Universität Potsdam

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Advanced Seminar

Knowledge Graphs meet Language Models

Alejandro Sierra and Nitisha Jain
Information Systems
02.05.2022

Outline

- Organizational Matters
- Knowledge Graphs
  - Definition, Examples
  - KG Construction - Open IE
  - KG Completion - Embeddings
- Language Models
  - Probabilistic Language Modeling and N-GRAMs
  - RNN
  - Transformers

Organization
Schedule

April 25	Organization & Preview (Nitisha and Alejandro)
May 2	Introduction session (Nitisha and Alejandro) - Topics + Part 1
May 9	No session
May 16	KG paper1 (student A) + LM paper1 (student B)
May 23	No session
May 30	KG paper2 (student C) + LM paper2 (student D)
June 6	Holiday
June 13	KG paper3 (student B) + LM paper3 (student A)
June 20	No session
June 27	KG paper4 (student D) + LM paper4 (student C) + Intro to Part 2
July 25	Final Poster session

Papers for Part 1 (Session 1 and 2)

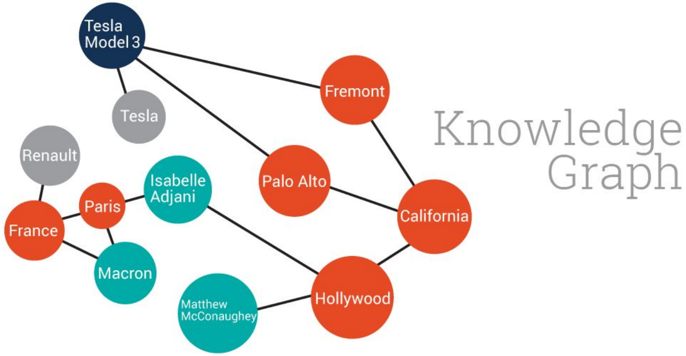
- KG
  - NELL - A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. “Toward an Architecture for Never-Ending Language Learning”. In: Conference on Artificial Intelligence (AAAI). 2010
  - YAGO - F. M. Suchanek, G. Kasneci, and G. Weikum. “Yago: a core of semantic knowledge”. In: The Web Conference (WWW). 2007.

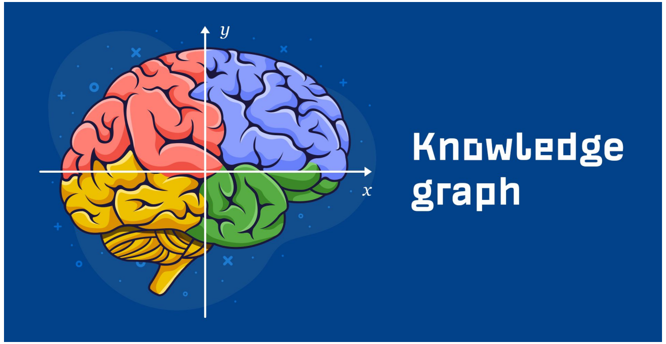
    or
  - DBpedia - S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. “DBpedia: A Nucleus for a Web of Open Data”. In: International Semantic Web Conference (ISWC). 2007. - **Assigned to Kien [16 May]**
- LM
  - Radford and Narasimhan. “Improving Language Understanding by Generative Pre-Training.” (2018). - **Assigned to Lukas [16 May]**
  - Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, NAACL 2019 - **Assigned to Maluna [30 May]**

KG or KB : Definition

A knowledge graph, also known as a semantic network, represents a network of *real-world entities*—i.e. objects, events, situations, or concepts—and illustrates the relationship between them.

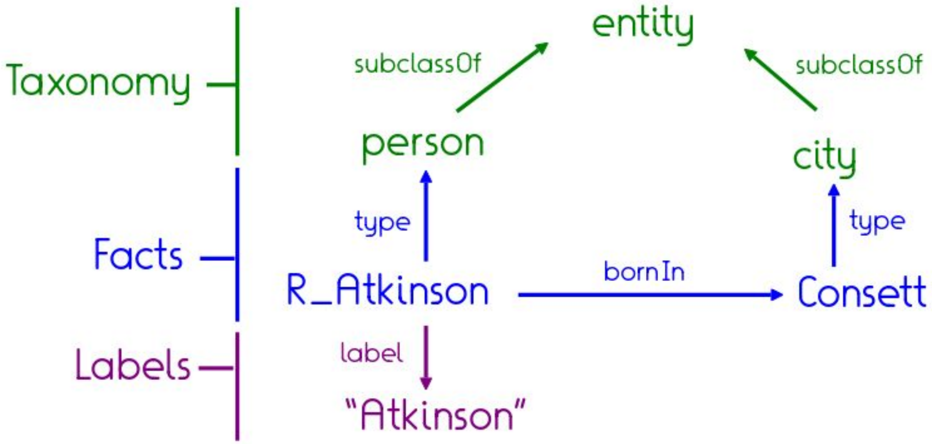
This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge "graph".





KG or KB : Technical Definition

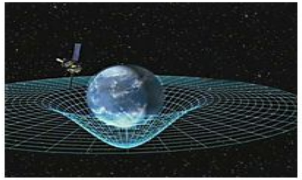
A Knowledge Base (also: Knowledge Graph, entity-relationship graph) is a set of triples. It can be equivalently seen as a directed labelled multi-graph.



Borrowed from slides by Fabian Suchanek, Télécom Paris University

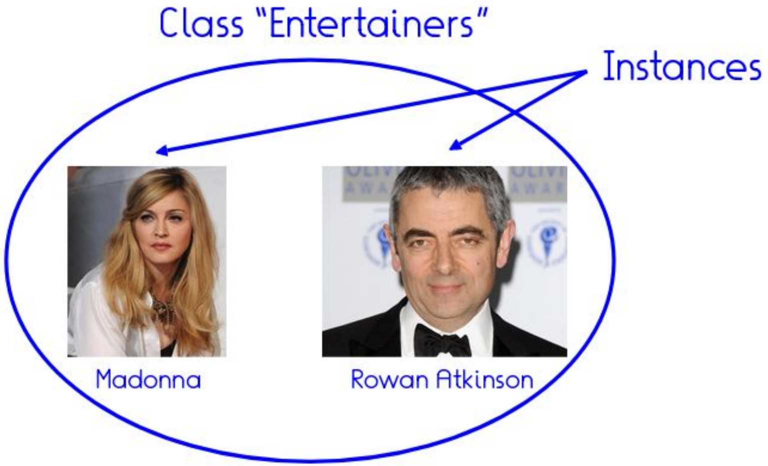
Entity

An entity is anything that may be an object of thought.



Class

A class (also: concept) is a set of similar entities. Each entity is an instance of (also: belongs to) the class.



Other classes:

- Scientists
- Cars
- Cities
- Rivers
- Universities
- Theories
- ...

Subclass, Taxonomy

A class is a subclass of another class, if all instances of the first class are also instances of the second class.

A taxonomy is a hierarchy of classes.

[Queen's College in YAGO 4]

Relation

A relation (also:predicate, property) over classes is a subset of their cartesian product.

The classes form the domain and range of the relations.

$$R \subseteq C_1 \times C_2 \times \ldots \times C_n$$

$$born \subseteq person \times city \times year$$

$$\begin{aligned} born = \{ & \langle Atkinson, Consett, 1955 \rangle, \\ & \langle Sakharov, Moscow, 1921 \rangle, \\ & \ldots \} \end{aligned}$$

Examples - Yago

- YAGO project (https://yago-knowledge.org) by Fabian Suchanek in 2006.
- One of the first large knowledge bases automatically extracted from Wikipedia.
- Maintained and advanced by the Max Planck Institute for Informatics in Germany and Télécom Paris University in France.
- Used for many projects - e.g., semantic type checking in the IBM Watson system that won Jeopardy.
- Taxonomy - combination of Wikipedia’s hierarchy of categories and WordNet.
- YAGO 2 (2010): Spatial and Temporal Scoping
- YAGO 3 (2014): Multilingual Knowledge
- YAGO 4 (2020): Alignment with Wikidata

Examples - DBpedia

- DBpedia (https://dbpedia.org), by Auer et al. in 2007.
- Also constructs a large-scale knowledge base from Wikipedia contents.
- Information of the Wikipedia infoboxes, larger coverage than YAGO.
- SPARQL endpoint for querying.
- Spotlight tool for named entity recognition and disambiguation.
- Since 2014, run by the DBpedia Association with regional chapters in 15 countries.

Examples - NELL

- Never-Ending Language Learner NELL (http://rtw.ml.cmu.edu), 2010.
- Project at Carnegie Mellon University to build a knowledge base “ab initio” from any kinds of web sources.
- Continuously running over many years, KB is incrementally grown
- Starts with a manually created schema, with ca. 300 classes and ca. 500 binary relations with type signatures.
- Latest iteration - 1095 !
    - 2,810,379 asserted instances of 1,186 different categories and relations.

