Introduction to Natural Language Processing

Ali Buğra Kanburoğlu

16.02.2022

Ali Buğra Kanburoğlu

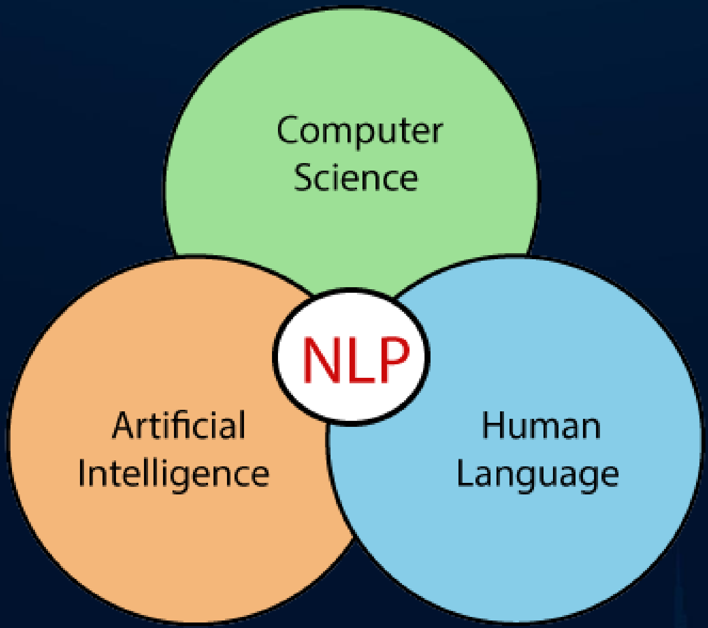
Senior Research Engineer at Huawei

Content

1. What is Artificial Intelligence?
2. What is Natural Language Processing?
3. From Linguistics to Natural Language Processing
4. Brief History of NLP
5. NLP Core Terminologies
6. Common NLP Tasks
7. What does an NLP Engineer do?
8. NLP in Huawei

What is Artificial Intelligence?

Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems.

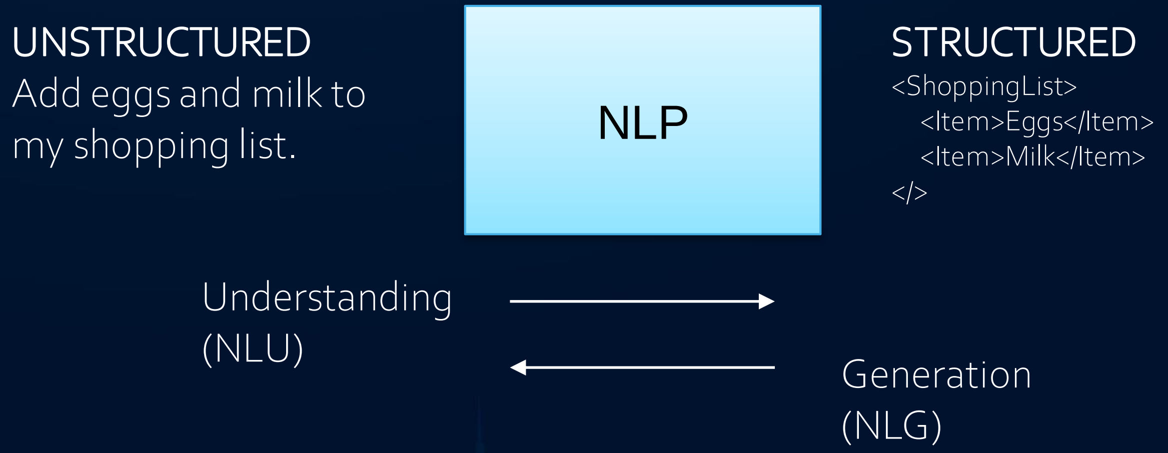


What is Natural Language Processing?

Natural Language Processing, or NLP for short, is a field in Artificial Intelligence (AI) devoted to creating computers that use natural language as input and/or output. (i.e., through text and speech)

Natural Language

- Signs
- Menus
- Email
- SMS
- Web Pages and so much more…



UNSTRUCTURED

Add eggs and milk to my shopping list.

