Utrecht University

DSCC Central Topic Seminar

Large-scale Transfer Learning for Natural Language Generation
–
Application to Conversational Agents

2019/01/11

thomas@huggingface.co

A quick introduction to Natural Language Processing

Human Language

Human Language

- A **distinctive feature** of Homo sapiens
  Sapiens: A Brief History of Humankind by Yuval Noah Harari

Human Language

- A **distinctive feature** of Homo sapiens
  Sapiens: A Brief History of Humankind by Yuval Noah Harari
- Very flexible and can be used to convey information about a very wide range of **objects and concepts**

Human Language

- A **distinctive feature** of Homo sapiens
  Sapiens: A Brief History of Humankind by Yuval Noah Harari
- Very flexible and can be used to convey information about a very wide range of **objects and concepts**
- The **storage format of** a huge portion of total *human knowledge*
  Scientific articles, books, encyclopedia, discourse transcriptions…

Human Language

- A **distinctive feature** of Homo sapiens
  Sapiens: A Brief History of Humankind by Yuval Noah Harari
- Very flexible and can be used to convey information about a very wide range of **objects and concepts**
- The **storage format of** a huge portion of total *human knowledge*
  Scientific articles, books, encyclopedia, discourse transcriptions…

We call the languages we (human beings) use to communicate together « **Natural Languages** » to distinguish them from *formal languages* like logic, mathematics, programming languages…

Natural Language Processing

Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

Natural Language Processing

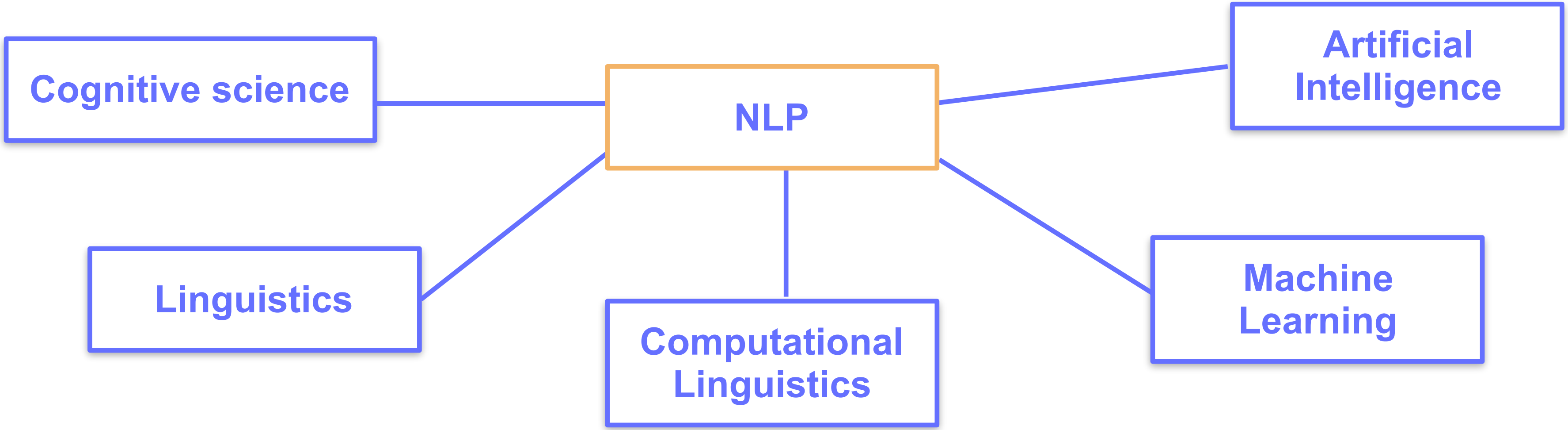
Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

NLP is an **engineering field** (like « building-planes ») standing on the shoulders of several **science** fields (like « fluid mechanics »):

Natural Language Processing

Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

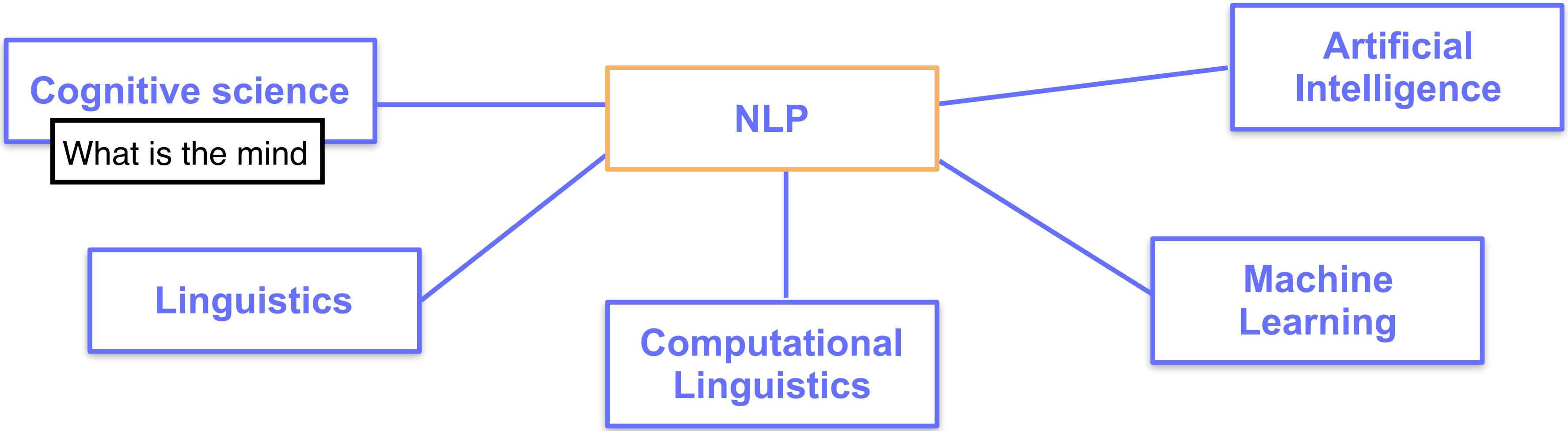
NLP is an **engineering field** (like « building-planes ») standing on the shoulders of several **science** fields (like « fluid mechanics »):



Natural Language Processing

Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

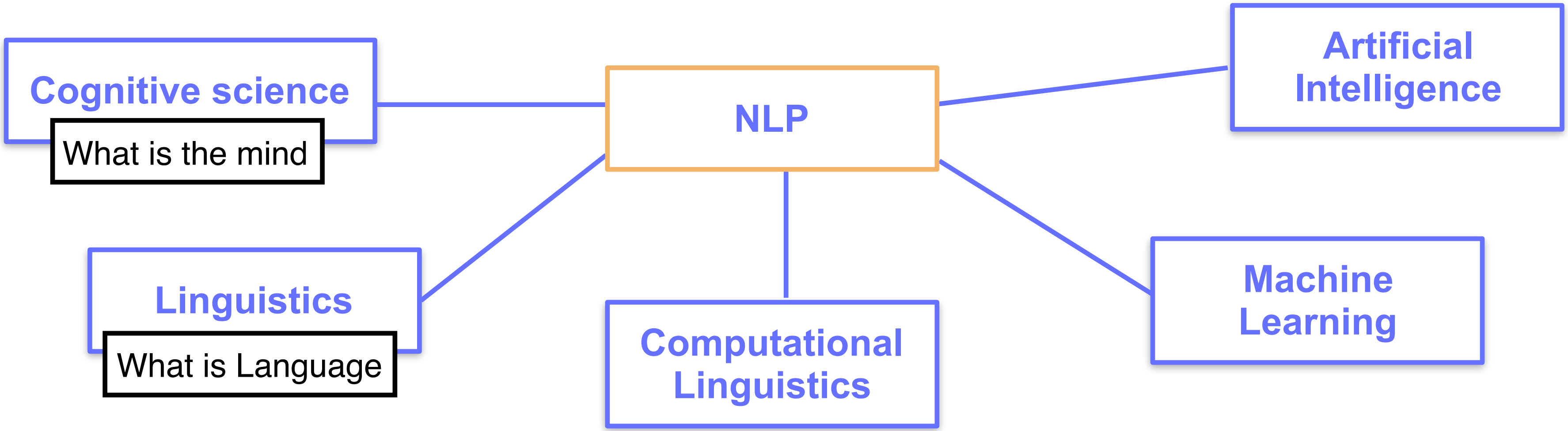
NLP is an **engineering field** (like « building-planes ») standing on the shoulders of several **science** fields (like « fluid mechanics »):



Natural Language Processing

Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

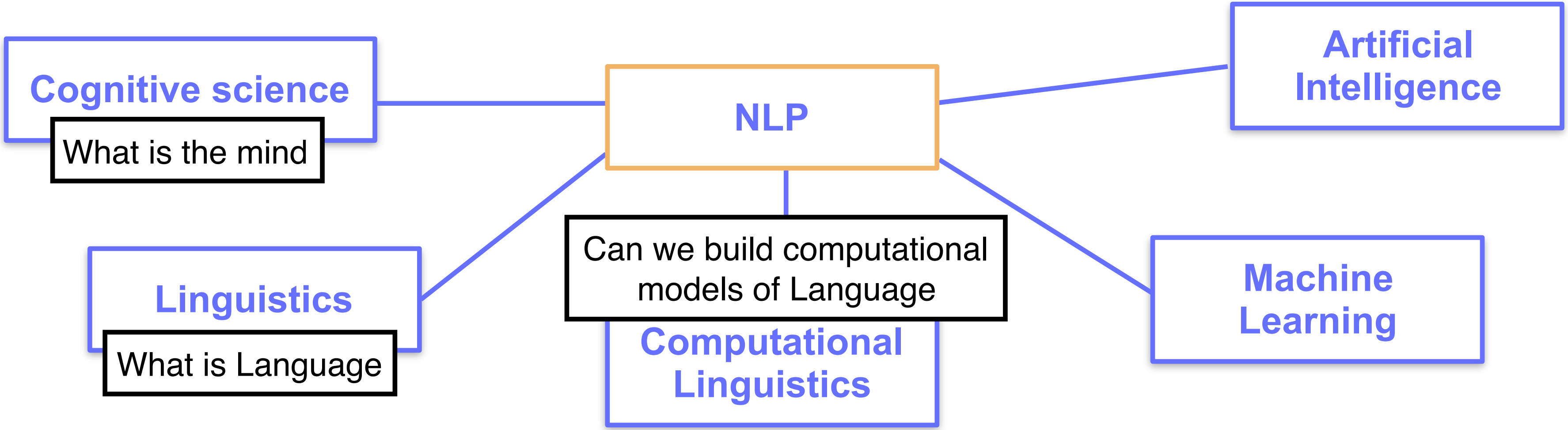
NLP is an **engineering field** (like « building-planes ») standing on the shoulders of several **science** fields (like « fluid mechanics »):



Natural Language Processing

Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

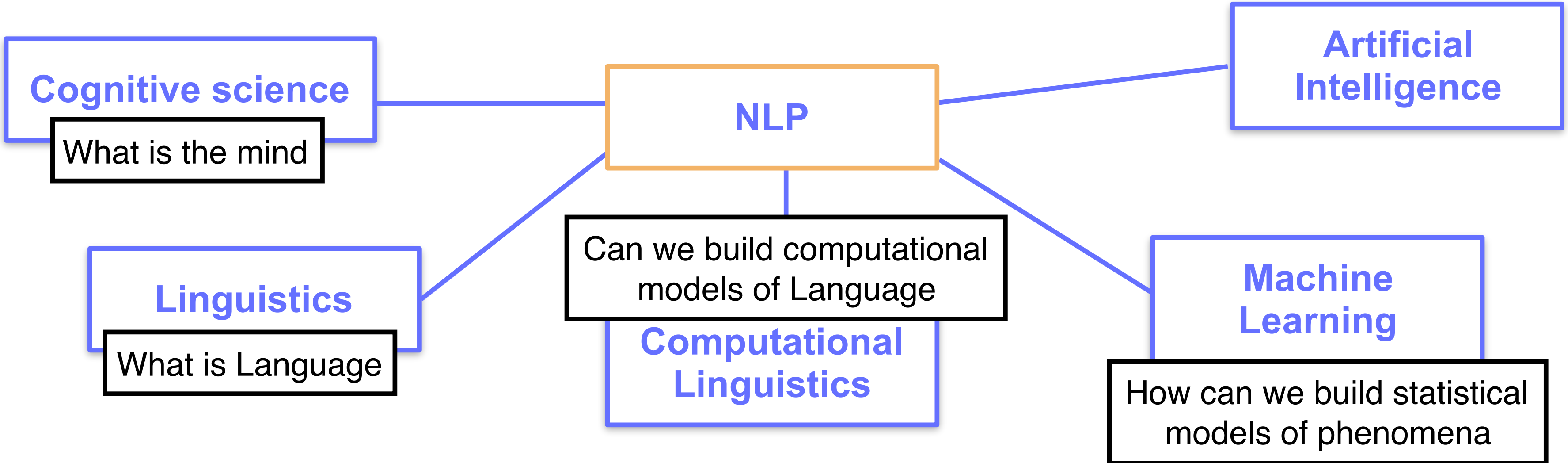
NLP is an **engineering field** (like « building-planes ») standing on the shoulders of several **science** fields (like « fluid mechanics »):



Natural Language Processing

Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

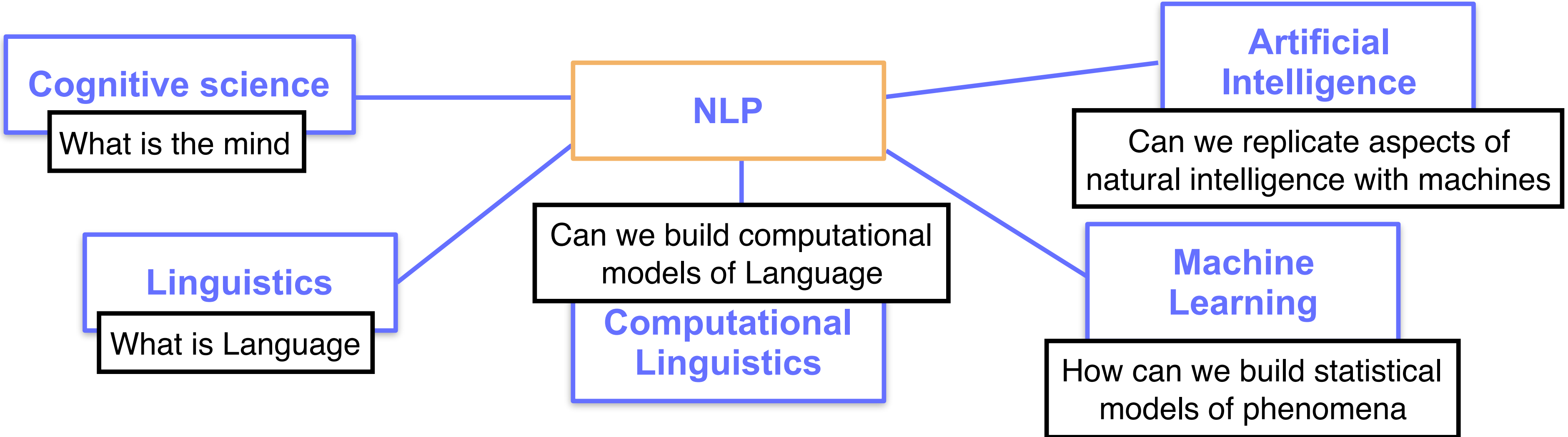
NLP is an **engineering field** (like « building-planes ») standing on the shoulders of several **science** fields (like « fluid mechanics »):



Natural Language Processing

Natural Language Processing (NLP): how to program computers to *process* and analyze (large amounts of) *natural language* data.

NLP is an **engineering field** (like « building-planes ») standing on the shoulders of several **science** fields (like « fluid mechanics »):



Natural Language Processing

Tasks in NLP that can be generally grouped as:

- **Natural Language Understanding** (NLU) (text is an *input*)
  - Information extraction (ex. from scientific publications)
  - Basis for down-stream systems that uses text as input
  - …
- **Natural Language Generation** (NLG) (text is an *output*)
  - Used to communicate with human beings
  - Store human-readable information
  - …

Natural Language Processing

Tasks in NLP that can be generally grouped as:

- **Natural Language Understanding** (NLU) (text is an *input*)
  - Information extraction (ex. from scientific publications)
  - Basis for down-stream systems that uses text as input
  - …
- **Natural Language Generation** (NLG) (text is an *output*)
  - Used to communicate with human beings
  - Store human-readable information
  - …

In my talk I’ll call:

- **Text** all natural language data
  - i.e. word/sentences/document written in a human readable language
- **Data** all the rest
  - i.e. numbers, graphs, formal languages, images, speech…

Natural Language Generation

Today we'll focus on **Natural Language Generation** (NLG):

Computer programs which
generate **human-readable text**
as their **output**.

Many theoretical reasons to study this task

- **General interest**:
  - Give hints to understand *human language* and *cognition*:
    - McDonald (2010) NLG: « the process by which thought is rendered into language »
    - Cognitive research on language production (Kukich 1987, Elman 1990, 1993, Chang et al. 2006)
    - Linguistics – Theories on the emergence/acquisition of language
  - Most of *human knowledge* is stored in natural language form in books/encyclopedia.

