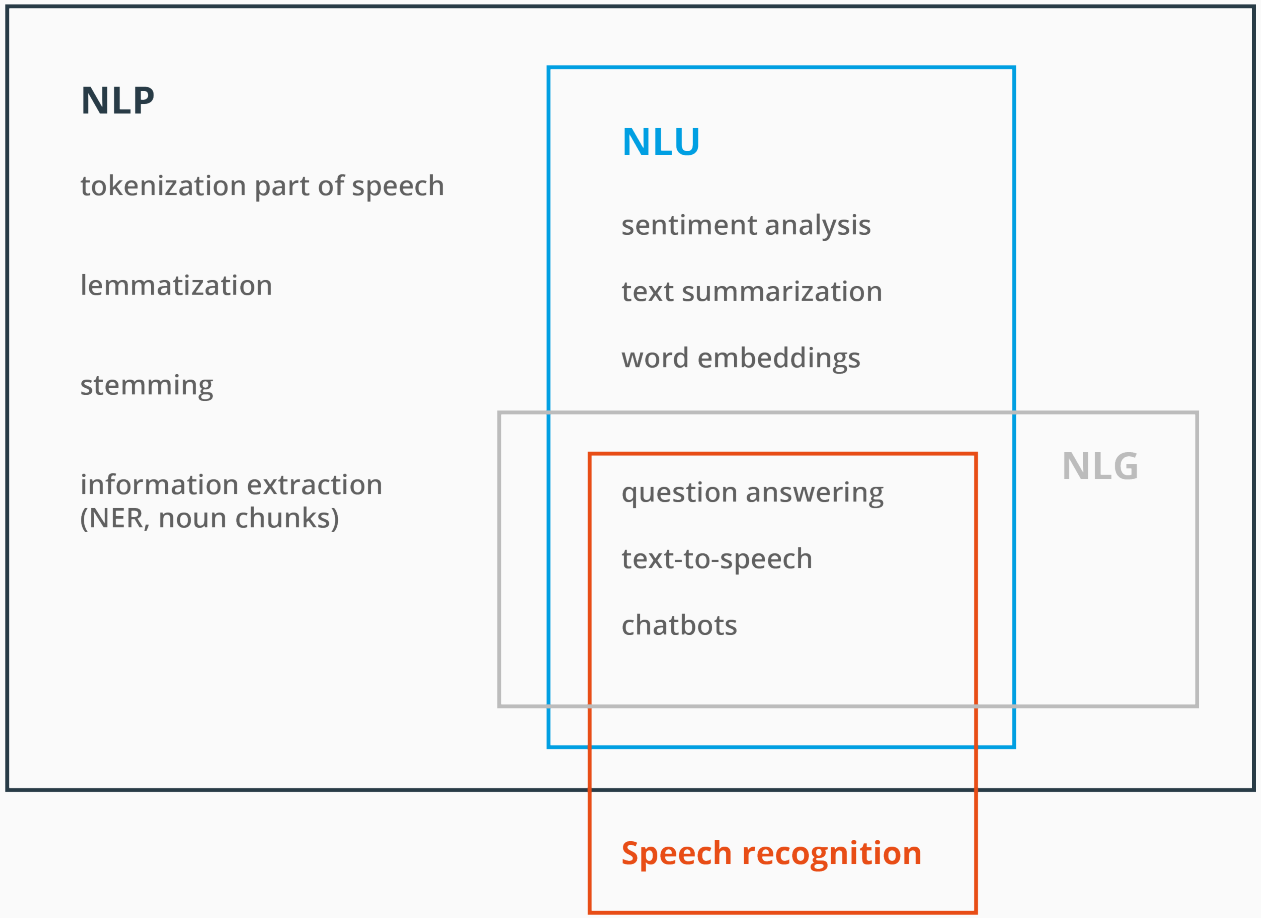
1. Introduction
2. Use cases
3. Text processing
4. Natural Language Processing
5. Text understanding
6. Chatbots
7. Text generation
8. Quality metrics

Introduction

Timeline


mid 1930s
First steps into machine translation
1950
Alan Turing introduced the Turing test
1954
First Machine translation implementation
1957
Noam Chomsky introduced the syntactic structures
1961
First question answering system was introduced
1966
First chatbot: ELIZA
1970
Augmented transition network
1983
First chatbot that generated English prose
1996
Introduction of Named Entity Recognition
2001
Word embeddings
2002
Sentiment analysis
2006
IBM Watson was used as a question answering system
2008
Multi-tasking used for NLP for the first time
2013
LSTM used for NLP
2013
Word2vec
2014
Seq2Seq and LSTM networks used for text generation
2015
Attention Networks used for NLP

Taxonomy

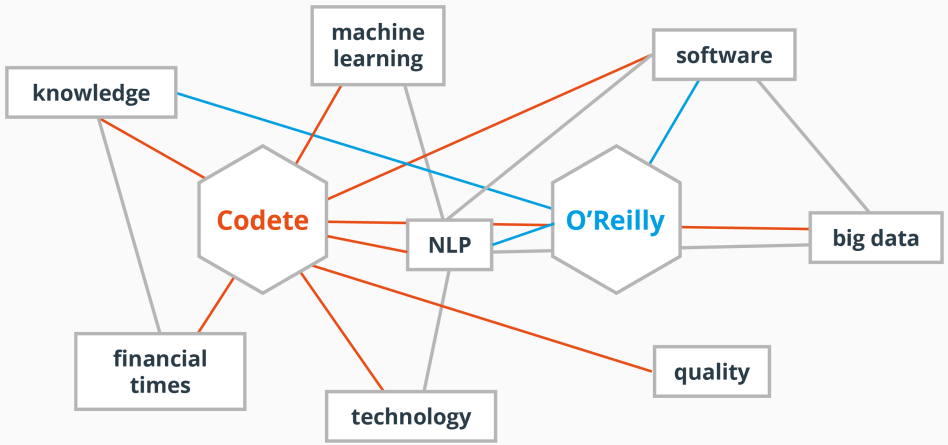

NLP
tokenization part of speech
lemmatization
stemming
information extraction
(NER, noun chunks)
NLU
sentiment analysis
text summarization
word embeddings
NLG
question answering
text-to-speech
chatbots
Speech recognition

Use cases

Use cases

There are many use cases of NLP methods, just to mention a few:

- term relationship,



Use cases

There are many use cases of NLP methods, just to mention a few:

- term relationship,
- text summarization,

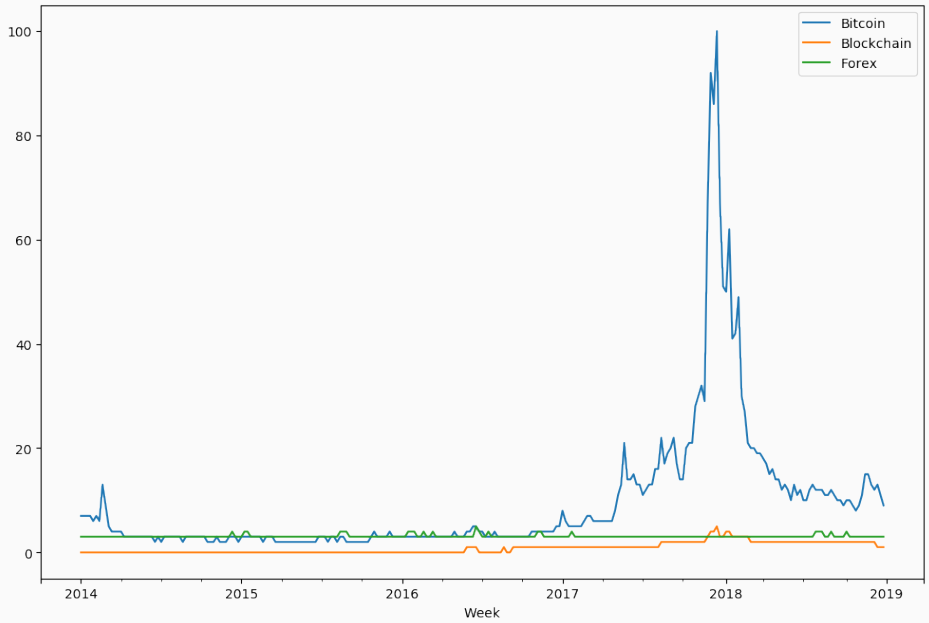

BBC
NEWS
Home | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts
Politics | Parliaments | Brexit
'Difficult' Brexit talks see no breakthrough, European Commission says
Brexit negotiations "have been difficult" and "no solution has been identified" to the Irish backstop, the European Commission has said.
It comes after the latest talks between UK ministers and EU Brexit negotiator Michel Barnier in Brussels.
Commission spokesman Margaritis Schinas said the talks had taken place in a "constructive atmosphere" but there had been no breakthrough.
The UK is pushing for legally-binding changes to the EU deal.
Mr Schinas was speaking after Mr Barnier briefed the European Commission's weekly meeting on the state of Brexit talks.

