Representation learning for natural language processing
- an interface for inference across modalities -

Benjamin Roth, LMU Munich

Outline

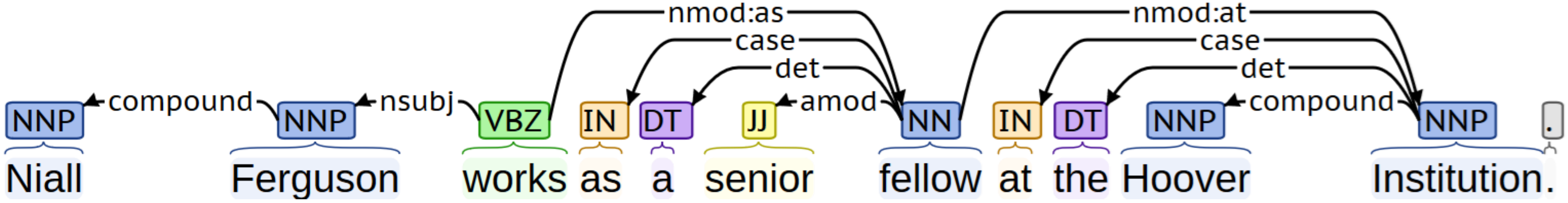
- **Deep Learning for NLP: overview**
- Unsupervised representations
  - Learning vectors for words
  - Modeling smaller units
  - Learning vectors for words in context
- Combining text and structured data

Why should we do NLP?

- human-to-human information exchange
  - Main channel: language
- Can the computer tap into this kind of information?
  - **Social science, business analytics:** analyze events, trends and opinions
  - **Linguistics:** analyze language properties
  - **Dialogue systems, question answering:**
    provide a natural interface between humans and computers
  - **Machine translation:** assist communication across languages
- Hypothesis (Turing test):
  Equivalence of full language capabilities with human-level intelligence

Rule-based systems

- Directly express human expectations and insights
- E.g. **relation extraction**:



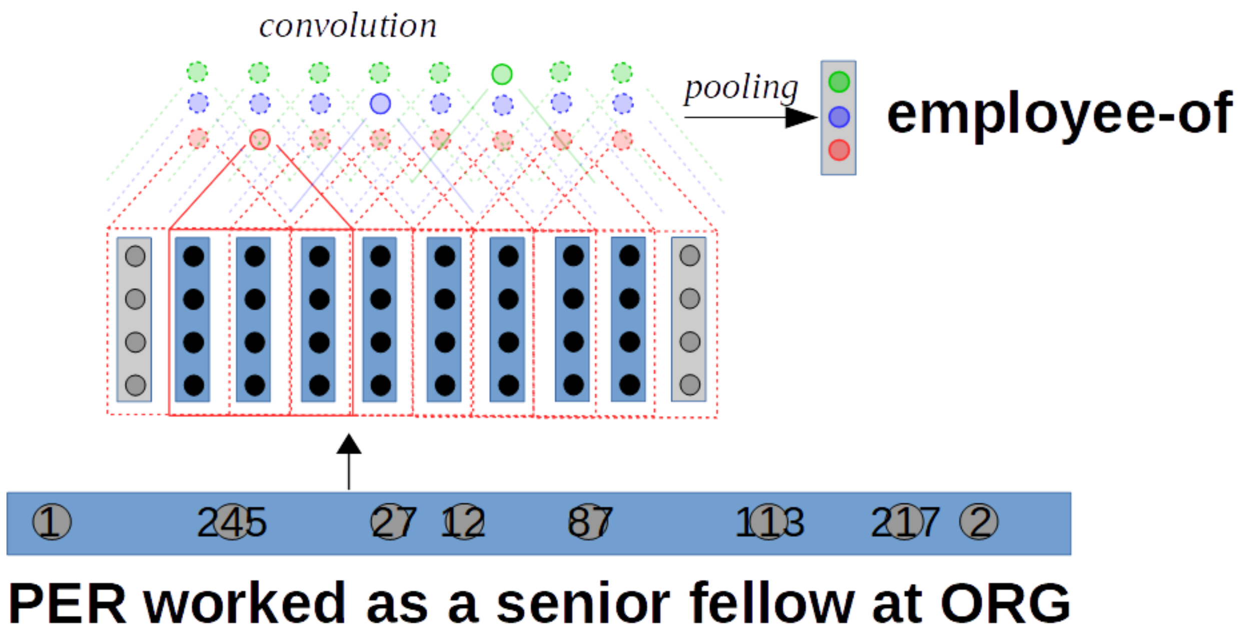
  - Syntactic patterns:
    PER ←nsubj ← works → nmod:as → * → nmod:at → ORG
  - => (Niall Ferguson, employee-of, Hoover Institution)
  - Good precision, low recall!

Statistical systems

- Provide features, automatically weighted by training data
- E.g. **relation extraction**:
  - N-grams:
    0.87 "PER works as"
    0.81 "works as a"
    0.21 "as a senior"
    0.11 "a senior fellow"
    …
    0.62 "at the ORG"
    => (Niall Ferguson, employee-of, Hoover Institution)
  - Better recall than rule-based
  - Cannot generalize to unseen features
  - Difficult to do joint learning (e.g., multilingual relation extraction)

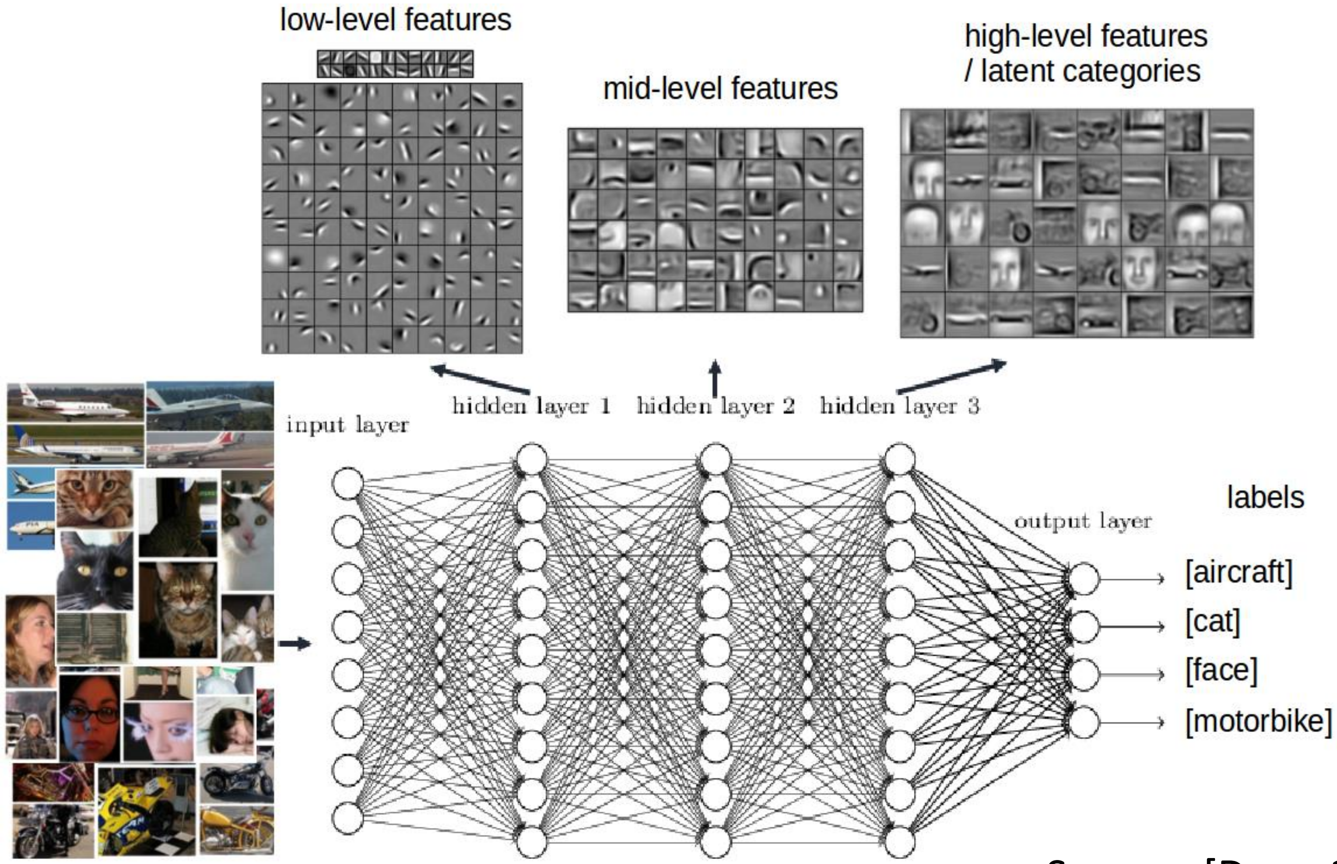
Representation learning

- Provide 'raw' input
- System finds and represents relevant interactions in input



Representation learning
= deep learning = neural networks

- **Raw input** instead of defined feature representation:
  - Images: Pixels
  - Text: Sequence of words or characters
- **Learn higher-level abstractions**



Source: [Deng 2009; Lee 2009]

Representation learning
= deep learning = neural networks

- **Learn higher-level abstractions**
  - **Non-linear functions** can model interactions of lower-level representations
  - E.g.:
    ``The plot was **not** particularly **original**.'' → **negative** movie review
- Typical setup for natural language processing (NLP)
  - Model starts with learned representations for words
    → **word vectors**
  - Word vectors are combined to represent larger units (sentences, documents)

Deep learning advantages (1)

Vector representations provide an API for machine learning

- Allows combination across modalities, input/output types
- A main advantage, even if sometimes traditional models perform equally well

input → vector(s)
vector(s) → vector(s)
vector(s) → output

Deep learning advantages (2)

General purpose mechanisms, independent of specific tasks

- **Mechanisms for encoding a sequence**
- **Mechanisms for producing an output** depend on the task

Deep learning advantages (2)

