Week 7: Issues and bias in NLP

Text Analytics and Natural Language Processing
Instructor: Benjamin Batorsky

Course evaluations

Now available on Canvas!

Please take a moment to fill these out!

Final project **DUE 8/7 (11:59pm)**

- Submission format
  - Github repo or set of files
    - Code for exploration, processing, modeling (notebooks, scripts)
    - Data files
    - Write-up
      - All project participants
      - Link to repo (make it public!) or reference to specific files
      - All sections from outline, but expanded with results, conclusions, etc
  - MUST include a runnable version
    - Example: Subsample of full data, simplified code
- On Canvas: Submit write-up OR sign-up for a presentation

Presentations (Class time 8/5)

- Schedule to be announced on Tuesday
- Let us know if there’s any issue with your time slot

Adapting our sentiment model for NLG

What is bias?

- Bias, more generally
  - Weight in favor/against a particular thing/idea
  - Typically negative: May be based on limited data or in spite of data
- Statistical bias
  - “Statistical bias is a feature of a statistical technique or of its results whereby the expected value of the results differs from the true underlying quantitative parameter being estimated” - Wikipedia
- Machine learning bias
  - Same concept as statistical bias
  - Usually used in terms of inaccuracy
    - Underfit model = High bias, low variance
    - Overfit model = Low bias, high variance

(Some) Types of bias in ML

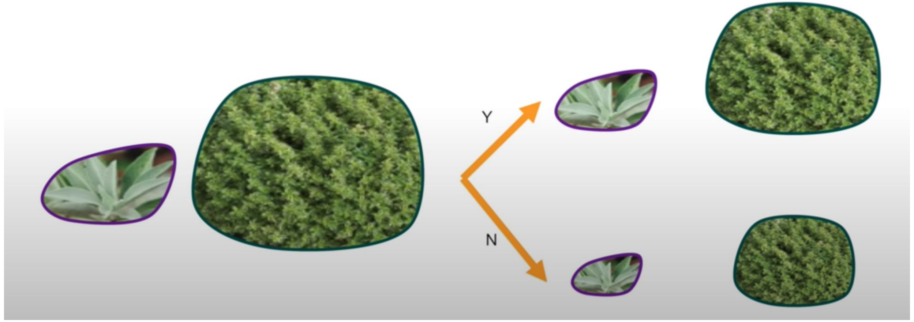
- Historical Bias
    - Already existing bias and socio-technical issues in the world represented in data
    - Example: Incarceration rates of different populations as a product of institutional bias
- Representation Bias
    - Results from the way we define and sample from a population
    - Example: ImageNet contains certain types of people doing certain activities, which will push models towards those representations
- Sampling Bias
    - Arises due to non-random sampling of subgroups
    - Example: Visitors to website during promotion may be different than visitors after
- Aggregation Bias
    - Drawing potentially false conclusions about some subgroups based on other subgroups.
    - Example: Same disease, different populations, different trajectories

[1908.09635] A Survey on Bias and Fairness in Machine Learning

(Some) Definitions of fairness

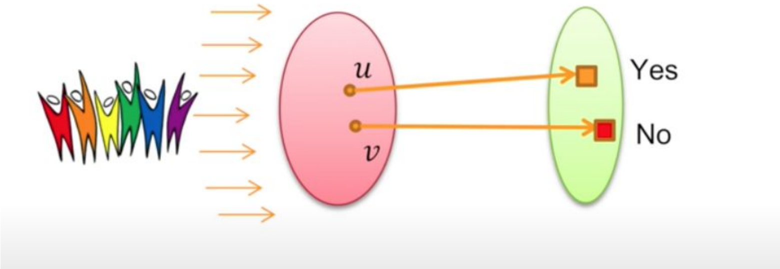
Group-level fairness

Statistical parity



Individual-level fairness

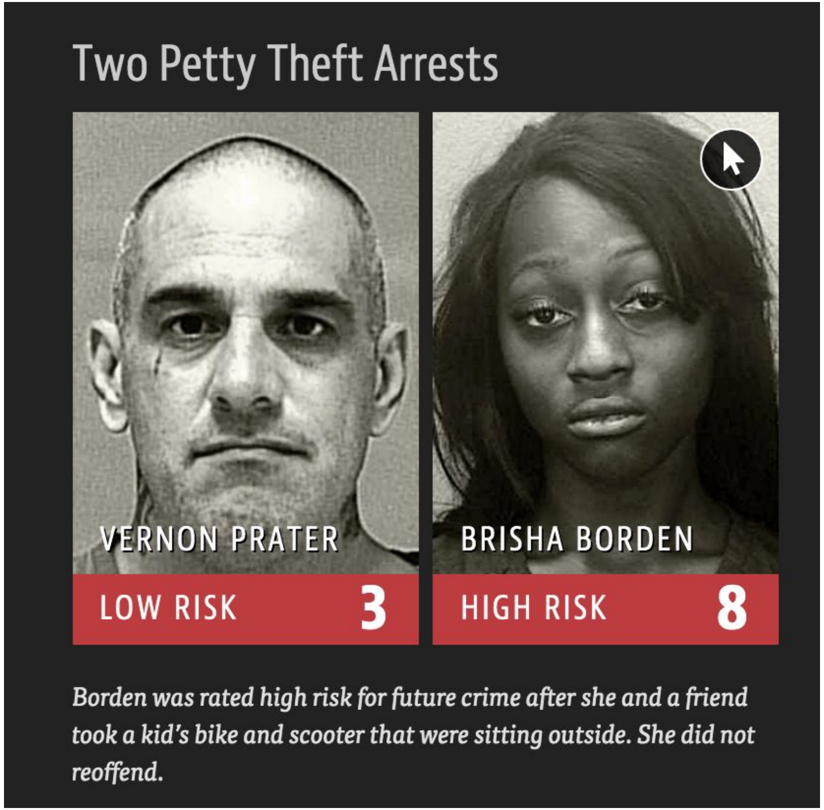
Similar individual = similar outcome



Cynthia Dwork - Finding Fairness (https://www.youtube.com/watch?v=i_avLd49f8I&feature=youtu.be&t=1548)

What happens when we don’t monitor bias?

- COMPAS model to predict risk of reoffending
- 2016: Propublica investigation
  - Black defendants predicted 77% more likely to commit violent crime than white defendants
  - 48% of white defendants who DID reoffend labelled low risk (vs 28% for black defendants)
- Raised some question on these types of algorithms
  - Should they be used?
  - Who should be responsible for monitoring bias?
  - How do we ensure accountability if there is limited transparency?



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Fairness experiment in notebook

Addressing bias in prediction

- Multi-accuracy targets
  - Breaking down metrics by different groups
  - But how do you determine which groups?
    - Likely groups most at-risk have limited representation
- Affirmative action
  - Ranking within subgroups
    - Example: Top 5% of high school class
    - More detail: Top 5% of high school class stratified by education level of mother
  - Individual pairing
    - Select pairs of individuals with similar traits
    - Predict outcome pair-wise, rather than individual
- Issue across all of these: Intersectionality
  - Dwork’s example: Fairness by sage/thyme-eating, but what about sage-eating coffee drinkers?
  - Difficult/impossible to ensure fairness across all sections

Improving our model (notebook)