Reading and resources

- Gerhard Weikum, Xin Luna Dong, Simon Razniewski and Fabian Suchanek (2021), "Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases", Foundations and Trends® in Databases: Vol. 10: No. 2-4, pp 108-490. http://dx.doi.org/10.1561/1900000064 - **Chapter 1, 2 , 9.1-9.3**

- NELL - A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. “Toward an Architecture for Never-Ending Language Learning”. In: Conference on Artificial Intelligence (AAAI). 2010

- YAGO - F. M. Suchanek, G. Kasneci, and G. Weikum. “Yago: a core of semantic knowledge”. In: The Web Conference (WWW). 2007.

- DBpedia - S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. “DBpedia: A Nucleus for a Web of Open Data”. In: International Semantic Web Conference (ISWC). 2007.

Knowledge Base Construction (KBC or KBP)

Knowledge Base Construction is the process of building the knowledge base i.e. populating facts in a structured manner from extracted information.

Why do it ? Structured information is useful for a many applications -

- Chatbots
- Recommendation systems
- Question Answering
- Search and exploration
- Entity and fact similarity
- …

Information Extraction

Information Extraction (IE) is the process of deriving structured information from digital text documents.

Barack Obama is an American politician

<Barack Obama, nationality, American>

<Barack Obama, job, politician>

Open IE Systems

- TextRunner[1]: original Open IE system
  - very generous in its extractions
- OLLIE[2] : extracts also non-verbal relations
  - *President Obama - > <Obama, is, president>*
- ReVerb[3] : uses hand-crafted patterns plus POS tagging
  - *Berlin is a city -> <Berlin, is, city>*
- ClausIE[4] : uses linguistically motivated patterns in dependency parses
  - *..she was born in Paris in 1996..” -> <she, was born in, Paris>*

    <she, was born in, 1996>

Borrowed from slides by Fabian Suchanek, Télécom Paris University

Open IE Systems - Recent ones

- OpenIE 5.0[5] : extracts more than subject, predicate, object
  - *..if he wins five key states, Republican candidate Mitt Romney will be elected President in 2008..*
  - *<Republican candidate Mitt Romney, will be elected, President; T: in 2008 >*

- Now, Open IE 6.0 - https://github.com/dair-iitd/openie6 ACL 2020.

Canonicalization

An entity or relation is canonic in a KG, if it has a single identifier in the KG.

Open IE provides non-canonic entities and relations.

Canonicalization is the task of bringing different mentions of the same relations or entities into one single form.

<*He*, ***married***, *Michelle Robinson*>
<*Michelle*, ***married***, *Barack*>
<*Michelle*, ***is wife***, *Obama*>
<*Barack*, ***is spouse***, *Ms. Robinson*>

Canonicalization

Canonicalization is essential for

- Counting
- Question answering
- Reasoning on KGs

Canonicalization

Canonicalization is essential for

- Counting
- Question answering
- Reasoning on KGs

<*He*, ***married***, *Michelle Robinson*>
<*Michelle*, ***married***, *Barack*>
<*Michelle*, ***is wife***, *Obama*>
<*Barack*, ***is spouse***, *Ms. Robinson*>

He, Barack, Obama	→	***Barack Obama***
Michelle Robinson, Michelle, Ms. Robinson	→	***Michelle Obama***
married, is spouse, is wife	→	***has_spouse***

References

1. Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In International Joint Conference on Artificial Intelligence (IJCAI), 2007, volume 7, pages 2670–2676.

2. Schmitz, Michael, et al. "Open language learning for information extraction." Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. 2012

3. A. Fader, S. Soderland, and O. Etzioni. “Identifying relations for open information extraction”. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). 2011.

4. Del Corro, L. and Gemulla, R., 2013, May. Clausie: clause-based open information extraction. In Proceedings of the 22nd international conference on World Wide Web (pp. 355-366).

5. Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 523–534, Jeju Island, Korea, July. Association for Computational Linguistics

Reading and Resources

- Gerhard Weikum, Xin Luna Dong, Simon Razniewski and Fabian Suchanek (2021), "Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases", Foundations and Trends® in Databases: Vol. 10: No. 2-4, pp 108-490. http://dx.doi.org/10.1561/1900000064 - **Chapter 7**

- Galárraga, L., Heitz, G., murphy, k., and Suchanek, F. M. (2014). Canonicalizing Open Knowledge Bases. In CIKM, Shanghai, France. ACM Press.

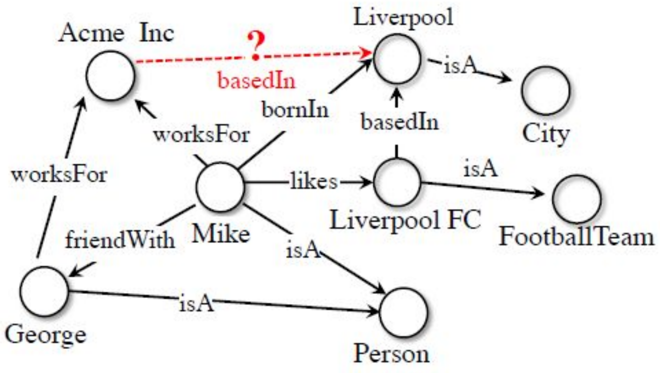
- Kolluru, K., Adlakha, V., Aggarwal, S., Mausam, and Chakrabarti, S. (2020). OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In The 58th Annual Meeting of the Association for Computational Linguistics (ACL), Seattle.

- Vashishth, S., Jain, P., and Talukdar, P. P. (2018). CESI: Canonicalizing open knowledge bases using embeddings and side information. Proceedings of the 2018 World Wide Web Conference.

Knowledge Graphs & The Open World Assumption

- Closed World Assumption (CWA): absence of a fact means it is necessarily false.
- Open World Assumption (OWA): absence of a fact does not imply fact is false. We simply do not know.

Knowledge Graphs adopt this assumption

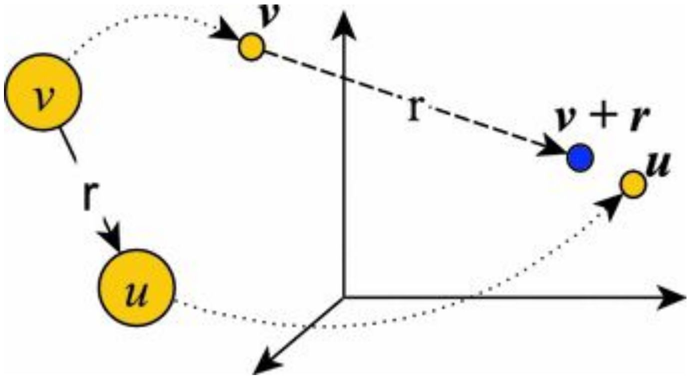


And so, Knowledge Graphs are inherently **incomplete** - Knowledge Graph Completion is an important task.

Knowledge Graph Embeddings

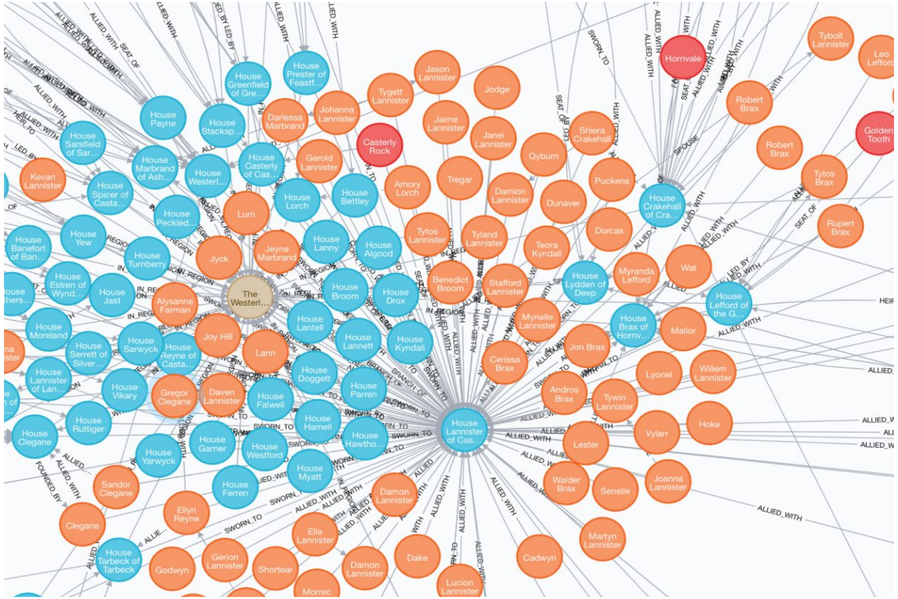
KG triple <v, r, u >

- Embed components of KG (entities, relations) into continuous vector spaces
- Allow easy manipulation of data while preserving inherent structure of KG
- Capture the interactions between entities of KG