Brexit: UK and UE negotiations failed

Use cases

There are many use cases of NLP methods, just to mention a few:

- term relationship,
- text summarization,
- time series analysis,



Use cases

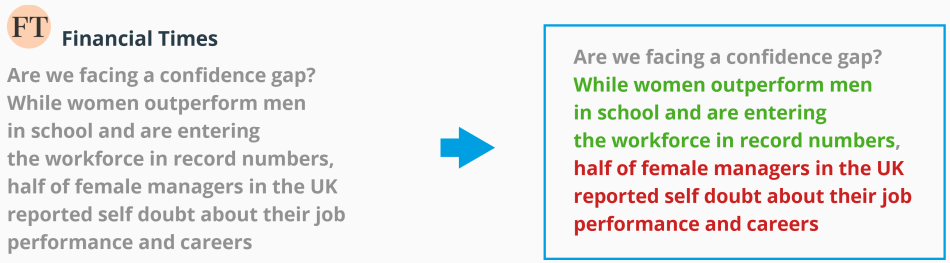
There are many use cases of NLP methods, just to mention a few:

- term relationship,
- text summarization,
- time series analysis,
- intent recognition,

Use cases

There are many use cases of NLP methods, just to mention a few:

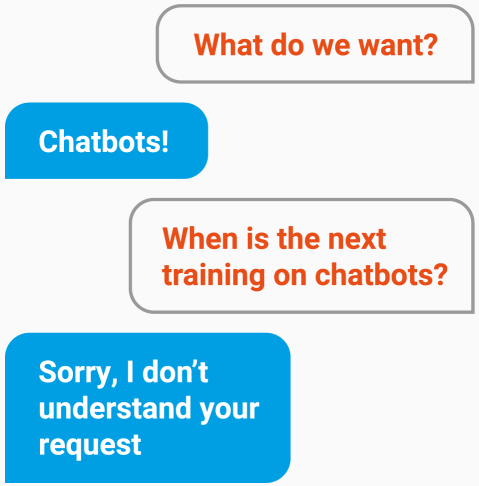
- term relationship,
- text summarization,
- time series analysis,
- intent recognition,
- sentiment analysis,



Use cases

There are many use cases of NLP methods, just to mention a few:

- term relationship,
- text summarization,
- time series analysis,
- intent recognition,
- sentiment analysis,
- text generation.



Text processing

Regular expressions in Python for string comparison

Regular expressions in Python are almost the same as in any other programming languages. We can use regex methods to:

- search – finds only the first occurrence of expression in text,
- match – finds all occurrences of expression in text,
- fullmatch – matches only if the whole string matches the regular expression,
- split – splits into a list based on the splitting expression,
- escape – replaces all characters in the pattern.

Regular expressions are used within many methods that we go through in the next slides.

Word and sentence comparison methods

String comparison methods available in Python:

- Levenshtein distance,
- Damerau-Levenshtein distance,
- Jaro distance,
- Jaro-Winkler distance,
- Match rating approach comparison,
- Hamming distance,
- Gestalt pattern matching.

You can use at least two libraries:

- `Difflib` – `https://docs.python.org/3.6/library/difflib.html`,
- `Jellyfish` – `https://pypi.org/project/jellyfish/`.

String comparison – Levenshtein Distance

The Levenshtein distance is a number of insertion, deletion or replacement changes that needs to be done to get the same strings.

It is a number that is equal or higher than 0. It can be normalized to get a number from 0 to 1.

compared words	word length
train*ing*	8
train*s*	6

The distance for both words is 3. After the normalization the distance is $\frac{3}{8} = 0.375$.

String comparison – Gestalt pattern matching

This solution can be formulated as:

$$G_{PM} = \frac{\#\text{same characters}}{\#\text{total characters}}.$$

For the same example we have 5 same characters in each word and four that are different. This makes the G_{PM} value:

$$G_{PM} = \frac{10}{14} = 0.7142.$$

SQL Like vs. Full-text search

The full-text search is in most cases much faster than a Like query.

$$\text{bm25}(D, Q) = -1 \sum_{i=1}^{n} \text{IDF}(q_i) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \frac{|D|}{\text{avgl}})},$$

where:

- $|D|$ is the number of tokens in the current document,
- k_1 and b are constants with values 1.2 and 0.75,
- avgl is the average number of tokens.

SQL Like vs. Full-text search

IDF is the inverse-document-frequency of query phrase i and is formulated as:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

where:

- N is the total number of rows in table,
- $n(q_i)$ is the total number of rows that contain at least one instance of phrase i.

$f(q_i, D)$ is the phrase frequency of phrase i:

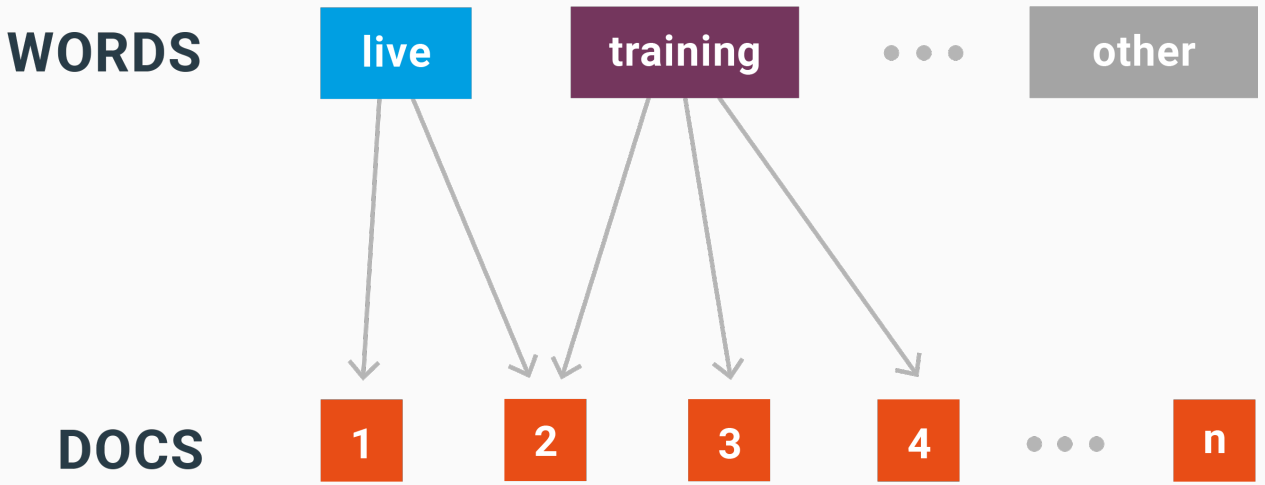
$$f(q_i, D) = \sum_{1}^{nc} w_c \cdot n(q_i, c),$$

where:

- w_c are the weights assigned to columns,
- $n(q_i, c)$is the number of occurrences of phrase i in column c of the current row.

SQL Full-text search

This is a **live** **training**.



Natural Language Processing

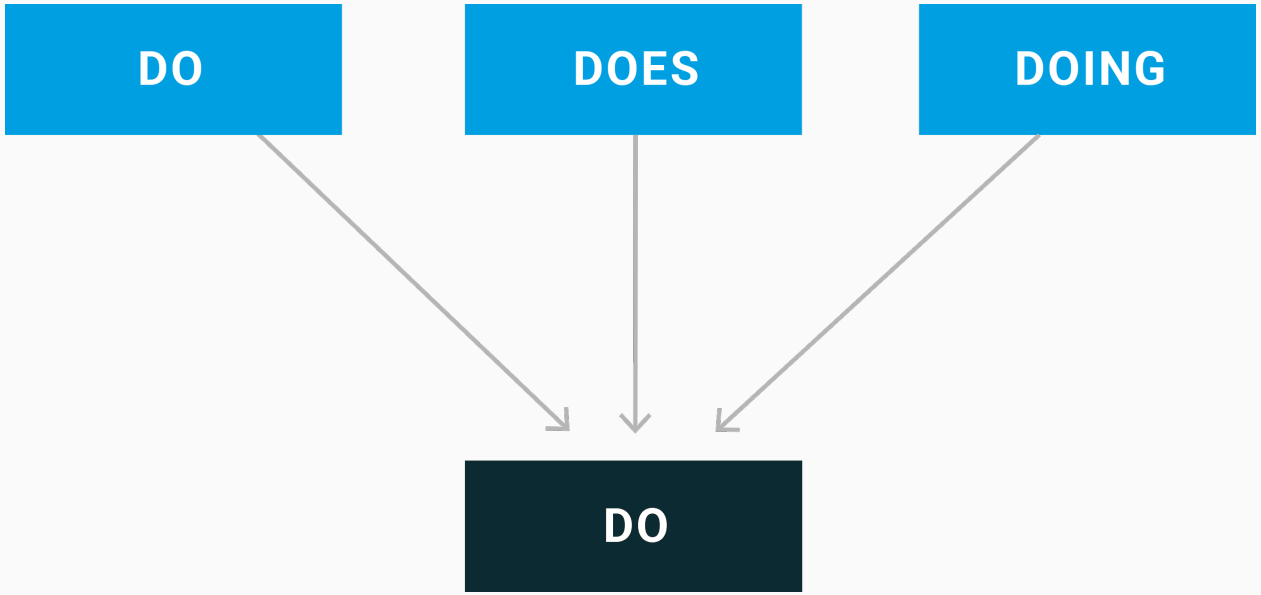
NLP methods used for sentence comparison

There are three popular methods that are used in NLP for text processing:

- tokenization,
- lemmatization,
- stemming.

Tokenization divides a sentence into separate words.