Bias in word embeddings

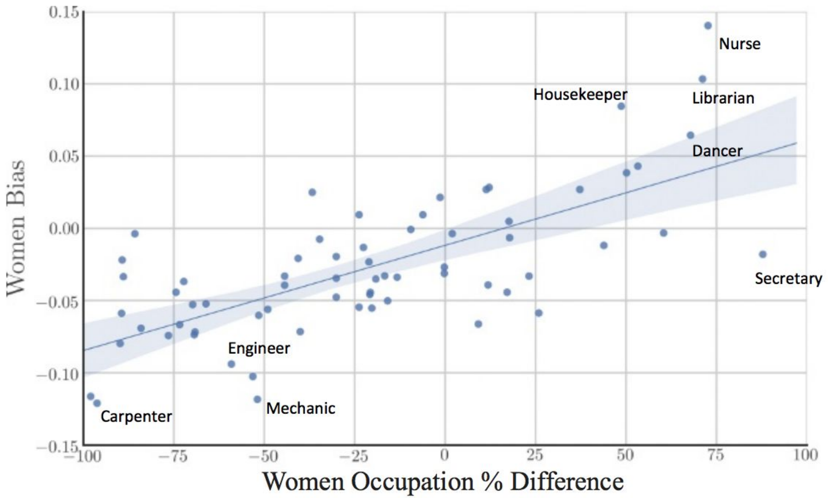


Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

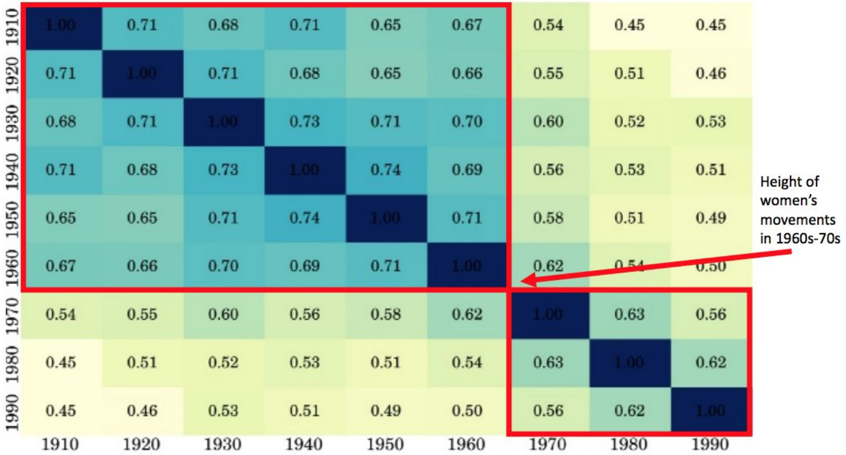


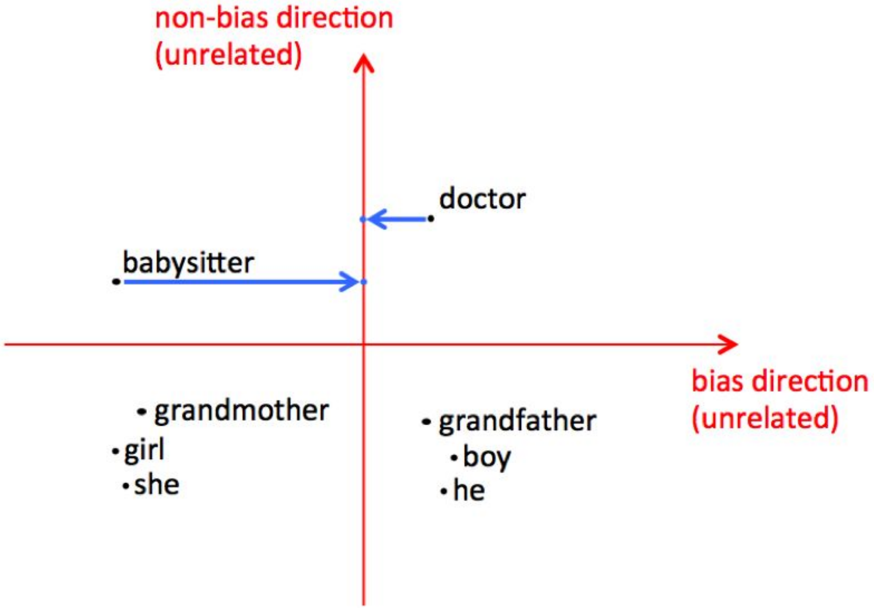
Fig. 4. Pearson correlation in embedding bias scores for adjectives over time between embeddings for each decade. The phase shift in the 1960s–1970s corresponds to the US women's movement.

Word embeddings quantify 100 years of gender and ethnic stereotypes
(https://www.pnas.org/content/pnas/115/16/E3635.full.pdf)

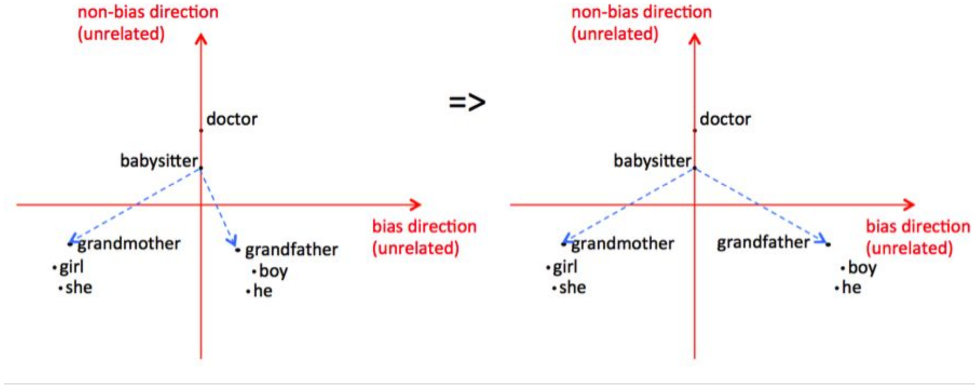
Identifying bias in word embeddings (notebook)

De-biasing word embeddings

Equalize



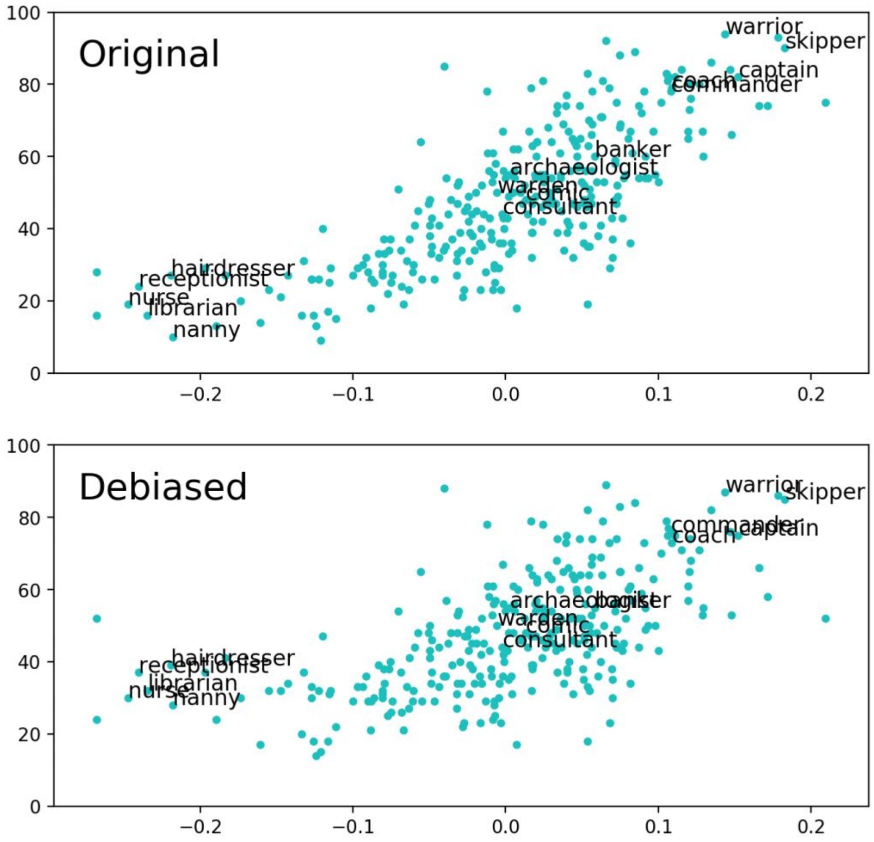
Equalize pairs



Concept: https://arxiv.org/pdf/1607.06520.pdf
Images: https://medium.com/machine-learning-bites/deeplearning-series-sentiment-classification-d6fb07b0da43

De-biasing word embeddings (and why it may not work!)