STRUCTURED

```
<ShoppingList>
   <Item>Eggs</Item>
   <Item>Milk</Item>
</>
```

Understanding (NLU)

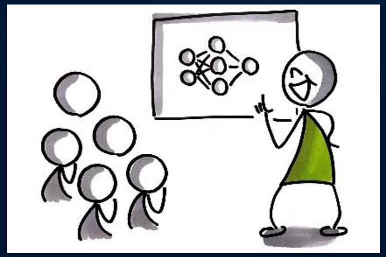
Generation (NLG)

From Linguistics to Natural Language Processing

Linguistics

- Scientific study of language
- Grammar
- Semantics
- Phonetics

Computational Linguistics



- Modern study of the linguistics
- Uses tools of computer science
- Testing of grammars

Natural Language Processing

- Works with text data
- Natural language understanding
- Natural language generation

6

Brief History of NLP

- In **1950**, Alan Turing published "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence.
- In **1954**, Georgetown experiment involved translation of Russian sentences into English.
- In **1957**, Noam Chomsky's Syntactic Structures revolutionized Linguistics with "universal grammar".
- In **1964**, ELIZA which is one of the first chatterbots created at MIT CSAIL.
- In **1980s**, ML algorithms introduced for natural language processing.
- In **1990s**, Early successes on statistical methods in NLP occurred in the field of machine translation.
- In **2000s**, Use of supervised and unsupervised learning algorithms.
- In **2010s and Today**, Representation learning and deep learning methods started implementing in NLP.
  - Siri,
  - Neural language models,
  - Word embeddings,
  - Seq-to-seq learning,
  - Attention,
  - Pre-trained models, etc.

NLP Core Terminologies

- Corpus
- Tokenization
- Stemming & Lemmatization
- Syntax & Semantics
- Parsing

8

Corpus

- A corpus is a collection of text
  - Newspaper, recipes, tweets
  - Often annotated in some way
  - Sometimes just lots of text
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged text
  - The Web: billions of words of who knows what
- What makes a Corpus **better**?

Corpus

- A corpus is a collection of text
  - Newspaper, recipes, tweets
  - Often annotated in some way
  - Sometimes just lots of text

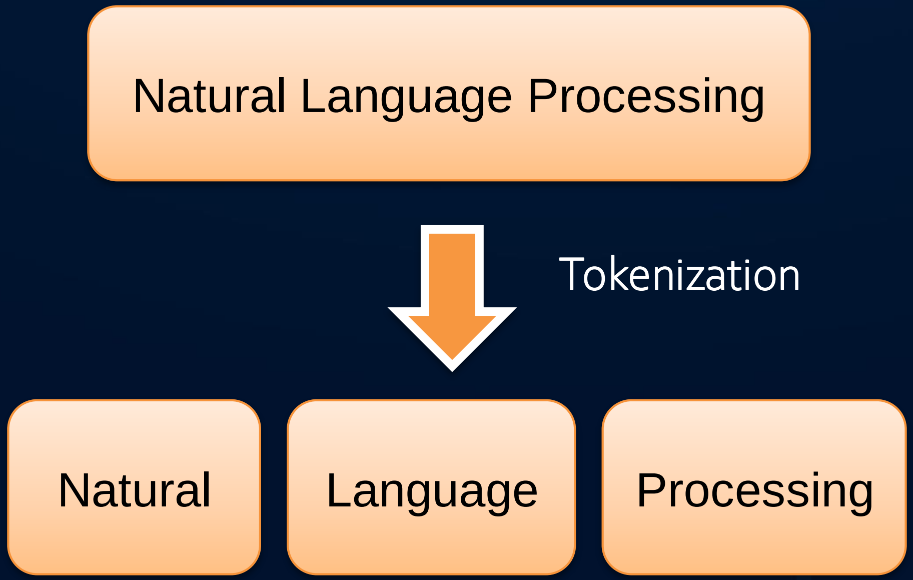
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged text
  - The Web: billions of words of who knows what

- What makes a Corpus **better**?
  - *Large, high quality, clean data, balanced*

10

Tokenization

- Tokenization is the process of tokenizing or splitting a string, text into a list of tokens.
  - Text into sentences
  - Sentences into words
  - Words into characters