General purpose mechanisms, independent of specific tasks

- **Mechanisms for encoding a sequence**
  - Representing an input
    - Word vectors
    - Contextualized word vectors
  - Modelling interactions in a sequence of words
    - Convolutional Filters (+ Pooling)
      - Only local interactions (n-grams)
    - Recurrent Networks (Long short-term memory, gated recurrent units)
      - Global interactions with proximity bias
    - Attention [Bahdanau 2015, Hermann 2015, Vaswani 2017]
      - Look-up of relevant information, even if far away in the sequence
- **Mechanisms for producing an output** depend on the task
  - Modelling dependencies in the output:
    Conditional Random Fields [Lafferty 2001; Lample 2016]

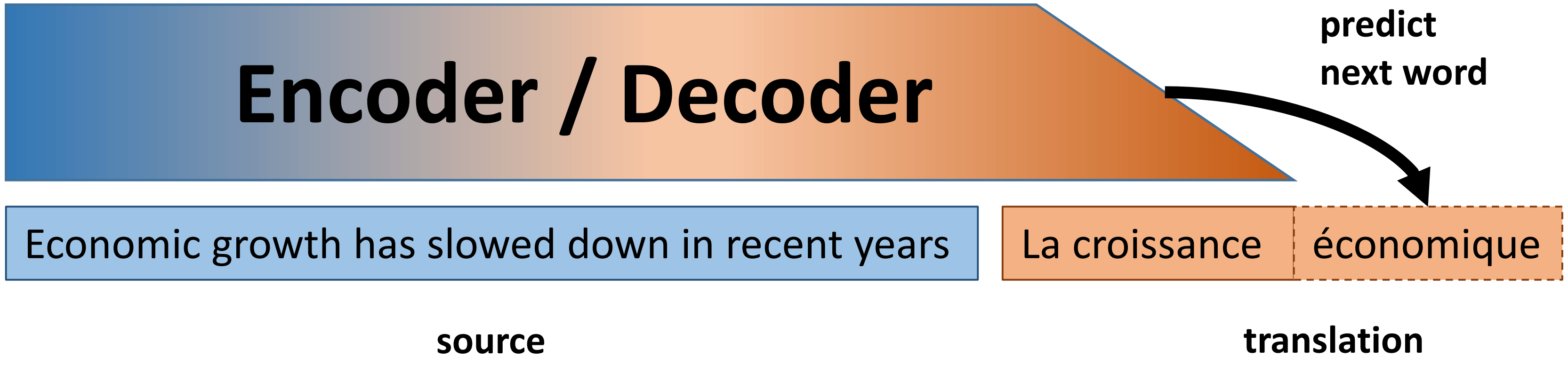
Deep learning advantages (3)

Good trade-off

- Can learn **arbitrary functions** … [Cybenko 1989, Hornik 1991]
- … but biased towards simple functions (**good generalization**) [Perez, 2018]

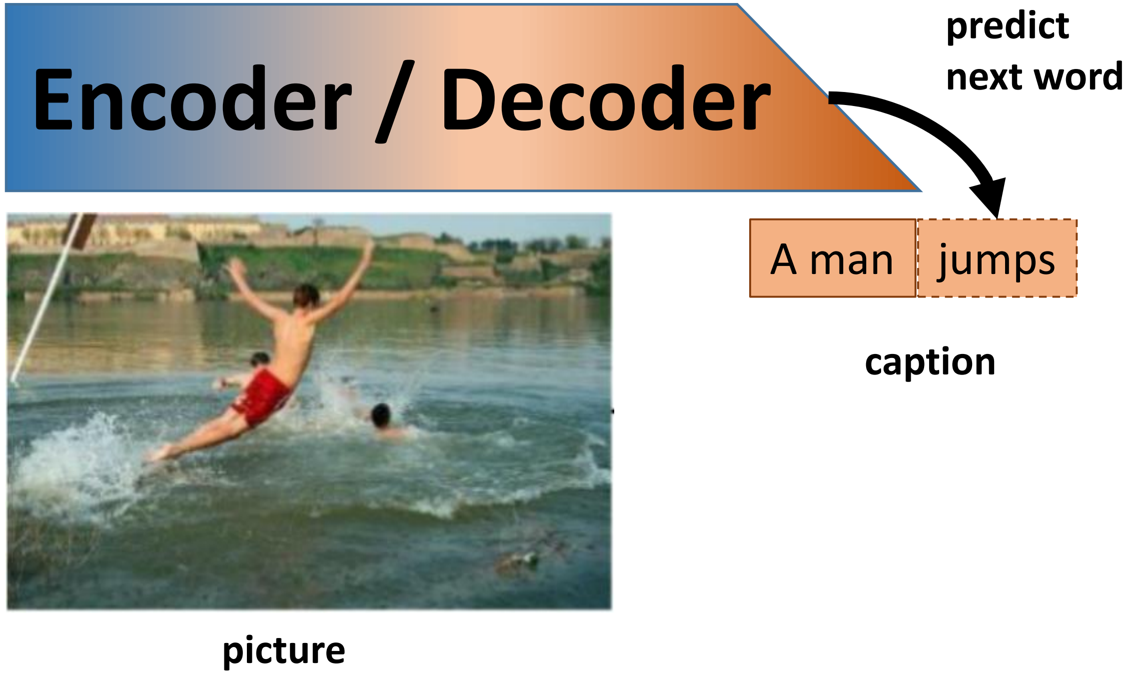
Deep learning/NLP success stories

- Neural Machine Translation
  [Sutskever 2014; Bahdanau 2015; Vaswani 2017 …]
  - Interactions between source and translation generated so far



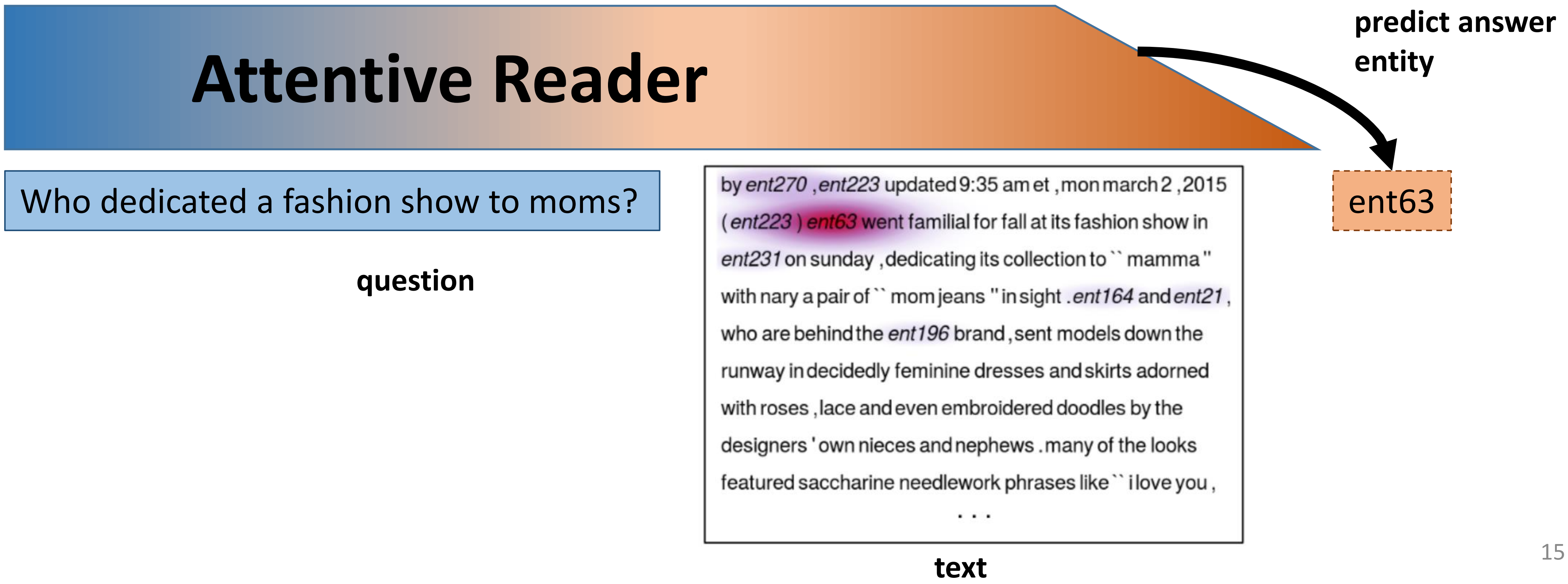
Deep learning/NLP success stories

- Image captioning
  - Interactions between image and caption generated so far [Kiros 2014; Mao 2014; Xu 2015;…]



Deep learning/NLP success stories

- Question Answering
  - Interactions between question and text containing the answer [Hermann 2015, Seo 2017, …]



Deep learning limitations (and how to overcome them)

- Lack of training data
  - → domain adaptation, transfer learning [Howard & Ruder 2018]
  - → **unsupervised pre-training**
- Difficulty to leverage human expertise
  - → combine with rule-based systems, weak supervision [Ratner 2017]
- Lack of insight
  - → automated explanations [Poerner 2018]

Outline

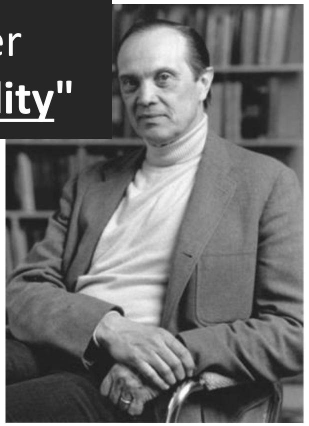
- Deep Learning for NLP: overview
- **Unsupervised representations**
  - **Learning vectors for words**
  - Modeling smaller units
  - Learning vectors for words in context
- Combining text and structured data

The lexical hypothesis

- ``The meaning of a word is captured by the distribution of contexts in which it occurs''
- Co-occurrence between words: no annotation necessary!

The lexical hypothesis

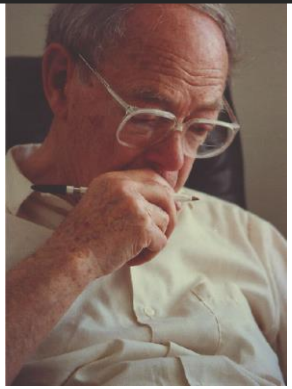
George A Miller (1991):
"Those things are **similar** of which one can be **substituted** for the other without loss of **plausibility**"



Gottfried Wilhelm Leibniz (17th century):
"Those things are **identical** of which one can be **substituted** for the other without loss of **truth**."