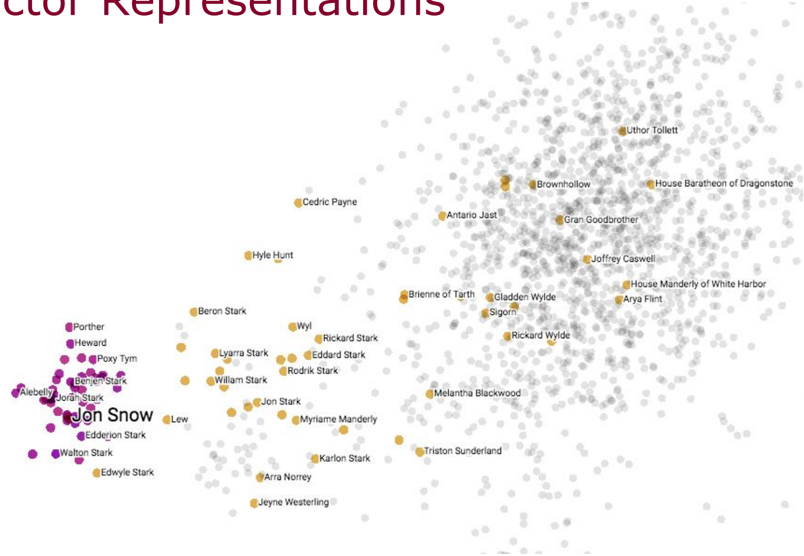
Translation based KG embedding

Knowledge Graph Embeddings



From Nodes and Edges …

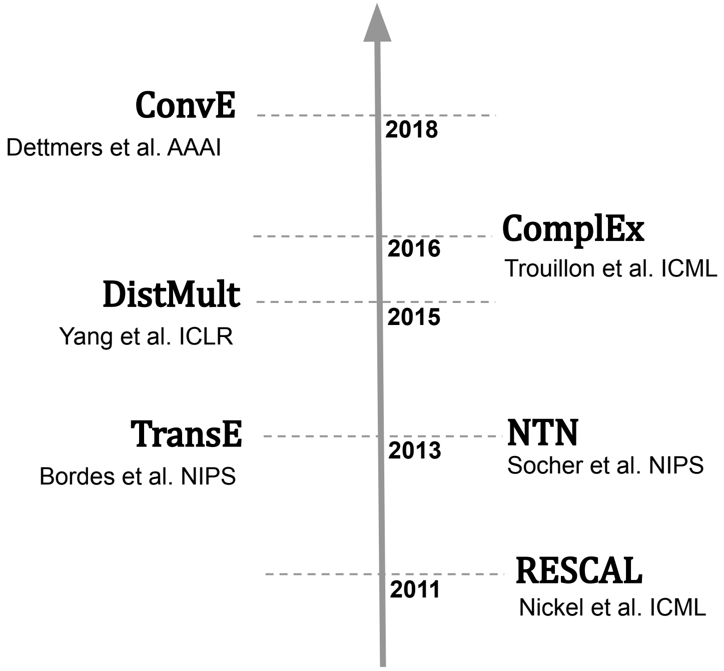
… to Semantically Meaningful Vector Representations



Borrowed from kge-tutorial-ecai2020.github.io

Popularity of KG embeddings

- Many embedding models
  - TransE
  - RESCAL
  - DistMult
  - ComplEx
  - ConvE
  - …
- Several new models being proposed every year ..

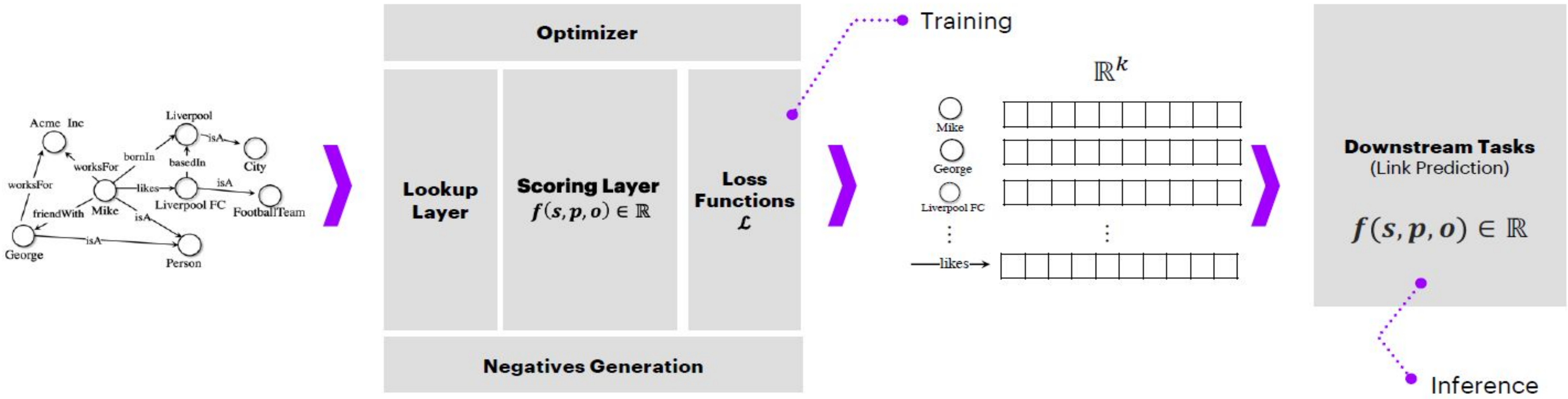


Applications are widespread..

- KG embeddings are being explored for various semantic tasks
  - Entity similarity (Sun et al. VLDB 2020)
  - Relation similarity (Kalo et al. ISWC 2019)
  - Conceptual clustering (Gad-Elrab et al. ISWC 2020)
  - Rule-based reasoning (Ho et al. 2018)

- All attempt to leverage semantic knowledge encoded in embeddings

At a Glance



Scoring function f : Assigns a score to a triple (s,p,o)

High score : high chances for the triple to be a true fact.

Examples - TransE

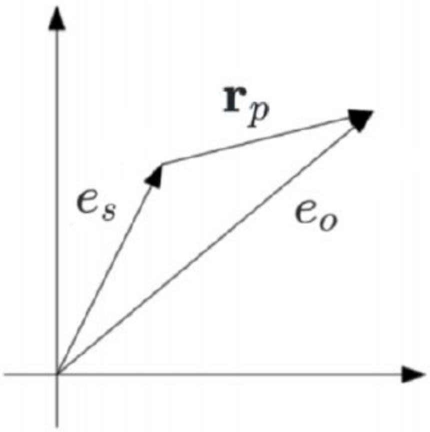
TransE - Translation-based scoring function

- representative translational distance model
- represents entities and relations as vectors in the same semantic space of dimension *d*
- *d* is the dimension of the target space with reduced dimension

In terms of vector computation, subject + relation ≈ object .

$$f_{TransE} = -||(\mathbf{e}_s + \mathbf{r}_p) - \mathbf{e}_o||_n$$

Bordes et al. 2013

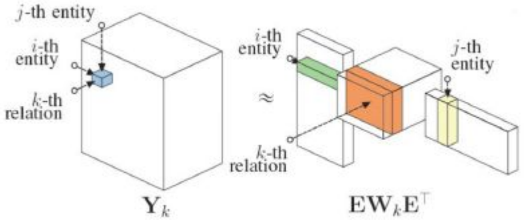


Other KG embedding models

- **RESCAL**: low-rank factorization with tensor product

[Nickel et al. 2011]

$$f_{RESCAL} = \mathbf{e}_s^T \mathbf{W}_r \mathbf{e}_o$$



- **DistMult**: bilinear diagonal model. Dot product.