Natural Language Processing

Tokenization

Natural

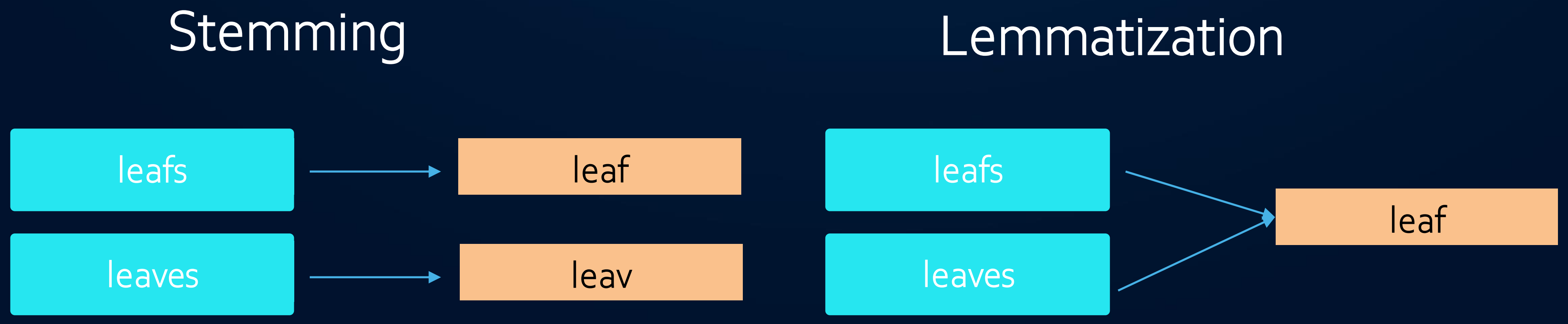
Language

Processing

11

Stemming & Lemmatization

- Stemming reduces **word-forms to stems** whereas lemmatization reduces the **word-forms to lemmas** (morphological stems).



Stemming

leafs → leaf

leaves → leav

Lemmatization

leafs → leaf

leaves → leaf

12

Syntax & Semantics

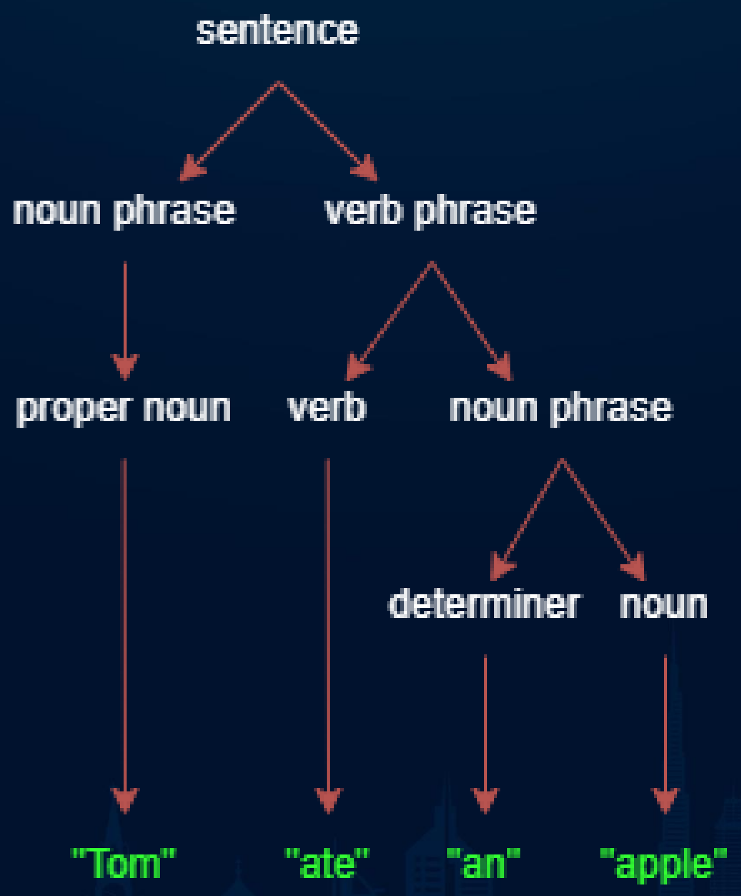
- Syntax concerns the proper **ordering of words** and its affect on meaning.
  - The dog bit the boy.
  - The boy bit the dog.
  - Bit boy dog the the.
- Semantics concerns the literal **meaning of words**, phrases and sentences.
  - "plant" as a photosyntetic organism
  - "plant" as a manufacturing facility
  - "plant" as the act of sowing

13

Parsing

- Parsing is the problem of constructing a **derivation tree** for an **input string** from a formal definition of a grammar.