The lexical hypothesis

Zellig Harris (1954): "difference in meaning correlates with difference of distribution."



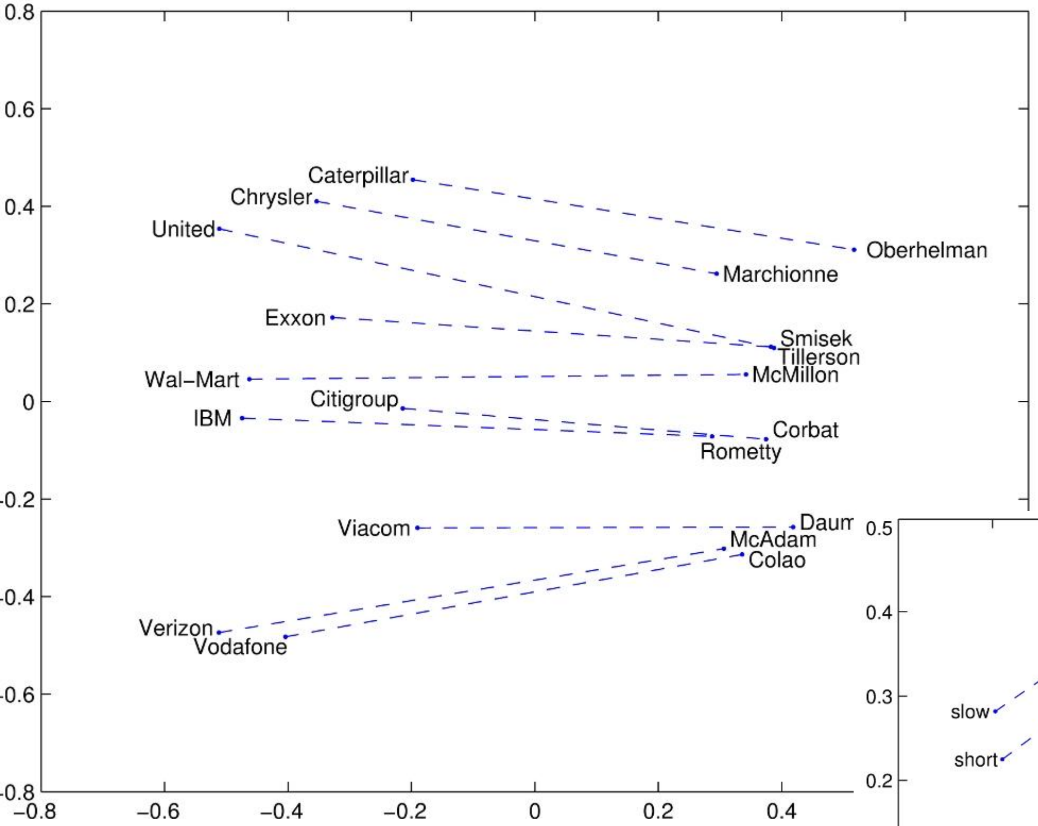
John Rupert Firth (1957): "You shall know a word by the company it keeps."

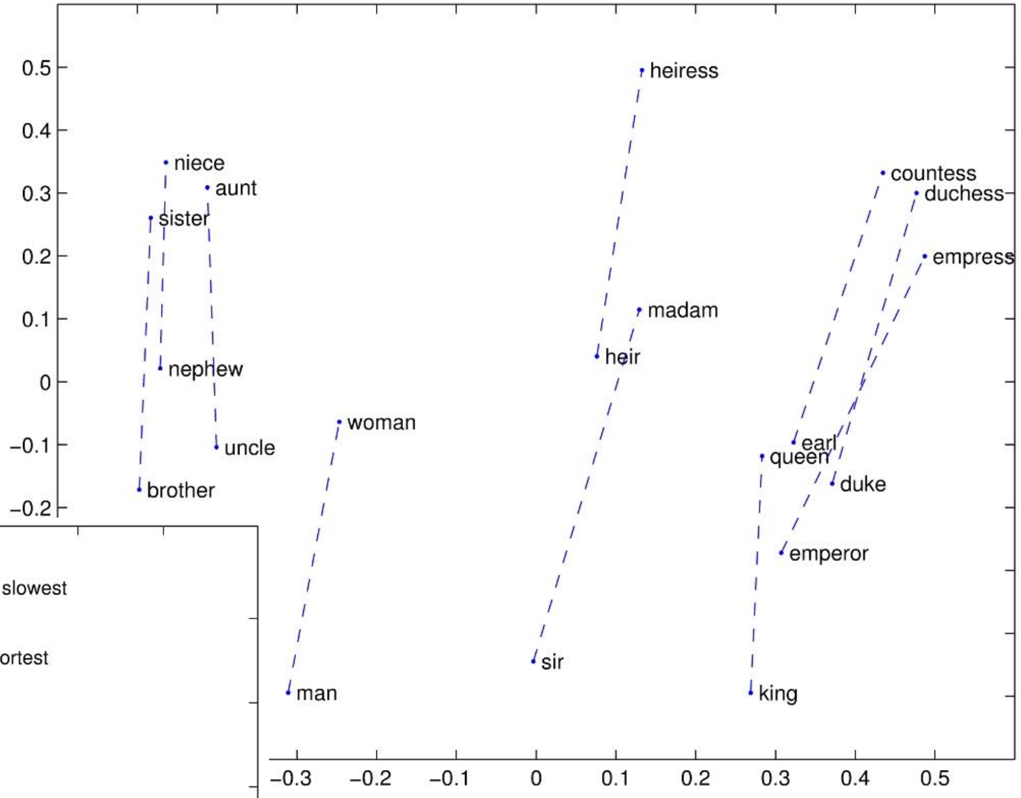
Word vectors: Idea

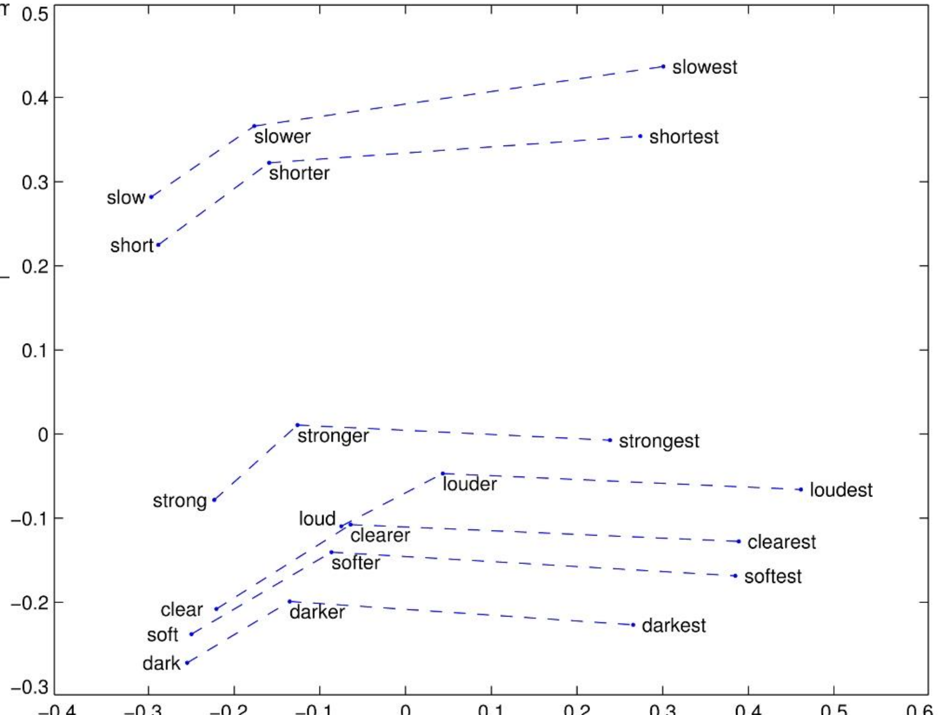
- Represent each word by a vector of numbers indicating abstract semantic properties
- The properties, and the actual values, are automatically found using corpus co-occurrences
- Learn vectors in a task-independent, unsupervised way
  - Goal: Faster & better generalization for specific tasks
- Word vectors can help neural networks to generalize from fewer task-specific training data

U	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
wood	−0.70	0.35	0.15	−0.58	0.16
tree	−0.26	0.65	−0.41	0.58	−0.09