Lemmatization and stemming

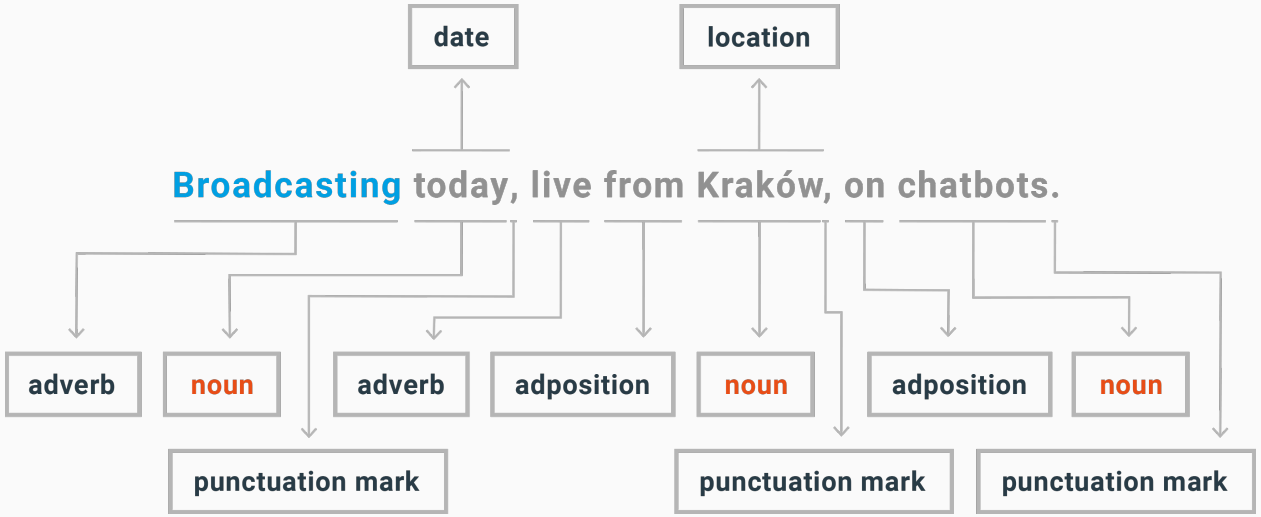


More advanced NLP methods

There are three popular NLP methods that make it easier to understand written text:

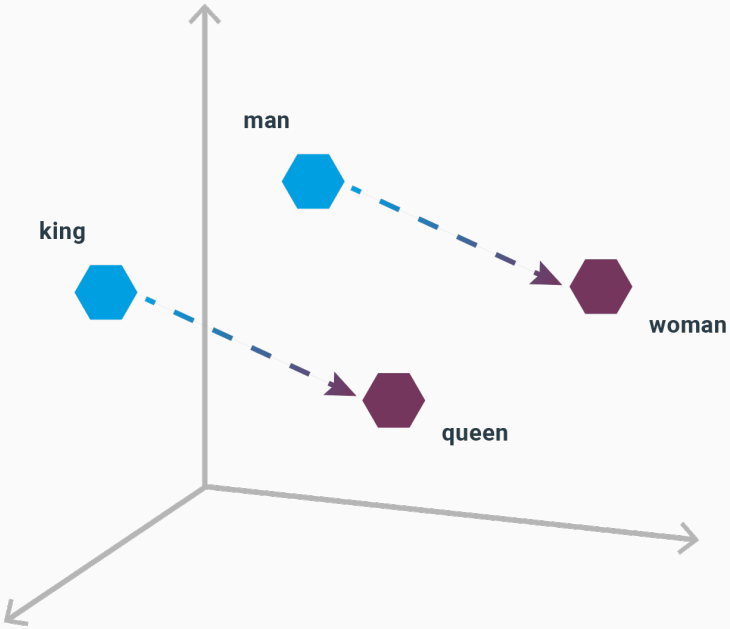
- part of speech,
- noun chunk,
- named entity recognition.

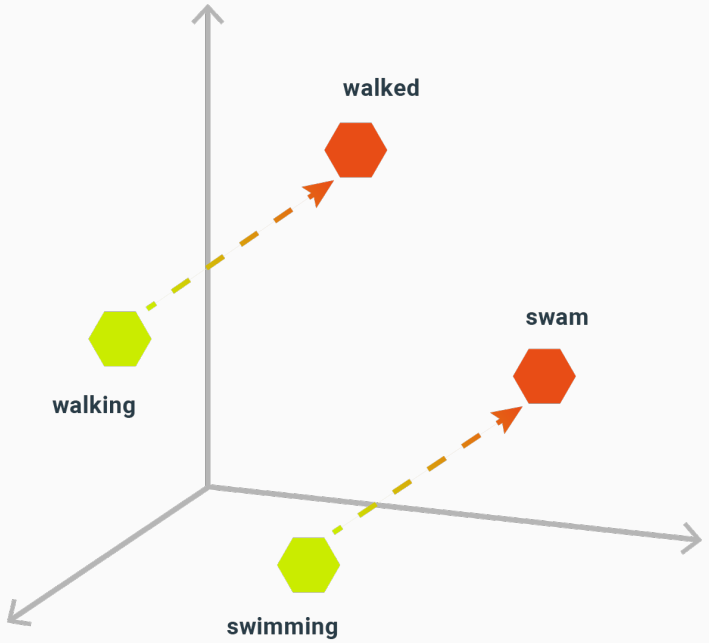
More advanced NLP methods



Text understanding

Word vectorization





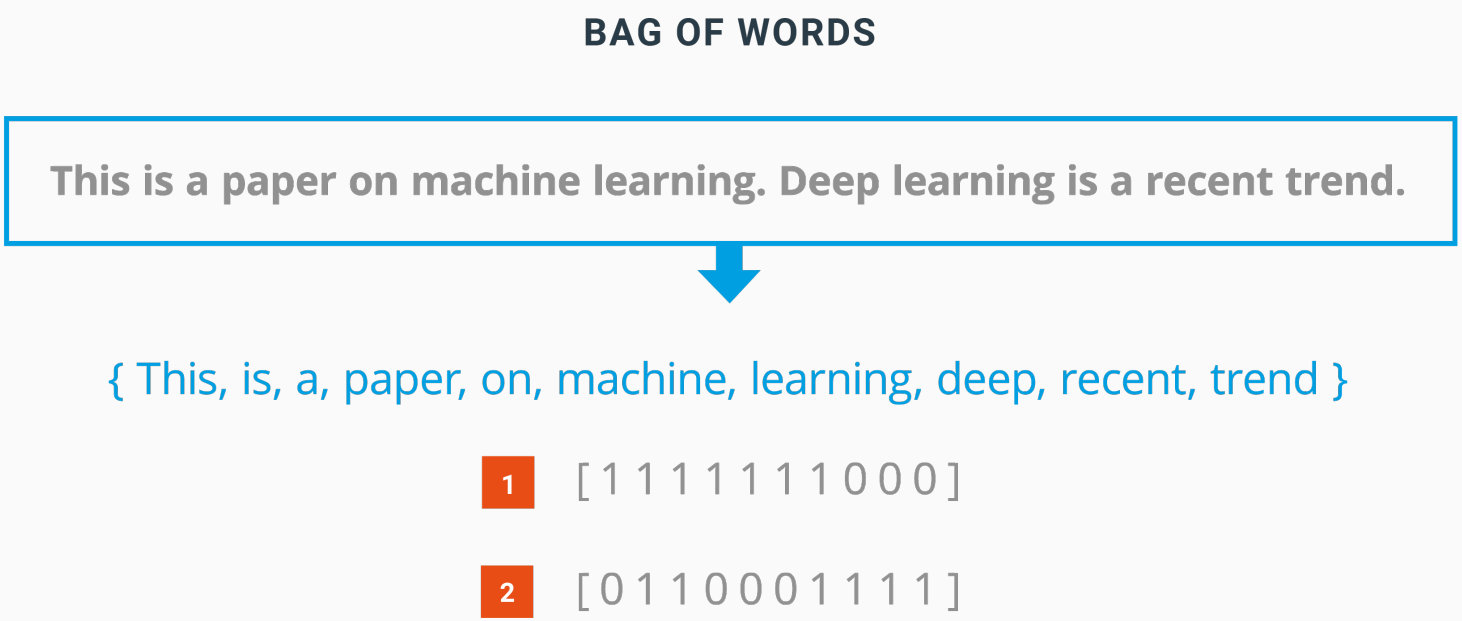
Word vectorization – concept

Word vectorization – methods

The most popular methods that are used to create a space of vectorized words are:

- bag of words,
- tf-idf,
- transfer learning,
- n-gram model,
- skip-thought vectors.

