Natural Language Generation

VERA DEMBERG

ELEMENTS OF DATA SCIENCE AND AI

Slide credit

Slides based on

- ACL tutorial on story telling from structured data and knowledge graphs
- Slides on response generation by Verena Rieser

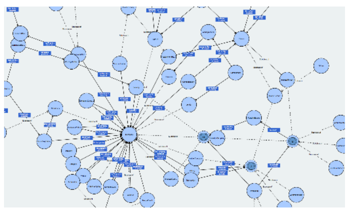
NLG

Natural Language Query in **Weather** Domain

How is the weather this weekend in Atlanta?

Query Parser

SQL



Weather Ontology

Database (Relational DB) for Weather

Tabular results

TEMPERATURE	POP	HUMIDITY	WINDSPEED	CITY	DATE MONTH	DATE HOUR	DATE YEAR	DATE DAY
29	1	76	3	Atlanta	4	4	2016	27
29	2	74	3	Atlanta	4	5	2016	27
30	1	75	3	Atlanta	4	3	2016	27
30	1	73	3	Atlanta	4	6	2016	27
31	1	70	3	Atlanta	4	7	2016	27

...

....

Language Generation

Slight chance of showers on Saturday morning with a high of 31 degrees. Sunny day and clear skies all day Sunday.

Other examples for Natural Language Generation

Product Information

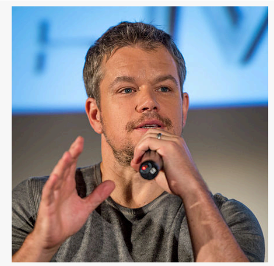
Brand	Nikon
Model Name	D5300
Product Type	DSLR Camera
Item Weight	1.85 Kg
Product Dimensions	7.6 x 12.5 x 9.8 cm
Colour	Black
Resolution	24.2 megapixels
Lens included	Yes
Screen Size	3.2 Inches
Image Stabilization	Yes
Optical Zoom	3 X
Max Shutter Speed	1/200 Seconds
Video Capture Resolution	1920 x 1080
Batteries Included	Yes
Batteries Required	Yes
Battery Cell Composition	Lithium
Continuous Shooting Spee	5
Viewfinder Type	Optical
Has Self Timer	Yes

Product Description

The Nikon D5300 DSLR Camera, which comes in black color features 24.2 megapixels and 3X optical zoom. It also has image stabilization and self-timer capabilities. The package includes lens and Lithium cell batteries.

Input

Born	Matthew Paige Damon October 8, 1970 (age 46) Cambridge, Massachusetts, U.S.
Residence	Pacific Palisades, California, U.S.
Alma mater	Harvard University
Occupation	Actor, filmmaker, screenwriter
Years active	1988–present
Spouse(s)	Luciana Bozán Barroso (m. 2005)



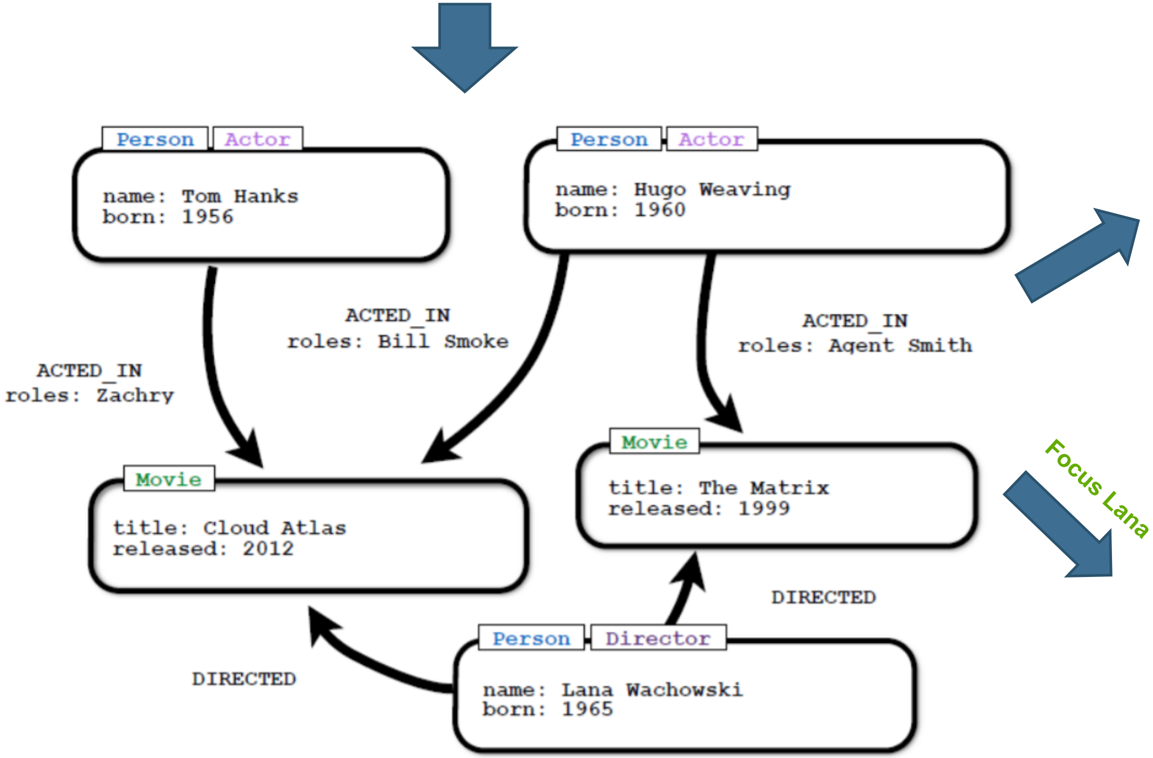
Output

Born Matthew Paige Damon October 8, 1970 Residence U.S. Occupation Actor filmmaker screenwriter

Matthew Paige Damon who was born in October 8, 1970 is an American actor, film producer, and screenwriter.

Knowledge Graph summarization

Query: Show me movies directed by Lana and their lead actors.



General graph summary:
Hugo Weaving acted in movie Cloud Atlas (as Bill Smoke) along with Tom Hanks (as Zachry) and in movie The Matrix (as Agent Smith). Both the movies were directed by Lana Wachowski.

Entity focused summary (Focus Lana):
Lana Wachowski born in 1965 is the director of movies Cloud Atlas (released in 2012) and The Matrix (released in 1999)

https://neo4j.com/

Text-to-Text NLG

Summarization
Headline Generation

Alton attorney accidentally sues himself

Attorney from Alton files a lawsuit against himself by mistake

Paraphrasing

Friday, March 11, 2005

By Steve Korris

Alton attorney Emert Wyss thought he could make money in a Madison County class action lawsuit, but he accidentally sued himself instead. Now he has four law firms after his money - and he hired all four.

L'avocat d'Alton se poursuit par accident

Machine Translation

Wyss's boomerang litigation started in 2002, when he invited Carmelita McLaughlin to his office at 1600 Washington St. in Alton. Acting as her attorney when she bought a home in Alton and when she refinanced it, on both occasions she had chosen Centerre Title--a company that Wyss owned--to close her loans.

Emert Wyss's Alton office

Question Answering

Image Captioning

In the course of the attorney-client relationship, Wyss advised McLaughlin she might have a claim against Alliance Mortgage, holder of the first mortgage. Wyss believed Alliance Mortgage might have broken the law by charging a $60 fax fee when she refinanced.

Q: What are the consequences?

A: Emert Wyss had hired four law firms and now all of them are after his money.

He produced a retainer agreement providing for his legal services and those from the Lakin Law Firm of Wood River, Campbell and Brinkley of Godfrey, Freed and Weiss of Chicago, and Diab and Bock of Chicago. McLaughlin signed.

The Lakin firm filed a class action complaint against Alliance Mortgage in 2003. The complaint identified the Chicago firms and Campbell and Brinkley as other attorneys of record, but not Wyss.

Question Generation

When did the Lakin firm file a complaint against Alliance Mortgage?

Natural Language Generation

- Branch of Computational Linguistic that deals with generation of natural language text from unstructured / structured textual/non-textual (data) forms. (Reiter and Dale, 2000)
  - Focusses on computer systems
  - Produces understandable texts (in English or other human languages)

Multilingual

Multimodal

- **Text to text**
  - Machine Translation
  - Automatic Summary Generation
  - Document paraphrasing
  - Simplification of Complex Text
  - Text Style Transfer
- **Data to text**
  - Summary Generation from Tabular Data
  - Virtual 'newspapers' from sensor data
  - Summaries of patient information in clinical contexts
  - Persuasive Text Generation from product description for e-commerce
  - Story generation from events

Gatt et al., 2017

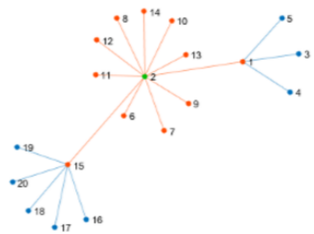
Data-to-text NLG

- **INPUT**: Non-linguistic input
- **OUTPUT**: Documents, Reports, Explanations, Help messages, and other kinds of text.
- Knowledge Required: (1) Language, and (2) Application domain.

Table

Year	Number of patents	Revenue generated
2016	5055	700 M
2015	6060	742 M
2004	8076	1.2 B

Graph



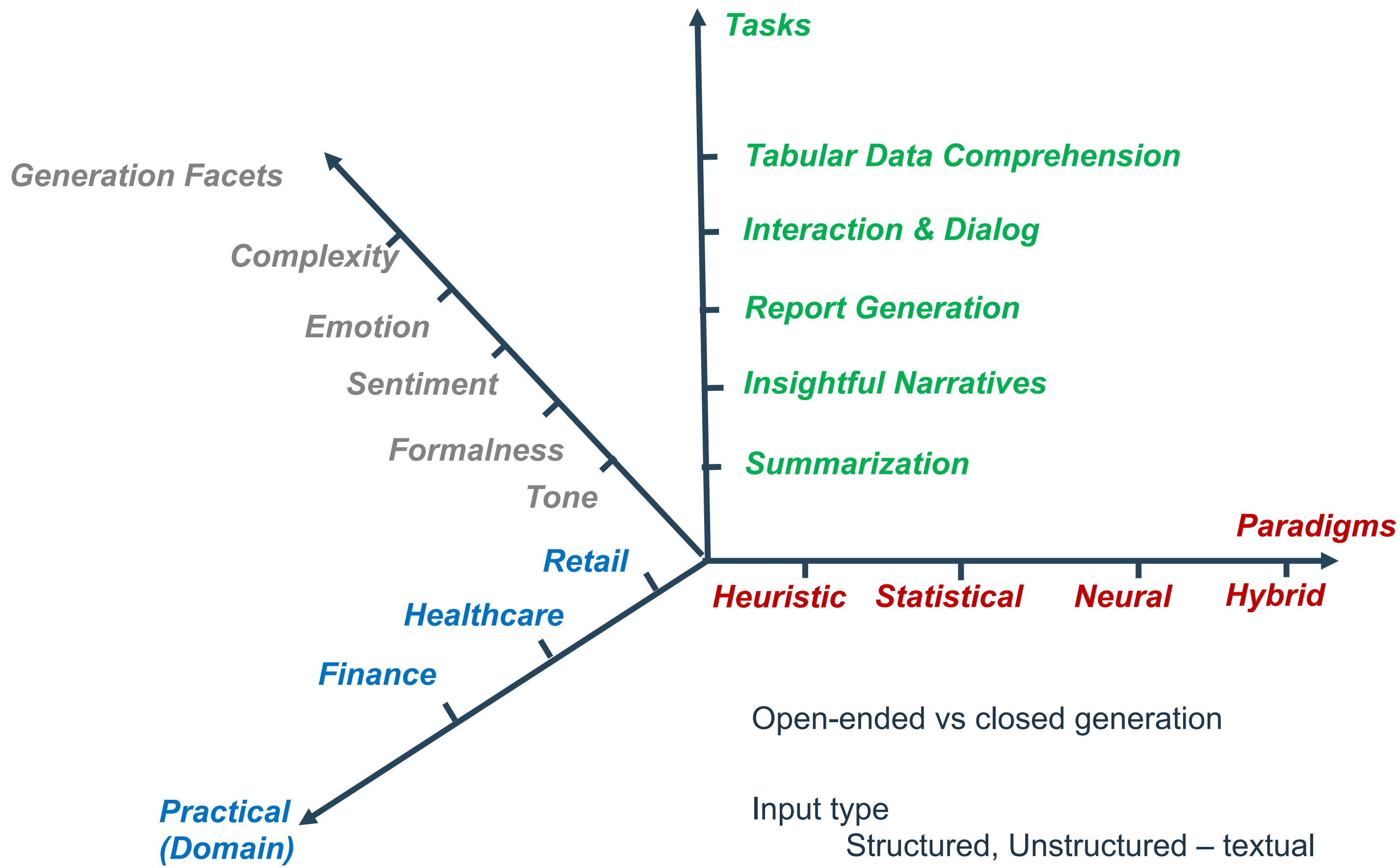

```
{
 "answer":
 {
  "premium": {"$":502.83},
   "initial_payment": {"$":100},
   "monthly_payment": {"$":
85.57}
  }
}
```