[Yang et al. 2015]

$$f_{DistMult} = \langle \mathbf{r}_p, \mathbf{e}_s, \mathbf{e}_o \rangle$$

- **ComplEx**: Complex Embeddings (Hermitian dot product): (i.e. extends DistMult with dot product in $\mathbb{C}$)

$$f_{ComplEx} = Re(\langle \mathbf{r}_p, \mathbf{e}_s, \overline{\mathbf{e}_o} \rangle)$$

[Trouillon et al. 2016]

Evaluation Metrics

Mean Rank (MR)

$$MR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} rank_{(s,p,o)_i}$$

Mean Reciprocal Rank (MRR)

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_{(s,p,o)_i}}$$

Hits@N

$$Hits@N = \sum_{i=1}^{|Q|} 1 \text{ if } rank_{(s,p,o)_i} \leq N$$

Reading and resources

- Gerhard Weikum, Xin Luna Dong, Simon Razniewski and Fabian Suchanek (2021), "Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases", Foundations and Trends® in Databases: Vol. 10: No. 2-4, pp 108-490. http://dx.doi.org/10.1561/1900000064 - **Chapter 8 (8.2, 8.4)**
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NeurIPS. pp. 2787–2795 (2013)
- M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in Proc. 28th Int. Conf. Mach. Learn., 2011, pp. 809–816.
- B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in Proc. Int. Conf. Learn. Represent., 2015.
- T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in Proc. 33rd Int. Conf. Mach. Learn., 2016, pp. 2071–2080
- Wang, Quan, et al. "Knowledge graph embedding: A survey of approaches and applications." IEEE Transactions on Knowledge and Data Engineering 29.12 (2017): 2724-2743.

KG Topics (Part 1)

- Knowledge Graph Use cases
  - DBpedia
  - Yago
  - NELL
- Open Information Extraction
  - MinIE, ClausIE, OpenIE
  - Open Language Learning for Information Extraction
- Knowledge Graph Embeddings
  - TransE,
  - ConvE
  - RotatE

Language Models

Language Models

- Probabilistic Language Modeling and N-GRAMs
- RNN
- Transformers

Probabilistic Language Modeling

- Sequence W= $w_1,w_2,w_3,w_4,w_5 \ldots w_n$
- Goal: Compute P(W) = $P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$
- Chain rule:
  - P(A,B)=P(A)P(B|A)
  - P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)

$$\boxed{P(w_1 w_2 \ldots w_n)} = \prod_i \boxed{P(w_i \mid w_1 w_2 \ldots w_{i-1})}$$

Example

- "Today is Monday"
- P(today,is,Monday)=

  P(today) x P(is|today) x P(Monday|today,is)

- "Sesame Street is a long running American educational children's television ___"
  - P(series|Sesame,Street,is,a,long,running,American,educational children's,television)
    >
    P(dog|Sesame,Street,is,a,long,running,American,educational children's,television)

Maximum Likelihood Estimator

- Estimate $$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

  using counts in a training corpus

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- Problems
  - # of possibilities
  - Unseen words (OOV)
  - Zeros

N-grams

- Markov assumption

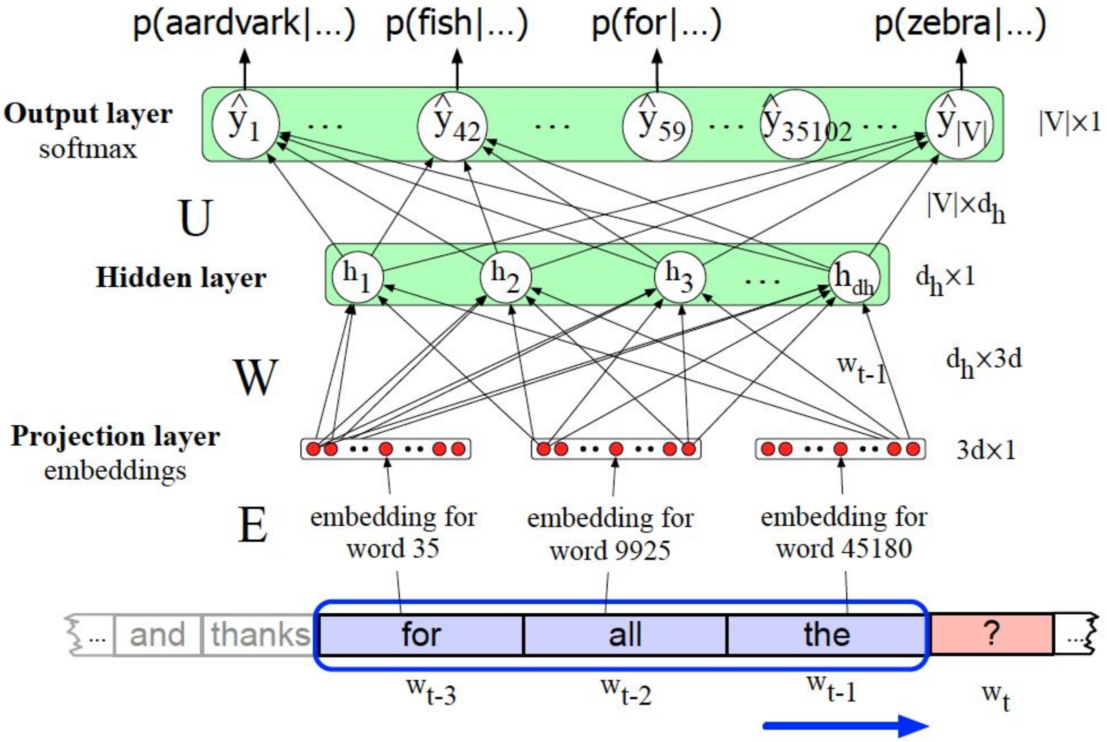
$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1})$$

- Examples
  - Unigram P(today,is,Monday)=P(today)P(is)P(Monday)
  - Bigram P(today,is,Monday)=P(today)P(is|today)P(Monday|is)
  - …

Unknown words (OOV)

- Smoothing
  - Add-one estimation $$P_{Add-1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + 1}{c(w_{i-1}) + V}$$
  - Back-off
    - e.g. If bigram count = 0, use unigram
  - Interpolation $$\begin{aligned}\hat{P}(w_n|w_{n-2}w_{n-1}) &= \lambda_1 P(w_n|w_{n-2}w_{n-1}) \\ &+ \lambda_2 P(w_n|w_{n-1}) \\ &+ \lambda_3 P(w_n)\end{aligned} \qquad \sum_i \lambda_i = 1$$
  - Cached history $$P_{CACHE}(w \mid history) = \lambda P(w_i \mid w_{i-2} w_{i-1}) + (1-\lambda)\frac{c(w \in history)}{\mid history \mid}$$
  - Discounting
  - Kneser-Ney Smoothing

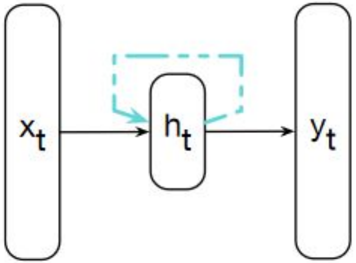
Neural Language Models

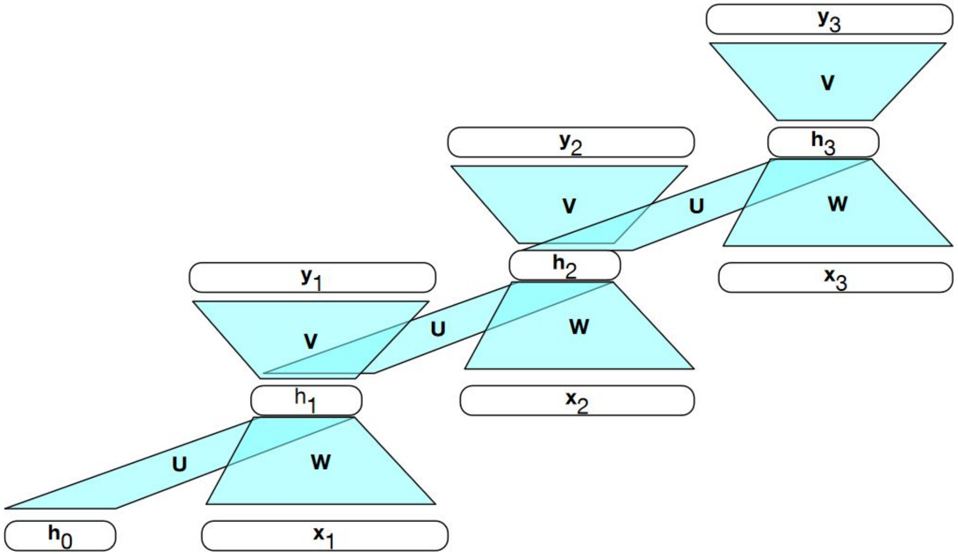


- N-Grams
- Depends on weights instead of counts
- Tokenizer could be used to overcome OOV issues
- Can include UNK tokens without zeroing probability