Bag of words

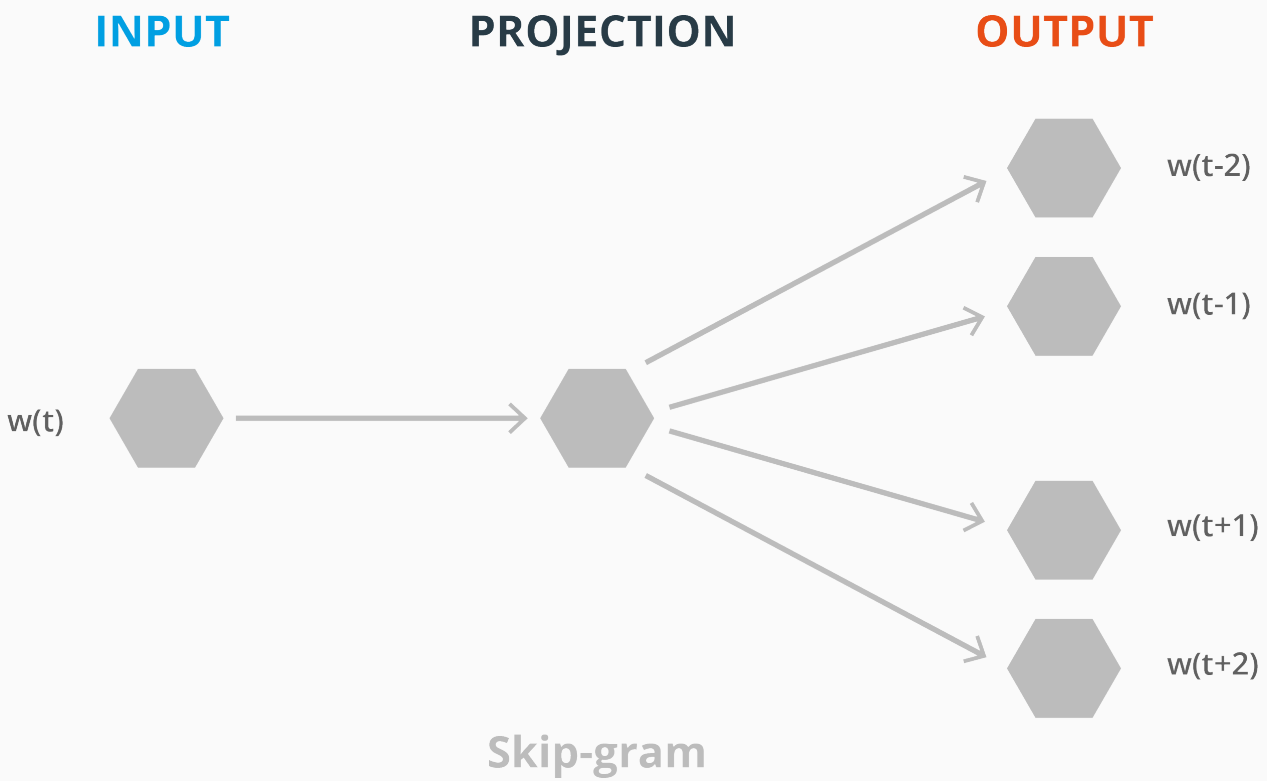

BAG OF WORDS
This is a paper on machine learning. Deep learning is a recent trend.
{ This, is, a, paper, on, machine, learning, deep, recent, trend }
1
[1 1 1 1 1 1 1 0 0 0]
2
[0 1 1 0 0 0 1 1 1 1]

Distance metrics

Also known as similarity or dissimilarity measures.

Measure name	equation	
Manhattan distance	$\rho_{Man}(x_r, x_s) = \sum_{i=1}^{n} \lvert x_{ri} - x_{si} \rvert$	(1)
Chebyshew distance	$\rho_{Ch}(x_r, x_s) = max_{1 \leq i \leq n} \lvert x_{ri} - x_{si} \rvert$	(2)
Frecht distance	$\rho(x_r, x_s) = \sum_{i=1}^{d} \frac{\lvert x_{ri} - x_{si} \rvert}{1 + \lvert x_{ri} + x_{si} \rvert} \frac{1}{2^i}$	(3)
Canberra distance	$\rho(x_r, x_s) = \sum_{i=1}^{d} \frac{\lvert x_{ri} - x_{si} \rvert}{\lvert x_{ri} + x_{si} \rvert}$	(4)
Post office distance	$\rho_{pos}(x_r, x_s) = \begin{cases} \rho_{Min}(x_r, 0) + \rho_{Min}(0, x_s), & \text{for } x_r \neq x_s, \\ 0, & \text{for } x_r = x_s \end{cases}$	(5)
Bray-Curtis distance	$\rho_{bc}(x_r, x_s) = \frac{\sum_{i=1}^{d} \lvert x_{ri} - x_{si} \rvert}{\sum_{i=1}^{d} (x_{ri} - x_{si})}$	(6)

Word vectorization – concept



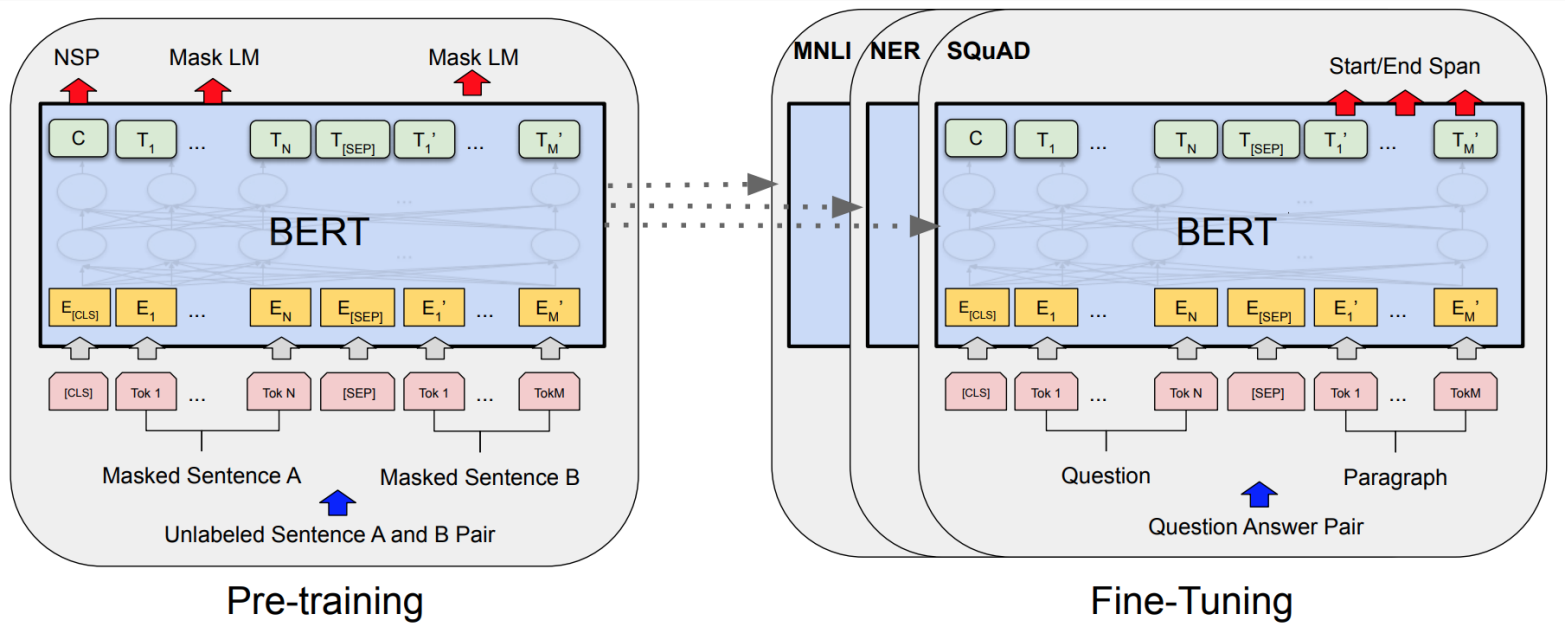
Modern word vectorization methods

The BoW or Tf-idf methods are simople, but currently more advanced methods are used like:

- Word2Vec – `https://arxiv.org/abs/1301.3781`,
- GloVe – `https://nlp.stanford.edu/projects/glove/`,
- BERT (VisualBERT) – `https://arxiv.org/abs/1810.04805`.

BERT

Source: https://arxiv.org/abs/1810.04805



Chatbots

Chatbots – a new interface

Bots are a new way of communication between the user and the app [1].

[1] Designing Bots, 1st Edition. *Amir Shevat*, O'Reilly Media 2017

Bot taxonomy

Bots can be divided into a few types, based on:

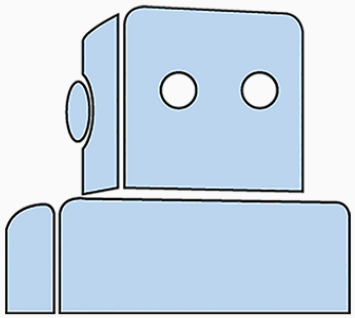
- interface – automation, audio or text,
- privacy – on-site and online,
- usage – superbots, domain-driven, etc.[2]