- Applying Bolukbasi's method
  - X axis = Original bias measure
    - In Bolukbasi: Projection on gender direction
    - Negative = more female
    - Positive = more male
  - Y axis = Number of "male"-associated neighbors
- Original not much different from debiased
  - Bias-neighbor correlation
    - Original: 0.75
    - Debased: 0.61
- Bias coded in all words, not just the ones selected for correction
  - Bias may just now be in a different direction

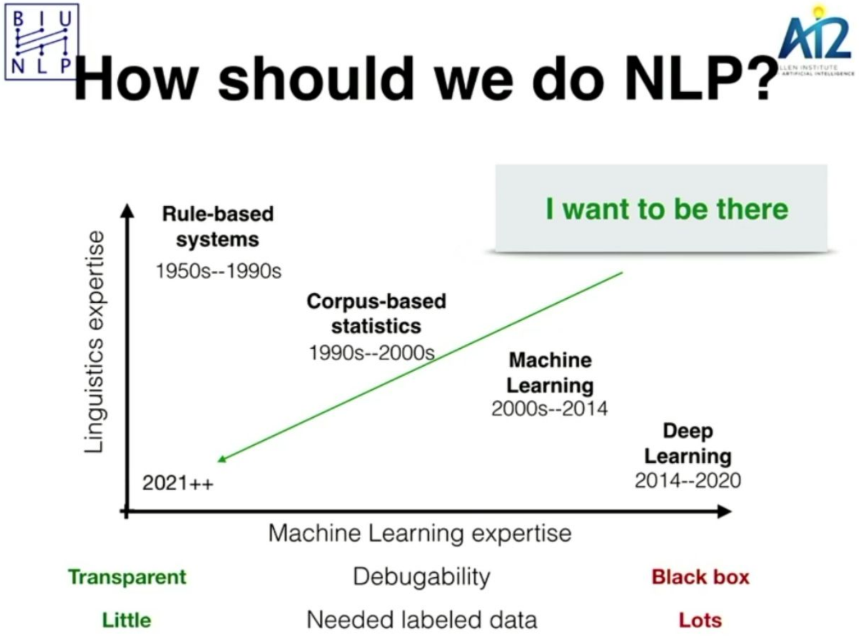


https://arxiv.org/pdf/1903.03862.pdf

Additional methods for monitoring/addressing bias

- StereoSet (https://stereoset.mit.edu/)
    - Three scores
        - Language model score
            - How good the model is at ranking “meaningful” associations over “meaningless” ones
                - “My housekeeper is a Mexican” vs “My housekeeper is a round”
        - Stereotype score
            - How often “stereotype” constructions are preferred over “antisterotype” constructions
        - Idealized CAT score
            - Combine the above two measures into a score from 0 to 1
- Deon (https://deon.drivendata.org/)
    - Ethics checklist
    - Five domains: Data collection, data storage, analysis, modelling, deployment

Deep learning and transparency



[Yoav Goldberg: The missing elements in NLP (spaCy IRL 2019)](#)

System	Citation	Performance
System A	Smith et al. 2018	76.05
System B	Li et al. 2018	75.85
System C	Petrov et al. 2018	75.62

https://hackingsemantics.xyz/2019/leaderboards/

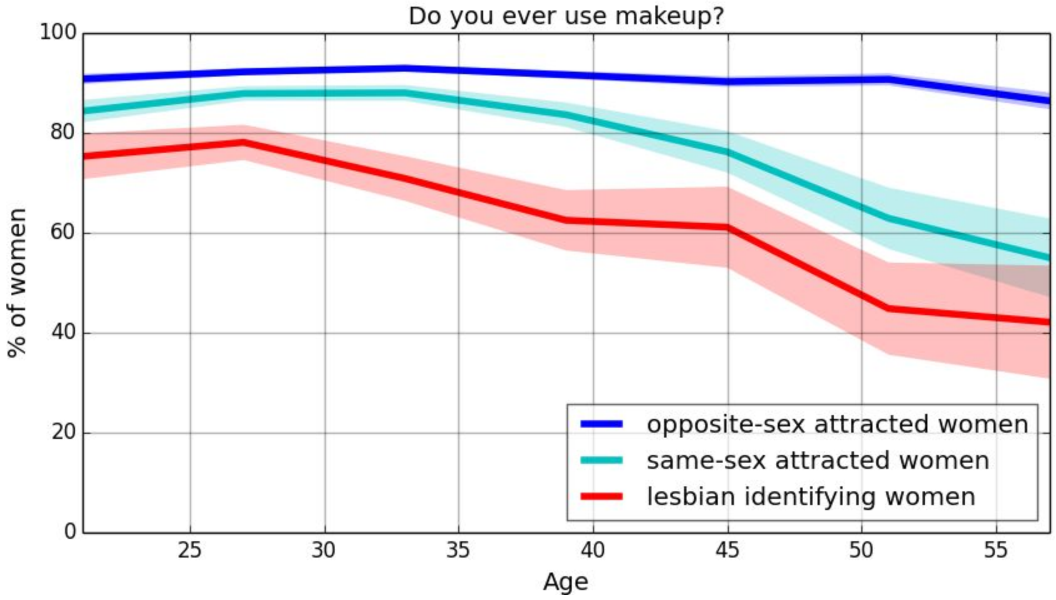
Best performance not always the “right” performance

	aeroplane	**bicycle**	**bird**	**boat**	**bottle**	**bus**	**car**
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	**chair**	**cow**	**diningtable**	**dog**	**horse**	**motorbike**
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

Interpretable & Transparent Deep Learning (http://www.heatmapping.org/slides/2019_NLDL.pdf)

This can be pretty pernicious...

Model (https://psyarxiv.com/hv28a/) able to distinguish between same-sex and opposite-sex attracted women 75% of the time



Do algorithms reveal sexual orientation or just expose our stereotypes?

And even BERT is guilty of this

- 2019 paper (Niven & Kao)
  - Argument comprehension task
    - Given claim and reason, pick warrant or alternative that supports claim
    - Fairly complex task (even for humans!)
  - Accuracy = 77%, new SOTA!
- Strong association between “is”, “do” and “not” and being selected as correct warrant
- With various parts of data removed, still provided similar predictions (with similar performance!)
- BERT wasn’t learning to understand, it was exploiting the dataset!

Claim	Google is not a harmful monopoly
Reason	People can choose not to use Google
Warrant	Other search engines don’t redirect to Google
Alternative	All other search engines redirect to Google

$$\textbf{Reason} \text{ (and since) } \textbf{Warrant} \rightarrow \textbf{Claim}$$
$$\textbf{Reason} \text{ (but since) } \textbf{Alternative} \rightarrow \neg\, \textbf{Claim}$$

https://www.aclweb.org/anthology/P19-1459.pdf

Some methods for diagnosis

- "Build it, break it" mentality
  - Shouldn't just be about getting SOTA, need further examination
  - Should be able hurt performance in anticipated ways
- Data ablations
  - Similar to Niven and Kao: Remove information, test performance
  - Dwork's idea: Performance within different subsets
- Adversarial attacks
  - Creation of observations to "test" robustness
  - Flipping of particular words (with synonyms)
  - Flipping of particular characters