- INPUT: Tom ate an apple



14

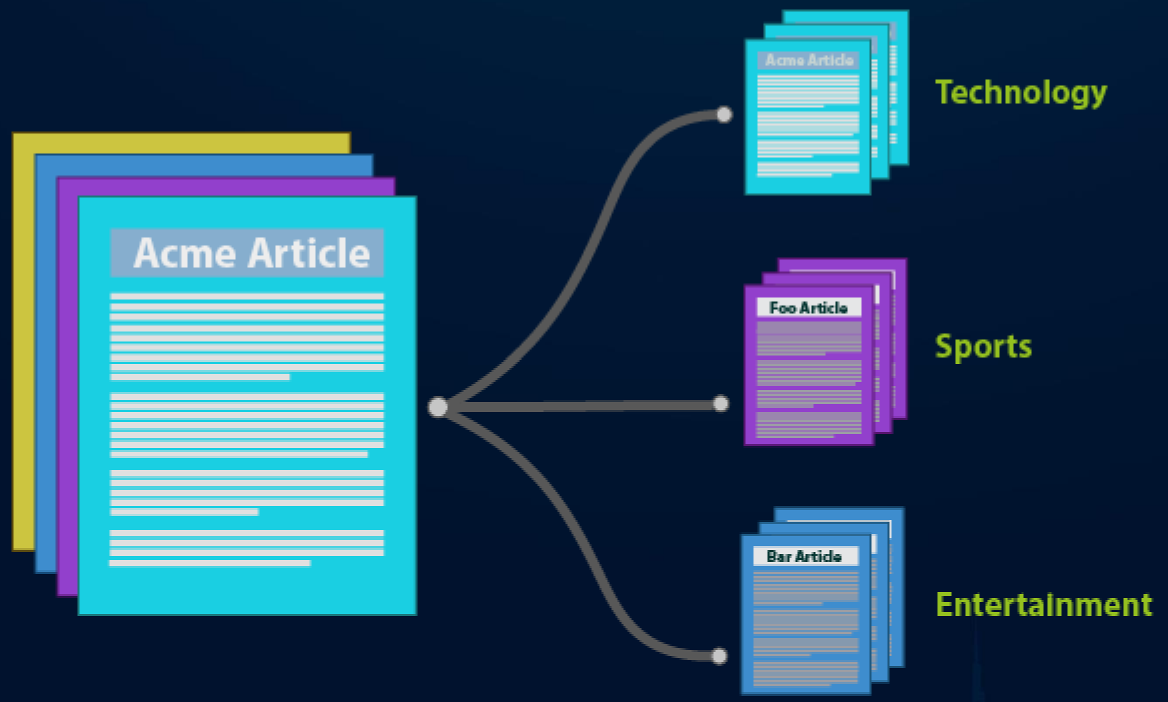
Common NLP Tasks

- Text Classification
- Sentiment Analysis
- Named Entity Recognition (NER)
- Morphological Analysis
- Part-of-Speech (POS) Tagging
- Question Answering
- Machine Translation

15

Text Classification

- Text classification is the process of assigning a labeled category, known as a class, to text.



16

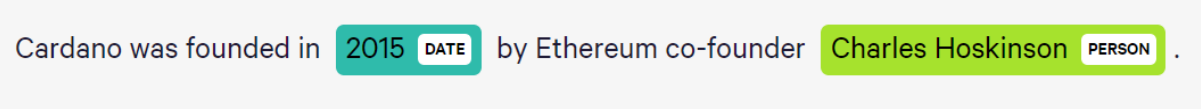
Sentiment Analysis

- Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments and emotions through opinions expressed in written texts.

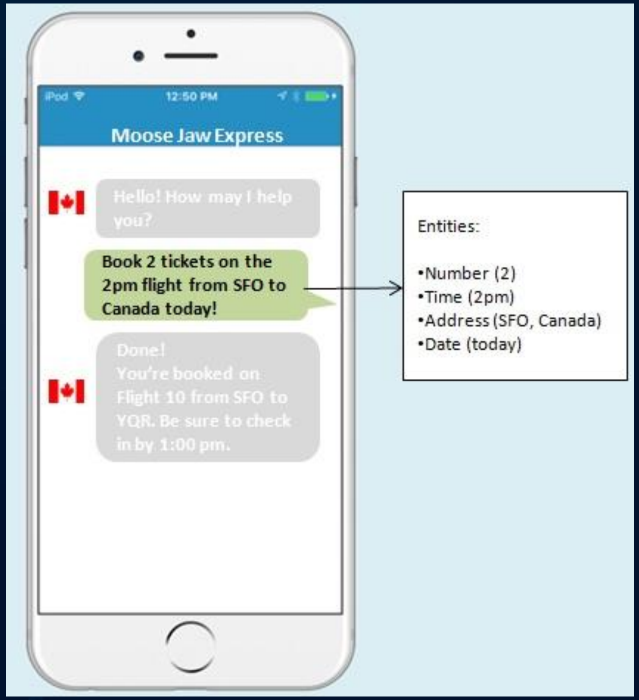
17

Named Entity Recognition

- Named Entity Recognition (NER) is the task which find names in a text such as; **people, organizations, locations, dates.**



Cardano was founded in 2015 DATE by Ethereum co-founder Charles Hoskinson PERSON .