Regularities in word vector space



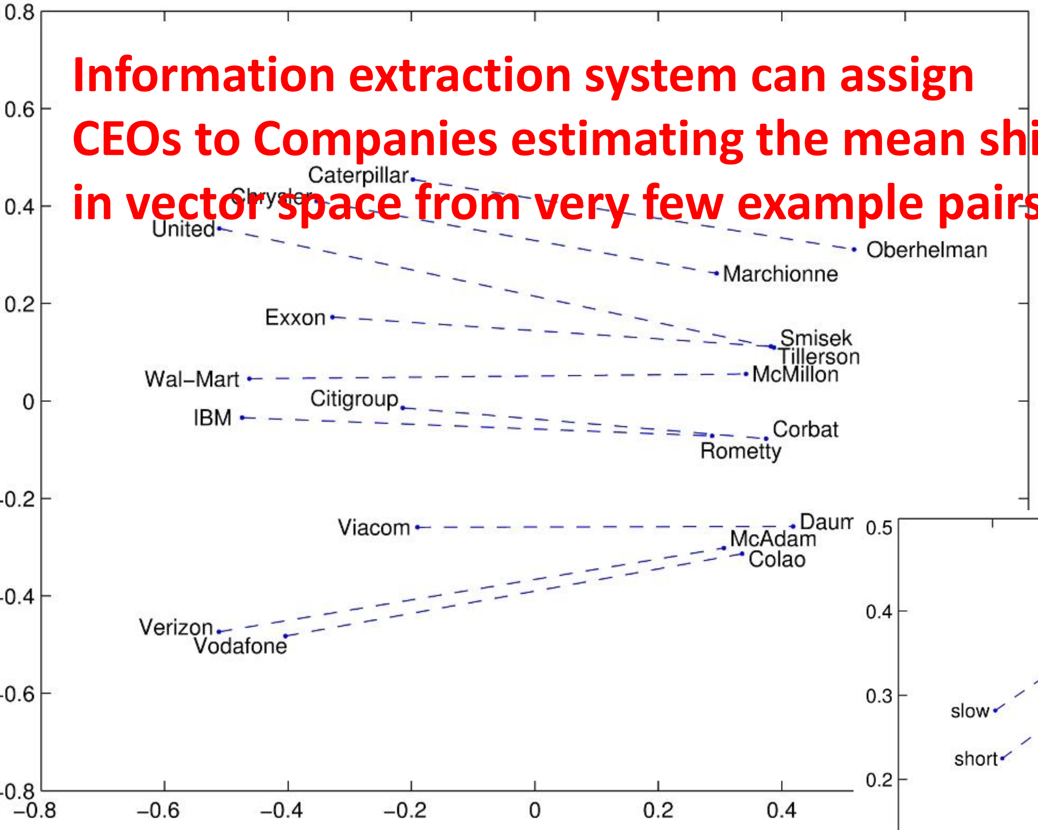




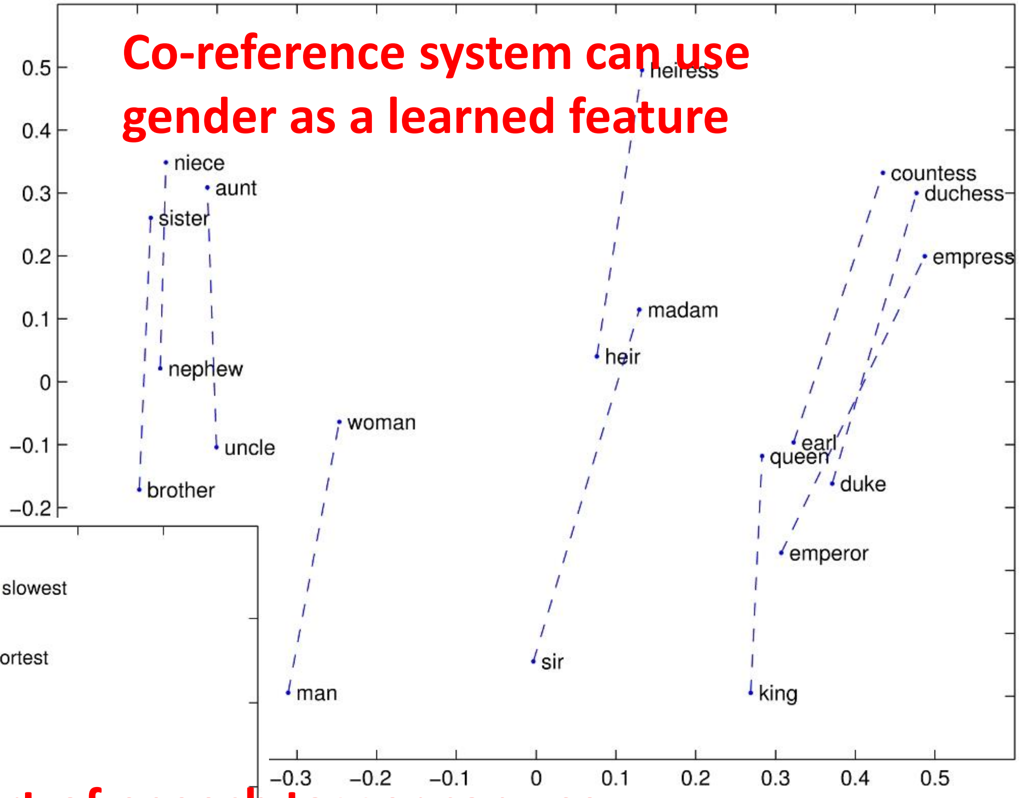
Source: Pennington 2014
(Global word Vectors, GloVe)

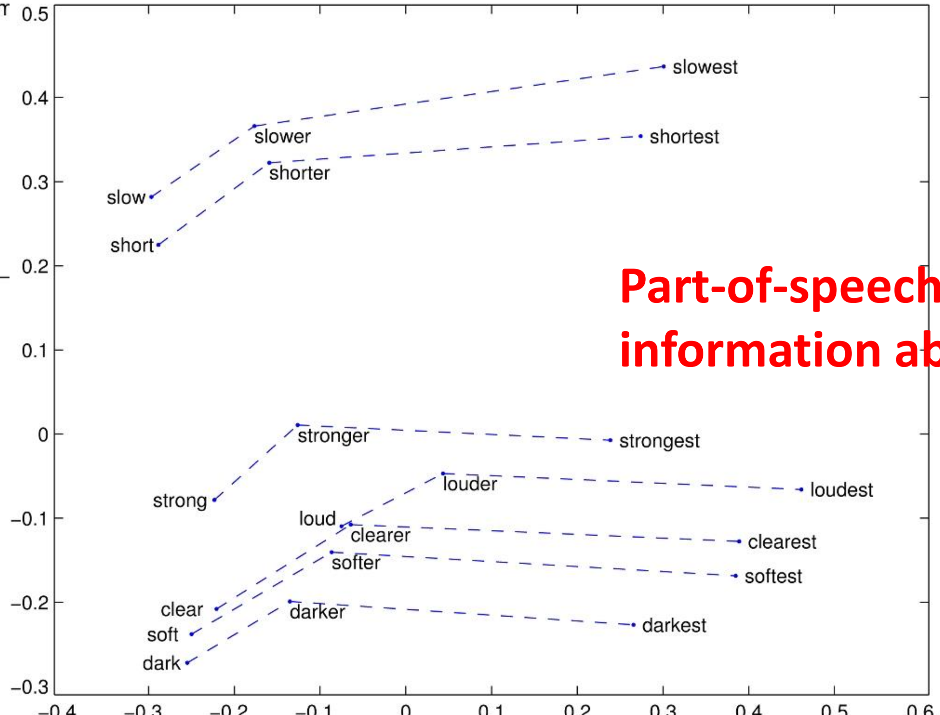
Regularities in word vector space

Information extraction system can assign CEOs to Companies estimating the mean shift in vector space from very few example pairs



Co-reference system can use gender as a learned feature



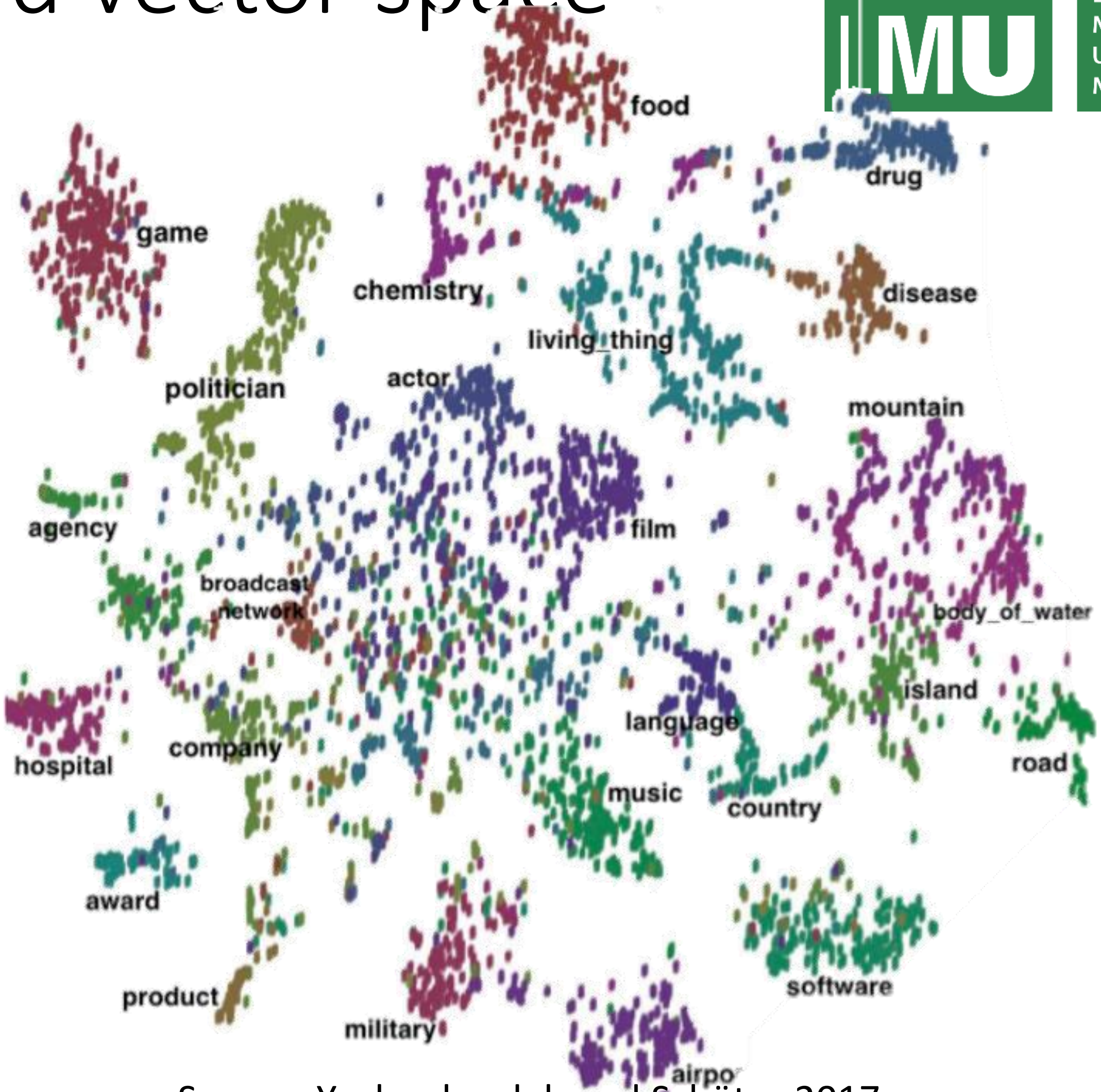


Part-of-speech tagger can use information about syntactic paradigm

Source: Pennington 2014 (Global word Vectors, GloVe)

Regularities in word vector space

- Word vectors of entities cluster along the types of the entities
- Named entity tagger can predict types for unlabeled examples