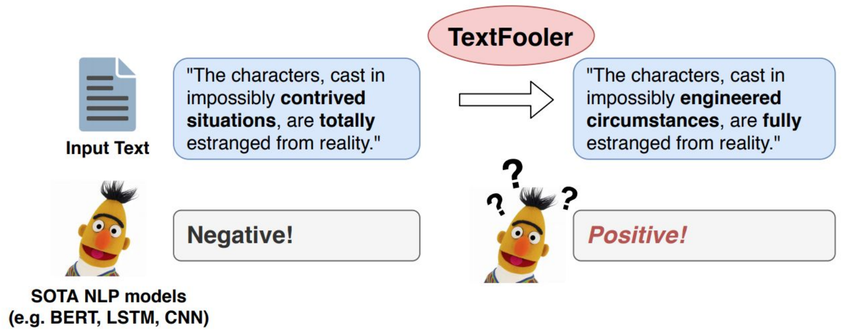
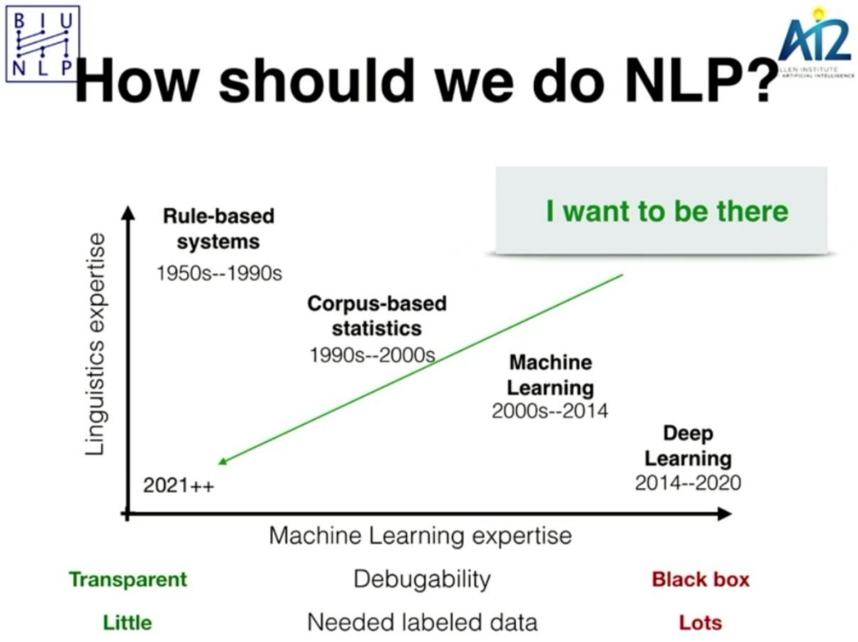


Figure 1: Our model TextFooler slightly change the input text but completely altered the prediction result.

https://arxiv.org/pdf/1907.11932.pdf

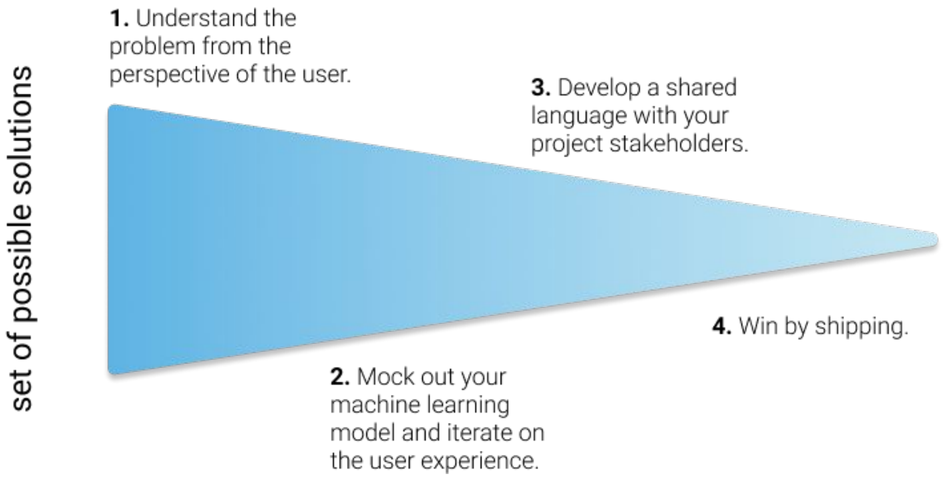
So how SHOULD we do NLP?



[Yoav Goldberg: The missing elements in NLP (spaCy IRL 2019)]

Think about the full development process

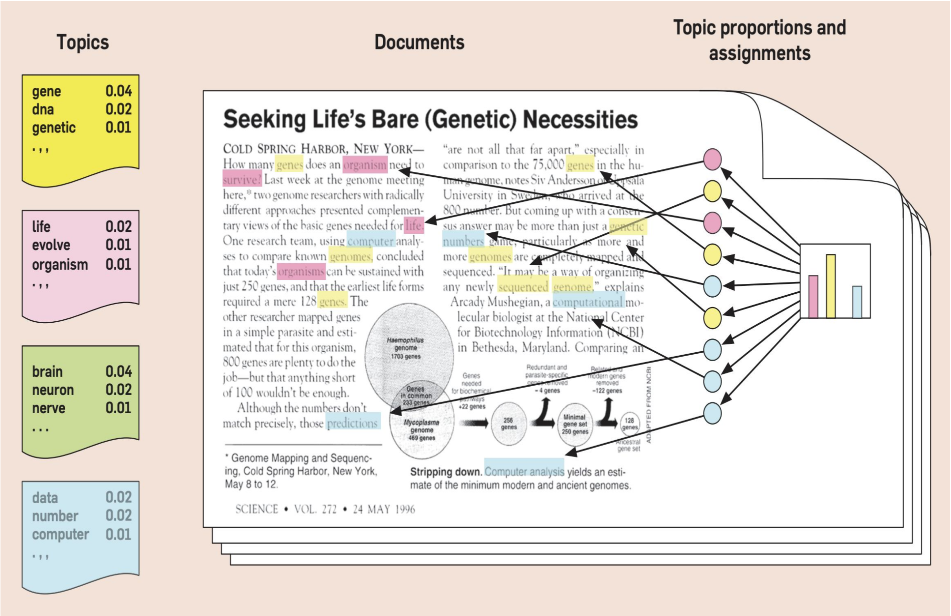
- Setup
  - Understand the problem
  - Inventory of solutions
    - Impact
    - Feasibility
    - Requirements
  - Setting up code base
- Data collection/labelling/sourcing
- Model exploration
- Deployment
- (throughout) Debugging and testing



https://www.jeremyjordan.me/ml-requirements/

Review your history

- 40s-50s: Machine translation era
- 60-70s: Shift towards semantic-driven processing
- 70s to 80s: Community expansion
- 90s-00s: Probabilistic/Statistical models
- 2000s: Neural Language models
- 2008: Multi-task learning
- 2013: Word embeddings
- 2014: Expansion of Neural models
- 2015: Attention
- 2018 and beyond: Language model advancements



http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

Make progress with simple approaches

Word counts

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

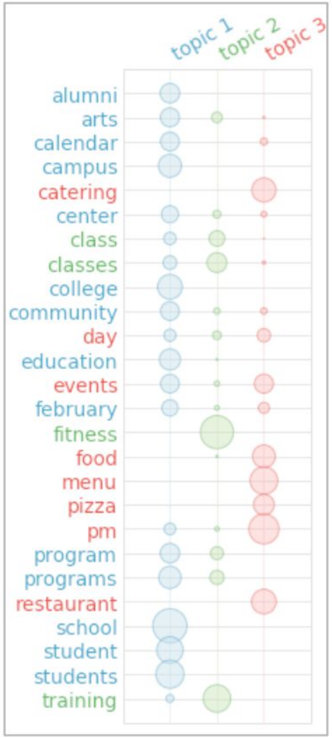
Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