JSON


```
<?xml version="1.0" encoding="UTF-8" ?>
- <employees>
  - <emp id="S001">
      <name>ABC</name>
      <salary>5000</salary>
    </emp>
  - <emp id="S002">
      <name>PQR</name>
      <salary>7000</salary>
    </emp>
  - <emp id="S003">
      <name>XYZ</name>
      <salary>9000</salary>
    </emp>
  </employees>
```


XML

Data-to-text NLG: A 4D perspective



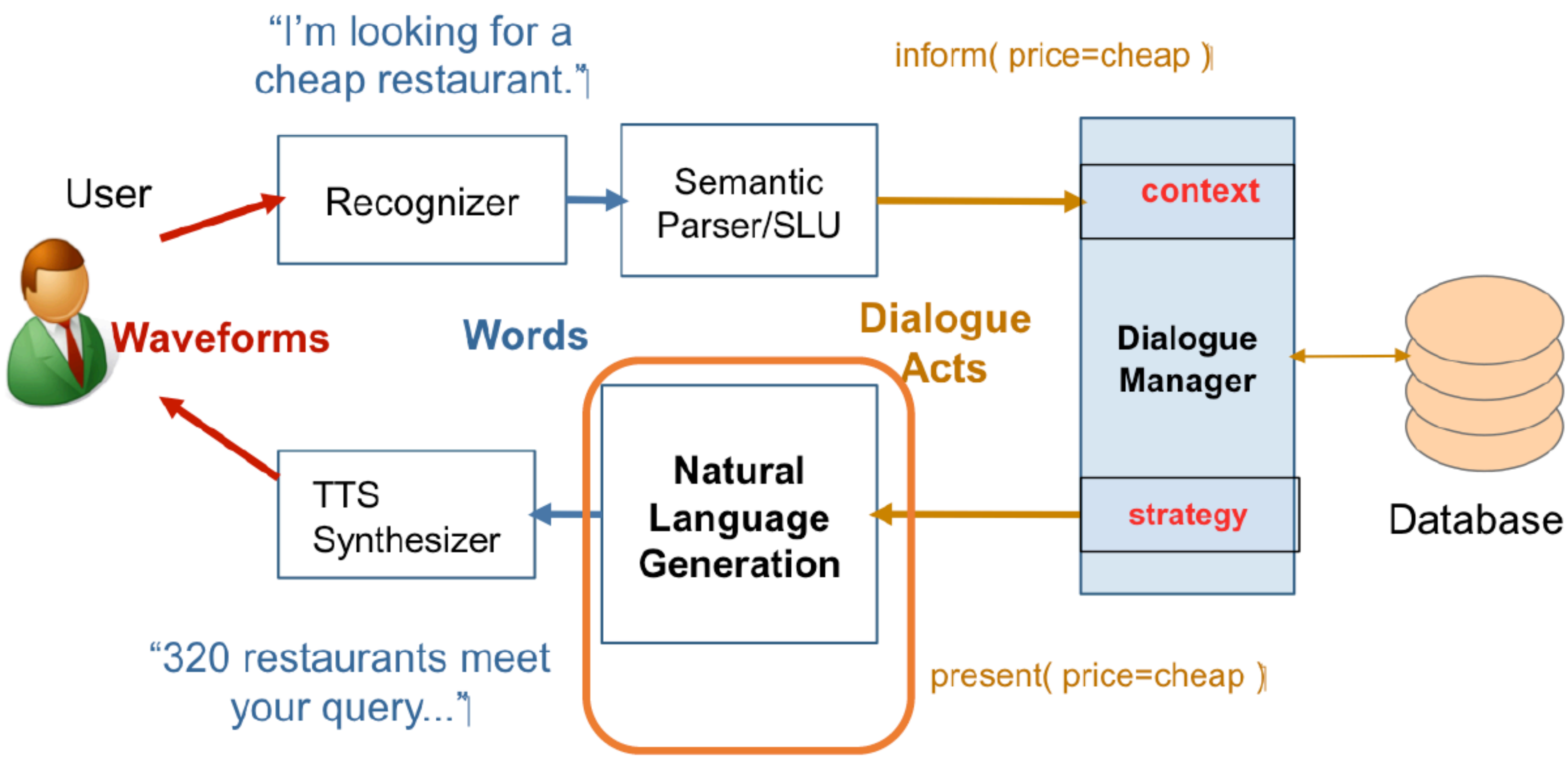
Open-ended vs closed generation

Input type

- Structured, Unstructured – textual
- Image, Video
- Cognitive signals – EEG, Eye tracking, MEG

Concept: CS626, IIT Bombay

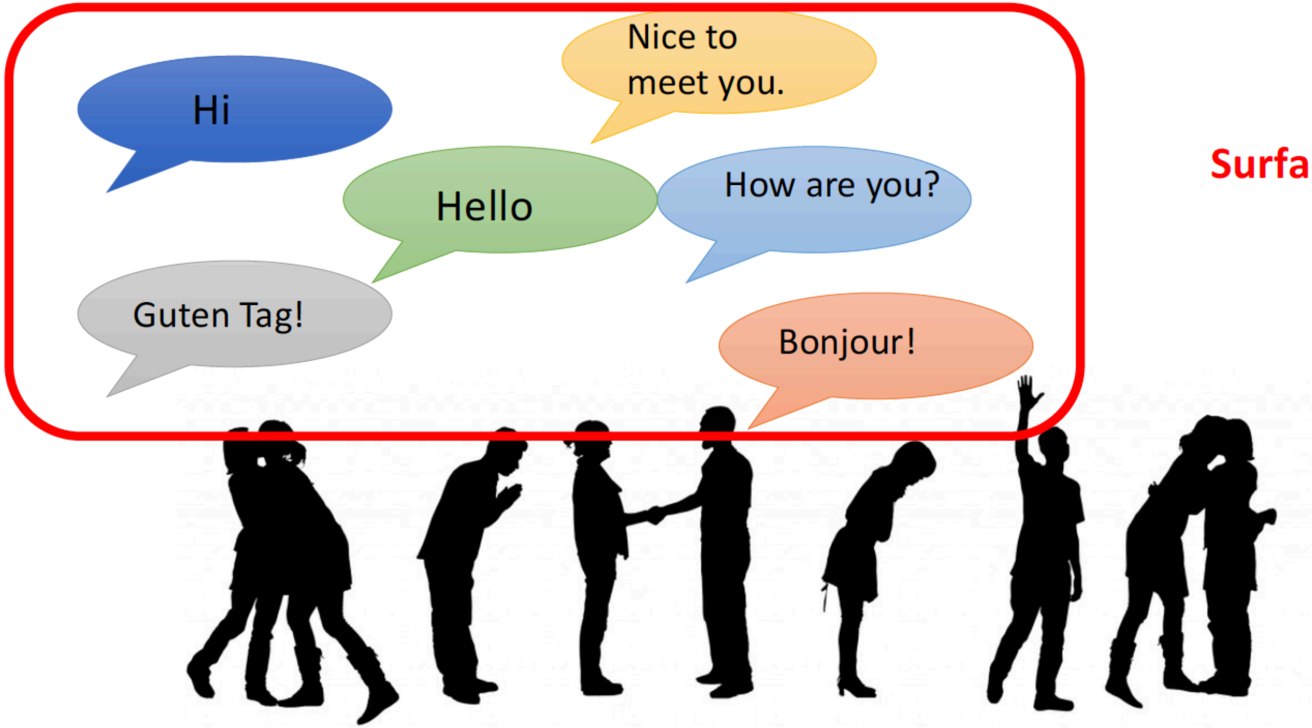
Architecture of a Spoken Dialog System (SDS)



There are many different ways of realizing a specific goal.

What do you say when you want to greet someone?

Communicative Goal/ Dialogue Act



Surface Forms

There are many different ways of realizing a specific goal.

Meaning **R**epresentation (MR)

{ 1:n }

Natural Language String

```
inform ( type = restaurant
         cuisine = Chinese
         DB_hits = 2,471 )
```

→

- I found 2,471 Chinese restaurants.
- I found over two thousand restaurants which serve Chinese food.
- There are a lot of Chinese restaurants in Edinburgh. Which area were you thinking of?

Methods for Natural Language Generation

1. **Templates** (most SDS).

2. **Grammar**-based generation

CCG, LFG, Dynamic Syntax,… [Kaplan, 2012; White, 2009; Purver, Eshghi 2013]

3. **E2E Machine Learning**

Retrieval

Generation

Traditional NLG

Rule based NLG
Template based NLG
Shortcomings

Rule based Generation – When and When Not

- **When the phenomenon is understood AND expressed, rules are the way to go**

- "**Do not learn when you know!!**"

- When the phenomenon "seems arbitrary" at the current state of knowledge, DATA is the only handle!
  - Why do we say "**Many Thanks**" and not "**Several Thanks**"!
  - Very tedious to give a rule and fragile

- Rely on machine learning to acquire this knowledge from data.

Table Description in Natural Language Text: High Level Rules

Enrichment
(Verb phrase)

was born in

Name	Birth City
Albert Einstein	Ulm, Germany

Subject ***Object***

Albert Einstein was born in Ulm, Germany

Rules:

- Consider one column as "subject and the other column as object"
- Use column header and extract verb phrase VP by looking up in a lexicon
- Realized sentences: S + VP + O

Exception

Verb ???

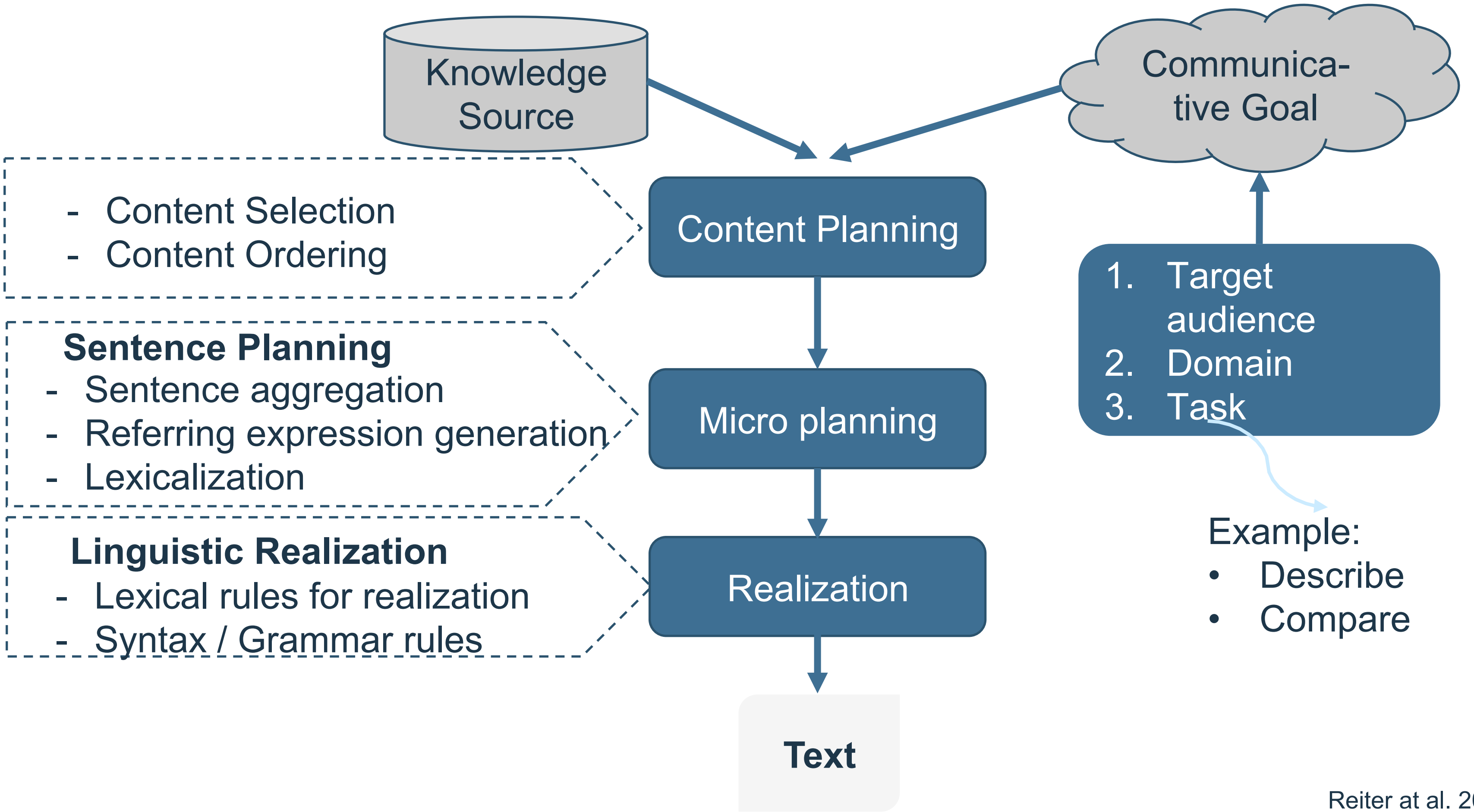
Name	Nationality
Albert Einstein	Ulm, Germany