[1] Designing Bots, 1st Edition. *Amir Shevat*, O'Reilly Media 2017

Implementations

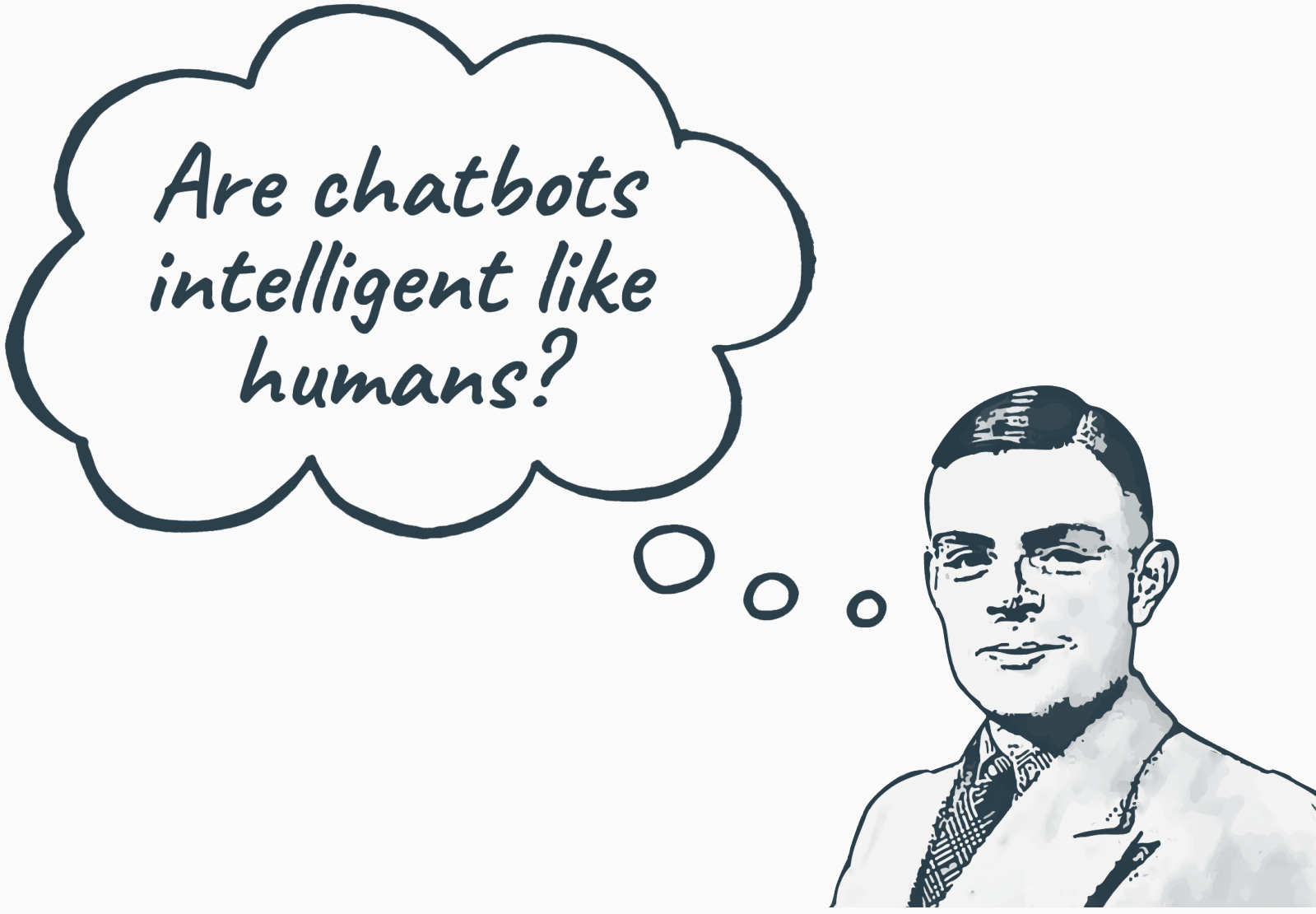
{LawGeex}

Tools

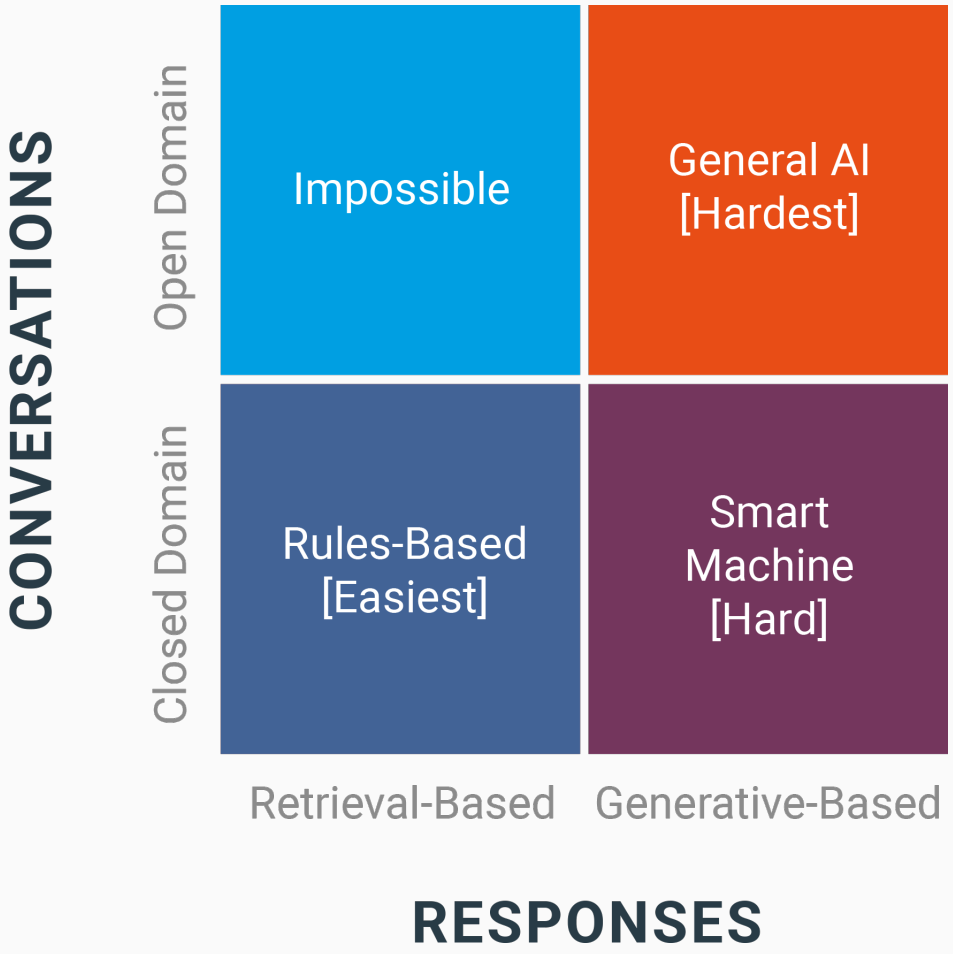


Integration

You can find a short explanation on how to start in the chatbots notebooks:
`https://github.com/codete/oreilly-intelligent-bots/blob/master/Chatbot_Integrations.ipynb`


Are chatbots intelligent like humans?

Bot matrix



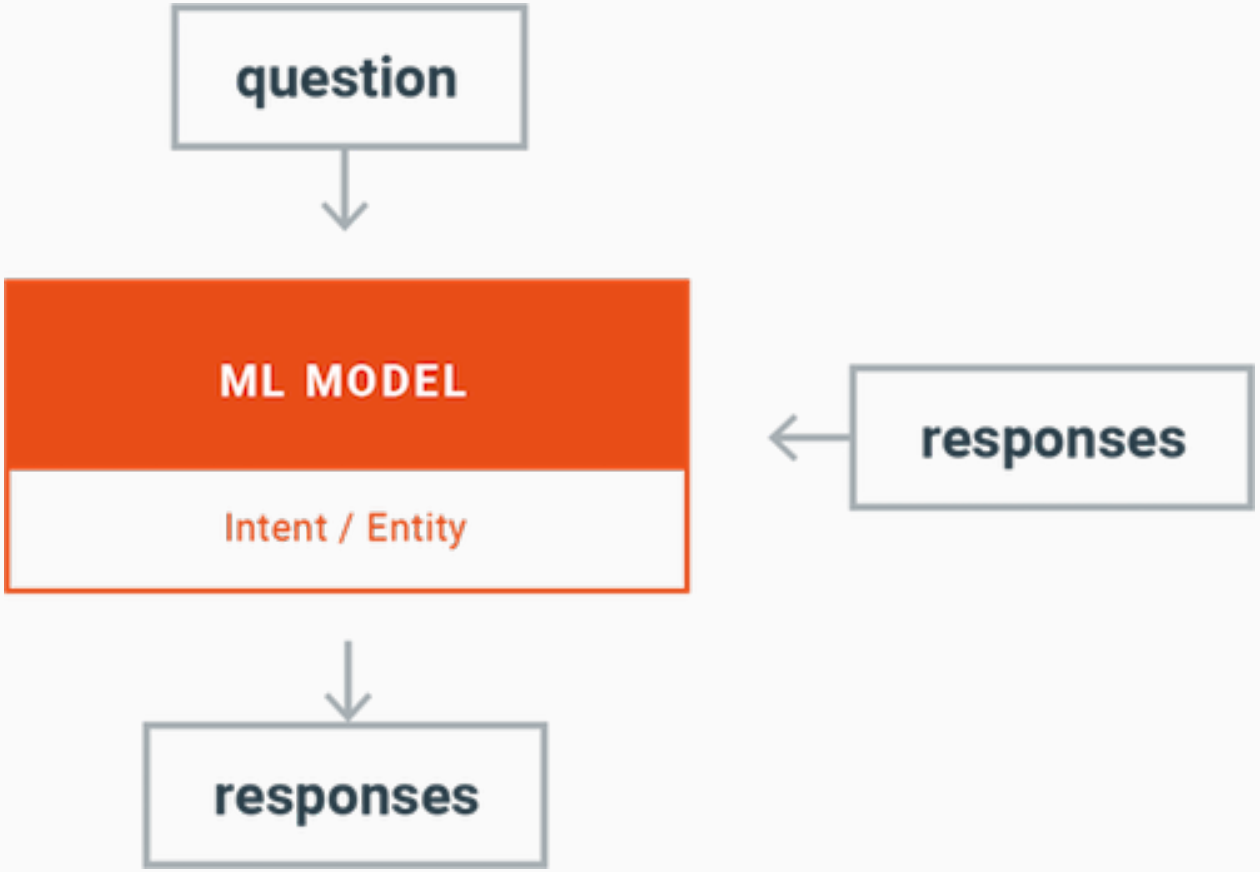
[1]Ultimate Guide to Leveraging NLP and Machine Learning for your Chatbot. Stefan Kojouharov, Chatbots Life 2016