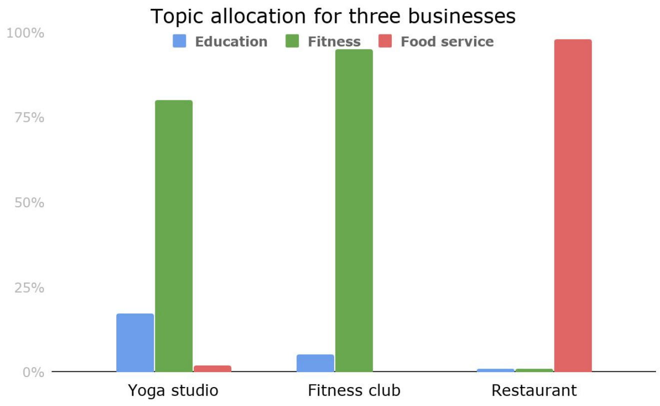
TF-IDF

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

Present and iterate on solutions

Product similarity

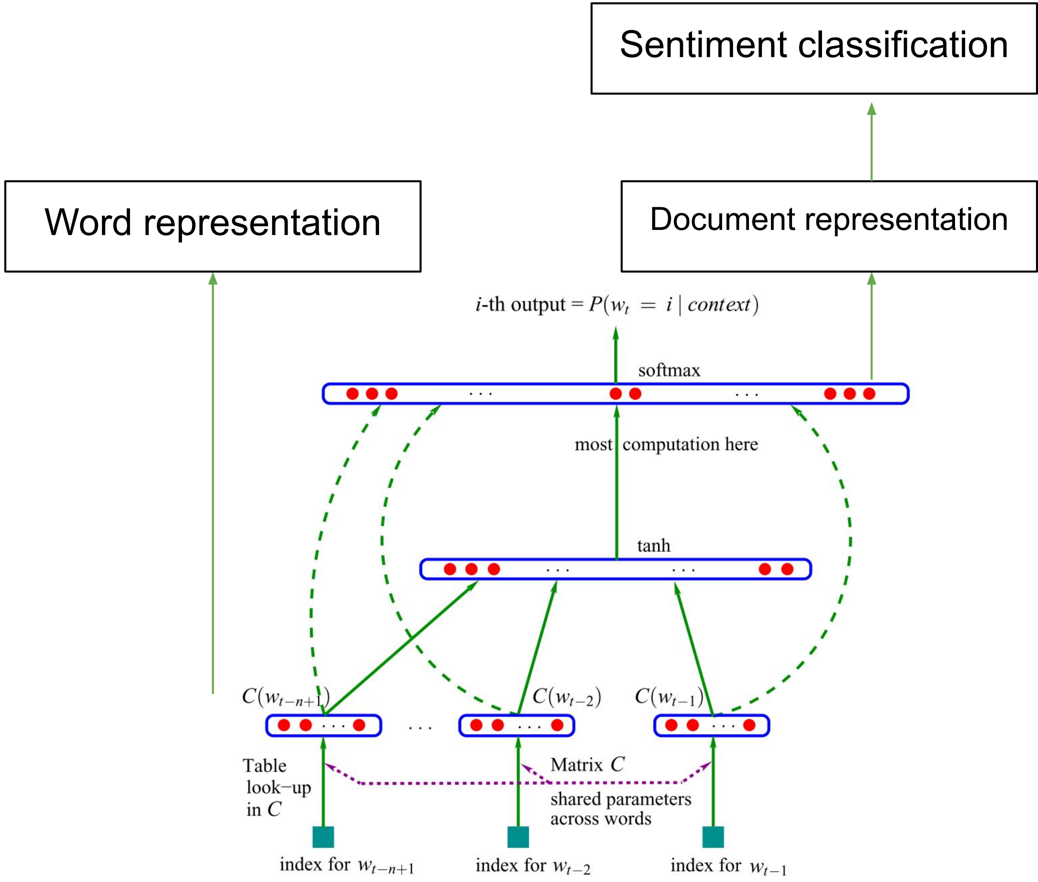




Circles are sized according to "relevance" to each topic

Review your history

- 40s-50s: Machine translation era
- 60-70s: Shift towards semantic-driven processing
- 70s to 80s: Community expansion
- 90s-00s: Probabilistic/Statistical models
- 2000s: Neural Language models
- 2008: Multi-task learning
- 2013: Word embeddings
- 2014: Expansion of Neural models
- 2015: Attention
- 2018 and beyond: Language model advancements



[A Neural Probabilistic Language Model](https://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model.pdf)
(https://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model.pdf)

Even vanilla models (RNN) can “learn” language

100 epochs

tyntd-iafhatawiaoihrdemot lytdws e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng

700 epochs

Aftair fall unsuch that the hall for Prince Velzonski's that me of her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort how, and Gogition is so overelical and ofter.

2000 epochs

"Why do what that day," replied Natasha, and wishing to himself the fact the princess, Princess Mary was easier, fed in had oftened him. Pierre aking his soul came to the packs and drove up his father-in-law women.

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

Think of the marginal impact vs the feasibility

CNN
LSTM
LSTM
CNN+LSTM

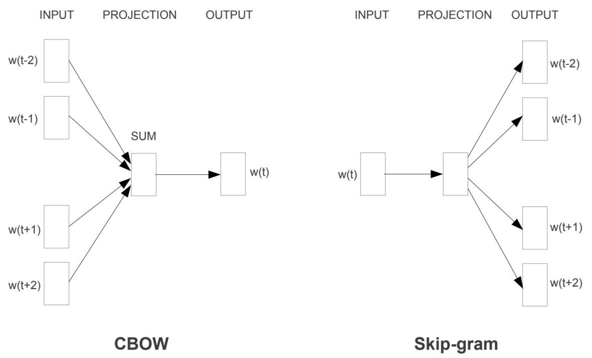
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

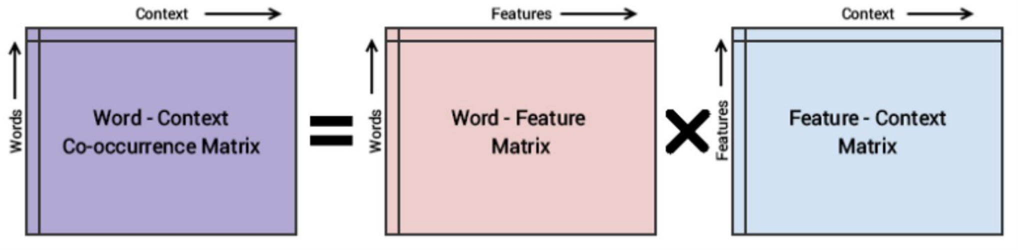
Vaswani, Ashish, et al. "Attention is all you need."
https://arxiv.org/pdf/1706.03762.pdf

Review your history

- 40s-50s: Machine translation era
- 60-70s: Shift towards semantic-driven processing
- 70s to 80s: Community expansion
- 90s-00s: Probabilistic/Statistical models
- 2000s: Neural Language models
- 2008: Multi-task learning
- 2013: Word embeddings
- 2014: Expansion of Neural models
- 2015: Attention
- 2018 and beyond: Language model advancements