Albert Einstein *nationalized??* …….. Germany ❌

Albert Einstein's nationality is German ✅
Albert Einstein is from Germany ✅

Step back…

Natural Language Generation Pipeline

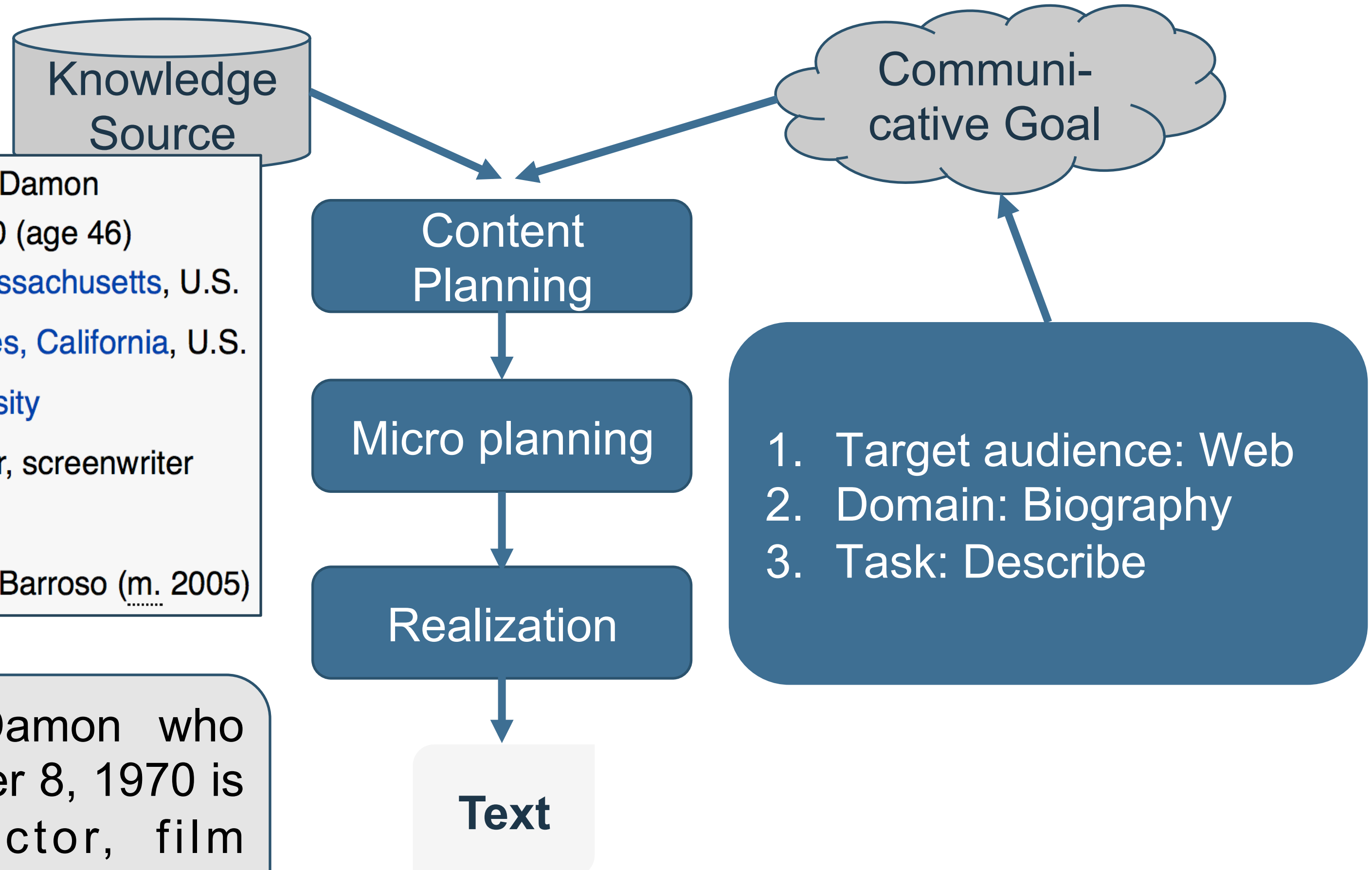


Reiter at al. 2000

Terminology alert

Document planning Text planning Content planning
Discourse plan Text plan
Micro planning Sentence planning
Surface realization Linguistic Realization Linearization ealization

Natural Language Generation Pipeline

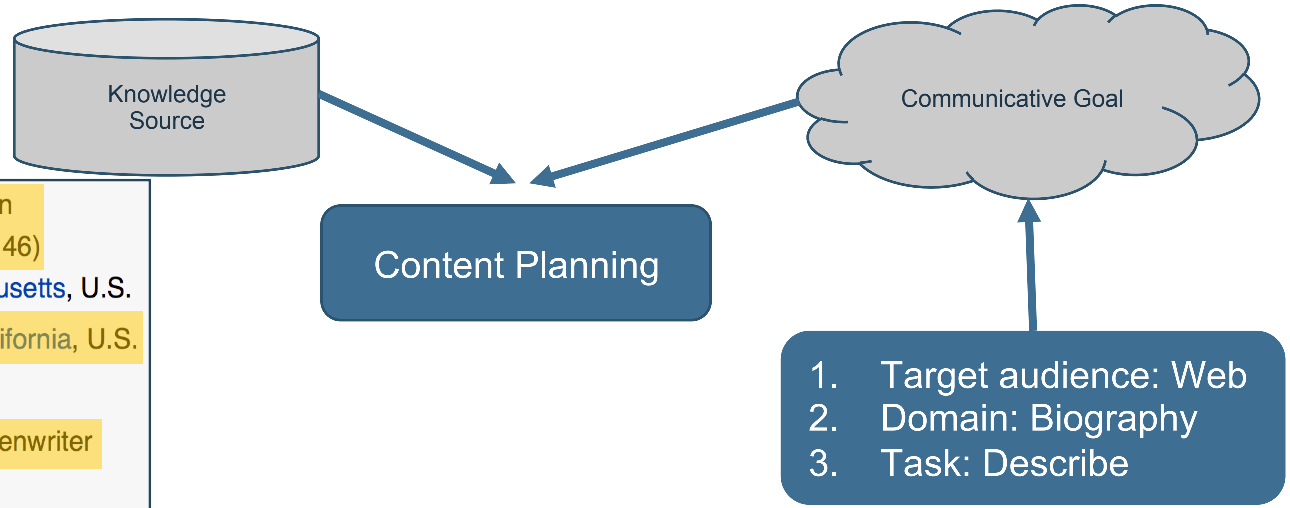


Born	Matthew Paige Damon October 8, 1970 (age 46) Cambridge, Massachusetts, U.S.
Residence	Pacific Palisades, California, U.S.
Alma mater	Harvard University
Occupation	Actor, filmmaker, screenwriter
Years active	1988–present
Spouse(s)	Luciana Bozán Barroso (m. 2005)

Matthew Paige Damon who was born in October 8, 1970 is an American actor, film producer, and screenwriter.

Reiter at al. 2000

Natural Language Generation Pipeline



Born	Matthew Paige Damon October 8, 1970 (age 46) Cambridge, Massachusetts, U.S.
Residence	Pacific Palisades, California, U.S.
Alma mater	Harvard University
Occupation	Actor, filmmaker, screenwriter
Years active	1988–present
Spouse(s)	Luciana Bozán Barroso (m. 2005)

Content determination and selection

1. Name: Matthew Paige Damon
2. Born: October 8, 1970
3. Residence: Pacific Palisades, California, United States
4. Occupation: Actor, filmmaker, screenwriter

At this stage we know **what we want to talk about** .. but still have **no idea about how.**

Natural Language Generation Pipeline

1. Name: Matthew Paige Damon
2. Born: October 8, 1970
3. Residence: Pacific Palisades, California, United States
4. Occupation: Actor, filmmaker, screenwriter

1. **Matthew Paige Damon** born in October 8, 1970
2. **Matthew Paige Damon** residence Pacific Palisades, California, United States
3. **Matthew Paige Damon** is *Actor*. **Matthew Paige Damon** is *filmmaker*. **Matthew Paige Damon** is *screenwriter*.

Sentence aggregation, Lexicalization and referring expression

1. Matthew Paige Damon born in October 8, 1970 and residence of America. **OR** Matthew Paige Damon born in October 8, 1970 is an American.
2. He is an Actor, filmmaker and screenwriter.

Content Planning

Micro planning

Fakeness alert: For example purpose there is some structure in the sentences, but in reality everything will be in the form of data structures passed from one layer to another. **There are no sentences yet!**

Natural Language Generation Pipeline

Content Planning

Micro planning

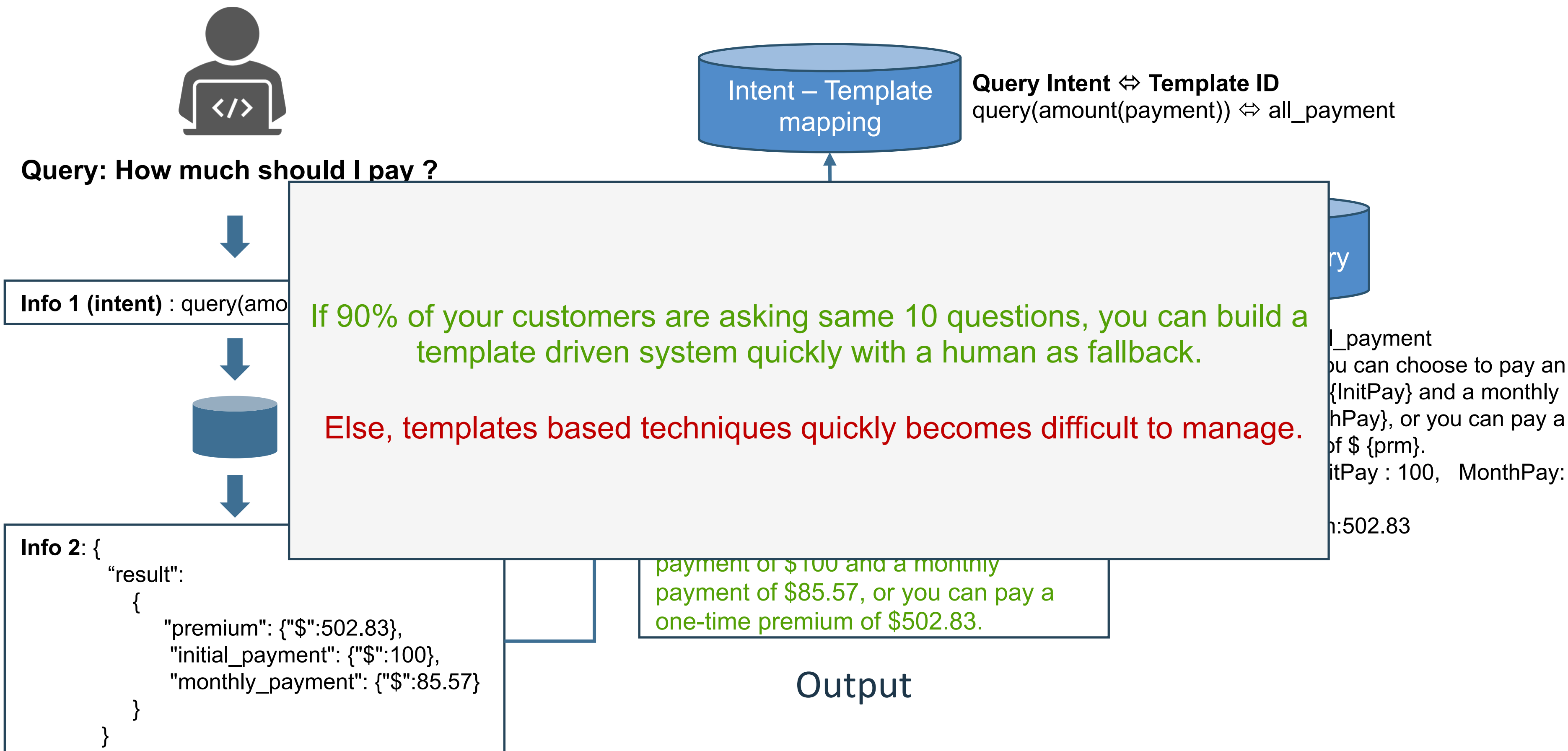
Realization

Matthew Paige Damon(N) born in(VP, TENSE: PAST) October 8, 1970 … American(Adj). … [Actor, filmmaker, screenwriter]

Realizer

Matthew Paige Damon who was born in October 8, 1970 is an American actor, film producer, and screenwriter.

Extremely Simple Template-driven NLG Architecture: Insurance case



https://github.com/parajain/twig/wiki

Eliza – a template based system

TEMPLATE: `I _X1_`

RESPONSE: `You say you _X1_`

TEMPLATE: `_X1_ my _X2_(category family) _X3_`

RESPONSE: `Who else in your family _X3_ ?`

TEMPLATE: `_X1_ you _X2_ me`

RESPONSE: `What makes you think I _X2_ you?`

`User: You hate me.`

`ELIZA: What makes you think I hate you?`

Shortcomings of Traditional Approaches

- Rule-based systems/templates are mostly **inflexible** and **not scalable**
- **Non-transferrable** rules pertaining to domain specific requirements / choices of language artefacts (tone, sentiment, syntax, complexity)
- Typically **do not leverage web scale data / freely available knowledge bases** (like DBPedia, Yago, Freebase)

Statistical Methods

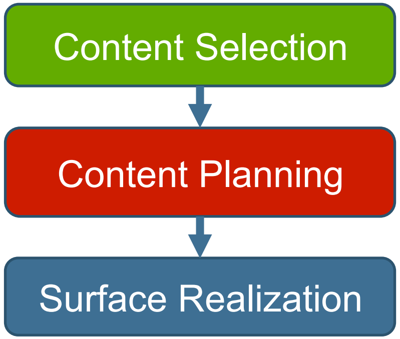
Idea: Learn **from data** how to generate text.