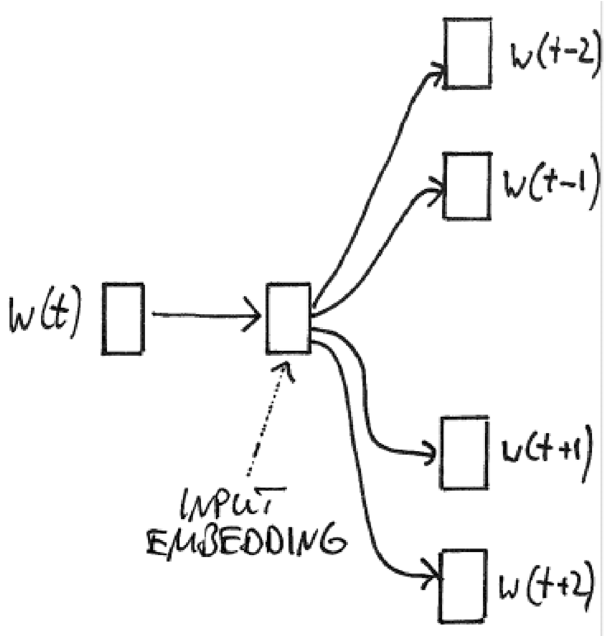
Source: Yaghoobzadeh and Schütze 2017

Word2Vec

- Word2vec [Mikolov, 2013]: predict context around words
  - **Bag-of-words**: no order of context words
  - **No hidden Layer**! Use word vectors directly
  - **Negative sampling, stochastic gradient descent:** scale to very large data sets
- Related to neural language models

  [Bengio 2003; Schwenk 2007; Mikolov, 2010]

  previous context → **hidden layer** → predict next word



Outline

- Deep Learning for NLP: overview
- Unsupervised representations
  - Learning vectors for words
  - **Modeling smaller units**
  - Learning vectors for words in context
- Combining text and structured data

Sub-word modeling

- Words are related through sharing and combining character subsequences
  - sing**ing** – danc**ing**
  - **encod**er – **encod**ing
- Exploit these regularities for better generalization
- Popular subword modeling approaches:
  - FastText [Bojanowski, 2016]: Use all character n-grams
  - Byte-Pair Encoding (BPE), [Sennrich, 2015], SentencePiece Model [Kudo, 2018]: Use most frequent subsequences instead of words
  - Character-level Recurrent Neural Networks [Akbik, 2018]

Sub-word units: FastText [Boyanowski, 2016]

- FastText is an extension of word2vec
- **It computes embeddings for character ngrams**
- A word's embedding is a weighted sum of its character ngram embeddings
- The embedding of the word ``encoder'' will be the sum of the following ngrams:
  - @encoder@ @en enc nco cod ode der er@ @enc enco ncod code oder der@ @enco encod ncode coder oder@ @encod encode ncoder coder@

Sub-word units: BPE

- Byte Pair Encoding (BPE) [Sennrich 2015]
  - Start with characters as the only segments in the corpus
  - Merge most frequent consecutive segments, until desired vocabulary size is reached

```
bpe_tokenize('BERT stands for Bidirectional Encoder Representations from Transformers')
```

```
['bert', 'stands', 'for', 'bid', '##ire', '##ction', '##al', 'en',
'##code', '##r', 'representations', 'from', 'transformers']
```

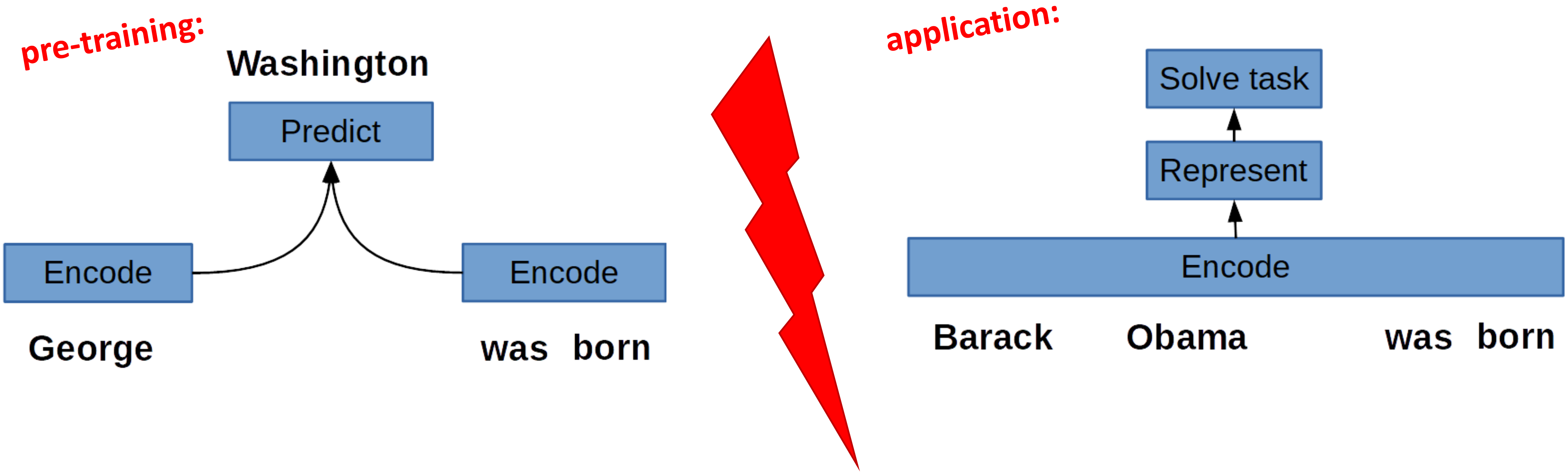
Outline

- Deep Learning for NLP: overview
- Unsupervised representations
  - Learning vectors for words
  - Modeling smaller units
  - **Learning vectors for words in context**
- Combining text and structured data

What about context?

- Part-of-speech for "**stick**"?
  - "Please stick to the topic!"
  - "How do you find the perfect drum stick?"
- Entity type of "**Washington**"?
  - "Washington was born on February 22, 1732, at his family's plantation on Pope's Creek in Westmoreland County"
  - "Some in Europe worry that Washington and Moscow will abandon the treaty."
- Context matters! [McCann, 2017]
- Traditional solution: learn context dependence from **annotated** training data.
- Can one learn contextualized word embeddings with an **unsupervised** objective?

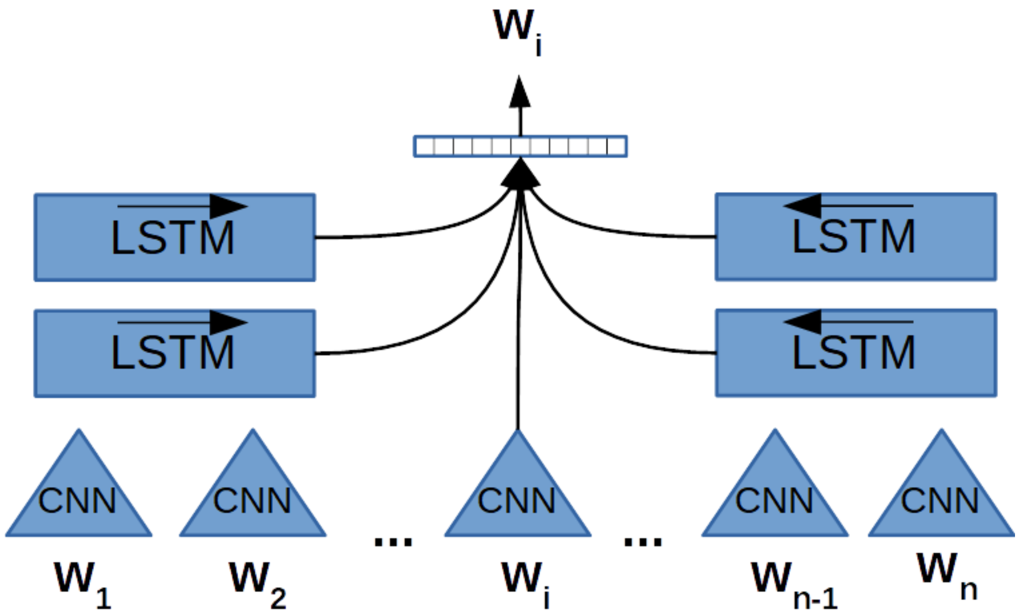
Contextualized word embeddings: Language model objective vs. downstream usage



- ELMO [Peters/AllenAi, 2018]
- BERT [Devlin/Google AI, 2018]
- GPT/GPT2 [Radford/OpenAI 2018, 2019]
- FLAIR (Akbik/Zalando, 2018)

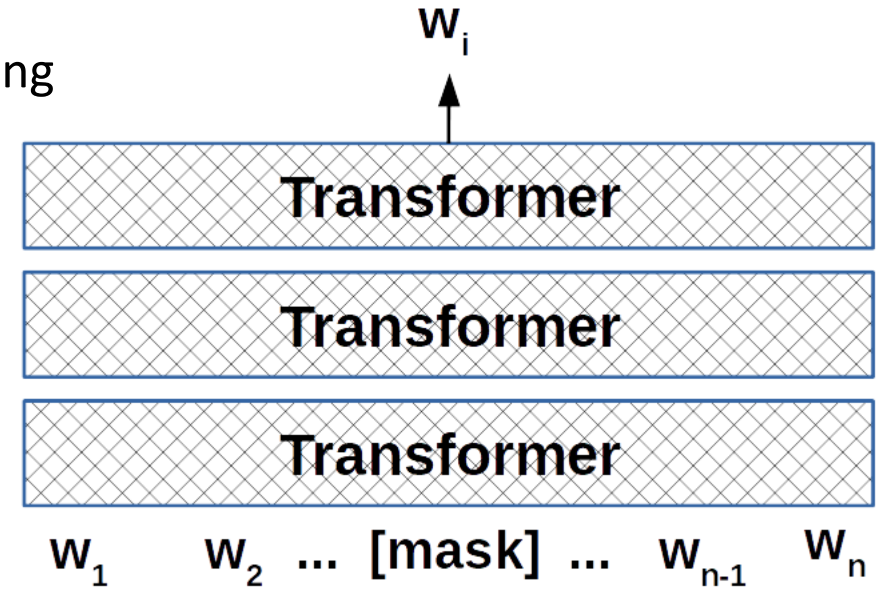
Contextualized word embeddings

- ELMo (*Embeddings from Language Models*) [Peters, 2018]:
  - Word representations:
    Character n-grams → CNN
  - Context representation:
    Bidirectional LSTM Layers → predict word from left and right context