- **In the field of Artificial Intelligence:**
  - *Debugging and understanding* our AI systems:
    - Strong incentive to make « black-box » AI models more interpretable
    - Human way to explain a decision is by using natural language
  - Enabling unsupervised learning (the problem of *data availability* in NLP):
    - Recent NLP/AI systems require huge datasets that are expensive to annotate (ex annotate entities in text, write translation)
    - Can we learn general concepts by learning to generate language?
    - This is called Transfer Learning

Ecosystem of Natural Language Generation

Science

- Computational Linguistics
- Cognitive Science
- Machine Learning

Natural Language Generation

Art

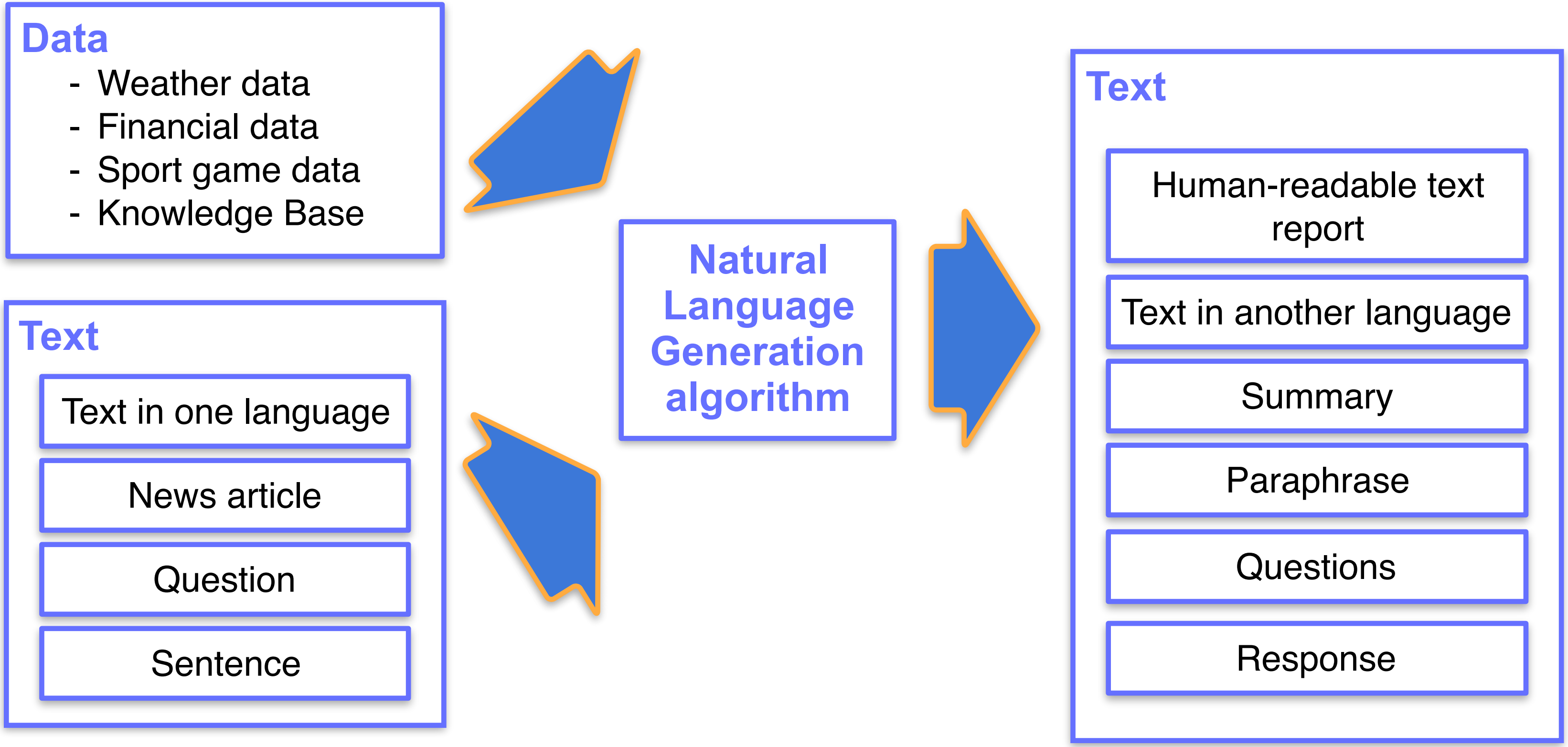
- Computational Creativity
- Interactive Art (improvisation…)
- AI Augmented Art (inspiration tool…)

Applications

- Machine Translation
- Image captioning
- Entertainment
- HCI
- Summarization
- Business Intelligence
- Customer support
- …

Anatomy of a Natural Language Generation System

Anatomy of an NLG system



So how do you generate text?

- **Modular approach**:
  - You think a lot and split the task in several sub-tasks:

- *Good things:*
  - You can iterate on each module and combine various methods
- *Bad things:*
  - Language generation doesn't work in a top-down fashion

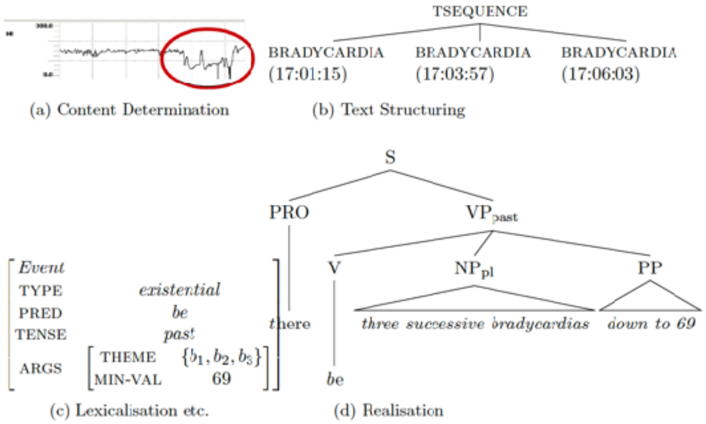
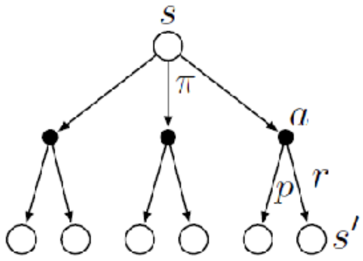


Figure 1: Tasks in NLG, illustrated with a simplified example from the neonatal intensive care domain. First the system has to decide what the important events are in the data (a, content determination), in this case, occurrences of low heart rate (bradycardias). Then it has to decide in which order it wants to present data to the reader (b, text structuring) and how to express these in individual sentence plans (c, aggregation, lexicalisation, reference). Finally, the resulting sentences are generated (d, linguistic realisation).

So how do you generate text?

- **Planning based approach**:
  - You are in a current state.
  - You are taking action by generating a word and end up in a new state.



Backup diagram for v_π

How you model the current state

Discrete state

Continuous dense vector

How you select the actions

Rule-based, planning-based approaches
Recent twist: Reinforcement learning

Neural networks

How do you learn to generate?

- Hand-based
  - You write the rules.
    - Works well for small and delimited fields
    - Cannot generalize to new fields
    - Brittle since you have to think about every corner case

- Statistical approaches
  - Recently became the prominent approach
  - Gather a dataset (as large as possible) of example by crowdsourcing
  - Use machine learning tools to learn the parameters of your system from this dataset

Let’s see that on a real example

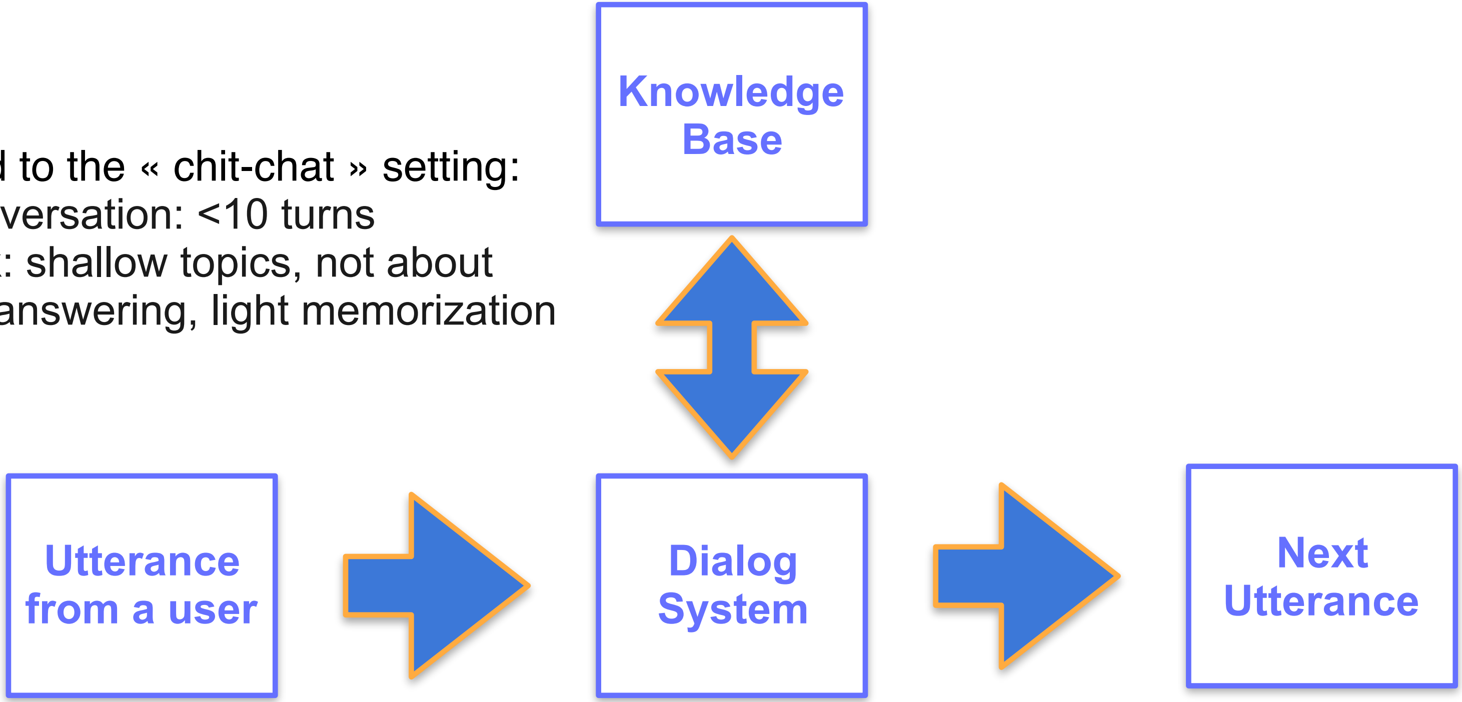
Open-Domain Conversational Agents

Open-Domain Conversational Agents

A conversational agent which can talk about any topic

Often restricted to the « chit-chat » setting:

- Short conversation: <10 turns
- Small talk: shallow topics, not about question-answering, light memorization



Open-Domain Conversational Agents

Example of training dataset – Evaluation dataset: PERSONA-CHAT (Zhang et al. 2018)

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

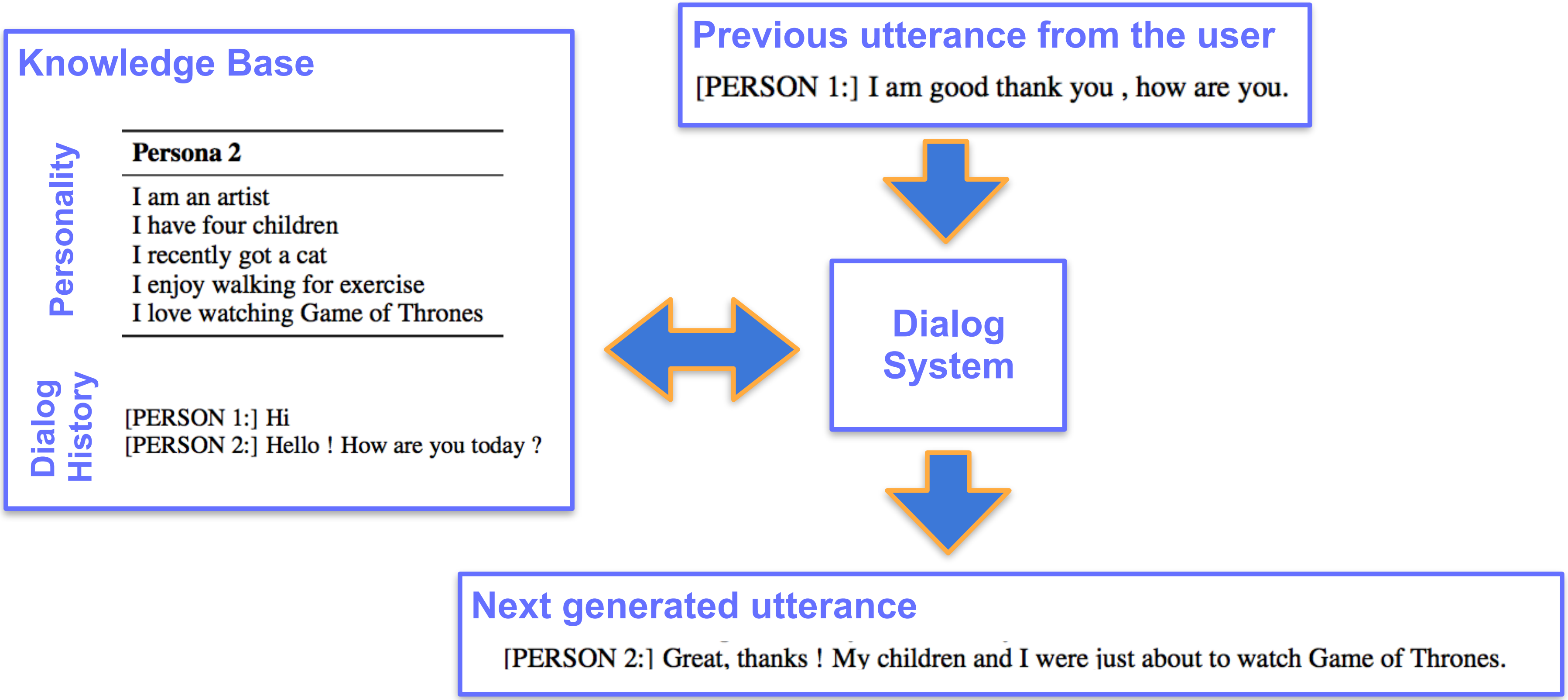
[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Example dialog from the PERSONA-CHAT dataset. Person 1 is given their own persona (top left) at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation.

- Amazon Mechanical Turkers were:
  - **paired** by two,
  - each given a **personality** comprising 4-5 simple sentences, and
  - asked to **talk** together in order to get to know each other.
- Resulted in a dataset of
  - **10,981 dialogs** comprising
  - **164,356 utterances** and about **1-2M words**
  - Average number of turns: **14**

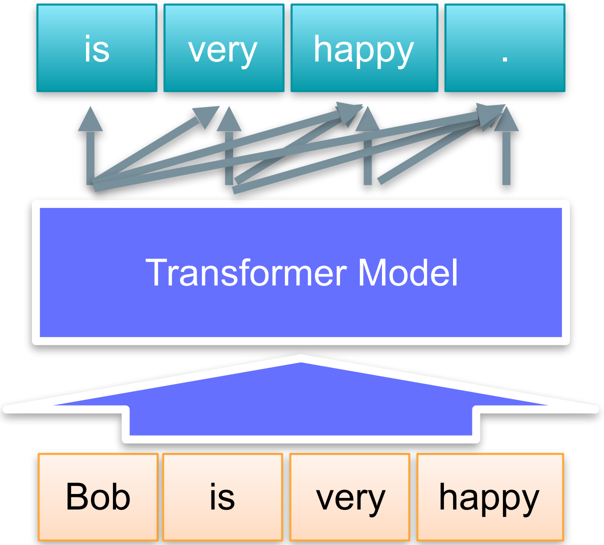
Open-Domain Conversational Agents

How does this work in our case?



Dialog System

- Our Dialog System has two elements:
  - A **generative model** which generate the words one by one given the context,
  - A **decoder** which controls the generative model.



The **generative model** is a **Transformer Model** which has recently became a major model in NLP.

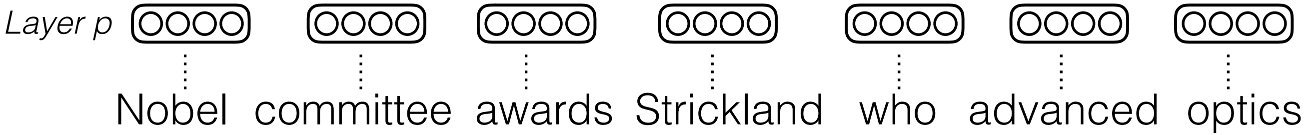
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]

Nobel committee awards Strickland who advanced optics

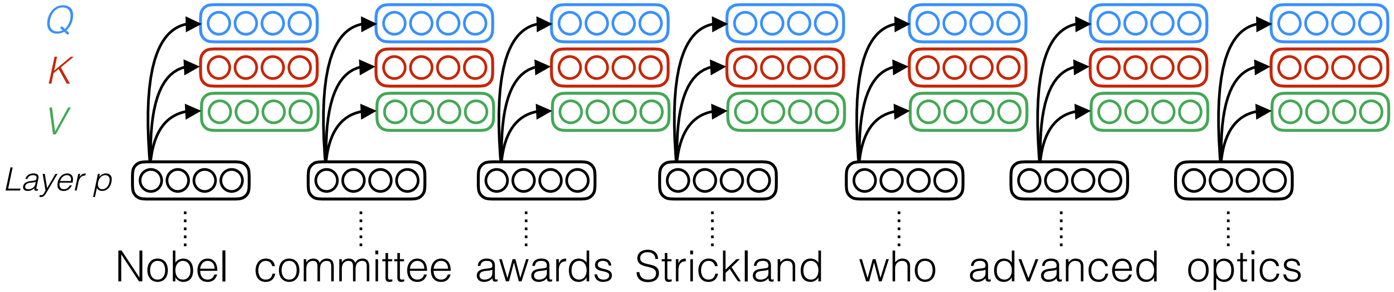
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



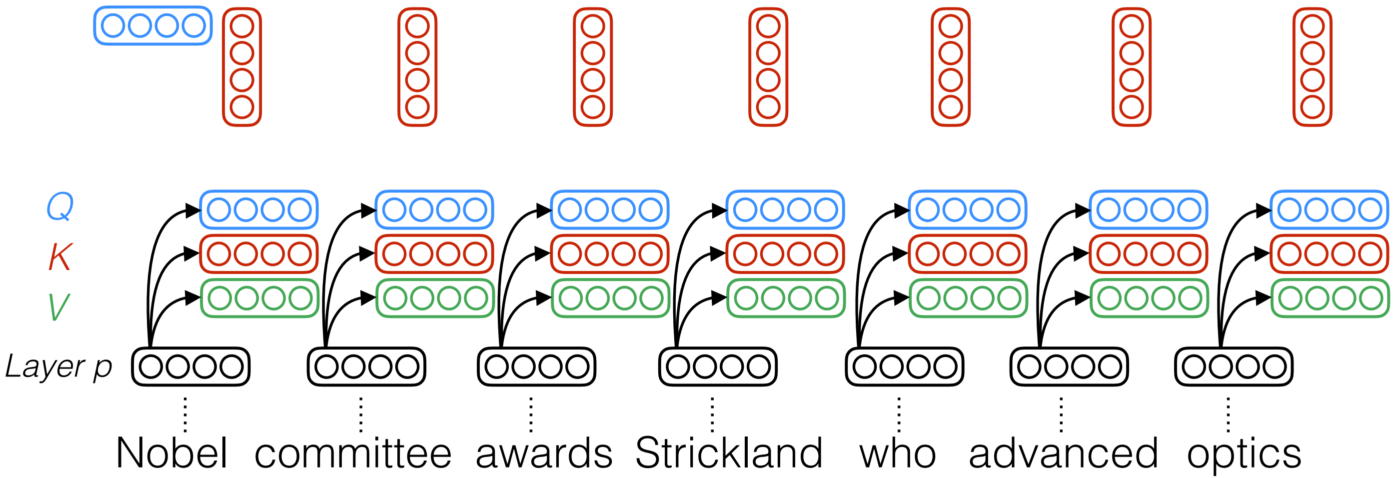
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