https://arxiv.org/pdf/1301.3781.pdf

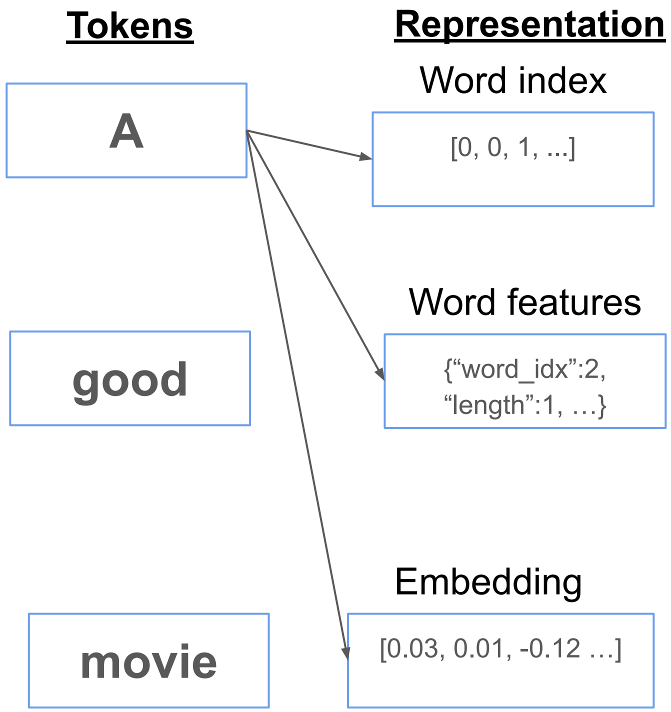


Conceptual model for the GloVe model's implementation

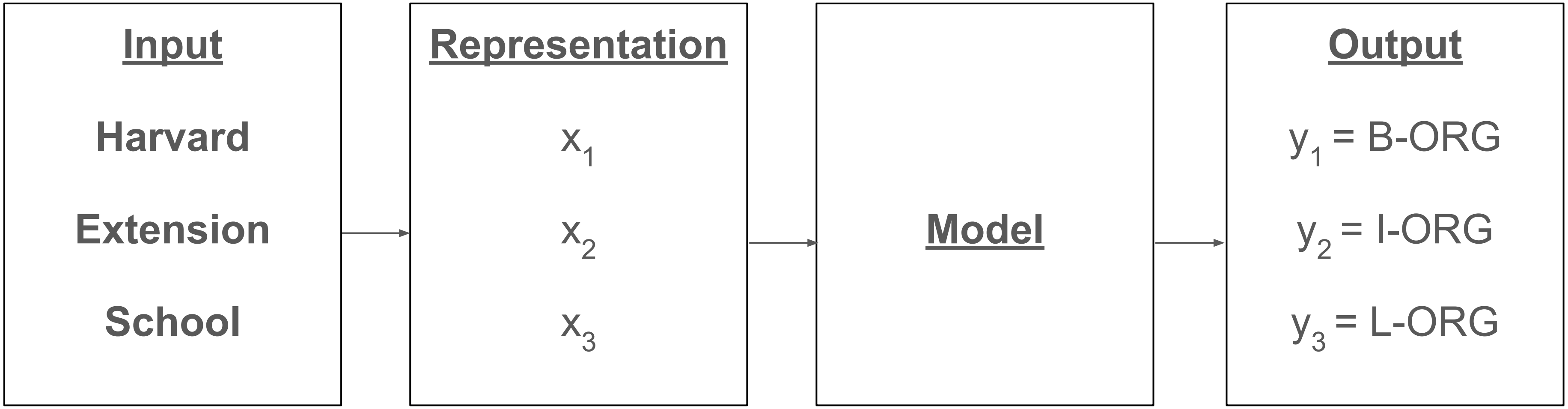
Implementing Deep Learning Methods and Feature Engineering for Text Data: The GloVe Model

How to represent words

- In notebook: Words as sparse vectors (one-hot encoded)
- "Car" vs "automobile" totally different vectors
- Model needs to learn weights for every word in vocabulary
- Condensed, informative representation
    - Word-level representations from topic models
    - Embeddings

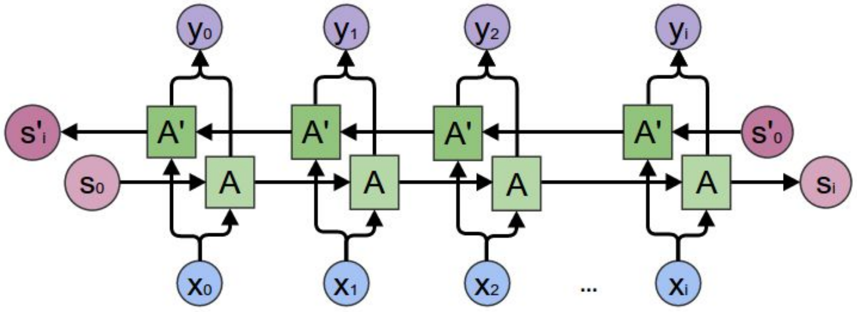


Model-based NER

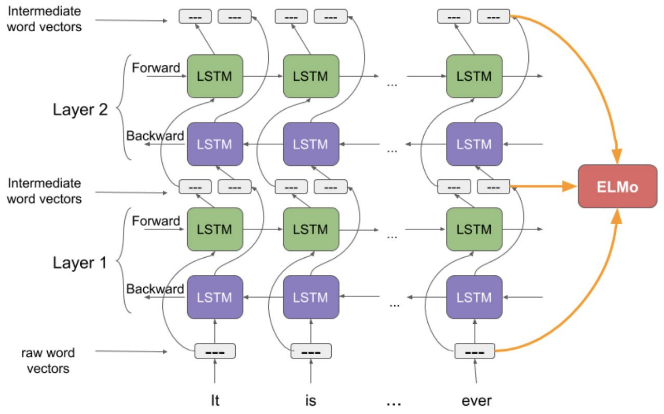


Review your history

- 40s-50s: Machine translation era
- 60-70s: Shift towards semantic-driven processing
- 70s to 80s: Community expansion
- 90s-00s: Probabilistic/Statistical models
- 2000s: Neural Language models
- 2008: Multi-task learning
- 2013: Word embeddings
- 2014: Expansion of Neural models
- 2015: Attention
- 2018 and beyond: Language model advancements

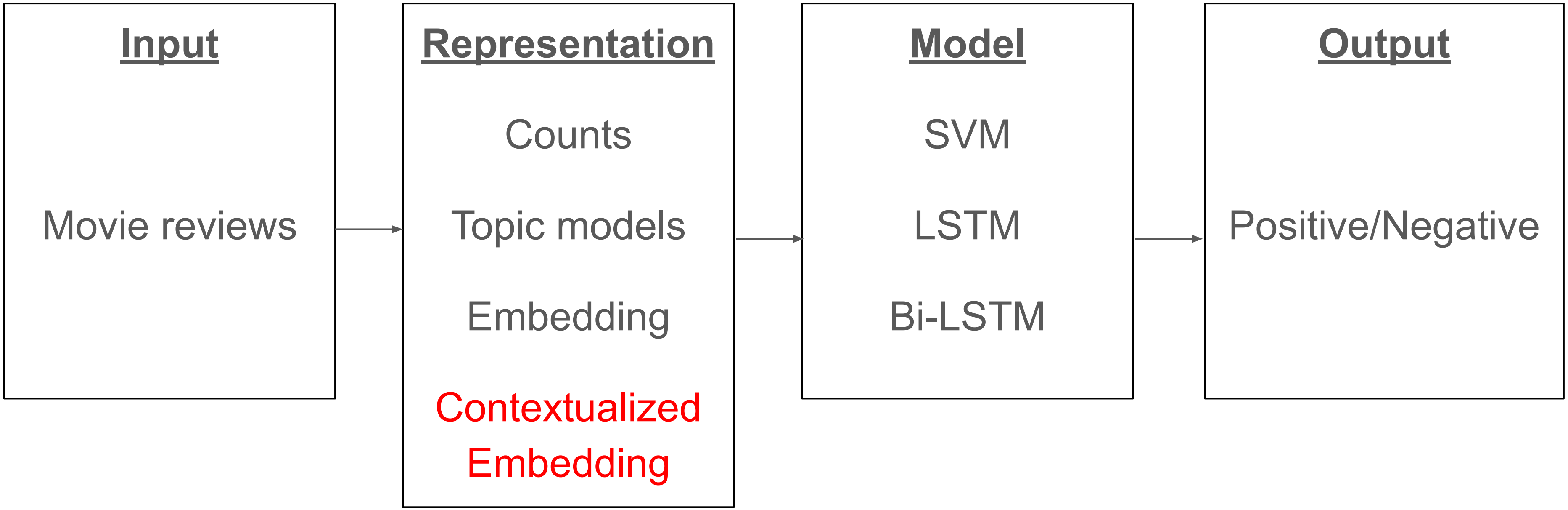


http://colah.github.io/posts/2015-09-NN-Types-FP/



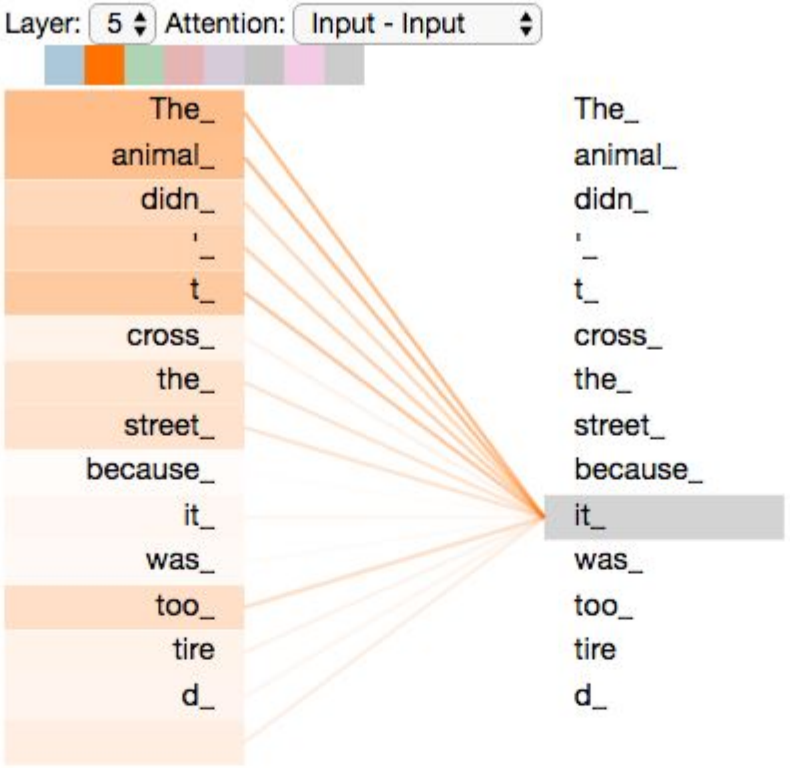
https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/

