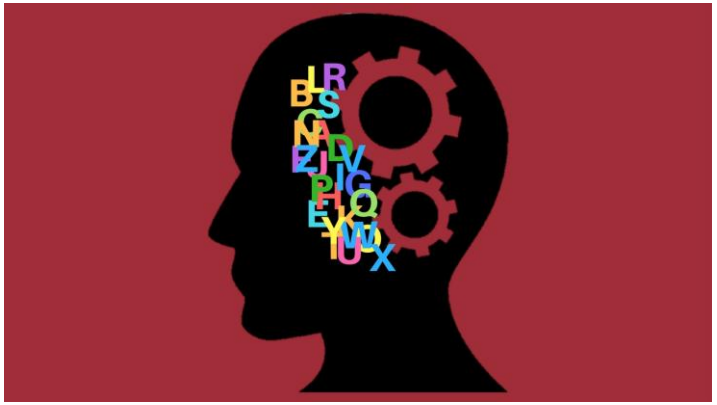
University of Ljubljana, Faculty of Computer and Information Science

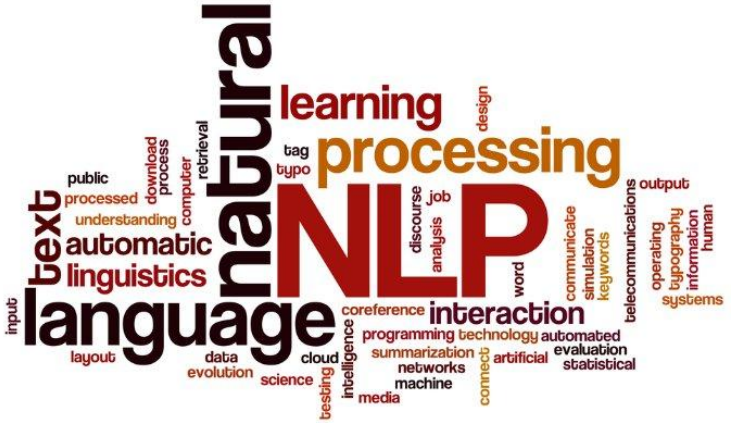
Natural language processing



Prof Dr Marko Robnik-Šikonja

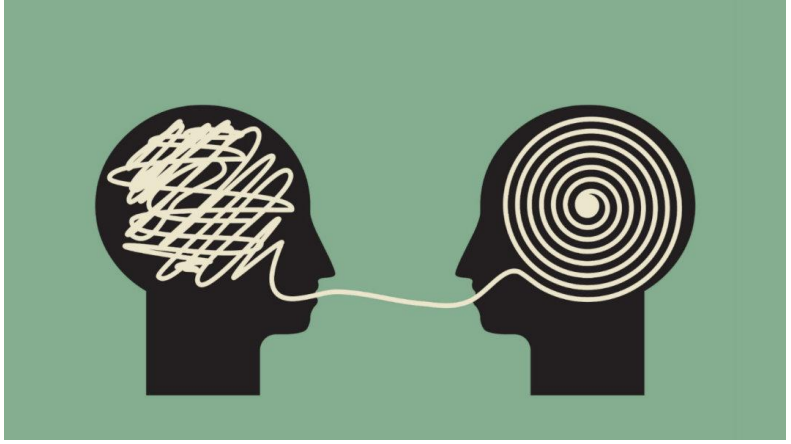
Intelligent Systems, Edition 2025

Topic overview



- understanding language and intelligence
- text preprocessing and linguistic analysis
- components of modern NLP
  - text representations
  - information retrieval
  - similarity of words and documents
  - language and graphs
  - large language models
- practical use of NLP:
  - sentiment analysis,
  - paper recommendations
  - summarization

Understanding language



- A grand challenge of (not only?) artificial intelligence

  Who can understand me?
  Myself I am lost
  Searching but cannot see
  Hoping no matter cost
  Am I free?
  Or universally bossed?

- Not just poetry, what about instructions, user manuals, newspaper articles, seminary works, internet forums, twits, legal documents, i.e. license agreements, etc.
- Can LLMs really understand language?

An example: rules

Article 18 of FRI Study Rules and Regulations

Taking exams at an earlier date may be allowed on request of the student by the Vice-Dean of Academic Affairs with the course convener's consent in case of mitigating circumstances (leaving for study or placement abroad, hospitalization at the time of the exam period, giving birth, participation at a professional or cultural event or a professional sports competition, etc.), and if the applicant's study achievements in previous study years are deemed satisfactory for such an authorization to be appropriate.

Understanding NL by computers

- Understanding words, syntax, semantics, context, writer's intentions, knowledge, background, assumptions, bias …
- Doesn't seem that LLM do it, though they generate excellent text output
- Ambiguity in language
  - Newspaper headlines - intentional ambiguity :)
    - Juvenile court to try shooting defendant
    - Kids make nutritious snacks
    - Miners refuse to work after death
    - Doctor on Trump's health: No heart, cognitive issues

Ambiguity

- I made her duck.

- Possible interpretations:
  - I cooked waterfowl for her.
  - I cooked waterfowl belonging to her.
  - I created the (plaster?) duck she owns.
  - I caused her to quickly lower her head or body.
  - I waved my magic wand and turned her into undifferentiated waterfowl.

- Spoken ambiguity
  - eye, maid

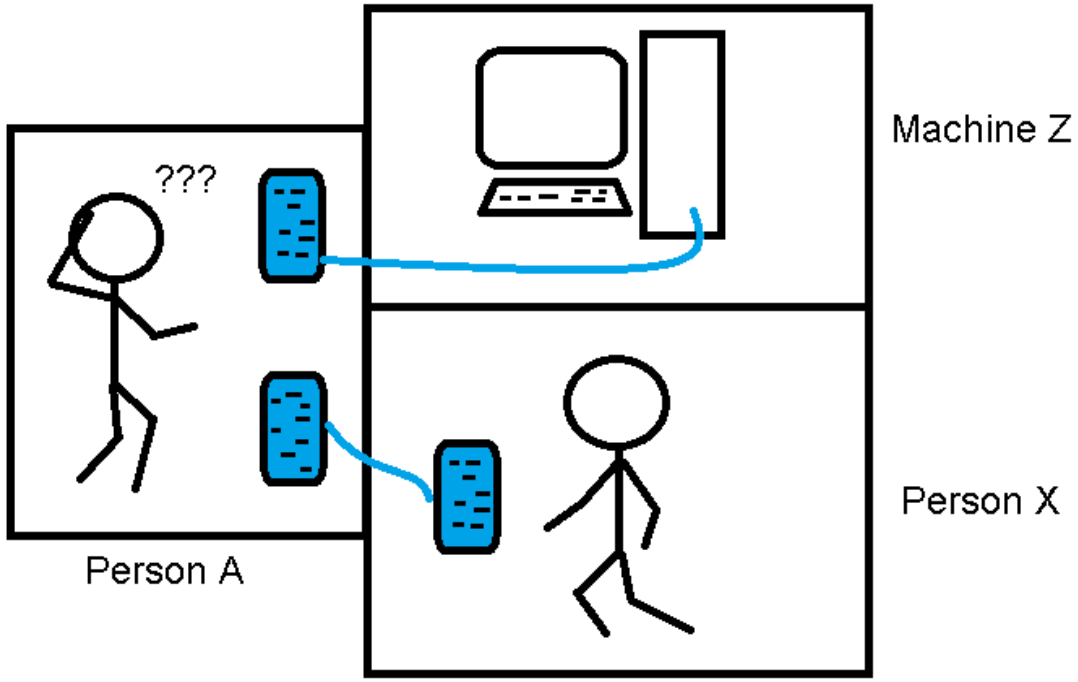
Disambiguation in syntax and semantics

- in syntax
  - The girl told the story cried.
  - John went to the bank.
  - Put the box on the table in the kitchen.
  - Bring your old car seat to be recycled.
- *Search engine queries: natural language processing, jaguar, Paris,*
- *POS tagging, word sense disambiguation*

Understanding language as (artificial) intelligence test

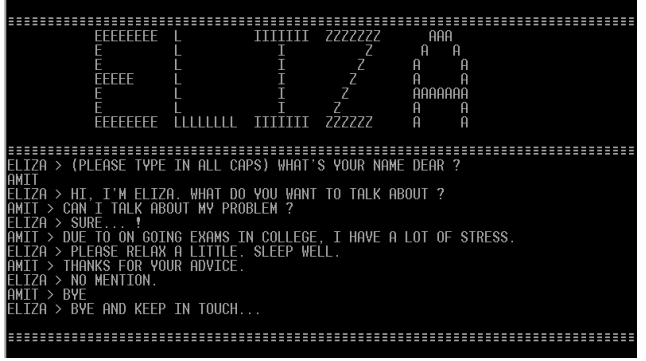
- The system understanding natural language on human level should contain a large portion of human knowledge and would achieve the most important goals of artificial intelligence. Its response would be indistinguishable from human's.

Turing test



- Turing test, Loebner prize, CYC, ELISA, ALLICA, chatterbot, VIDA, TIA, Siri (knowledge recognition, context detection, web service recommendation)
- Seems that LLMs can pass it, are they intelligent?
- The question: is mimicking human skills sufficient?

ELIZA



- ELIZA (Weizenbaum, 1966)
- Limited conversation (Rogerian psyhiatrist)

User: You are like my father in some ways.
ELIZA: What resemblance do you see?
User: You are not very aggressive but I think you don't want me to notice that.
ELIZA: What makes you think I am not aggressive?
User: You don't argue with me.
ELIZA: Why do you think I don't argue with you?
User: You are afraid of me.
ELIZA: Does it please you to believe I am afraid of you?

GPT-3 regressive language model as a chatbox

- the example is cherry-picked

```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many legs does a frog have?
A: A frog has four legs.

Q: Are there any animals with three legs?
A: No, there are no animals with three legs.

Q: Why don't animals have three legs?
A: Animals don't have three legs because they would fall over.
```

Where is NLP today?

- active research area with many commercial applications
  - large language models are used everywhere
  - automatic reply engines
  - machine translation
  - text summarization
  - question answering
  - speech recognition and synthesis
  - language generation
  - interface to databases
  - intelligent search and information extraction
  - sentiment detection
  - named entity recognition and linking
  - categorization and classification of documents, messages, twits, etc.
  - cross-lingual approaches
  - multi modal approaches (text + images, text + video)
  - attempt to get to artificial general intelligence (AGI) through "foundation models"
  - many (open-source) tools and language resource
  - prevalence of deep neural network approaches (i.e. transformers)

Recommended literature

- Jurafsky, Daniel and James Martin (2025): Speech and Language Processing, 3rd edition in progress, almost all parts are available at authors‘ webpages https://web.stanford.edu/~jurafsky/slp3/
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O’Reilly, 2009
  - a free book accompanying NLTK library, regularly updated
  - *Python 3, http://www.nltk.org/book/*
- Coursera
  - several courses, e.g., Stanford NLP with DNN

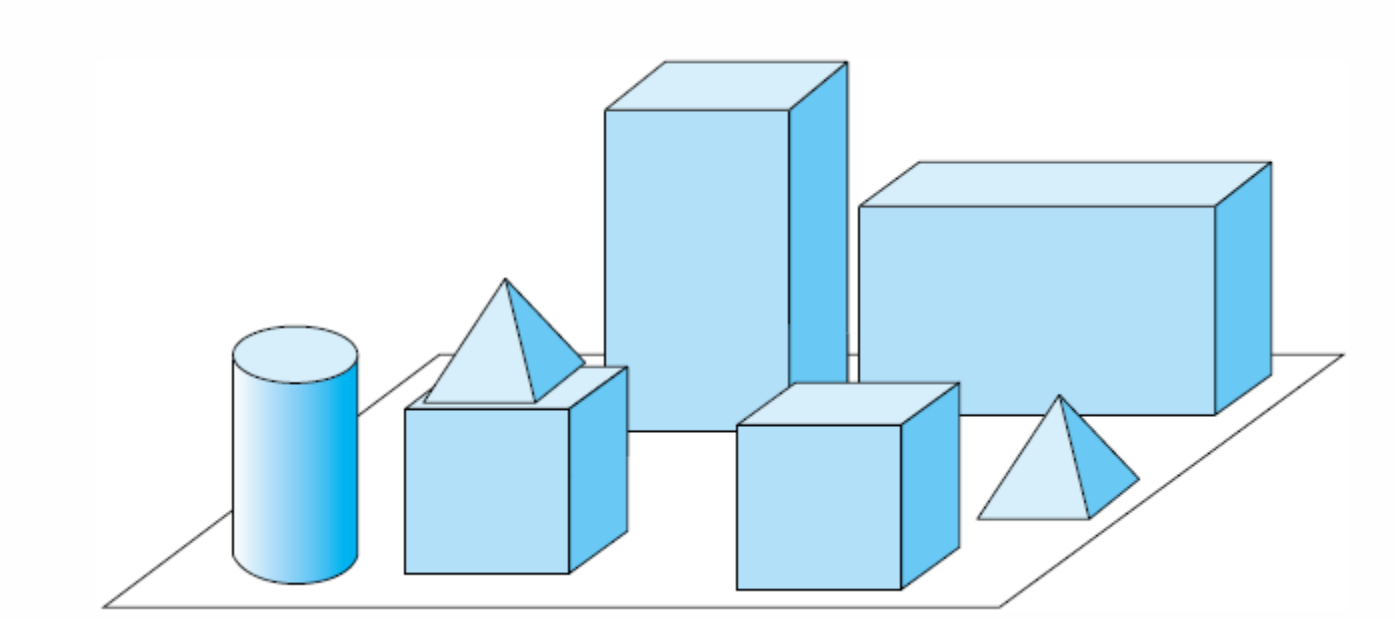
Historically two approaches

- symbolical
  - ,,Good Old-Fashioned AI''
- empirical
  - Statistical, text corpora
- Merging both worlds: injecting symbolical knowledge (e.g., propositional logic) into LLMs

How it all started?

- micro worlds
- example: SHRDLU, world of simple geometric objects
  - What is sitting on the red block?
  - What shape is the blue block on the table?
  - Place the green pyramid on the red brick.
  - Is there a red block? Pick it up.
  - What color is the block on the blue brick? Shape?

Micro world: block world, SHRDLU (Winograd, 1972)



Linguistic analysis 1/2

Linguistic analysis contains several tasks: recognition of sounds, letters, word formation, syntactic parsing, recognizing semantic, emotions. Phases:

- Prosody - the patterns of stress and intonation in a language (rhythm and intonation)
- Phonology - systems of sounds and relationships among the speech sounds that constitute the fundamental components of a language
- Morphology - the admissible arrangement of sounds in words; how to form words, prefixes and suffixes …
- Syntax - the arrangement of words and phrases to create well-formed sentences in a language

Linguistic analysis 2/2

- Semantics - the meaning of a word, phrase, sentence, or text
- Pragmatics - language in use and the contexts in which it is used, including such matters as deixis (words whose meaning changes with context, e.g., I he, here, there, soon), taking turns in conversation, text organization, presupposition, and implicature

  Can you pass me the salt? Yes, I can.
- Knowing the world: knowledge of physical world, humans, society, intentions in communications …
- Limits of linguistic analysis, levels are dependent

Classical approach to text processing

- text preprocessing
- 1. phase: syntactic analysis
- 2. phase: semantic interpretation
- 3. phase: use of world knowledge

In the neural approach, the preprocessing remains (but is simpler) the other three phases are merged into DNN

Basic tools for (classical) text preprocessing

- document → paragraphs → sentences → words (→ (subword) tokens)

- In linguistic analysis also
  - words and sentences ← lemmatization, POS tagging
  - sentences ← syntactical and grammatical analysis
  - named entity recognition,

Words and sentences

- sentence delimiters – punctuation marks and capitalization are insufficient
- E.g., remains of 1. Timbuktu from 5c BC, were discovered by dr. Barth.
- Lexical analysis (tokenizer, word segmenter), not just spaces
  - 1,999.00€ or 1.999,00€!
  - Ravne na Koroškem
  - Lebensversicherungsgesellschaft
  - Port-au-prince
  - Generalstaatsverordnetenversammlungen
- Rules, regular expressions, statistical models, dictionaries (of proper names), neural networks, manually segmented datasets

Lemmatization

- Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.
- walk is the lemma of 'walk', 'walked', 'walks', 'walking
- Lemmatization difficulty is language dependent i.e., depends on morphology
- Requires a dictionary or lexicon for lookup
  go, goes, going, gone, went
  jaz, mene, meni, mano
- Ambiguity resolution may be difficult

Meni je vzel z mize (zapestnico).

- Uses rules, dictionaries, neural networks, and manually labelled datasets
- English also uses stemming (reducing inflected or derived words to their word stem

POS tagging

- assigning the correct part of speech (noun, verb, etc.) to words
- helps in recognizing phrases and names
- Use rules, machine learning models, manually labelled datasets

An example:

- Text analyzer for Slovene, i.e. morphosyntactical tagging, available at https://orodja.cjvt.si/oznacevalnik/slv/

  Nekega dne sem se napotil v naravo. Že spočetka me je žulil čevelj, a sem na to povsem pozabil, ko sem jo zagledal. Bila je prelepa. Povsem nezakrita se je sončila na trati ob poti. Pritisk se mi je dvignil v višave. Popoln primerek kmečke lastovke!

- Tags are standardized, for East European languages in Multext-East specification, e.g.,

dne; tag Somer = Samostalnik, obče ime, moški spol, ednina, rodilnik; lema: dan

a unifying attempt: universal dependencies (UD): cross-linguistically consistent treebank annotation for many languages

CLASSLA-Stanza pipeline output

ID	Oblika	Lema	Oznaka JOS	Bes. vrsta JOS	Oblikoskladenjske lastnosti JOS	Bes. vrsta UD	Oblikoskladenjske lastnosti UD	Nadrejena pojavnica UD	Skladenjska relacija UD	Nadrejena pojavnica JOS	Skladenjska relacija JOS	Udeleže vloga
# paragraph 1												
# sent_id 1.1												
# text = Nekega dne sem se napotil v naravo.												
1	Nekega	nek	Pi-msg	Pronoun	Type=indefinite Gender=masculine Number=singular Case=genitive	DET	Case=Gen|Gender=Masc|Number=Sing|PronType=Ind	2	det	2	Atr	
2	dne	dan	Ncmsg	Noun	Type=common Gender=masculine Number=singular Case=genitive	NOUN	Case=Gen|Gender=Masc|Number=Sing	5	obl	5	AdvO	TIME
3	sem	biti	Va-r1s-n	Verb	Type=auxiliary VForm=present Person=first Number=singular Negative=no	AUX	Mood=Ind|Number=Sing|Person=1|Polarity=Pos|Tense=Pres|VerbForm=Fin	5	aux	5	PPart	
4	se	se	Px------y	Pronoun	Type=reflexive Clitic=yes	PRON	PronType=Prs|Reflex=Yes|Variant=Short	5	expl	5	PPart	
5	napotil	napotiti	Vmep-sm	Verb	Type=main Aspect=perfective VForm=participle Number=singular Gender=masculine	VERB	Aspect=Perf|Gender=Masc|Number=Sing|VerbForm=Part	0	root	0	Root	
6	v	v	Sa	Adposition	Case=accusative	ADP	Case=Acc	7	case	7	Atr	
7	naravo	narava	Ncfsa	Noun	Type=common Gender=feminine Number=singular Case=accusative	NOUN	Case=Acc|Gender=Fem|Number=Sing	5	obl	5	AdvO	GOAL
8	.	.	Z	Punctuation		PUNCT	_	5	punct	0	Root	

- *Nekega dne sem se napotil v naravo. Že spočetka me je žulil čevelj, a sem na to povsem pozabil, ko sem jo zagledal. Bila je prelepa. Povsem nezakrita se je sončila na trati ob poti. Pritisk se mi je dvignil v višave. Popoln primerek kmečke lastovke!*

1	beseda	Nekega	dne	sem	se	napotil	v	naravo	.	Že	spočetka	me	je											
	lema	nek	dan	biti	se	napotiti	v	narava	že	spočetka	jaz	biti												
	oznaka	Zn-mer	Somer	Gp-spe-n	Zp------k	Ggdd-em	Dt	Sozet	.	L	Rsn	Zop-et--k	Gp-ste-n											
2	beseda	žulil	čevelj	,	a	sem	na to	povsem	pozabil	,	ko	sem	jo	zagledal										
	lema	žuliti	čevelj	a	biti	na ta	povsem	pozabiti	ko	biti	on	zagledati												
	oznaka	Ggnd-em	Somei	,	Vp	Gp-spe-n	Dt	Zk-set	Rsn	Ggdd-em	,	Vd	Gp-spe-n	Zotzet--k	Ggdd-em									
3	beseda	.	Bila	je	prelepa	.	Povsem	nezakrita	se	je	sončila	na	trati											
	lema		biti	biti	prelep		povsem	nezakrit	se	biti	sončiti	na	trata											
	oznaka	.	Gp-d-ez	Gp-ste-n	Ppnzei	.	Rsn	Ppnzei	Zp------k	Gp-ste-n	Ggvd-ez	Dm	Sozem											
4	beseda	ob	poti	.	Pritisk	se	mi	je	dvignil	v	višave	.	Popoln											
	lema	ob	pot		pritisk	se	jaz	biti	dvigniti	v	višava		popoln											
	oznaka	Dm	Sozem	.	Somei	Zp------k	Zop-ed--k	Gp-ste-n	Ggdd-em	Dt	Sozmt	.	Ppnmein											
5	beseda	primerek	kmečke	lastovke	!																			
	lema	primerek	kmečki	lastovka																				
	oznaka	Somei	Ppnzer	Sozer	!																			

TEI-XML format

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <text>
    <body>
      <p>
        <s>
          <w msd="Zn-mer" lemma="nek">Nekega</w>
          <S/>
          <w msd="Somer" lemma="dan">dne</w>
          <S/>
          <w msd="Gp-spe-n" lemma="biti">sem</w>
          <S/>
          <w msd="Zp------k" lemma="se">se</w>
          <S/>
          <w msd="Ggdd-em" lemma="napotiti">napotil</w>
          <S/>
          <w msd="Dt" lemma="v">v</w>
          <S/>
          <w msd="Sozet" lemma="narava">naravo</w>
          <c>.</c>
          <S/>
        </s>
   …