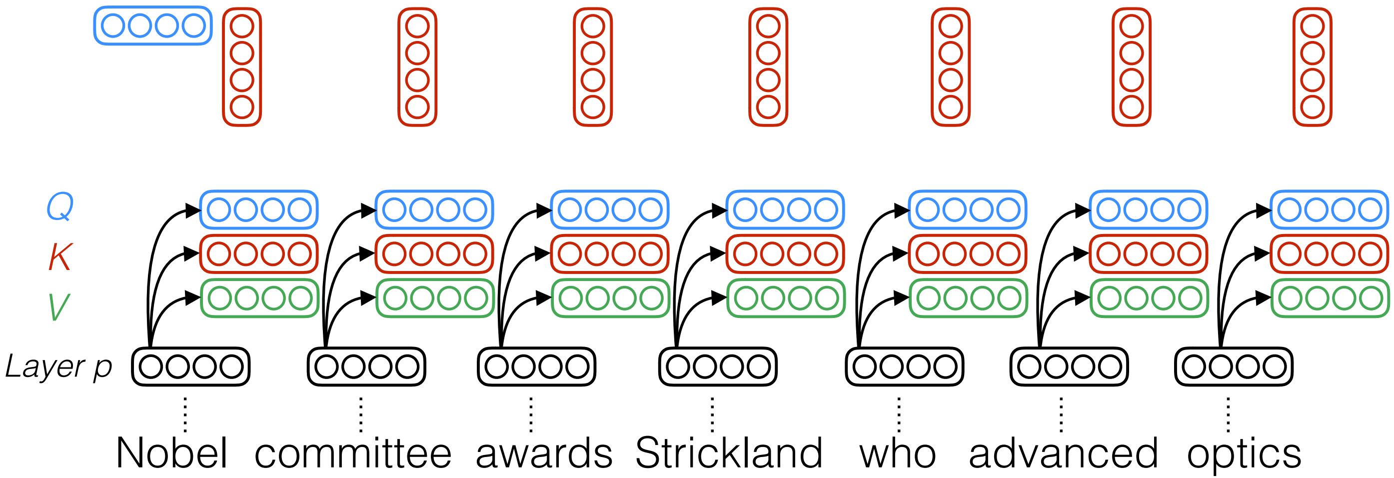
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



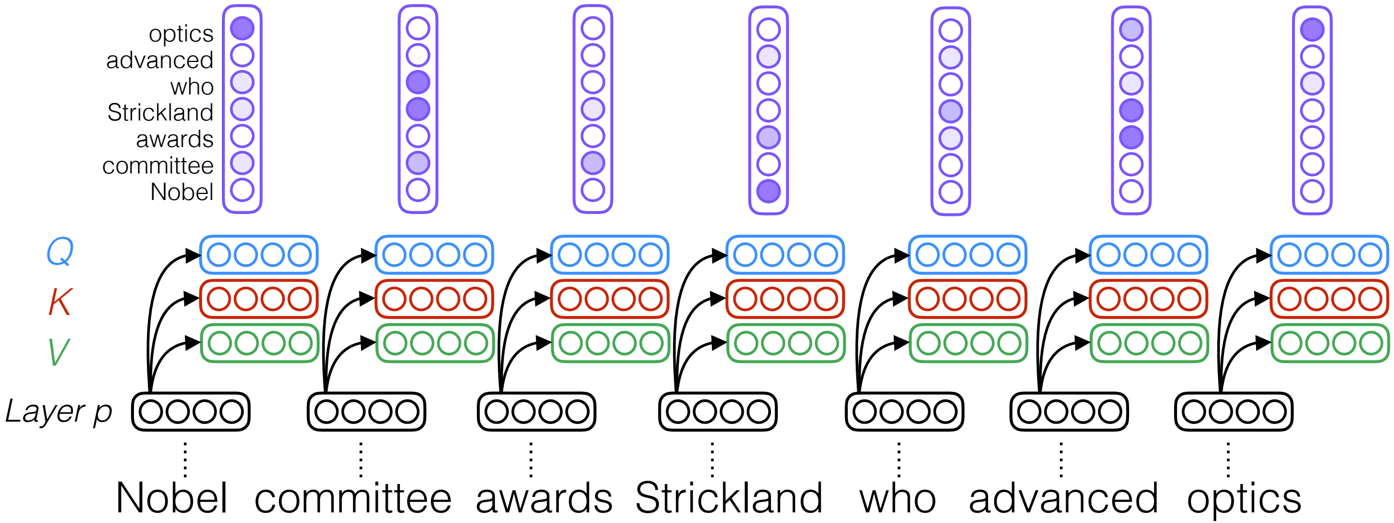
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



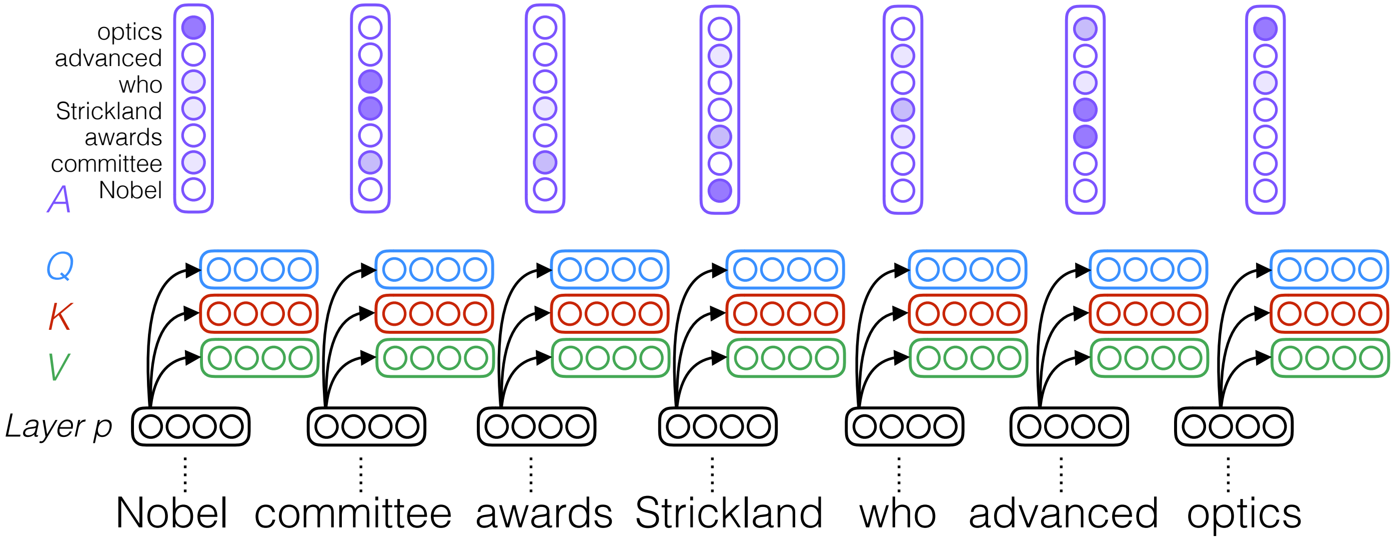
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



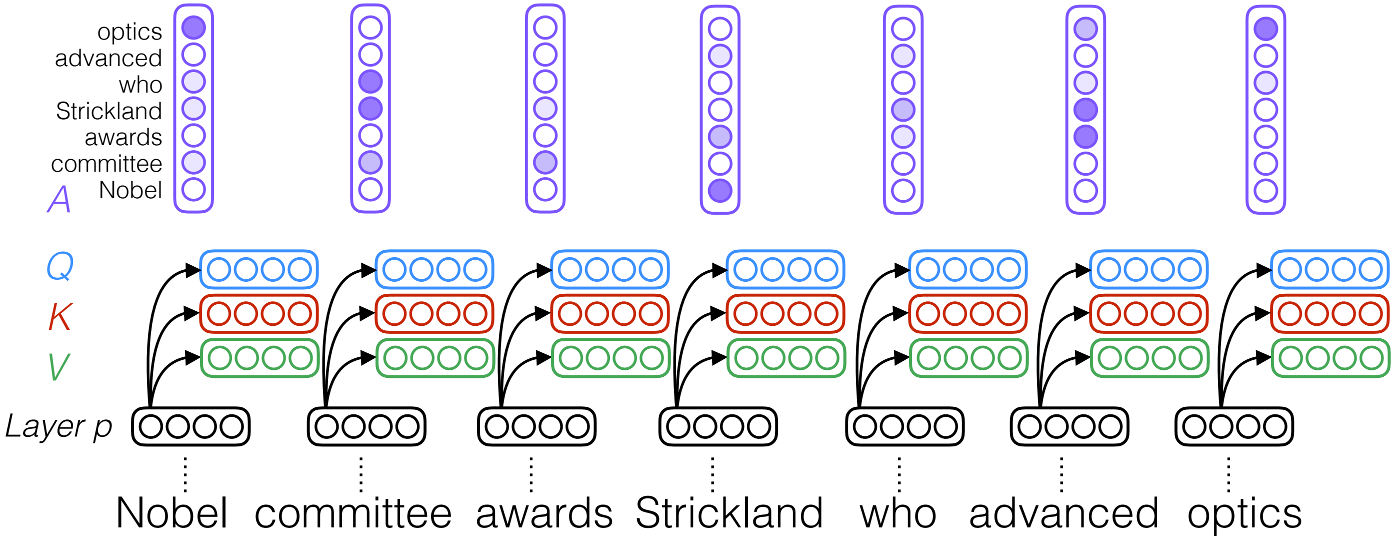
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



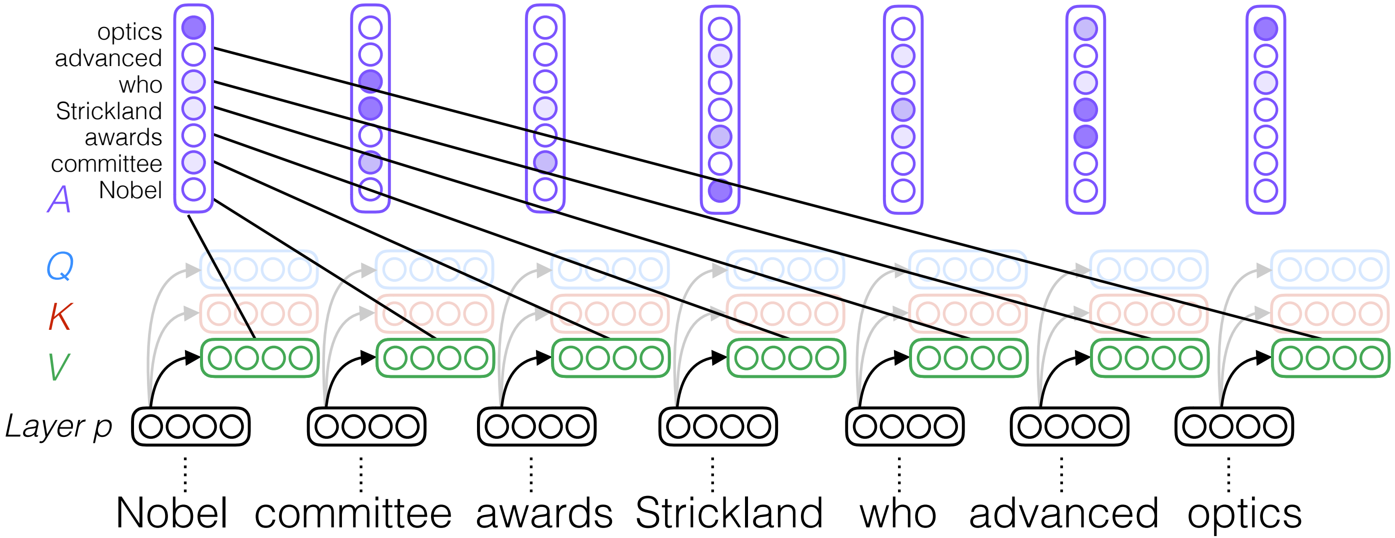
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



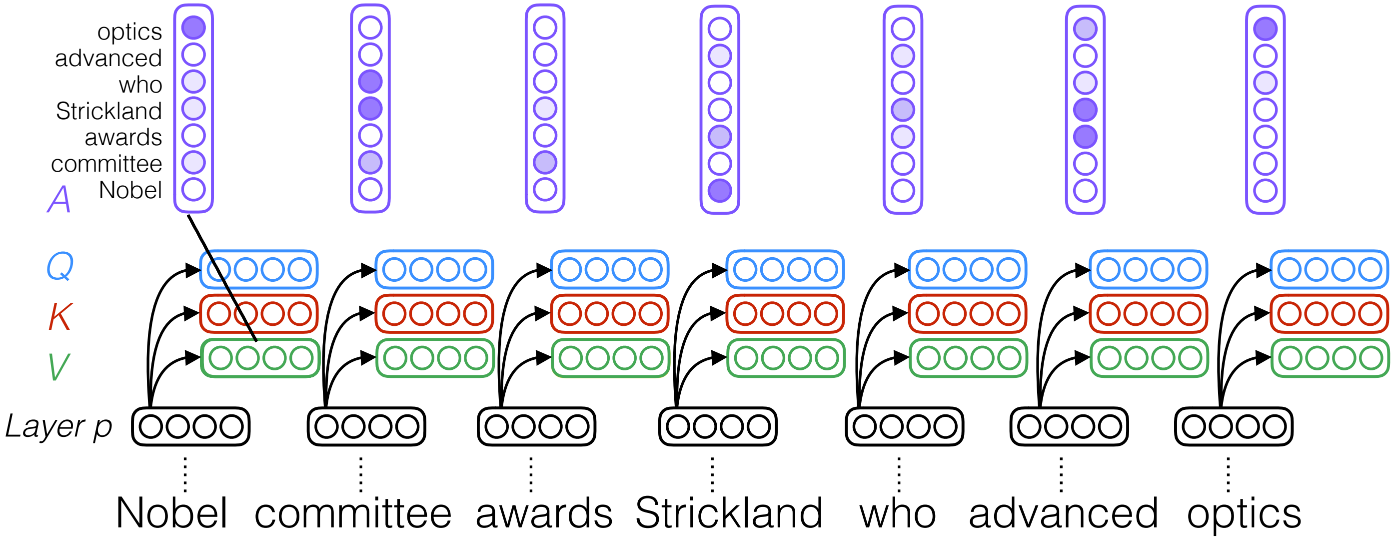
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



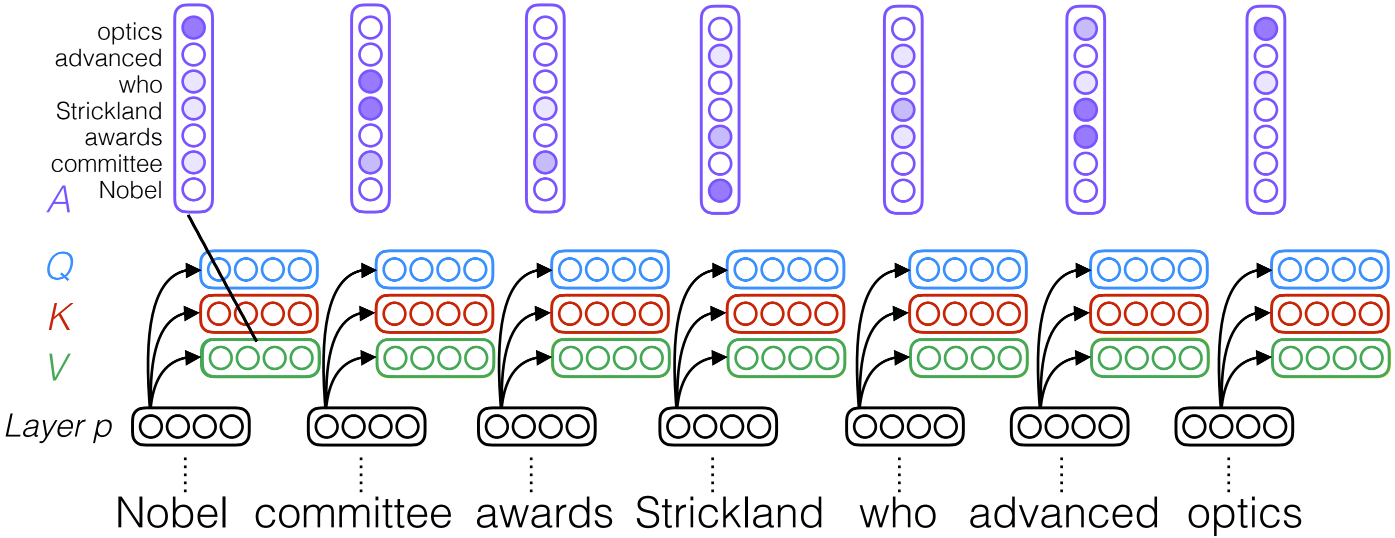
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