Rule-based – basics

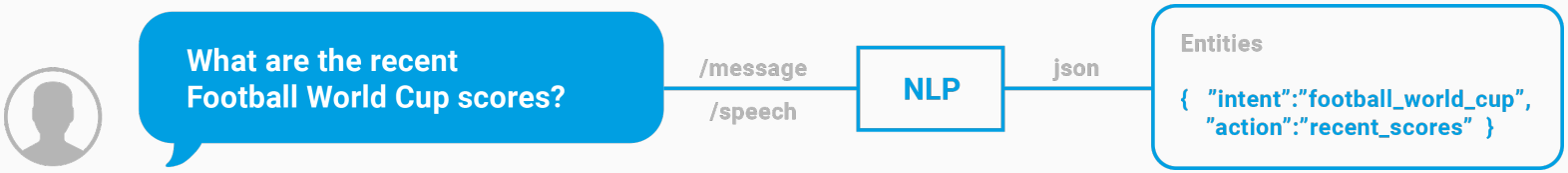
Phrases list	Answers list
Show status of recruitment.	We have currently X candidates.
What is the weather in Berlin?	It is <current weather>.
Hi!	Hi. How are you?
Hire candidate <name>.	Sent an email to <email>.

No valid phrase found

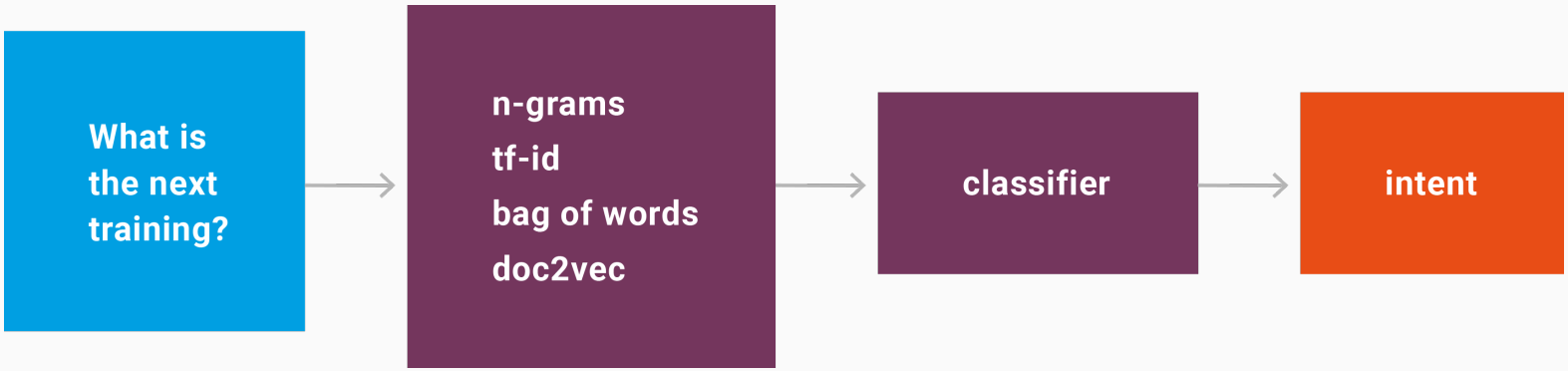
Retrieval-based – basics



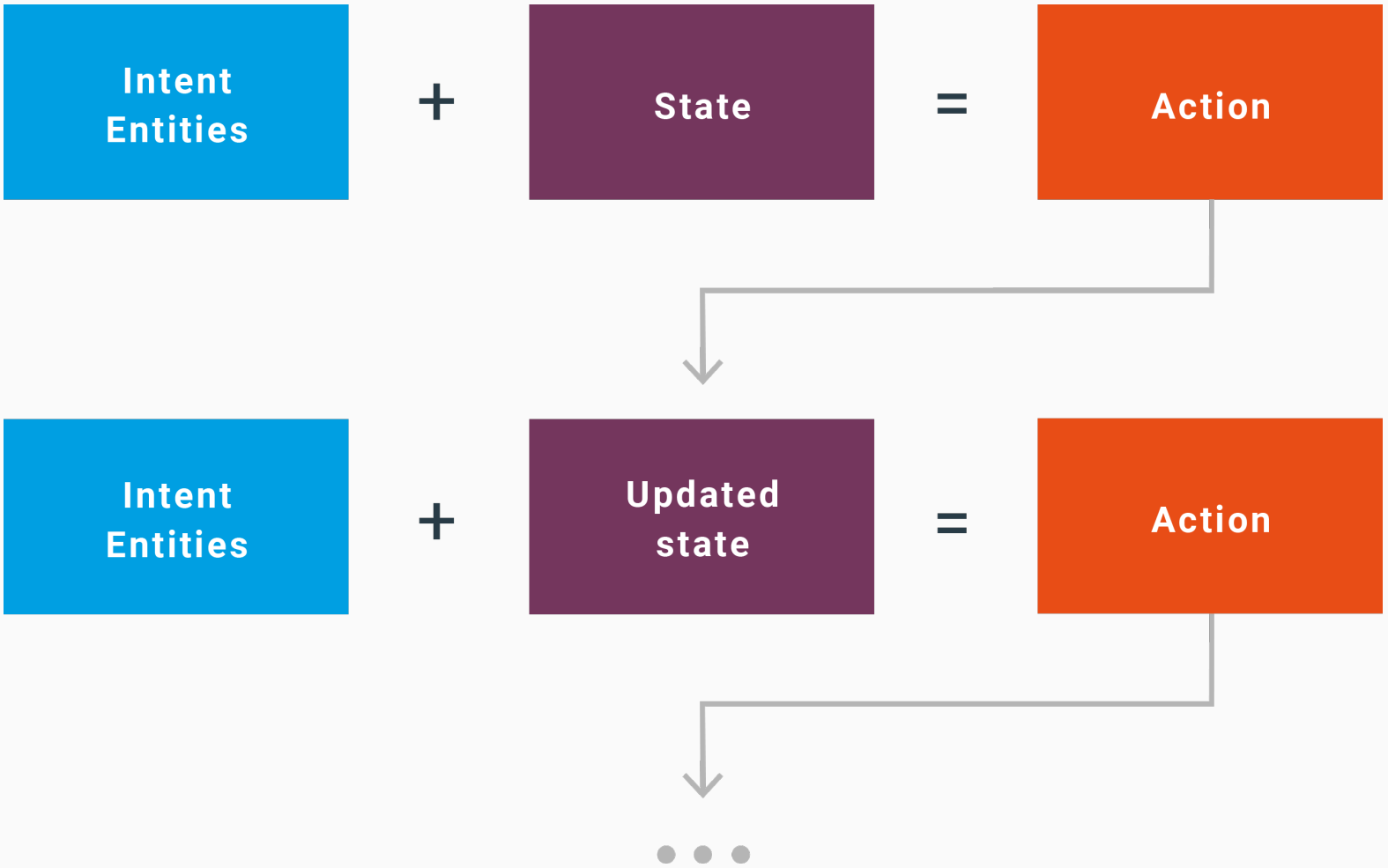
Entities and intents



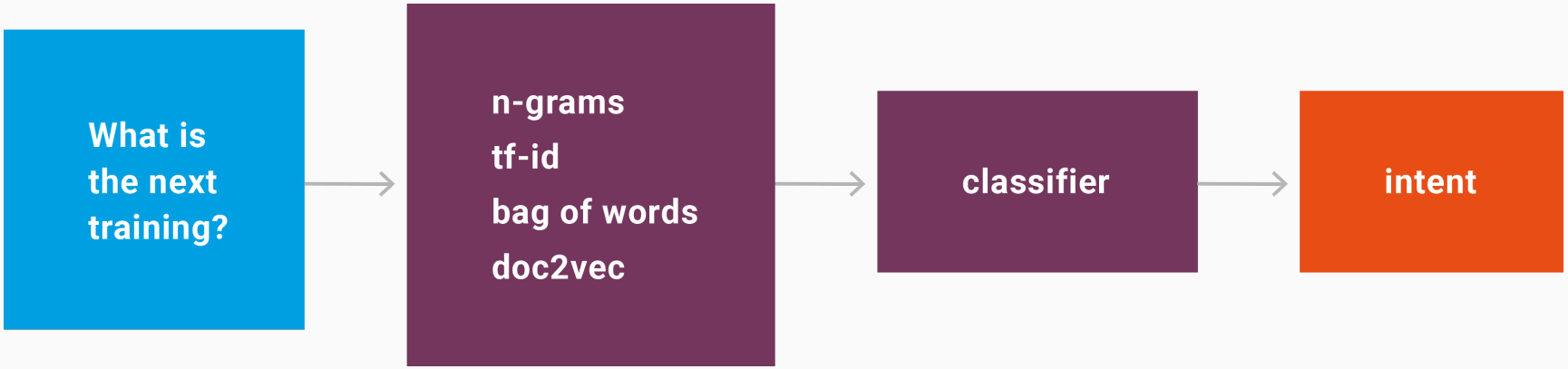
Entities and intents



Rasa NLU engine



Rasa intent learning process



Advantages

Rule-based chatbots:

- predictable,
- clear principles,
- cheap.