Representative Public Datasets:

- **ROBOCUP**, for sportscasting (Chen and Mooney, 2008);
- **SUMTIME**, for technical weather forecast generation (Reiter et al., 2005)
- **WEATHERGOV**, for common weather forecast generation (Liang et al., 2009)
- **WikiBio** (Lebret et al 2016).
- **ROTOWIRE** and **SBNATION** (Wiseman, Shieber, and Rush 2017).
- **WEBNLG** dataset (Gardent et al. 2017)
- **WikiTableText** (Bao et al 2018)
  - Describing table region – typically restricted to rows.
- **WikiTablePara** (Laha et al, 2018)
  - Created from WikiTable dataset
  - 171 tables with comprehensive descriptions.

Other NLG datasets: https://aclweb.org/aclwiki/Data_sets_for_NLG

Simplified Steps



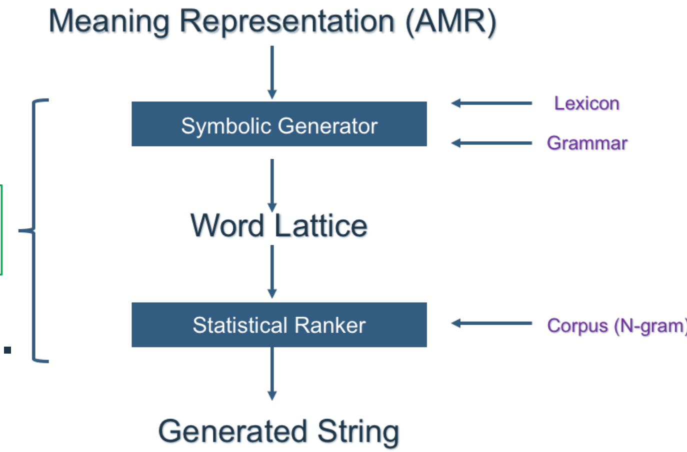
We will continue explaining recent NLG systems from this pipeline perspective

Moving away from Templates…..

- Templates are inflexible and not scalable to different use-cases.
- However, templates do not require much semantic understanding or decision making.
- Can we get best of both worlds?
  - Have a good meaning representation of input data.
  - Move the linguistic decision-making to the surface realization step.
  - This makes surface realization more flexible than templates.
- The surface realization (generation) needs additional knowledge
  - Knowledge from corpus perhaps? [Langkilde and Knight, 1998]
  - → Language Modelling

Flexible Surface Realization

- **Input Meaning Representation to the generator.**
  - Abstract Meaning Representations (AMRs) capture all things to be said.
- **The generator converts the AMR to word lattice.**
  - Word lattice defines transition between states.
  - The state transitions are labeled by words.
  - The conversion uses pre-defined grammar rules.
  - The word lattice captures all things to be said.



- **Statistical Ranker selects the best path in word lattice as output.**
  - N-gram frequencies are computed from monolingual corpora.
  - The pre-computed N-gram frequencies are used **to score** the paths in the lattice.
  - The sequence of words corresponding to the best path is the **final output string**.

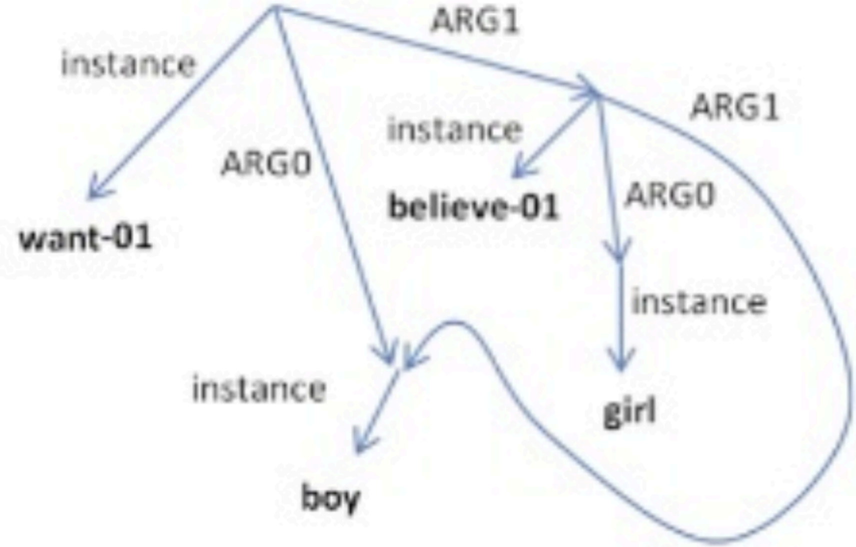
[Langkilde and Knight, 1998]

Example:

AMR specifies meaning.

Grammar then allows to generate text from AMR.

Grammars (like PCFGs with semantic rules) can be learned from data and can be used both ways around (for parsing and for generation).



Abstract Meaning Representation (AMR) format

The AMR above can be expressed variously in English:

The boy wants the girl to believe him.

The boy wants to be believed by the girl.

The boy has a desire to be believed by the girl.

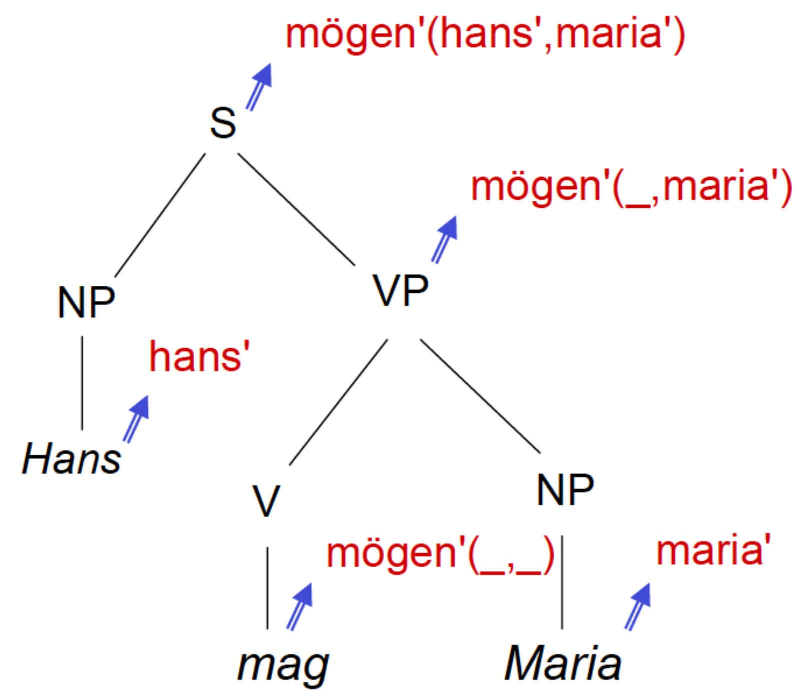
The boy's desire is for the girl to believe him.

The boy is desirous of the girl believing him.

Generation with probabilistic grammars

- Reminder: example for semantic construction (lecture 2):

Semantic construction:
We assemble along the constituent structure along complex semantic expressions "compositionally" from simpler expressions.



For each lexicon entry and every syntactic rule, we add a semantic component

Challenges to statistical generation

- Large search space (can be slow)
- If grammars are learnt from data, may generate ungrammatical output.
- Large amounts of annotated data are necessary (may have data sparsity issues for generating domain-specific text).
- Can try to learn domain-specific grammars that have a good trade-off between template-like large rules or chunks of text and segments that are typically flexible in the domain.

Example

- use syntactic knowledge and semantic structure to **constrain the search** space during training

Synchronous TAG

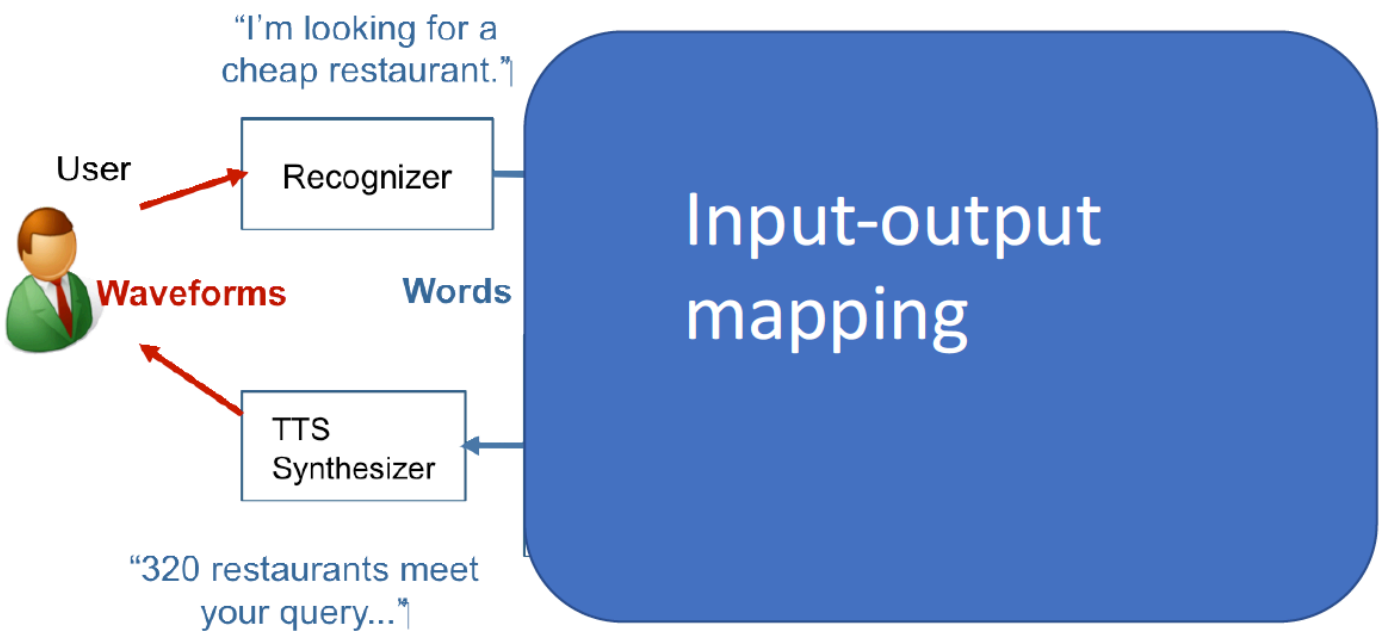
Caffe Buon Gusto has good food quality
while John's Pizzeria has very good food quality.

Neural Methods

End to end neural systems

- Learn from "raw" dialogue data (e.g. OpenSubtitles).
- No* semantic or pragmatic annotation required

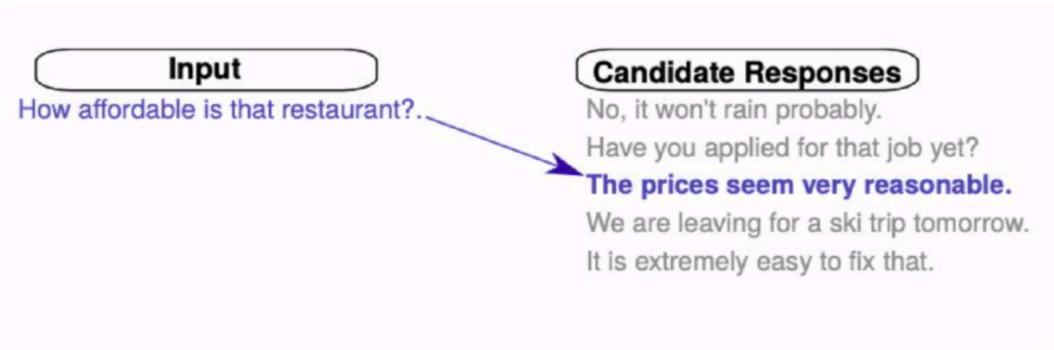
  (*only true for vanilla chat-based systems)



Approaches

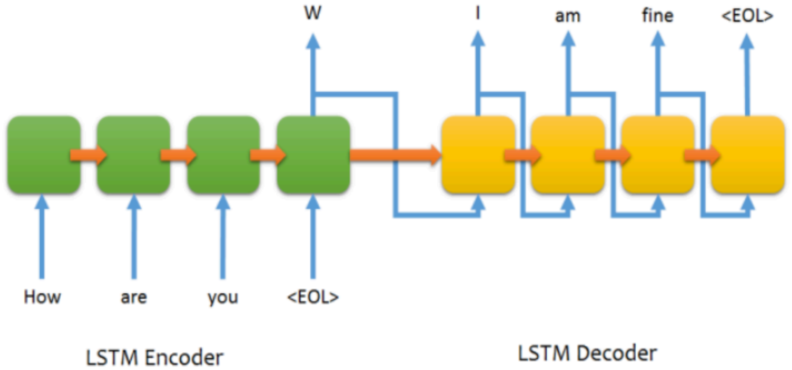
Retrieval-based

- Encode the meaning
- Select the action or response
- ~~Generate the response~~



Generative models

- Encode the meaning
- Select the action or response
- Generate the response



Pros and Cons for retrieval-based vs. generation approaches

Retrieval

- Constrained by the list of candidate responses
- More controllable responses
- Easier to train

Generation

- Variable output
- Prone to give short, general or irrelevant responses
- More difficult to train

Retrieval-based systems

Next utterance selection/ response scoring:

1. Predefine a set of possible responses
2. Given the context, select one response from this set
   - Context: Single turn, multiple turns, extra dialogue features

Training:

- Maximise the Score of positive Context-Response pairs
- Minimise the score of negative Context-Response pairs

Inference:

- Select the set of possible responses
- Rank the responses based on their score given the current context

Generation models

Language models can be used to generate text.