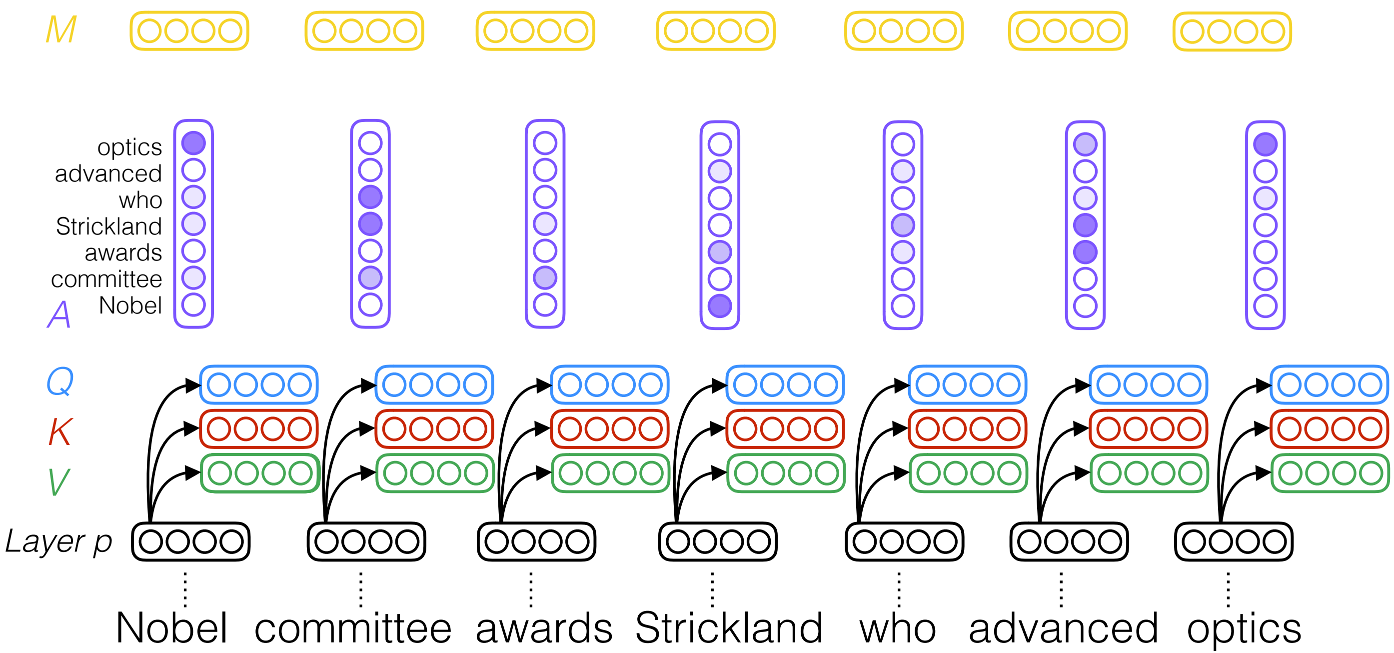
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



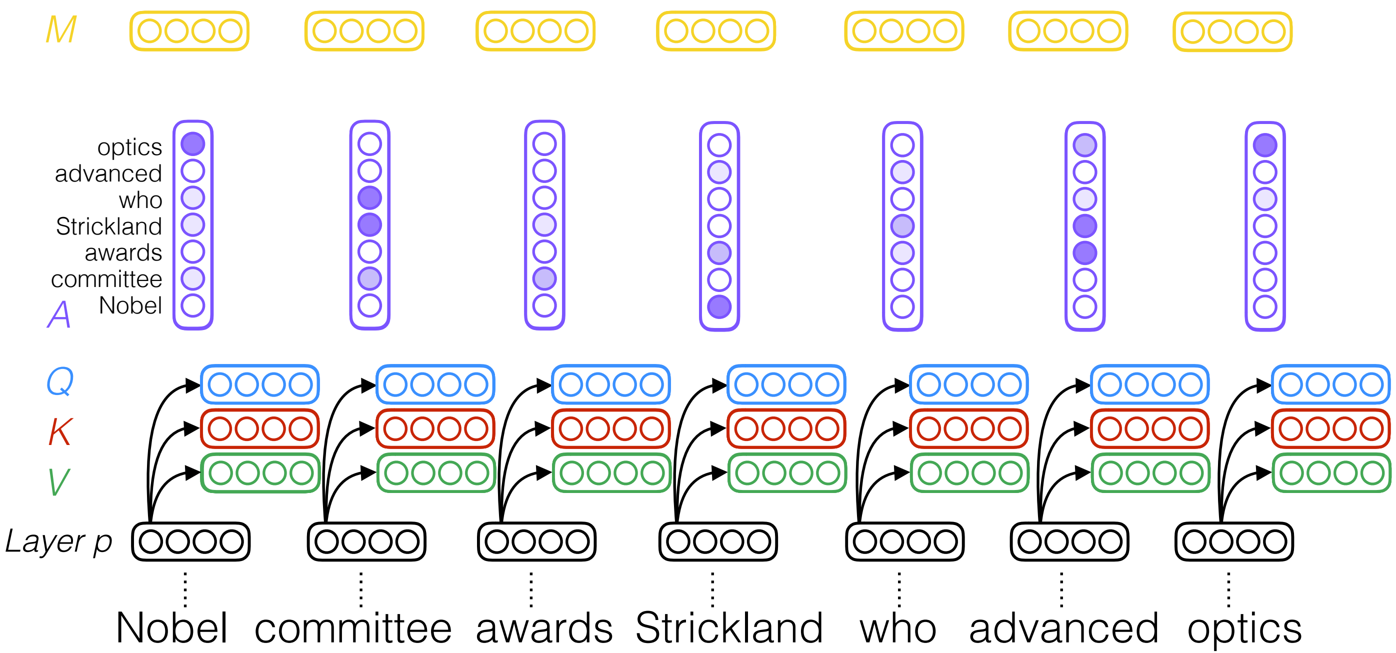
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



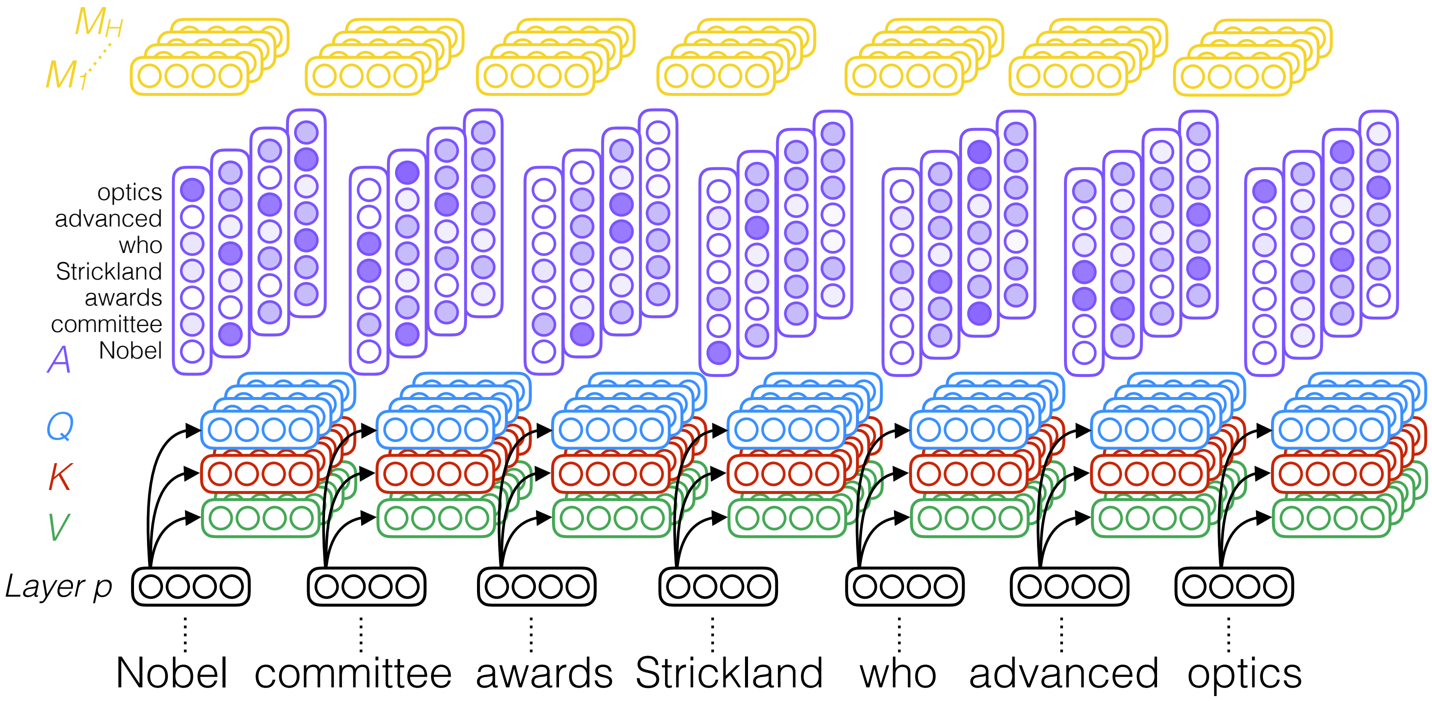
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



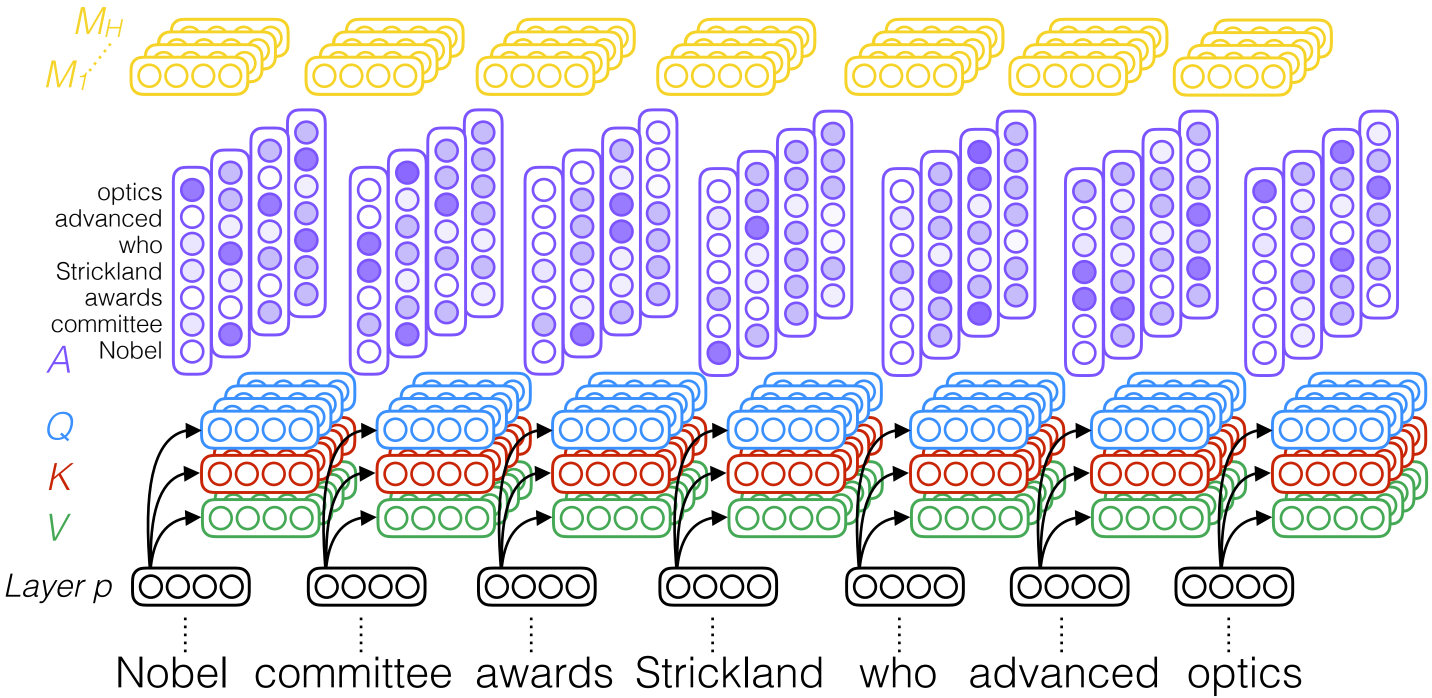
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]



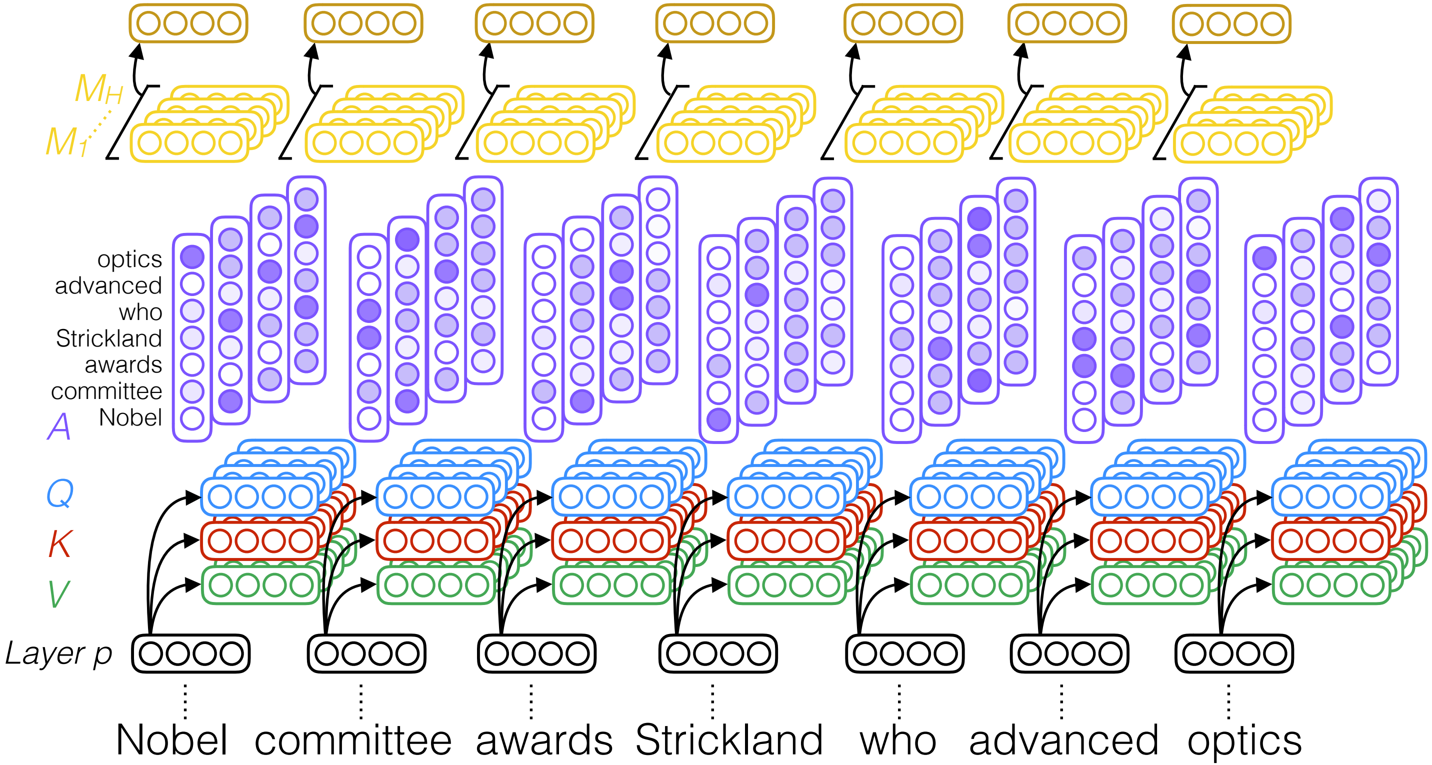
Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]

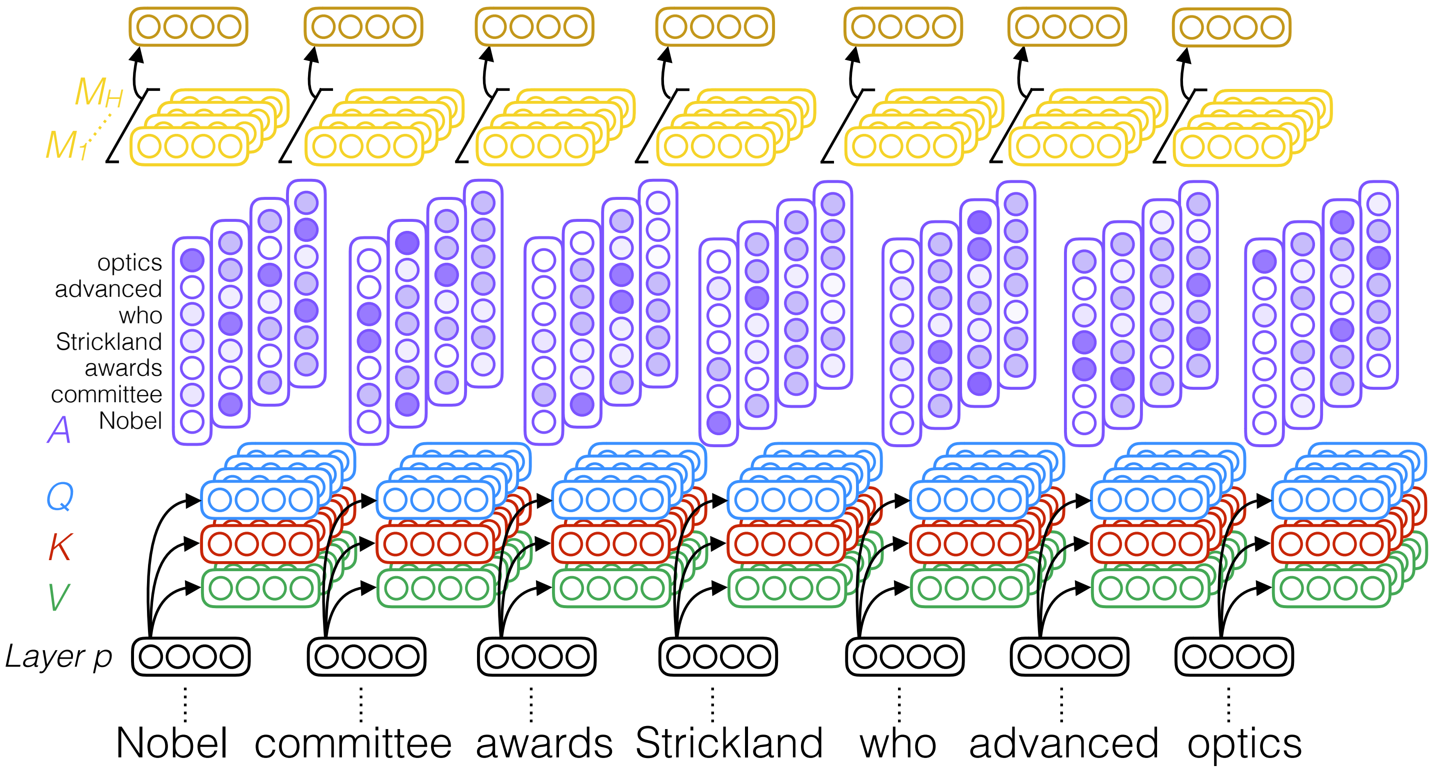


Transformer Model

[Slides by Emma Strubbell – EMNLP 2018]