Hello! How may I help you?

Book 2 tickets on the 2pm flight from SFO to Canada today!

Done!
You're booked on Flight 10 from SFO to YQR. Be sure to check in by 1:00 pm.

Entities:

- Number (2)
- Time (2pm)
- Address (SFO, Canada)
- Date (today)

Morphological Analysis

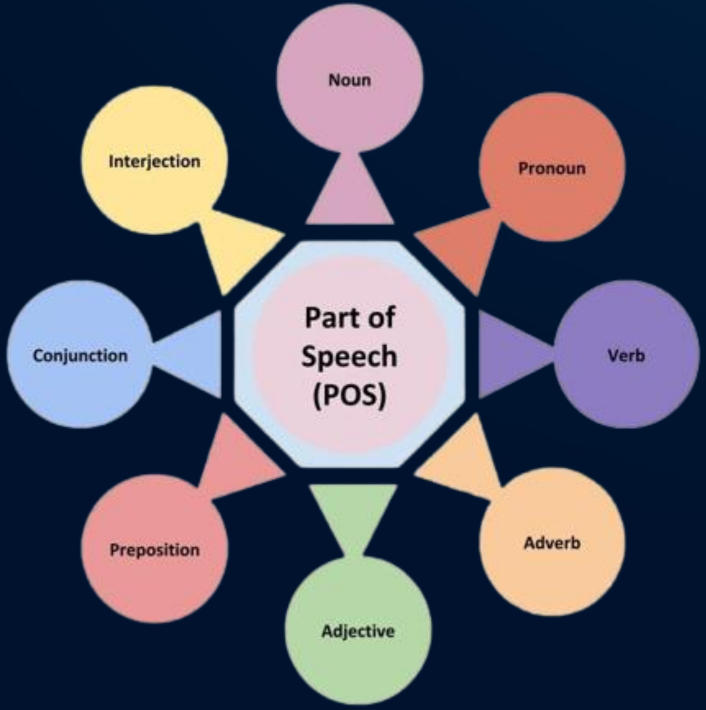
- Morphological analysis is the task of **segmenting a word** into its **morphemes**.
  - A **morpheme** is the smallest linguistic unit that has semantic meaning.

For example, the word **"books"** can be divided into two morphemes; ***'book'*** and ***'s'***, where the meaning of **'s'** is as a ***plural suffix***.

19

Part of Speech (POS) Tagging

- Part-of-Speech (POS) tagging is the process of assigning one of the parts of speech to the given word.

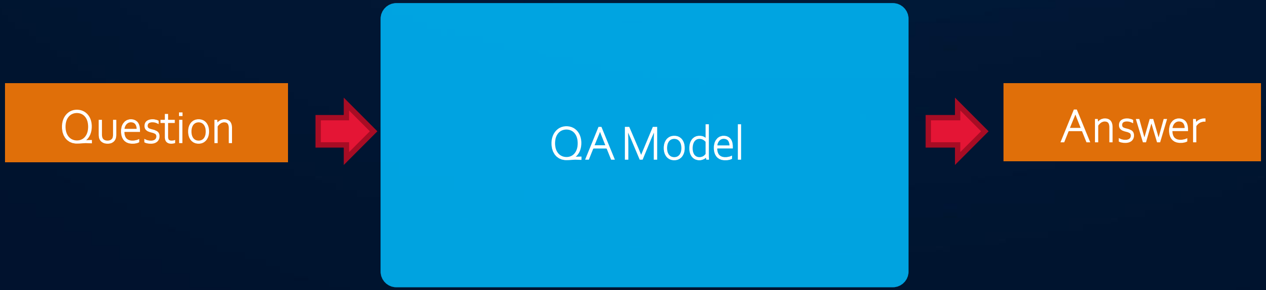


I (PRON) want (VERB) an (DET) apple (NOUN)