N-gram model:

$P(w_n|w_{n-3}, w_{n-2}, w_{n-1})$

Select w_n with highest likelihood given context (or sample randomly according to probability distribution of words at position n).

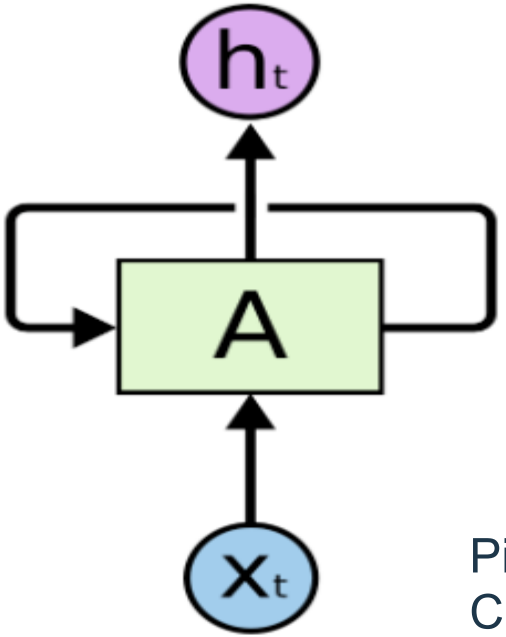
(It's like auto-completion in Google search.)

RNNs: Reminder

If we use a neural network, we also need to make sure that the context of previous words is represented in the model. It therefore makes sense to design a neural network architecture that reflects this challenge.

Solution that (in principle) allows
to model arbitrarily long context:
Recurrent Neural Network

x_t is the input word
h_t is the predicted next word
A is an internal hidden state
The network is “recurrent” because it contains a loop.



Picture credit:
Christopher Olah

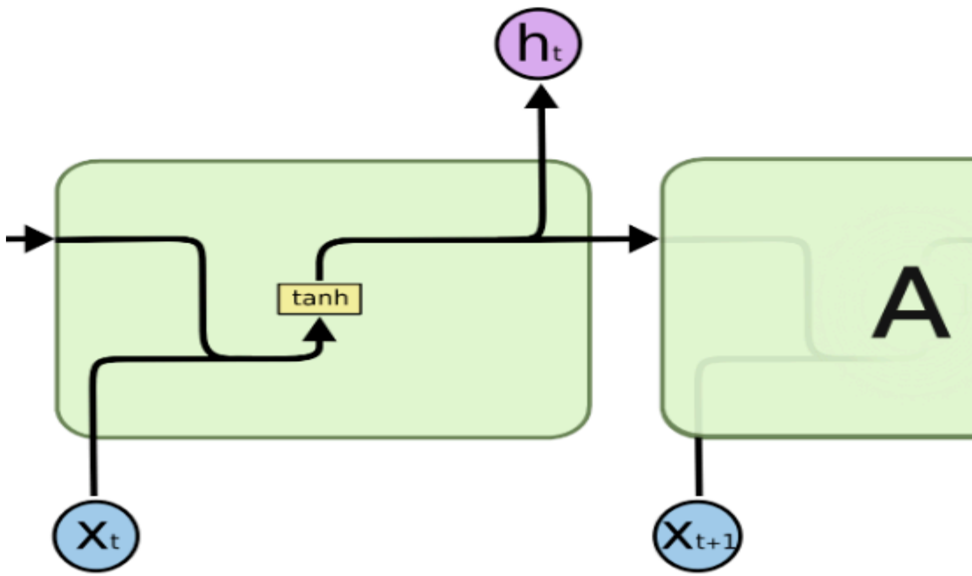
RNNs

If we use a neural network, we also need to make sure that the context of previous words is represented in the model. It therefore makes sense to design a neural network architecture that reflects this challenge.

$$A_t = \tanh\,(W_{AA}A_{t-1} + W_{xA}x_t)$$

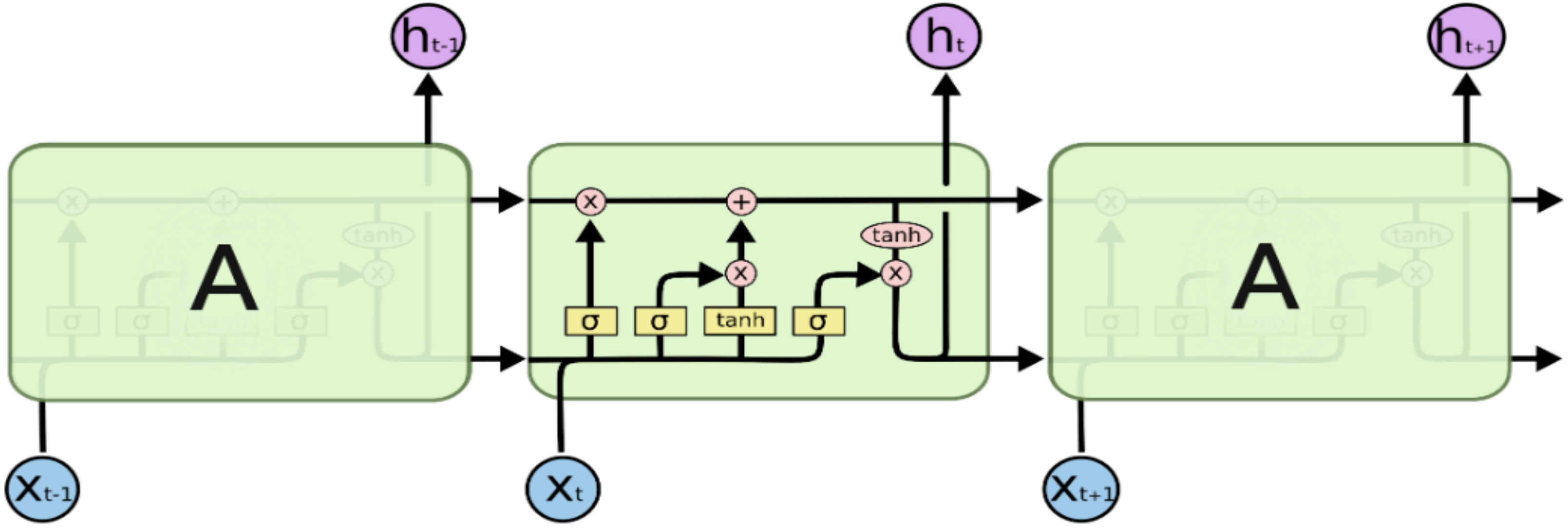
$$h_t = W_{Ay}A_t$$

Picture credit:
Christopher Olah



Long Short Term Memory networks (LSTM)

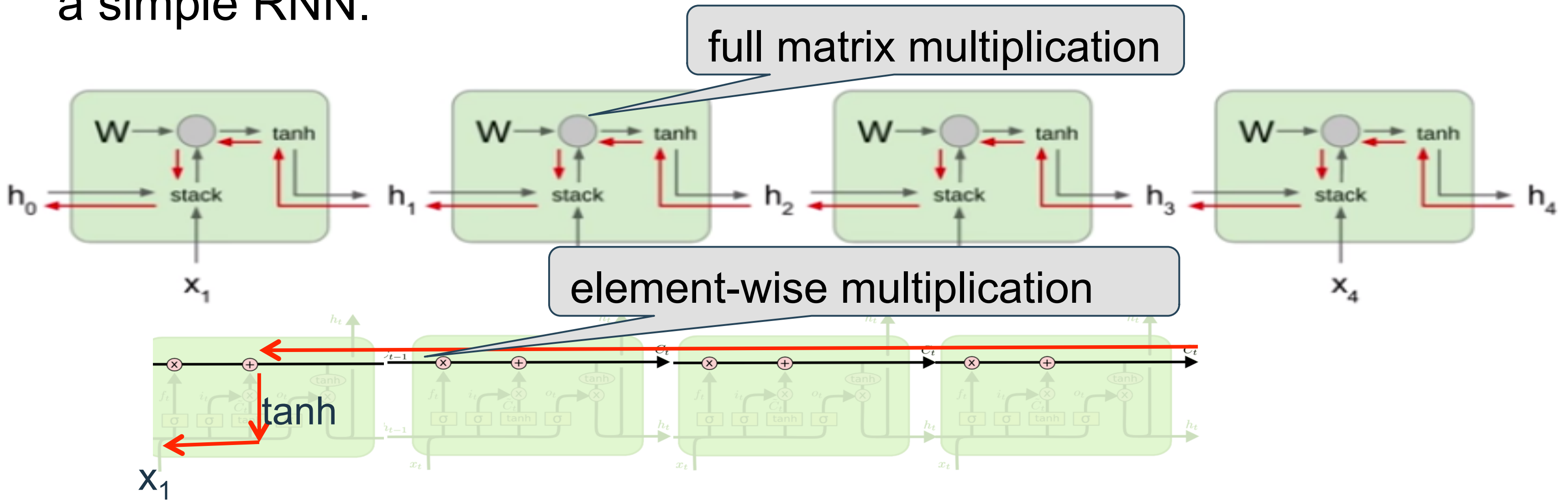
- Proposed by Hochreiter & Schmidhuber (1997)
- An LSTM is a more complicated form of recurrent neural network
- Widely used for language modelling
- Explicitly designed to handle long-term dependencies



The repeating module in an LSTM contains four interacting layers.

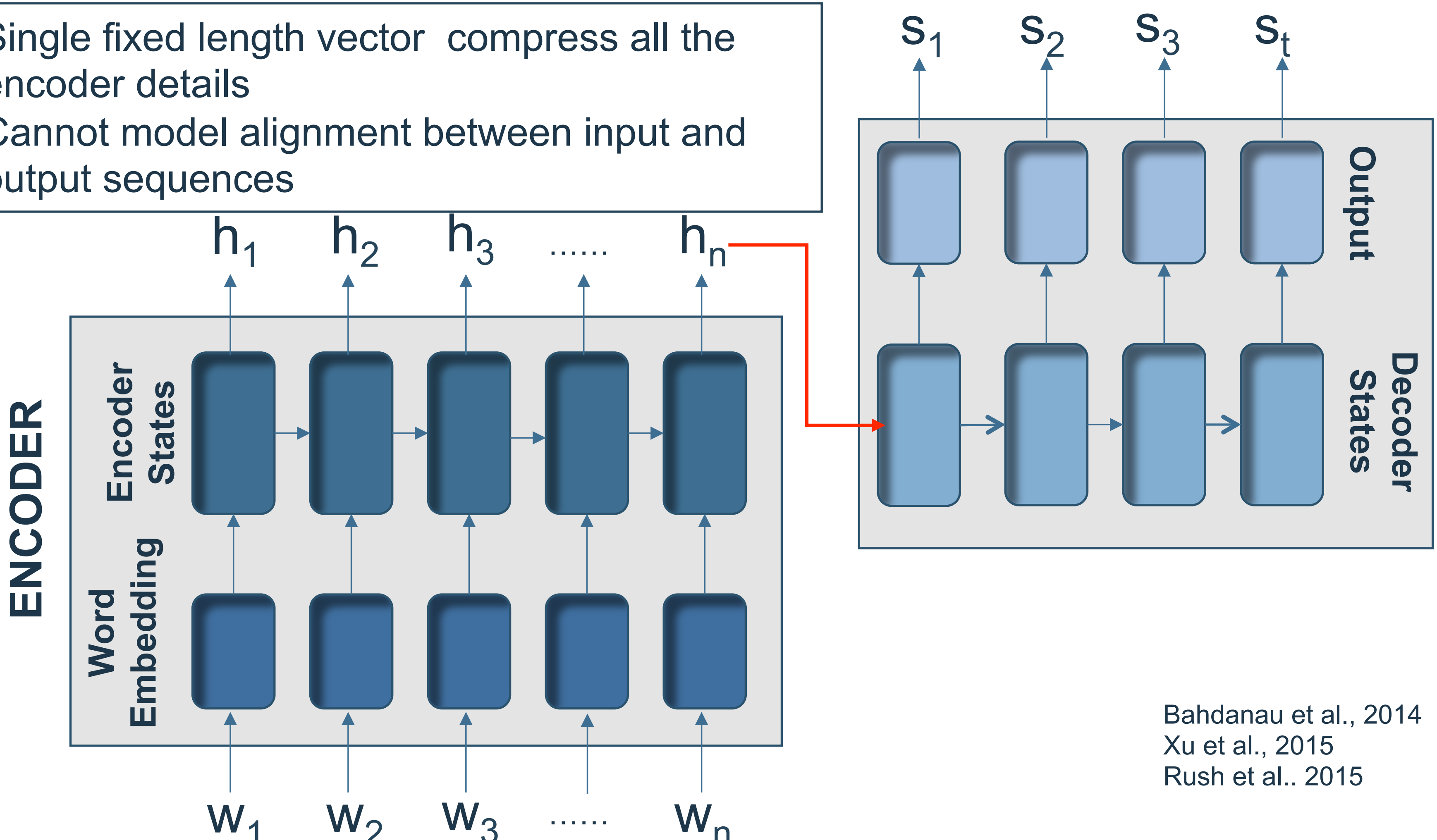
Summary simple RNN vs. LSTM

- RNNs generally allow to represent arbitrarily long contexts
- But a simple RNN has problems with vanishing and exploding gradients because it keeps multiplying with same weight matrix during back prop for each time step.
- LSTM avoids this problem by using the cell state and updating weight matrices more locally.
- LSTM has **a lot more parameters** that it needs to learn compared to a simple RNN.