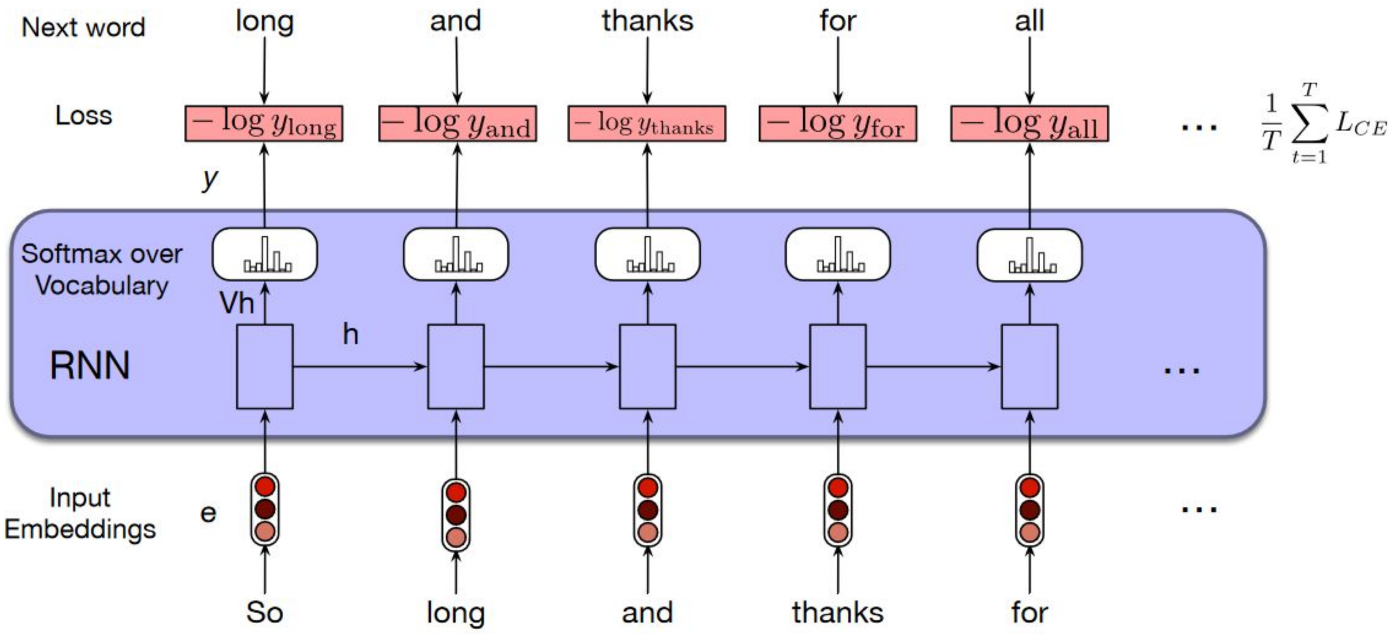
Recurrent Neural Networks RNNs

- Simple structure
- Arbitrary length sequences
- One item (token) at a time



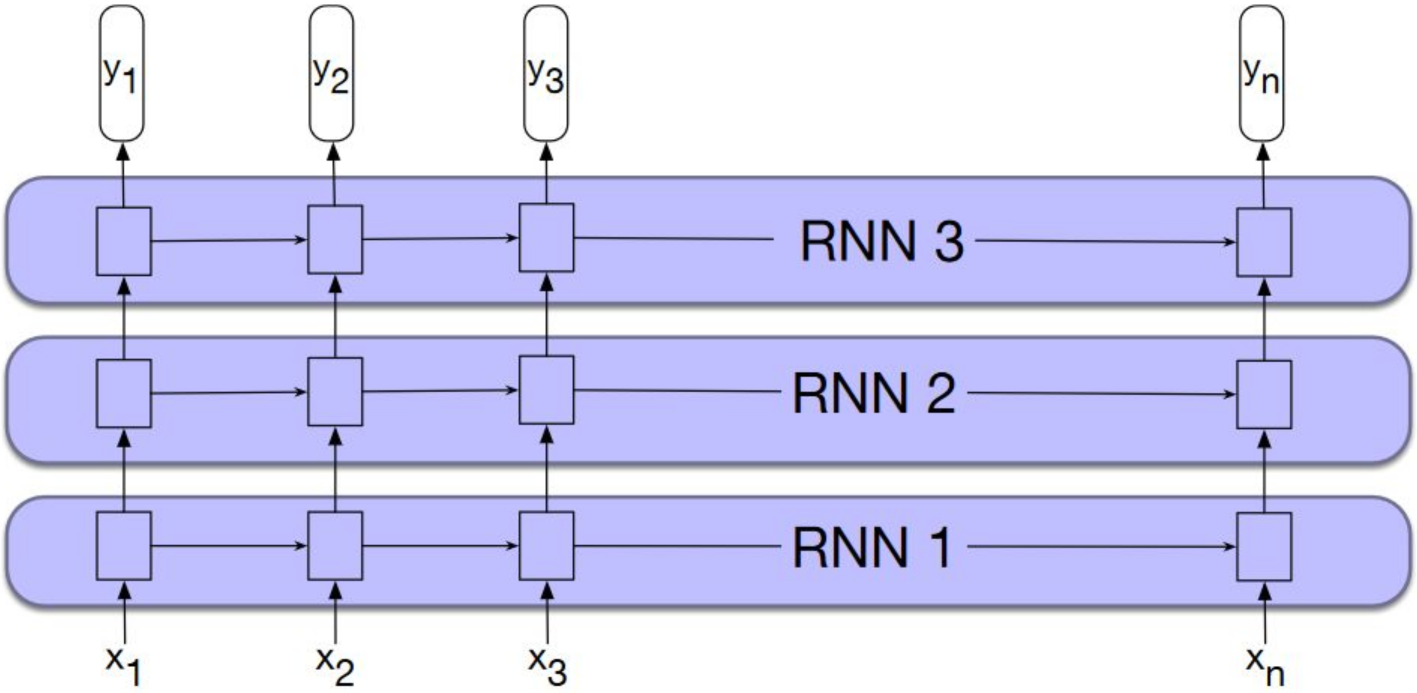


RNN for LM

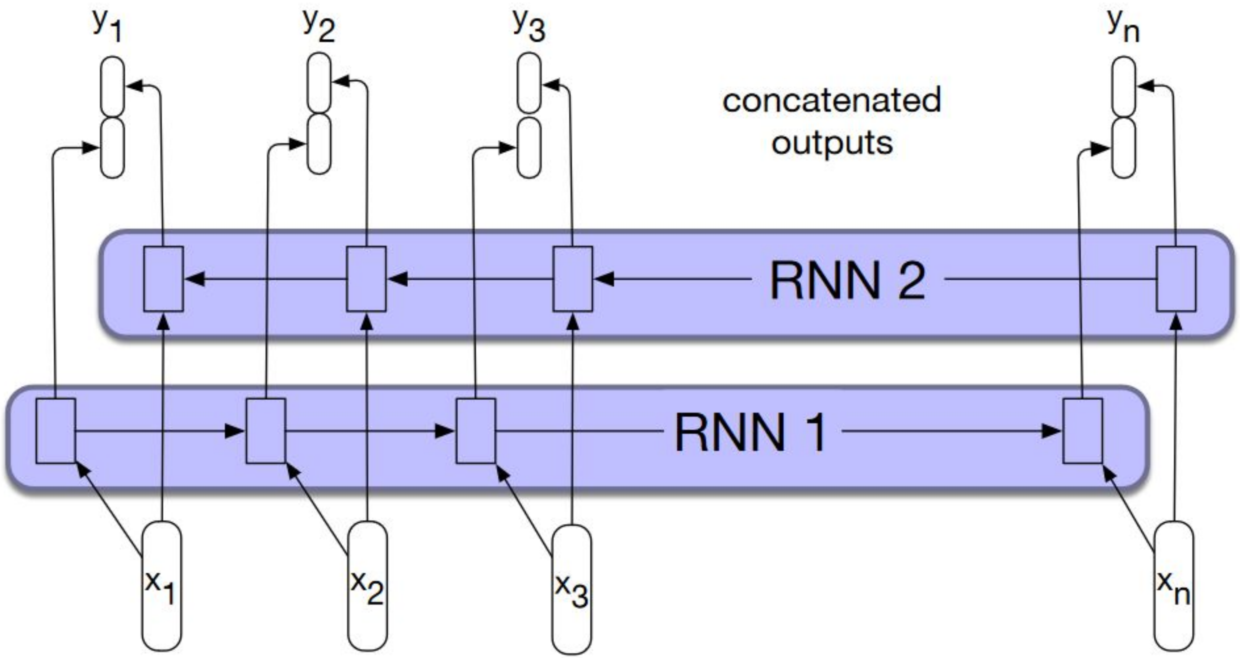


Next word
long
and
thanks
for
all
Loss
$-\log y_{\text{long}}$
$-\log y_{\text{and}}$
$-\log y_{\text{thanks}}$
$-\log y_{\text{for}}$
$-\log y_{\text{all}}$
...
$\frac{1}{T}\sum_{t=1}^{T} L_{CE}$
y
Softmax over Vocabulary
Vh
h
RNN
...
Input Embeddings
e
...
So
long
and
thanks
for