20

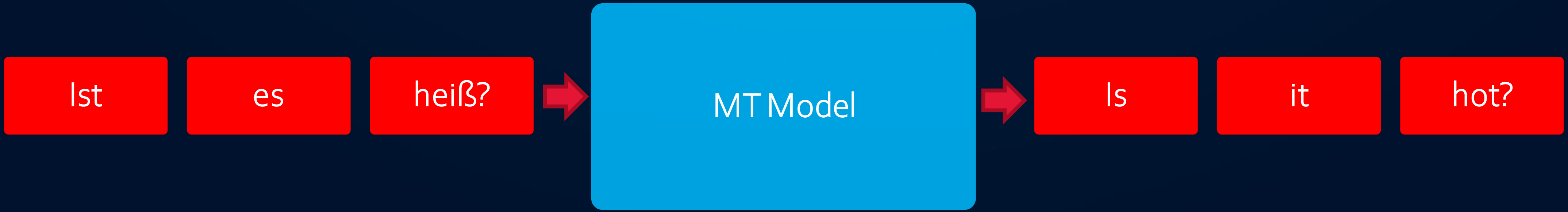
Question Answering

- In Question Answering tasks, the model receives a **question** regarding text content and is required to mark the beginning and end of the **answer** in the text.



Machine Translation

- Machine Translation (MT) is the task of automatically **converting one natural language into another**, **preserving the meaning** of the input text, and **producing fluent text** in the output language.



Ist es heiß? → MT Model → Is it hot?

22

What does an NLP Engineer do?

Responsibilities

- Design and develop NLP systems
- Define appropriate datasets for language learning

Skills

- Deep understanding of text representation techniques
- Experience with machine learning frameworks and libraries
- An analytical mind with problem-solving abilities

Tasks

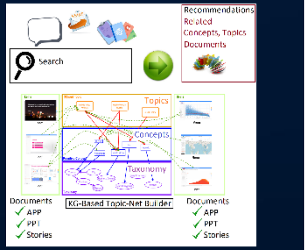
- Use effective text representations to transform natural language into useful features
- Develop NLP systems according to requirements
- Train the developed model and run evaluation experiments
- Find and implement the right algorithms and tools for NLP tasks

23

NLP in Huawei

Huawei App Gallery Search

Intelligent Operation Systems



Recommendations
Related
Concepts, Topics
Documents

Search

Topics

Concepts

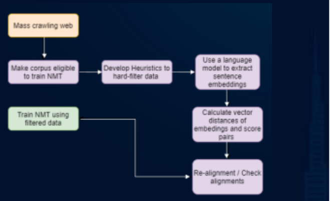
Taxonomy

KG-Based Topic-Net Builder

Documents
✓ APP
✓ PPT
✓ Stories

Documents
✓ APP
✓ PPT
✓ Stories

Neural Machine Translation



Mass crawling web

Make corpus eligible to train NMT

Develop Heuristics to hard-filter data

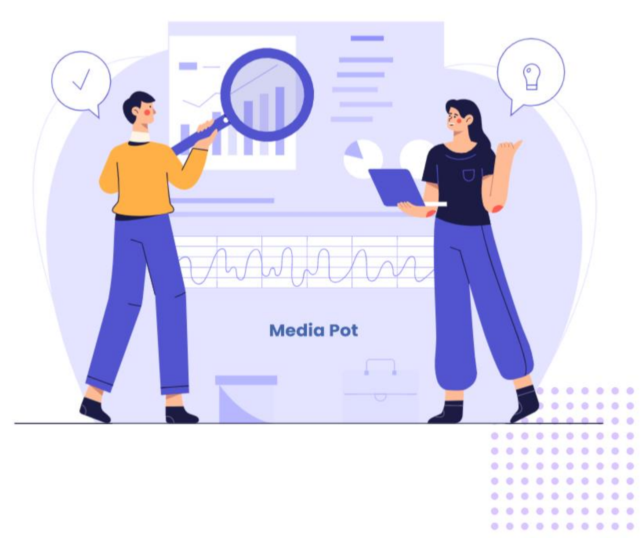
Use a language model to extract sentence embeddings

Calculate vector distances of embedings and score pairs

Re-alignment / Check alignments

Train NMT using filtered data

Media Pot



Resource: https://mediapot.net

Huawei App Gallery Search

- Turkey Team is responsible for Russia, Africa, Asia-Pacific, Latin America regions.