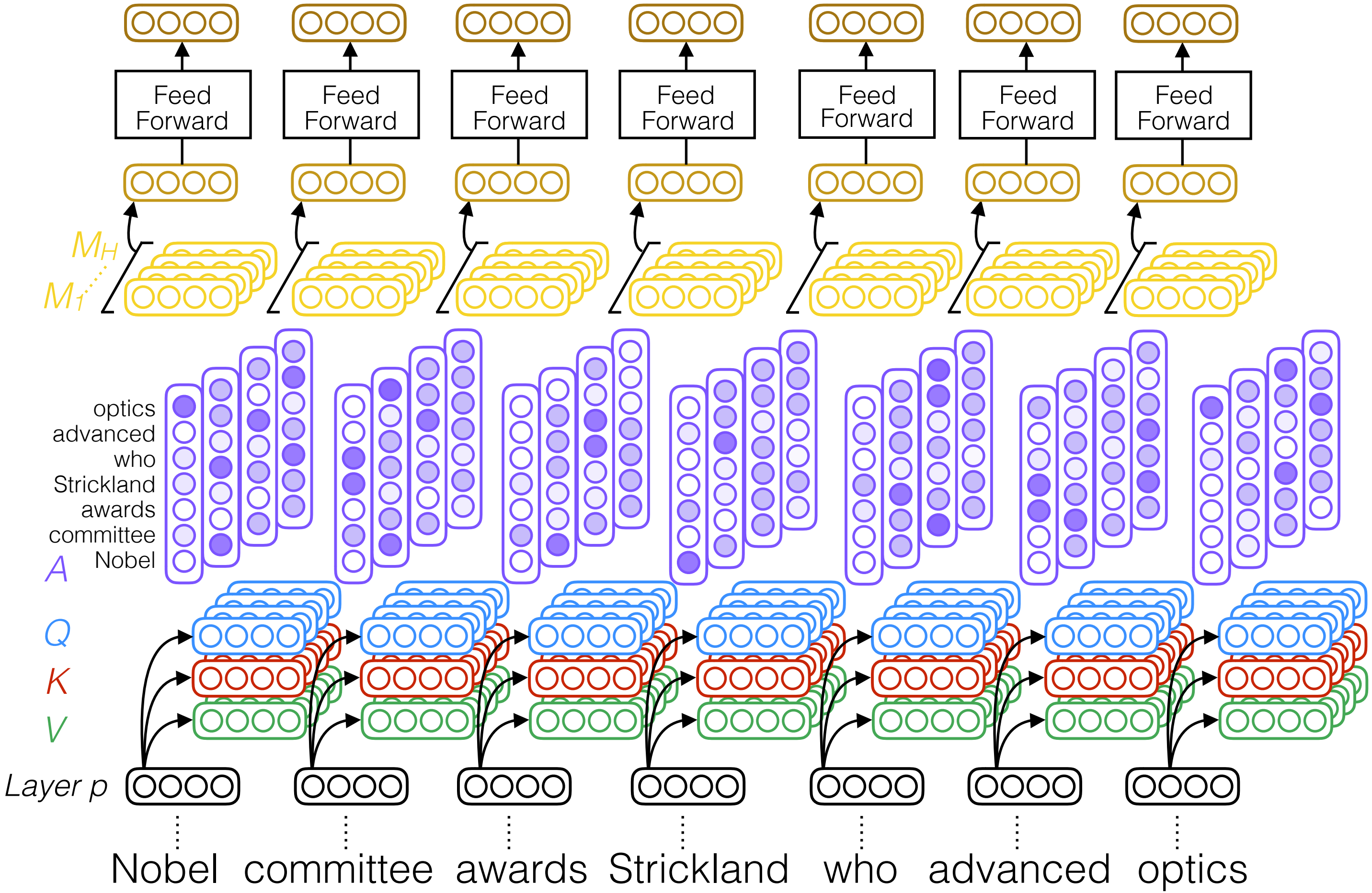


Transformer Model



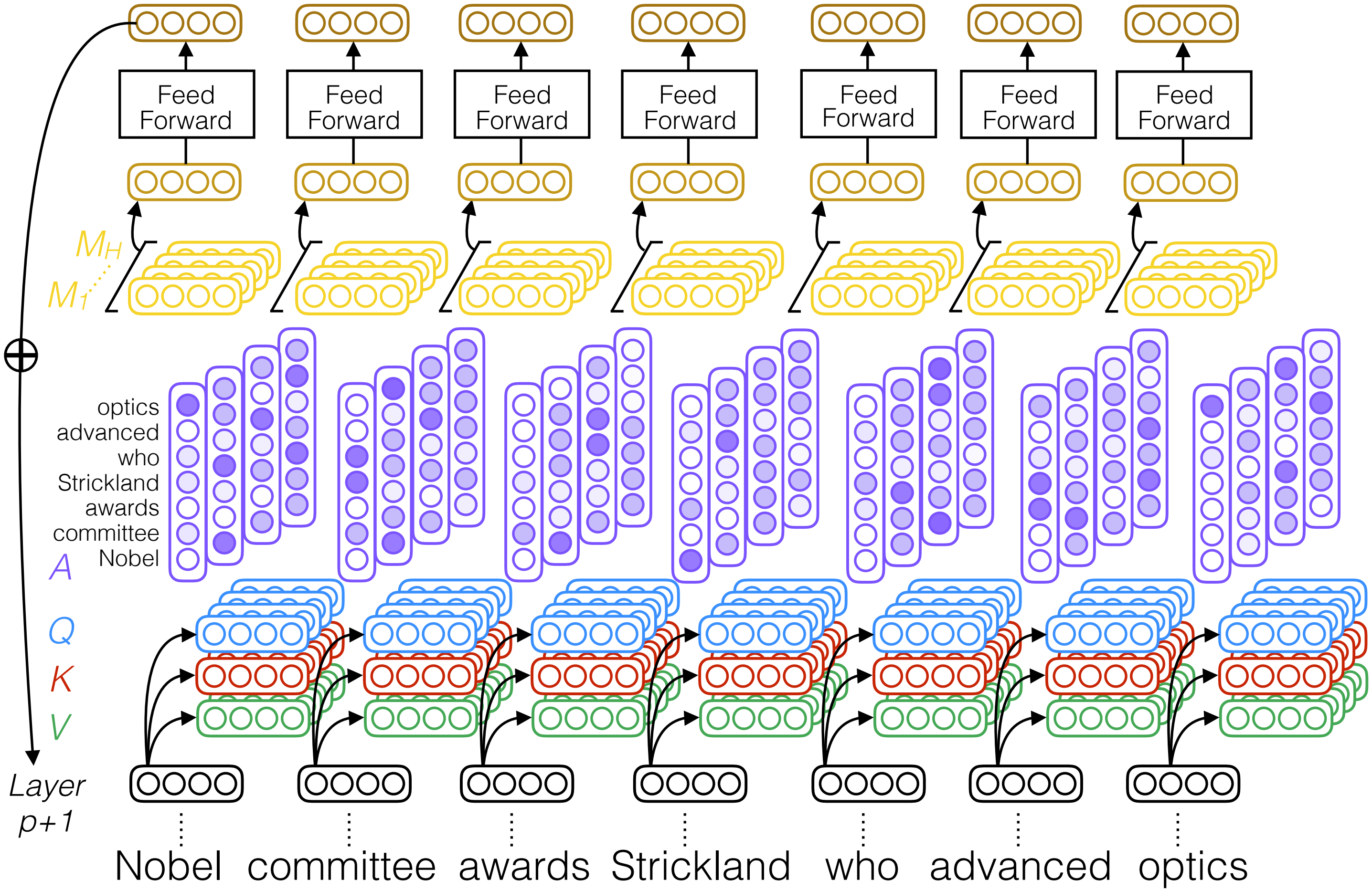
[Slides by Emma Strubbell – EMNLP 2018]

Transformer Model



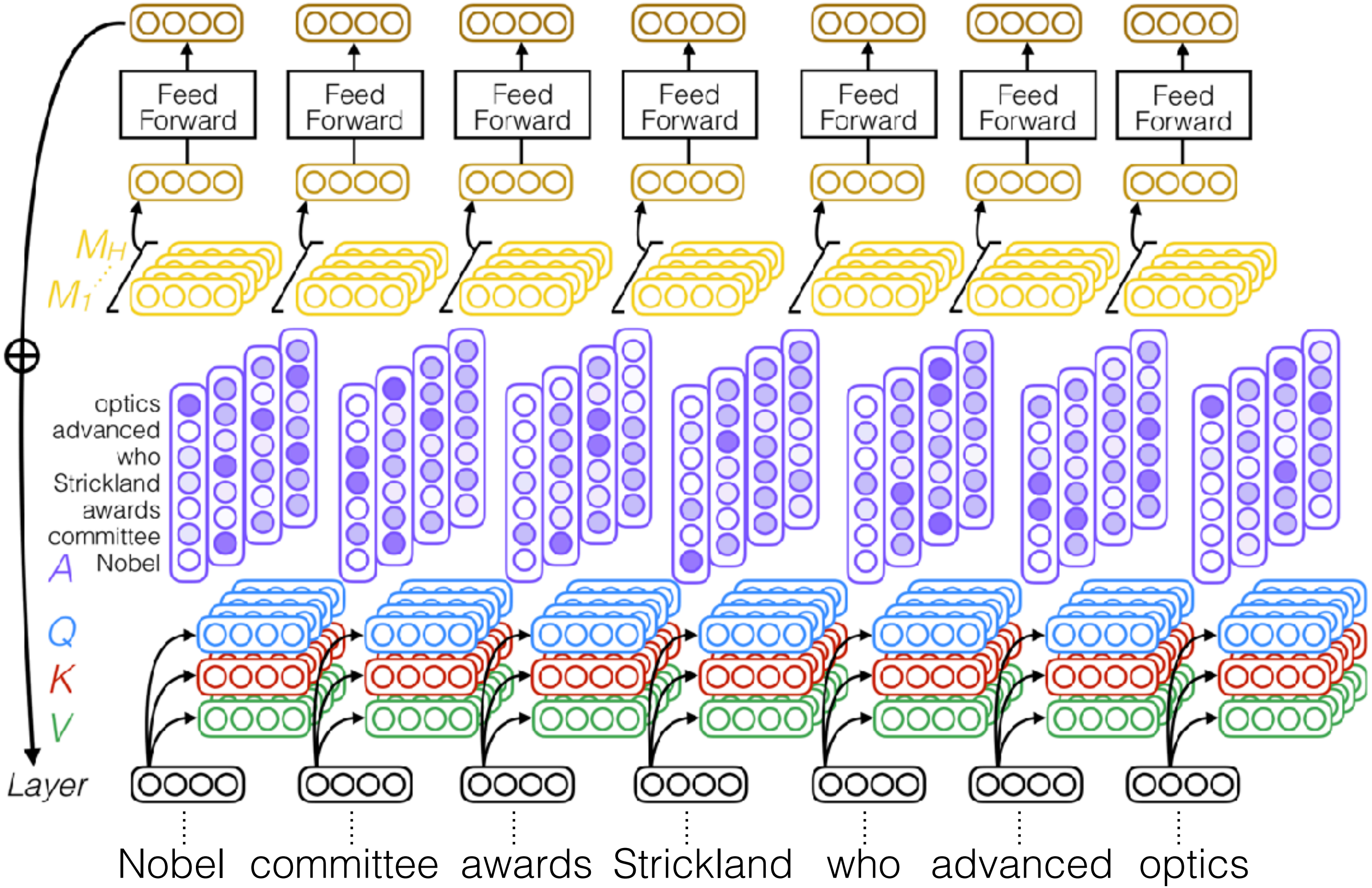
[Slides by Emma Strubbell – EMNLP 2018]

Transformer Model



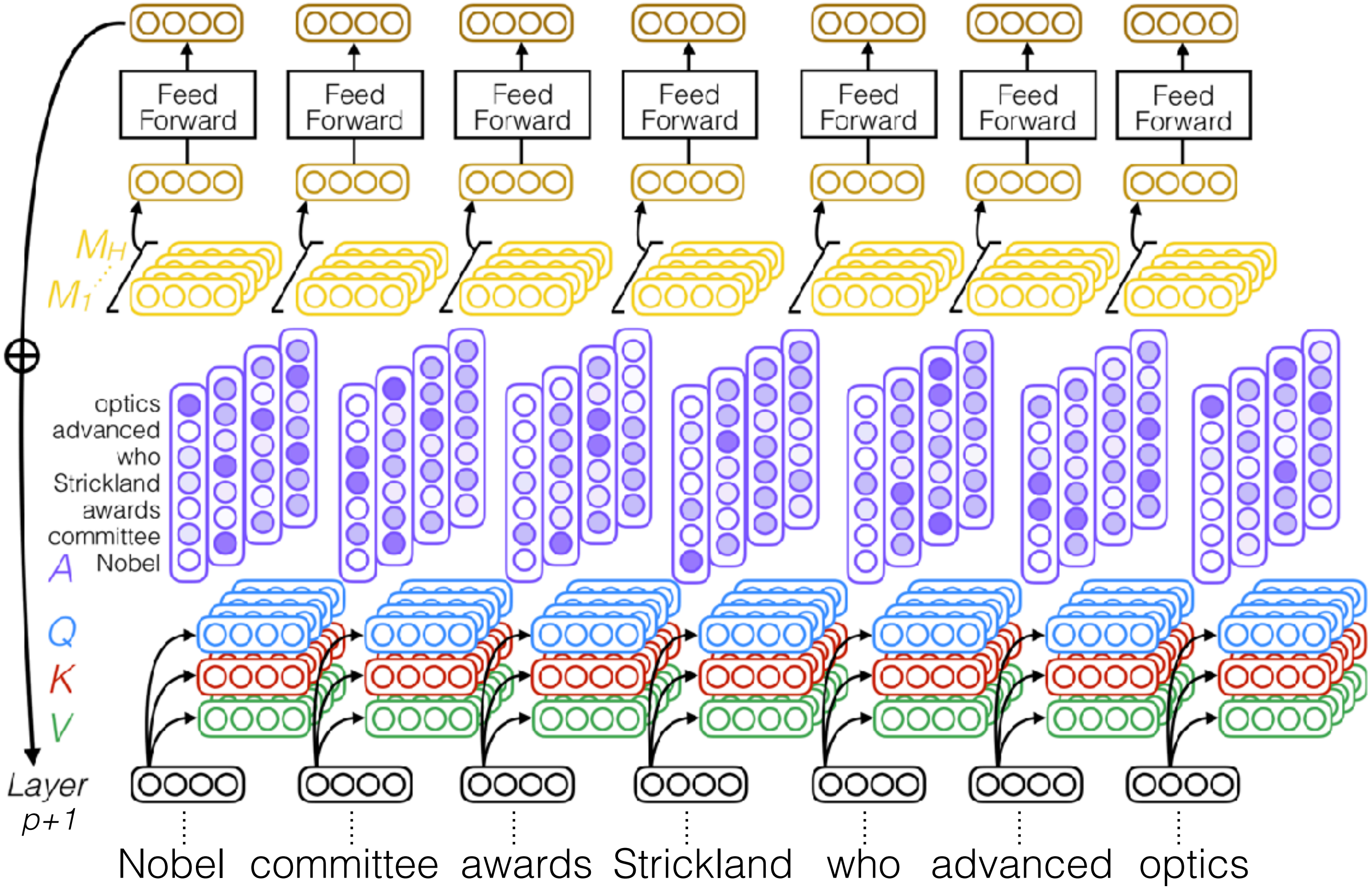
[Slides by Emma Strubbell – EMNLP 2018]

Transformer Model



[Slides by Emma Strubbell – EMNLP 2018]

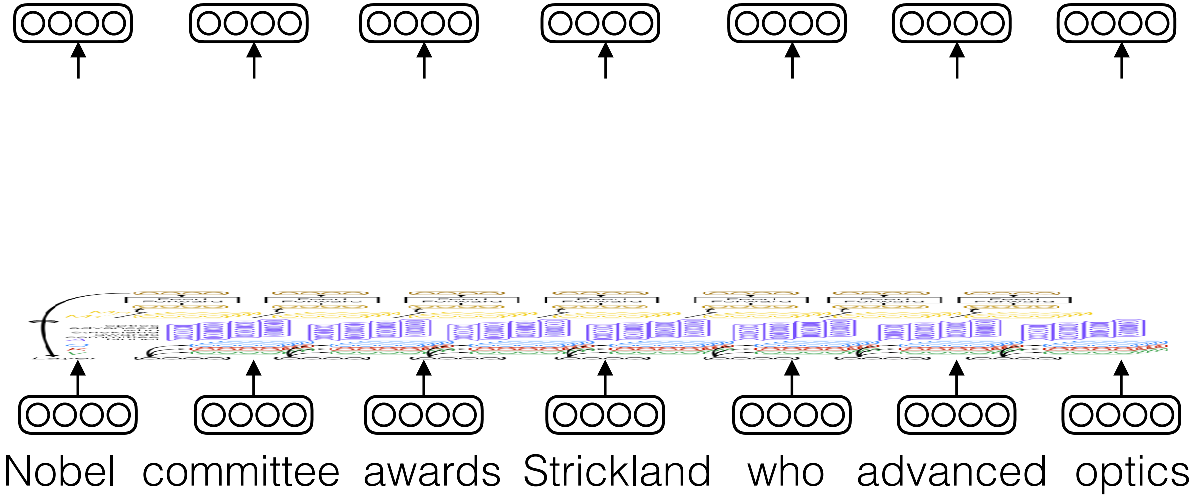
Transformer Model



[Slides by Emma Strubbell – EMNLP 2018]