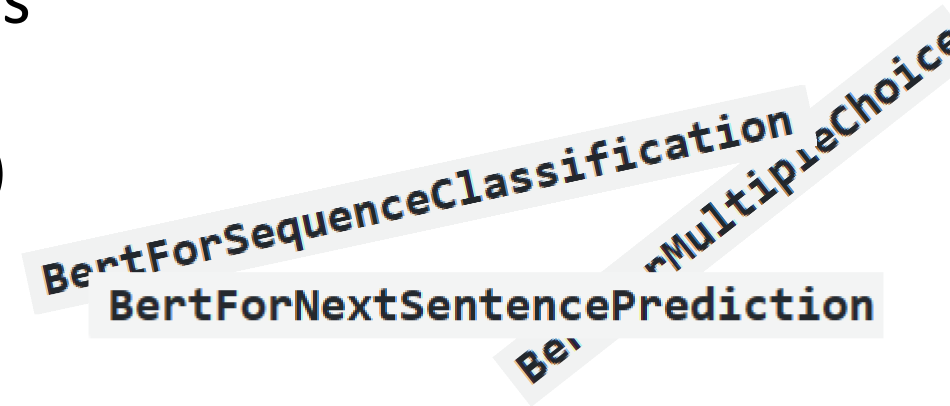
Contextualized word embeddings

- BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin, 2018]
  - BPE word pieces
  - Use *Transformer* [Vaswani 2017] instead of BiLSTM:
    - Every element can interact with every other element
    - Random elements of the input are masked, objective during training is to reconstruct them
  - Clever encoding of different tasks and inputs
    - sequence labelling (tagging)
    - sentence classification
    - sentence pair classification
  - Multilingual: trained on union of different language corpora



Using BERT (or similar)

- Using contextualized pre-trained models is very easy!
- Standard cases covered by pre-trained models
  - Text classification
  - Classification of text pairs (similarity, relatedness)
  - Sequence Labelling
  - …
- Contextualized encoding can be combined with larger architecture/other inputs



The BERT* revolution

*[ELMO/GPT/FLAIR/…]

- Across tasks, current state-of-the-art results are achieved using contextualized word embeddings
  - Machine translation [Lample & Conneau, 2019]
  - Language modelling [Radford 2019]
  - Question answering [Devlin 2018]
  - Named entity recognition [Akbik 2018, Baevski et al., 2019]
  - Sentiment analysis [Liu et al., 2019]
  - Natural language inference [Zhang et al., 2018]
- Simply fine-tuning BERT on task-specific training data is a very strong baseline! [Peters 2019]

Do we still need annotated training data?

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

- Web-sized corpora contain **information** about a range of NLP tasks that can be **elicited from language models** without task-specific fine-tuning
- From the GPT-2 paper: [Radford, 2019]
  - Better than other unsupervised methods for sentence completion in translation contexts:

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”

Outline

- Deep Learning for NLP: overview
- Unsupervised representations
  - Learning vectors for words
  - Modeling smaller units
  - Learning vectors for words in context
- **Combining text and structured data**

How knowledge is stored

- Humans communicate using language: **unstructured**
- Very relevant information is stored in **structured** form
  - spreadsheets, curated knowledge bases (KBs)
  - interface human - computer
- Other data sources
  - sensory data
  - images/video
  - logging data
  - …

How knowledge is stored

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- Humans communicate using language: **unstructured**
- Very relevant information is stored in **structured** form
  - spreadsheets, curated knowledge bases (KBs)
  - interface human - computer
- Other data sources
  - sensory data
  - images/video
  - logging data
  - …

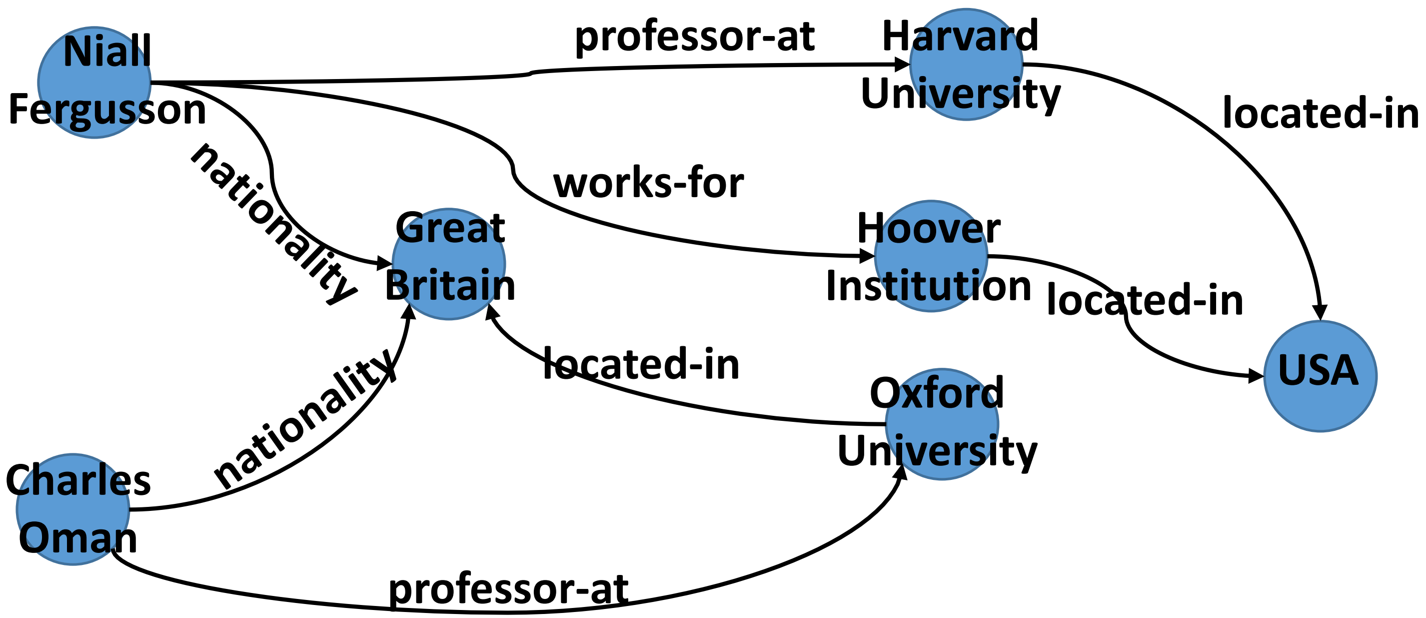
How knowledge is stored

- Humans communicate using language: **unstructured**
- Very relevant information is stored in **structured** form
  - spreadsheets, curated knowledge bases (KBs)
  - interface human - computer
- Other data sources
  - sensory data
  - images/video
  - logging data
  - …

Structured data

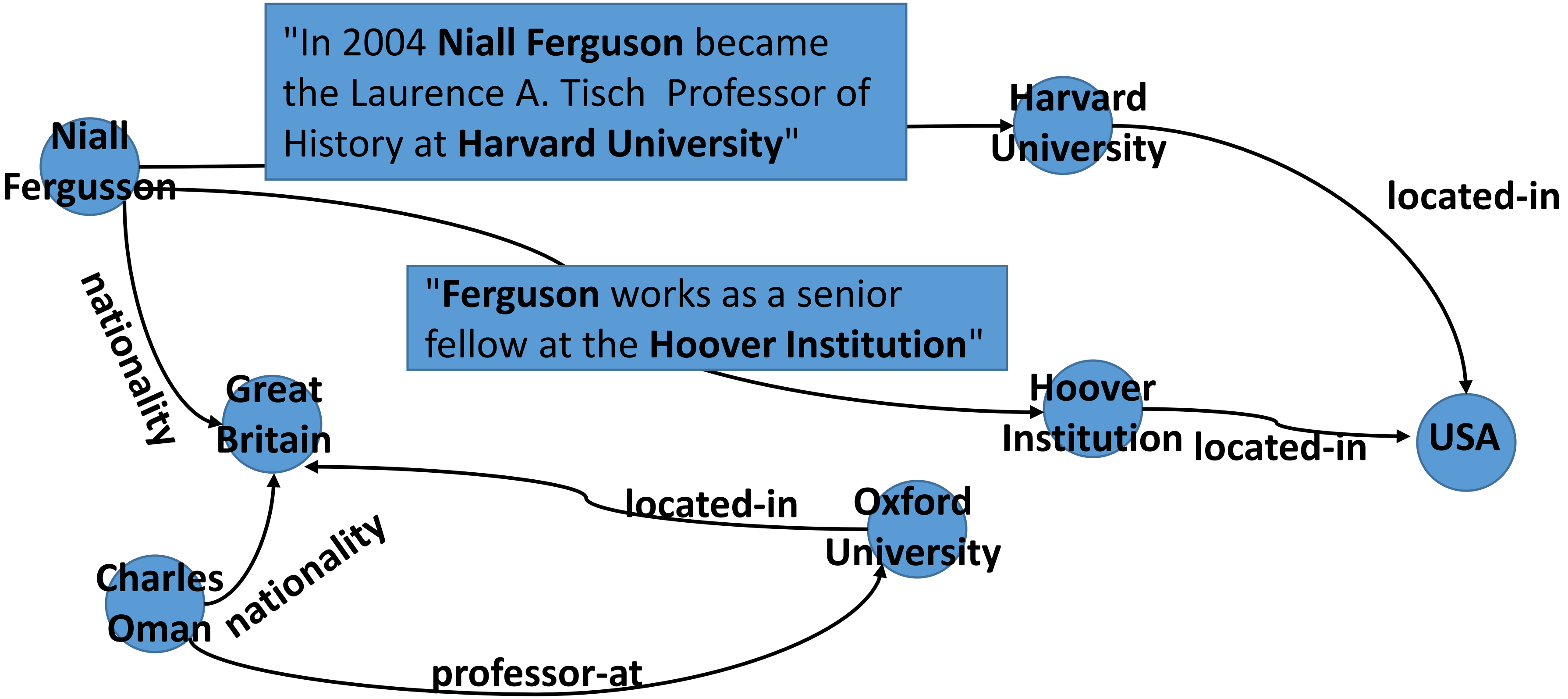
- Structured data:
  - Tables
  - Graphs
  - RDF-Tuples