- Research topics
  - Learning to rank
  - Natural language processing
    - Unsupervised keyword extraction & query enrichment
    - Query category prediction
    - Text classification
    - Language identification

https://appgallery.huawei.com

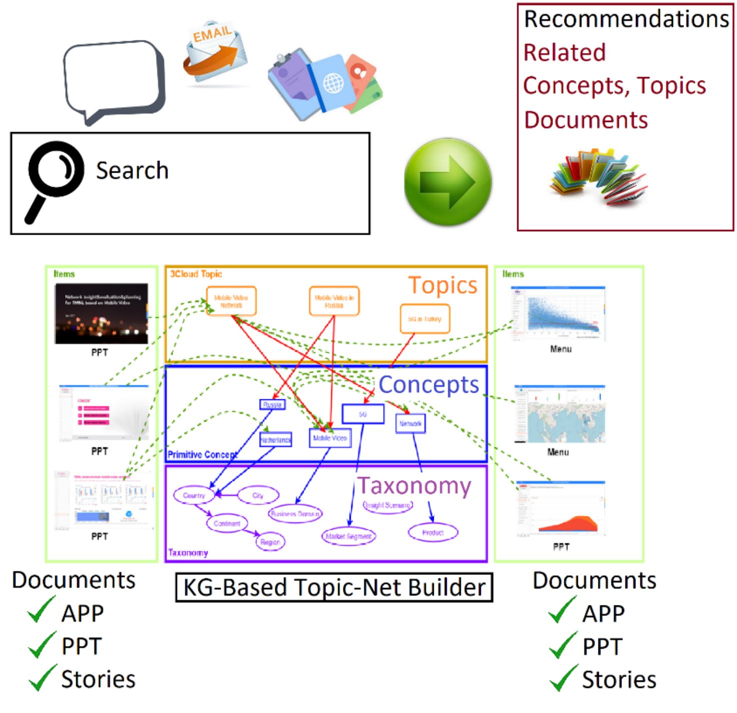
25

Intelligent Operation Systems

- This system is used to keep data in structured and retrieval related business concepts, business topics and presentation materials when documents, e-mails, or a simple natural text is used as the query.

- Our system will help marketing experts to search solution stories and presentation materials quickly.

- KG-based search systems enables 3Clouds users to do exclusive search for materials in the Topic-Net. During view of a specific content, search for related topics are also enabled. The KG-pipeline provides user-friendly experience for the system users.