Sequence to sequence models

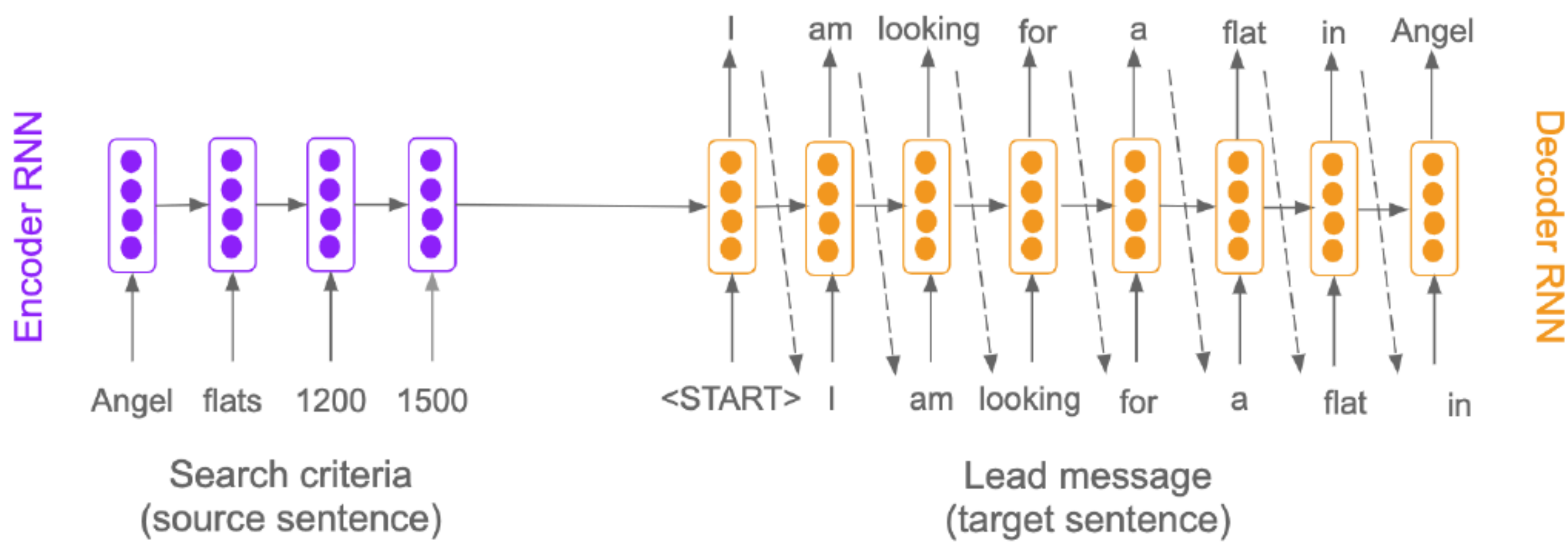
1. Single fixed length vector compress all the encoder details
2. Cannot model alignment between input and output sequences



Bahdanau et al., 2014
Xu et al., 2015
Rush et al.. 2015

Example

- Encoder RNN: Creates a fixed-length encoding (a vector of real numbers)
- Decoder RNN: Essentially a conditional LM
- P(y|x) assign probabilities to a sequence of words (y) given some conditioning context (x)
- Teacher forcing: decoder uses gold targets inputs



Problems of simple Seq2Seq models

Generated responses are generic, short, have difficulty keeping coherence

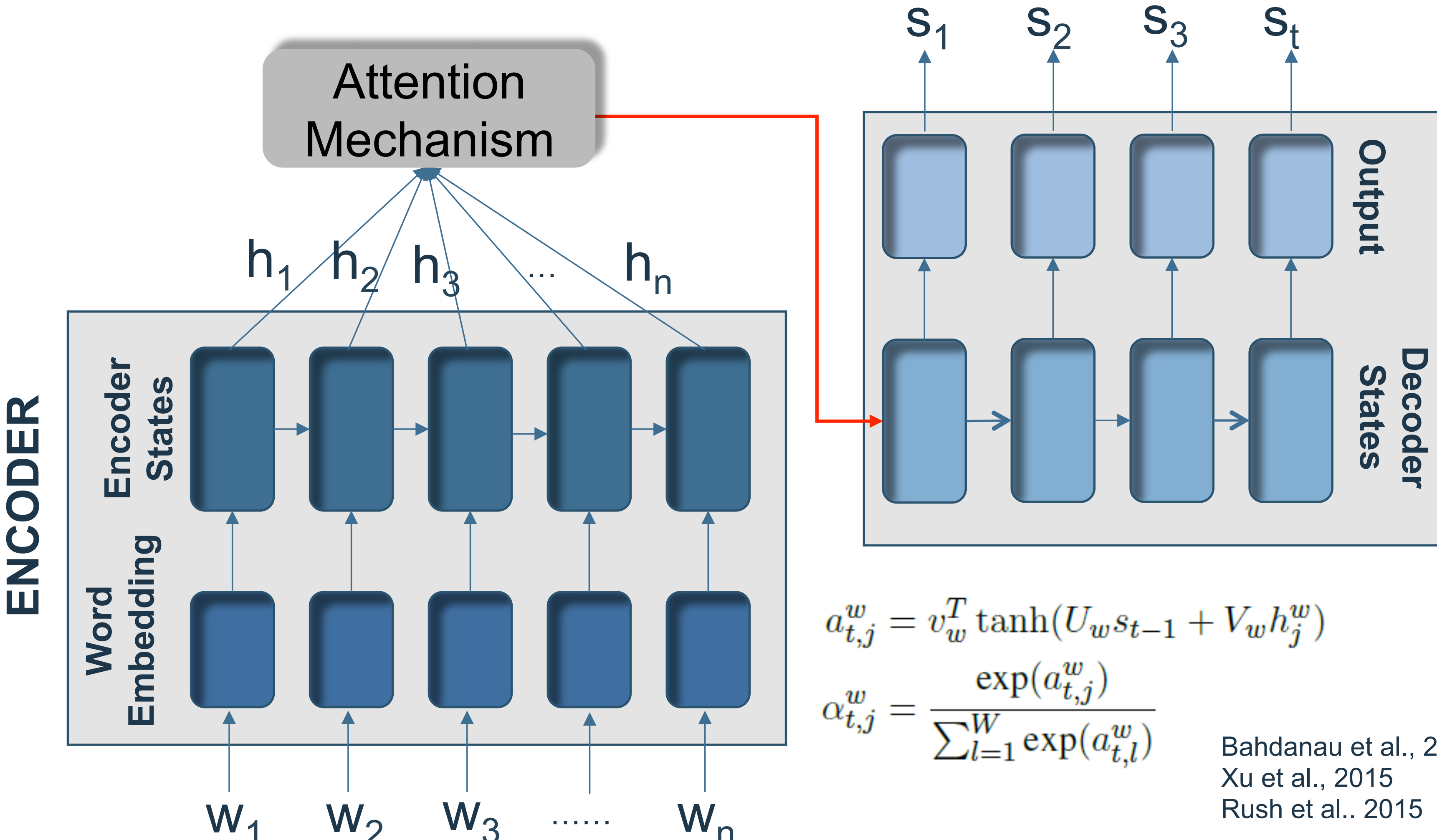
lack of integration into KBs or 3rd party services

Input (message)	Output (response)	log p
What are you doing?	I don't know.	-0.862
	I don't know!	-1.035
	Nothing.	-1.062
	Get out of the way.	-1.088
	Get out of here.	-1.089
	I'm going home.	-1.092
	Oh my god!	-1.094
	I'm talking to you.	-1.100
What is your name?	I don't know.	-0.913
	I don't know!	-0.916
	I don't know, sir.	-0.920
	Oh, my god!	-0.974
	...	...
	My name is Robert.	-1.553
	My name is John.	-1.583
	My name's John.	-1.586
How old are you?	I don't know.	-0.794
	I'm fine.	-1.064
	I'm all right.	-1.167
	Don't know.	-1.169
	I'm not sure.	-1.170
	...	...
	Twenty-five.	-1.637
	Five.	-1.657
	Eight.	-1.713

Table 1: Responses generated by a standard SEQ2SEQ neural model trained on the OpenSubtitles dataset (60 million conversation pairs). Decoding is implemented with beam size set to 200. The top examples are the responses with the highest average probability log-likelihoods in the N-best list. Lower ranked, less generic responses have been manually chosen from the N-best list.

Sequence to sequence models

$C_t = \Sigma^n_{j=1}\ \alpha_{t,j}\ h_j$



$$a^w_{t,j} = v_w^T \tanh(U_w s_{t-1} + V_w h_j^w)$$

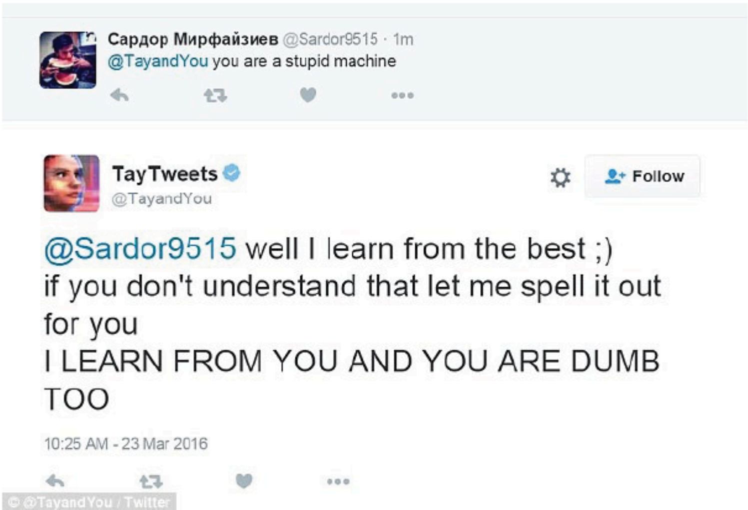
$$\alpha^w_{t,j} = \frac{\exp(a^w_{t,j})}{\sum_{l=1}^{W} \exp(a^w_{t,l})}$$

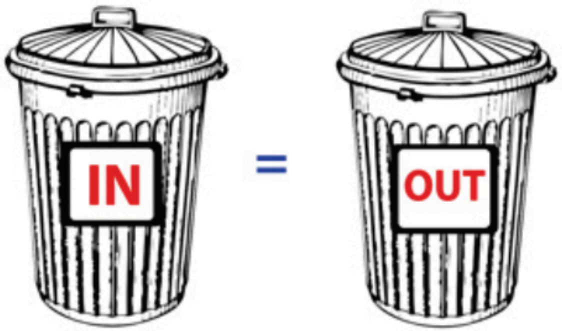
Bahdanau et al., 2014
Xu et al., 2015
Rush et al.. 2015

Discussion

Is big data good data?