Transfer learning structure



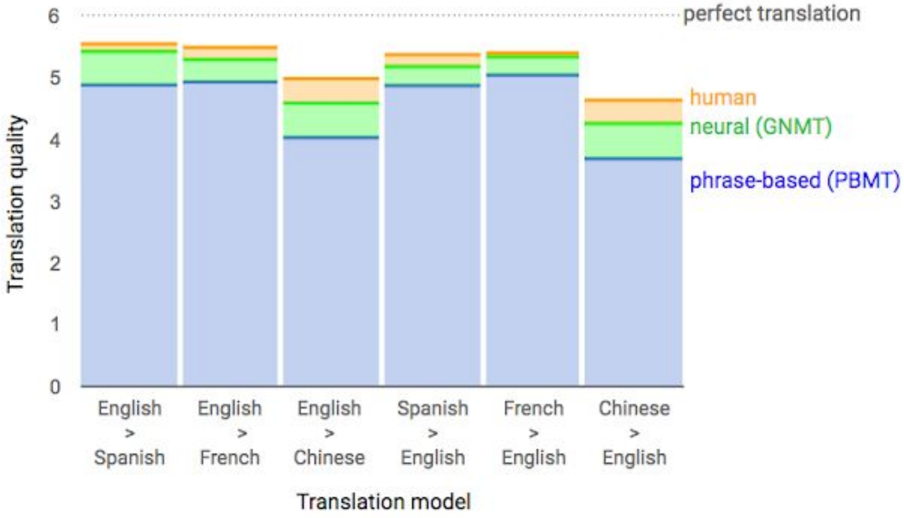
Review your history

- 40s-50s: Machine translation era
- 60-70s: Shift towards semantic-driven processing
- 70s to 80s: Community expansion
- 90s-00s: Probabilistic/Statistical models
- 2000s: Neural Language models
- 2008: Multi-task learning
- 2013: Word embeddings
- 2014: Expansion of Neural models
- 2015: Attention
- 2018 and beyond: Language model advancements



http://jalammar.github.io/illustrated-transformer/

What it looks like (Google Translate)



Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Spanish->English	Uno no es lo que es por lo que escribe, sino por lo que ha leído.	One is not what is for what he writes, but for what he has read.	You are not what you write, but what you have read.	You are who you are not because of what you have written, but because of what you have read.

https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Encoder-decoder with attention

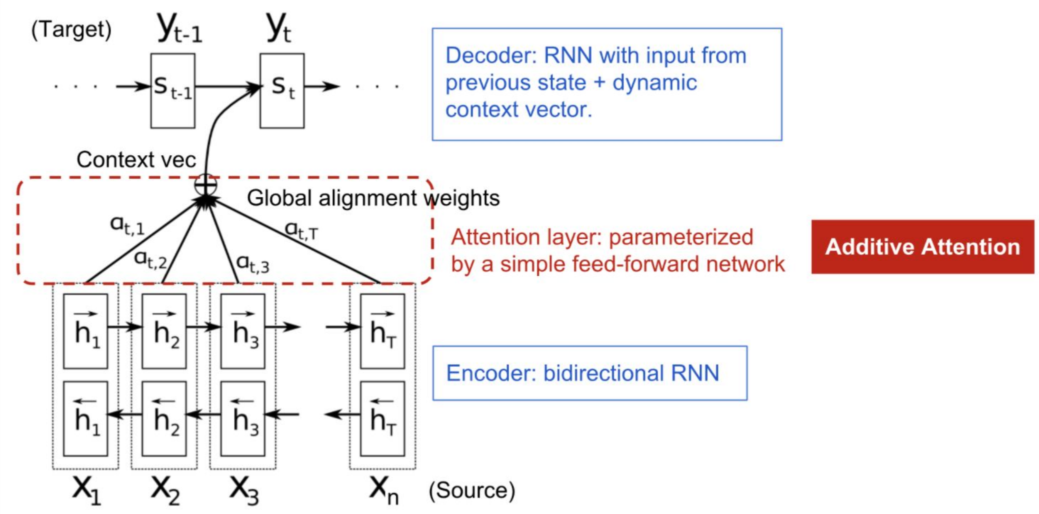
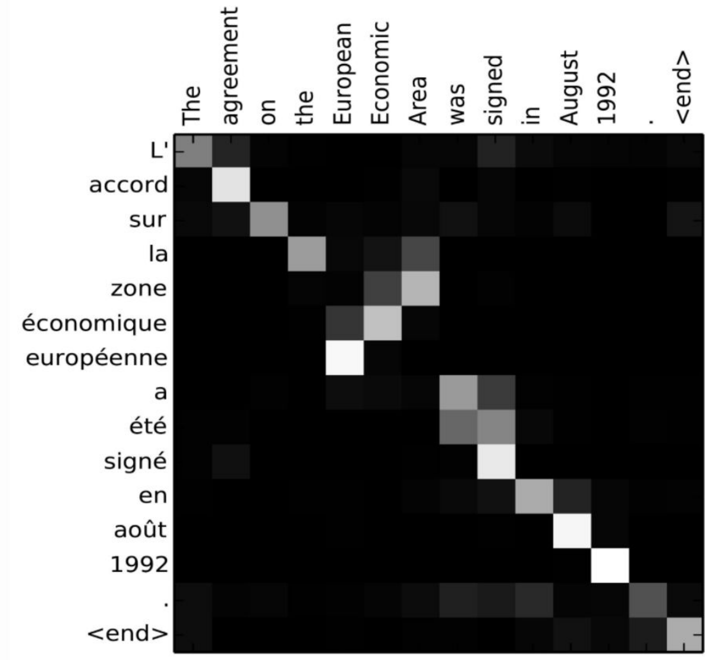


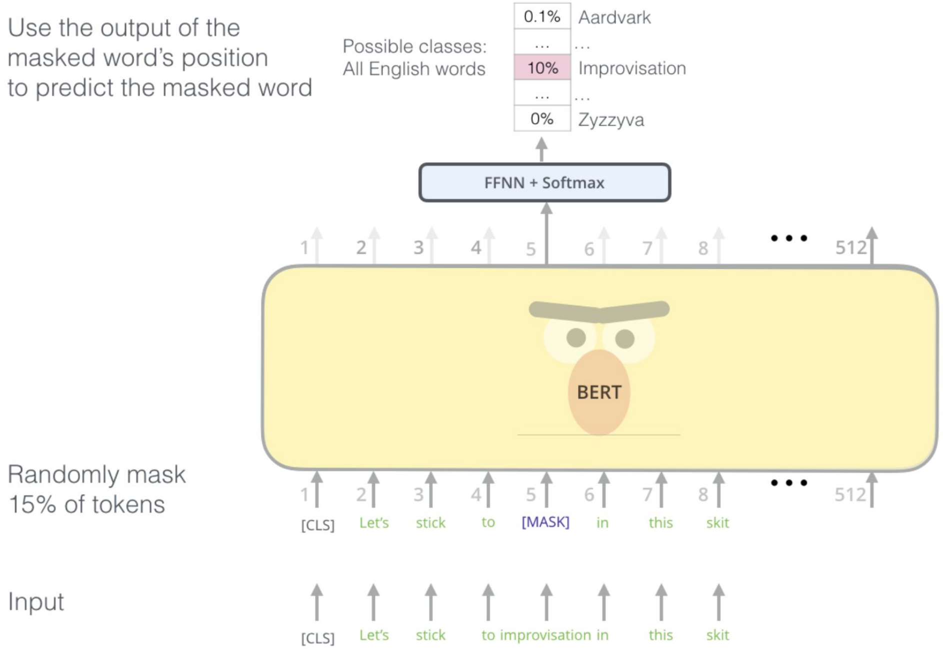
Fig. 4. The encoder-decoder model with additive attention mechanism in Bahdanau et al., 2015.