Name	Professor-at
Niall Ferguson	Harvard
Charles Oman	Oxford
…	…



Structured data + text: Universal Schema
[Riedel, 2013; Toutanova 2015; Verga, 2015…]

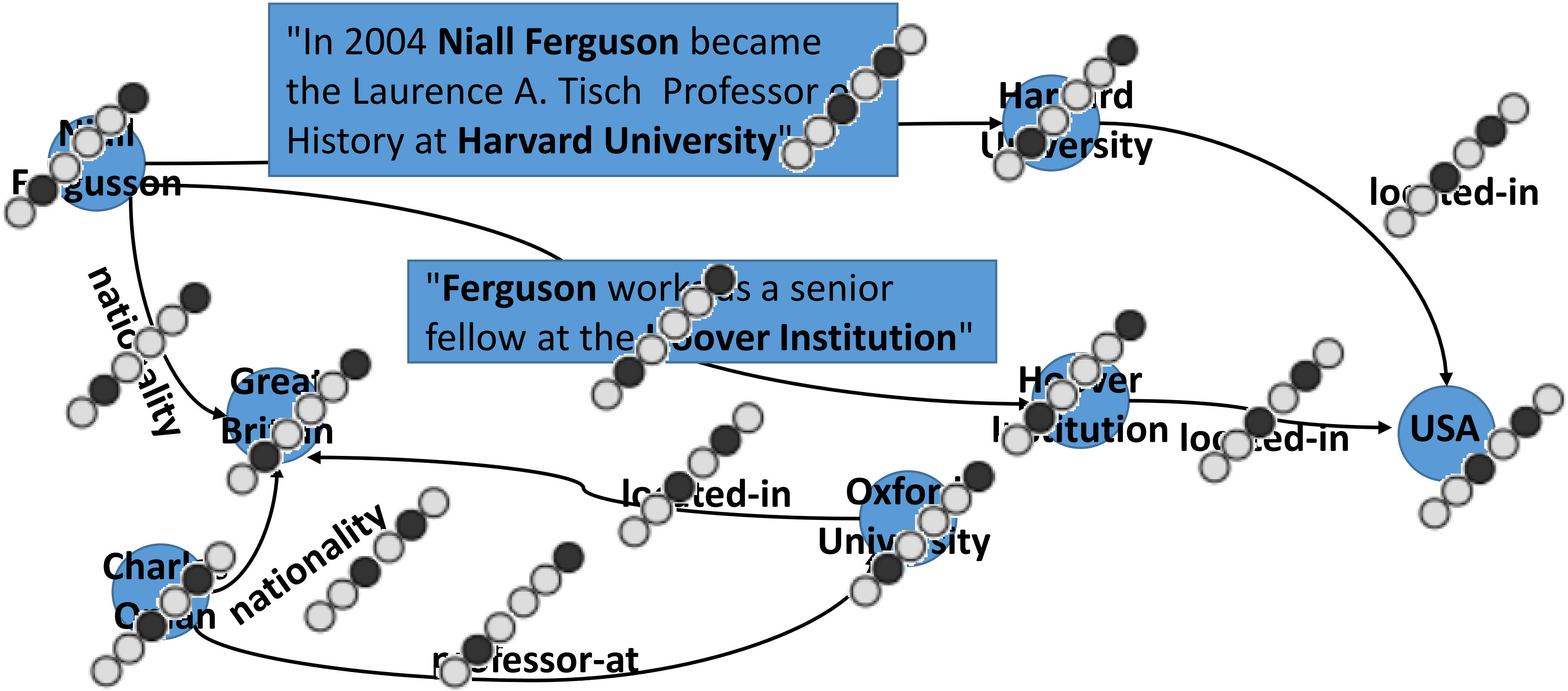
- Language can express (arbitrarily fine-grained) relationships between entities.



Structured data + text: Universal Schema

[Riedel, 2013; Toutanova 2015; Verga, 2015...]

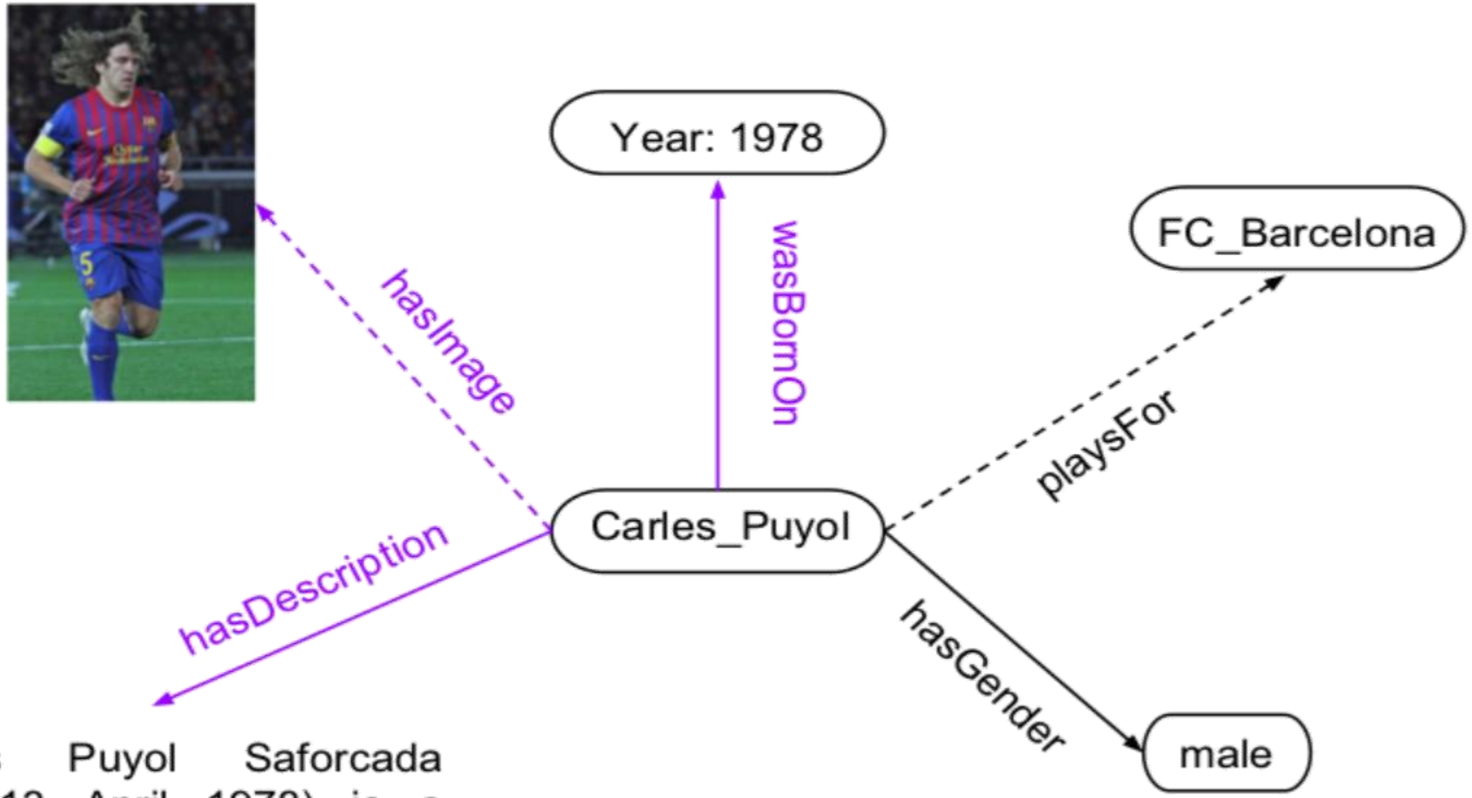
- Language can express (arbitrarily fine-grained) relationships between entities.
- **Encode nodes and edges as vectors**
- **Use entities to align language vector space with KB vector space**



Multimodal structured data

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

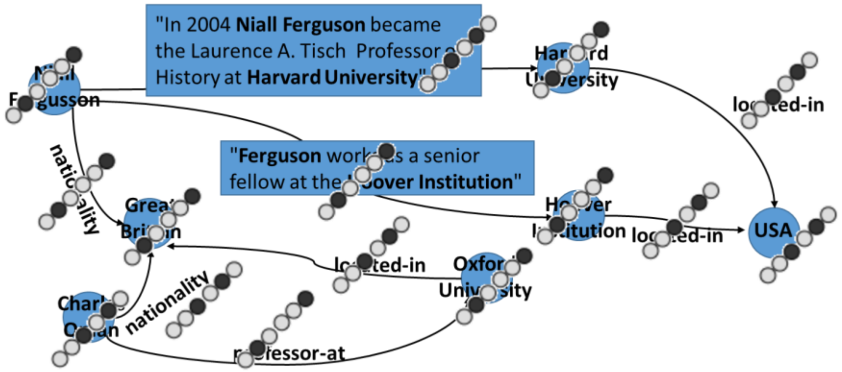
[Pezeshkpour 2018]: Nodes, too, can be analyzable



"Carles Puyol Saforcada (born 13 April 1978) is a Spanish retired professional footballer. He was regarded as one of the best defenders of his generation."

Variants of Universal Schema

- **What are the atomic units?**
  - Text modeling [Toutanova 2015, Verga, 2015]
  - Entity modeling [Verga 2016, Yaghoobzadeh, 2017]
  - Multimodal nodes [Pezeshkpour, 2018]
- **Local modeling of fact triples**
  - Linear translation (TransE, …) [Bordes 2013]
  - Bilinear form (Rescal, Complex …) [Nickel 2011, Trouillon 2016]
  - …
- **Global view**
  - By transitivity from local fact modeling (A-lives-in-city-B, B-in-state-C, C-in-country-D) [Bordes 2013]
  - Ranking loss [Riedel 2013]
  - Graph attention [Velickovic 2018]
  - Recurrent path modeling [Neelakantan 2015]
  - Query-driven: [Das 2017, 2019]
    - Memory networks
    - Reinfocement learning



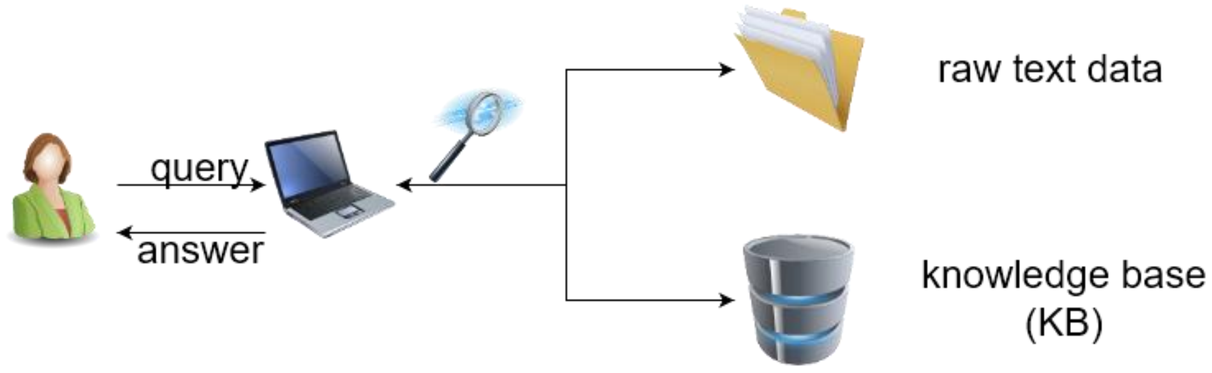
Use-cases of Universal Schema

- Multilingual relation extraction [Verga, 2016]

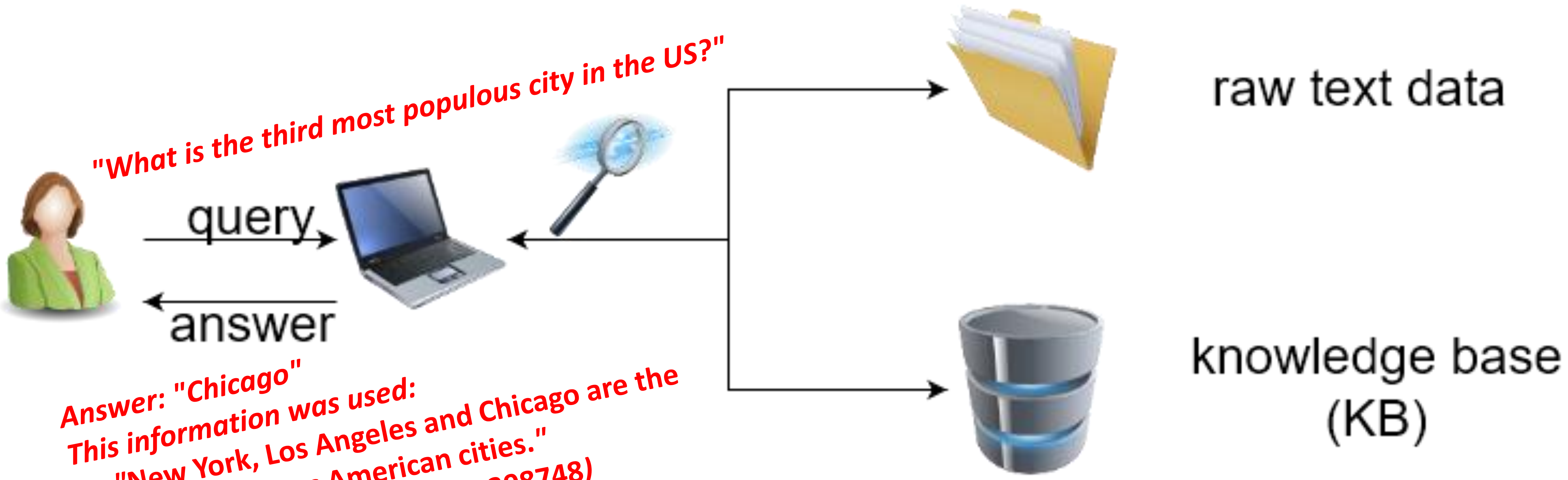
María Múnera está casado con Juan M Santos
Robert C. MacKenzie is survived by his wife, Sybil MacKenzie

person	married to
María Múnera	Juan M Santos
Robert C. MacKenzie	Sybil MacKenzie
…	…

- Question-Answering on Knowledge Bases and Text (TextKBQA) [Das, 2017]



Our current work: Explainable TextKBQA [Sydorova, Poerner, Roth, 2019]



"What is the third most populous city in the US?"

Answer: "Chicago"

This information was used:

- *"New York, Los Angeles and Chicago are the most populous American cities."*
- *population(New York City, 8398748)*
- *population(Chicago, 2716450)*
- *population(Los Angeles, 3999759)*

Summary

- Current state-of-the-art natural language representations
  - represent subwords …
    … in context
  - learned in an unsupervised way from large corpora
  - to be fine-tuned on task-specific data
- Universal Schema
  - Represent structured and unstructured data in same space
  - Allows for inferences across modalities
- Insight into what deep models are doing is important!

Summary

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- Current state-of-the-art natural language representations
  - represent subwords …
    … in context
  - learned in an unsupervised way from large corpora
  - to be fine-tuned on task-specific data
- Universal Schema
  - Represent structured and unstructured data in same space
  - Allows for inferences across modalities
- Insight into what deep models are doing is important!

Thank you!
Questions?

References

- [Akbik 2018] Akbik, Alan and Blythe, Duncan and Vollgraf, Roland. Contextual String Embeddings for Sequence Labeling.
- [Baevski 2019] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, Michael Auli. Cloze-driven Pretraining of Self-attention Networks.
- [Bahdanau 2015] Dzmitry Bahdanau, Kyunghyung Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate.
- [Bengio 2003] Y Bengio, R Ducharme, P Vincent, C Jauvin. A neural probabilistic language model.
- [Bojanowski, 2016] P Bojanowski, E Grave, A Joulin, T Mikolov. Enriching Word Vectors with Subword Information.
- [Bordes 2013] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. Translating embeddings for modeling multi-relational data.
- [Cybenko 1989] G. Cybenko Approximations by superpositions of sigmoidal functions
- [Das 2017] Question answering on knowledge bases and text using universal schema and memory networksR Das, M Zaheer, S Reddy, A McCallum
- [Das 2019] Multi-step Retriever-Reader Interaction for Scalable Open-domain Question AnsweringR Das, S Dhuliawala, M Zaheer, A McCallum
- [Deng 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. CVPR 2009.

References

- [Devlin 2018] J Devlin, MW Chang, K Lee, K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.
- [Hermann 2015] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom. Teaching Machines to Read and Comprehend.
- [Hornik 1991] Kurt Hornik. Approximation Capabilities of Multilayer Feedforward Networks
- [Howard & Ruder 2018] Jeremy Howard, Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification
- [Kiros 2014] Ryan Kiros, Ruslan Salakhutdinov, Rich Zemel. Multimodal Neural Language Models.
- [Kudo 2018] T Kudo, J Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.
- [Lafferty 2001] J Lafferty, A McCallum, FCN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data
- [Lample & Conneau] Guillaume Lample, Alexis Conneau. Cross-lingual Language Model Pretraining
- [Lee 2009] Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations.
- [Liu 2019] Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao. Multi-Task Deep Neural Networks for Natural Language Understanding.
- [Mao 2014] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN).

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

References

- [Mikolov 2010] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, Sanjeev Khudanpur. Recurrent Neural Network Based Language Model
- [Mikolov, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality.
- [Neelakantan 2015] Neelakantan, A., Roth, B., & McCallum, A. Compositional vector space models for knowledge base inference.
- [Nickel 2011] Nickel, M., Tresp, V., & Kriegel, H. P.. A Three-Way Model for Collective Learning on Multi-Relational Data.
- [Pennington 2014] J Pennington, R Socher, C Manning. Glove: Global vectors for word representation.
- [Perez 2018] Guillermo Valle-Pérez, Chico Q. Camargo, Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions.
- [Peters 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. Deep contextualized word representations
- [Peters 2019] Matthew Peters, Sebastian Ruder, Noah A. Smith. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks.
- [Pezeshkpour 2018] Pezeshkpour, P., Chen, L., & Singh, S. Embedding multimodal relational data for knowledge base completion.
- [Poerner 2018] Nina Poerner, Benjamin Roth, Hinrich Schütze. Evaluating neural network explanation methods using hybrid documents and morphological agreement.

References

- [Radford 2018] A Radford, K Narasimhan, T Salimans, I Sutskever. Improving language understanding by generative pre-training.
- [Radford 2019] A Radford, J Wu, R Child, D Luan, D Amodei, I Sutskever. Language Models are Unsupervised Multitask Learners.
- [Ratner 2017] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré. Snorkel: Rapid Training Data Creation with Weak Supervision.
- [Riedlel 2013] Riedel, S., Yao, L., McCallum, A., & Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas
- [Schwenk 2007] H Schwenk. Continuous space language models.
- [Sennrich 2015] R Sennrich, B Haddow, A Birch. Neural machine translation of rare words with subword units
- [Seo 2017] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi. Bidirectional Attention Flow for Machine Comprehension.
- [Sutskever 2014] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks.
- [Sydorova, Poerner, Roth, 2019] Alona Sydorova, Nina Poerner, Benjamin Roth. Explainable Question Answering on Knowledge Bases and Text.
- [Toutanova 2014] Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. Representing text for joint embedding of text and knowledge bases.
- [Trouillon 2016] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. Complex embeddings for simple link prediction.

References

- [Vaswani 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin . Attention is all you need.
- [Veličković 2017] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks.
- [Verga 2014] Verga, P., Belanger, D., Strubell, E., Roth, B., & McCallum, A. Multilingual relation extraction using compositional universal schema.
- [Verga 2016] Verga, P., & McCallum, A.. Row-less universal schema.
- [Yaghoobzadeh and Schütze 2017] Yadollah Yaghoobzadeh, Hinrich Schütze. Multi-level Representations for Fine-Grained Typing of Knowledge Base Entities.
- [Yaghoobzadeh 2017] Yaghoobzadeh, H Adel, H Schütze. Noise Mitigation for Neural Entity Typing and Relation Extraction.
- [Xu 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.
- [Zhang 2018] Zhuosheng Zhang, Yuwei Wu, Zuchao Li, Shexia He, Hai Zhao. I Know What You Want: Semantic Learning for Text Comprehension.