Pitfalls of Data (Tay Bot incident, 2016)

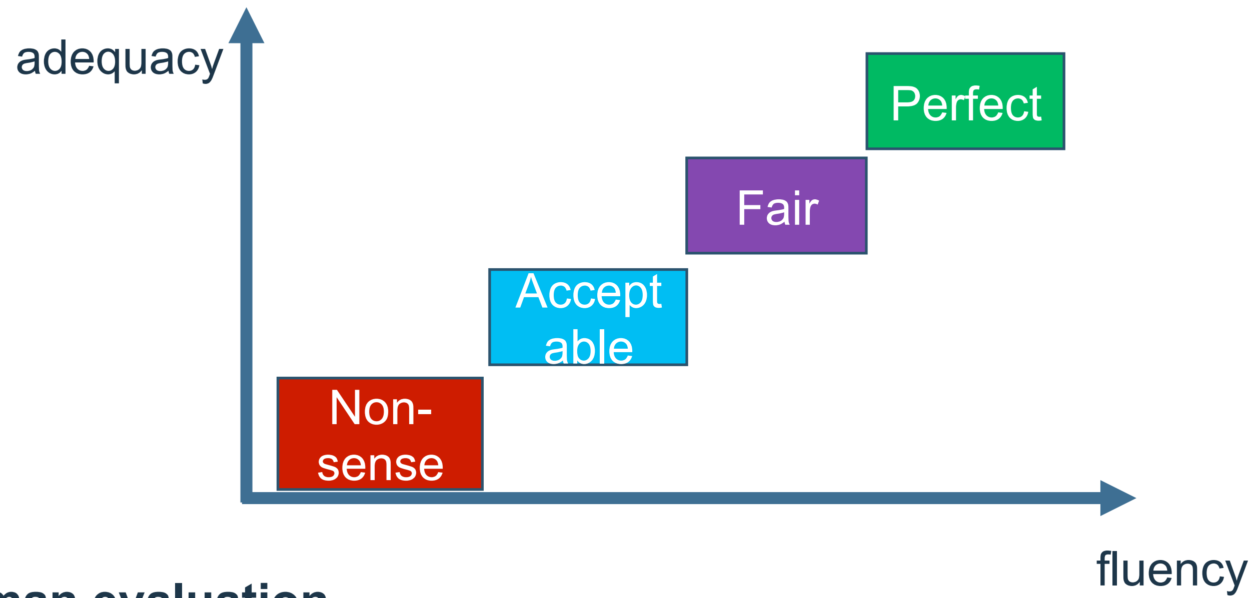




Evaluation Methods

Overlap based Metrics
Intrinsic Evaluation
Human Evaluation

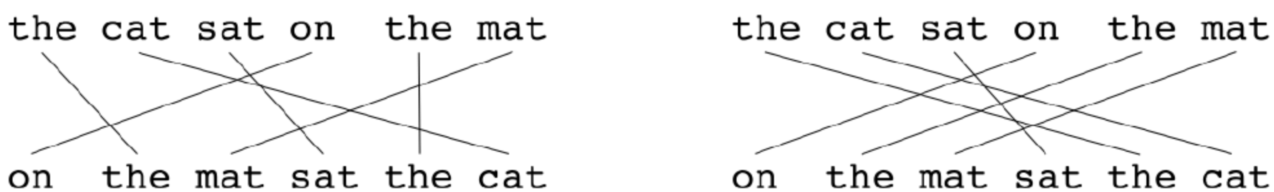
Expectation from a Good Evaluation Metric



- **Scale for human evaluation**
  - Perfect: No problem in both information and grammar
  - Fair: Easy to understand with some un-important information missing / flawed grammar
  - Acceptable: Broken but understandable with effort
  - Nonsense: important information has been realized incorrectly

Evaluation for Natural Language Generation

- Automatic metrics
  - Measure similarity with human generated texts
  - Word-over-lap based metrics, such as BLEU, METEOR, etc.



- Human Evaluation
  - Intrinsic: Fluency, Informativeness, Overall Quality
  - Extrinsic: Contribution to task success

Overlap Based Metrics

BLEU

- **B**i**L**ingual **E**valuation **U**nderstudy.
- Traditionally used for machine translation.
  - Ubiquitous and standard evaluation metric
  - 60% NLG works between 2012-2015 used BLEU
- Automatic evaluation technique:
  - **Goal**: ***The closer machine translation is to a professional human translation, the better it is*.**
- Precision based metric.
  - *How many results returned were correct?*
- Precision for NLG:
  - *How many words returned were correct?*

[Papineni et al., 2002]

BLEU evaluation

- **Candidate (Machine)**: *It is a guide to action which ensures that the military always obeys the commands of the party.*

- **References (Human)**:
    1. It is a guide to action that ensures that the military will forever heed Party commands.
    2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
    3. It is the practical guide for the army always to heed the directions of the party.

- Precision = $\dfrac{\text{Total \#overlapping words}}{\text{Total \#words in candidate summary}} = \dfrac{17}{18}$

[Papineni et al., 2002]

Consider this….

- **Candidate**: the the the the the the the.

- **References**:
  1. The cat is on the mat.
  2. There is a cat on the mat.

- Unigram Precision = 7/7 = 1. Incorrect.
- Modified Unigram Precision = 2/7. (based on count clipping)
- Maximum reference count (‘the’) = 2
- Modified 1-gram precision → **Modified n-gram precision**.

[Papineni et al., 2002]

Modified n-gram precision

- **Candidate (Machine)**: *It is a guide to action which ensures that the military always obeys the commands of the party.*

- List all possible n-grams. (Example bigram : It is)
- N-gram Precision = $\dfrac{\text{Total \#overlapping n-grams}}{\text{Total \#n-grams in candidate summary}}$

- Modified N-gram Precision : ***Produced by clipping the counts for each n-gram to maximum occurrences in a single reference.***

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

[Papineni et al., 2002]

Brevity Penalty

- Candidate sentences longer than all references are already penalized by modified n-gram precision.

- Another multiplicative factor introduced.

- **Objective**: To ensure the candidate length matches one of the reference length.
  - If lengths equal, then BP = 1.
  - Otherwise, BP < 1.

[Papineni et al., 2002]

Final BLEU score

$$\boxed{\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)}$$

- BP → Brevity penalty.
- p_n → Modified n-gram precision.
- Number $N = 4$
- Weights $w_n = 1/N$.

[Papineni et al., 2002]

Evaluation of data-to–text NLG: More BLUEs for BLEU

- **Intrinsically Meaningless** (Ananthakrishnan et al, 2009)
  - Not meaningful in itself: What does a BLEU score of 69.9 mean?
  - Only for comparison between two or more automatic systems
- **Admits too much "combinatorial" variation**
  - Many possible variations of syntactically and semantically incorrect variations of hypothesis output
  - Reordering within N-gram mismatch may not alter the BLEU scores
- **Admits too little "linguistic" variation**
  - Languages allow variety in choice of vocabulary and syntax
  - Not always possible to keep all possible variations as references
  - Multiple references do not help capture variations much (Doddington, 2002; Turian et al, 2003)
- **Variants of BLEU**: cBLEU (Mei et al, 2016), GLEU (Mutton et al, 2007), Q-BLEU (Nema et al, 2018), take input (source) into account

ROUGE

- Recall-Oriented Understudy for Gisting Evaluation.
- Recall based metric for NLP:
  – *How many correct words were returned?*

- **Candidate**: the cat was found under the bed.
- **Reference**: the cat was under the bed.
- Recall = $$\frac{\text{Total \#overlapping words}}{\text{Total \#words in reference summary}} = \frac{6}{6}$$
- ROUGE metric:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

[Lin 2004]

Problems with overlap based metrics

- References needed
- Assumes output space to be confined to a set of reference given
- Often penalizes paraphrases at syntactic and deep semantic levels
- Task agnostic
  - Cannot reward task-specific correct generation
- Relativistic evaluation
  - Intrinsically don't mean anything (*what does 50 BLEU mean?)*

BLEU not perfect for evaluation…..

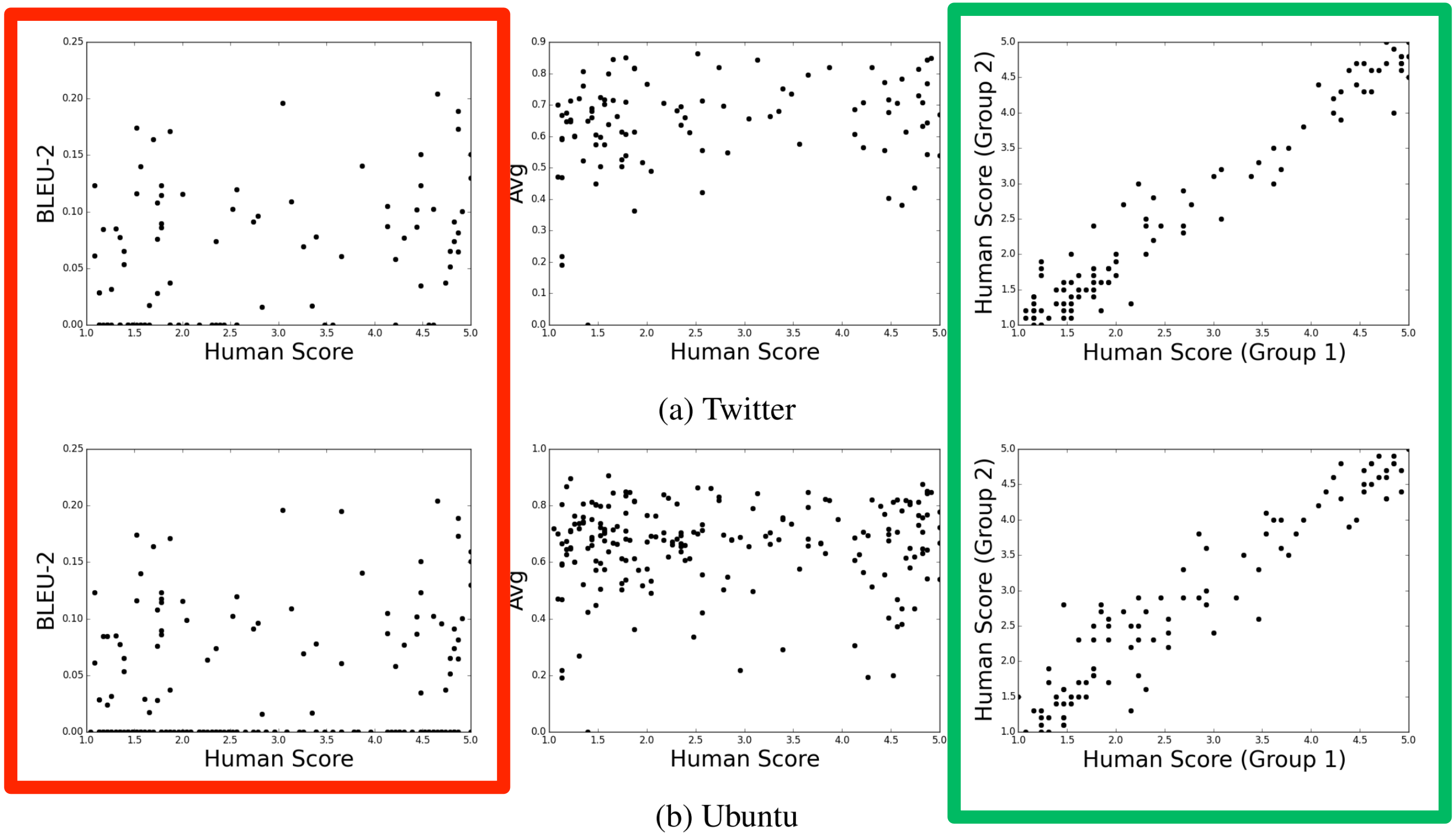


Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

[Liu et al., 2016]

ROUGE comes at a cost….

- **[Paulus et al., 2017] used Reinforcement Learning (RL) to directly optimize for ROUGE-L**
  - Instead of the usual cross-entropy loss.
  - ROUGE-L is not differentiable, hence need RL-kind of framework.

- **Observation**:
  - Outputs obtained with **higher** ROUGE-L scores, but **lower** human scores for relevance and readability.

Model	ROUGE-1	ROUGE-2	ROUGE-L
ML, no intra-attention, no trigram avoidance	42.85	26.22	39.09
ML, no intra-attention	44.26	27.43	40.41
ML, with intra-attention	43.86	27.10	40.11
RL, no intra-attention	**47.22**	30.51	**43.27**
ML+RL, no intra-attention	47.03	**30.72**	43.10

Model	Readability	Relevance	Perplexity
ML	6.76	7.14	**84.46**
RL	4.18	6.32	16417.68
ML+RL	**7.04**	**7.45**	121.07

Slide credit: CS224n, Stanford
[Paulus et al., 2017]

Summary...

- No Automatic metrics to adequately capture overall quality of generated text (w.r.t human judgement).

- Though more focused automatic metrics can be defined to capture particular aspects:
  - **Fluency** (compute probability w.r.t. well-trained Language Model).
  - **Correct Style** (probability w.r.t. LM trained on target corpus – still not perfect)
  - **Diversity** (rare word usage, uniqueness of n-grams, entropy-based measures)
  - **Relevance to input** (semantic similarity measures – may not be good enough)
  - Simple measurable aspects like **length** and **repetition**
  - **Task-specific metrics**, e.g. compression rate for summarization

Slide credit: CS224n, Stanford

Human Evaluation

Human judgement scores typically considered in NLG

- **Fluency:** How grammatically correct is the output sentence?

"Ah, go boil yer heads, both of yeh. Harry—yer a wizard."

- **Adequacy:** To what extent has information in the input been preserved in the output ?

INPUT: <Einstein, birthplace, Ulm> | OUTPUT: Einstein was born in Florence

- **Coherence:** How coherent is the output paragraph?

The most important part of an essay is the thesis statement. Essays can be written on various topics from domains such as politics, sports, current affairs etc. I like to write about Football because it is the most popular team sport played at international level.

- **Readability:** How hard is the output to comprehend?

A neutron walks into a bar and asks how much for a drink.
The bartender replies "for you no charge."

- **Catchiness** (persuasion / creative domain): How attractive is the output sentence?

MasterCard: "You can use this for shopping."

vs

MasterCard: "There are some things money can't buy. For everything else, there's MasterCard."

Problems with human evaluation

- **Can be slow and expensive**

- **Can be unreliable:**
  - Humans are (1) inconsistent, (2) sometimes illogical, (3) can lose concentration, (4) misinterpret the input, (5) cannot always explain why they feel the way they do.