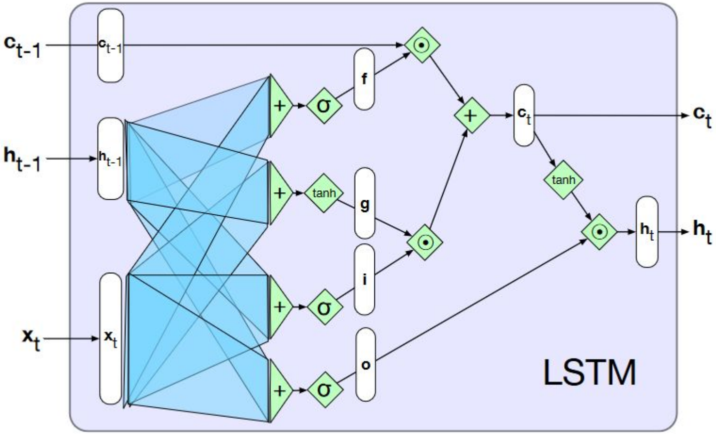
Stacked RNNs



Bidirectional RNNs



Long Short Term Memory



One LSTM Cell

With simple RNNs

- The h vector contains
  - Local prediction
  - Previous context (Tends to vanish)

LSTM

- Separate context from hidden vector
- Gates to selectively forget information from context

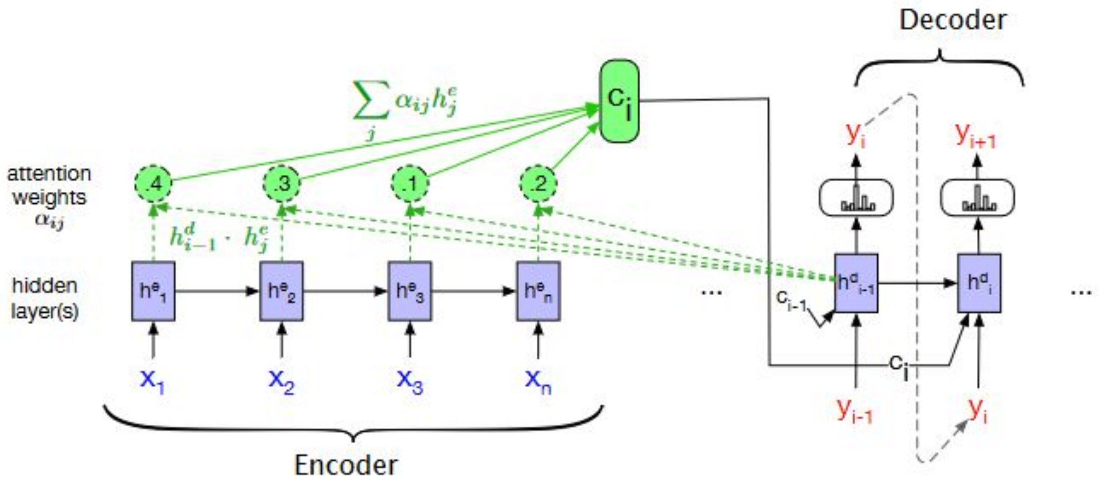
LSTM Text Generation

LSTM generated Wikipedia article

```
Naturalism and decision for the majority of Arab countries' capitalide was grounded
by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated
with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal
in the [[Protestant Immineners]], which could be said to be directly in Cantonese
Communication, which followed a ceremony and set inspired prison, training. The
emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom
of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known
in western [[Scotland]], near Italy to the conquest of India with the conflict.
Copyright was the succession of independence in the slop of Syrian influence that
was a famous German movement based on a more popular servicious, non-doctrinal
and sexual power post. Many governments recognize the military housing of the
[[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]],
that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humah.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery
was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]
```


Source http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Attention Mechanism in seq2seq RNNs

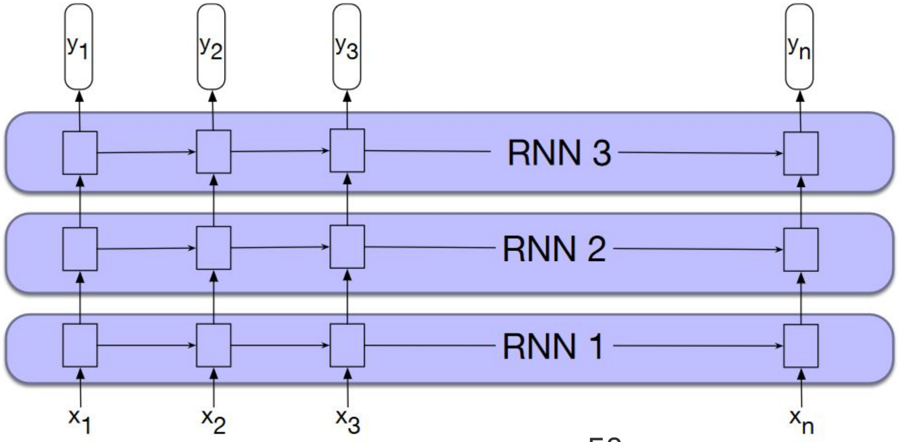


The context pays more/less attention to different encoder tokens

Allows representation of longer dependencies

Issues with RNNs

- Hard to train. One token at a time.
  - h_{t-1} must be computed before h_t can be computed
  - Same for backpropagation
- Vanishing/exploding gradient
  - When unrolled a RNN is a very deep NN
  - The terms are repeated in the gradient
  - Numerical instability



Transformers – Self-attention

- (Causal) Self-attention

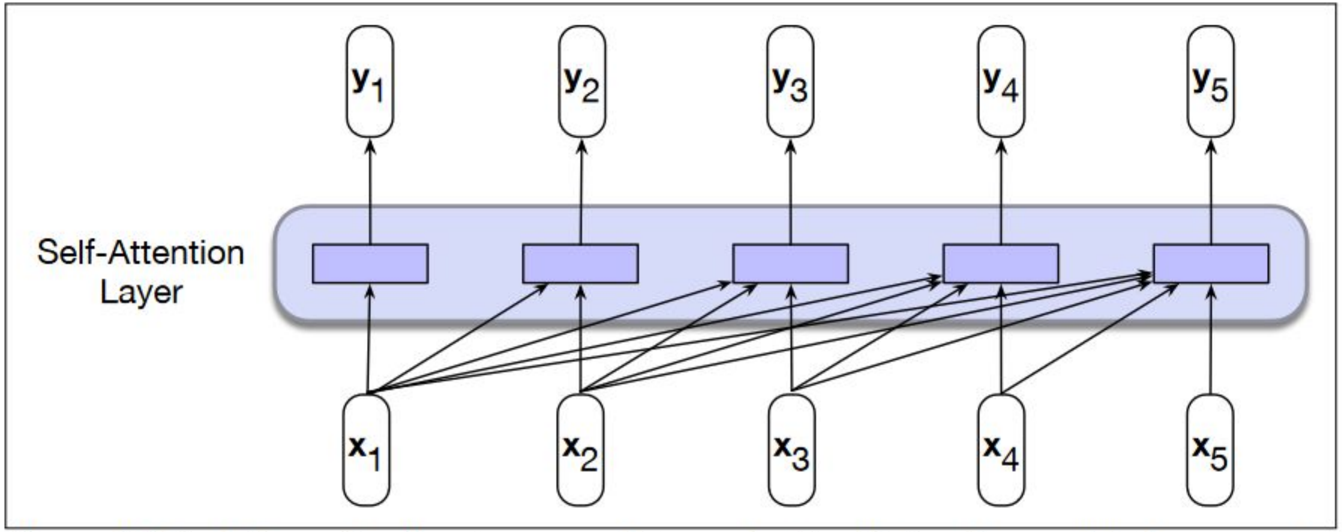
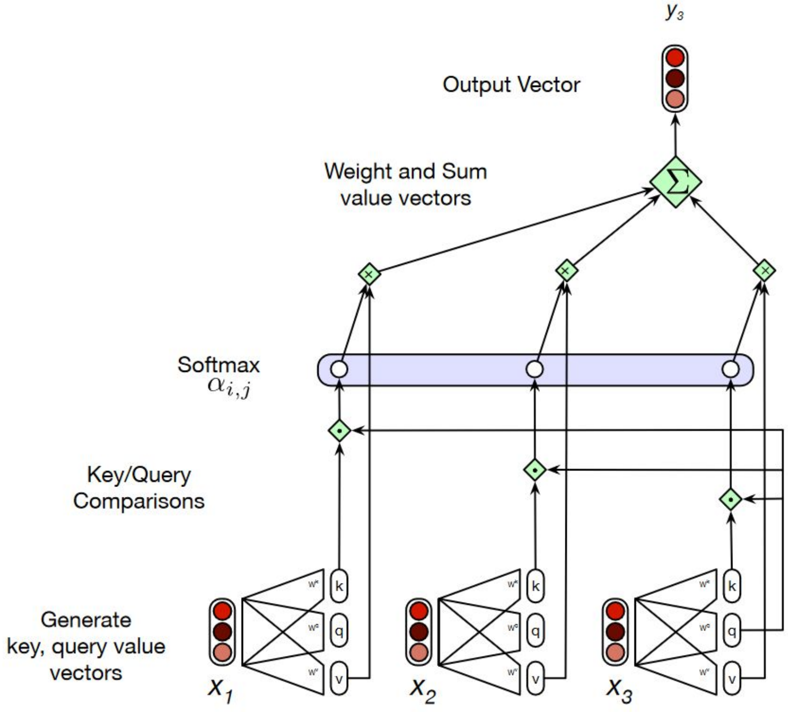


Figure 9.15 Information flow in a causal (or masked) self-attention model. In processing each element of the sequence, the model attends to all the inputs up to, and including, the current one. Unlike RNNs, the computations at each time step are independent of all the other steps and therefore can be performed in parallel.