Retrieval-based chatbots:

- identify the intent,
- usually easy to train,
- do not need too many questions/answers,
- more intelligent than rule-based.

Generative-based chatbots:

- generic, intelligent answers,
- raw data as training data set.

Bottlenecks

Rule-based chatbots:

- too simple for most cases,
- not really intelligent.

Retrieval-based chatbots:

- limited to questions/answers
- not a generic solution.

Generative-based chatbots:

- usually take longer to train,
- needs a dataset, usually a huge one,
- sometimes unpredictable.

Text generation

NLG

Natural Language Generation is a part of Natural Language Processing. The goal of NLG is to generate a sentence or the whole document that has a logical sense, follows the grammar and answers the question properly if we deal with a bot.

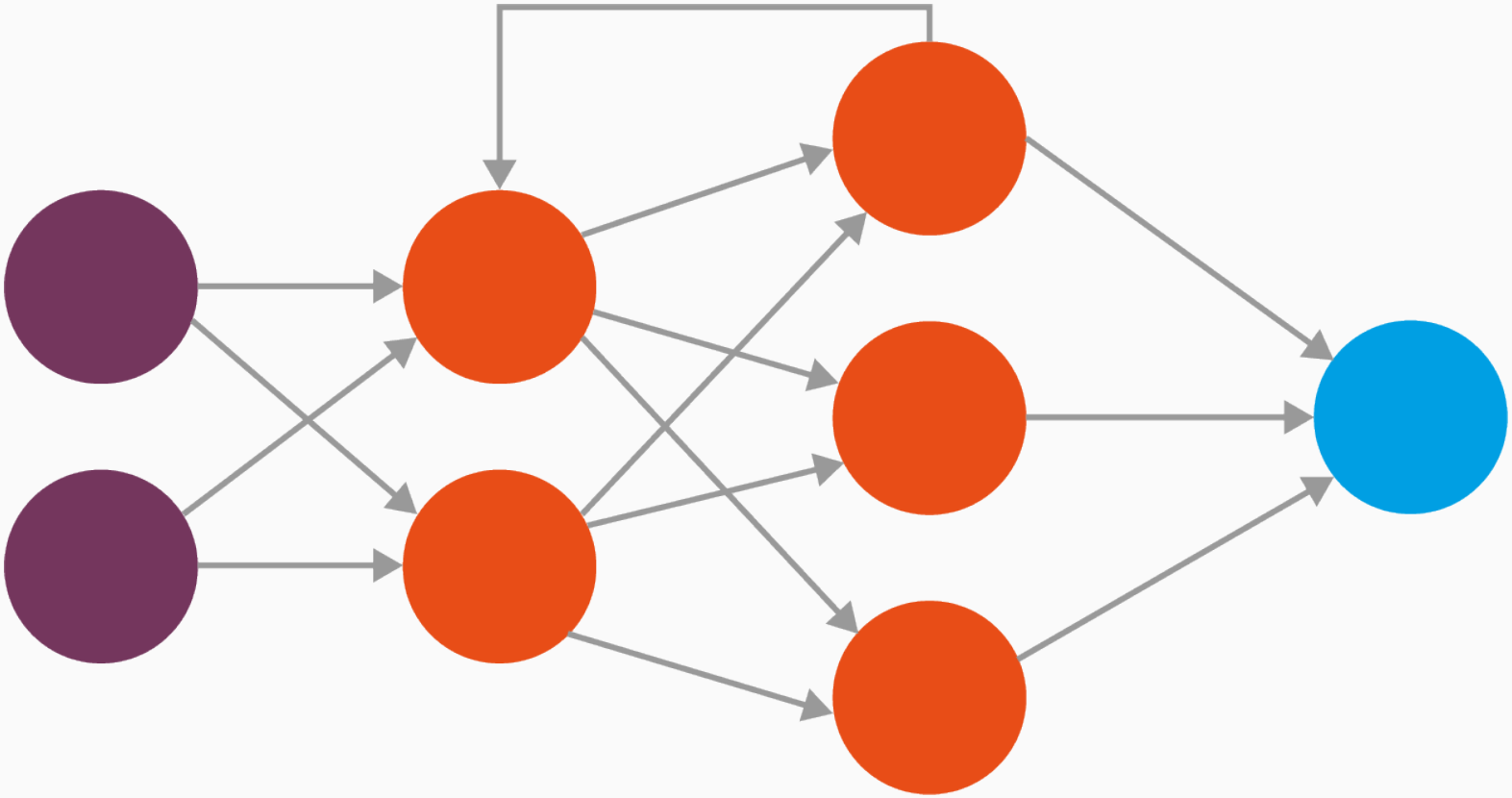
There are plenty of methods that can be used for text generation. The most popular are:

- n-gram model,
- recurrent neural network,
- autoencoders,
- generative adversarial network.

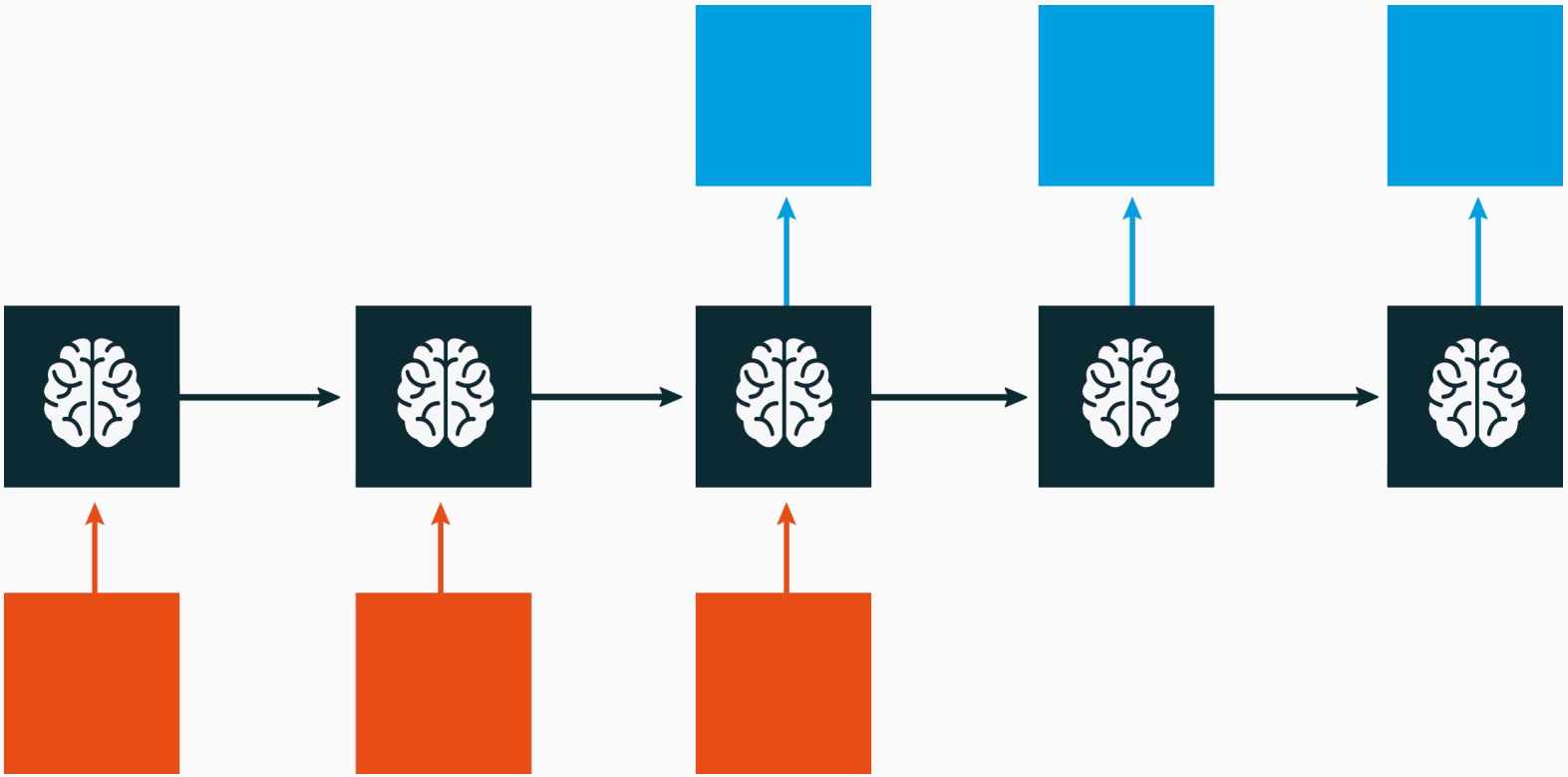
N-gram model


This paper is on natural language processing.
This paper is on natural language processing.
This paper is on natural language processing.
This paper is on natural language processing.
This paper is on natural language processing.