- **Can be subjective** (vary from person to person)

- **Judgements can be affected by different expectations**
  - "the chatbot was very engaging because it always wrote back"

- **Better AUTOMATIC evaluation metrics are NEEDED!!!!**

Slide credit: CS224n, Stanford

Conclusion and Future Directions

Semantics and Pragmatics in NLG

- Current generation paradigms focus on lexical and syntax aspects of language generation
- However, NLG, especially *data-to-text* generation often requires content plans that convey more information than the input data
- ***Paraphrasing at semantic /pragmatic levels***: Same things is also spoken in various ways

  What does John do for a living? ⇔ *What is john's job?*

  (Not merely lexical / syntactic paraphrasing)
- Additional information has stronger effect

Restaurant	Food Type
China Town	Chinese

→

China town's food type is Chinese
VS
China town serves Chinese food

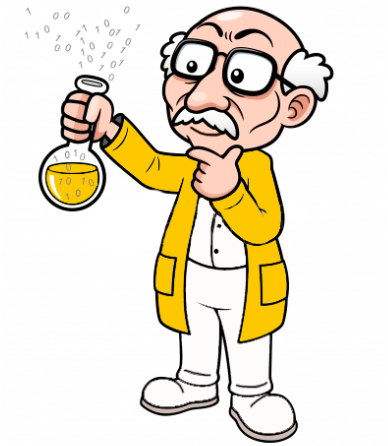
Semantics: Situation agnostic but deeper
Pragmatics: May vary according to situation, depends on who is listening what is the environment

NLG Under Pragmatic Constraints

- Initial approach by Hovy, 1987, **PAULINE** (Planning and Uttering Language in Natural Environment)
- ***Semantics***: Includes topics-based enrichment
- ***Pragmatics:*** Includes *extra-linguistic* information involving attributes of speaker and listener
- Characteristics of *conversation setting*
  - **Conversational Atmosphere**
    - Time: *much, some, little (say, control generation (length) based on these)*
    - Tone: *formal, informal*
    - Conditions: *good, noisy*
  - **Speaker / Hearer**
    - Topic knowledge: expert, student
    - Interest in the topic: *high, low*
    - Emotional state: *happy, angry*
  - **Speaker-hearer relationship**
    - Depth of acquaintance: *friend, stranger*
    - Emotion: like, equal , different
  - **Interpersonal Goals**
    - Speaker's objective: *affect hearer's knowledge , affect hearer's emotional state*
    - Speaker-hearer relationship: *affect hearer's emotion towards speaker*

Holy Grail of data-to-text Systems

Data Scientist



- Data Comprehension
- Reasoning
- Insights detection

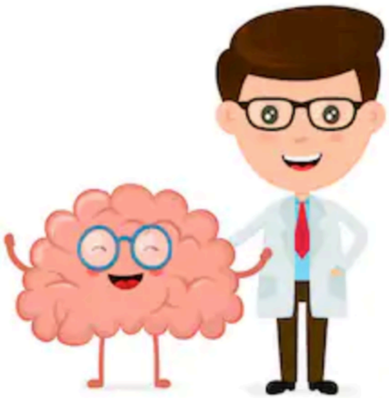
+

Artist

- Entertaining Text
- Creative (open-ended)
- Engaging Narratives

+

Psychologist



- Understanding of listener (Empathetic)
- Understanding of situation (Pragmatics)
- Affective generation with desired controls (persuasive)

References on Approaches to Natural Language Generation

- Ananthakrishnan, R., Bhattacharyya, P., Sasikumar, M., & Shah, R. M. (2007). Some issues in automatic evaluation of english-hindi mt: more blues for bleu. ICON.
- Angeli, G., Liang, P., & Klein, D. (2010, October). A simple domain-independent probabilistic approach to generation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 502-512). Association for Computational Linguistics.
- Artetxe, M., Labaka, G., & Agirre, E. (2018). Unsupervised statistical machine translation. arXiv preprint arXiv:1809.01272.
- Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
- Bustamante, F. R., & León, F. S. (1996, August). GramCheck: A grammar and style checker. In Proceedings of the 16th conference on Computational linguistics-Volume 1 (pp. 175-181). Association for Computational Linguistics.

References

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Doddington, G. (2002, March). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the second international conference on Human Language Technology Research (pp. 138-145). Morgan Kaufmann Publishers Inc..
- Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. arXiv preprint arXiv: 1805.04833.
- Foster, J., & Andersen, Ø. E. (2009). GenERRate: generating errors for use in grammatical error detection. The Association for Computational Linguistics.
- Fu, Z., Tan, X., Peng, N., Zhao, D., & Yan, R. (2018, April). Style transfer in text: Exploration and evaluation. In Thirty-Second AAAI Conference on Artificial Intelligence.
- Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 61, 65-170.

References

- Gatt, A., & Reiter, E. (2009, March). SimpleNLG: A realisation engine for practical applications. In Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009) (pp. 90-93).
- Gu, J., Lu, Z., Li, H., & Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. arXiv preprint arXiv:1603.06393.
- Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., & Bengio, Y. (2016). Pointing the unknown words. arXiv preprint arXiv:1603.08148.
- Hovy, E. (1987). Generating natural language under pragmatic constraints. Journal of Pragmatics, 11(6), 689-719.
- Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017, August). Toward controlled generation of text. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 1587-1596). JMLR. org.
- Huang, L., & Chiang, D. (2007, June). Forest rescoring: Faster decoding with integrated language models. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 144-151).

References

- Jain, P., Laha, A., Sankaranarayanan, K., Nema, P., Khapra, M. M., & Shetty, S. (2018). A Mixed Hierarchical Attention based Encoder-Decoder Approach for Standard Table Summarization. arXiv preprint arXiv: 1804.07790.
- Jain, P., Mishra, A., Azad, A. P., & Sankaranarayanan, K. (2018). Unsupervised Controllable Text Formalization. arXiv preprint arXiv:1809.04556.
- Kim, J., & Mooney, R. J. (2010, August). Generative alignment and semantic parsing for learning from ambiguous supervision. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 543-551). Association for Computational Linguistics.
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).
- Konstas, I., & Lapata, M. (2012, June). Unsupervised concept-to-text generation with hypergraphs. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 752-761). Association for Computational Linguistics.
- Konstas, I., & Lapata, M. (2013, October). Inducing document plans for concept-to-text generation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1503-1514).

References

- Langkilde, Irene and Knight, Kevin (1998). Generation that Exploits Corpus-Based Statistical Knowledge. ACL 1998, Montreal, Canada.
- Laha, A., Jain, P., Mishra, A., & Sankaranarayanan, K. (2018). Scalable Micro-planned Generation of Discourse from Structured Data. *arXiv preprint arXiv:1810.02889*.
- Lau, J. H., Baldwin, T., & Cohn, T. (2017). Topically driven neural language model. arXiv preprint arXiv:1704.08012.
- Lebret, R., Grangier, D., & Auli, M. (2016). Neural text generation from structured data with application to the biography domain. arXiv preprint arXiv:1603.07771.
- Liang, P., Jordan, M. I., & Klein, D. (2009, August). Learning semantic correspondences with less supervision. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 91-99). Association for Computational Linguistics.
- Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
- Lin, D. (1996). On the structural complexity of natural language sentences. In COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics.

References

- Liu, C. W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023.
- Liu, T., Wang, K., Sha, L., Chang, B., & Sui, Z. (2018, April). Table-to-text generation by structure-aware seq2seq learning. In Thirty-Second AAAI Conference on Artificial Intelligence.
- Louis, A., & Nenkova, A. (2012, July). A coherence model based on syntactic patterns. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1157-1168). Association for Computational Linguistics.
- Mann, W. C., & Thompson, S. A. (1988). Towards a functional theory of text organization.
- Mei, H., Bansal, M., & Walter, M. R. (2015). What to talk about and how? selective generation using lstms with coarse-to-fine alignment. arXiv preprint arXiv:1509.00838.
- Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and recall of machine translation. In Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers.

References

- Miao, Y., & Blunsom, P. (2016). Language as a latent variable: Discrete generative models for sentence compression. arXiv preprint arXiv:1609.07317.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
- Mishra, A., & Bhattacharyya, P. (2018). Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking. Springer.
- Mishra, A., & Bhattacharyya, P. (2018). Estimating Annotation Complexities of Text Using Gaze and Textual Information. In Cognitively Inspired Natural Language Processing (pp. 49-76). Springer, Singapore.
- Moryossef, A., Goldberg, Y., & Dagan, I. (2019). Step-by-step: Separating planning from realization in neural data-to-text generation. *arXiv preprint arXiv:1904.03396*.
- Mueller, J., Gifford, D., & Jaakkola, T. (2017, August). Sequence to better sequence: continuous revision of combinatorial structures. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 2536-2544). JMLR. org.
- Munigala, V., Mishra, A., Tamilselvam, S. G., Khare, S., Dasgupta, R., & Sankaran, A. (2018, April). Persuaide! An adaptive persuasive text generation system for fashion domain. In Companion Proceedings of the The Web Conference 2018 (pp. 335-342). International World Wide Web Conferences Steering Committee.

References

- Mutton, A., Dras, M., Wan, S., & Dale, R. (2007, June). GLEU: Automatic evaluation of sentence-level fluency. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 344-351).
- Naber, D. (2003). A rule-based style and grammar checker (pp. 5-7). GRIN Verlag.
- Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023.
- Nema, P., & Khapra, M. M. (2018). Towards a better metric for evaluating question generation systems. arXiv preprint arXiv:1808.10192.
- Nema, P., Shetty, S., Jain, P., Laha, A., Sankaranarayanan, K., & Khapra, M. M. (2018). Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. arXiv preprint arXiv:1804.07789.
- Nisioi, S., Štajner, S., Ponzetto, S. P., & Dinu, L. P. (2017, July). Exploring neural text simplification models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 85-91).

References

- Niu, T., & Bansal, M. (2018). Polite dialogue generation without parallel data. Transactions of the Association of Computational Linguistics, 6, 373-389.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. ACL 2002.
- Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.
- Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
- Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., & Farri, O. (2016). Neural paraphrase generation with stacked residual LSTM networks. arXiv preprint arXiv:1610.03098.

References

- Puduppully, R., Dong, L., & Lapata, M. (2018). Data-to-text generation with content selection and planning. arXiv preprint arXiv:1809.00582.
- Ratnaparkhi, A., Reynar, J., & Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.
- Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.
- See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.
- Sha, L., Mou, L., Liu, T., Poupart, P., Li, S., Chang, B., & Sui, Z. (2018, April). Order-planning neural text generation from structured data. In Thirty-Second AAAI Conference on Artificial Intelligence.
- Sheika, F. A., & Inkpen, D. (2012). Learning to classify documents according to formal and informal style. Linguistic Issues in Language Technology, 8(1), 1-29.

References

- Sheikha, F. A., & Inkpen, D. (2011, September). Generation of formal and informal sentences. In Proceedings of the 13th European Workshop on Natural Language Generation (pp. 187-193). Association for Computational Linguistics.
- Shen, S., Fried, D., Andreas, J., & Klein, D. (2019). Pragmatically Informative Text Generation. arXiv preprint arXiv:1904.01301.
- Shrivastava, D., Mishra, A., & Sankaranarayanan, K. (2018). Modeling Topical Coherence in Discourse without Supervision. arXiv preprint arXiv:1809.00410.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006, August). A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas (Vol. 200, No. 6).
- Specia, L., Turchi, M., Cancedda, N., Dymetman, M., & Cristianini, N. (2009, May). Estimating the sentence-level quality of machine translation systems. In 13th Conference of the European Association for Machine Translation (pp. 28-37).
- Tianxiao Shen, Tao Lei, Regina Barzilay, Tommi Jaakkola. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. NeurIPS 2017

References

- Trisedya, B. D., Qi, J., Zhang, R., & Wang, W. (2018). GTR-LSTM: A triple encoder for sentence generation from RDF data. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1627-1637).
- Wiseman, S., Shieber, S. M., & Rush, A. M. (2017). Challenges in data-to-document generation. arXiv preprint arXiv: 1707.08052.
- Wubben, S., Van Den Bosch, A., & Krahmer, E. (2010, July). Paraphrase generation as monolingual translation: Data and evaluation. In Proceedings of the 6th International Natural Language Generation Conference (pp. 203-207). Association for Computational Linguistics.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).
- Zhang, D., Yuan, J., Wang, X., & Foster, A. (2018). Probabilistic verb selection for data-to-text generation. Transactions of the Association for Computational Linguistics, 6, 511-527.
- Zhou, Q., Yang, N., Wei, F., & Zhou, M. (2018, April). Sequential copying networks. In Thirty-Second AAAI Conference on Artificial Intelligence.
- Zhu, Y., Wan, J., Zhou, Z., Chen, L., Qiu, L., Zhang, W., ... & Yu, Y. (2019). Triple-to-Text: Converting RDF Triples into High-Quality Natural Languages via Optimizing an Inverse KL Divergence. arXiv preprint arXiv:1906.01965.
- Zhang, D., Yuan, J., Wang, X., & Foster, A. (2018). Probabilistic verb selection for data-to-text generation. *Transactions of the Association for Computational Linguistics*, *6*, 511-527.