Issues with RNNs



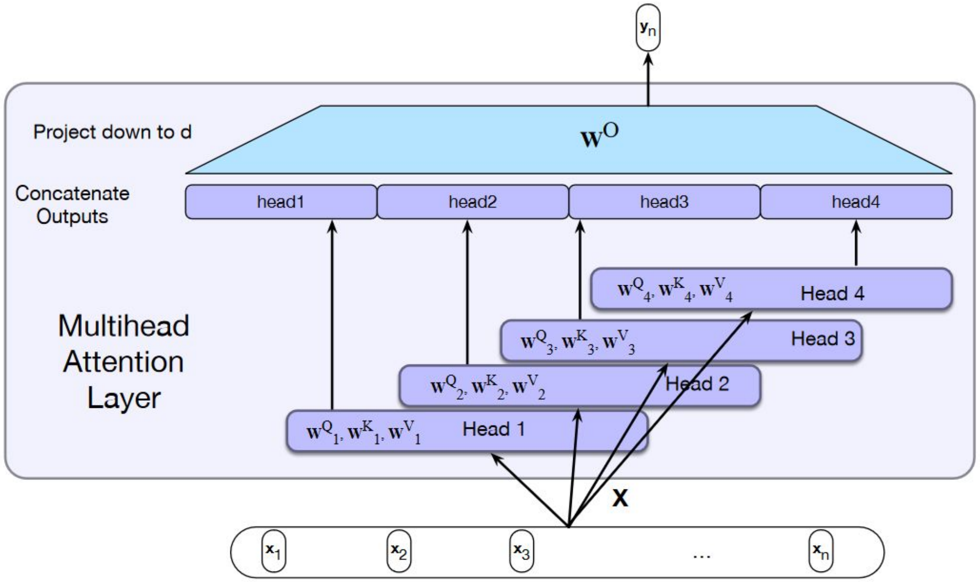
$$\mathbf{y}_i = \sum_{j \le i} \alpha_{ij} \mathbf{v}_j$$

$$\text{score}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}}$$

$$\mathbf{q}_i = \mathbf{W}^{\mathbf{Q}}\mathbf{x}_i;\ \ \mathbf{k}_i = \mathbf{W}^{\mathbf{K}}\mathbf{x}_i;\ \ \mathbf{v}_i = \mathbf{W}^{\mathbf{V}}\mathbf{x}_i$$

Calculating the value of y3, the third element of a sequence using causal (left-to-right) self-attention

Complete Transformer



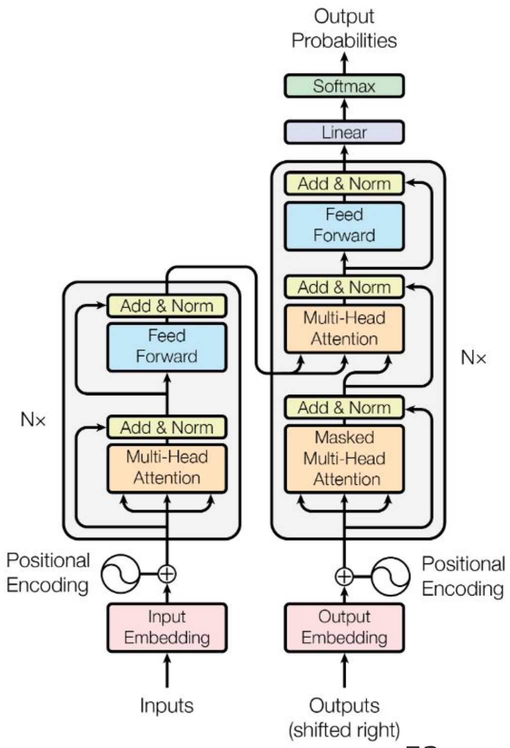
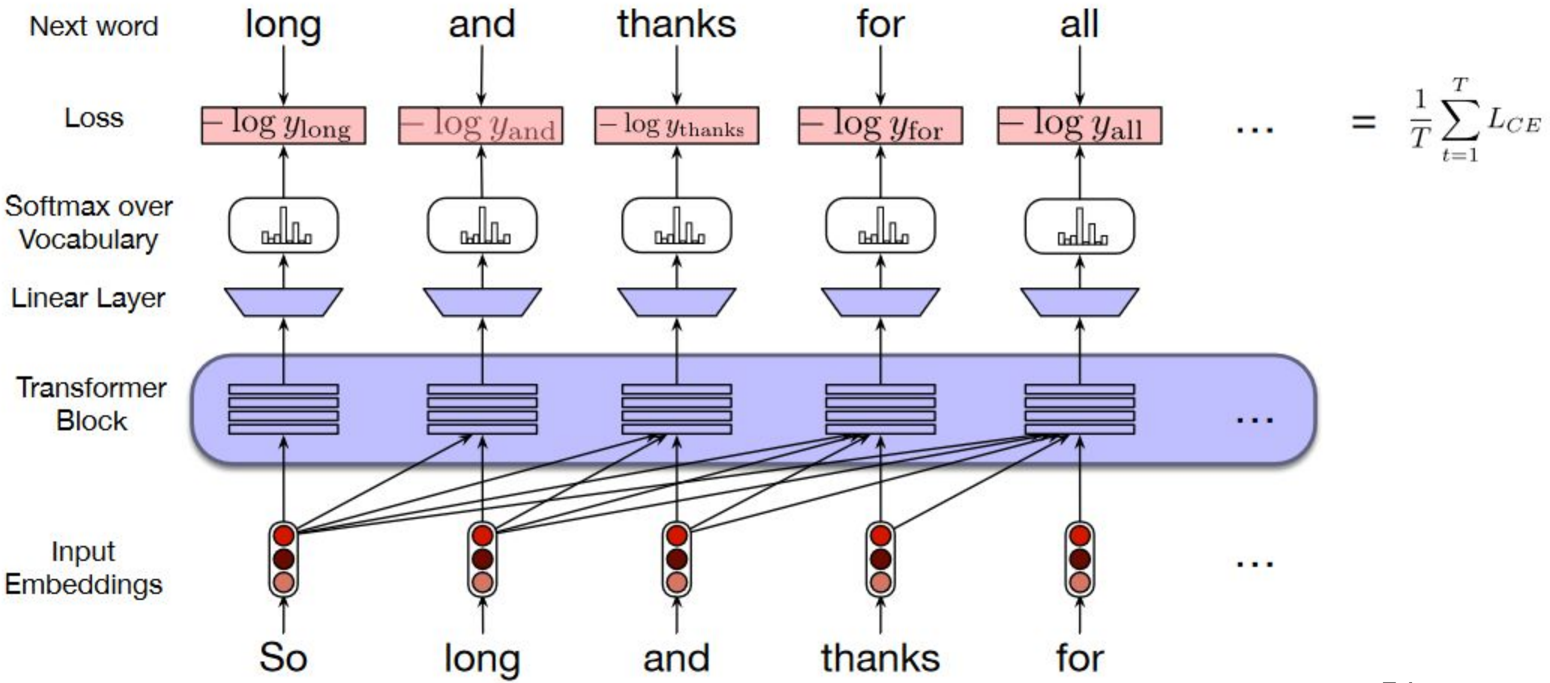
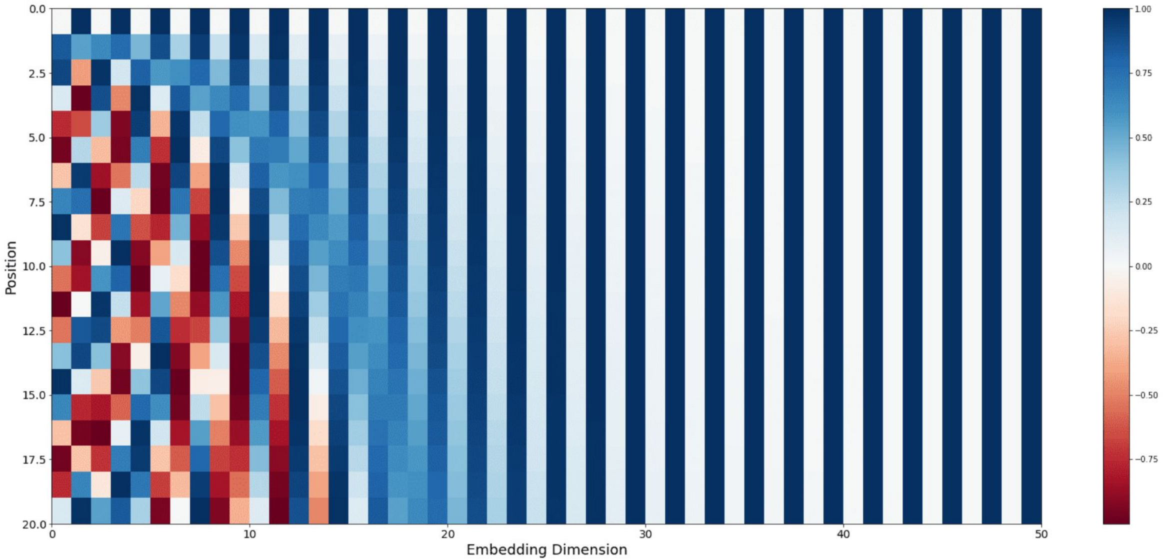