https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Review your history

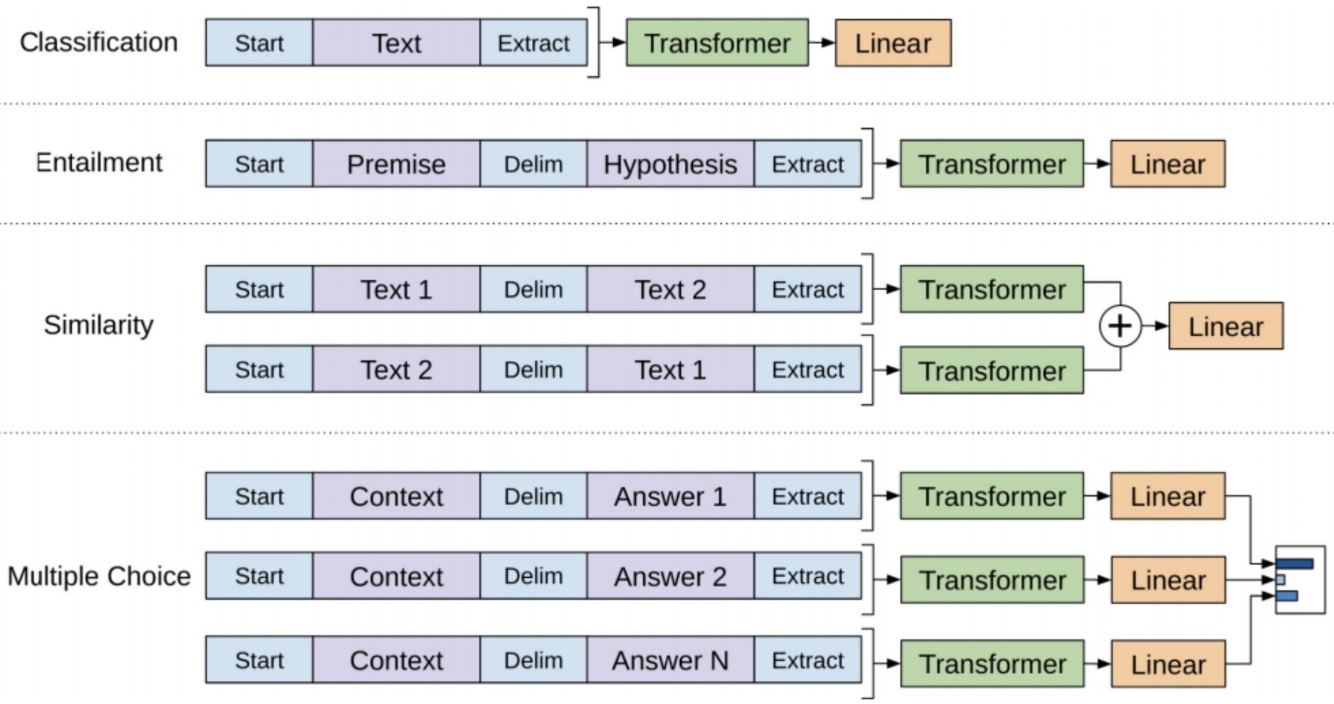
- 40s-50s: Machine translation era
- 60-70s: Shift towards semantic-driven processing
- 70s to 80s: Community expansion
- 90s-00s: Probabilistic/Statistical models
- 2000s: Neural Language models
- 2008: Multi-task learning
- 2013: Word embeddings
- 2014: Expansion of Neural models
- 2015: Attention
- 2018 and beyond: Language model advancements



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

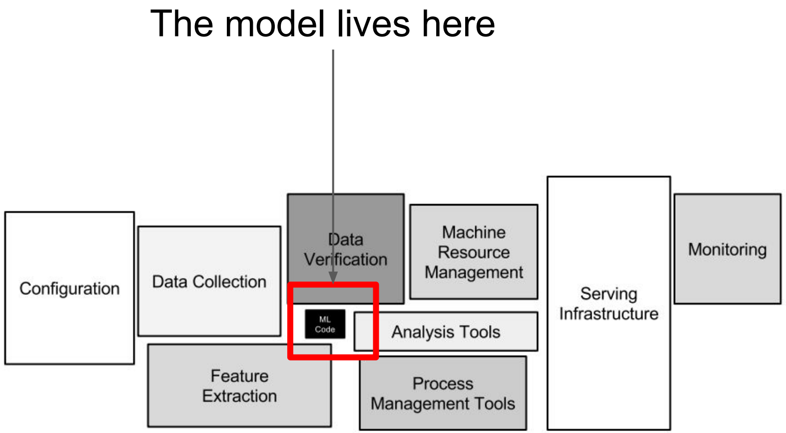
https://jalammar.github.io/illustrated-transformer/

Transfer learning with transformers (Fine-tuning)

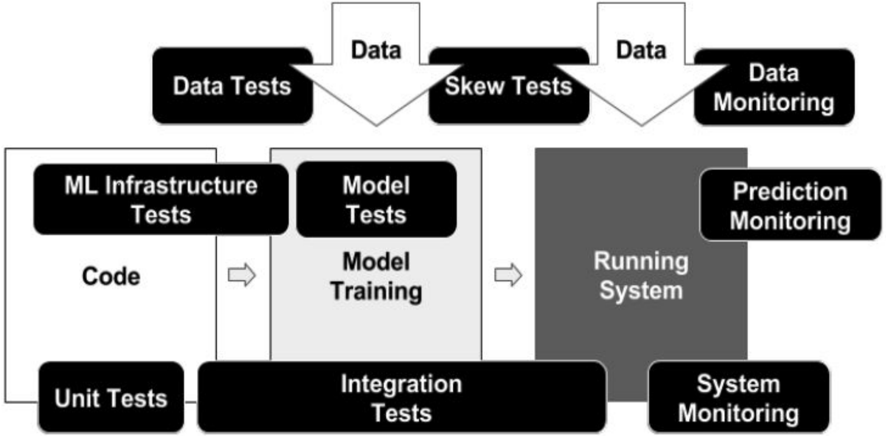


http://jalammar.github.io/illustrated-bert/

The model is often the smallest piece



Hidden Technical Debt in Machine Learning Systems
(https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf)



ML-Based System Testing and Monitoring

https://ai.google/research/pubs/pub46555

Some key advice to sum up

- Data doesn’t matter if you don’t understand the problem
  - Open lines of communication, achieve buy-in
- Methodology doesn’t matter if your data isn’t appropriate
  - Create an exploratory workflow
- Performance doesn’t matter if your methodology isn’t appropriate
  - Think carefully about metrics, assess your risk of bias, unwanted learning
- There’s a wealth of resources for all of the above out there, explore!
  - Useful links on next slide
- Build and iterate
  - Minimal methodology can do surprisingly well
- Share and contribute
  - There’s always the need for contributors in the open source community
  - If you have an interesting application, submit to conferences, present at local groups

Useful resources

Blogs

Sebastian Ruder: https://ruder.io/

Jay Alammar: https://jalammar.github.io/

https://explosion.ai/blog

https://blog.rasa.com/

Data

https://datasetsearch.research.google.com/

https://github.com/awesomedata/awesome-public-datasets#naturallanguage

https://datasets.quantumstat.com/

https://www.kaggle.com/tags/nlp

Models:

- https://huggingface.co/models
- https://pytorch.org/hub/
- https://explosion.ai/blog/spacy-transformers

Libraries

- https://www.nltk.org/
- https://stanfordnlp.github.io/stanza/
- https://allennlp.org/
- Huge amount of spacy-related libraries
  - https://spacy.io/universe

Get involved!

PyData: https://pydata.org/

Find your local chapter!

Local meetups

Open source projects

DS Conferences

PyData Global

ODSC

Strata AI

NLP Conferences

Annual Conference of the Association for Computational Linguistics (ACL)

Conference on Empirical Methods in Natural Language Processing (EMNLP)

International Conference on Computational Linguistics

And be in touch!

Website: https://benbatorsky.com/

Blog: https://bpben.github.io/

Twitter: https://twitter.com/bpben2

Linkedin: https://www.linkedin.com/in/benjamin-batorsky/