Search

Recommendations
Related
Concepts, Topics
Documents

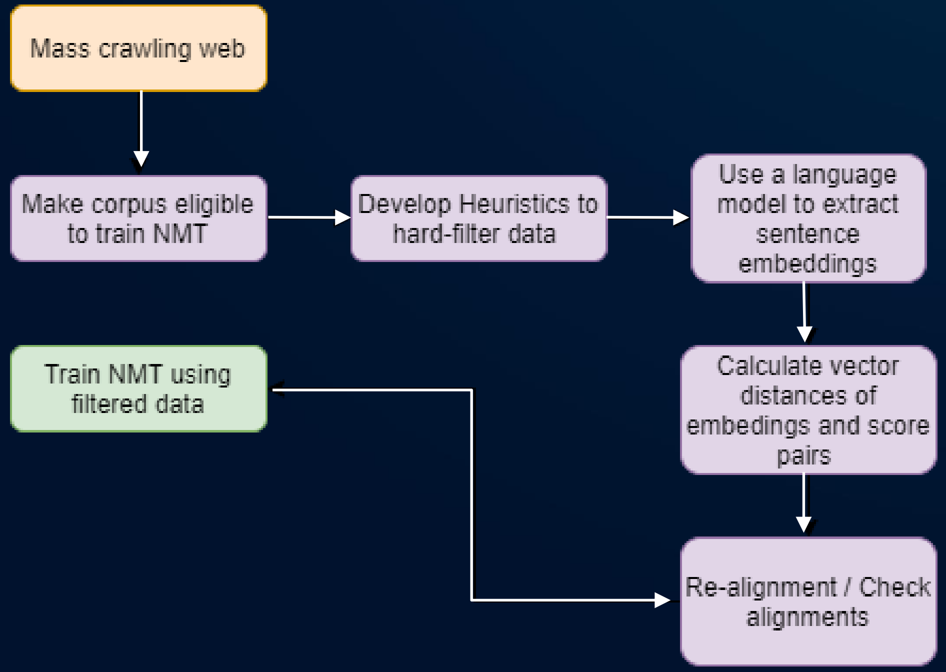
Documents
✓ APP
✓ PPT
✓ Stories

KG-Based Topic-Net Builder

Documents
✓ APP
✓ PPT
✓ Stories

26

Neural Machine Translation: Parallel corpus filtering

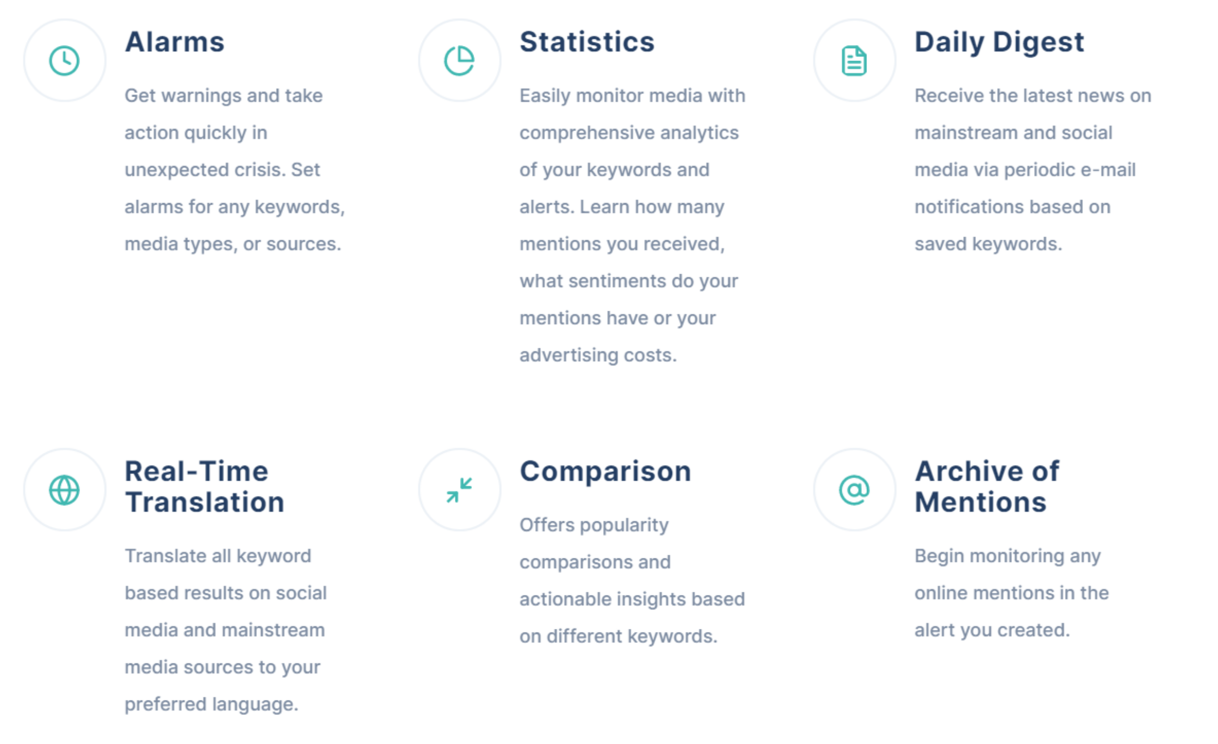

Mass crawling web
Make corpus eligible to train NMT
Develop Heuristics to hard-filter data
Use a language model to extract sentence embeddings
Calculate vector distances of embedings and score pairs
Re-alignment / Check alignments
Train NMT using filtered data

- Parallel corpus filtering is essentially a pre-processing step to build datasets to train neural machine translation systems with high accuracy.
- Focus is on low-resource languages such as Pashtu and Kimeri.
- We used BERT language model to embed sentences and calculated distance on those embeddings. The smaller the distance, the higher the chances to end with a good translation pair.
- Published our work to EMNLP contest, paper can be found in: Acarcicek, Haluk, et al. "Filtering Noisy Parallel Corpus using Transformers with Proxy Task Learning." Proceedings of the Fifth Conference on Machine Translation, 2020.

27

Media Pot

- Media Pot is **an AI powered media monitoring and listening platform** that allows you to easily follow **trends, moods or competitors** by analyzing the flow of information and emotions in the mainstream media and social media in real time.



Alarms

Get warnings and take action quickly in unexpected crisis. Set alarms for any keywords, media types, or sources.

Statistics

Easily monitor media with comprehensive analytics of your keywords and alerts. Learn how many mentions you received, what sentiments do your mentions have or your advertising costs.

Daily Digest

Receive the latest news on mainstream and social media via periodic e-mail notifications based on saved keywords.

Real-Time Translation

Translate all keyword based results on social media and mainstream media sources to your preferred language.

Comparison

Offers popularity comparisons and actionable insights based on different keywords.

Archive of Mentions

Begin monitoring any online mentions in the alert you created.

28

Huawei NLP Training

BTK Academy NLP Training:

29

Thank you.

ali.bugra.kanburoglu@huawei.com

Bring digital to every person, home and organization for a fully connected, intelligent world.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI