Welcome!

9 00 – 9 05  Introduction
9 05 – 9 30  Information extraction
9 30 – 9 40  Breakout 1
9 40 – 9 50  Summary
9 50 – 10 00  Coffee break
10 00 – 10 40  Attention and Transformers
10 40 – 10 55  Breakout 2
10 55 – 11 00  Wrap up
Accelerating the use of AI for the benefit of our society, our competitiveness, and for everyone living in Sweden
Deep dive sessions

Support by targeting key roles

Provide a neutral arena where you can meet and discuss similar challenges

An outside-in perspective, inspiration & meet with other in the same situation
Data Scientist or similar function

Role description:
A Data Scientist needs to be able to:

• Autonomously and flexibly carry out advanced analytics in multiple domains and environments.
• Run advanced modeling on data in order to extract knowledge and/or predict future events.
• Autonomously create and develop concepts.
• Understand the complete analytics chain from storing, structuring, modeling, to visualizing and translating data into actionable insights.
• Review, advice and communicate in analytics use cases.
Attention, the heart of transformers, to improve Information extraction

Severine Verlinden
AI Developer, Language Technology
AI Sweden
Breakout session 1

Questions

• Which problems can be solved with Information Extraction in your area of work?
• Main points of the discussion
Please fill the survey:

https://docs.google.com/forms/d/e/1FAIpQLSeJcFpnNyL3362G152O8qR5NqITJQjPCK7fq58g/viewform
Breakout session 2

Questions

• How do you think Attention or Transformers can be used in your area of work?

• How would you want to use language models?
Next step

• The General Survey
• Networking on Slack
We need your feedback!

https://www.surveymonkey.com/r/MGBT6HR
Thank you!
Deep Dive
Transformers

youtube.com/c/oisweden
linkedin.com/company/oisweden
What is Information Extraction?

Read more on this topic:

Thread: I know there's a lot of scary stuff in the world rn, but this is something I've been thinking about that I can't get out of my head.

9:56 PM · 01 Nov 17

8,240 Retweets 17K Likes

Kumail Nanjiani @kumailn
As a cast member on a show about tech, our job involves visiting tech companies/conferences etc. We meet ppl eager to show off new tech.

Kumail Nanjiani @kumailn
Often we'll see tech that is scary. I don't mean cyber threats. I mean disasters.
Thread: I know there's a lot of scary stuff in the world rn, but this is something I've been thinking about that I can't get out of my head.

9:55 PM - 01 Nov 17

8,240 Retweets 17K Likes

Kumail Nanjiani @kumailn

As a co-host member of a show about tech, our job involves visiting tech companies/conferences etc. We meet ppl eager to show off new tech.

Kumail Nanjiani @kumailn

Often we'll see tech that is scary. I don't mean zeros and ones. I mean actual people...
Today ...

Thread: I know there's a lot of scary stuff in the world rn, but this is something I've been thinking about that I can't get out of my head.
9:56 PM - 01 Nov 17
8,240 Retweets 17K Likes

Kumail Nanjiani @kumailn
As a cast member on a show about tech, our job entails visiting tech companies/conferences etc. We meet ppl eager to show off new tech.

Kumail Nanjiani @kumailn
Often we'll see tech that is scary. I don't mean disaster, I mean chicken.

Hi there,
I found some cool new hiking trails that are not too far from here. One of them is fairly close to your place and supposedly has a spring running through. Want to check it out?
Today ...

Thread: I know there's a lot of scary stuff in the world rn, but this is something I've been thinking about that I can't get out of my head.

As a cast member on a show about tech, our job is to schedule tech companies/conferences etc. We meet so eager to show off new tech.

I found some cool new hiking trails that are not too far from here. One of them is fairly close to your place and supposedly has a spring running through. Want to check it out?
Today ...

Thread: I know there's a lot of scary stuff in the world rn, but this is something I've been thinking about that I can't get out of my head.

Hi there,

I found some cool new hiking trails that are not too far from here. One of them is fairly close to your place and supposedly has a spring running through. Want to check it out?
Information Extraction =

- Process of extracting relevant information from text or data sources.
- Typically involves natural language processing techniques.
- Aim is to automate the identification of specific data from unstructured or semi-structured formats.

Diagram illustration:
- Documents or data sources are fed into a funnel.
- The funnel represents the process of information extraction.
- The output is structured data, blocks, or a checklist, indicating validation or correct extraction.
Concrete example of information extraction

"We are very excited for Harry and Meghan. It has been wonderful getting to know Meghan and to see how happy she and Harry are together," Clarence House said in a tweet.

- Named Entity recognition: Mentions: Harry, Meghan, Clarence House...
- Coreference Resolution: Cluster: (Harry, Harry), (Meghan, Meghan), ...
- Relation Extraction: Relation: Meghan $\rightarrow$ in relation with $\rightarrow$ Harry
Concrete example of information extraction

"We are very excited for Harry and Meghan. It has been wonderful getting to know Meghan and to see how happy she and Harry are together," Clarence House said in a tweet.

- Named Entity recognition
  - Mentions: Harry, Meghan, Clarence House...

- Coreference Resolution
  - Cluster: (Harry, Harry), (Meghan, Meghan), ...

- Relation Extraction
  - Relation: Meghan--> in relation with -->Harry
Concrete example of information extraction

"We are very excited for Harry and Meghan. It has been wonderful getting to know Meghan and to see how happy she and Harry are together," Clarence House said in a tweet.

- **Named Entity recognition**
  - Mentions: Harry, Meghan, Clarence House...

- **Coreference Resolution**
  - Cluster: (Harry, Harry), (Meghan, Meghan), ...

- **Relation Extraction**
  - Relation: Meghan→ in relation with →Harry
Concrete example of information extraction

"We are very excited for Harry and Meghan. It has been wonderful getting to know [Meghan] and to see how happy she and [Harry] are together," Clarence House said in a tweet.

• Named Entity recognition
  Mentions: Harry, Meghan, Clarence House...

• Coreference Resolution
  Cluster: (Harry, Harry), (Meghan, Meghan), ...

• Relation Extraction
  Relation: Meghan→ in relation with ? → Harry
"We are very excited for Harry and Meghan. It has been wonderful getting to know Meghan and to see how happy she and Harry are together," Clarence House said in a tweet.
"We are very excited for Harry and Meghan. It has been wonderful getting to know Meghan and to see how happy she and Harry are together," Clarence House said in a tweet.
Concrete example of information extraction

"We are very excited for Harry and Meghan. It has been wonderful getting to know Meghan and to see how happy she and Harry are together," Clarence House said in a tweet.

=> Wikipedia, Wikidata
Problem Statement

Improvement of the current IE algorithm

Including external knowledge of a knowledge base.
- Wikipedia
- Wikidata
<table>
<thead>
<tr>
<th>WORD</th>
<th>EMBEDDING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woman</td>
<td>0</td>
</tr>
<tr>
<td>Queen</td>
<td>0</td>
</tr>
<tr>
<td>Apple</td>
<td>0</td>
</tr>
<tr>
<td>Man</td>
<td>0</td>
</tr>
<tr>
<td>King</td>
<td>1</td>
</tr>
<tr>
<td>Orange</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Royal</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woman</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>Queen</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>Apple</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Man</td>
<td>-1</td>
<td>0.01</td>
</tr>
<tr>
<td>King</td>
<td>-0.95</td>
<td>0.93</td>
</tr>
<tr>
<td>Orange</td>
<td>0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

$|V| \times 1$ $|V| \times k$
WORD EMBEDDING

**CBOW** (continuous bag-of-words) objective:
- average
- learning
- never
- the
- mind
- exhasts

**skip-gram** objective:
- learning
- never
- the
- mind
- exhasts
- "context word"
- "target words"
WORD EMBEDDING

learning
the
... never
... mind

|V| X k

|V| X 1

exhausts

average

CBOW (continuous bag-of-words) objective
How? (*)

- NER
- RE
- CR

=> End-to-end
=> Jointly
=> Document-level

How? (*)

- NER
- RE
- CR

=> End-to-end
=> Jointly
=> Document-level

"We are very excited for Harry and Meghan..."

Contextualised Embedding

[0.67, 0.9, 0.11, 0.01,...]

How? (*)

- NER
- RE
- CR

=> End-to-end
=> Jointly
=> Document-level

"We are very excited for Harry and Meghan...

Contextualised Embedding

[0.67, 0.9, 0.11, 0.01, ...]

Spans creation

(We, are) (Meghan) ... (Clarence, House)

"We are very excited for Harry and Meghan..."

We are

Contextualised Embedding

Clarence

House

Spans creation

(We, are)

(Meghan)

...(Clarence, House)

Relation scorer

(Charles) is parent of (Harry)

Coreference scorer

(Harry)

NER scorer

("tags":
  "gender::male",
  "type::person",
  "type::royalty"
  ...)


How? (*)

- NER
- RE
- CR

=> End-to-end

=> Jointly

=> Document-level
"We are very excited for Harry and Meghan...

Contextualised Embedding

Spans creation

We are... [0.67, 0.9, 0.11,0.01,...]

Coreference scorer

Relation scorer

Charles is parent of Harry

(We, are) ➔ (Meghan) ➔ (Harry)

NER scorer

(We, are) ➔ Clarence ➔ (Clarence, House)

(We, are) ➔ Clarence, House

How? (*)

- NER
- RE
- CR

=> End-to-end
=> Jointly
=> Document-level

External Knowledge

- Wikipedia
- Wikidata

External Knowledge

Meghan, Duchess of Sussex

From Wikipedia, the free encyclopedia

Meghan, Duchess of Sussex (/ˈmɛɡən/; born Rachel Meghan Markle, August 4, 1981), is an American member of the British royal family and a former actress.

Markle was born and raised in Los Angeles, California. Her acting career began while she was studying at Northwestern University. She attributed early career difficulties to her biracial heritage. Her most significant acting role is that of Rachel Zane in the American legal drama Suits, in which she starred for seven seasons (2011–2018). She simultaneously profited from a strong social media presence, including a lifestyle blog, The Tig (2014–2017). Through The Tig she gained recognition for her fashion sense, which led to creating and releasing two lines of clothing in 2015–2016. During her acting career, Markle became involved in charity work, focusing primarily on women's issues and social justice.
External Knowledge

Meghan, Duchess of Sussex

From Wikipedia, the free encyclopedia

Meghan, Duchess of Sussex (/məˈɡɛn/; born Rachel Meghan Markle, August 4, 1981), is an American member of the British royal family and a former actress.

Markle was born and raised in Los Angeles, California. Her acting career began while she was studying at Northwestern University. She attributed early career difficulties to her biracial heritage. Her most significant acting role is that of Rachel Zane in the American legal drama Suits, in which she starred for seven seasons (2011–2018). She simultaneously profited from a strong social media presence, including a lifestyle blog. Through The Tig she gained creating and releasing two career, Markle became involved in women's issues and social justice.

Multiracial people

From Wikipedia, the free encyclopedia

“Mixed race” redirects here. For the album by Tricky, see Mixed Race (album). For more information, see Multiracialism.
External Knowledge

Candidates

1. "Meghan_Trainor"
2. "Meghan_McCain"
3. "Meghan,_Duchess_of_Sussex"
4. "Megan"
5. "Meghan_Allen"

Prior

0.4
0.3
0.2
0.1
0.0
External Knowledge

Candidates
1. "Meghan_Trainor"
2. "Meghan_McCain"
3. "Meghan,_Duchess_of_Sussex"
4. "Megan"
5. "Meghan_Allen"

Prior
0.4
0.3
0.2
0.1
0.0

KB-text: [0.67, 0.9, 0.11, 0.01, ...]
Different configurations

Candidates

1. "Meghan_Trainor"
2. "Meghan_McCain"
3. "Meghan,_Duchess_of_Sussex"
4. "Megan"
5. "Meghan_Allen"
Different configurations

Candidates

1. "Meghan_Trainor"
2. "Meghan_McCain"
3. "Meghan,_Duchess_of_Sussex"
4. "Megan"
5. "Meghan_Allen"
Different configurations

Candidates
1. "Meghan_Trainor"
2. "Meghan_McCain"
3. "Meghan,_Duchess_of_Sussex"
4. "Megan"
5. "Meghan_Allen"
Attention

- Basic dot-product attention:

- Multiplicative attention:

- Additive attention:
Attention

- Basic dot-product attention:

- Multiplicative attention:

- Additive attention:
Different configurations $\alpha$?

Possible names:
- "Meghan_Trainor"
- "Meghan_McCain"
- "Meghan_Duchess"
- "Meghan"
- "Meghan_Allen"

(Meghan)
Different configurations $\alpha$?
Different configurations $\alpha$?
Results
Part 03
Q3: How does external knowledge influence the prediction results?

CR

THE CLUSTERING BECOMES MORE FINE GRAINED
Q3: How does external knowledge influence the prediction results?

The clustering becomes more fine grained.
Q3: How does external knowledge influence the prediction results?

<table>
<thead>
<tr>
<th>CR</th>
<th>NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE CLUSTERING BECOMES MORE FINE GRAINED</td>
<td>IMPROVES THE PREDICTION OF RARE ENTITY TYPES</td>
</tr>
</tbody>
</table>

Google VS Google street view
Q3: How does external knowledge influence the prediction results?

CR
THE CLUSTERING BECOMES MORE FINE GRAINED

NER
IMPROVES THE PREDICTION OF RARE ENTITY TYPES

Google VS street view

Graph showing the improvement in prediction results with external knowledge.
Q3: How does external knowledge influence the prediction results?

CR
THE CLUSTERING BECOMES MORE FINE GRAINED

NER
IMPROVES THE PREDICTION OF RARE ENTITY TYPES

Google VS Google street view

The occurrence of the tag in the train set.

Mention-level NEK performance

Baseline
Best configuration
The difference between both.

10.4%
1. Which problems can be solved with Information Extraction in your current work?
Transformers =

- Language Models
- long range dependencies
- Omnidirectional
Mission logic specification
language

Mission = (s, g, path, robot)

Simple mission = (g, robot)

Fleet Manager
Language Modeling = Models the probability of text

Input = *Where are we*
Output = *going*

Use-cases: auto-complete, spelling correction
Baseline = text characteristics + data in abundance
"Performance depends strongly on scale, weakly on model shape"
GPT3:

175B parameters

cost of 4.6 M€ as lower bound per training run; between 11.5 and 27.6 M€ total development cost
Adapting

First: Pre-Training LM
Then: Fine-Tuning (Transfer Learning) and/or prompting
Prompting for Knowledge Acquisition

What does a GPT model know? EleutherAI - text generation testing UI

**One-shot:** New Delhi is the capital of

**Few-shot:**
1. Stockholm is the capital of Sweden.
2. Brussels is the capital of Belgium.
3. Taipei is the capital of Taiwan.
4. Kampala is the capital of Uganda
5. New Delhi is the capital of
Architecture
The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)
I am a student

Je suis étudiant
I am a student.
$e_{ij} = q_i^T k_j$

Compute key-query affinities

I am a student

Je suis étudiant

0.5
\[ \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})} \]

Compute attention weights from affinities (softmax)

I am a student

<\text{s}> Je suis étudiant

Value V

Query Q

Q: I
V: am
a
student
Output = $V_1 \times 0.5 + V_2 \times 0.3 + V_3 \times 0.1 + V_4 \times 0.1$
ATTENTION

= Long-Range Dependencies
The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)
SELF-ATTENTION

Embedding  General  Intelligence
SELF-ATTENTION

Score
Value
Key
Query
Embedding

General

Q₁ * K₁ = 112

Intelligence

Q₁ * K₂ = 96
SELF-ATTENTION

General Intelligence Embedding

<table>
<thead>
<tr>
<th>Softmax</th>
<th>Score</th>
<th>Value</th>
<th>Key</th>
<th>Query</th>
<th>Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q1 * K1 = 112</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q1 * K2 = 96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SELF-ATTENTION

Softmax X Value

Softmax 0.88

Score Q₁ * K₁ = 112

Value

Key

Query

Embedding

General

Intelligence

Q₁ * K₂ = 96
General Intelligence Embedding

<table>
<thead>
<tr>
<th></th>
<th>Query</th>
<th>Value</th>
<th>Key</th>
<th>Score</th>
<th>Softmax</th>
<th>Softmax X Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>🔵</td>
<td>🔴</td>
<td>🔴</td>
<td>🔵</td>
<td>0.88</td>
<td>🔴 🔴 🔴 ⬆️</td>
</tr>
<tr>
<td>Q1*</td>
<td>🔵</td>
<td>🔴</td>
<td>🔴</td>
<td>🔵</td>
<td>Q1 * K1 = 112</td>
<td>🔵 🔴 🔴 🔴</td>
</tr>
<tr>
<td>Q1*</td>
<td>🔵</td>
<td>🔴</td>
<td>🔴</td>
<td>🔵</td>
<td>Q1 * K2 = 96</td>
<td>🔴 🔴 🔴 ⬆️</td>
</tr>
<tr>
<td></td>
<td>🔵</td>
<td>🔴</td>
<td>🔴</td>
<td>🔵</td>
<td>0.12</td>
<td>🔴 🔴 🔴 🔴</td>
</tr>
</tbody>
</table>

Sum

Intelligence

Self-Attention
ATTENTION

= Omni-Directional
MULTI-HEAD

General Intelligence

Embedding

Query

Key

Value

Softmax

Score

Q1 * K1 = 112

Q1 * K2 = 96

Softmax X

Value

Sum

Score

= 96

X

Value

Sum
I am a student | Je ...
POSITION EMBEDDINGS

- Learned embeddings
- Sinusoidal position representations

Order?
Types:

Sequence-to-sequence models

- T5

Encoders

- BERT, roBERTa, KG-BERT, BART, ...

Decoders

- GPT-3, GPT-j, ...
Seq-To-Seg

For example:

- **machine translation**: “translate English to German: That is good.”
  “Das ist gut.”

- **natural language inference**: “mnt premise: I hate pigeons. hypothesis: My feelings towards pigeons are filled with animosity.”
  “entailment”

- **Winograd challenge**: “The city councilmen refused the demonstrators a permit because “they” feared violence.”
  “The city councilmen”

**INPUT**

**ENCODER**

**DECODER**

**OUTPUT**
Goal: [CLS] my dog is cute [SEP] he likes play ##ing

Input: [CLS] my [MASK] is cute [SEP] he [MASK] play ##ing
The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)
SWEDISH NLP
Overview: Swedish NLP

Datasets & Evaluation
- Text Corpora
- Labeled datasets
- SuperLIM

Models
- Encoders
- Decoders
- Seq2Seq

Use Cases
Swedish Datasets - Corpora

- Swedish Wikipedia (~2GB)
- Litteraturbanken, Swedish literature (< 1GB)
- Oscar, crawled corpus filtered for Swedish (~30GB)
- Swedish Forums (> 10GB)
- National Library (KB), OCR & Radio etc. (~ a lot)
- Data Collection, work in progress...
Swedish Datasets - Labeled Data

- SUC 3.0, Swedish NER Corpus, Swe-NERC, Swedish Medical NER (PoS & NER)
- 😊 Swedish Reviews (Sentiment Classification)
- Machine translated datasets (experimental)
- and more
Swedish Datasets - SuperLIM

- Swedish SuperGLUE evaluation suite
- Evaluates performance and bias
Swedish Datasets - SuperLIM

- Swedish SuperGLUE evaluation suite
- Evaluates performance and bias
- 13 test sets

Overview

<table>
<thead>
<tr>
<th>Resource</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect-Based Sentiment Analysis (Immigration)</td>
<td>Label the sentiment that the author of a text expressed towards immigration on the 1--5 scale</td>
</tr>
<tr>
<td>Dial-A3</td>
<td>Determine whether a sentence is correct Swedish or not</td>
</tr>
<tr>
<td>Swedish FAQ (mismatched)</td>
<td>Match the question with the answer within a category</td>
</tr>
<tr>
<td>SweSAT synonyms</td>
<td>Select the correct synonym or description of a word or expression</td>
</tr>
<tr>
<td>Swedish Analogy test set</td>
<td>Given two word pairs A:B and C:D, capture that the relation between A and B is the same as between C and D</td>
</tr>
<tr>
<td>Swedish Test Set for SemEval 2020 Task 1: Unsupervised Lexical Semantic Change Detection</td>
<td>Determine whether a given word has changed its meaning during a hundred year period</td>
</tr>
<tr>
<td>SweFreCas</td>
<td>Determine to what extent a given word has changed its meaning during a hundred year period</td>
</tr>
<tr>
<td>SweWinograd</td>
<td>Given the question and the premises, choose the suitable answer</td>
</tr>
<tr>
<td>SweWinograd</td>
<td>Resolve pronouns to their antecedents in items constructed to require reasoning (Winograd Schemata)</td>
</tr>
<tr>
<td>SweDiagnosics</td>
<td>Find the correct antecedent of a personal pronoun, avoiding the gender bias</td>
</tr>
<tr>
<td>SweParaphrase</td>
<td>Determine the logical relation between the two sentences</td>
</tr>
<tr>
<td>SuperSim</td>
<td>Determine how similar two sentences are</td>
</tr>
<tr>
<td>SweWIC</td>
<td>Predict semantic word similarity and/or relatedness between words out of context.</td>
</tr>
</tbody>
</table>

https://spraakbanken.gu.se/en/resources/superlim
Swedish Models - Decoders

- GPT-SW3

Swedish Models - Seq2Seq

- Being explored
Swedish Models - Encoders

- AF-BERT (Swedish Public Employment Service)
- KB-BERT
- KB-ELECTRA
- KB-SBERT[1]
- Bigger & better encoders desired

[1]: https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/
Swedish Models - Use Cases

Skatteverket: Anonymization

Severine likes good weather.

LINK:
https://www.ai.se/sites/default/files/content/swebert_best_practices_2021-02-03.pdf
Swedish Models - Use Cases

Skatteverket: Anonymization

Severine likes good weather.

LINK:
https://www.ai.se/sites/default/files/content/swebert_best_practices_2021-02-03.pdf
FUTURE CHALLENGES
Future:

**AI & ETHICS**
Data based, Reflects bias in society

**EVALUATION IN NLP**
Does the results of the benchmark improve because of task understanding?

**GENERALIZATION**
GPT-3, GPT-j, …
Acknowledgements

Material based on
https://jalammar.github.io/illustrated-transformer/
https://jalammar.github.io/illustrated-bert/
https://jalammar.github.io/illustrated-gpt2/
https://jalammar.github.io/how-gpt3-works-visualizations-animations/

Material based on NLP course of Stanford Stanford CS 224N | Natural Language Processing with Deep Learning
1. How do you think that Attention/Transformers can be used in your working field?