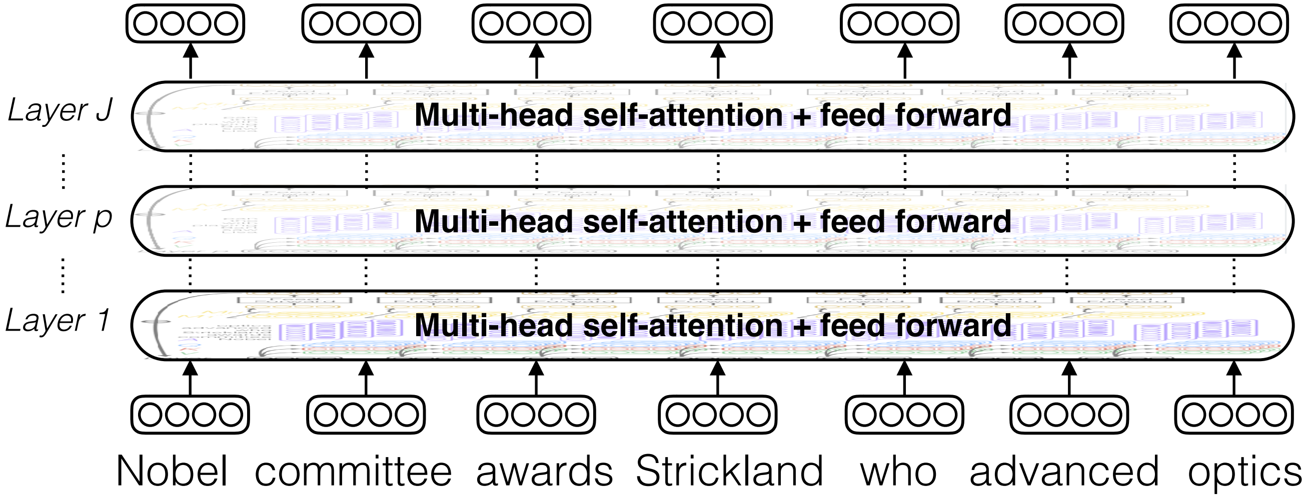
Transformer Model

[Adapted from slides by Emma Strubbell – EMNLP 2018]



Transformer Model

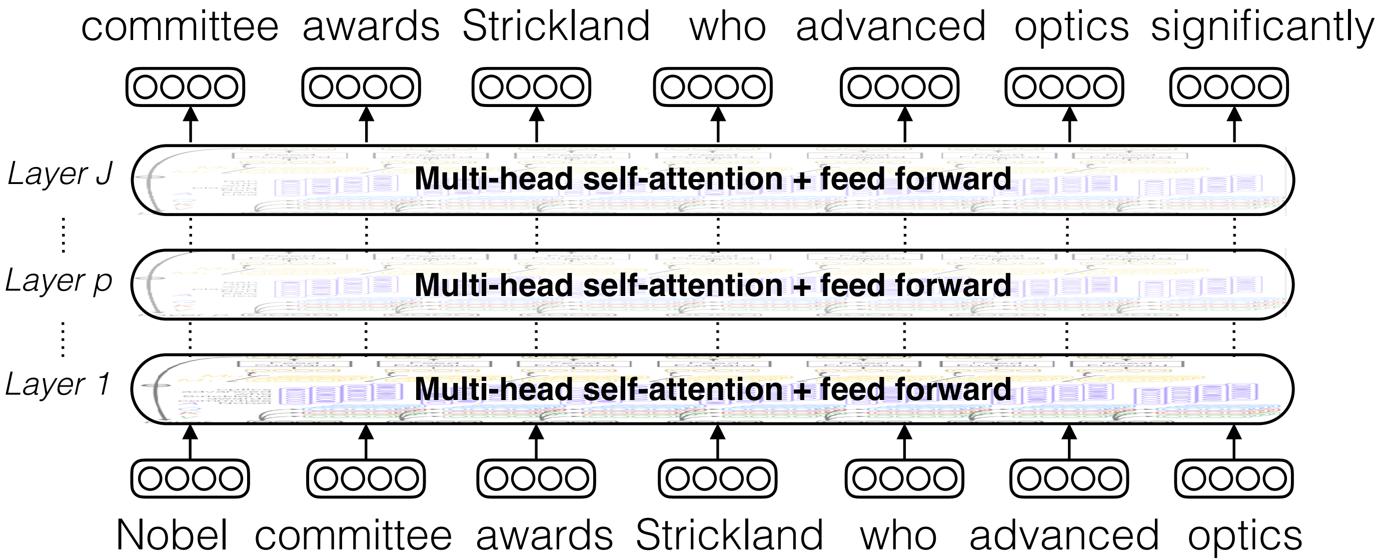
[Adapted from slides by Emma Strubbell – EMNLP 2018]



Transformer Model

[Adapted from slides by Emma Strubbell – EMNLP 2018]

The Transformer is trained to predict the next words given the history.

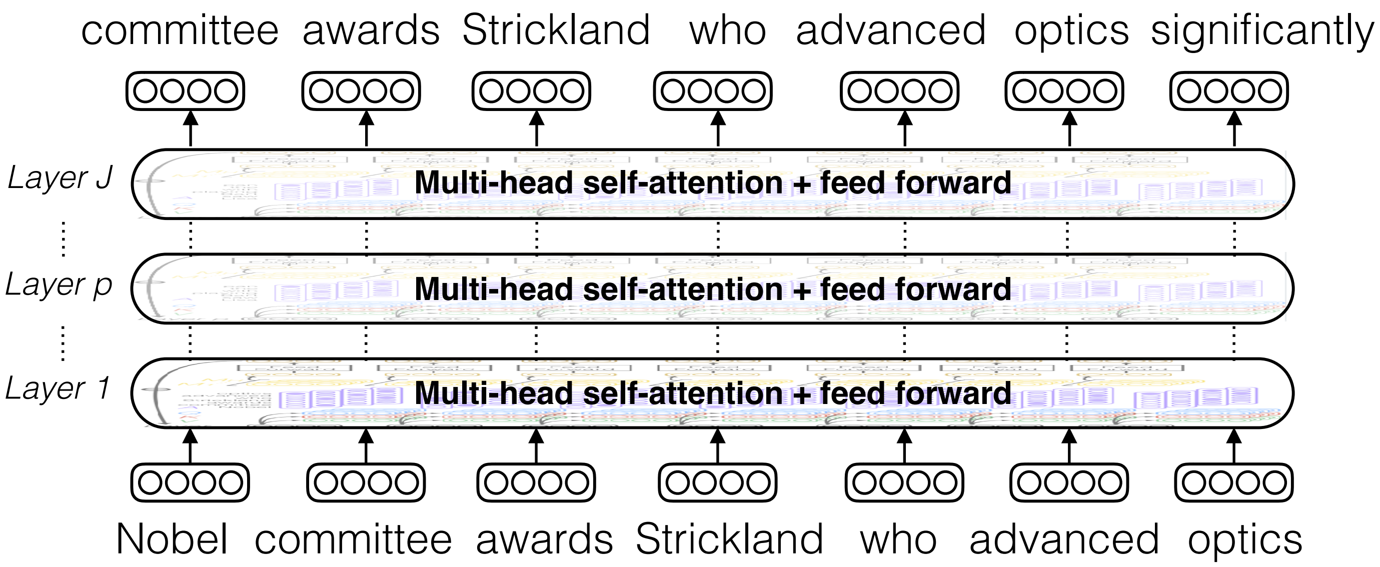


Transformer Model

[Adapted from slides by Emma Strubbell – EMNLP 2018]

The Transformer is trained to predict the next words given the history.

We use a mask so that each word is only « mixed" with the previous words (and not the following)



Transformer Model

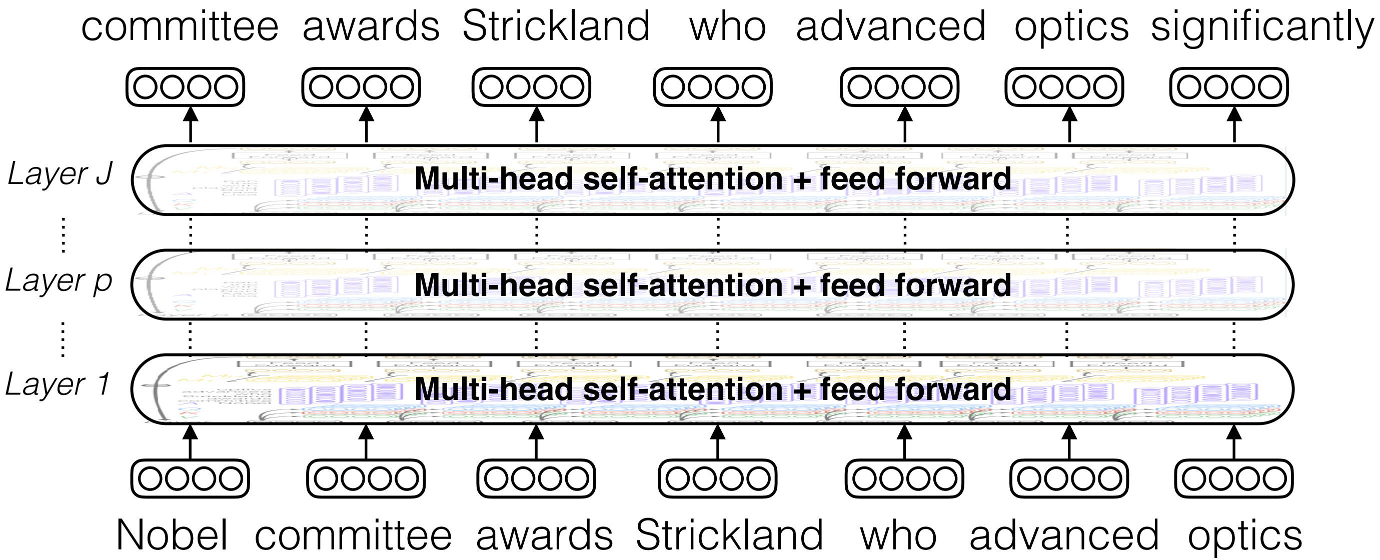
[Adapted from slides by Emma Strubbell – EMNLP 2018]

The Transformer is trained to predict the next words given the history.

We use a mask so that each word is only « mixed" with the previous words (and not the following)

This is called Language Modeling

(we learn a model of the probability of language)



Transformer Model

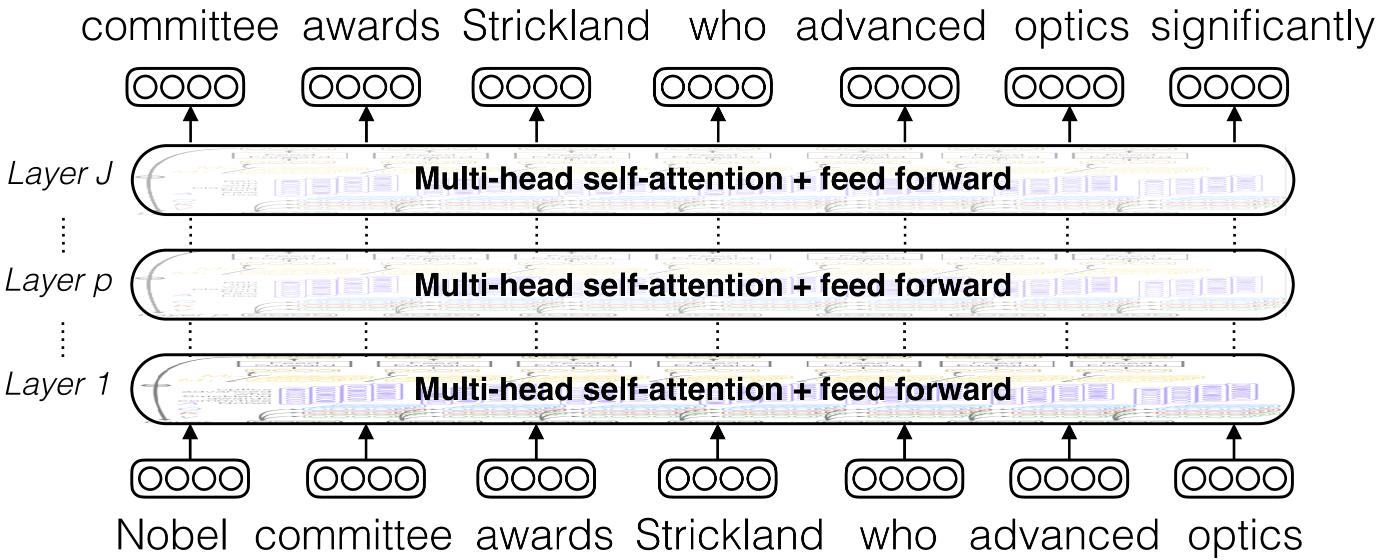
[Adapted from slides by Emma Strubbell – EMNLP 2018]

The Transformer is trained to predict the next words given the history.

We use a mask so that each word is only « mixed" with the previous words (and not the following)

This is called Language Modeling
(we learn a model of the probability of language)

$$p(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i | w_1, \ldots, w_{i-1})$$



Transformer Model

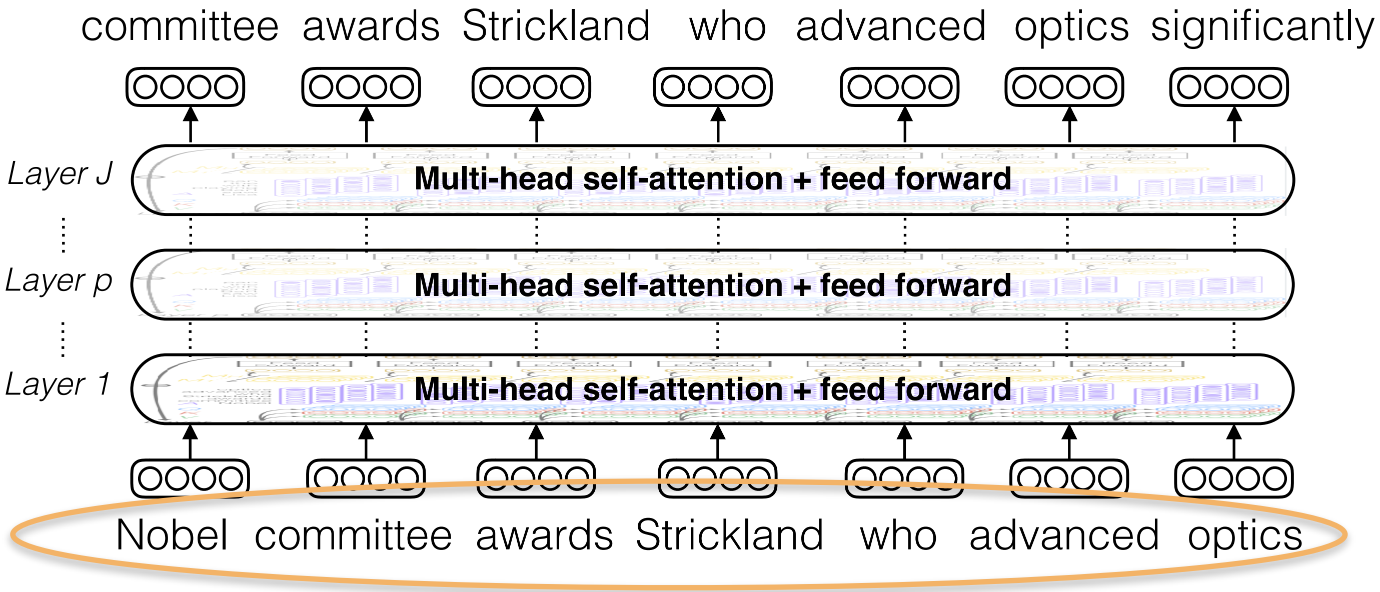
[Adapted from slides by Emma Strubbell – EMNLP 2018]

The Transformer is trained to predict the next words given the history.

We use a mask so that each word is only « mixed" with the previous words (and not the following)

This is called Language Modeling
(we learn a model of the probability of language)

$$p(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i | w_1, \ldots, w_{i-1})$$



Transformer Model

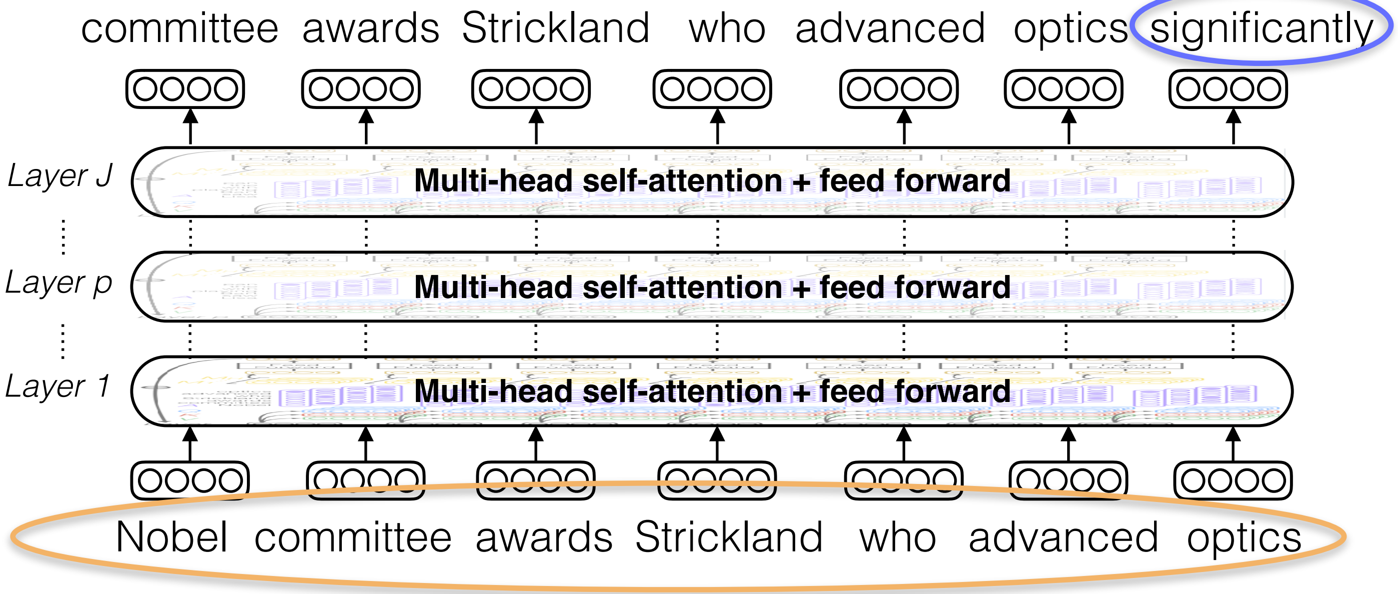
[Adapted from slides by Emma Strubbell – EMNLP 2018]

The Transformer is trained to predict the next words given the history.