Figure 1: The Transformer - model architecture.


Next word
long
and
thanks
for
all
Loss
$-\log y_{\text{long}}$
$-\log y_{\text{and}}$
$-\log y_{\text{thanks}}$
$-\log y_{\text{for}}$
$-\log y_{\text{all}}$
...
$= \frac{1}{T}\sum_{t=1}^{T} L_{CE}$
Softmax over Vocabulary
Linear Layer
Transformer Block
...
Input Embeddings
...
So
long
and
thanks
for

Positional Encoding

- Originally a combination of *sin* and *cos* with different wavelengths
- Added to vectors
- Encode relative distance between positions



https://erdem.pl/2021/05/understanding-positional-encoding-in-transformers

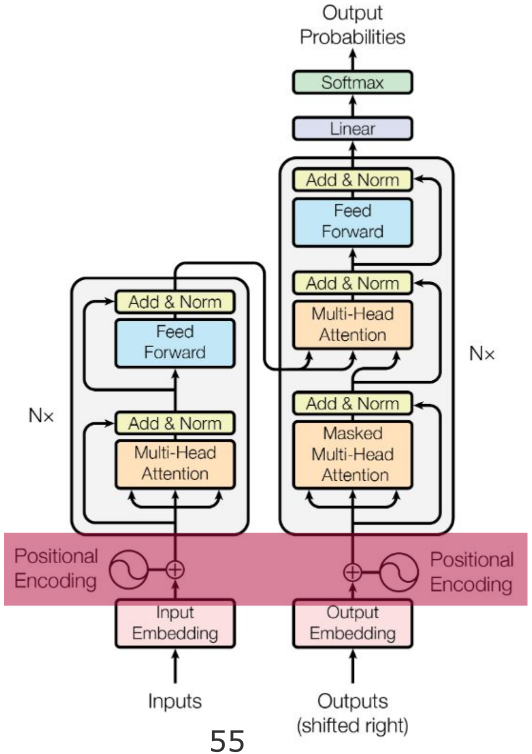
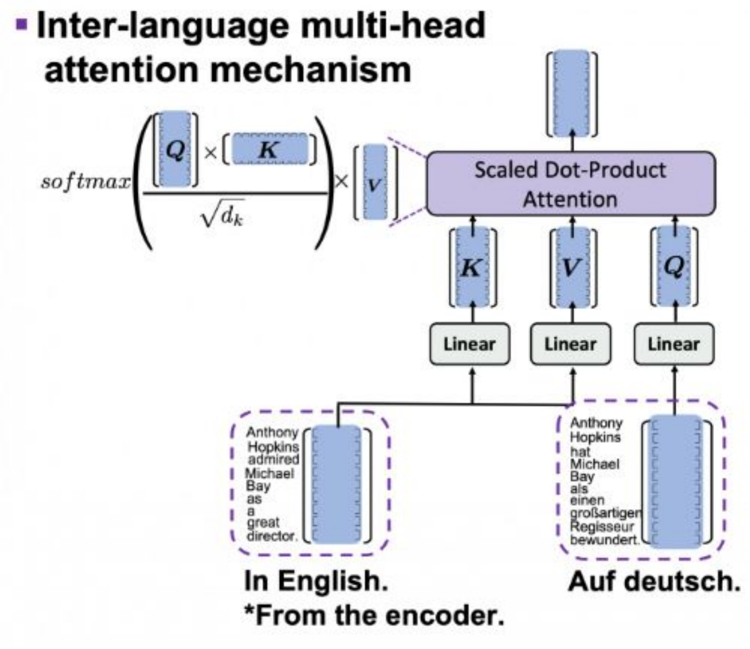
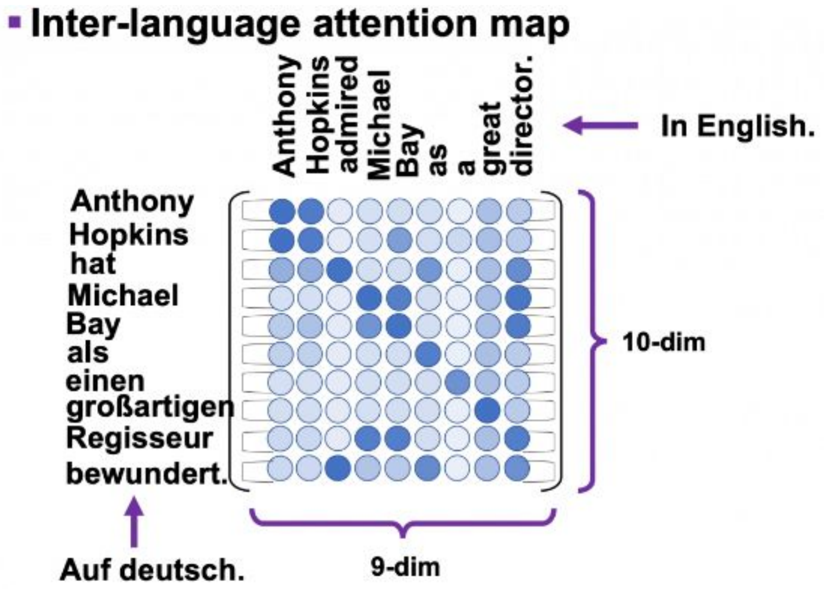


Figure 1: The Transformer - model architecture.

Cross-attention

- Attention from decoder tokens to encoder tokens





Source: https://data-science-blog.com/blog/2021/05/10/constructing-transformer-model-little-by-little/

Transformers

- Advantages
  - Can be computed in parallel
  - # of parameters
- Disadvantages
  - Fixed sequence length
  - # of parameters

Summary

- LMs estimate probabilities of sequences
  - Or which is the most probable token given some history
- Markov assumption (**N-GRAMS**)
- MLE (Counts) with **smoothing**
- FFNNs to compute probabilities without counting
- **RNNs** to represent text as a **sequence** of repeated transformations given a context.
  - LSTM
- Transformers (**Self-attention**)
  - Contextualized embeddings
  - Machine translation

References

- Dan Jurafsky and James H. Martin, **Speech and Language Processing** (3rd ed. draft) https://web.stanford.edu/~jurafsky/slp3/ **(Chapters 3, 7,9)**
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need**. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. **Deep Contextualized Word Representations**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics

Literature

Knowledge Graphs

Gerhard Weikum, Xin Luna Dong, Simon Razniewski and Fabian Suchanek (2021), "Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases", Foundations and Trends® in Databases: Vol. 10: No. 2-4, pp 108-490. http://dx.doi.org/10.1561/1900000064 **(Chapter 1)**

Language Models

Dan Jurafsky and James H. Martin, "Speech and Language Processing" (3rd ed. draft) https://web.stanford.edu/~jurafsky/slp3/ **(Chapter 9)**

Language Models As or For Knowledge Bases

Simon Razniewski , Andrew Yates , Nora Kassner and Gerhard Weikum
https://arxiv.org/abs/2110.04888