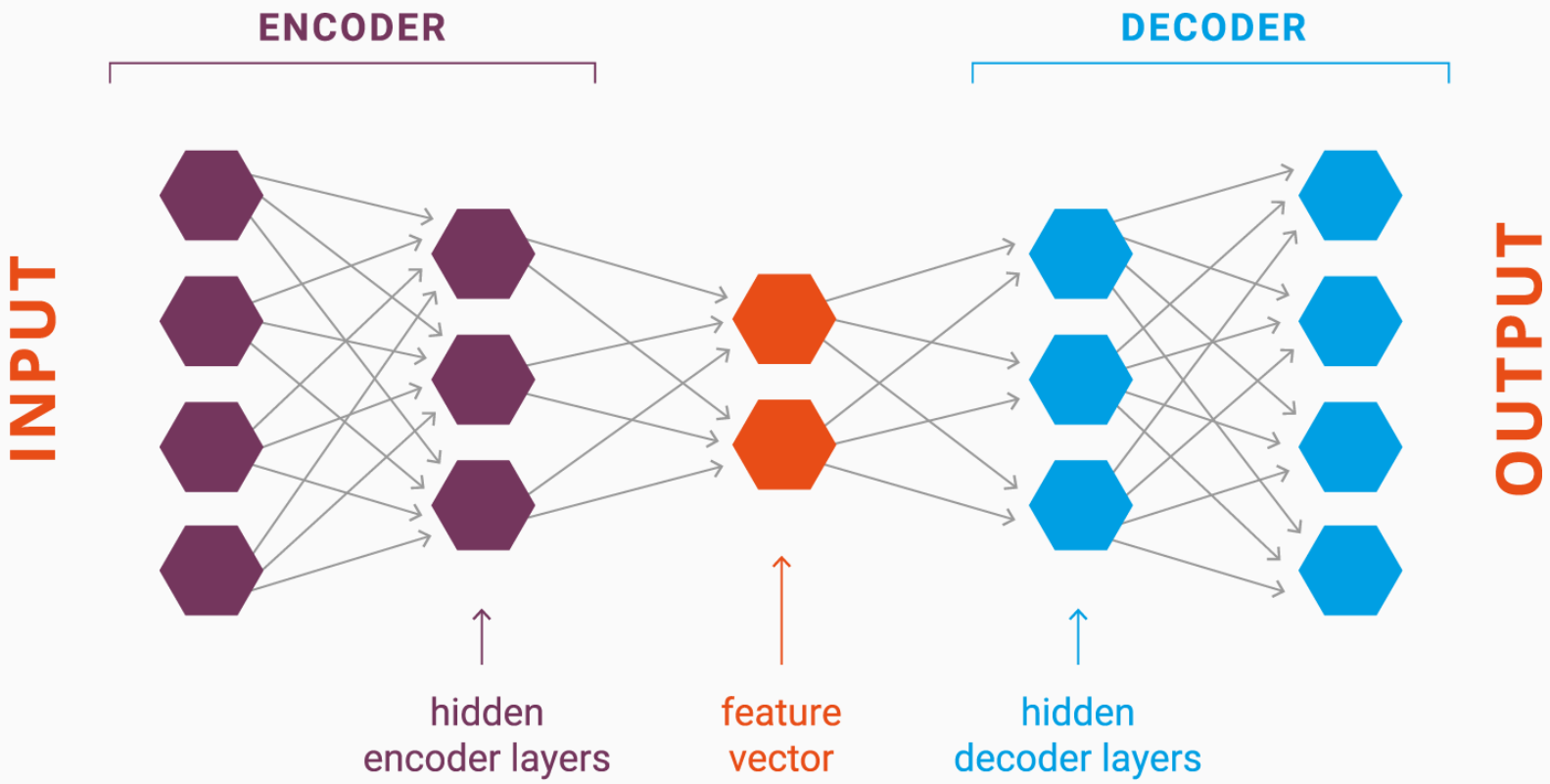
RNN



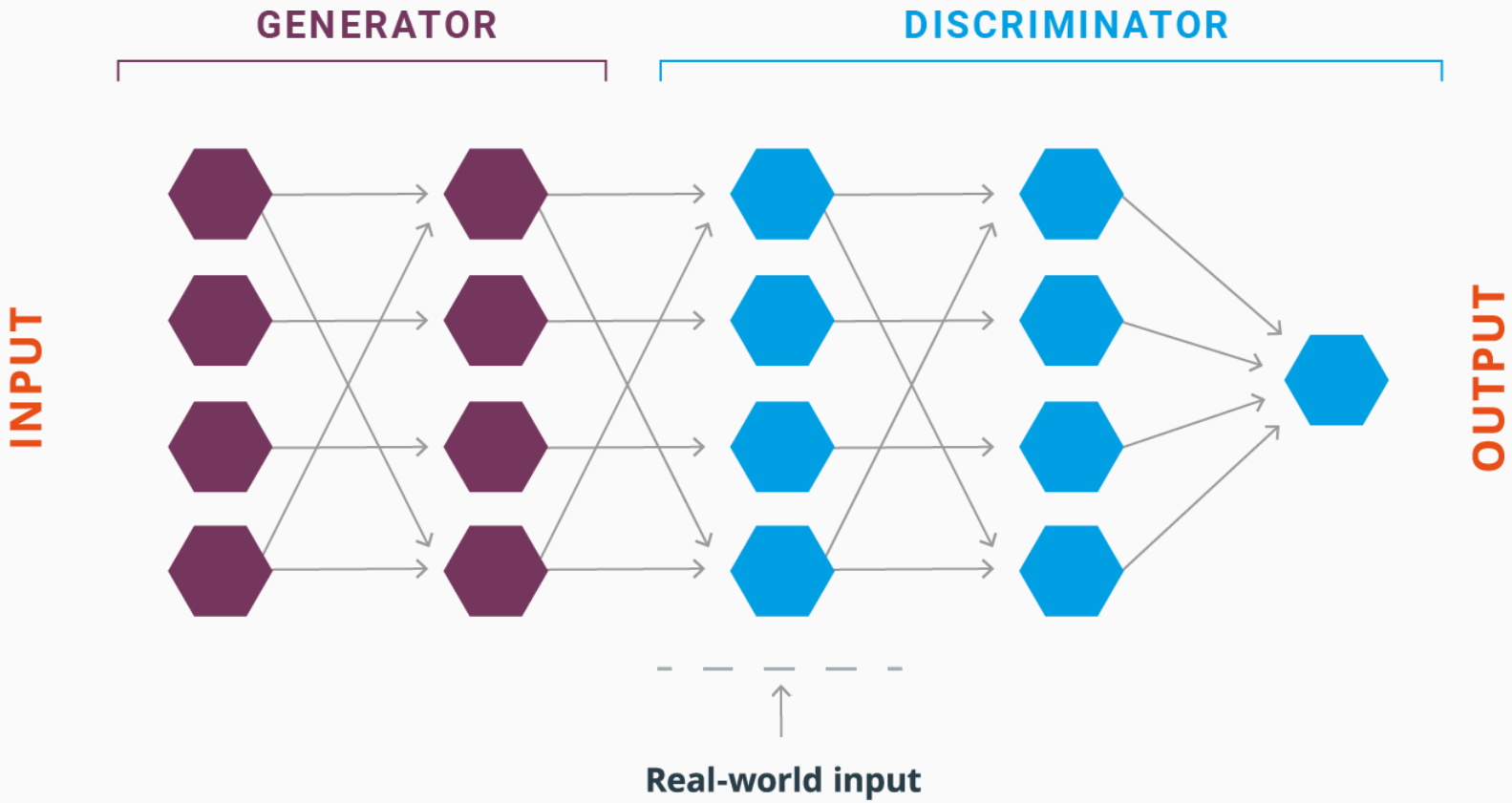
LSTM



Autoencoders



Generative Adversarial Networks



Working examples

There are many open source trained models available. Here are a few worth mentioning:

- chatterbot – a chatbot implementation `http://chatterbot.readthedocs.io/`
- DeepQA – uses RNN and has a web interface `https://github.com/Conchylicultor/DeepQA`
- Generative Conversational Agents – uses LSTM, RNN and GAN `https://github.com/oswaldoludwig/Adversarial-Learning-for-Generative-Conversational-Agents`.

Research datasets

A few datasets useful for your research:

- **SQuAD** – reading comprehension dataset, consists of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text, `https://rajpurkar.github.io/SQuAD-explorer/`,
- **Cornell Movie Dialogs Corpus** – movie dialogs, `https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html`,
- **DeepMind datasets** – AQua is a dataset of questions and answers, `https://github.com/deepmind/AQuA`, more datasets from DeepMind: `https://deepmind.com/research/open-source/open-source-datasets/`,
- **DMQA** – Daily Mail and CNN articles data sets, `https://cs.nyu.edu/~kcho/DMQA/`,
- **MS MARCO** – Microsoft MAchine Reading COmprehension Dataset, `http://www.msmarco.org/dataset.aspx`.

Quality metrics

Quality metrics for NLP

In most cases we use the same quality metrics as in regular machine learning models, if any machine learning model is used for NLP.

Other methods that can be used for NLP methods validation, especially, text generation:

- spell check,
- grammer check,
- typo check.

Questions?