We use a mask so that each word is only « mixed" with the previous words (and not the following)

This is called Language Modeling
(we learn a model of the probability of language)

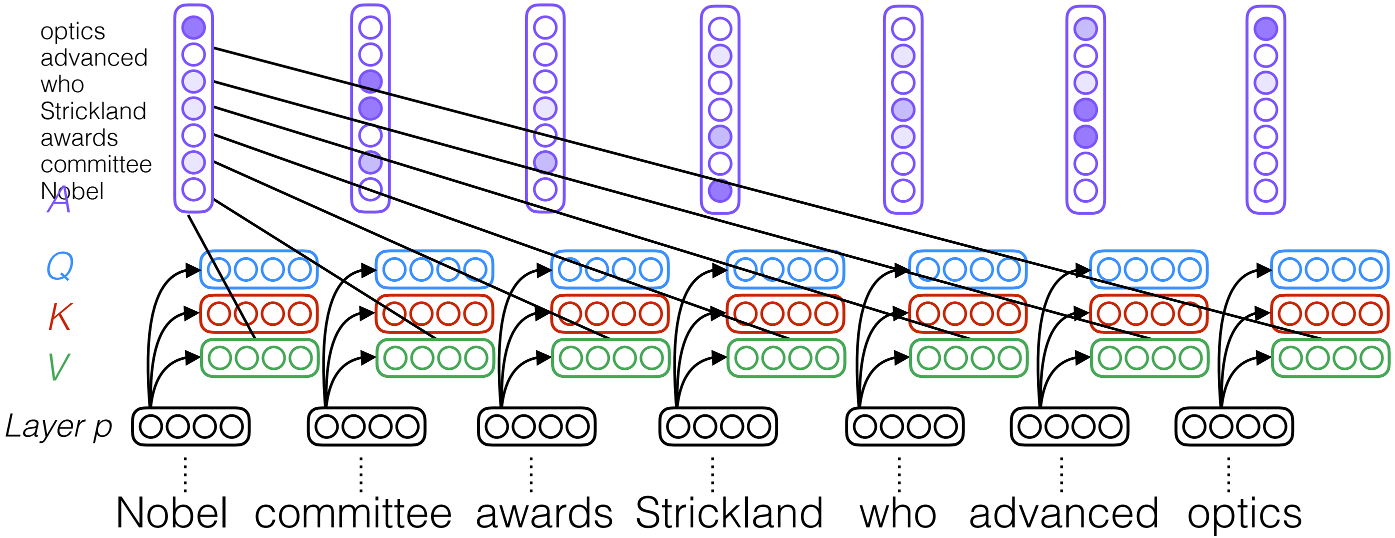
$$p(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i | w_1, \ldots, w_{i-1})$$



Encoding Position in Transformers

[Slides by Emma Strubbell – EMNLP 2018]

There is one <u>important</u> « caveat » with Transformers:

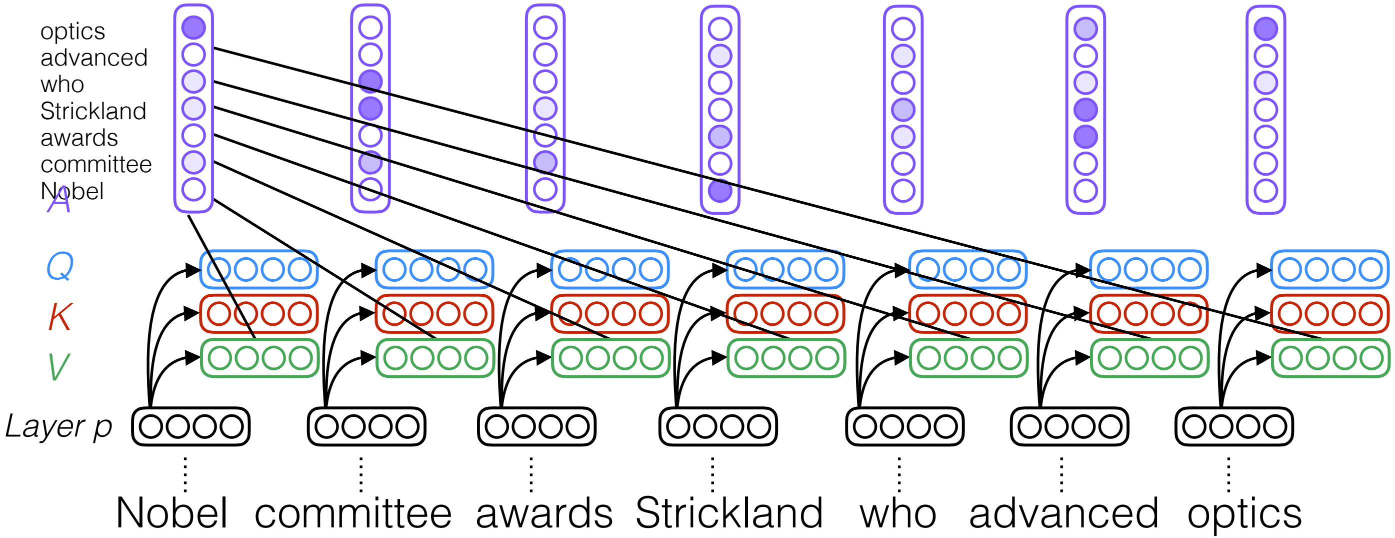


Encoding Position in Transformers

[Slides by Emma Strubbell – EMNLP 2018]

There is one important « caveat » with Transformers:

The operations they perform (weighted sum) are **invariant to word order**.
The model thus has no notion of word ordering.

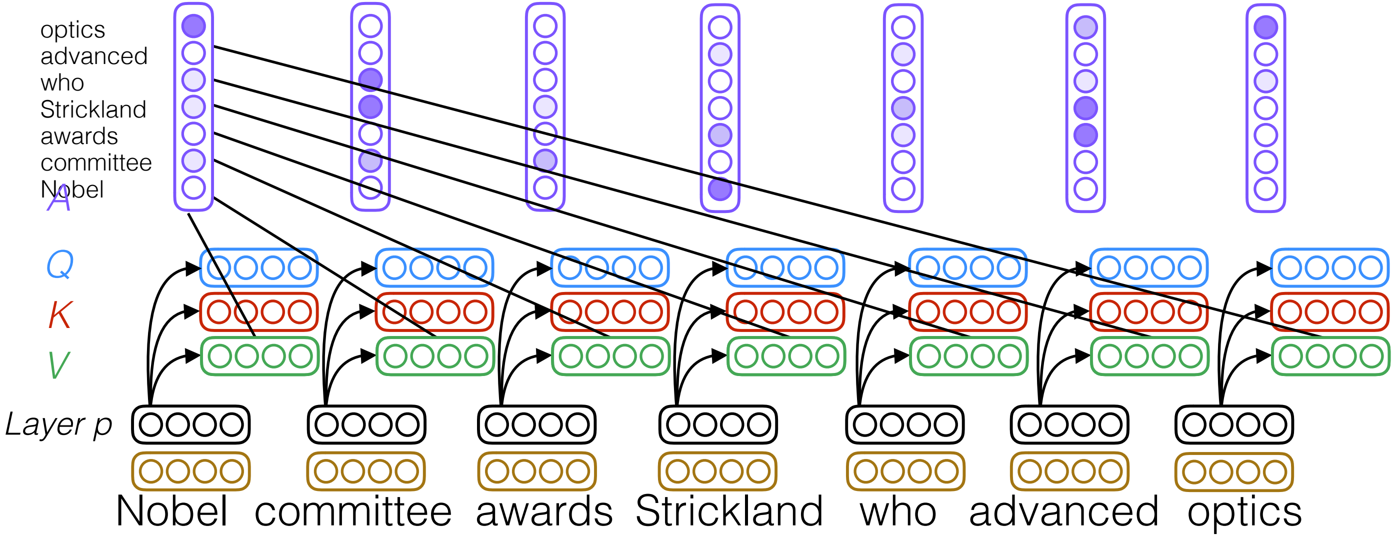


Encoding Position in Transformers

[Slides by Emma Strubbell – EMNLP 2018]

There is one important « caveat » with Transformers:

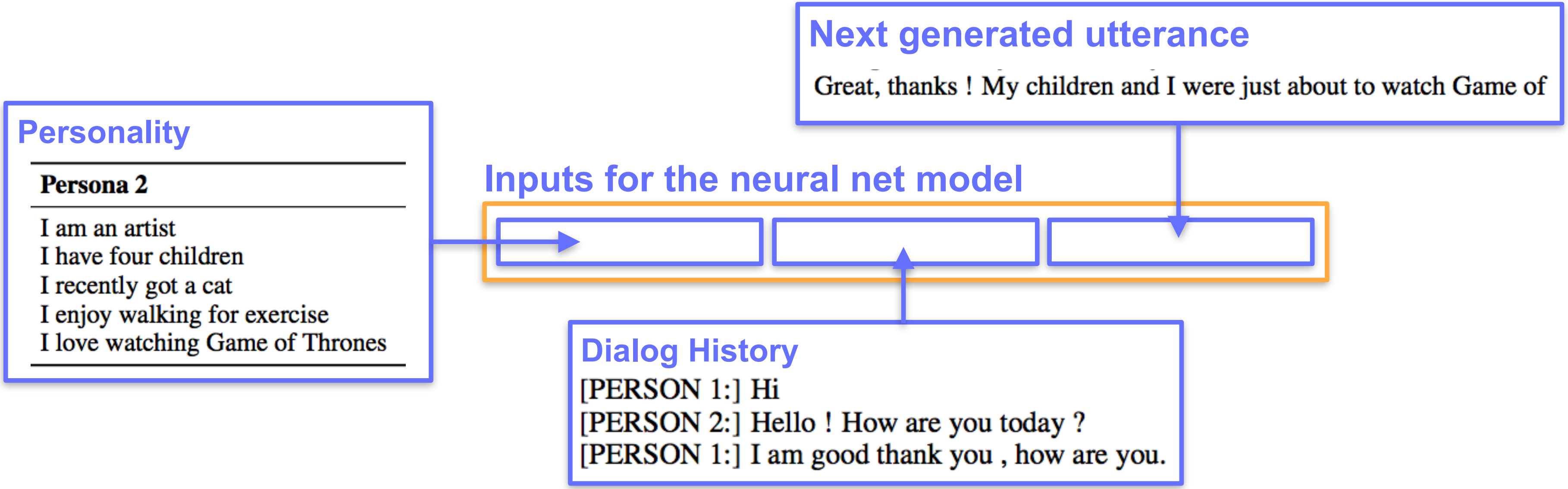
The operations they perform (weighted sum) are **invariant to word order**.
The model thus has no notion of word ordering.
To solve that we provide **position embeddings** that indicate the **position** of each token in the sentence



Now let’s train our Model

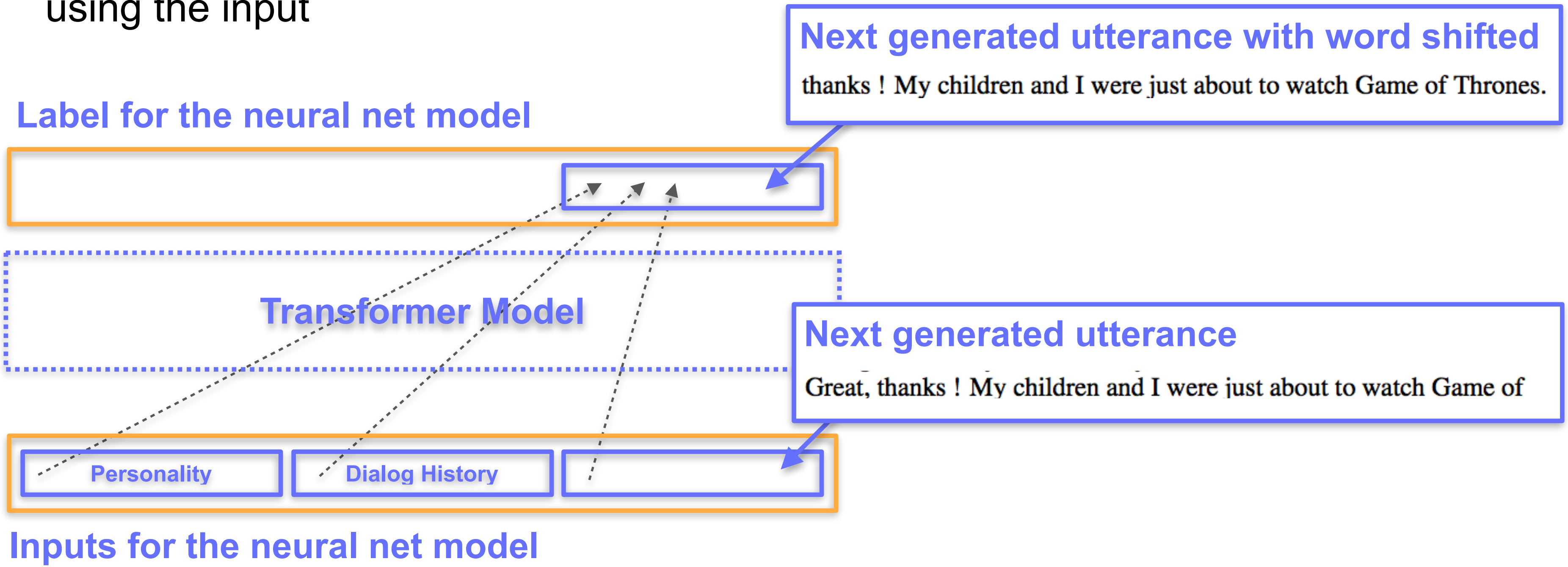
Training with Language Modeling

- For each utterance in our dialog, we create a **sequential input** by concatenating:
  - The personality sentences juxtaposed one to the other,
  - The history of the dialogue up to the current utterance,
  - The current utterance.



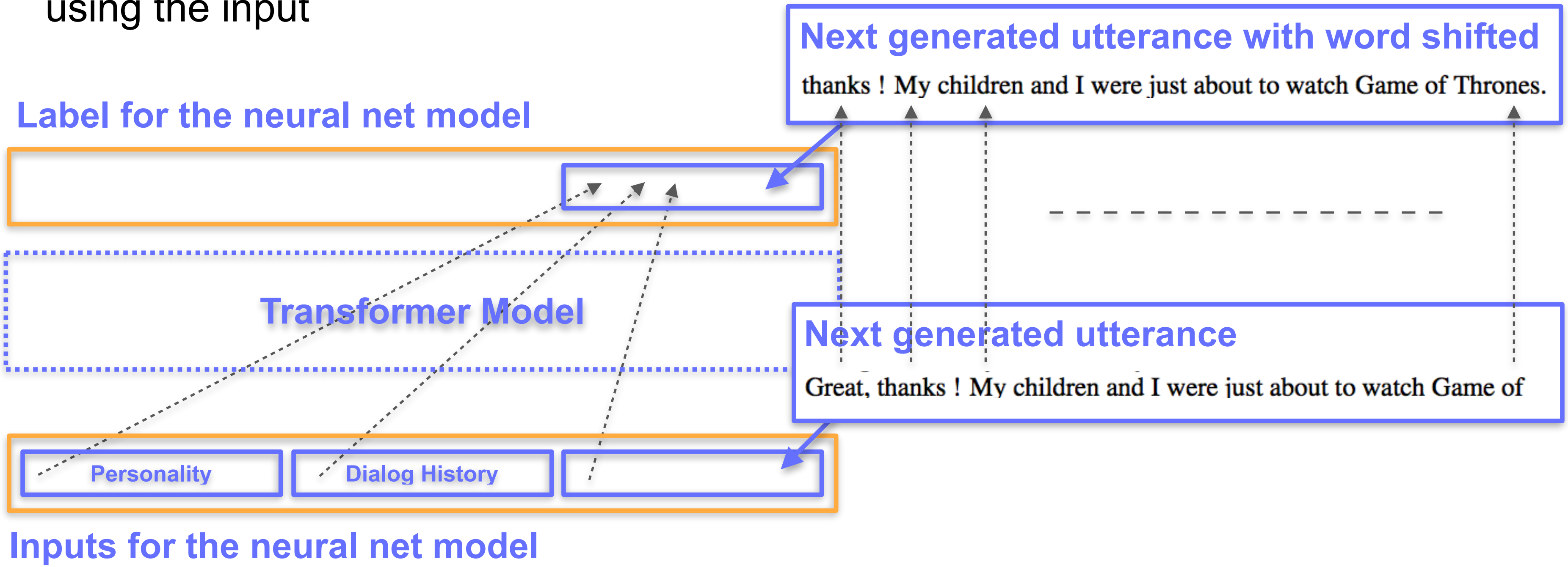
Training with Language Modeling

- For each utterance in our dialog, we create a **target (or label)** which is:
  - The current utterance with all the words shifted to the left
- The model is trained to generate the labels (ie. all the next words in parallel) using the input



Training with Language Modeling

- For each utterance in our dialog, we create a **target (or label)** which is:
  - The current utterance with all the words shifted to the left
- The model is trained to generate the labels (ie. all the next words in parallel) using the input



Limitations of the dataset

- PERSONA-CHAT is **one of the biggest** multi-turn dialog dataset :
  - 164,356 utterances and about 1-2M words
  - Average number of turns: 14

Limitations of the dataset

- PERSONA-CHAT is **one of the biggest** multi-turn dialog dataset :
  - 164,356 utterances and about 1-2M words
  - Average number of turns: 14
- But it is still **small** for training a deep learning model:
  - 1B words in the Billion Words dataset
  - ~1M sentences in CoNLL 2012 (used for training co-reference systems)

Limitations of the dataset

- PERSONA-CHAT is **one of the biggest** multi-turn dialog dataset :
  - 164,356 utterances and about 1-2M words
  - Average number of turns: 14

- But it is still **small** for training a deep learning model:
  - 1B words in the Billion Words dataset
  - ~1M sentences in CoNLL 2012 (used for training co-reference systems)

- And generating an engaging open-domain dialogue requires:
  - topic-coherence,
  - dialogue-flow,
  - common-sense,
  - short term memory,
  - co-reference resolution,
  - sentimental analysis,
  - textual entailment…

Why this is a problem? 👻

The Conversational Intelligence Challenge 2 « ConvAI2 » (NeurIPS 2018 competition)

Model	Creator	PPL	Hits@1	F1
	🤗 (Hugging Face)	20.47🍎	74.7🍎	17.52🍎
	High Five	-	65.9	-
	Little Baby	-	63.4	-
	Happy Minions	32.94	52.1	14.76
	Cats'team	-	35.9	-
	loopAI	-	25.6	-
	Mohd Shadab Alam	29.94	13.8	16.91
	1st-contact	31.98	13.2	16.42
	Tensorborne	38.24	12.0	15.94
	Team Dialog 6	40.35	10.9	7.27
	NEUROBOTICS	35.47	-	16.68
	Sonic	33.46	-	16.67
topicSeq2seq	Team Pat	-	-	16.11
	Roboy	-	-	15.83
	Lost in Conversation	55.84	-	15.74
	flooders	-	-	15.47
	IamNotAdele	66.47	-	13.09
	Salty Fish	38.86	-	-
	Pinta	37.85	-	-
Seq2Seq + Attention	ParlAI team	29.8	12.6	16.18
Language Model	ParlAI team	46.0	-	15.02
KV Profile Memory	ParlAI team	-	55.2	11.9

First submission results

(The 2nd had a +8 points improvement in Hits@1)

Validation set (public) Leaderboard – Test set (hidden) Leaderboard

Model	Creator	PPL	Hits@1	F1
	🤗 (Hugging Face)	23.05🍎	74.3🍎	17.85🍎
	Team Pat	-	-	17.85
	Pinta	-	51.4	17.25
	Mohd Shadab Alam	35.57	14.8	16.94
	Sonic	38.87	-	16.88
	NEUROBOTICS	39.7	-	16.82
	Happy Minions	34.57	68.1	16.72
	1st-contact	36.54	13.3	16.58
	Tensorborne	44.64	12.1	16.13
	flooders	-	-	15.96
	Lost in Conversation	62.83	-	15.91
	High Five	59.83	78.2	15.34
	Little Baby	-	72.9	-
	loopAI	-	29.7	-
	Salty Fish	42.3	-	-

Model	Creator	PPL	Hits@1	F1
	🤗 (Hugging Face)	20.47🍎	74.7🍎	17.52🍎
	Little Baby	-	61.0	-
	Happy Minions	32.94	52.1	14.76
	High Five	52.8	50.3	13.73
	Pinta	-	44.4	16.52
	loopAI	-	25.6	-
	Mohd Shadab Alam	30.97	14.4	16.44
	1st-contact	31.98	13.2	16.42
	Tensorborne	38.24	12.0	15.94
	Team Dialog 6	40.35	10.9	7.27
	NEUROBOTICS	35.47	-	16.68
	Sonic	33.46	-	16.67
	Lost in Conversation	55.84	-	15.74
	flooders	-	-	15.47
	Team Pat	-	-	13.23
	Salty Fish	45.87	-	-
Seq2Seq + Attention	ParlAI team	29.8	12.6	16.18
Language Model	ParlAI team	46.0	-	15.02
KV Profile Memory	ParlAI team	-	55.2	11.9

- Small dataset =>
  - Large models are overfitting
  - Small models are underfitting

What can we do? 👻

Transfer Learning 🦄

(Sequential) Transfer Learning

A two-stage procedure

1. *Pre-train* the model on a **large** dataset:
    - which is **not** the dataset you will use in the end,
    - but on which you hope to **learn general concepts** that will help in your case
2. *Fine-tune* the model on your **small** dataset:
    - to make it perform **well on your task**.

Pre-training

1. We pre-trained our model on
    - a **large dataset** of **contiguous** span of texts (Toronto Book Corpus: **~7000 books**)
    - with a *Language Modeling* objective (as we've just seen).

- Learns initial parameters of the neural network model.
- Provide the model with
    - some **kind of world knowledge** and
    - an ability to **build coherent sentences** by processing long-range dependencies.

- In our experiments, we started from the pre-trained model of Radford et al. 2018.

A Simple Method for Commonsense Reasoning by Trinh & Le (2018), *Improving, Language Understanding by Generative Pre-Training* by Radford et al. (2018), *Universal Language Model Fine-tuning for Text Classification* by Howard and Ruder (2018), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* by Jacob Devlin et al (2018)

Fine-tuning

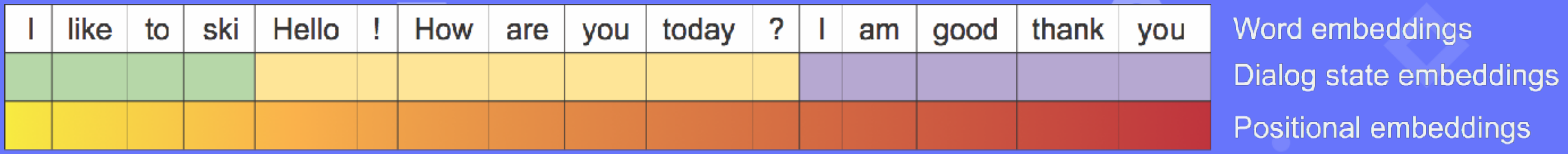
2. We fine-tune the model the dataset of dialog (PERSONA-CHAT) by
    - **adapting the inputs of the model** to the dialog setting
    - using a **multi-task fine-tuning scheme** in which the model is trained jointly on several objectives:
        - *Language Modeling*: to adapt to the vocabulary used in the dialog dataset
        - *Next Sentence Prediction*: to learn how to hold a conversation

Let’s quickly see these two operations

Encoding a Dialog and a Persona

- **After pre-training** we have a model with basic common-sense and co-reference capabilities, now we need to teach it the specificities of dialog:
  - Alternating utterances
  - Dialog flow (« speech/dialog acts »)
  - Conditioning on a personality

- How to build a sequential inputs for our model from a conditioned dialog?
  - Transformers don’t possess a natural notion of sequentiality and position
  - We already have positional embeddings to incorporate sequentiality
  - We add special embeddings related to utterances and personas



Encoding a Dialog and a Persona

- We can play with these embeddings to manipulate the notion of a sequence

Repeating specific embeddings to control positioning information

I	like	to	ski	I	hate	mexican	food	I	like	to	eat	cheetos

- We can also augment the dataset to bias towards positional invariance

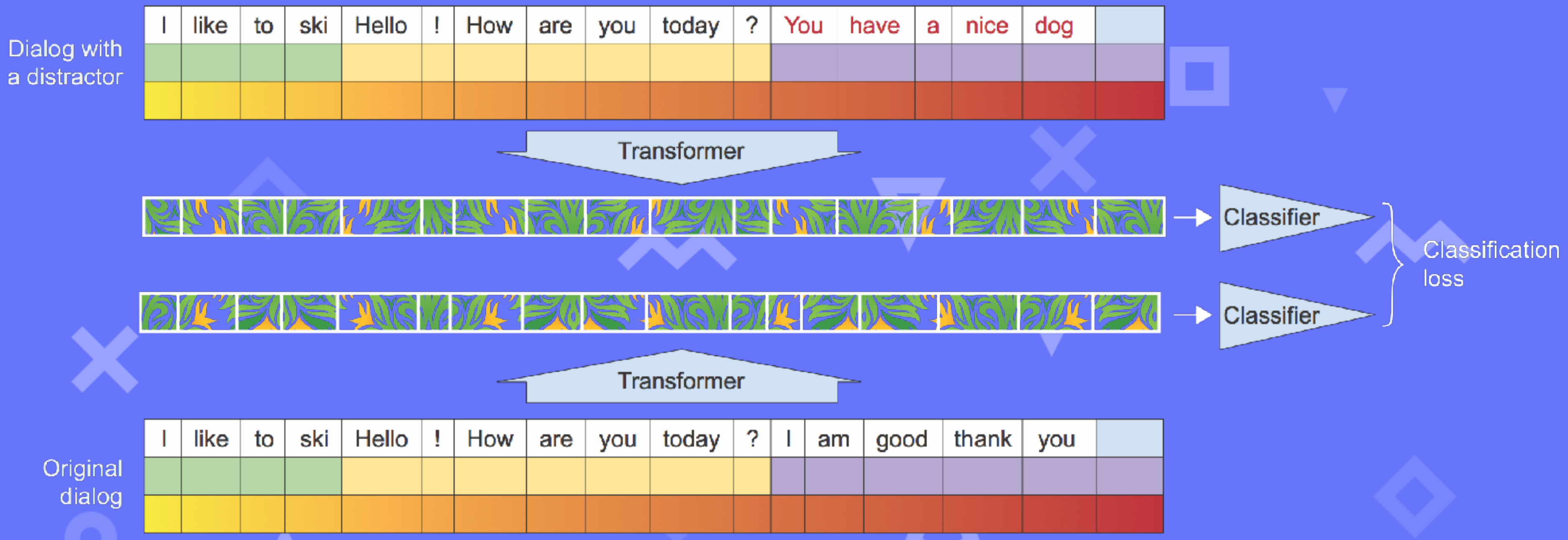
I	hate	mexican	food	I	like	to	eat	cheetos	I	like	to	ski

I	like	to	ski	I	hate	mexican	food	I	like	to	eat	cheetos

Permutation augmented dataset to bias towards positional invariance

Semantic Learning on Dialog Utterances

- Learning to distinguish a real answer from a distractor.



Combined with language modeling fine-tuning in a multi-task fashion

Very strong Results on the Automatic Metrics

Validation set (public) Leaderboard – <u>Test set (hidden) Leaderboard</u>

Model	Creator	PPL	Hits@1	F1
	🤗 (Hugging Face)	17.51🍎	82.1	19.09
	Happy Minions	29.85	-	17.79
	ADAPT Centre	22.57	-	20.3🍎
	Lost in Conversation	-	17.3	17.79
	Khai Mai Alt	-	85.3🍎	17.64
	Pinta	23.86	-	17.27
	Mohd Shadab Alam	34.12	13.4	17.08
	Sonic	38.87	-	16.88
	NEUROBOTICS	39.7	-	16.82
	1st-contact	36.54	13.3	16.58
topicSeq2seq	Team Pat	-	-	16.58
	Roboy	-	-	16.25
	Tensorborne	44.64	12.1	16.13
	flooders	-	-	15.96
	Clova Xiaodong Gu	-	-	15.39
	IamNotAdele	53.46	-	12.85
	Little Baby(AI小奶娃)	-	83.0	-
	High Five	-	79.1	-
	Sweet Fish	-	75.6	-
	Cats'team	-	43.4	-
	loopAI	-	29.7	-
	Salty Fish	33.46	-	-

Rank	Creator	PPL	Hits@1	F1
1 🍐	🤗 (Hugging Face)	16.28🍎	80.7🍎	19.5🍎
2 🍐	ADAPT Centre	31.4	-	18.39
3 🍐	Happy Minions	29.01	-	16.01
4 🍐	High Five	-	65.9	-
5 🍐	Mohd Shadab Alam	29.94	13.8	16.91
6 🍐	Lost in Conversation	-	17.1	17.77
7 🍐	Little Baby(AI小奶娃)	-	64.8	-
8	Sweet Fish	-	45.7	-
9	1st-contact	31.98	13.2	16.42
10	NEUROBOTICS	35.47	-	16.68
11	Cats'team	-	35.9	-
12	Sonic	33.46	-	16.67
13	Pinta	32.49	-	16.39
14	Khai Mai Alt	-	34.6	13.03
15	loopAI	-	25.6	-
16	Salty Fish	34.32	-	-
17	Team Pat	-	-	16.11
18	Tensorborne	38.24	12.0	15.94
19	Team Dialog 6	40.35	10.9	7.27
20	Roboy	-	-	15.83
21	IamNotAdele	66.47	-	13.09

Now what do humans think about this model?

👷 🕵️ 👩‍⚕️ 👨‍🎤 👩‍🎓 👩‍🌾 👨‍🚀

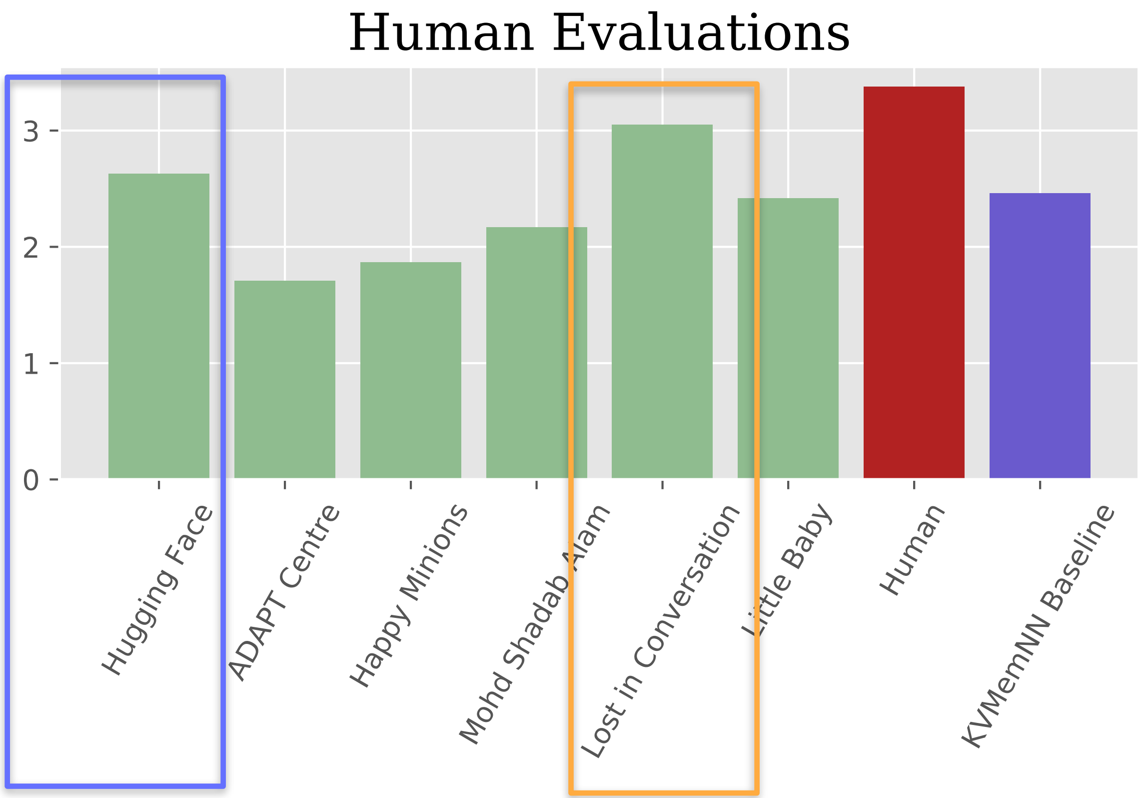
Human Evaluation

Using Amazon Mechanical Turk

- 100 evaluations per model
- Mechanical Turk worker and model were each assigned a persona and chat for 4-6 dialog turns each
- After the chat, the worker is asked:
- How much did you enjoy talking to this user?
  - Choices: not at all, a little, somewhat, a lot => 1, 2, 3, 4
- Next, the worker is shown the model's persona + a random persona, and asked:
- Which prompt (character) do you think the other user was given for this conversation?

[Adapted from slides by Emily Dinan / Jason Weston – NeurIPS 2018]

We were good… but not the best



[Adapted from slides by Emily Dinan / Jason Weston – NeurIPS 2018]

How does the conversations look

Hello there, I like rock climbing do you have any hobbies?

i like nature. do you? i like nature.

I do love nature, especially when going on walks

what is your favorite rock? mine is rock.

Well I like rock music, does that count?

i think it does. what do you do for fun?

Well I work at a restaurant that I love, and rock climbing

that sounds fun! i'm into nature.

Cool, so am I

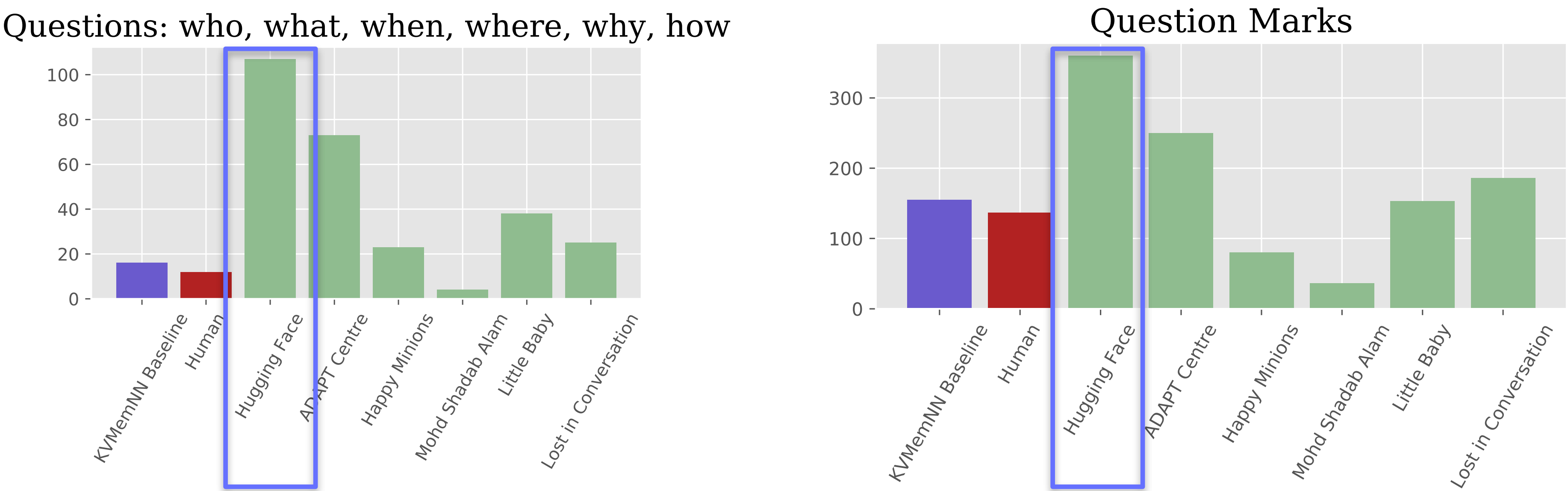
do you have any pets? do you have pets?

BOT IN BLUE

User feedbacks

[Adapted from slides by Emily Dinan / Jason Weston – NeurIPS 2018]

How does the conversations look



asks too many questions!

[Adapted from slides by Emily Dinan / Jason Weston – NeurIPS 2018]

Evaluating a Natural Language Generation System

An Open Research Question

- **Automatic metrics** don't correlate well with **human evaluations**
- We (together with Microsoft, University of Washington, Stanford and Facebook) are organizing a workshop on this topic this summer in Minneapolis:

NeuralGen 2019: Methods for Optimizing and Evaluating Neural Language Generation

That's it for today
Thanks for listening!

thomas@huggingface.co