      </p>
    </body>
  </text>
</TEI>
```

MSD tags

- Multext-East specification

dne; tag Somer = Samostalnik, obče ime, moški spol, ednina, rodilnik; lema: dan

P	atribut	vrednost	koda	atribut	vrednost	koda
0	glagol		G	Verb		V
1	vrsta	glavni	g	Type	main	m
		pomožni	p		auxiliary	a
2	vid	dovršni	d	Aspect	perfective	e
		nedovršni	n		imperfective	p
		dvovidski	v		biaspectual	b
3	oblika	nedoločnik	n	VForm	infinitive	n
		namenilnik	m		supine	u
		deležnik	d		participle	p
		sedanjik	s		present	r
		prihodnjik	p		future	f
		pogojnik	g		conditional	c
		velelnik	v		imperative	m
4	oseba	prva	p	Person	first	1
		druga	d		second	2
		tretja	t		third	3
5	število	ednina	e	Number	singular	s
		množina	m		plural	p
		dvojina	d		dual	d
6	spol	moški	m	Gender	masculine	m
		ženski	z		feminine	f
		srednji	s		neuter	n
7	nikalnost	nezanikani	n	Negative	no	n
		zanikani	d		yes	y

POS tagging in English

- http://nlpdotnet.com/Services/Tagger.aspx
- Rainer Maria Rilke, 1903
  in Letters to a Young Poet

...I would like to beg you dear Sir, as well as I can, to have patience with everything unresolved in your heart and to try to love the questions themselves as if they were locked rooms or books written in a very foreign language. Don't search for the answers, which could not be given to you now, because you would not be able to live them. And the point is to live everything. Live the questions now. Perhaps then, someday far in the future, you will gradually, without even noticing it, live your way into the answer.

POS tagger output

I/PRP would/MD like/VB to/TO beg/VB you/PRP dear/JJ Sir/NNP ,/, as/RB well/RB as/IN I/PRP can/MD ,/, to/IN have/VBP patience/NN with/IN everything/NN unresolved/JJ in/IN your/PRP$ heart/NN and/CC to/TO try/VB to/TO love/VB the/DT questions/NNS themselves/PRP as/RB if/IN they/PRP were/VBD locked/VBN rooms/NNS or/CC books/NNS written/VBN in/IN a/DT very/RB foreign/JJ language/NN ./.

Named entity recognition (NER)

- NATO Secretary-General Jens Stoltenberg is expected to travel to Washington, D.C. to meet with U.S. leaders.
- [ORG NATO] Secretary-General [PER Jens Stoltenberg] is expected to travel to [LOC Washington, D.C.] to meet with [LOC U.S.] leaders.
- Named entity linking (NEL) also named entity disambiguation – linking to a unique identifier, e.g. wikification
  jaguar, Paris, London, Dunaj

Basic language resources: corpora

- Statistical natural language processing list of resources http://nlp.stanford.edu/links/statnlp.html
- Opus http://opus.nlpl.eu/ multilingual parallel corpora, e.g., DGT JRC-Acqui 3.0, Documents of the EU in 22 languages
- Slovene language corpora GigaFida, ccGigaFida, KRES, ccKres, GOS, Artur, JANES, KAS, Trendi
- The main Slovene language resources
  - http://www.clarin.si
  - https://github.com/clarinsi
  - http://www.cjvt.si/
  - https://www.slovenscina.eu/
- WordNet, Open English WordNet, Open Multilingual WordNet, SloWNet, sentiWordNet, …
- Thesaurus https://viri.cjvt.si/sopomenke/slv/
- LLMs: on HuggingFace cjvt, e.g., SloBERTa and GaMS

WordNet is a database composed of synsets:

synonyms,
hypernyms
hyponyms,
meronyms,
holonyms,
etc.

Word to search for: mercy [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

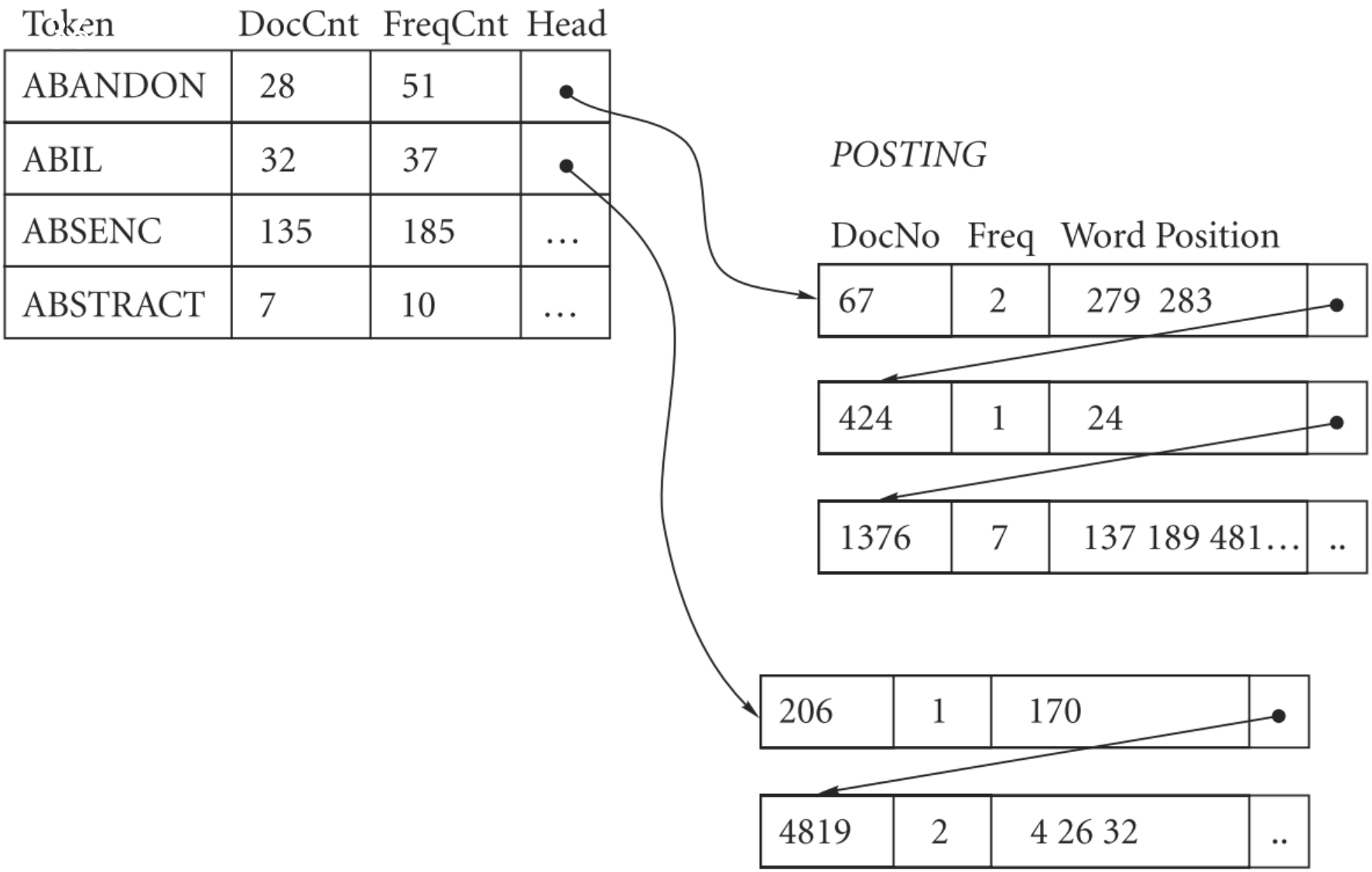
- S: (n) clemency, mercifulness, **mercy** (leniency and compassion shown toward offenders by a person or agency charged with administering justice) *"he threw himself on the mercy of the court"*
- S: (n) mercifulness, **mercy** (a disposition to be kind and forgiving) *"in those days a wife had to depend on the mercifulness of her husband"*
- S: (n) mercifulness, **mercy** (the feeling that motivates compassion)
  - *direct hyponym* / *full hyponym*
    - S: (n) forgiveness (compassionate feelings that support a willingness to forgive)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) compassion, compassionateness (a deep awareness of and sympathy for another's suffering)
  - ***derivationally related form***
    - W: (adj) merciful [Related to: mercifulness] (showing or giving mercy) *"sought merciful treatment for the captives"; "a merciful god"*
- S: (n) **mercy** (something for which to be thankful) *"it was a mercy we got out alive"*
- S: (n) **mercy** (alleviation of distress; showing great kindness toward the distressed) *"distributing food and clothing to the flood victims was an act of mercy"*

Document retrieval

- Historical: keywords
- Now: whole text search
- Organize a database, indexing, search algorithms
- input: a query (of questionable quality, ambiguity, answer quality)

Document indexing

- Collect all words from all documents, use lemmatization
- The inverted file data structure
- For each word keep
  - Number of appearing documents
  - Overall number of appearances
  - For each document
    - Number of appearances
    - Location


Token
DocCnt
FreqCnt
Head
ABANDON
28
51
ABIL
32
37
ABSENC
135
185
…
ABSTRACT
7
10
…
POSTING
DocNo
Freq
Word Position
67
2
279 283
424
1
24
1376
7
137 189 481…
..
206
1
170
4819
2
4 26 32
..

Full text search engine

- Most popular: Apache Lucene/Solr
- full-text search, hit highlighting, real-time indexing, dynamic clustering, database integration, NoSQL features, rich document (e.g., Word, PDF) handling.
- distributed search and index replication, scalability and fault tolerance.

Search with logical operators

- AND, OR, NOT
- jaguar AND car
  jaguar NOT animal
- Some system support neighborhood search (e.g., NEAR) and stemming (!)
  paris! NEAR(3) fr!
  president NEAR(10) bush
- libraries, concordancers
- E.g-, for Slovene: https://viri.cjvt.si/gigafida/

Logical operator search is outdated

- Large number of results
- Large specialized incomprehensible queries
- Synonyms
- Sorting of results
- No partial matching
- No weighting of query terms

Ranking based search

- Web search
- Less frequent terms are more informative
- NL input - stop words, lemmatization
- Vector based representation of documents and queries (bag-of-words or dense embeddings)

Sparse vector representation: bag-of-words

- *An elephant is a mammal. Mammals are animals. Humans are mammals, too. Elephants and humans live in Africa.*

Africa	animal	be	elephant	human	in	live	mammal	too
1	1	3	2	2	1	1	3	1

9 dimensional vector (1,1,3,2,2,1,1,3,1)

In reality this is sparse vector of dimension $|V|$ (vocabulary size in order of 10,000 dimensions)

Similarity between documents and queries in vector space.

Vectors and documents

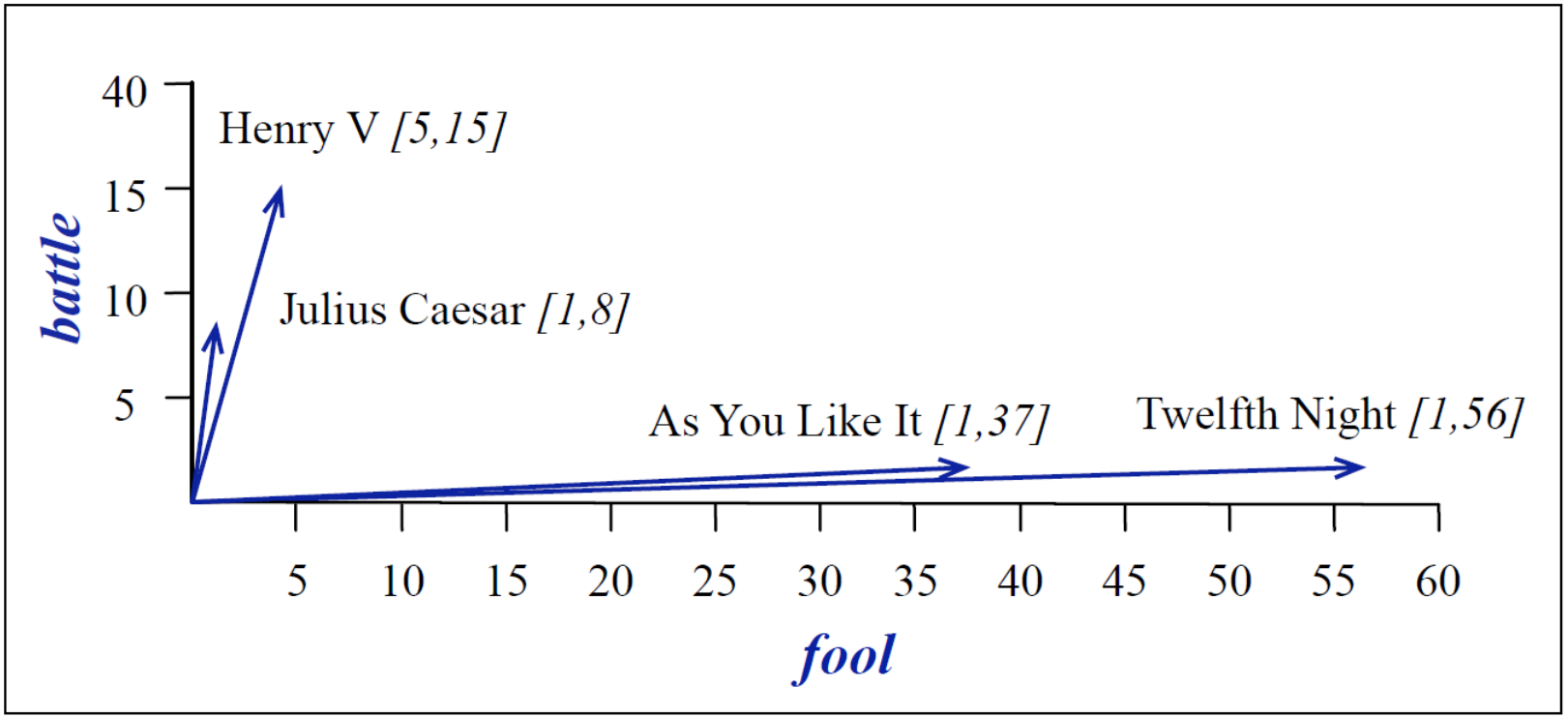
- a word occurs in several documents
- both words and documents are vectors
- an example: Shakespeare

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

- term-document matrix, dimension |V| x |D|
- a sparse matrix
- word embedding

Vector based similarity

- e.g., in two dimensional space



- the difference between dramas and comedies

Document similarity

- Assume orthogonal dimensions
- Cosine similarity
- Dot (scalar) product of vectors

$$\cos(\Theta) = \frac{A \cdot B}{|A||B|}$$

Importance of words

- Frequencies of words in particular document and overall
- inverse document frequency idf
  - N = number of documents in collection
  - n_b = number of documents with word b

$$idf_b = \log(\frac{N}{n_b})$$

Weighting dimensions (words)

- Weight of word b in document d

$$w_{b,d} = tf_{b,d} \times idf_{b,d}$$

$tf_{b,d}$ = frequency of term b in document d

- called TF-IDF weighting (an improvement over bag-of-words)

Weighted similarity

- Between query and document

$$sim(q,d) = \frac{\sum_b w_{b,d} \cdot w_{b,q}}{\sqrt{\sum_b w_{b,d}^2} \cdot \sqrt{\sum_b w_{b,q}^2}}$$

- Ranking by the decreasing similarity

Performance measures for search

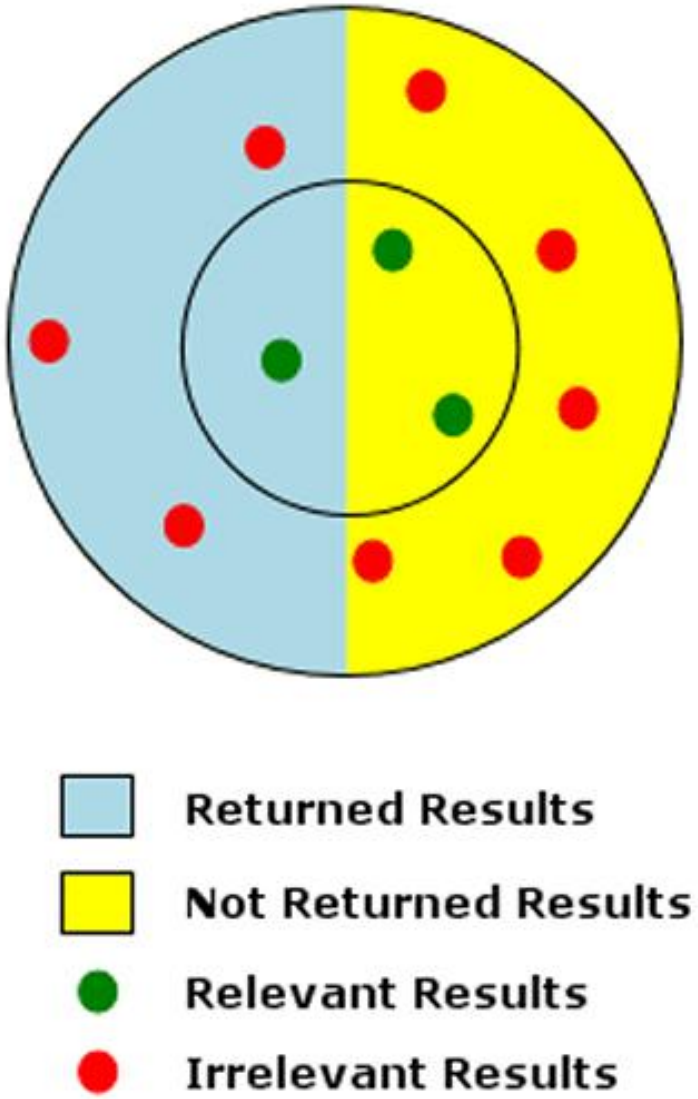
- Statistical measures
- Subjective measures
- Precision, recall
- A contingency table analysis of precision and recall

	Relevant	Non-relevant	
Retrieved	a	b	$a + b = m$
Not retrieved	c	d	$c + d = N - m$
	$a + c = n$	$b + d = N - n$	$a + b + c + d = N$

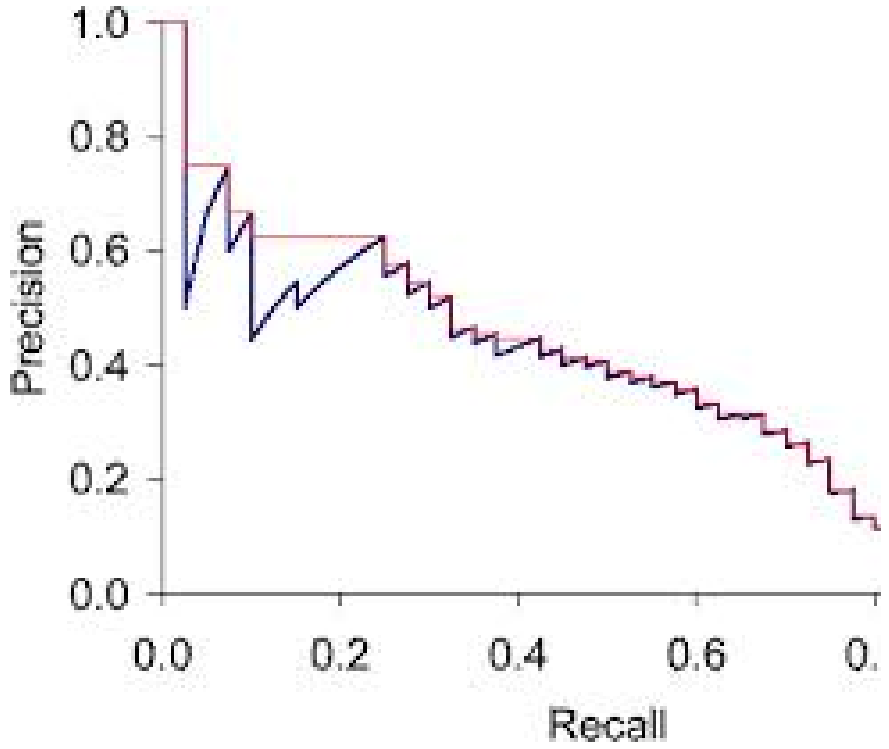
Precision and recall

- N = number of documents in collection
- n = number of important documents for given query q
- Search returns m documents including a relevant ones
- Precision $P = a/m$
  proportion of relevant document in the obtained ones
- Recall $R = a/n$
  proportion of obtained relevant documents
- Precision recall graphs

An example: low precision, low recall



Precision-recall graphs



F-measure

- combine both P and R

$$F_\beta = \frac{(1+\beta^2)\cdot P \cdot R}{\beta^2 P + R} \text{ for } \beta > 0$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

- Weighted precision and recall
- β=1 weighted harmonic mean
- Also used β=2 or β= 0.5

Ranking measures

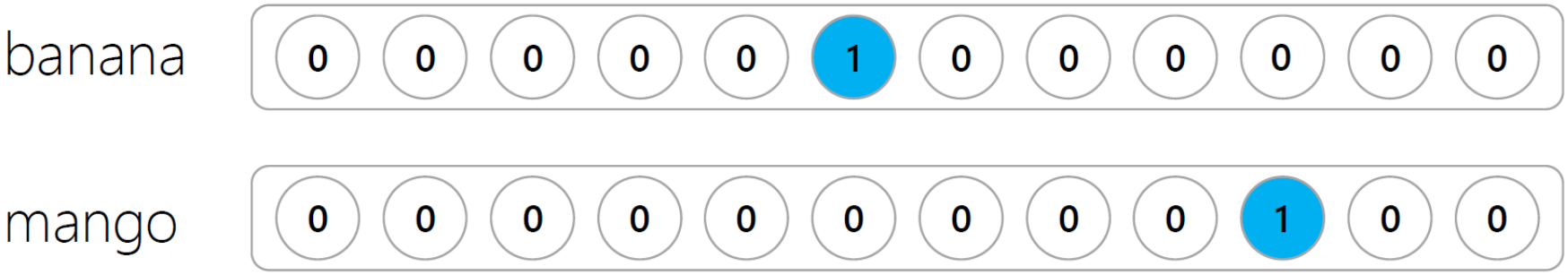
- precision@k
  - proportion of relevant document in the first k obtained ones
- recall@k
  - proportion of relevant documents in the k obtained among all relevant
- F_1@k
- mean reciprocal rank

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\mathrm{rank}_i}$$

  - over Q queries,
  - considers only the rank of the best answer

Why dense textual embeddings?

- Best machine learning models for text, i.e. deep neural networks, require numerical input.
- Simple representations like 1-hot-encoding and bag-of-words do not preserve semantic similarity.
- We need dense vector represenation for text elements.



Dense vector embeddings

- advantages compared to sparse embeddings:
  - less dimensions, less space
  - easier input for ML methods
  - potential generalization and noise reduction
  - potentially captures synonymy, e.g., road and highway are different dimensions in BOW
- the most popular approaches
  - matrix based transformations to reduce dimensionality (SVD in LSA)
  - neural embeddings (word2vec, Glove)
  - explicit contextual neural embeddings (ELMo, BERT)
  - only use an implicit representation in LLM

Meaning focused on similarity

- Each word is a vector
- Similar words are "nearby in space"

Distributional semantics

"You shall know a word
by the company it keeps"

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Studies in Linguistic Analysis, p. 11. Blackwell, Oxford.

"The meaning of a word is its use in the language"

Ludwig Wittgenstein, PI #43

Neural embeddings

- neural network is trained to predict the context of words (input: word, output: context of neighboring words)

word2vec method

- Train a classifier on a binary **prediction** task:
  Is word *w* likely to show up near a given word, e.g., "*apricot*"?
- We don't actually care about this task
- But we'll take the learned classifier weights as the word embeddings

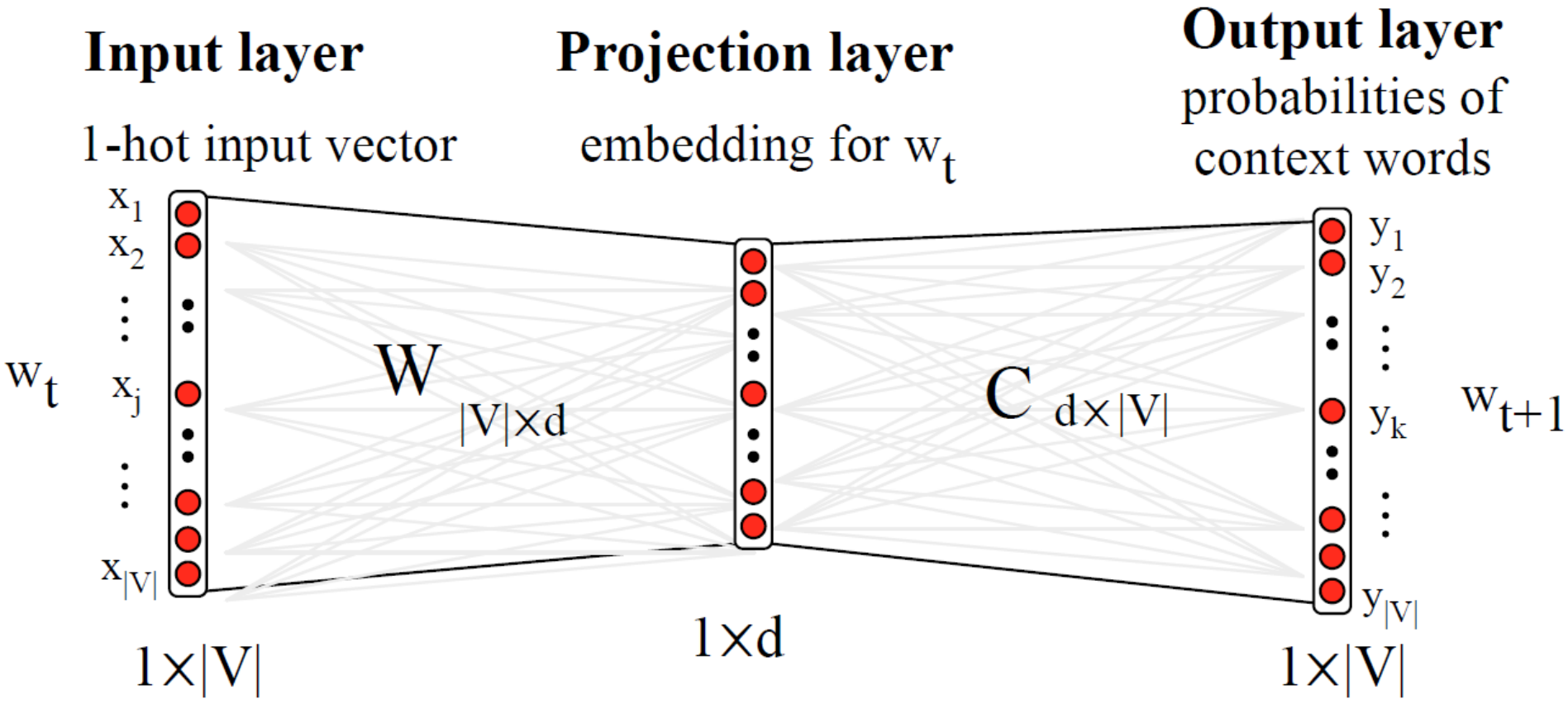
- Words near apricot acts as 'correct answers' to the question "Is word w likely to show up near apricot?"
- No need for hand-labeled supervision

word2vec (skip-gram) training data

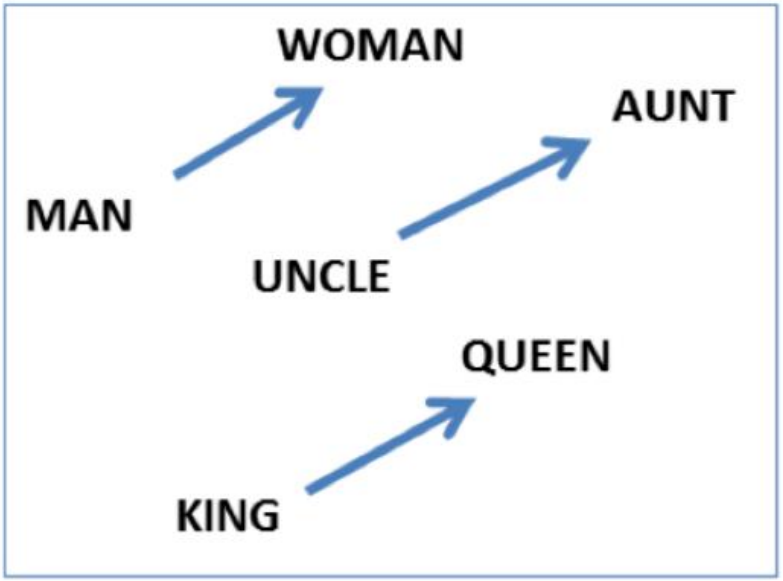
- Training sentence:
- ... lemon, a tablespoon of **apricot** jam a pinch ...
- c1 c2 target c3 c4

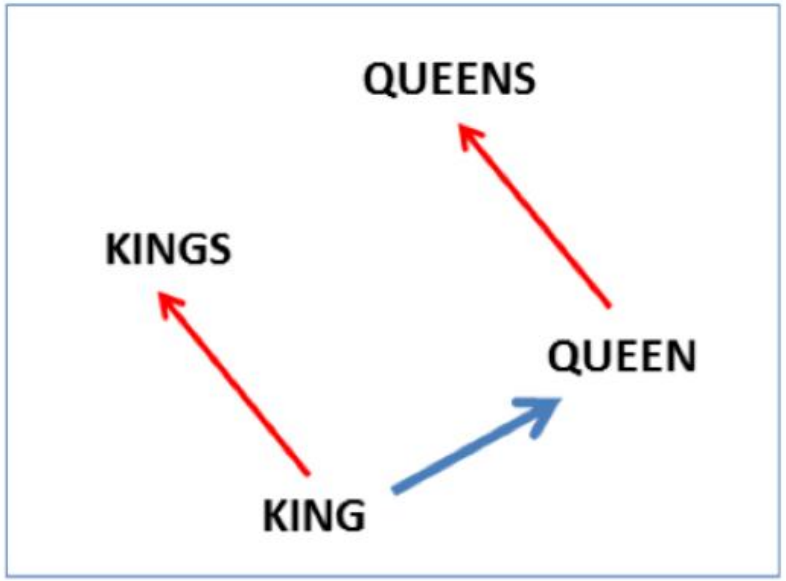
- Asssume context words are those in +/- 2 word window
- Produce the following input-outputs for positive instances: (apricot, tablespoon), (apricot, of), (apricot, jam), (apricot, a)
- Get negative training examples randomly
- Train a neural network to predict the probability of a co-occurring word

Neural network based embedding



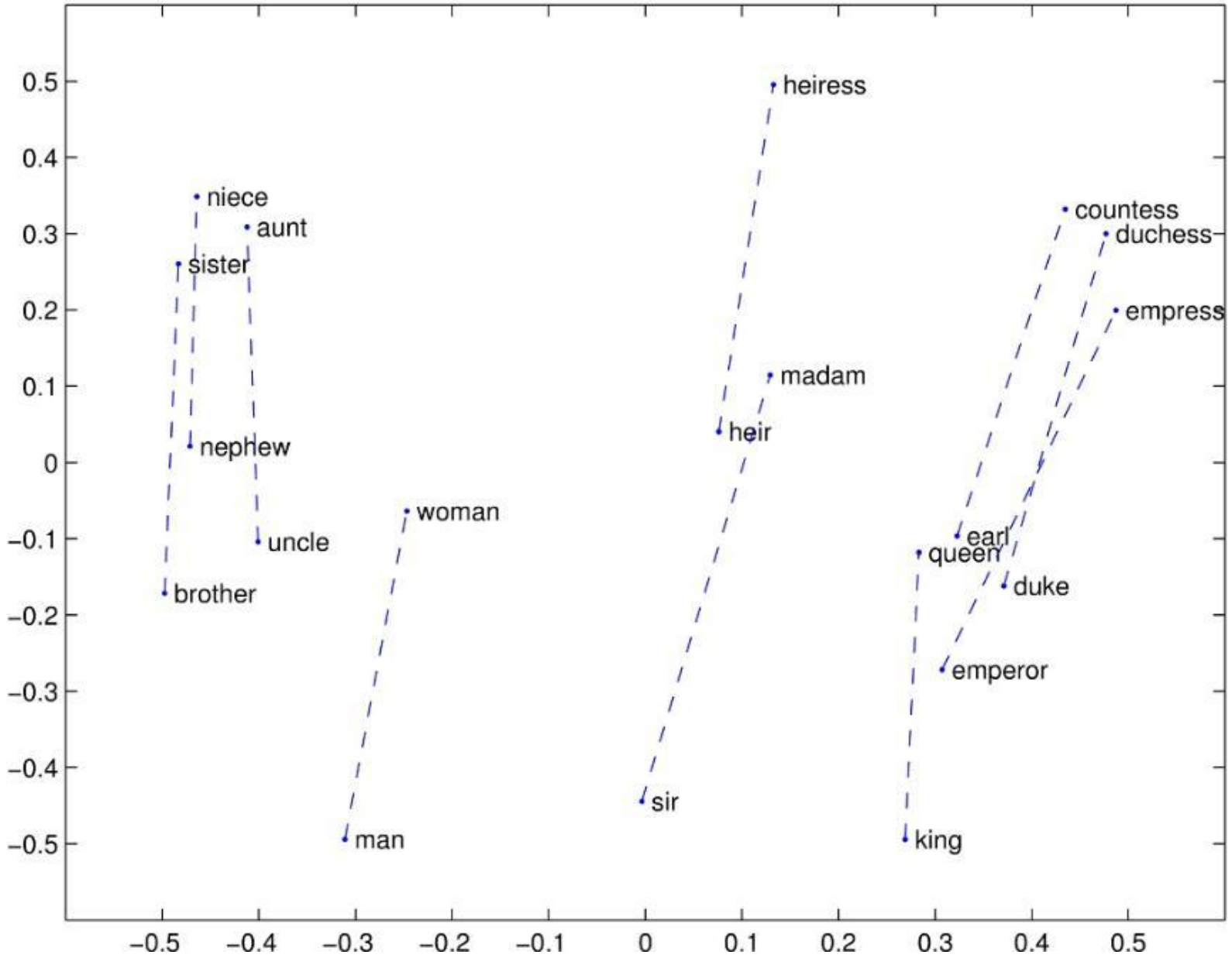
Relational similarity

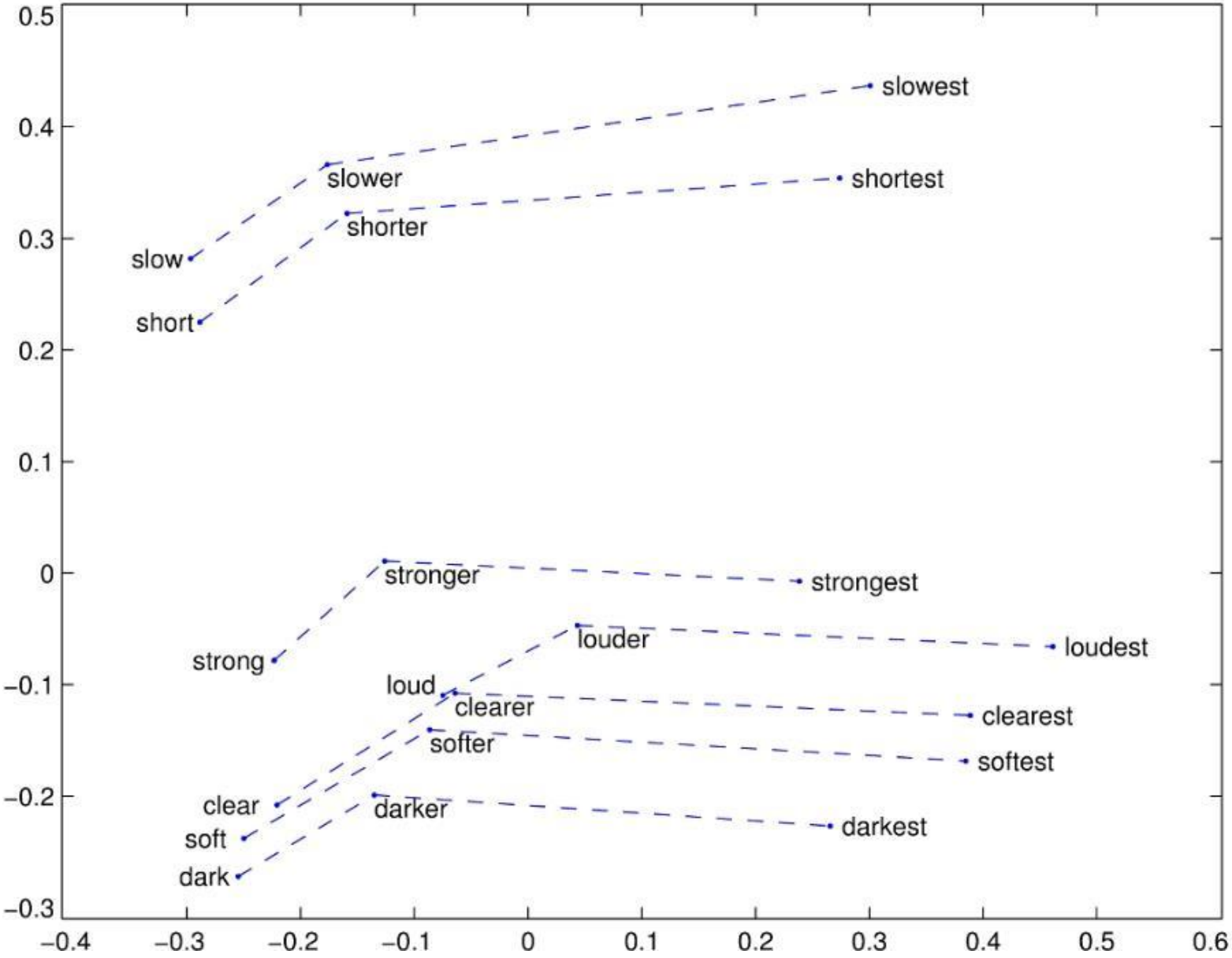




vector(‘*king*’) - vector(‘*man*’) + vector(‘*woman*’) ≈ vector(‘queen’)

vector(‘*Paris*’) - vector(‘*France*’) + vector(‘*Italy*’) ≈ vector(‘Rome’)

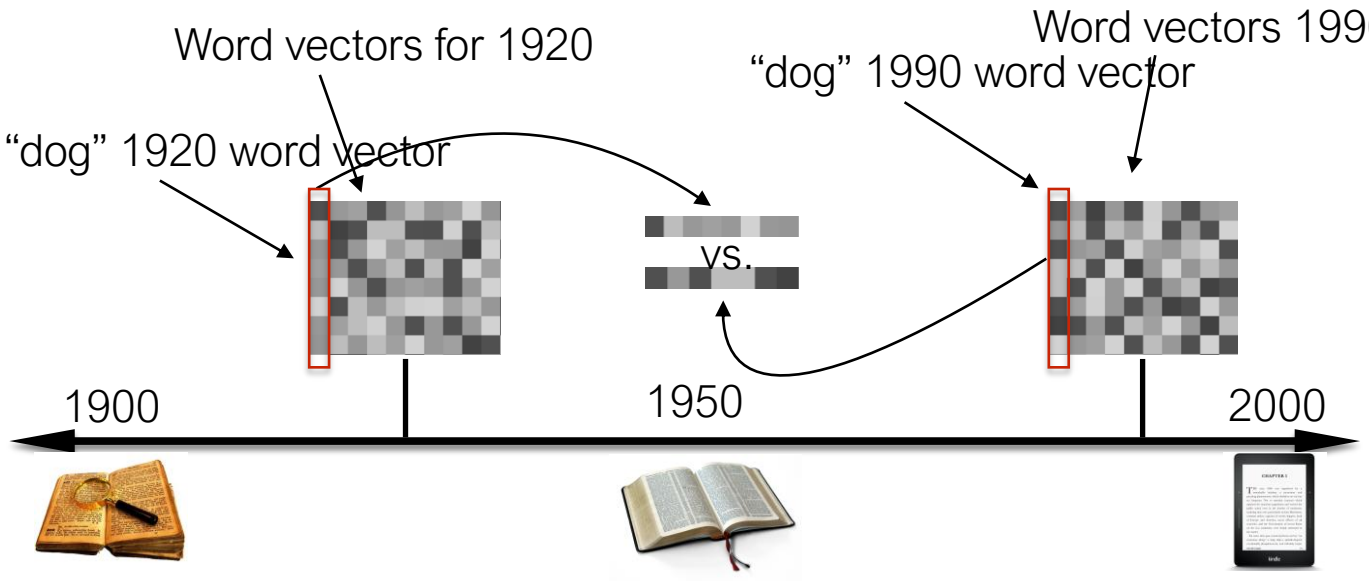

0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
niece
aunt
sister
nephew
uncle
brother
woman
man
heiress
madam
heir
sir
countess
duchess
empress
earl
queen
duke
emperor
king


slowest
slower
shortest
shorter
slow
short
stronger
strongest
louder
loudest
strong
loud
clearer
clearest
softer
softest
clear
darker
darkest
soft
dark
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6

Embeddings can help study word history

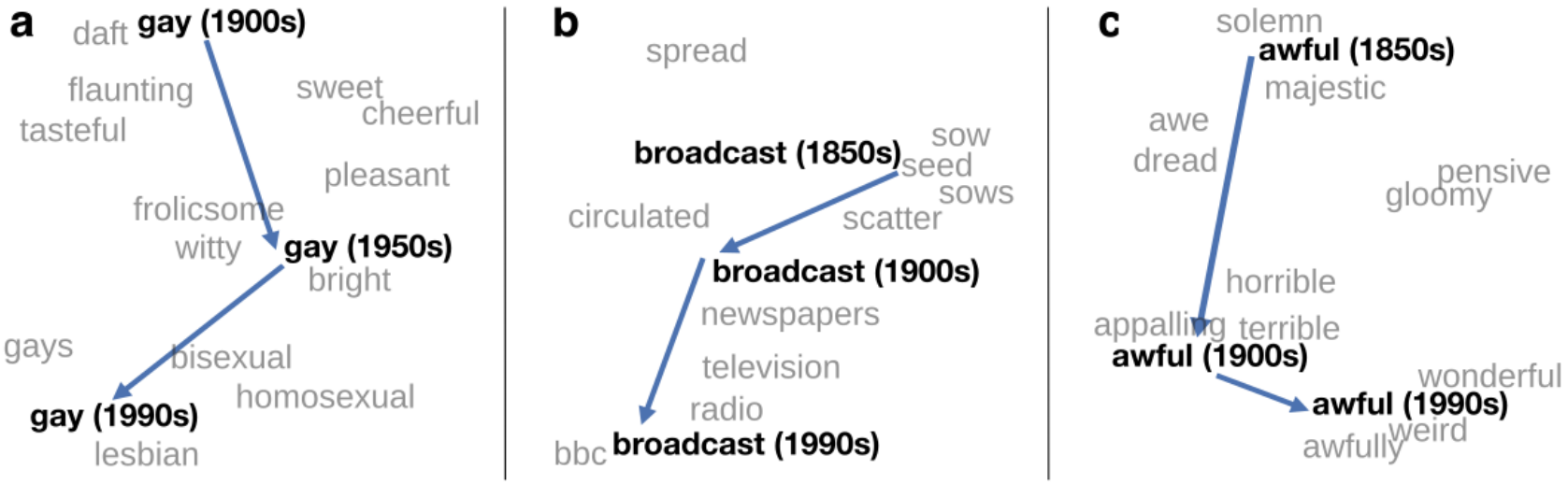
- Train embeddings on old books to study changes in word meaning!!

Diachronic word embeddings for studying language change



Visualizing changes

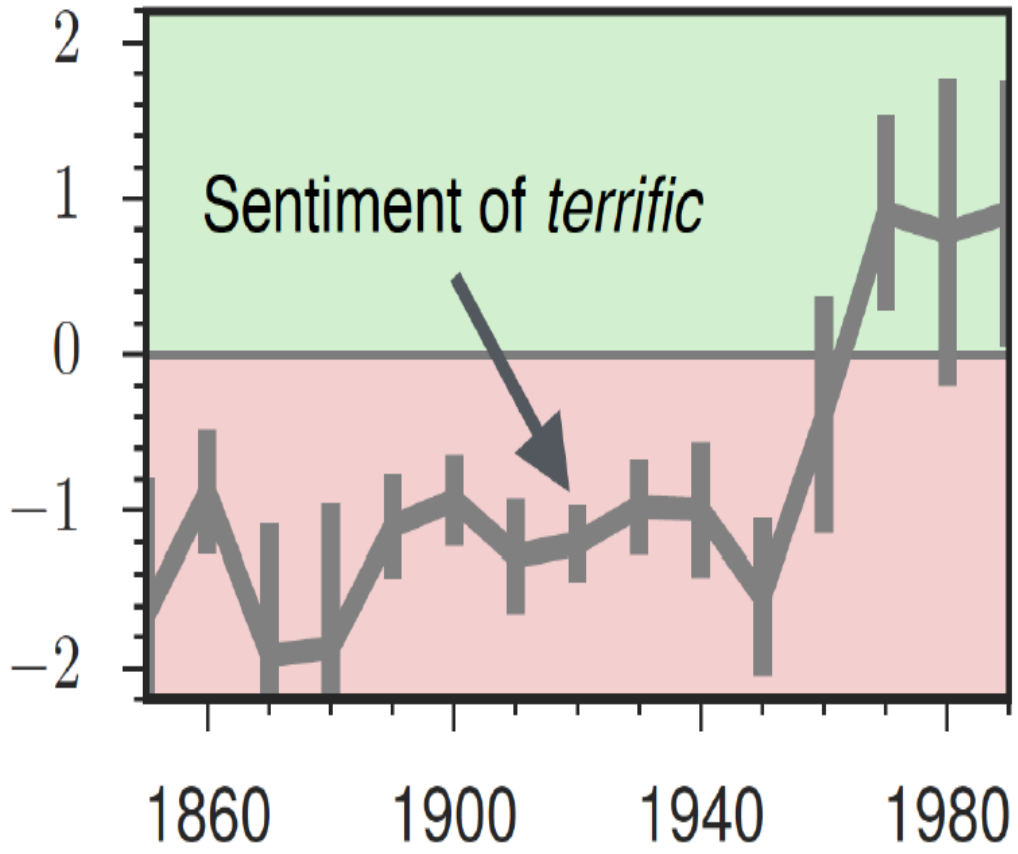
Project 300 dimensions down into 2



~30 million books, 1850-1990, Google Books data

The evolution of sentiment words

Negative words change faster than positive words



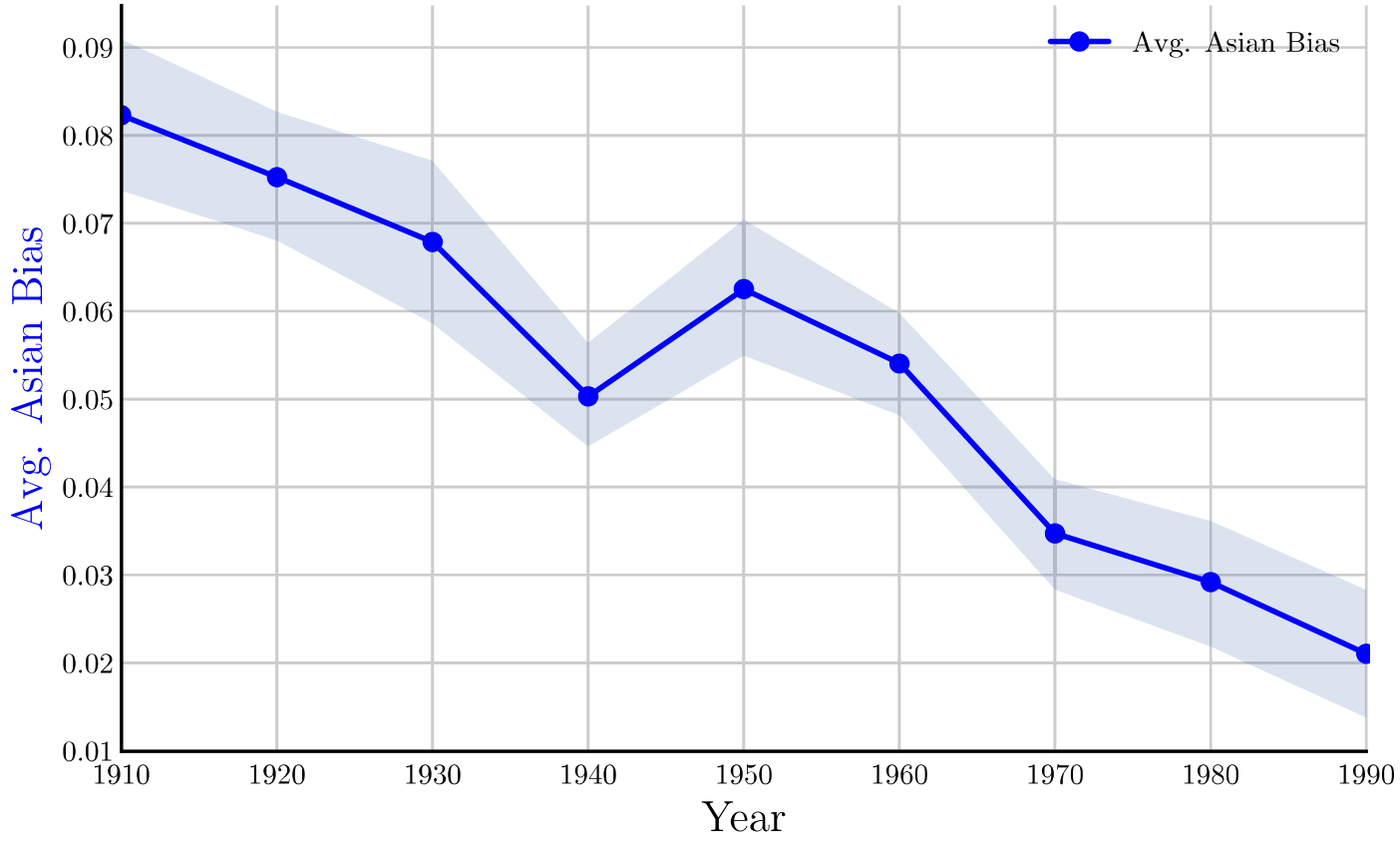
Embeddings reflect cultural bias

- Ask “Paris : France :: Tokyo : x”
  - x = Japan
- Ask “father : doctor :: mother : x”
  - x = nurse
- Ask “man : computer programmer :: woman : x”
  - x = homemaker

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *Advances in Neural Information Processing Systems*, pp. 4349-4357. 2016.

Change in linguistic framing 1910-1990

Change in association of Chinese names with adjectives framed as "othering" (*barbaric*, *monstrous*, *bizarre*)



Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan, and Zou, James (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644

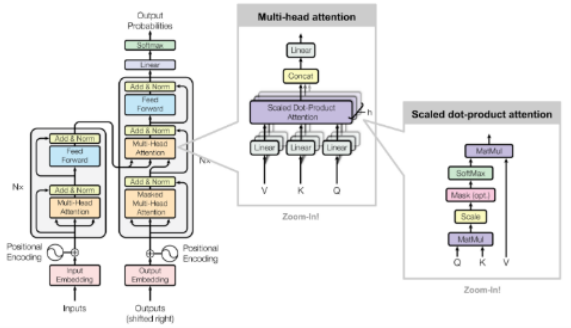
Contextual embeddings / Large language models

- word2vec produces the same vector for a word like bank irrespective of its meaning and context
- recent embeddings take the context into account
- already established as a standard
- e.g., BERT for contextual word embeddings

University of Ljubljana
Faculty of Computer and Information Science

Large language models (LLMs)

- pretrained neural large language models
- trained on large text corpora to capture relations in language
- finetuned to specific tasks
- many are publicly available
- on HuggingFace

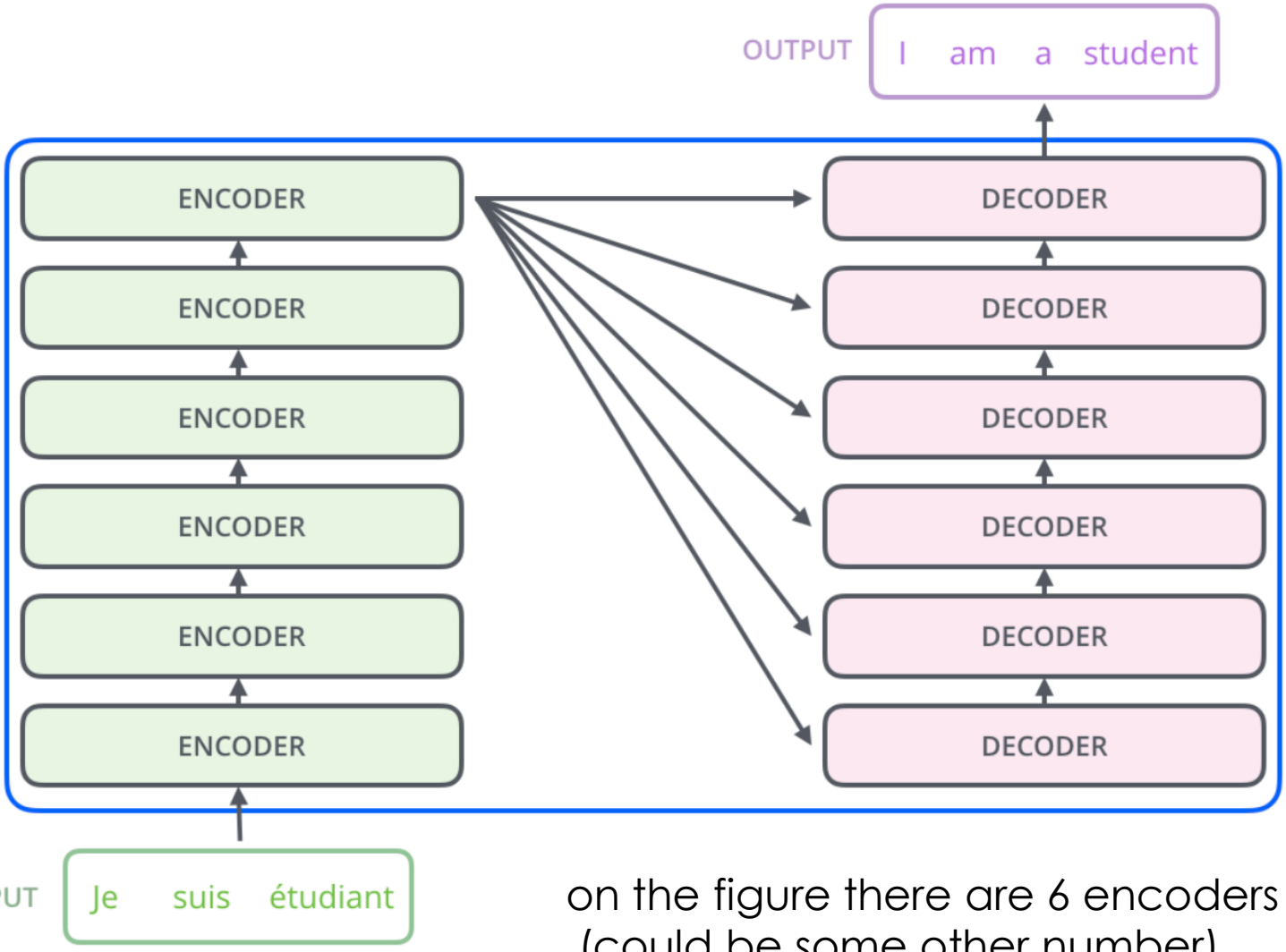


Transformer architecture of NNs

- currently the most successful DNN
- non-recurrent
- architecturally it is an encoder-decoder model
- fixed input length
- can be parallelized
- adapted for GPU (TPU) processing
- based on extreme use of attention

- https://github.com/dair-ai/Transformers-Recipe

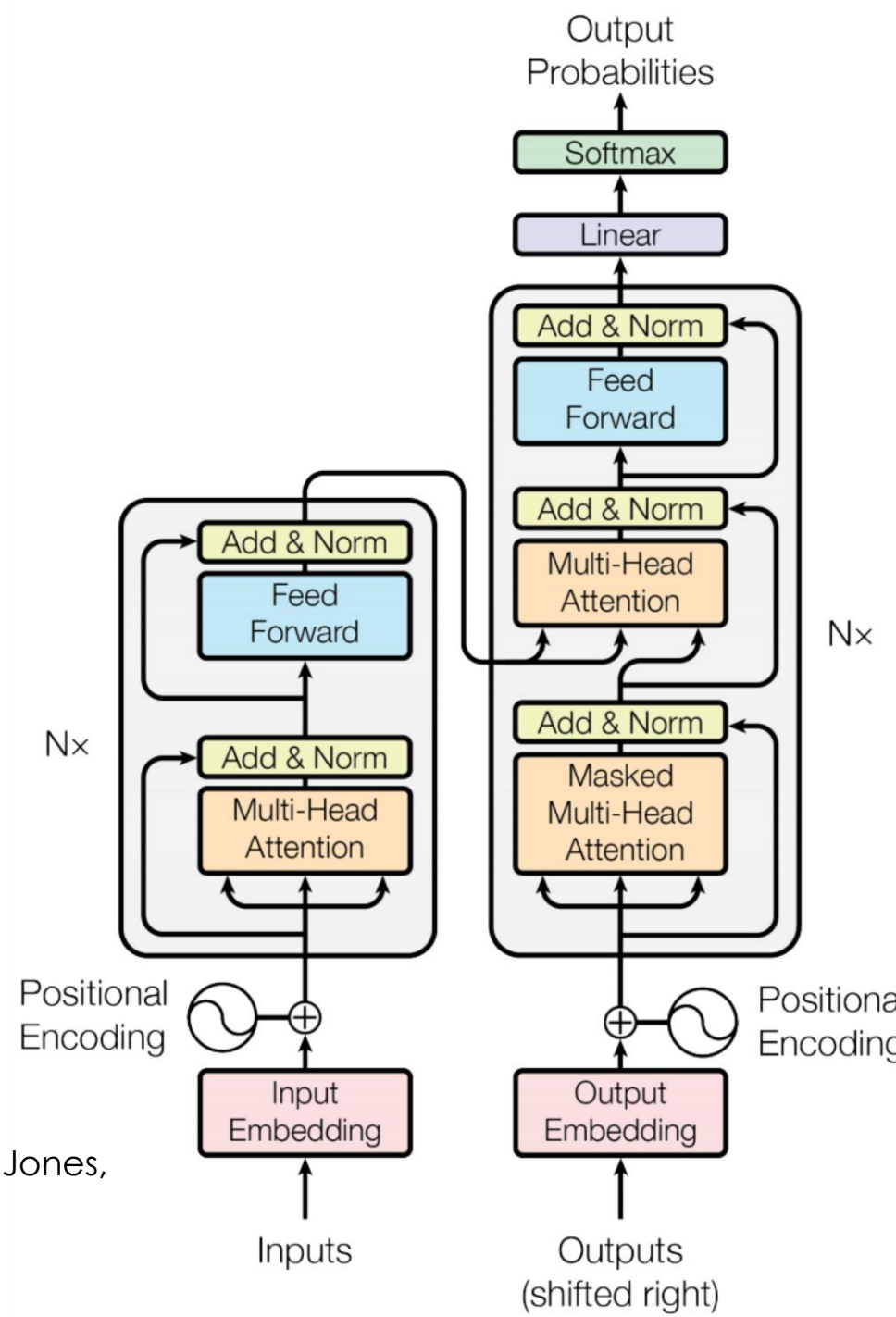
Transformer is an encoder-decoder model



on the figure there are 6 encoders and 6 decoders
(could be some other number)

Transformer overview

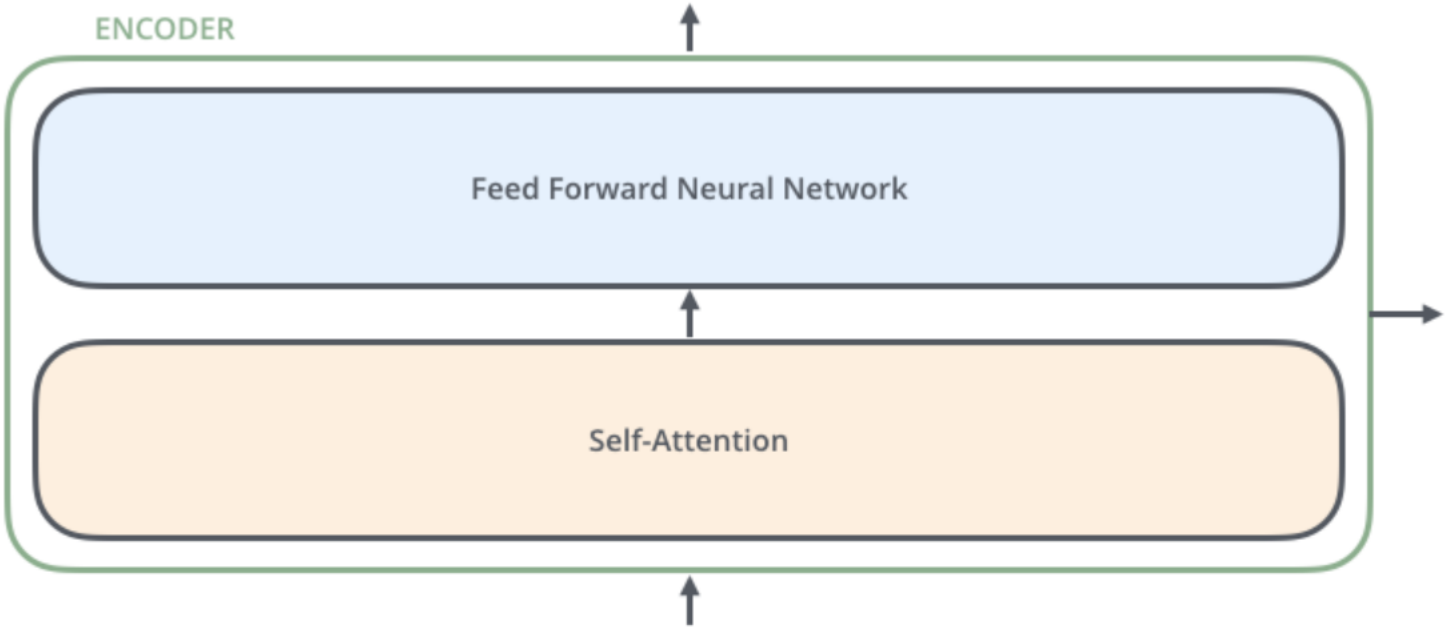
- Initial task: machine translation with parallel corpus
- Predict each translated word
- Final cost/loss/error function is standard cross-entropy error on top of a softmax classifier



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

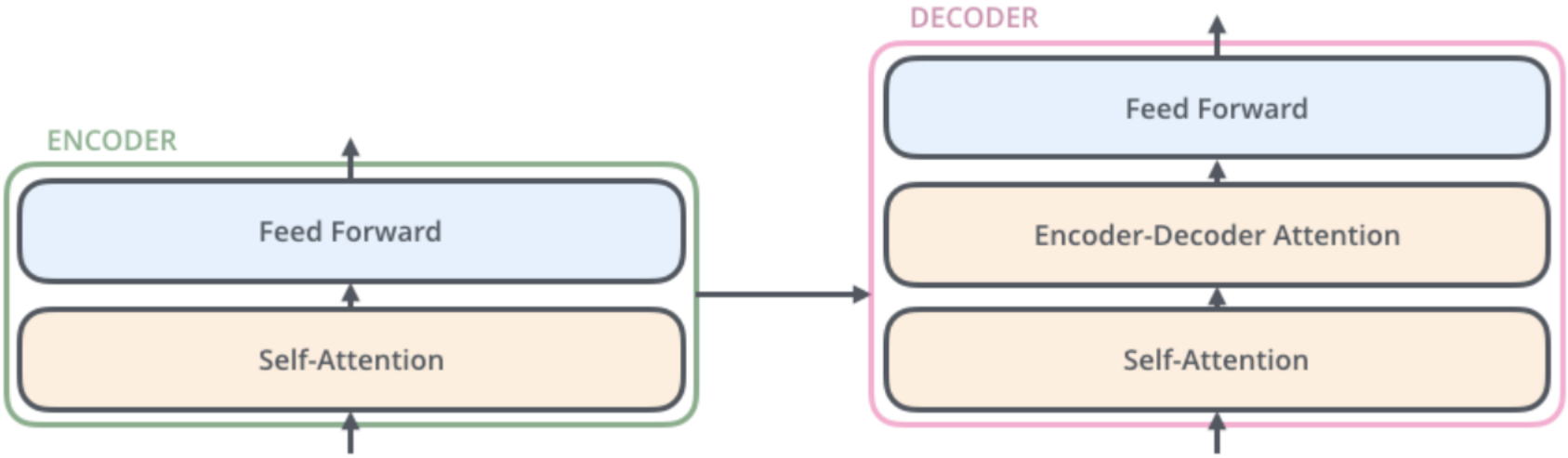
Transformer: encoder

- Two layers
- No weight sharing between different encoders
- Self-attention helps to focus on relevant part of input
- The illustration shows one encoder block

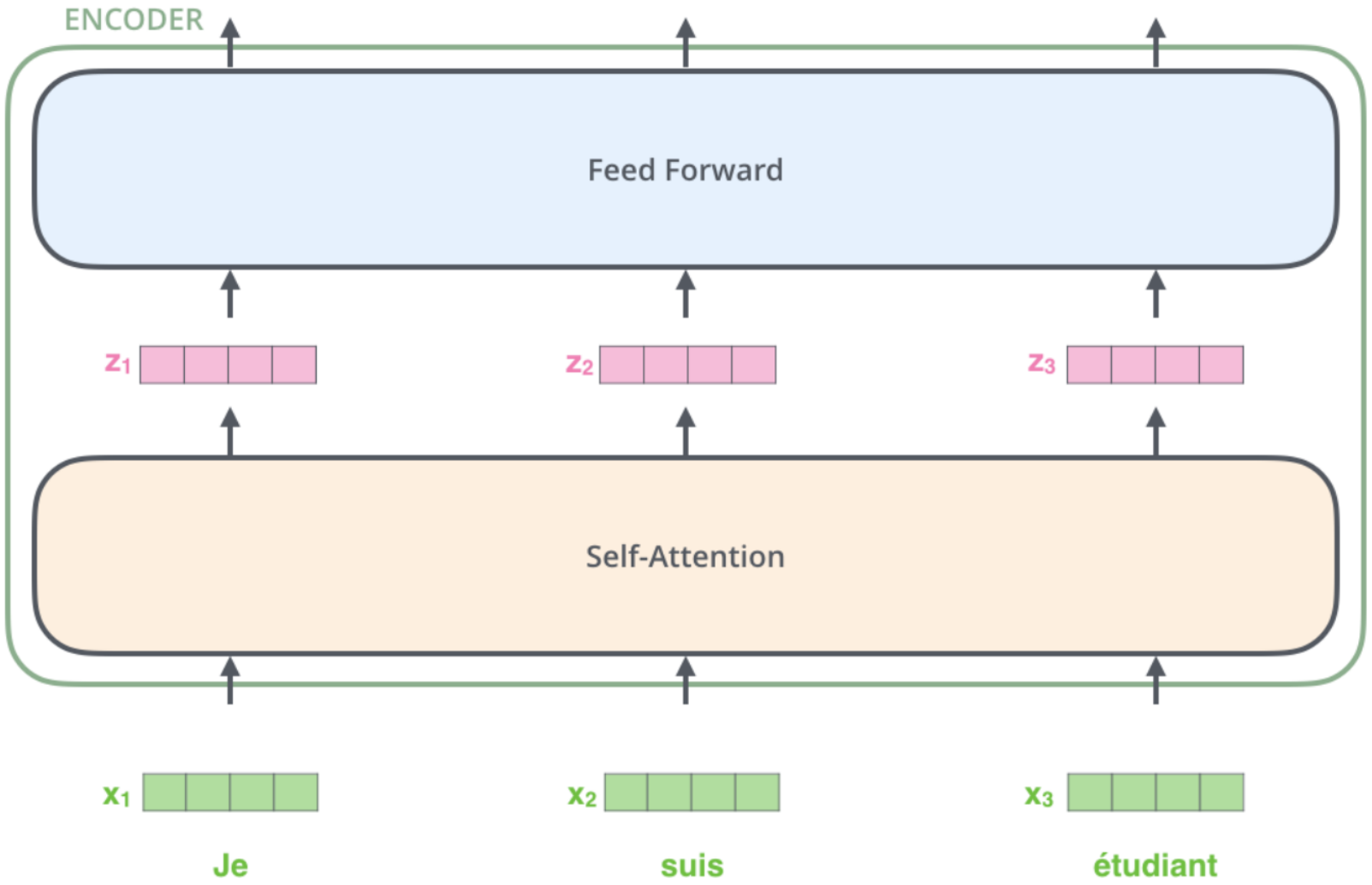


Transformer: decoder

- The same as encoder but with an additional attention layer in between, receiving input fom encoder
- Illustration shows one encoder and one decoder block (usually there are many more encoder and decoder blocks)



Start with embeddings

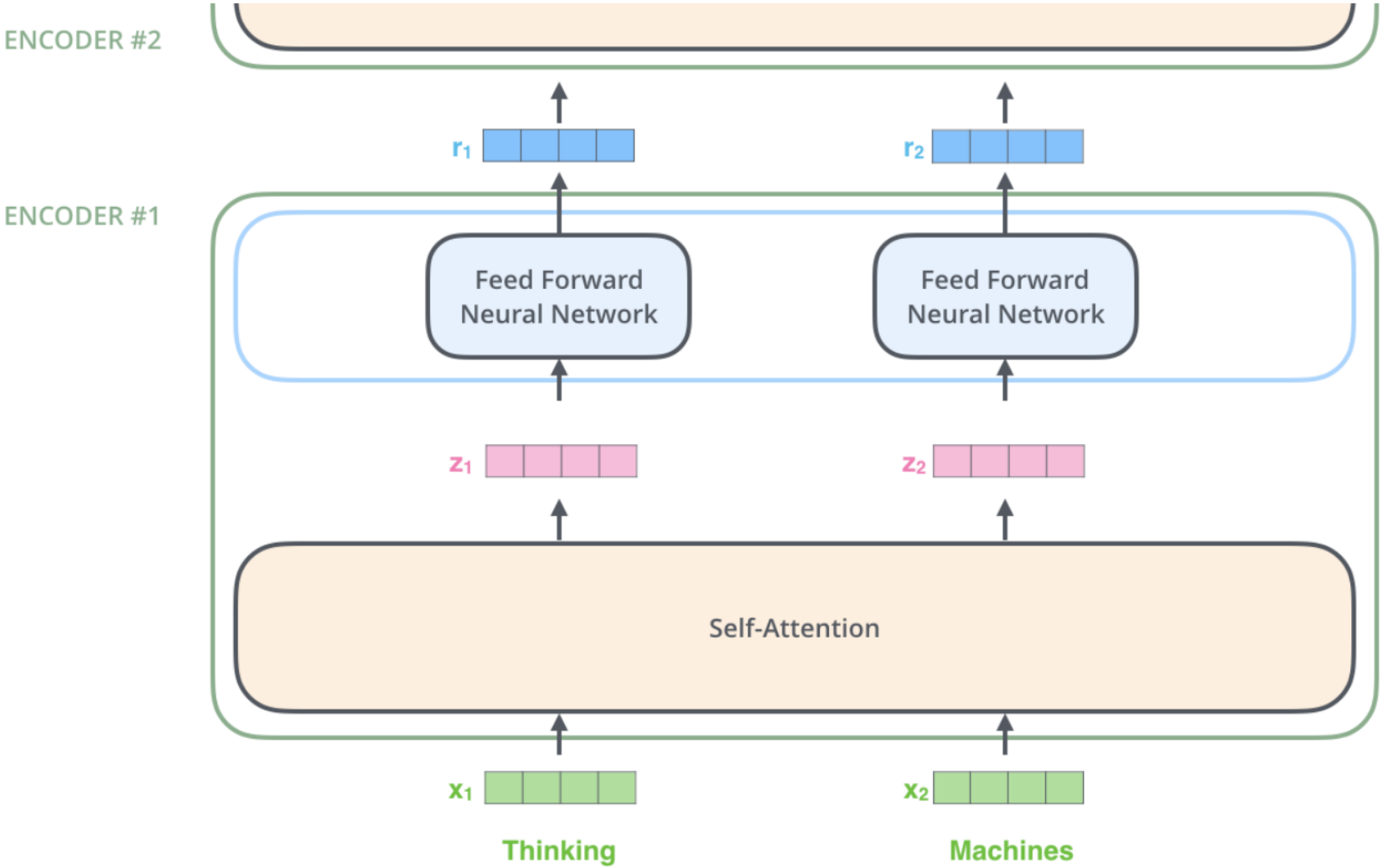


Input to transformer

- embeddings, e.g., 512 dimensional vectors (special, we will discuss that later)
- fixed length, e.g., 128 tokens (in modern architectures at least 8192)
- dependencies between inputs are only in the self-attention layer, no dependencies in feed-forward layer – good for parallelization

- Let us first present the working of the transformer with an illustration of the prediction

Encoding


ENCODER #2
r1
r2
ENCODER #1
Feed Forward Neural Network
Feed Forward Neural Network
z1
z2
Self-Attention
x1
x2
Thinking
Machines

Self-attention

- As the model processes each word (each position in the input sequence), self-attention allows it to look at other positions in the input sequence for clues that can help lead to a better encoding for this word

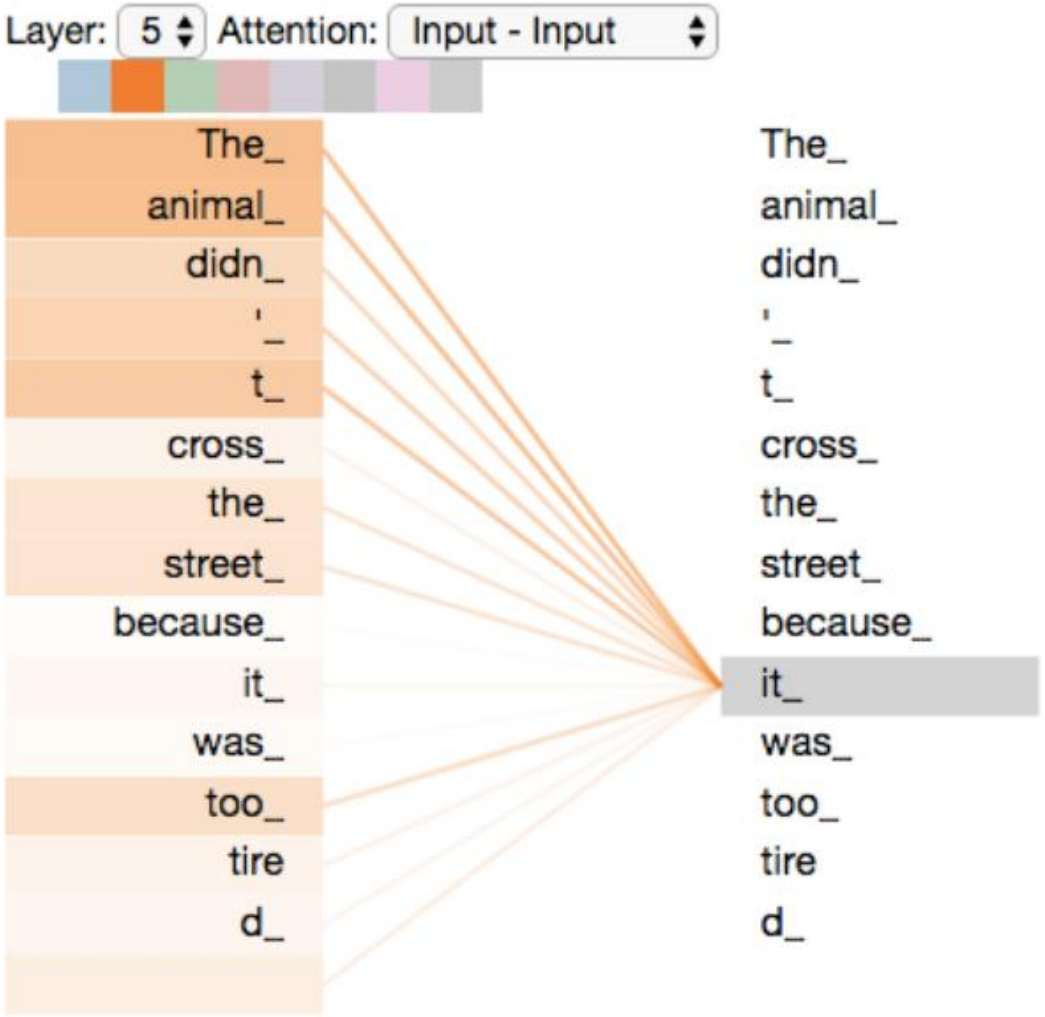
  ”The animal didn't cross the street because it was too tired”

- What does “it” in this sentence refer to? Is it referring to the street or to the animal? It’s a simple question to a human, but not as simple to an algorithm.
- *”The animal didn't cross the street because it was too wide”*

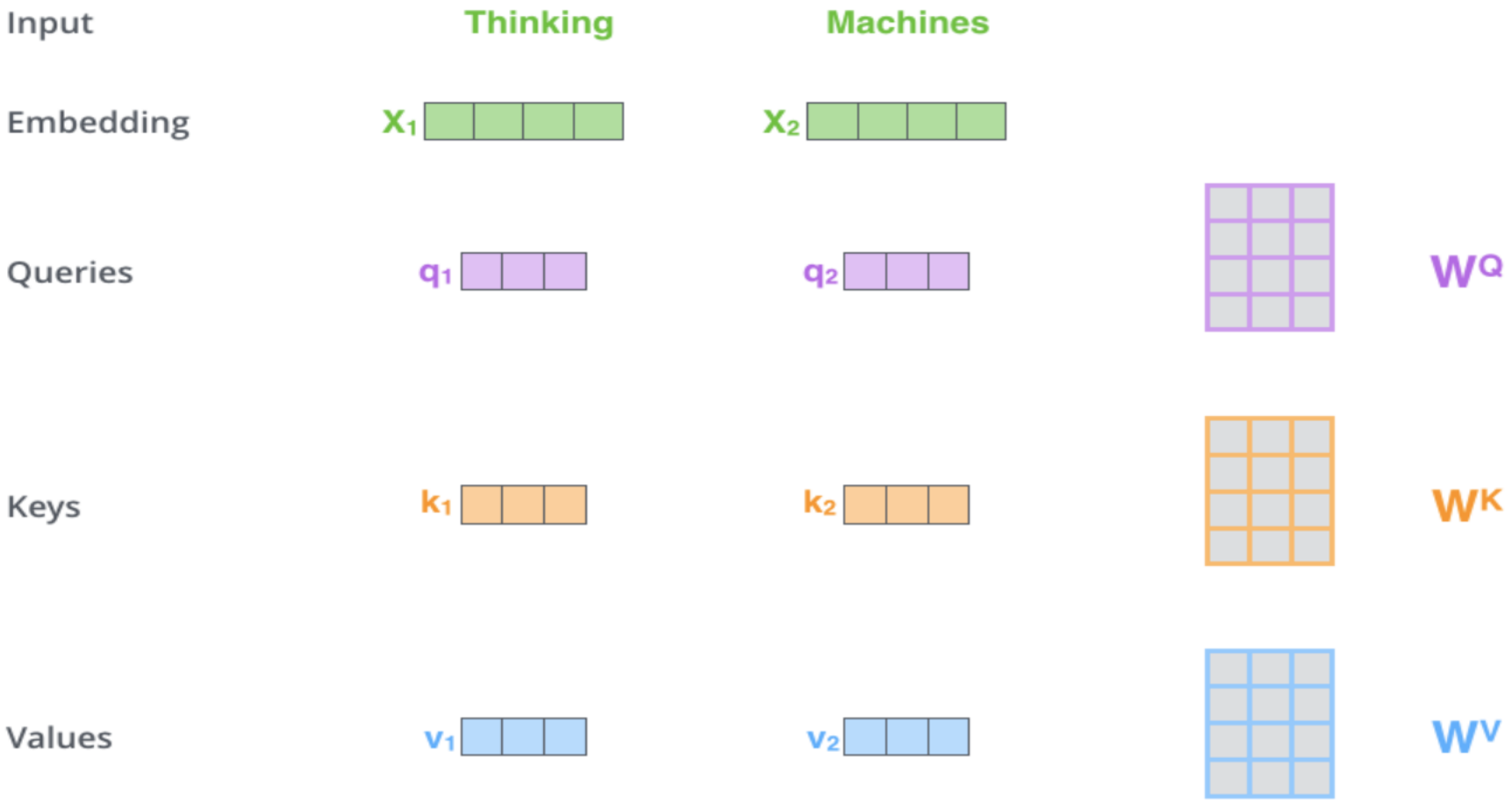
- When the model is processing the word “it”, self-attention allows it to associate “it” with “animal”.

Illustrating self-attention

- As we are encoding the word "it" in encoder #5 (the top encoder in the stack), part of the attention mechanism was focusing on "The Animal", and baked a part of its representation into the encoding of "it".

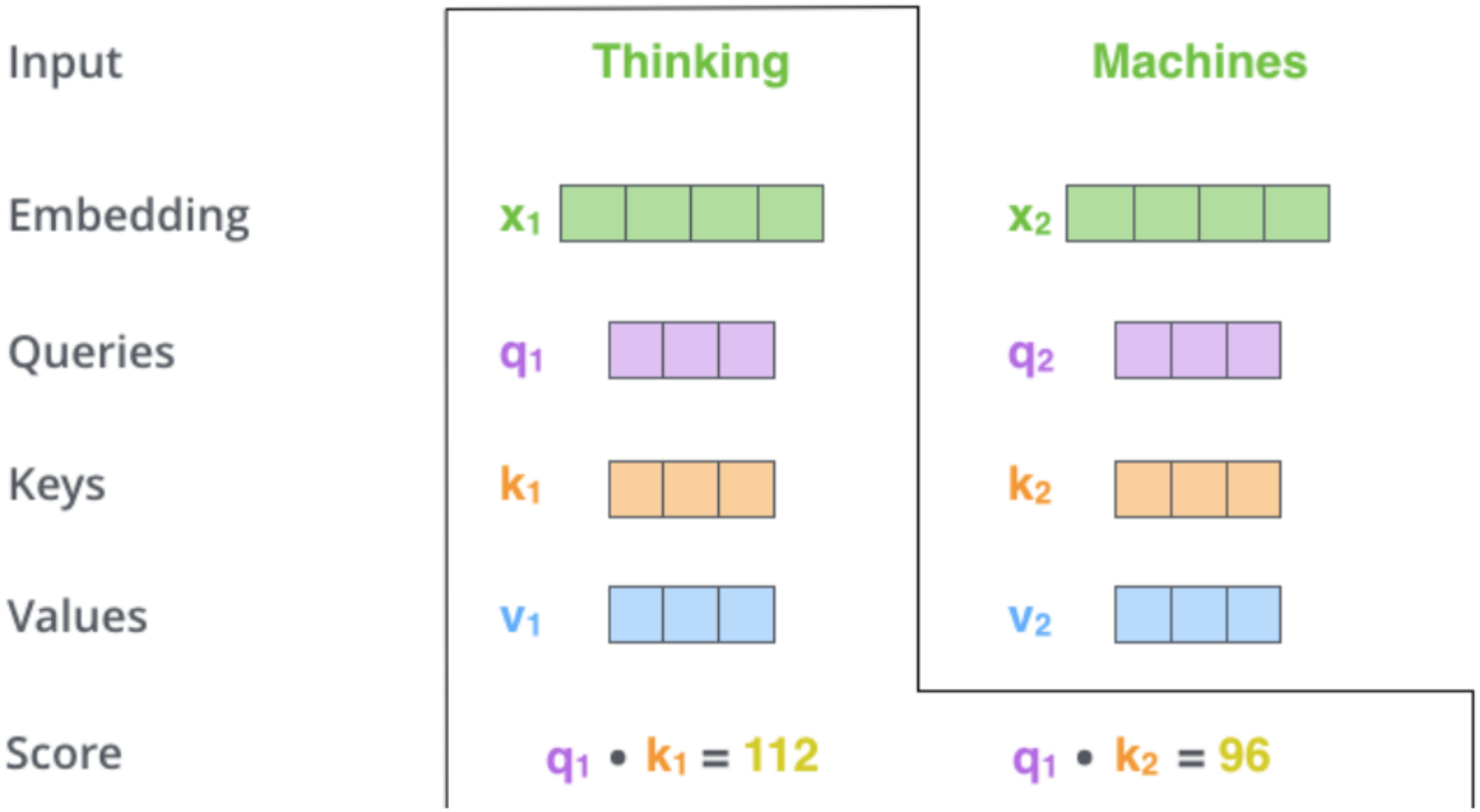


Self-attention details 1/4

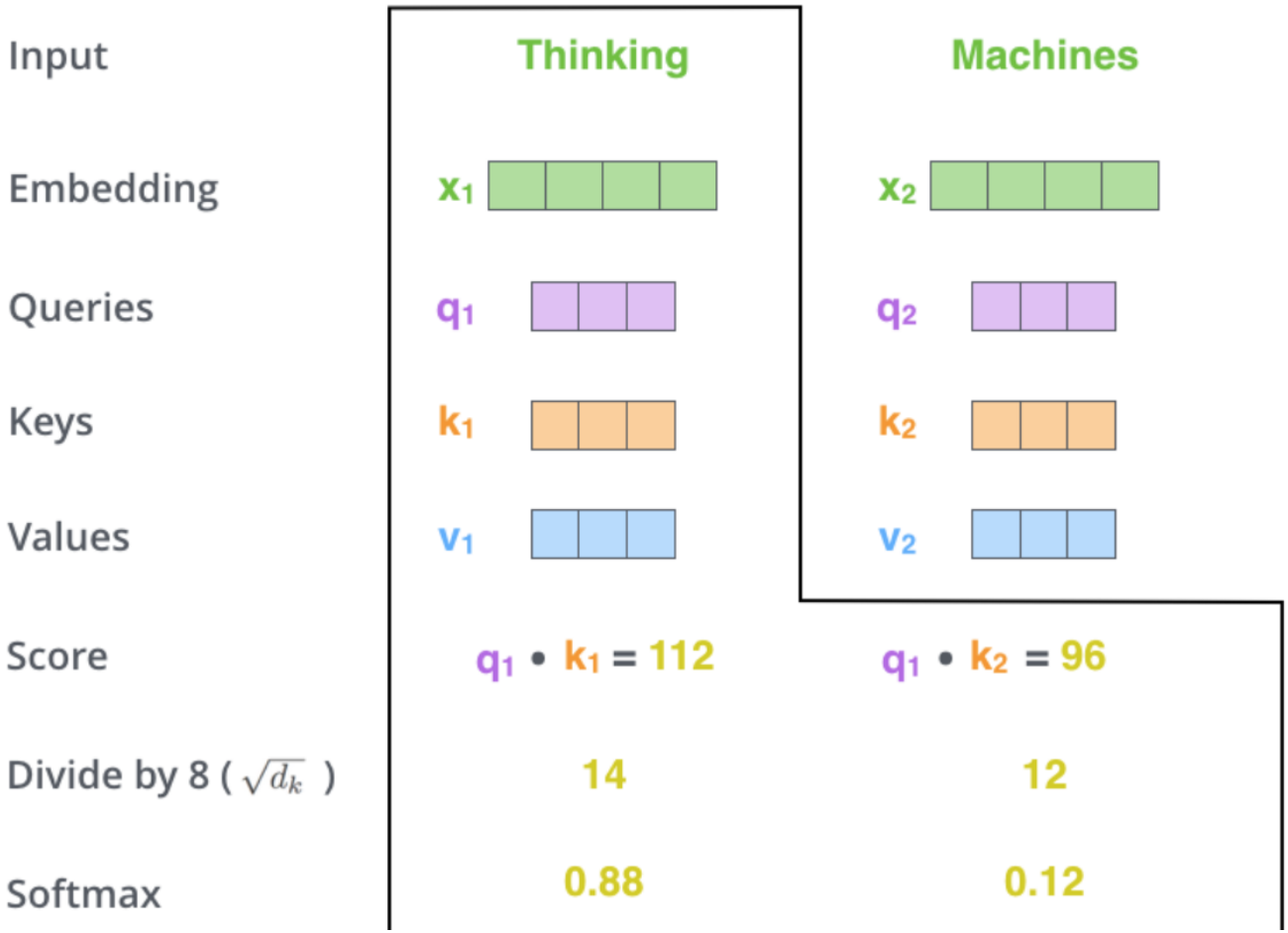


Multiplying x_1 by the W^Q weight matrix produces q_1, the "query" vector associated with that word. We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.

Details 2/4: scoring



Details 3/4: normalization of scores



Details 4/4: self-attention output

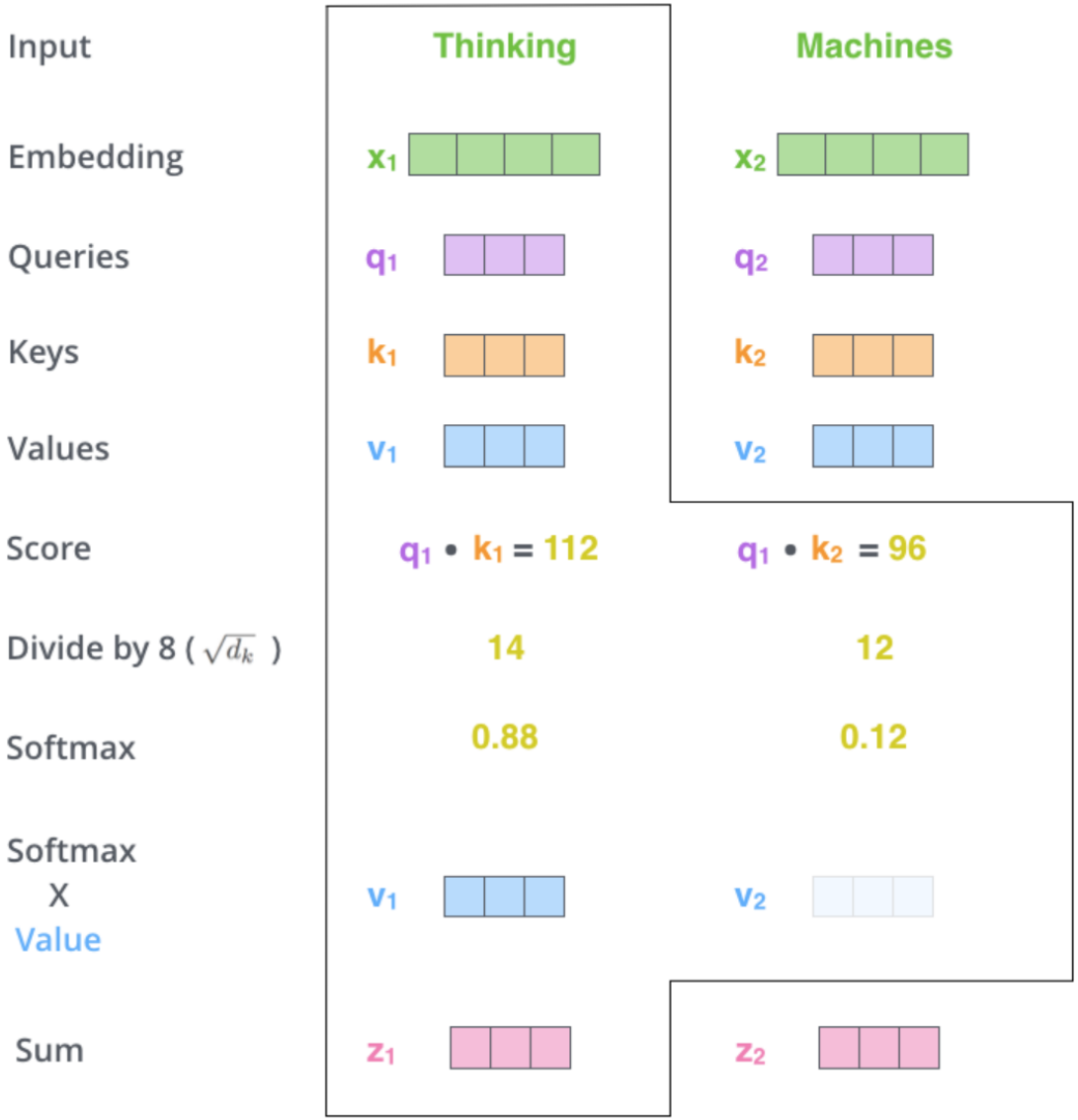
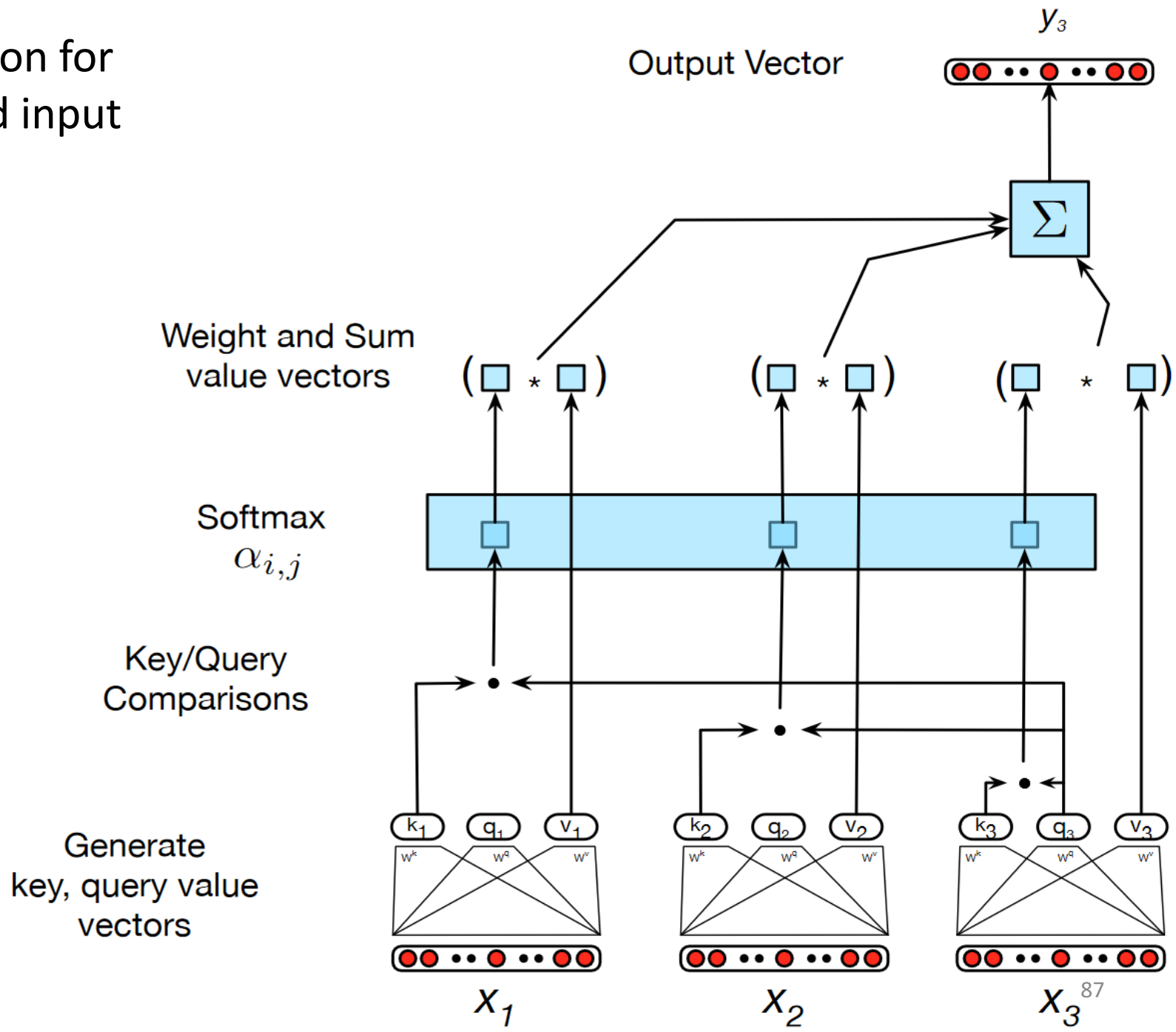
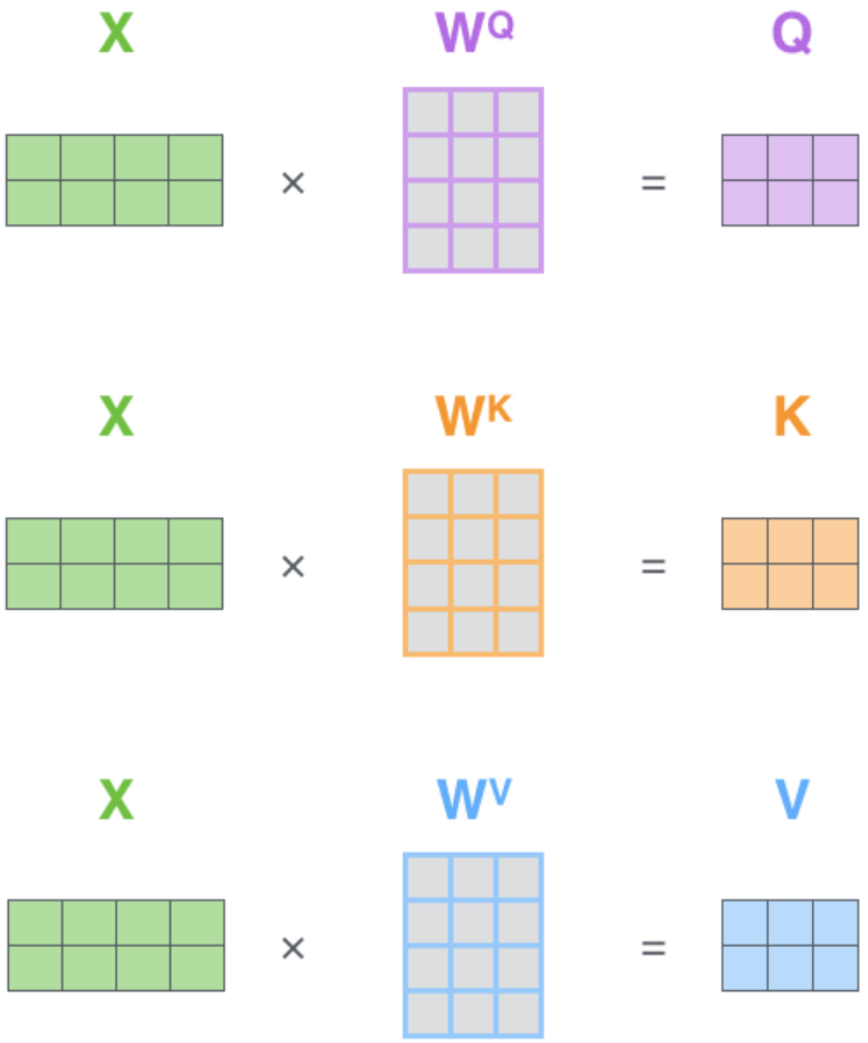


illustration for
the third input



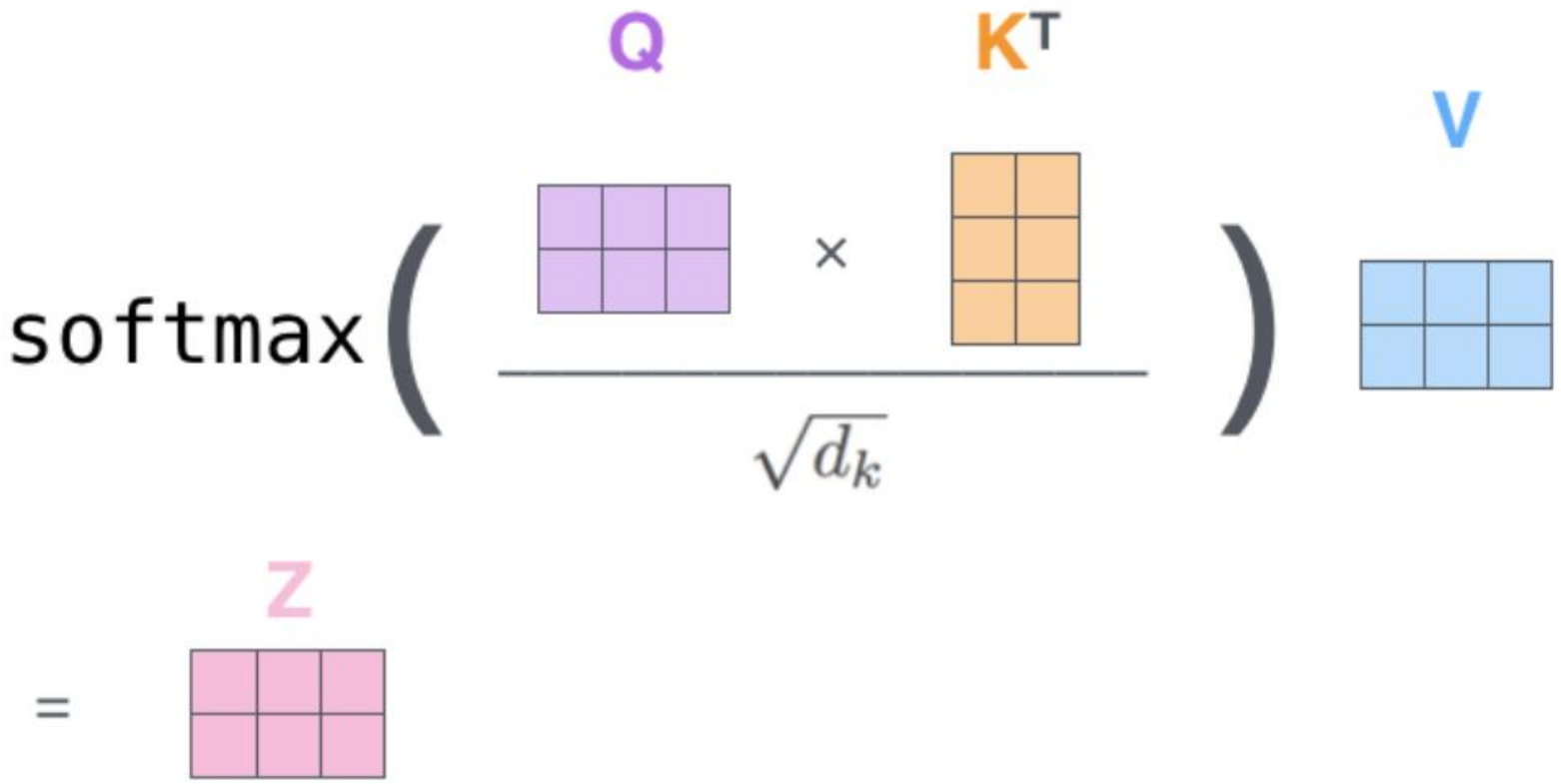
Matrix calculation of self-attention 1/2

Every row in the X matrix corresponds to a word in the input sentence.
The embedding vector x (512) is larger then the q/k/v vectors (64)

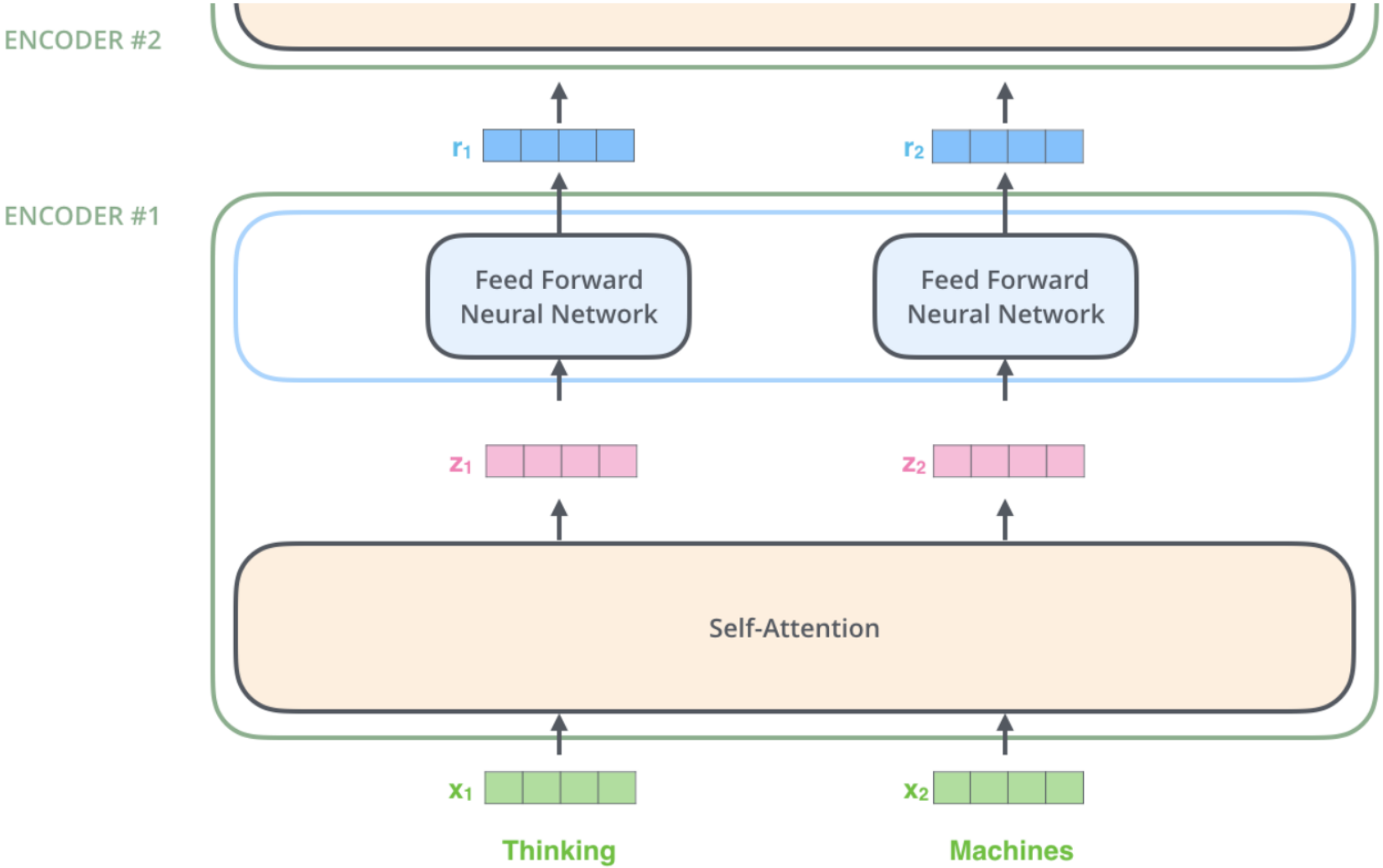


Matrix calculation of self-attention 2/2

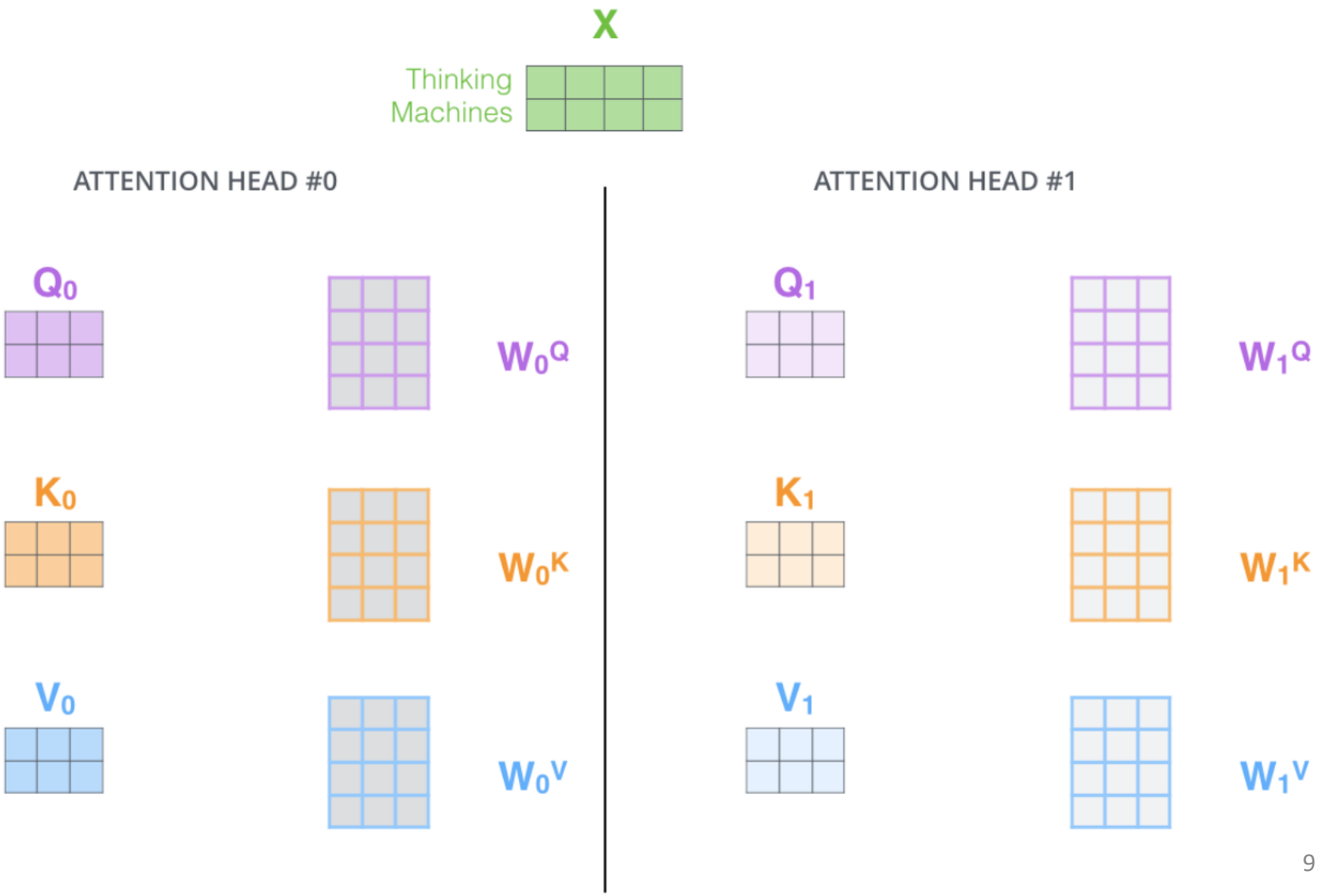
- final calculation



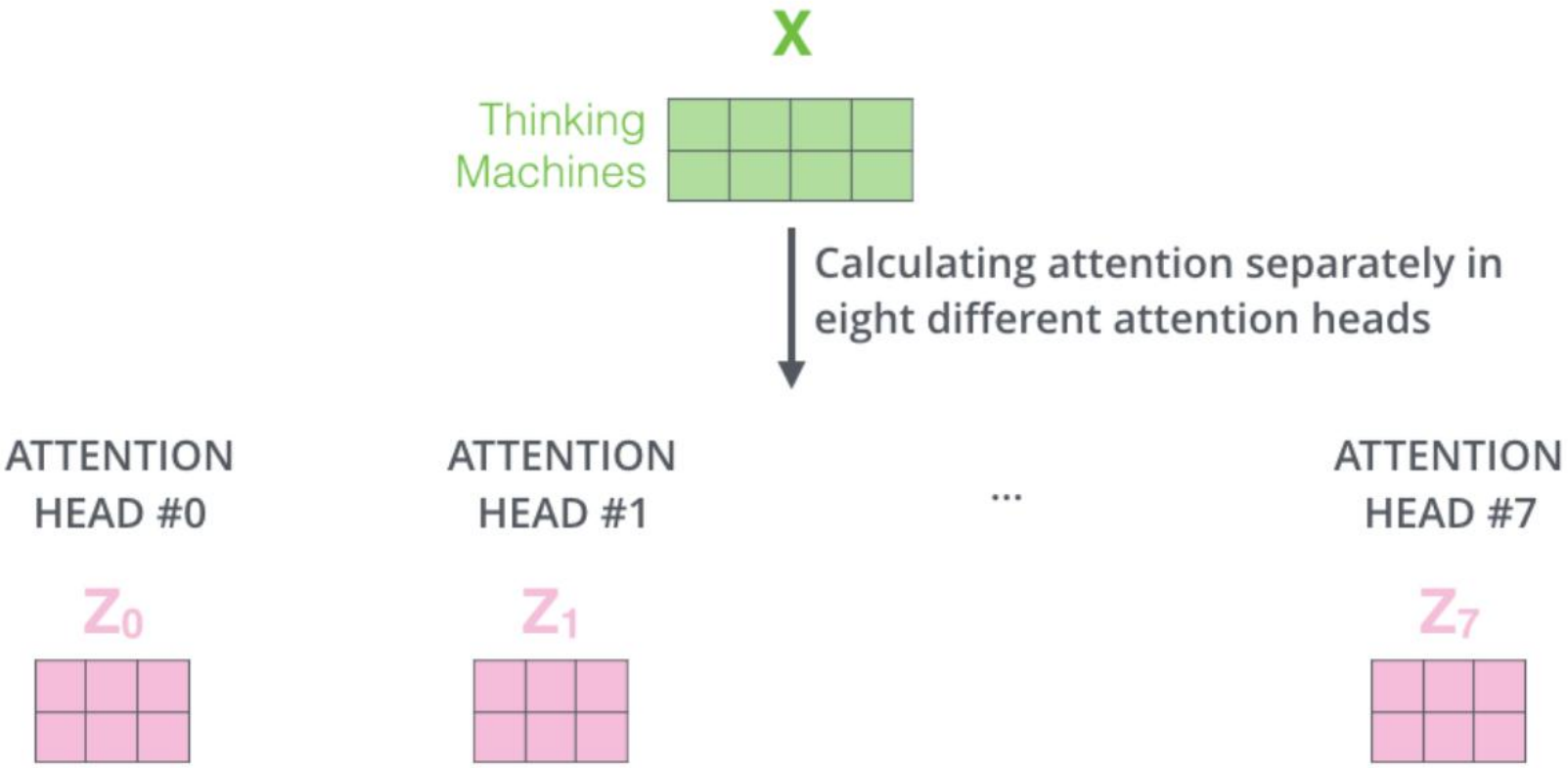
Encoding


ENCODER #2
ENCODER #1
r1
r2
Feed Forward Neural Network
Feed Forward Neural Network
z1
z2
Self-Attention
x1
x2
Thinking
Machines

Example: two attention heads



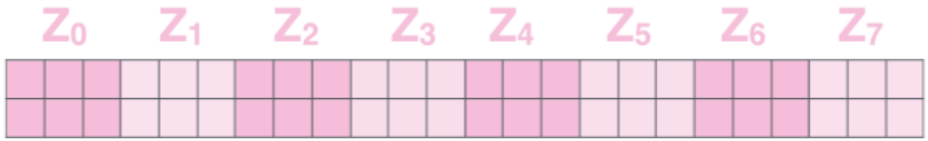
Example: 8 att. heads



- What to do with 8 Z matrices, the feed-forward layer is expecting a single matrix (one vector for each word). We need to condense all attention heads into one matrix.

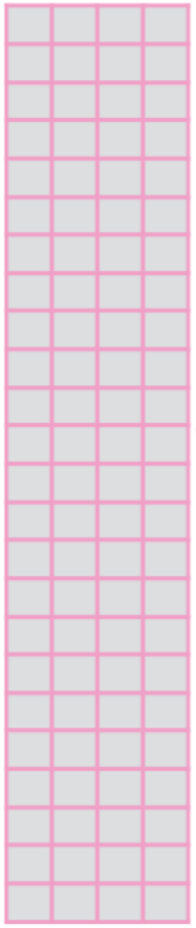
Condensation of attention heads

1) Concatenate all the attention heads



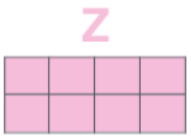
2) Multiply with a weight matrix W^O that was trained jointly with the model

X



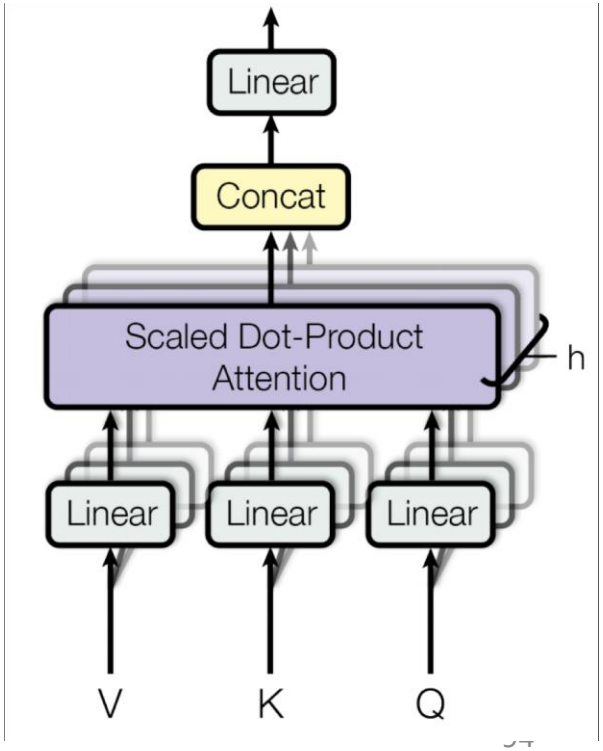
3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN

=



Computing multi-head attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Summary of self-attention

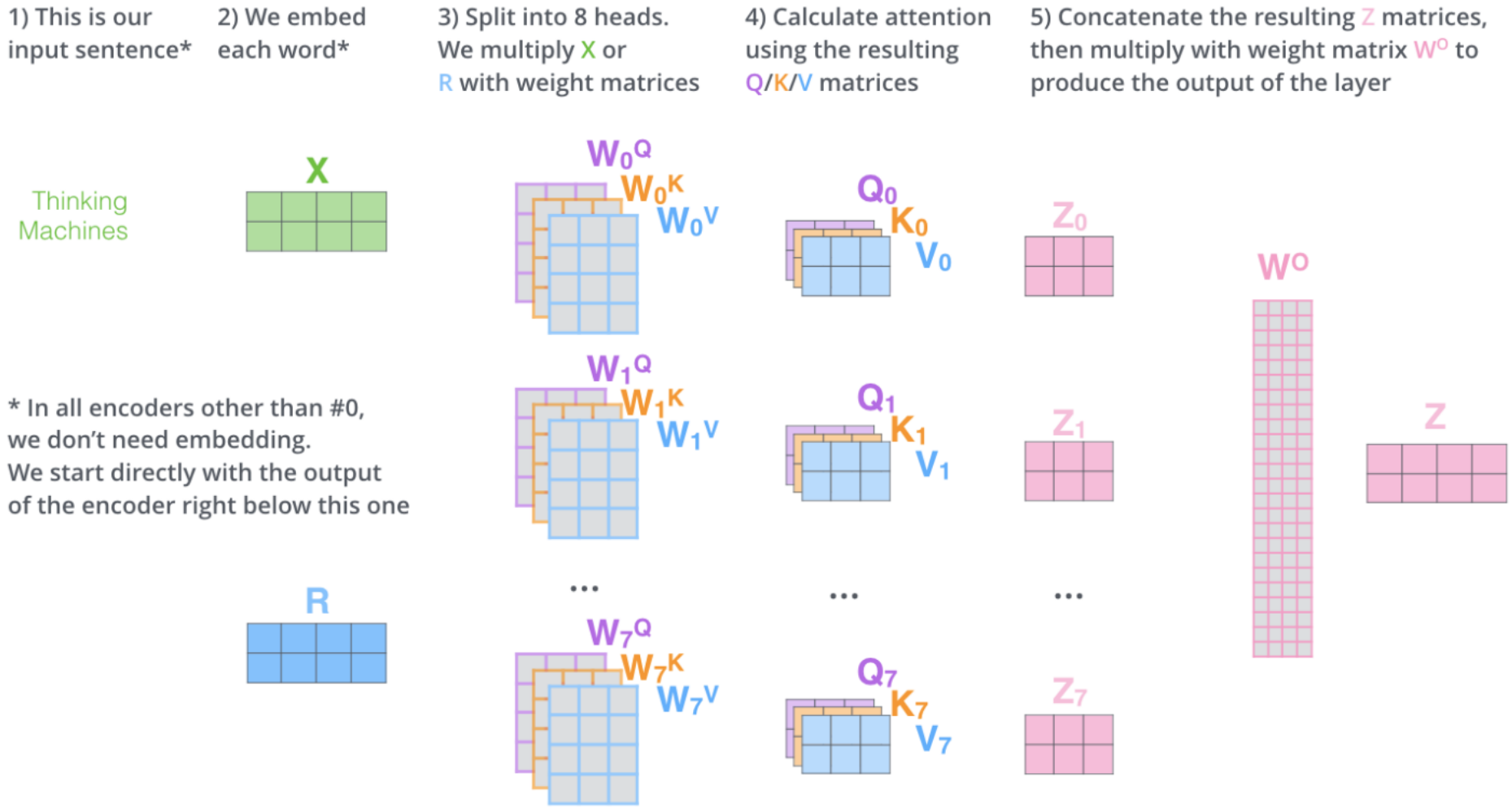


Illustration of self-attention: 1 head

- encoder #5 (the top encoder in the stack)
- As we encode the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired" -- in a sense, the model's representation of the word "it" bakes in some of the representation of both "animal" and "tired".

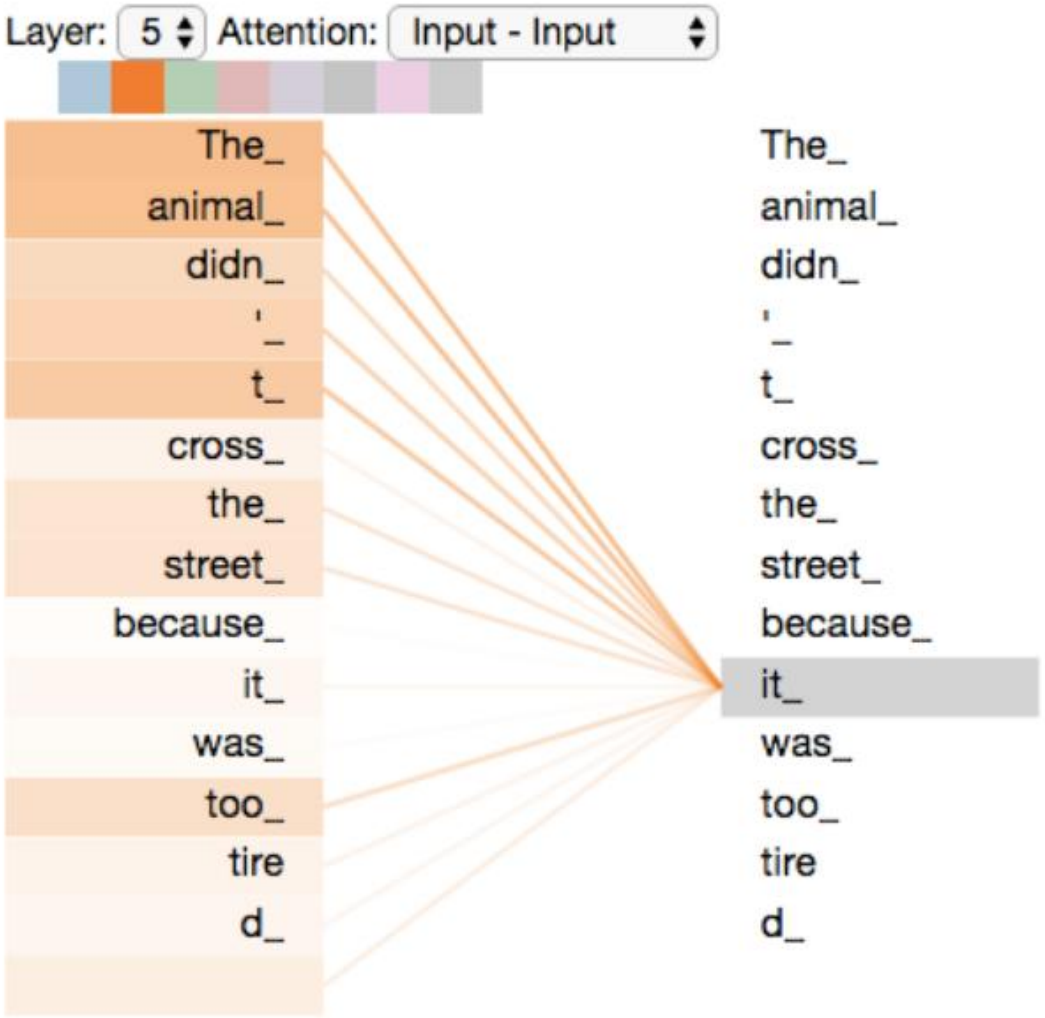
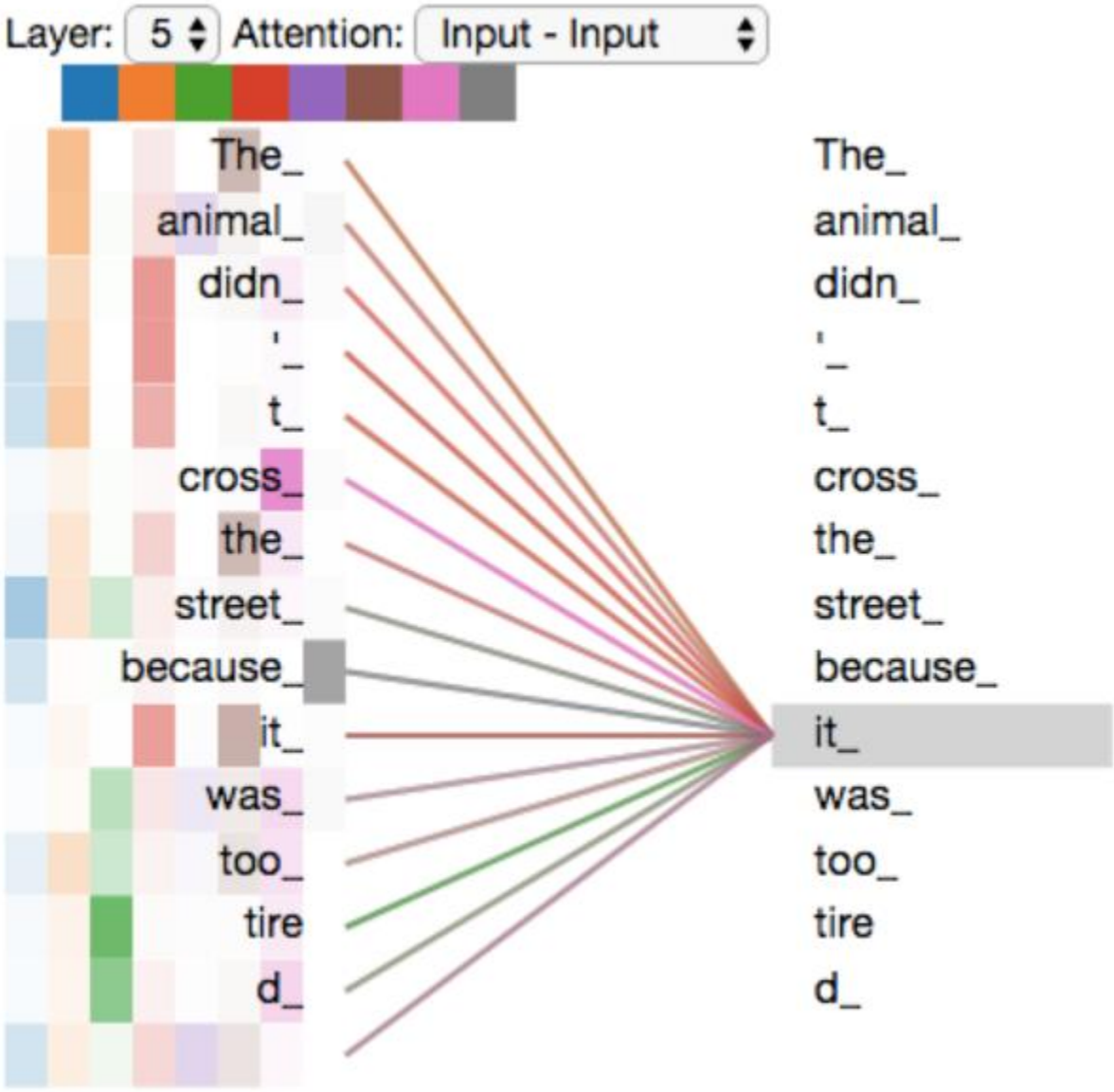
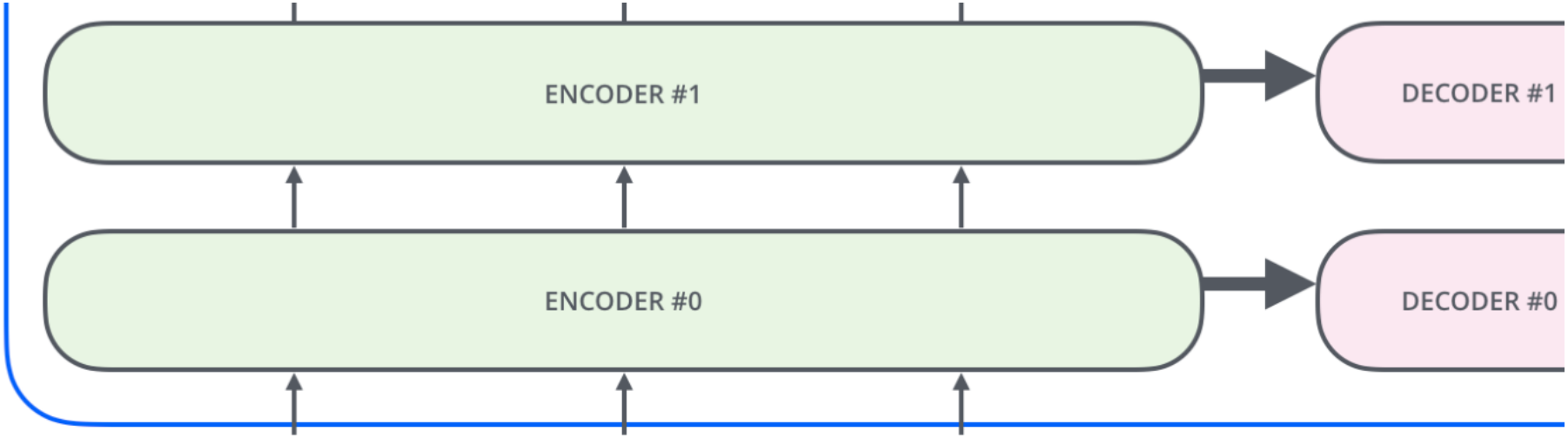


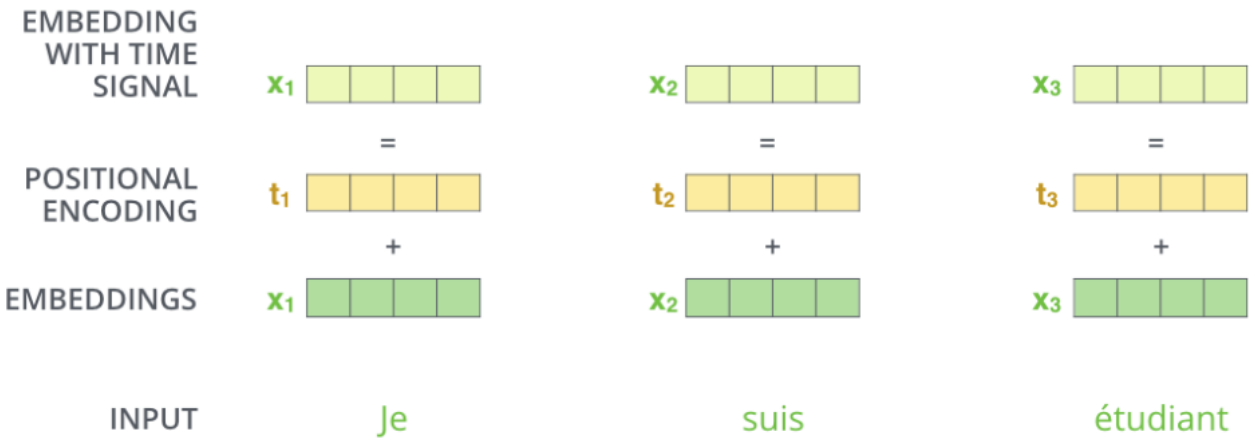
Illustration of self-attention: all heads

- all the attention heads in one picture are harder to interpret



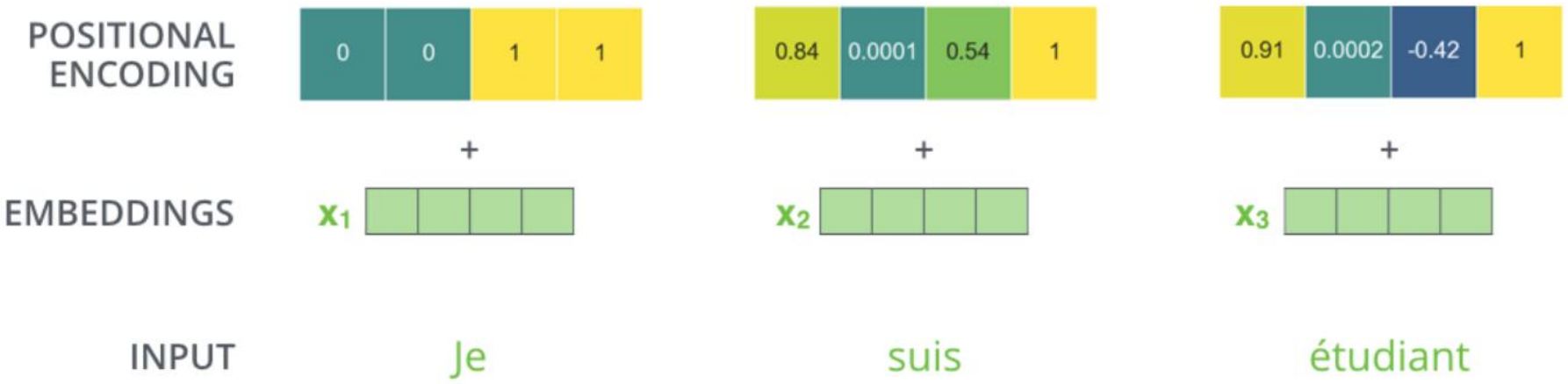
Adding position encoding



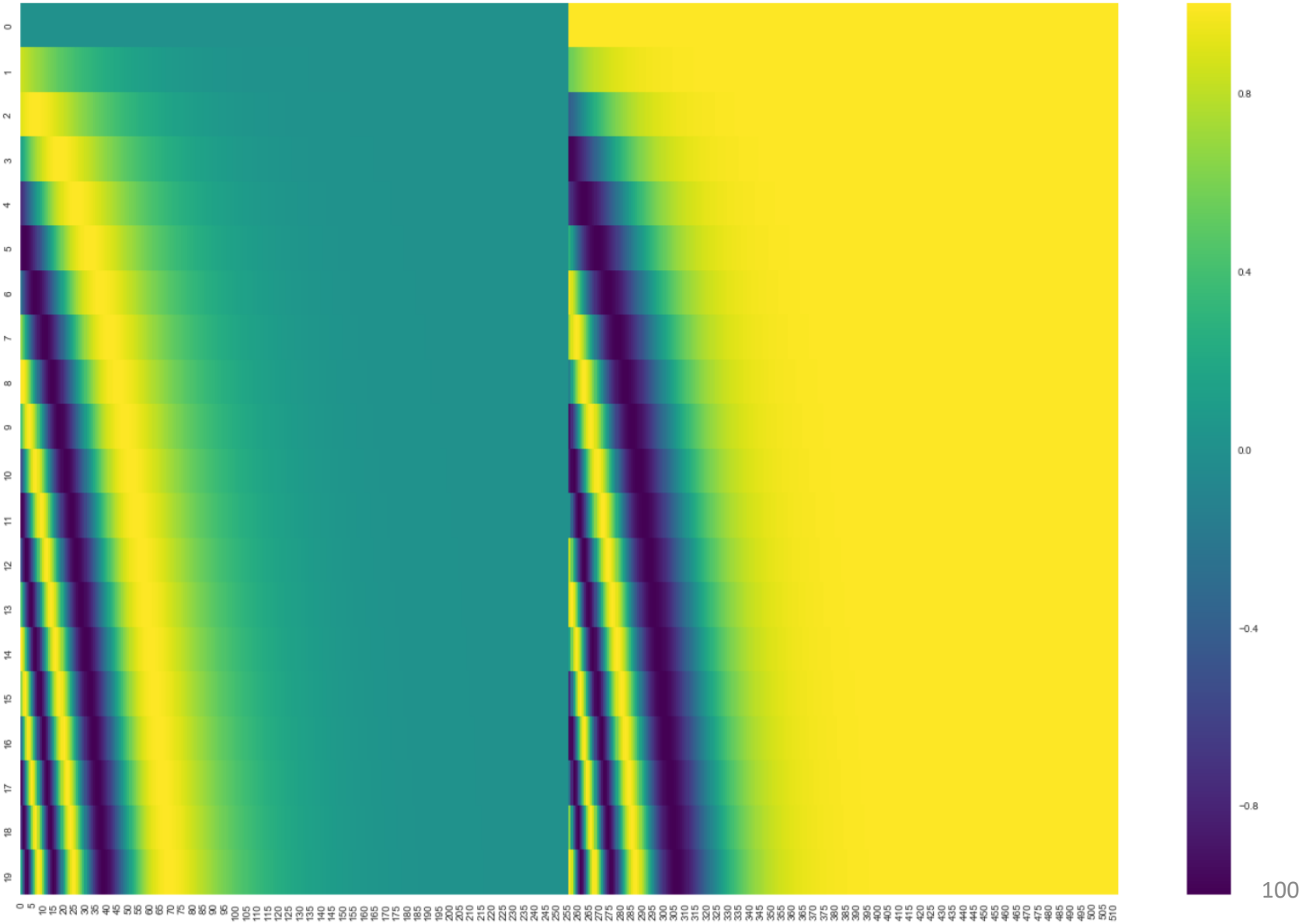


Example: encoding position

- the values of positional encoding vectors follow a specific pattern.



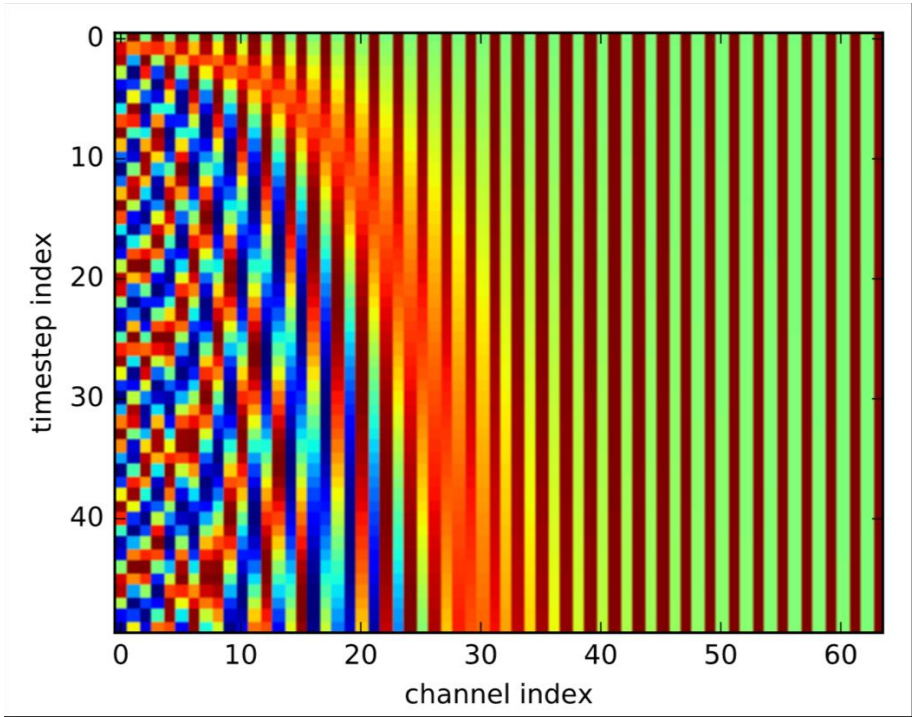
Example of positional encoding



Another positional encoding

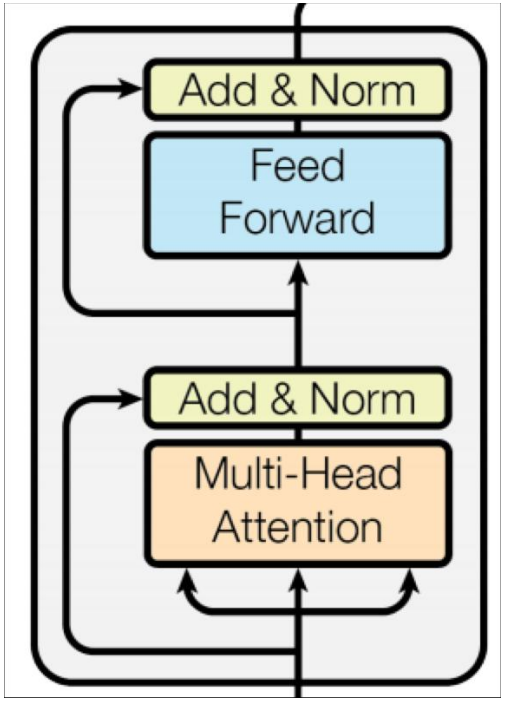
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

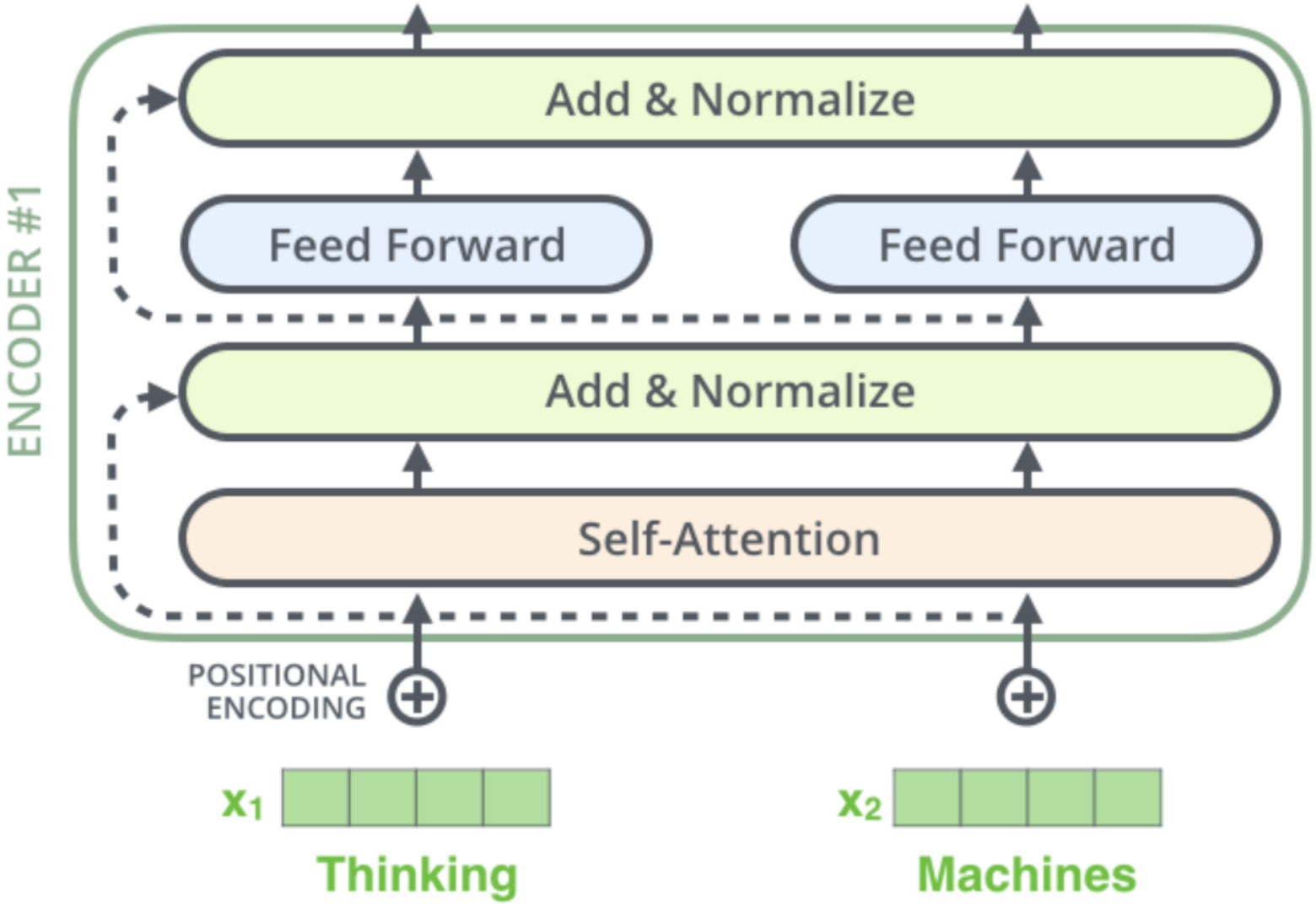


Encoder blocks

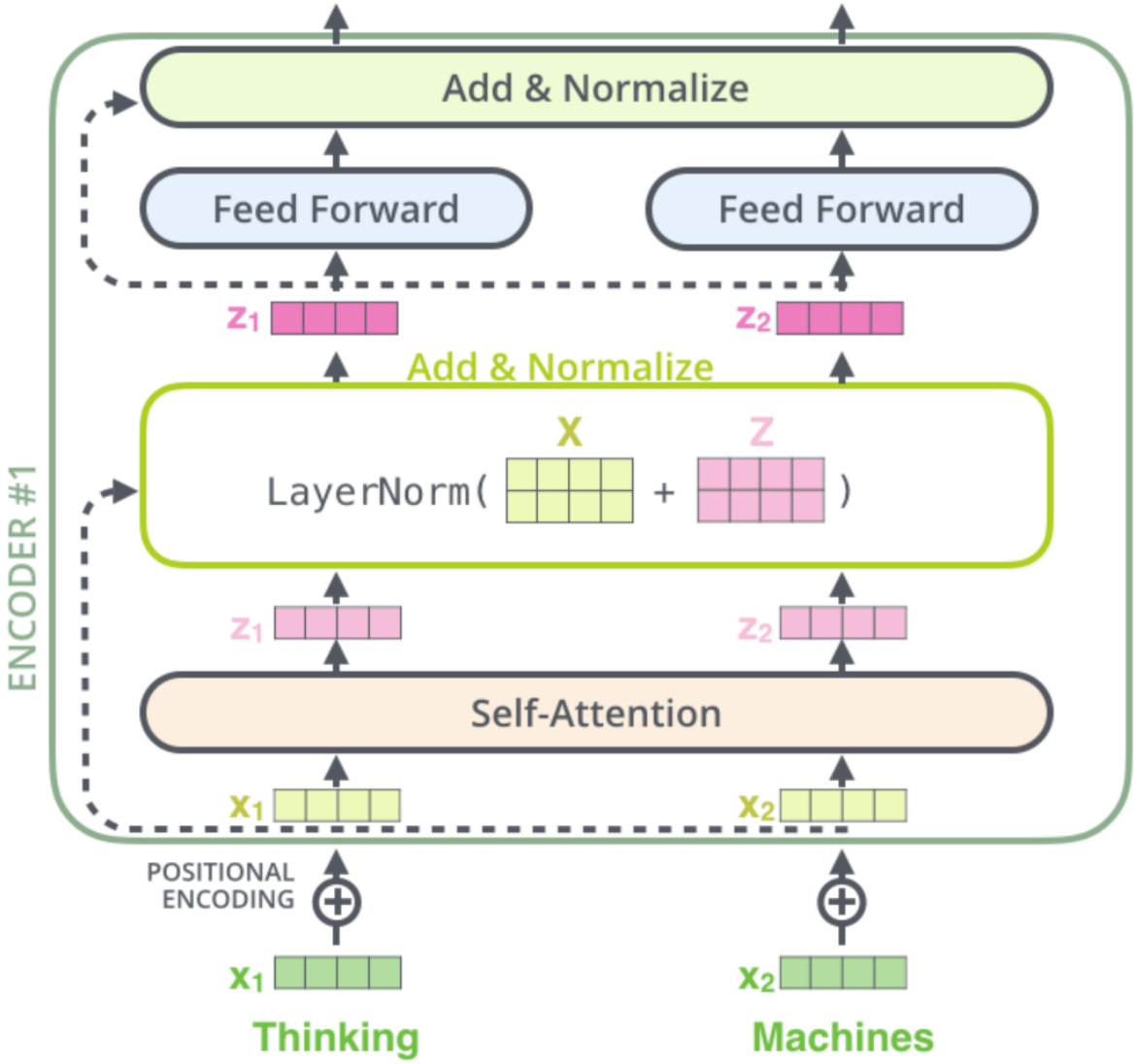
- Each block has two “sublayers”
  - Multihead attention
  - 2-layer feed-forward neural network (with ReLU)
- Each of these two steps also has a residual (short-circuit) connection and LayerNorm, i.e.:
  - LayerNorm(x + Sublayer(x))



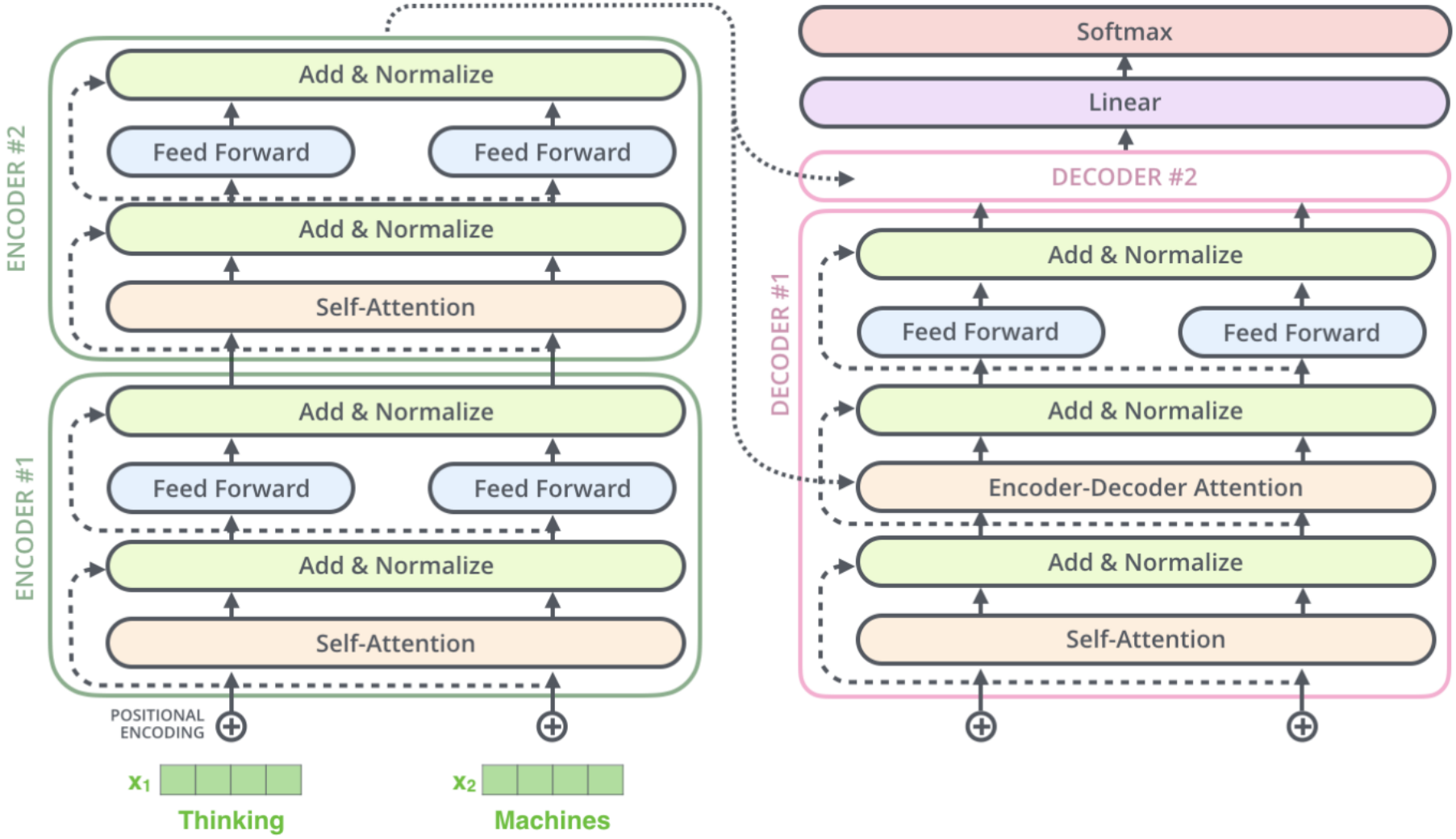
Architecture with residual connection – top level view



Architecture with residual connection – example

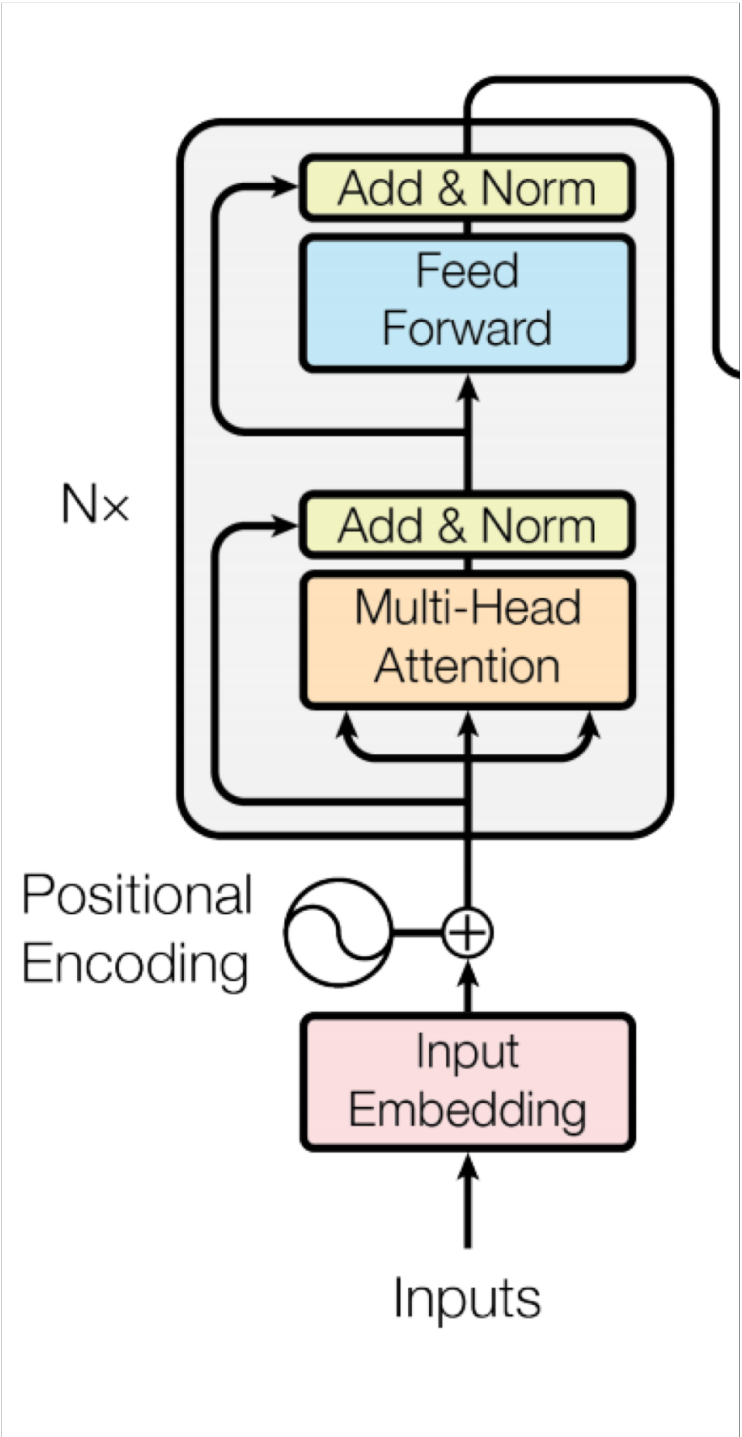


Example: 2 stacked transformer



Complete encoder

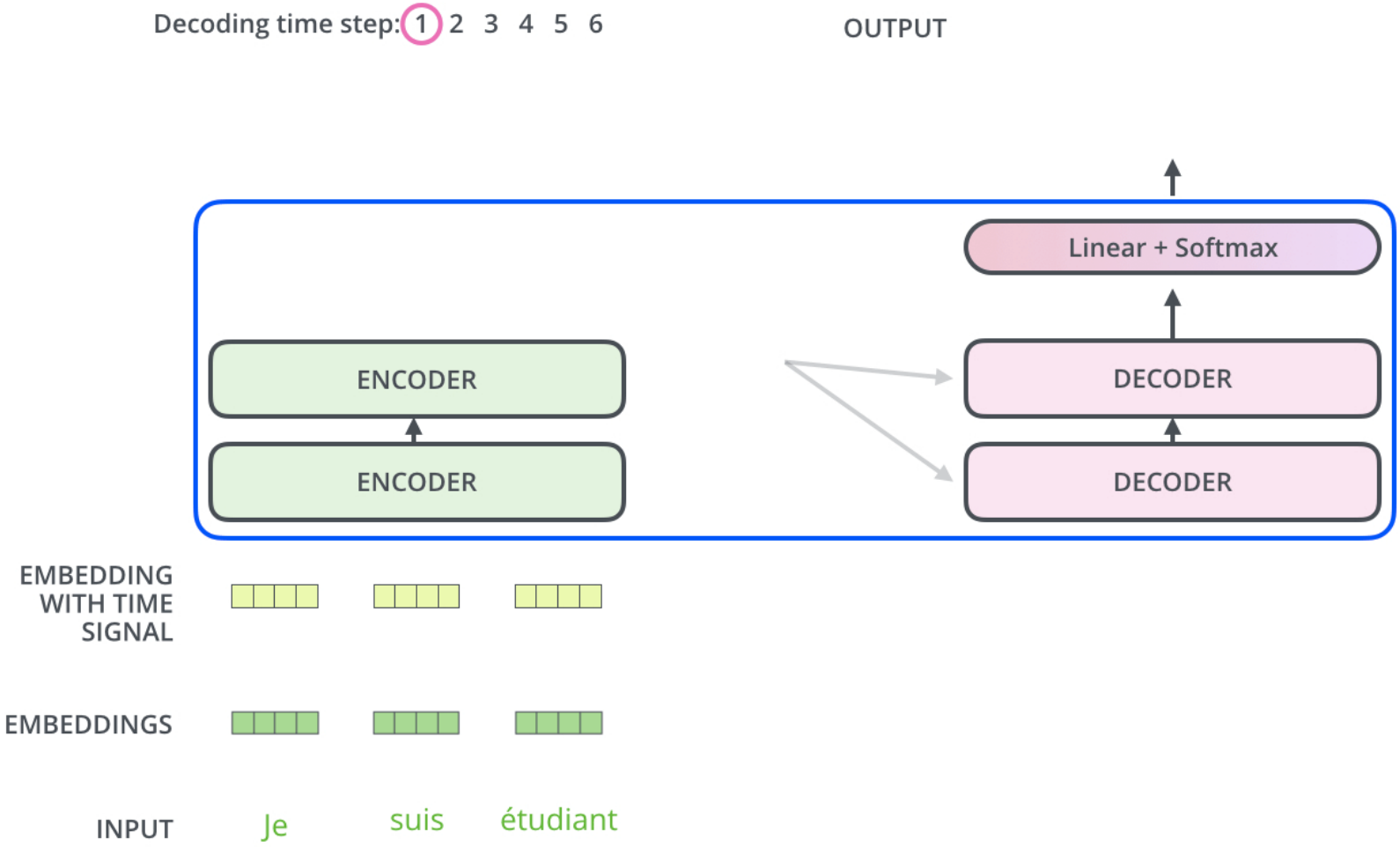
- each encoder block is repeated several times, e.g., 6 times



Decoder

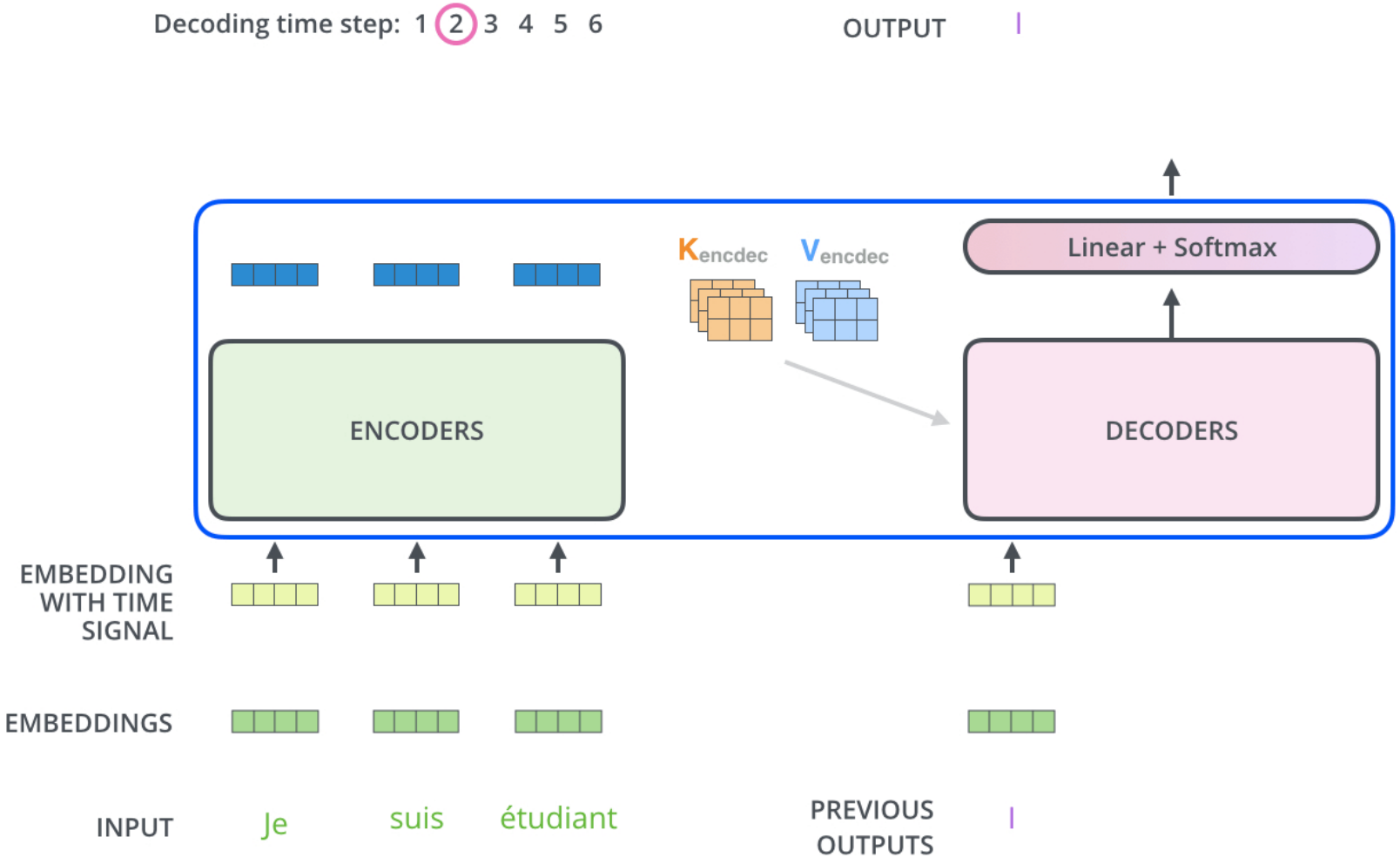
- Decoders have the same components as encoders
- An encoder starts by processing the input sequence.
- The output of the top encoder is transformed into a set of attention vectors K and V.
- These are used by each decoder in its “encoder-decoder attention” layer which helps the decoder to focus on appropriate places in the input sequence.

Encoder-decoder in action 1/2



After finishing the encoding phase, we begin the decoding phase. Each step in the decoding phase outputs an element from the output sequence (the English translation sentence in this case).

Encoder-decoder in action 2/2



The steps repeat until a special symbol indicating the end of output is generated. The output of each step is fed to the bottom decoder in the next time step. We add positional encoding to decoder inputs to indicate the position of each word.

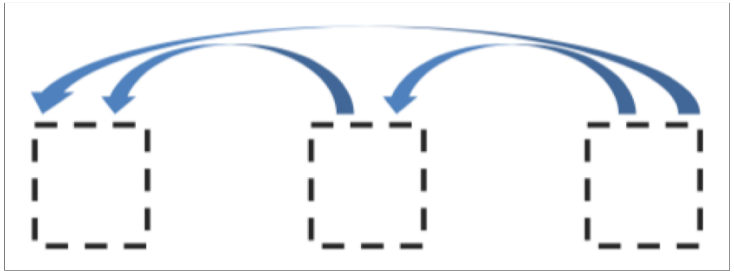
Self-attention and encoder-decoder attention in the decoder

- In the decoder, the self-attention layer is only allowed to attend to itself and earlier positions in the output sequence (to maintain the autoregressive property).
- This is done by masking future positions (setting them to -inf) before the softmax step in the self-attention calculation.

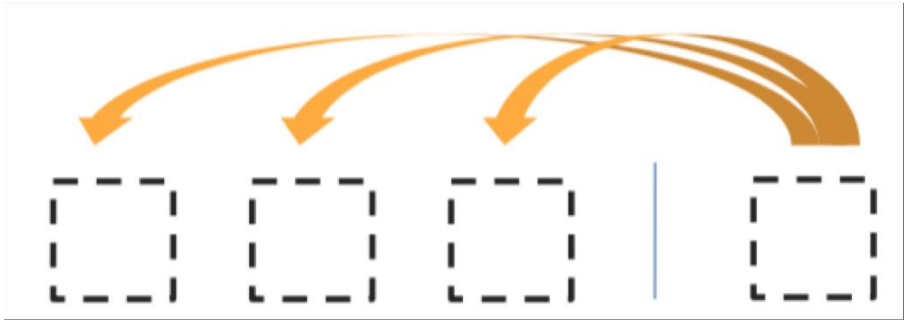
- The “Encoder-Decoder Attention” layer works just like multiheaded self-attention, except it creates its Q (queries) matrix from the layer below it, and takes the K (keys) and V (values) matrix from the output of the encoder stack.

Attentions in the decoder

1. Masked decoder self-attention on previously generated outputs

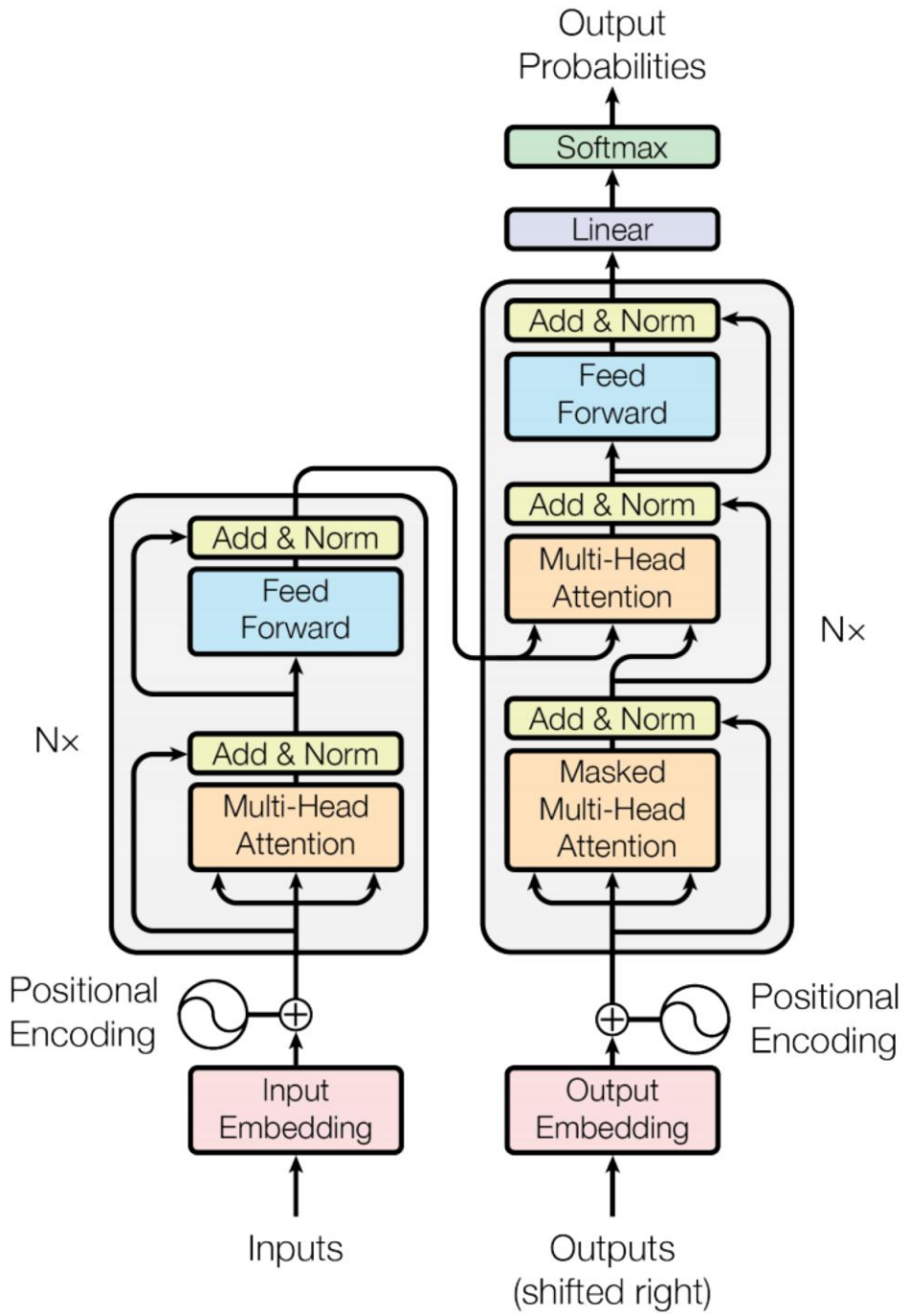


2. Encoder-Decoder Attention, where queries come from previous decoder layer and keys and values come from output of the encoder



Animated workings of attention in transformer

One encoder-decoder block



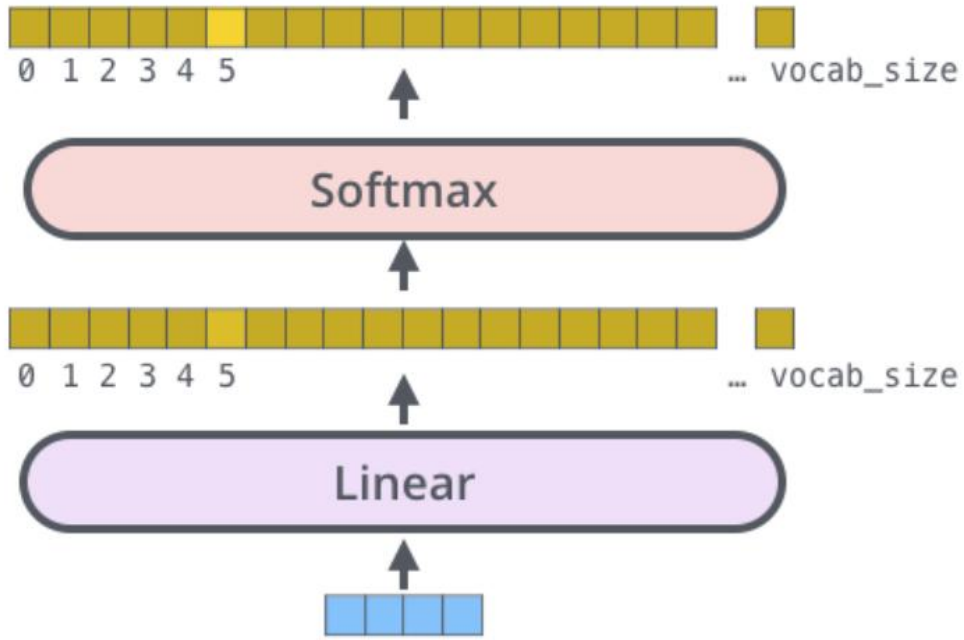
Producing the output words

Which word in our vocabulary is associated with this index?

am

Get the index of the cell with the highest value (**argmax**)

5



Training the transformer

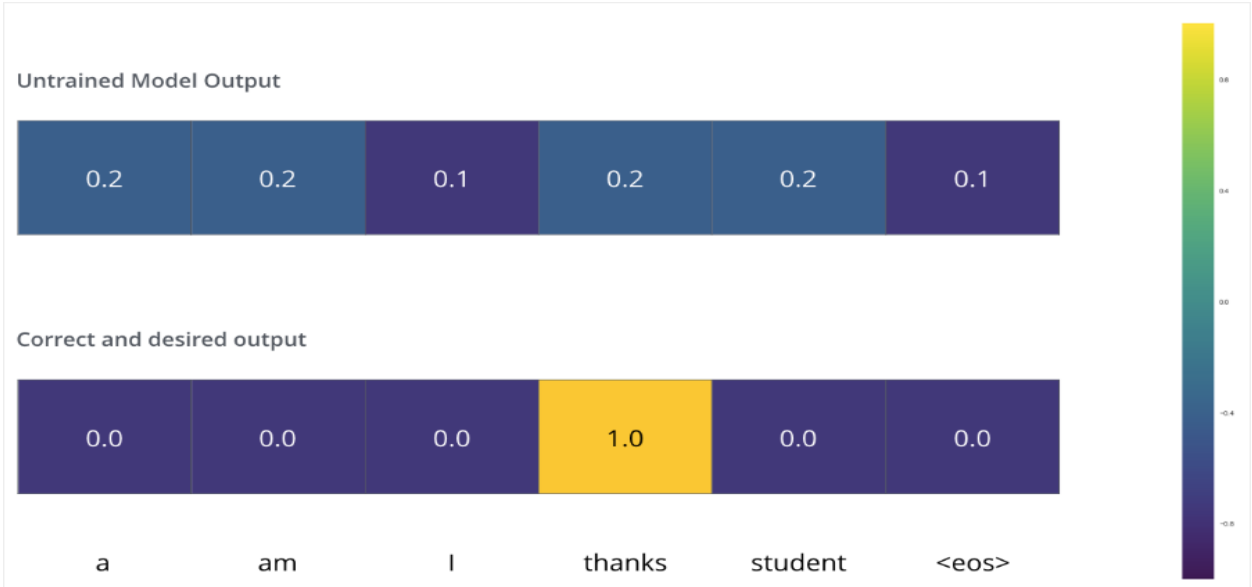
- During training, an untrained model would go through the exactly the same forward pass. But since we are training it on a labeled training dataset, we can compare its output with the actual correct output.
- For illustration, let's assume that our output vocabulary only contains six words(a, am, i, thanks, student, <eos>)
- The input is typically in the order of 10^4 (e.g., 30 000)

Output Vocabulary

WORD	a	am	I	thanks	student	<eos>
INDEX	0	1	2	3	4	5

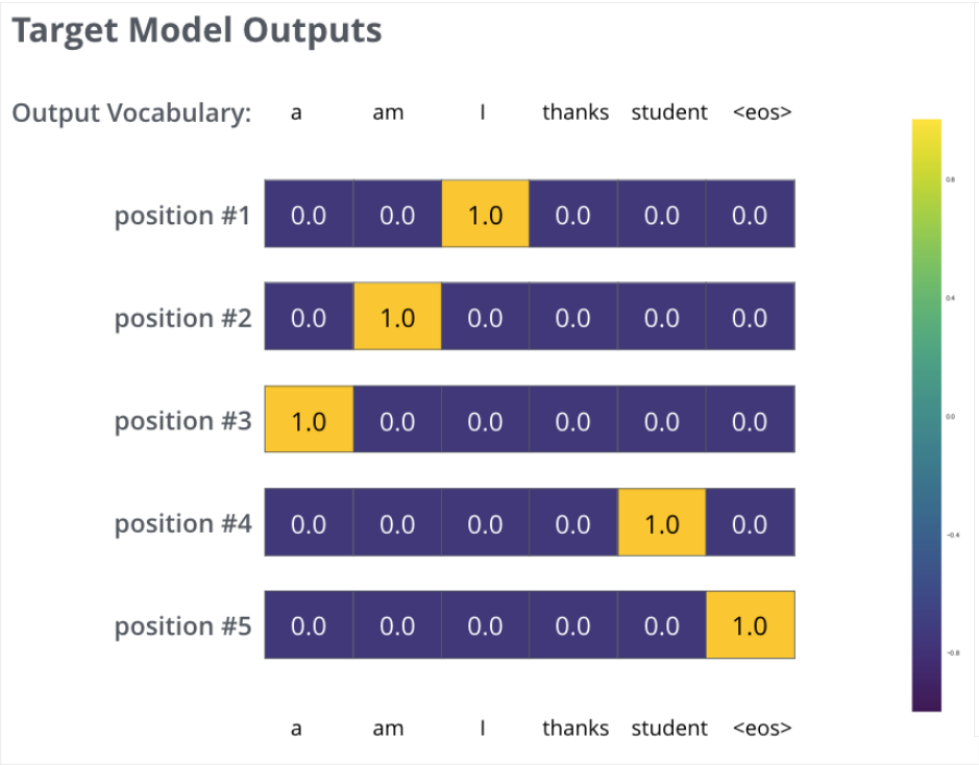
The Loss Function

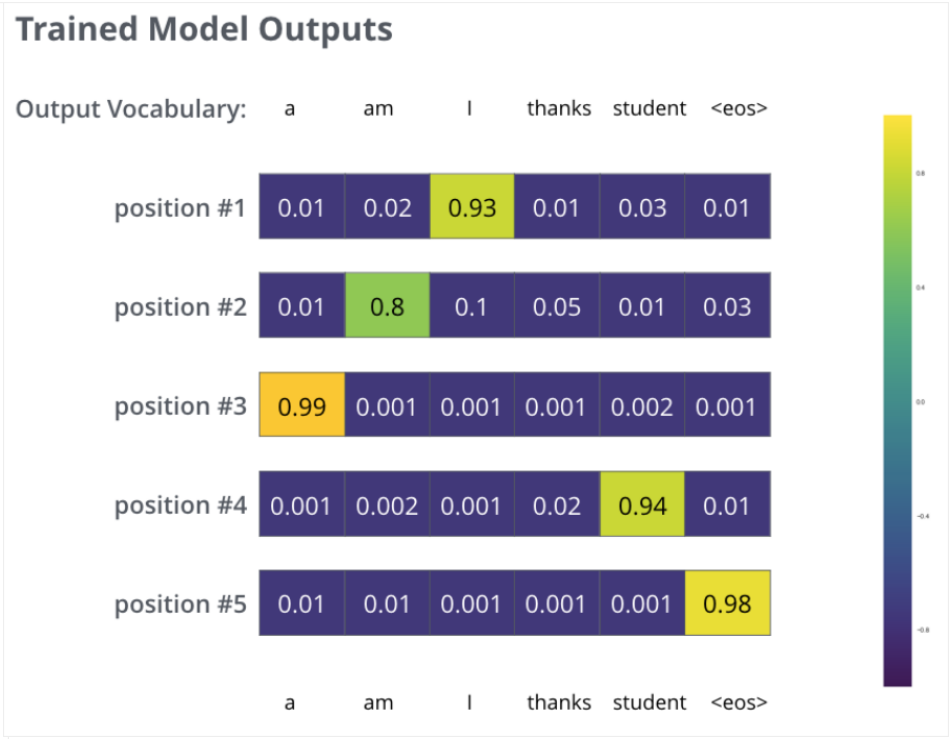
- Evaluates the difference between the true output and the returned output
- Transformer typically uses cross-entropy or Kullback–Leibler divergence.
- The model output is a probability distribution, the true output is 1-hot encoded, e.g.,



Loss evaluation for sequences

- Loss function has to be evaluated for the whole sentence, not just a single word.
- Transformers use greedy decoding or beam search.





Three flavours of transformers

- encoder only (BERT)
- encoder-decoder (machine translation, T5)
- decoder only (GPT)

BERT

- Combines several tasks
- Predicts masked words in a sentence
- Also predicts the order of sentences: is sentence A followed by sentence B or not?
- Combines several hidden layers of the network
- Uses transformer neural architecture, only the encoder part
- Uses several fine-tuned parameters
- Multilingual variant supports 104 languages by training on Wikipedia
- Many publicly available variants.

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 4171–4186

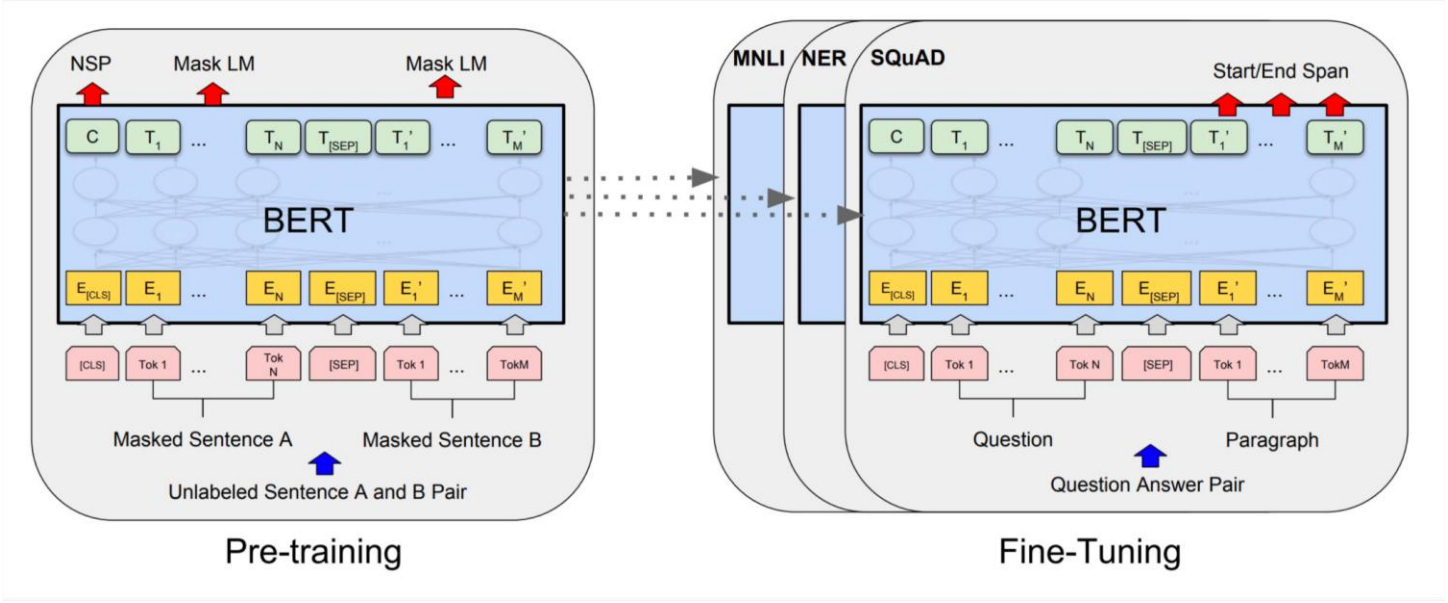
Many BERT-like embeddings

- XLM-R was trained on 2.5 TB of texts in 100 languages
- For Slovene: fastText (a variant of word2vec), ELMo, SloBERTa
- Trilingual BERT – CroSloEngual
- On HuggingFace
- Hundreds of papers investigating BERT-like models in major ML & NLP conferences

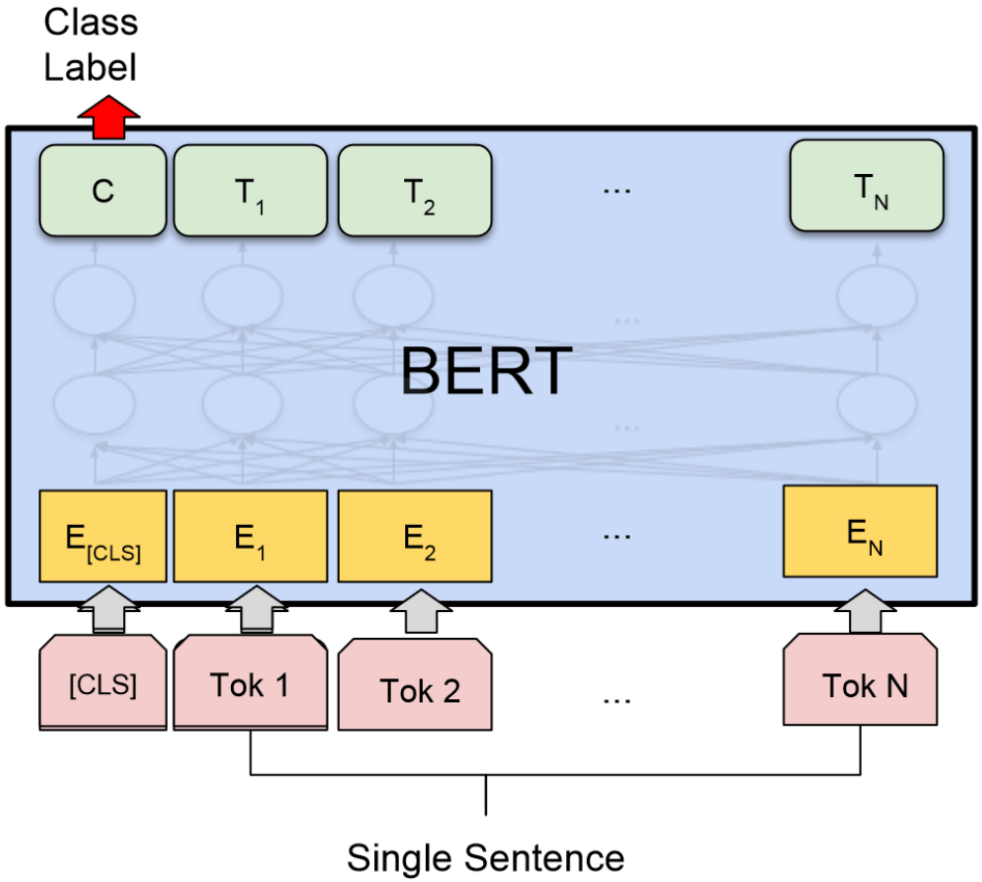
Ulčar, Matej and Marko Robnik-Šikonja. FinEst BERT and CroSloEngual BERT: less is more in multilingual models.
In Proceedings of Text, Speech, and Dialogue, TSD2020, 2020.

Use of BERT

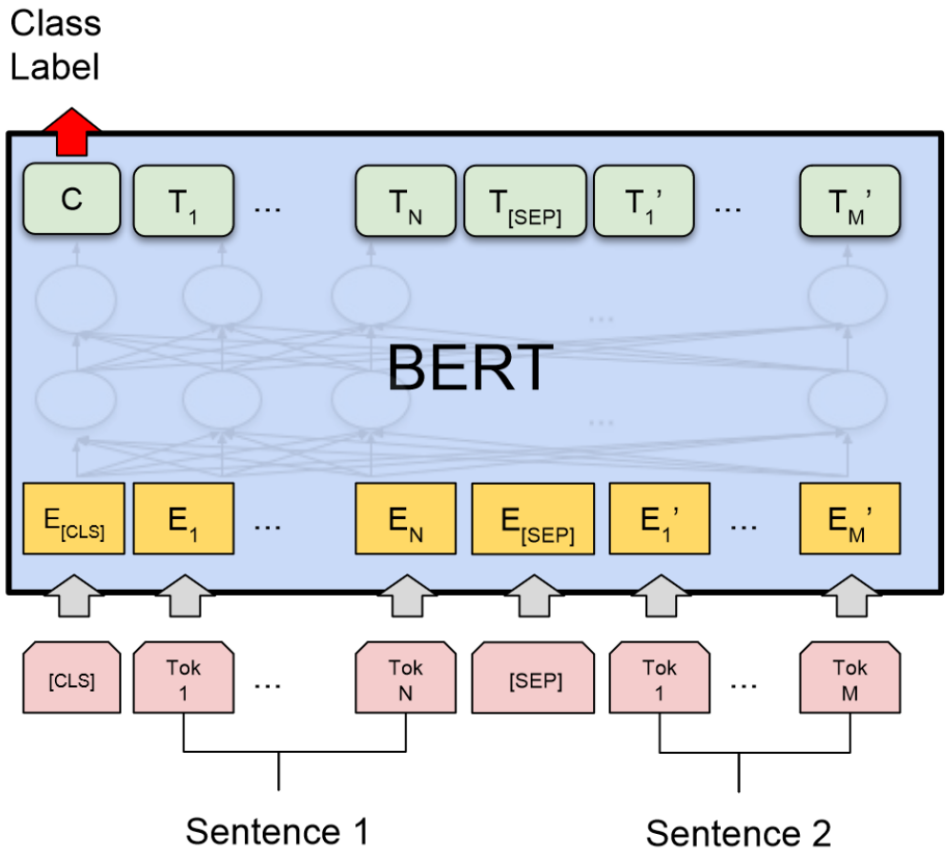
- train a classifier built on the top layer for each task that you fine tune for, e.g., Q&A, NER, inference
- achieves state-of-the-art results for many tasks
- GLUE and SuperGLUE tasks for NLI



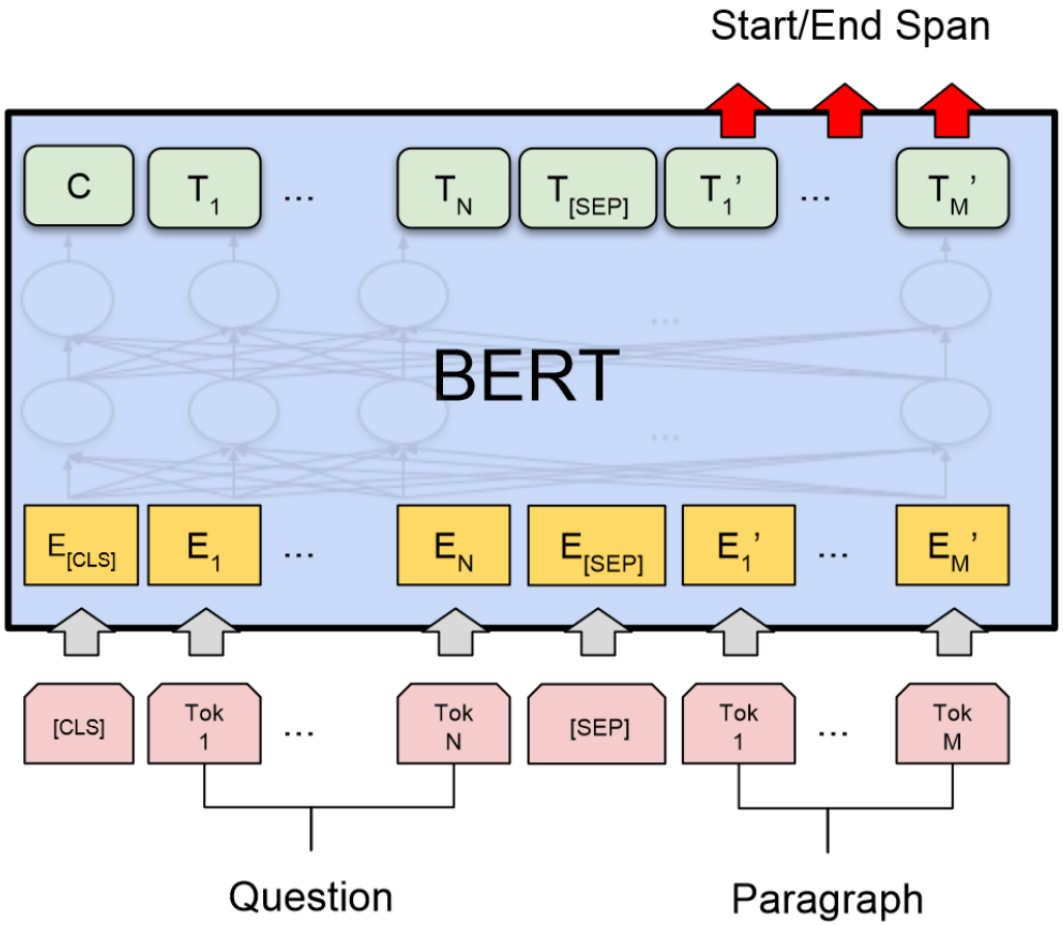
Sentence classification using BERT – sentiment, grammatical correctness



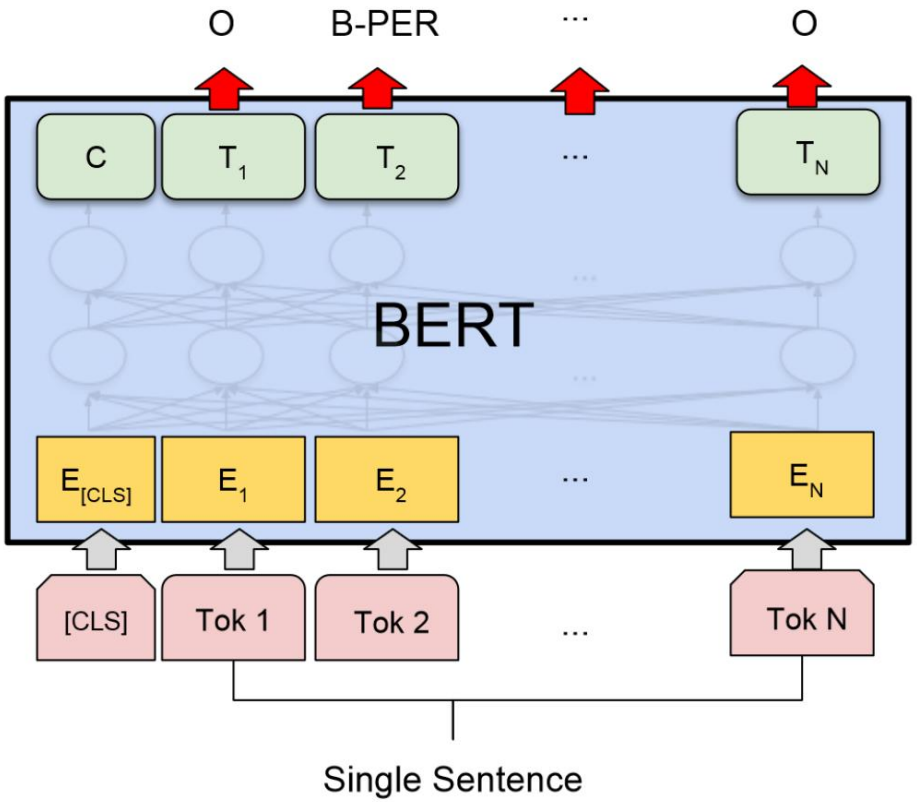
Two sentence classification using BERT-inference



Questions and answers with BERT

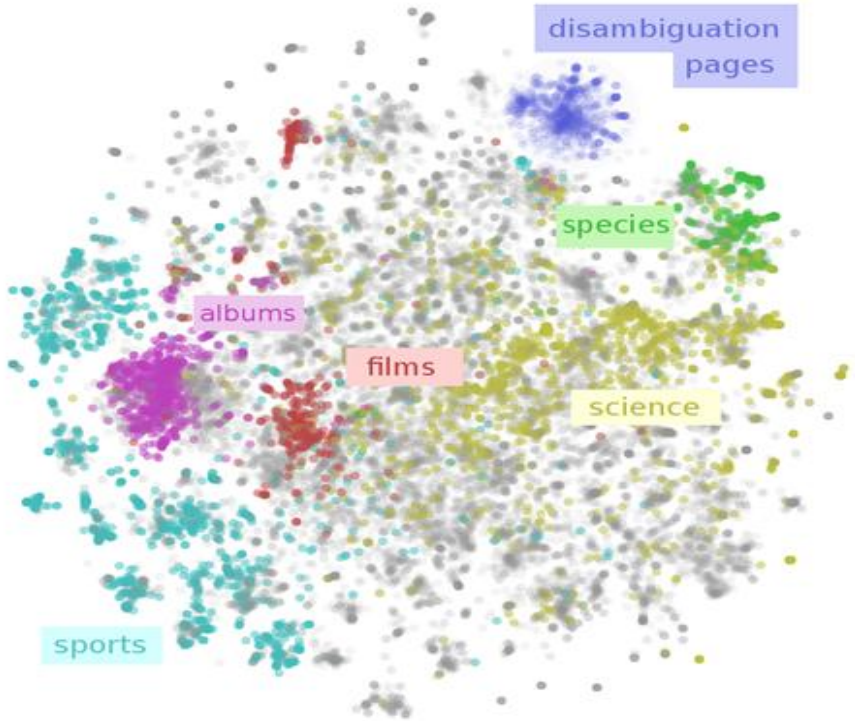


Sentence tagging with BERT-
NER, POS tagging, SRL



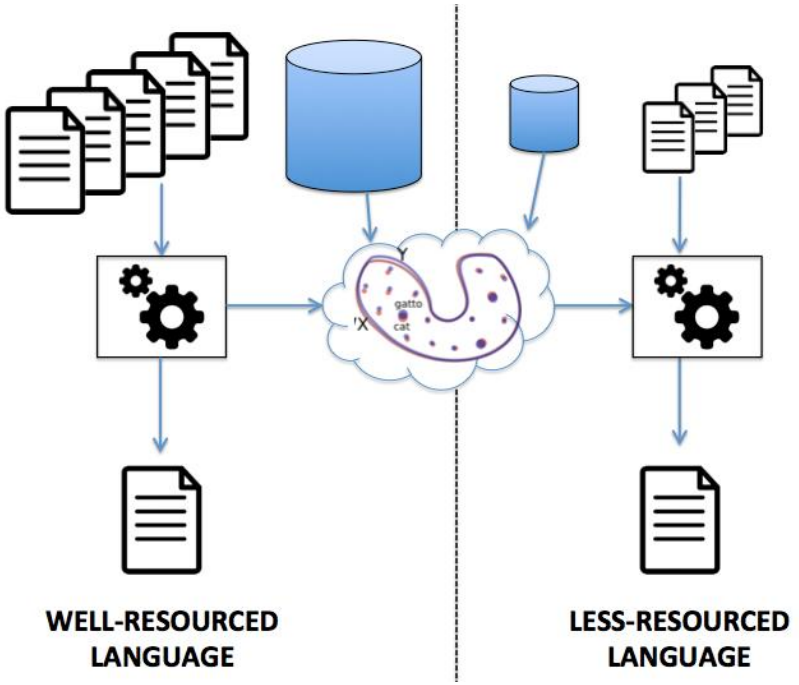
Cross-lingual embeddings

- mostly, embeddings are trained on monolingual resources
- words of one language form a cloud in high-dimensional space
- clouds for different languages can be aligned
  - $W S \approx E$ or
  - $W_1 S \approx W_2 E$



Cross-lingual model transfer based on embeddings

- Transfer of tools trained on mono-lingual resources



mostly not used anymore, superseded by multilingual LLMs

University of Ljubljana
Faculty of Computer and Information Science

Vocabulars in LLMs

- Tokenization depends on the dictionary
- The dictionary is constructed statistically (SentencePiece algorithm)

- Sentence: “Letenje je bilo predmet precej starodavnih zgodb.”
- SloBERTa:

'▁Le', 'tenje', '▁je', '▁bilo', '▁predmet', '▁precej', '▁staroda', 'vnih', '▁zgodb', '.’

- mBERT:

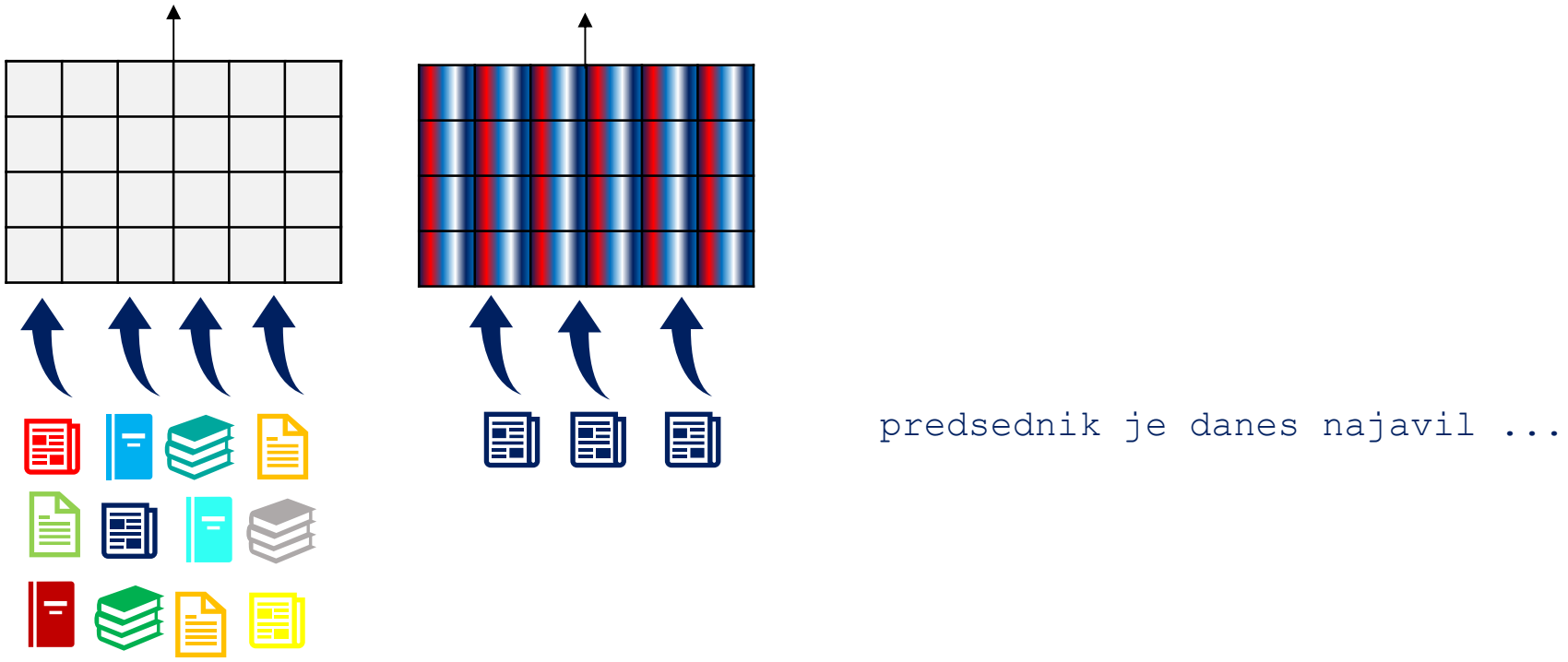
'Let', '##en', '##je', 'je', 'bilo', 'pred', '##met', 'pre', '##cej', 'star', '##oda', '##vnih', 'z', '##go', '##d', '##b', '.'

Ulčar, M., & Robnik-Šikonja, M. (2021) Training dataset and dictionary sizes matter in BERT models: The case of Baltic languages. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 162-172)

Multilingual LLMs

- Pretrained on multiple languages simultaneously
- multilingual BERT supports 104 languages by training on Wikipedia
- XLM-R was trained on 2.5 TB of texts
- allow cross-lingual transfer
- often solve the problem of insufficient training resources for less-resourced languages

Using multilingual models



Zero-shot transfer and few-shot transfer

What LLMs learn?

- We would like to travel to [MASK], ki je najlepši otok v Mediteranu.

SloBERTa: ..., Slovenija, I, Koper, Slovenia

CSE-BERT: Hvar, Rab, Cres, Malta, Brač

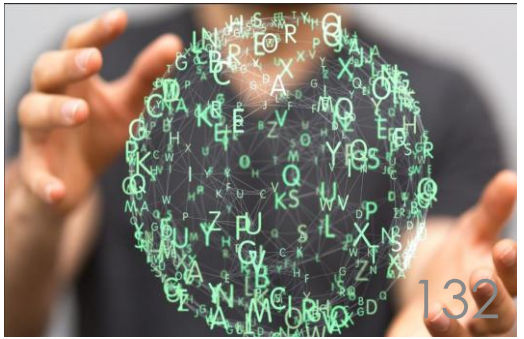
XLM-R: Mallorca, Tenerife, otok, Ibiza, Zadar

mBERT: Ibiza, Gibraltar, Tenerife, Mediterranean, Madeira

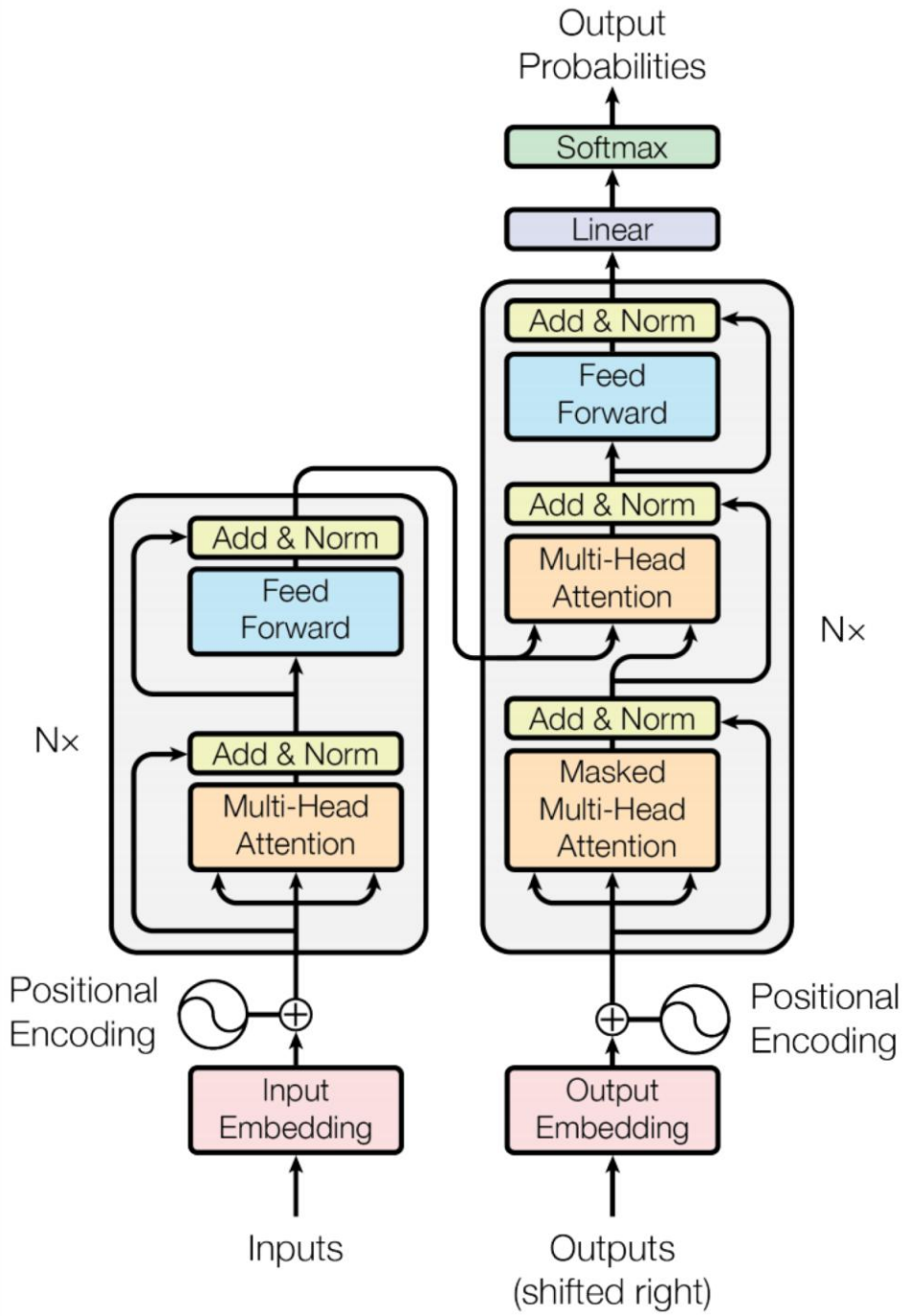
BERT (en): Belgrade, Italy, Serbia, Prague, Sarajevo

Embed all the things!

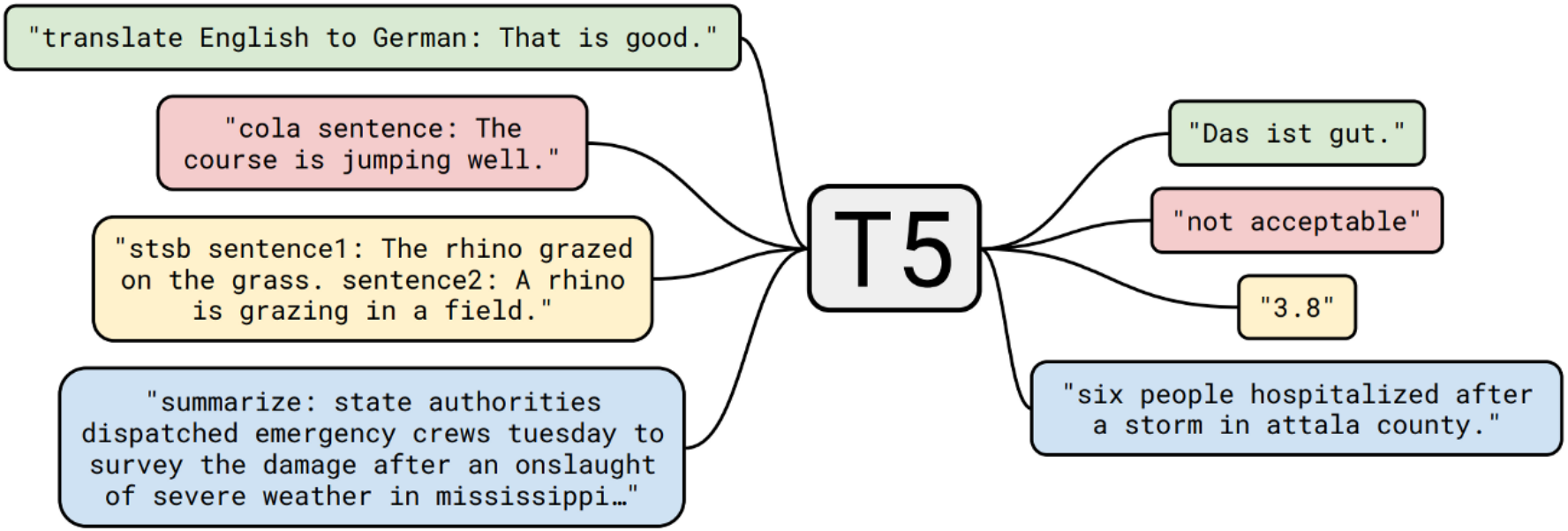
- Neural networks require numeric input
- Embedding shall preserve relations from the original space
- Representation learning is a crucial topic in nowadays machine learning
- Lots of applications whenever enough data is available to learn the representation
- In text, BERT-like models dominate for embeddings
- Similar ideas applied to texts, speech, graphs, electronic health records, relational data, time series, etc.



Transformer



T5 (Text-To-Text Transfer Transformer) models



Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y, Li, W. & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(140), 1-67.

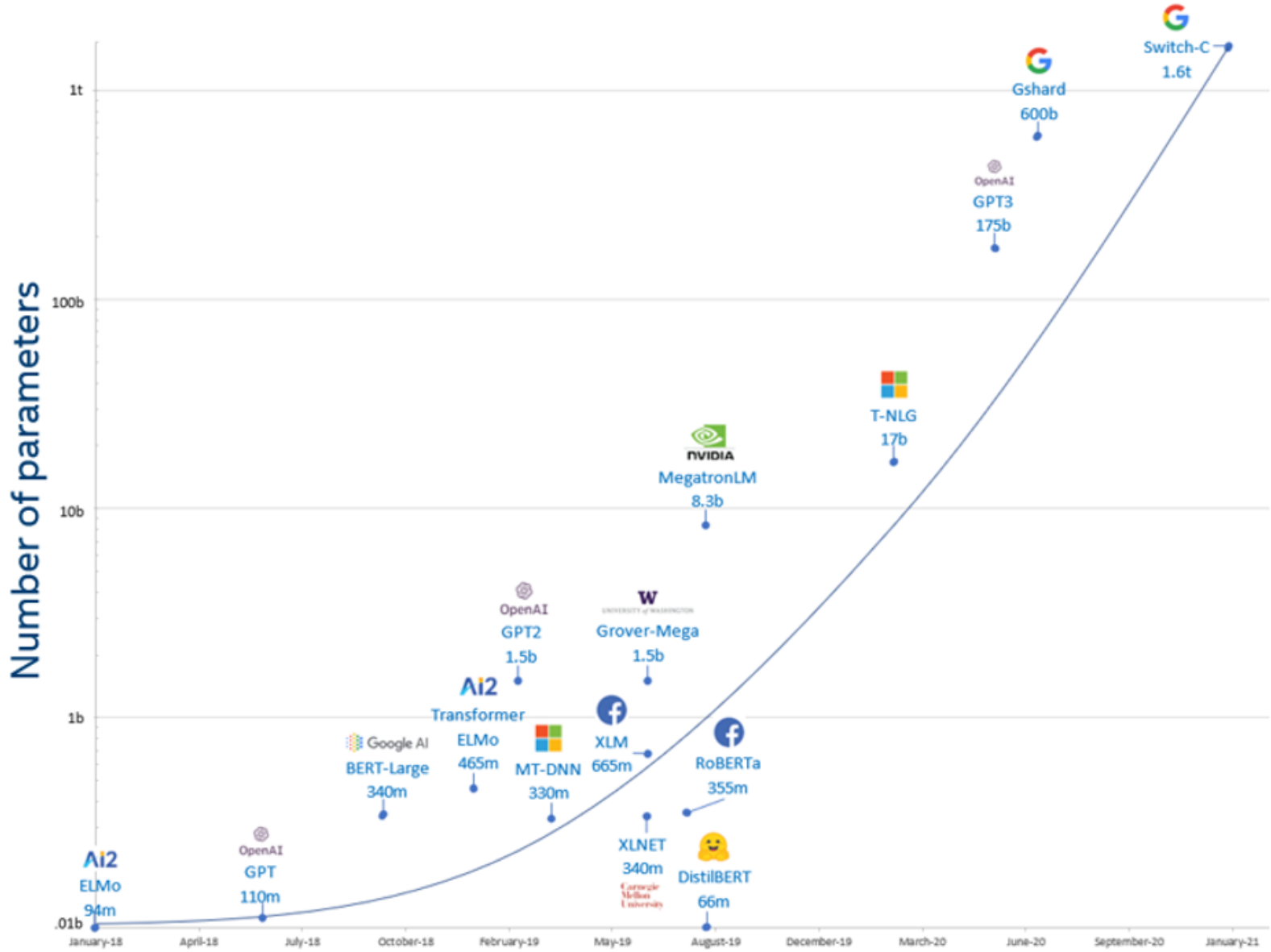
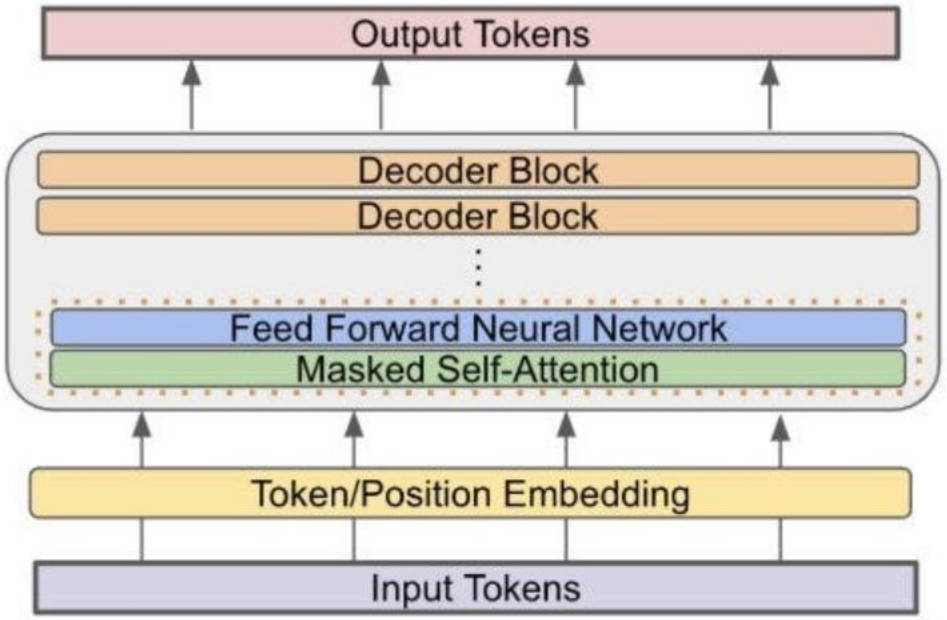


Figure 1: Exponential growth of number of parameters in DL models

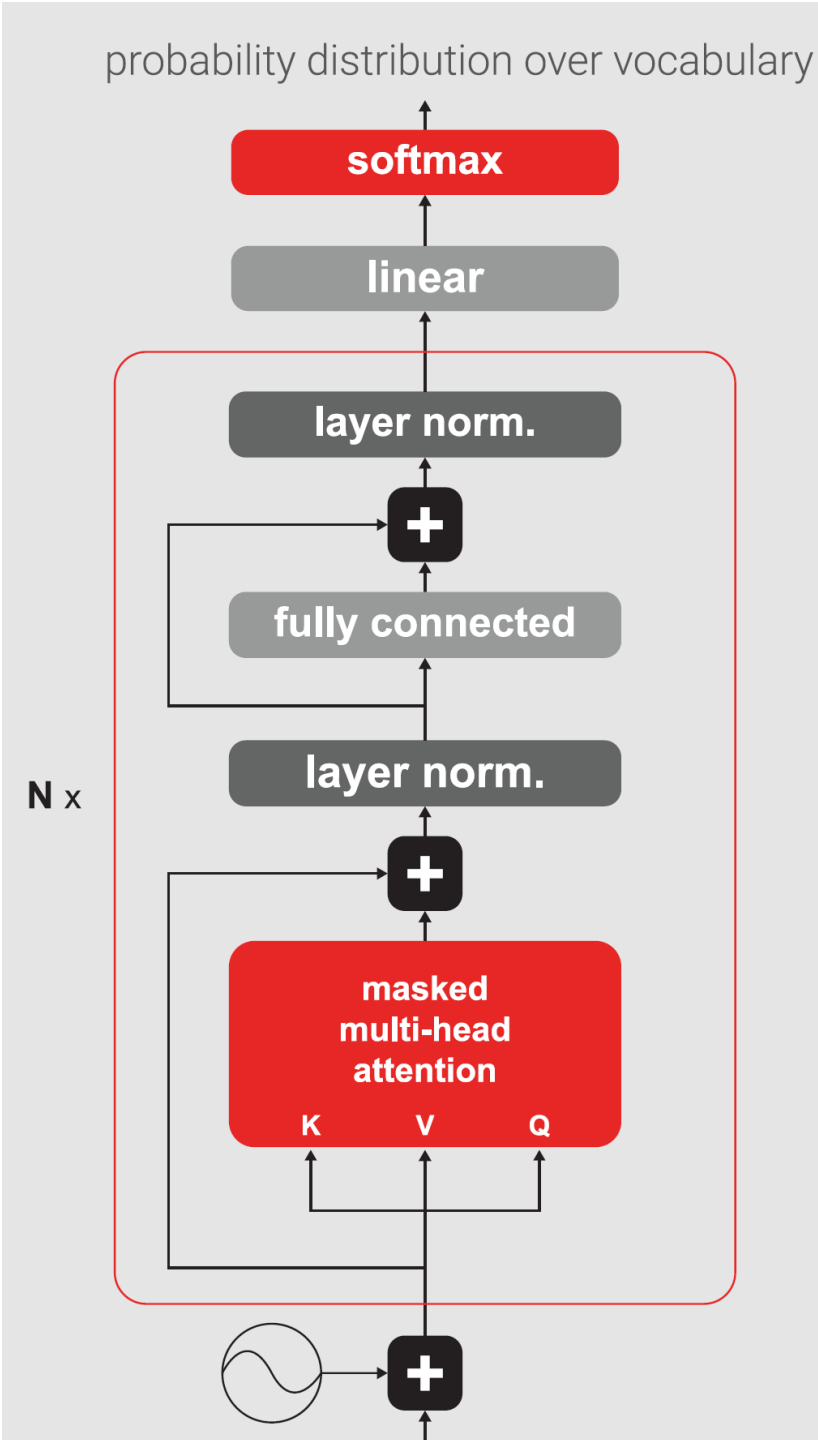
Decoder only models

- GPT, GPT-2, GPT-3, ChatGPT, GPT-4
- LLaMA, LLaMA-2, LLaMa-3
- MPT, Falcon
- Mistral and Mixtral
- OPT, Bloom
- Gemma and Gemini
- Olmo
- GaMS

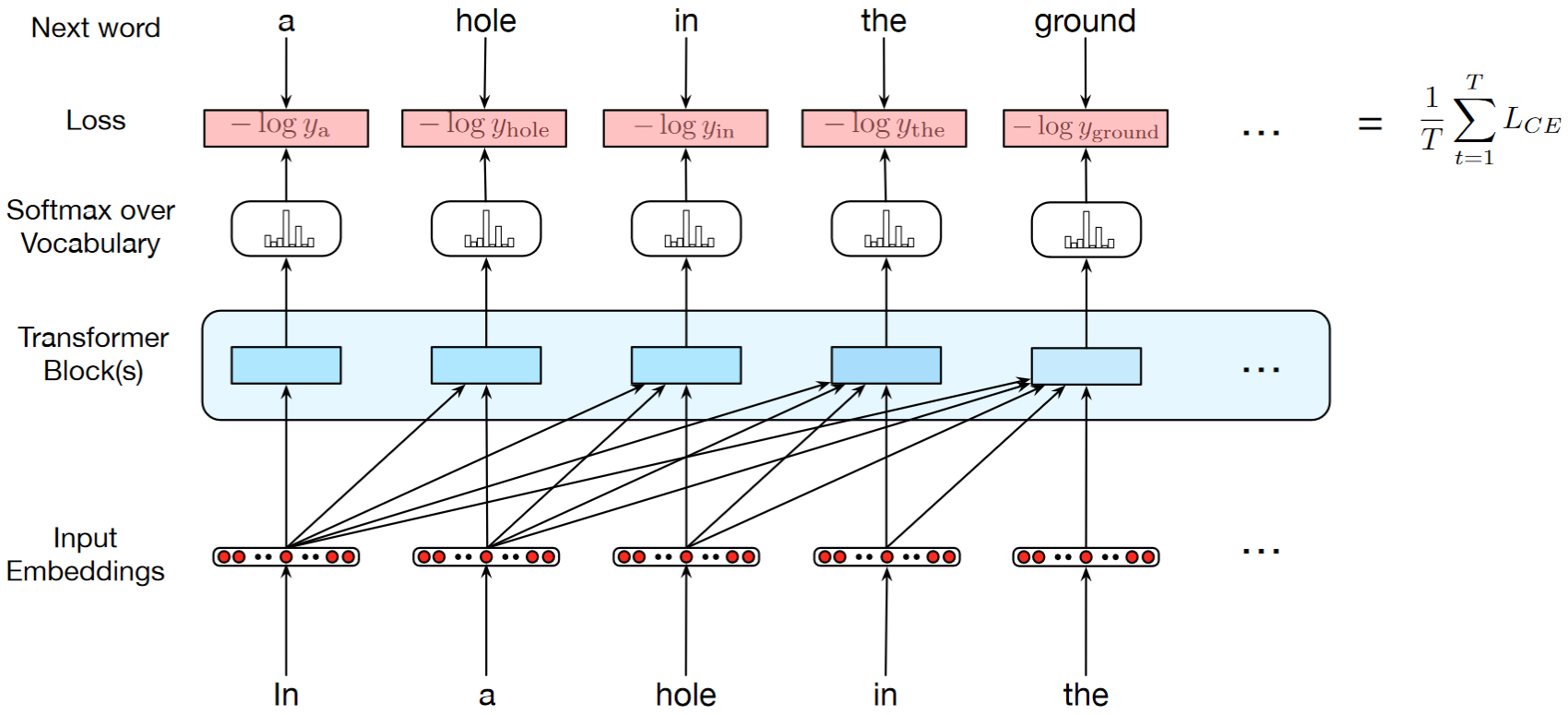


GPT family

- GPT: Generative Pre-trained Transformers
- use only the decoder part of transformer
- pretrained for language modeling (predicting the next word given the context)
- Potential shortcoming: unidirectional, does not incorporate bidirectionality
- “What are those?” he said while looking at my crocs.



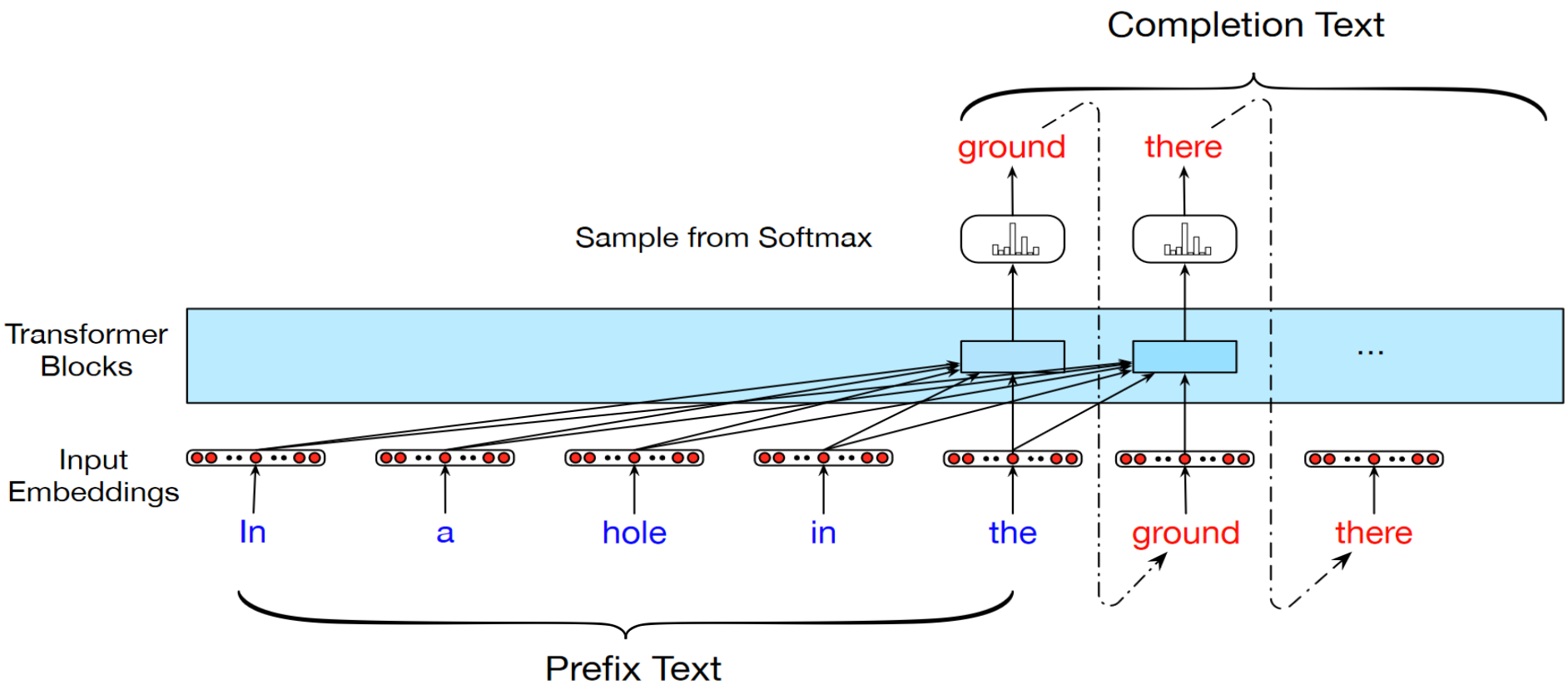
Transformer as a language model



- Can be computed in parallel

Autoregressive generators

- priming the generator with the context



- can be used also in summarization, QA and other generative tasks

GPT-2 and GPT-3

- few architectural changes, layer norm now applied to input of each subblock
- GPT-3 also uses some sparse attention layers
- more data, larger batch sizes (GPT-3 uses batch size of 3.2M)
- the models are scaled:

GPT-2:
48 layers, 25 heads
$d_m = 1600$, $d = 64$
context size = 1024
~ 1.5B parameters

GPT-3:
96 layers, 96 heads
$d_m = 12288$, $d = 128$
context size = 2048
~ 175B parameters

[1] Radford et al.: Language Models are Unsupervised Multitask Learners, 2019.
[2] Brown et al.: Language Models are Few-Shot Learners, 2020.

- see demos at https://transformer.huggingface.co/

Huge generative language models

- ChatGPT, OpenAI, Nov. 2022
  based on GPT-3.5 with additional training for dialogue
- uses instruction-following training and preference adaptation
- demo: https://chatgpt.com/
- huge public impact, disruptive for writing professions, learning, teaching, and scientific writing
- GPT-4, 5: even larger, allows longer context, image input and output, at least 8 x 220B parameters, MoE, reasoning
- LLaMa, Mistral, Gemma, DeepSeek, Qwen, EuroLLM, GaMS: smaller, open, very competitive LLMs

In-context learning

- GPT-2 and GPT-3 were the first models where one can ditch the “pre-train and fine-tune” training paradigm of GPT;
- GPT-2 explores unsupervised zero-shot learning, whereas in GPT-3 the authors expand the idea into in-context learning;
- Use text input to condition the model on task description and some examples with ground truth.
- Uses zero-shot learning, one-shot learning, few-shot learning (as many examples as they can fit into the context, usually 10-100)
- No gradient updates are performed.
- A sort of associative lookup

In-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

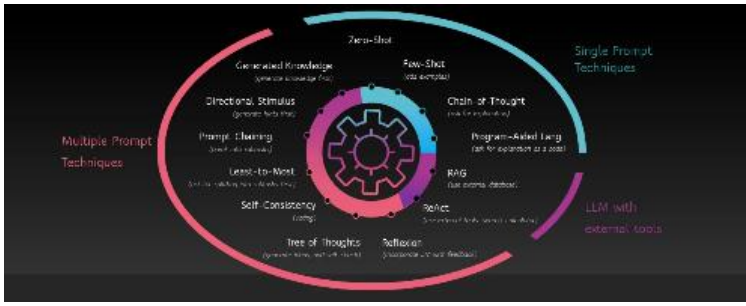
Figure source: Brown et al.: Language Models are Few-Shot Learners, 2020.

- GPT-3 is still a language model and can be used for text generation

- only 12% of respondents correctly classified this as not written by a human

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Prompt engineering (PE)



- Prompt engineering is a relatively new discipline for developing and optimizing prompts to efficiently use LLMs for a wide variety of tasks.
- Prompt engineering can help design robust and effective prompting techniques that interface with LLMs and other tools
- PE is a set of soft guidelines on how to design effective prompts
- PE is becoming an automatic approach to design effective prompts
- See https://www.promptingguide.ai/

Preference optimization of LLMs

- instruction-tuning
- Preference optimization
- ChatGPT uses RLHF (reinforcement learning with human feedback)

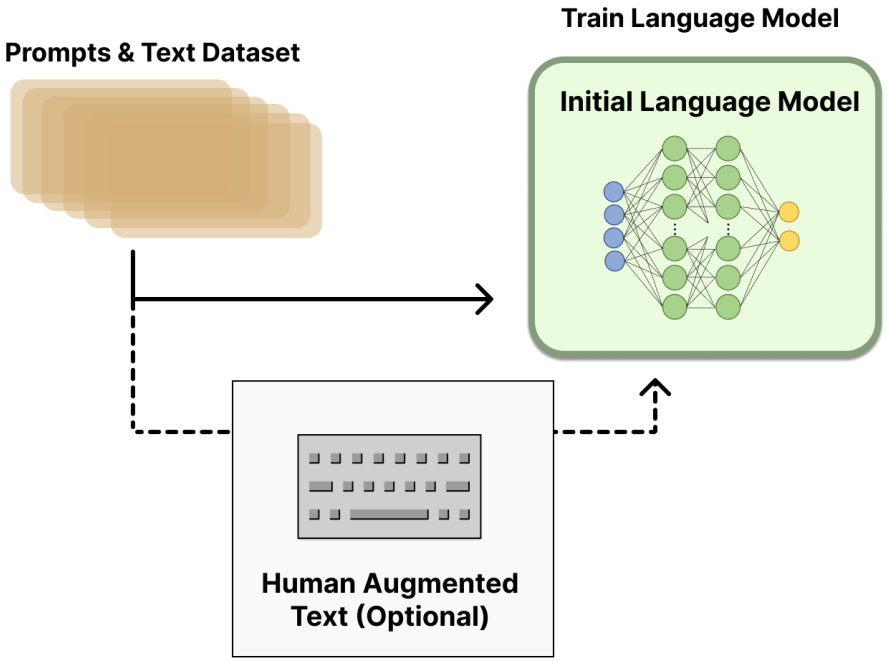
RLHF: idea

- Reinforcement Learning with Human Feedback
- A problem: Human feedback is not present during training
- Idea: Train a separate model on human feedback, this model can generate a reward to be used during training of LLM
- Three stages:
  1. Pretraining a language model (LM),
  2. Gathering data and training a reward model, and
  3. Fine-tuning the LM with reinforcement learning.

Nathan Lambert, Louis Castricato, Leandro von Werra, Alex Havrilla: "Illustrating Reinforcement Learning from Human Feedback (RLHF)", Hugging Face Blog, 2022.

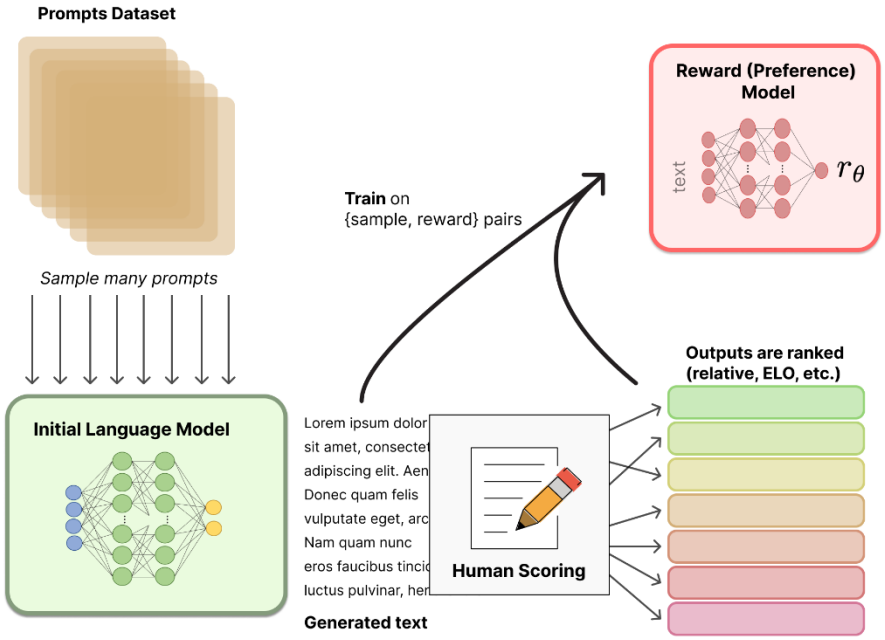
RLHF: the reward model 1/2

- input: a sequence of text, e.g., produced by LM and optionally improved by humans
- output: a scalar reward, representing the human preference of the text (e.g., a rank of the answer)
- the reward model could be an end-to-end LM, or the model ranks outputs, and the ranking is converted to reward



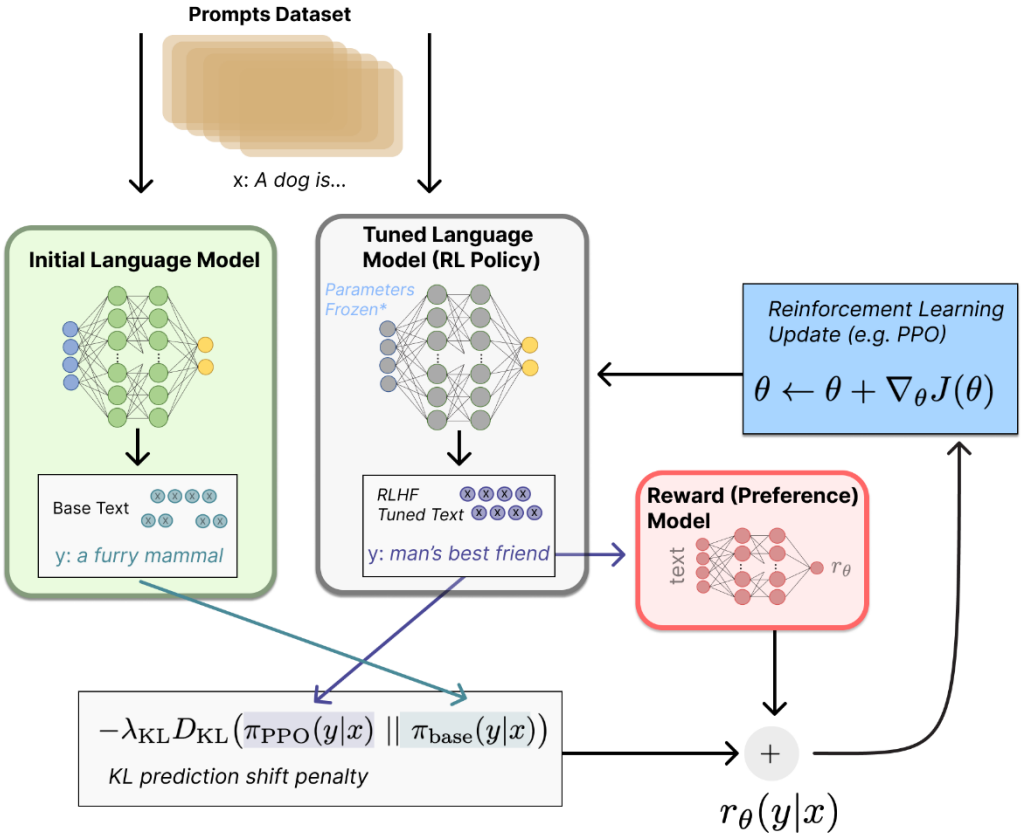
RLHF: the reward model 2/2

- the training dataset are pairs of prompts and (human improved) LM responses, e.g., 50k instances
- humans rank the responses instead of producing the direct reward as this produces better calibrated scores



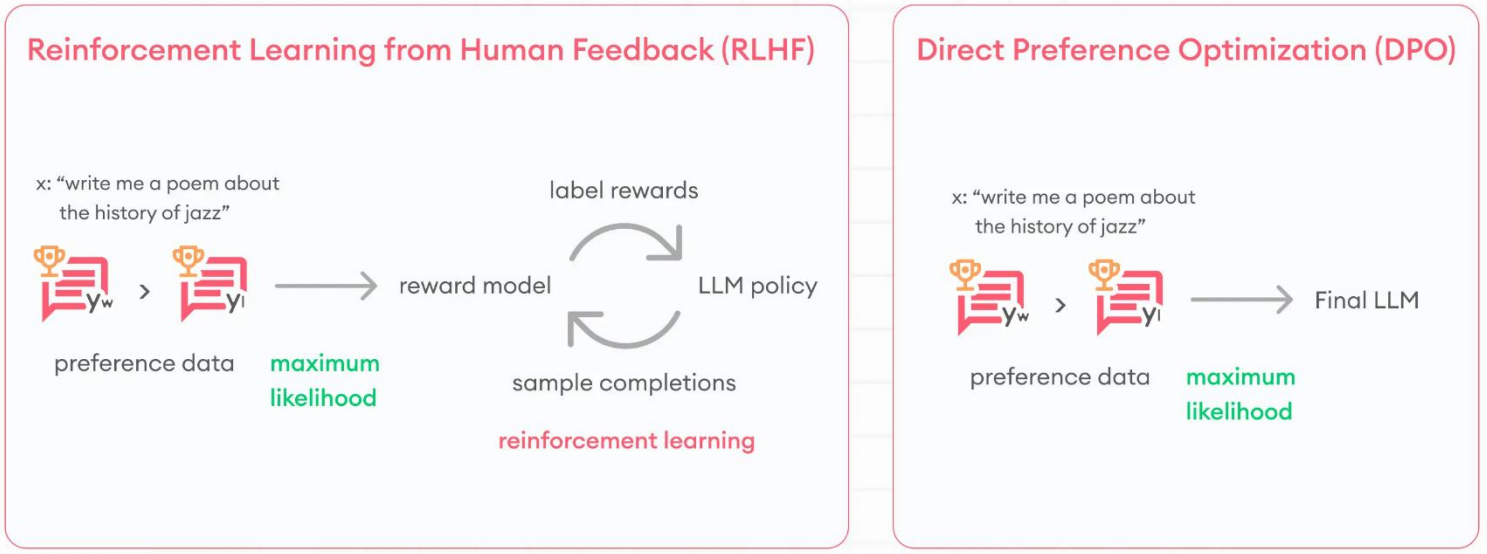
RLHF: fine-tuning with RL

- RL does not change all parameters, most of parameters are frozen
- the algorithm: Proximal Policy Optimization (PPO)



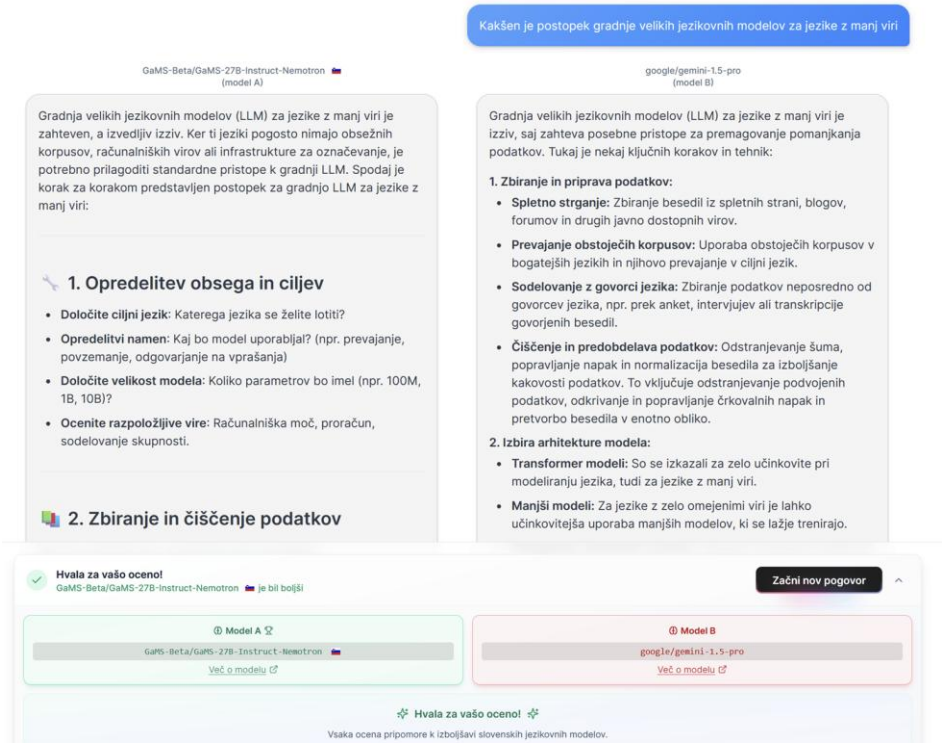
Direct preference optimization (DPO)

- DPO is a reinforcement learning approach, used as an alternative to classical RLHF with PPO
- Simpler, easier to collect data via LLM-Arena



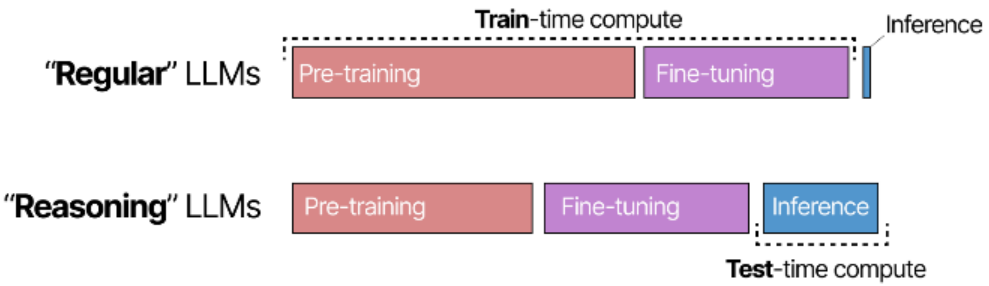
LLM Arena – feedback about model response

- A prompt is responded to by two models simultaneously
- The user chooses the better answer
- In the Slovenian LLM arena, the most successful models for Slovene are compared
- Collection and analysis of user responses
- First findings: GaMS is just a little behind the 1000x larger Gemini 2.5 Pro and is better than other models
- https://arena.cjvt.si/sl

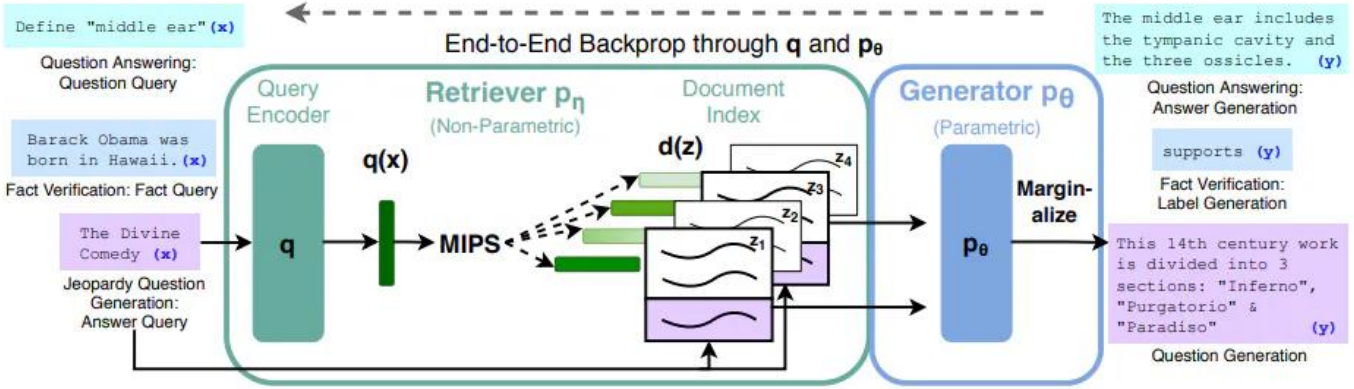


New ways to train and use LLMs

- Many prompt engineering techniques
- In-context learning: prepare context relevant for the given problem and is helpful in generating a useful response
- RAG, retrieval augmented generation
- Chain-of-thought reasoning
- Auto-GPT: breaks the goal in subgoals, prepares a sequence of actions for each subgoal, which are described and send to submodels or external programs
- Inference time reasoning
- Agent architectures



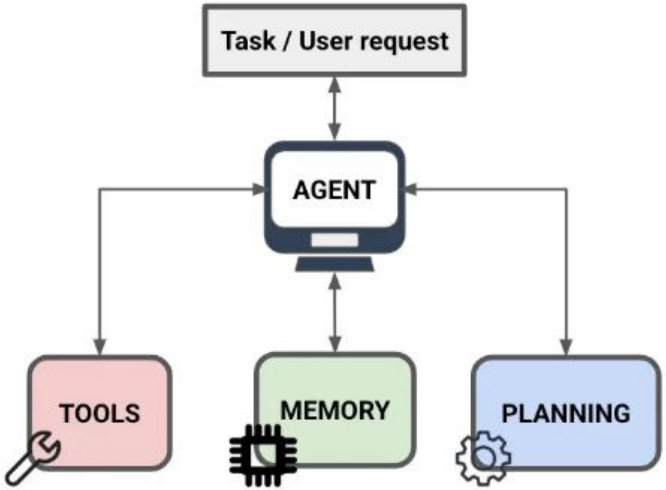
RAG details



Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, pp.9459-9474.

LLM Agents

- A relatively new approach utilizing LLMs for various complex tasks



GaMS – generativni model za slovenščino

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

ODPRTODOSTOPEN

- Vaši podatki ostanejo pri vas

- Učenje verzije 9B: 3 dni na 32 GPU vozliščih (128 GPU) na Leonardo Booster HPC
- Sledenje navodilom: 25.000 primerov, prilagoditev parametrov, 1 dan, 1 GPU vozlišče

GaMS:
Generativni model za slovenščino
Generative Model for Slovene

VEČ RAZLIČIC

- 2B
- 9B
- 27B

DODATNO UČEN NA SLOVENSKIH BESEDILIH

- Osnova: Gemma 2
- Trenutna različica: 8 milijard slovenskih “žetonov” (6 milijard besed)

NAJBOLJŠI ODPRT MODEL ZA SLOVENŠČINO

- SloBench
- Nadgradnja – več besedil

UL | FRI

GAMS :” 9B

- Based on Gemma 2 9B.
- Pretraining on Slovene data using 15B tokens (9B Slovene tokens)
- Pretraining: 3 days on 32 GPU nodes (128 GPUs) on Leonardo Booster super-computer
- Instruction following: 25.000 instances, updating all parameters, 1 day, 1 GPU node
- The most successful open model for Slovene
- Evaluations:
  - SloBENCH : SuperGLUE, machine translation
  - SlovenianEval

Rank	Title	Authors and Affiliations	Average
1	GaMS-27B	Domen Vreš, FRI	0.7601
2	PrešernGPT 0.1	Žan Knafelc	0.7568
3	Gemma 2 27B	Google, Domen ...	0.7546
4	GaMS-9B	Domen Vreš, FRI	0.7309
5	GaMS-9B-Instru...	Domen Vreš, FRI	0.6997
6	Gemma 2 9B	Google, Domen ...	0.6980
7	SlovenianGPT-C...	Domen Vreš, FRI	0.6436
8	CroSloEngual BE...	Aleš Žagar, FRI ...	0.6078
9	GaMS-1B-Chat-f...	Domen Vreš, FRI	0.5806
10	OPT_GaMS-1B-...	Domen Vreš, FRI	0.5645

https://povejmo.si

Gemma 2 -> GaMS:”

- Phase 1: language alignment
  - parallel English-Slovene corpora
- Phase 2: high-quality Slovene data
  - Wikipedia + MetaFida + KAS
  - without: web-crawl, tweets
  - news sorted by increasing time

GaMS: “

Evaluation: SloBench - SuperGLUE

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

Leaderboard

Slovene SuperGLUE 1.0

Related Leaderboards

Slovene SuperGLUE (v1)

Description

This leaderboard is a Slovene (translated) version of an existing English version. The task initially consists of multiple tasks, which test a system's ability to understand and reason about texts in Slovene. All the tasks should be solvable by most college-educated Slovene speakers.

The original SuperGLUE benchmark consists of 8 tasks, while 6 tasks are included in the Slovene evaluation. These tasks are BoolQ, CB, COPA, MultiRC, RTE and WSC.

Leaderboard Editors

Aleš Žagar, ales.zagar@fri.uni-lj.si
Slavko Žitnik, slavko@zitnik.si

https://slobench.cjvt.si/

Leaderboard Submissions

Rank	Title	Authors and Affiliations	Average	BoolQ Accuracy	CB Accuracy	CB F1 Score	CB Average	COPA Accuracy	MultiRC EM	MultiRC F1a Score	MultiRC Average	RTE Accuracy	WSC Accuracy	Tag
1	GaMS-27B	Domen Vreš, FRI	0.7601	0.8333	0.6440	0.5864	0.6152	0.9540	0.3904	0.7504	0.5704	0.7931	0.7945	
2	PrešernGPT 0.1	Žan Knafelc	0.7568	0.8333	0.8520	0.5868	0.7194	0.9740	0.4985	0.8061	0.6523	0.8276	0.5342	
3	Gemma 2 27B	Google, Domen ...	0.7546	0.8333	0.6680	0.5972	0.6326	0.9140	0.4174	0.7295	0.5735	0.8276	0.7466	
4	GaMS-9B	Domen Vreš, FRI	0.7309	0.7000	0.8400	0.7955	0.8178	0.9000	0.3243	0.6551	0.4897	0.7931	0.6849	
5	GaMS-27B-Instr...	Domen Vreš, FRI	0.7038	0.8333	0.7200	0.6322	0.6761	0.9400	0.0511	0.5813	0.3162	0.7931	0.6644	
6	GaMS-9B-Instru...	Domen Vreš, FRI	0.6997	0.8000	0.7960	0.7128	0.7544	0.8140	0.0721	0.6174	0.3447	0.7931	0.6918	
7	Gemma 2 9B	Google, Domen ...	0.6980	0.8333	0.8240	0.5683	0.6962	0.8700	0.2282	0.5310	0.3796	0.7241	0.6849	

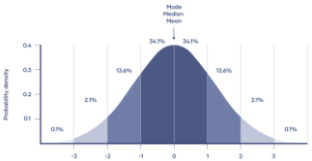
Injecting knowledge into large generative models

- Slovenian dataset for Common-Sense Reasoning KE-WS
- Slovenian dataset for Natural Language Inference SI-NLI
- Prepared datasets for logical reasoning and mathematics
- Prepared a Slovene safety dataset
- In working: A dataset with ethical norms
- Methodology for the transfer of linguistic and lexicographic knowledge to LLMs
- To be included in the next version of GaMS models
- New applications based on the GaMS models, e.g., grammar error correction, machine translation, summarization and simplification of texts, etc.
- GaMS 3.0 is almost ready (based on Gemma 3.0)

HPC related lessons

- Building LLM from scratch is computationally extremely costly
  - LLaMa-2 7B > 360,000 GPU hours,
  - SLING tender: Vega GPU offers 80,000 node hours = 320.000 GPU hours annually
  - GaMS 3: a combination of Leonardo Booster, Nvidia Sovereign LLMs support and UL FRI FRIDA cluster
- Building an LLM by further training an existing LLM:
  - Data quantity issues (web crawl, collection, machine translation): Hoffman scaling laws
  - Data quality issues
  - Model vocabulary issues,
  - Platform issues
  - Model parallelization (within 1 GPU node, i.e. 4 or 8 A/H100),
  - Data parallelization (across nodes), reduces inter-node communication
  - Support for bf16 (1b sign, 8b exponent, 7b significand)
  - QLORA: NF4, 1 node, needs instruction-tuning datasets,
  - Building LLM requires lots of HPC expertise
- Using platforms like the NVidia NeMo framework:
  - +speed
  - +easier parallelization
  - -less models available
  - -needs model and code tweaking
- In deployment: squeeze everything on a single computational node (if possible)

Statistical evaluation of LLMs' results



- LLMs can answer any question, even questions about complex problems, such as word sense definition

"You are a lexicographer who is writing definitions for a general dictionary of Slovene. The dictionary is targeted at native speakers of Slovene. Here are some sample definitions for a selection of nouns/adjectives/verbs/adverbs:
"file.txt"
"Write a dictionary definition for the selected sense of the noun/adjective/verb/adverb X. Below is some contextual information. There is only one sense, therefore write only one definition. Write the definition only; do not add any justifications, and do not repeat the provided contextual information."
…

- We have no guarantee that the answers are correct
- LLMs do not abstain; we almost always get an answer
- To statistically analyze the problem and usability of LLMs (either prompts or fine-tuned LLMs) we typically do a preliminary test
- To trust the results and use them, LLMs' results have to be verified first

Verification of LLMs outputs

- Using LLMs for complex questions:
  - Ask LLM
  - Verify the answers:
  - If the accuracy is high, use LLM; otherwise, this is senseless, revise
- Three verification paths:
  1. Comparison with human gold-standard answers
  2. Human verification of LLM's answers
  3. LLM-as-a-judge

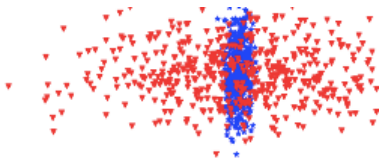
Verification 1: Comparison with human gold-standard answers

- It might be time-consuming or expensive to get the gold-standard answers
- For classification, use metrics such as accuracy, F_1, precision, recall, etc.
- In contrast to classification, the comparison is more difficult for generated texts, but methods exist
  - n-gram based matching (BLEU, ROUGE, etc),
  - sentence encoders and cosine distance
  - LLM as a judge (stronger LLM, has to be verified first, potentially circular)

Verification 2:
Human verification of LLM's answers

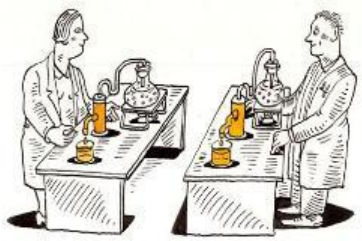
- Humans check the answers of LLMs (correctness, style, format, consistency)
- Note: Checking the answers is a weaker form of verification than producing the same or similar answers (the priming effect).

Reliability



- **Variance** in answers (humans' or LLMs') is always present
- Humans: different background, education, political orientation, sense of humour, mood, etc.
- LLMs: different prompts, different LLMs, temperature parameters, etc.
- If possible, compute
  - Inter-annotator agreement
  - Fleiss' kappa,
    Landis and Koch scale: $\kappa > 0.80$ (almost perfect), $0.60 < \kappa \leq 0.80$ (substantial), etc.
  - Annotator's self-agreement
- To improve reliability: repeat and average, show variance

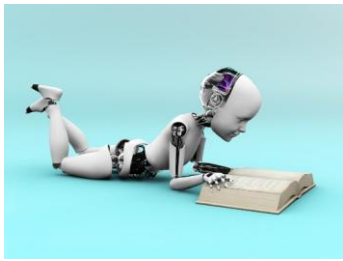
Reproducibility



- **Reproducibility** concerns:
  - The research in LLMs is very fast
  - Commercial LLMs have a life span of around a year
  - However, there are many (less capable) open-source models

Where LLMs go?

- thousand of new applications
- multimodal models
- Knowledge injection, e.g., BioRDF
- external knowledge incorporation, e.g., ToolFormer
- explanation and ethical use



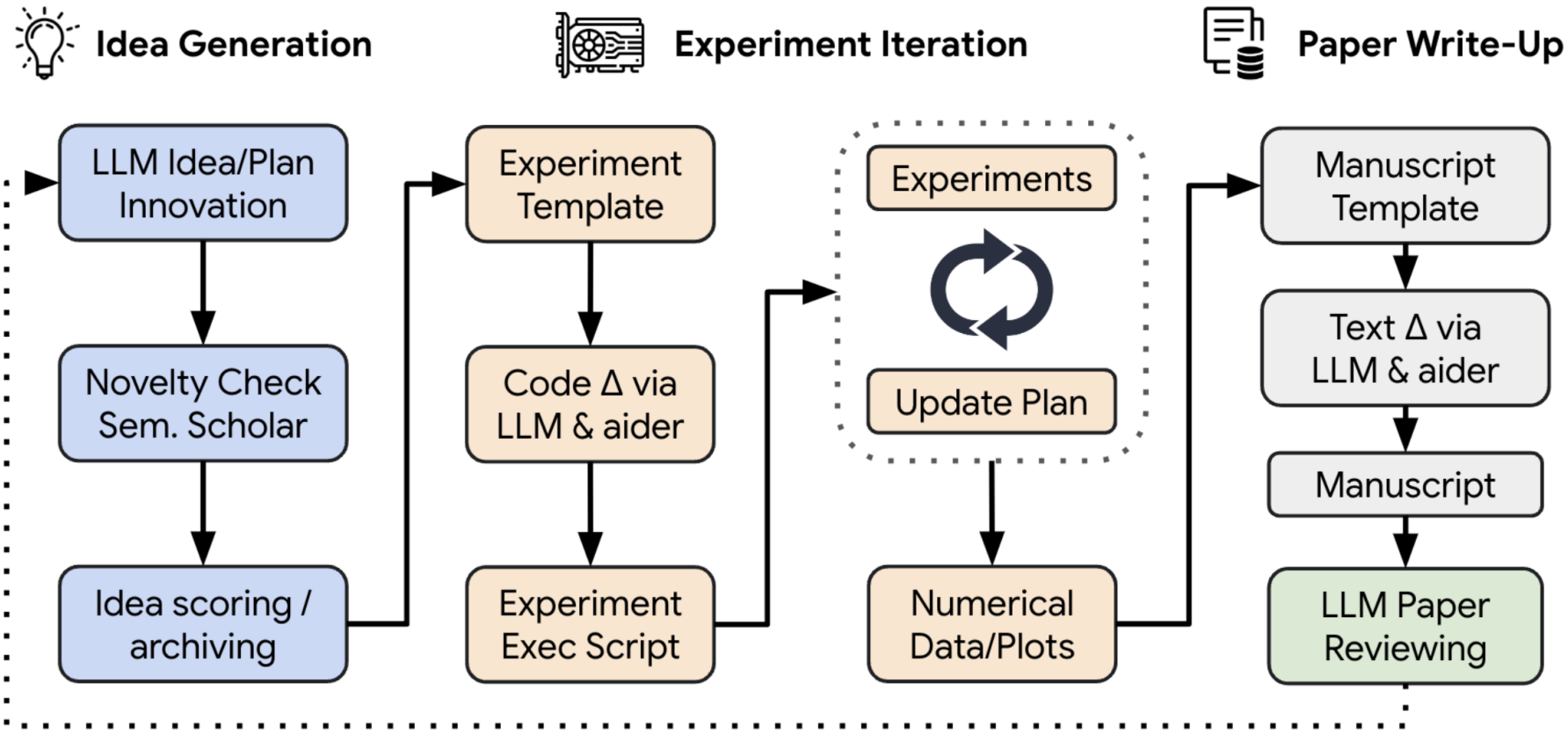
UL FRI

2024-8-16

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

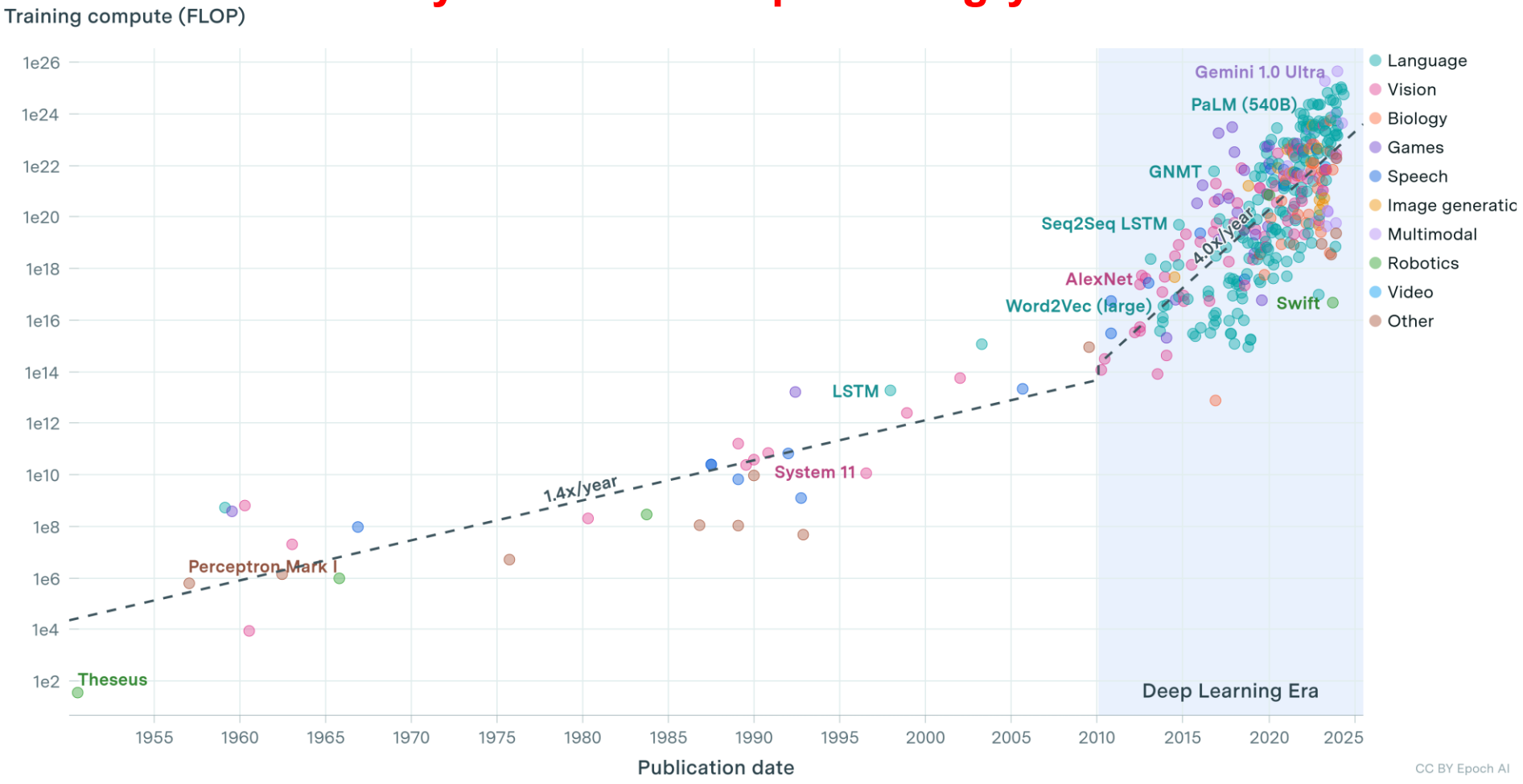
Chris Lu[1,2,*], **Cong Lu**[3,4,*], **Robert Tjarko Lange**[1,*], **Jakob Foerster**[2,†], **Jeff Clune**[3,4,5,†] **and David Ha**[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

One of the grand challenges of artificial general intelligence is developing agents capable of conducting scientific research and discovering new knowledge. While frontier models have already been used as aides to human scientists, e.g. for brainstorming ideas, writing code, or prediction tasks, they still conduct only a small part of the scientific process. This paper presents the first comprehensive framework for fully *automatic scientific discovery*, enabling frontier large language models (LLMs) to perform research independently and communicate their findings. We introduce THE AI SCIENTIST, which generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, and then runs a simulated review process for evaluation. In principle, this process can be repeated to iteratively develop ideas in an open-ended fashion and add them to a growing archive of knowledge, acting like the human scientific community. We demonstrate the versatility of this approach by applying it to three distinct subfields of machine learning: diffusion modeling, transformer-based language modeling, and learning dynamics. Each idea is implemented and developed into a full paper at a meager cost of less than $15 per paper, illustrating the potential for our framework to democratize research and significantly accelerate scientific progress. To evaluate the generated papers, we design and validate an automated reviewer, which we show achieves near-human performance in evaluating paper scores. THE AI SCIENTIST can produce papers that exceed the acceptance threshold at a top machine learning conference as judged by our automated reviewer. This approach signifies the beginning of a new era in scientific discovery in machine learning: bringing the transformative benefits of AI agents to the *entire* research process of AI itself, and taking us closer to a world where *endless affordable creativity and innovation* can be unleashed on the world's most challenging problems. Our code is open-sourced at https://github.com/SakanaAI/AI-Scientist.


Idea Generation
Experiment Iteration
Paper Write-Up
LLM Idea/Plan Innovation
Novelty Check Sem. Scholar
Idea scoring / archiving
Experiment Template
Code Δ via LLM & aider
Experiment Exec Script
Experiments
Update Plan
Numerical Data/Plots
Manuscript Template
Text Δ via LLM & aider
Manuscript
LLM Paper Reviewing

UL FRI

AI systems are compute hungry



LLMs are data hungry

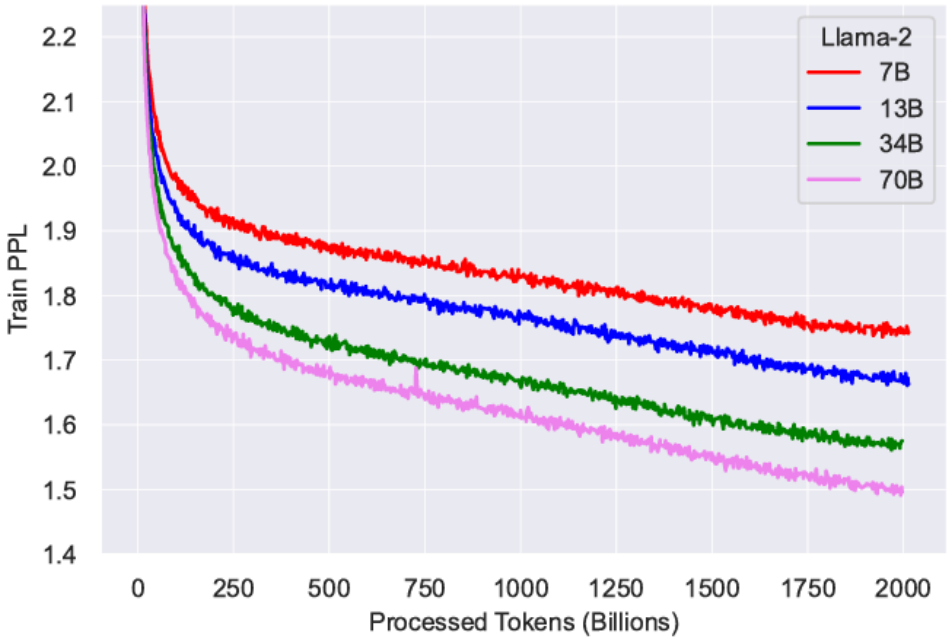


Figure 5: Training Loss for Llama 2 models. We compare the training loss of the Llama 2 family of models. We observe that after pretraining on 2T Tokens, the models still did not show any sign of saturation.

Ethical dilemmas of using AI

- bias and fairness: social and data biases
- quality of training data
- privacy and security of data
- responsibility of owners and developers
- transparency and explainability
- updating models (legal, content-based)

Responsible use of AI in research

EU prepared a draft of the guidelines

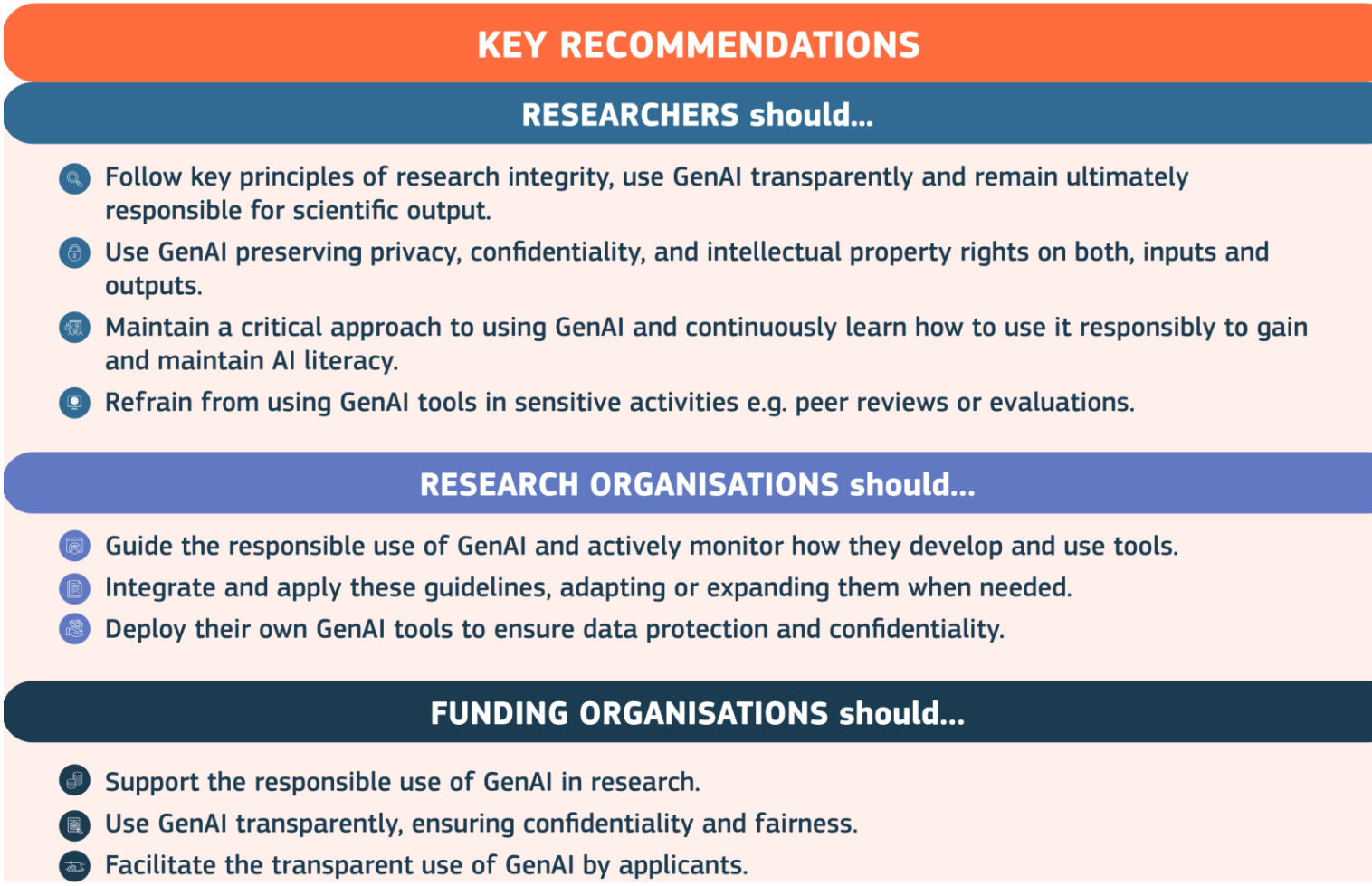
1. The researcher remains ultimately responsible for scientific output.
2. The researcher ensures fully transparency in the use of Generative AI (including tool, version, prompts, etc.).
3. The researcher pays attention to privacy and the sensitive/protected information they could share with AI tools.
4. As in their regular research activities, researchers must respect legislation applicable when using Generative AI.
5. Researchers need to learn and be trained regularly in the proper use of Generative AI tools in order to maximize the benefits from these tools.

UL FRI

EU guidelines for responsible use of AI in research

Live document

EU AI Act



EU AI Act

- Non-commercial research is exempt from regulation
- For research that can lead to marketable products, AI systems are classified into 4 categories:
  - **Unacceptable risk**: AI apps that pose a serious threat to security or fundamental rights are prohibited.
  - **High risk:** AI systems that can have a significant impact on health, safety or fundamental rights must meet strict requirements, including quality, transparency, human oversight, and safety standards.
  - **Limited risk**: These systems have certain transparency requirements so that users know that they are interacting with artificial intelligence.
  - **Minimal risk**: Most AI applications fall into this category and face minimal regulatory requirements.
- AI research targeting high-risk applications must take these requirements into account at an early stage of development.
- The act also includes provisions for general-purpose AI systems, such as underlying models (e.g., ChatGPT). These systems are subject to transparency requirements, and those with higher risk must undergo more detailed assessments. Researchers developing such models should be mindful of these obligations, especially if their system could have a major impact in different sectors.
- Researchers need to be aware of risk classifications and anticipate what requirements will need to be met if they want to **commercialize** their AI system.
- The act emphasizes the importance of **ethical development of AI**, so researchers must consider ethical considerations during development.

Slovenska tovarna umetne intelligence
Slovene AI factory - SLAIF

- namen: razširiti ekosistem umetne inteligence v Sloveniji
- povezava s superračunalniškim omrežjem Euro HPC
- strojna oprema v Sloveniji: superračunalnik Vega 2
- podporni projekt Slo/EU za prenos znanja v znanost in industrijo v okviru programa Obzorje 2020, 2025–2028
- razvoj talentov in znanja
- izobraževalni programi
- dostop do podatkov in upravljanje podatkov
- podpiranje znanosti in podjetij pri uporabi umetne inteligence
- UL FRI je vodilni partner na področju HPC, LLM in računalniškega vida

NLP application examples

Sentiment analysis (SA)

- Definition: a computational study of opinions, sentiments, emotions, and attitudes expressed in texts towards an entity.
- Purpose: detecting public moods, i.e. understanding the opinions of the general public and consumers on social events, political movements, company strategies, marketing campaigns, product preferences etc.
- Part of Affective Computing (emotion, mood, personality traits, interpersonal stance, attitude)
- Can be target-based or general

SA: getting and preprocessing data

- Frequent data sources:
  - Twitter-X, forum comments, product review sites, company's Facebook pages
- Data cleaning
  - quality assessment, annotator (self-) agreement
  - preprocessing: tokenization, emojis, links, hashtags, etc,

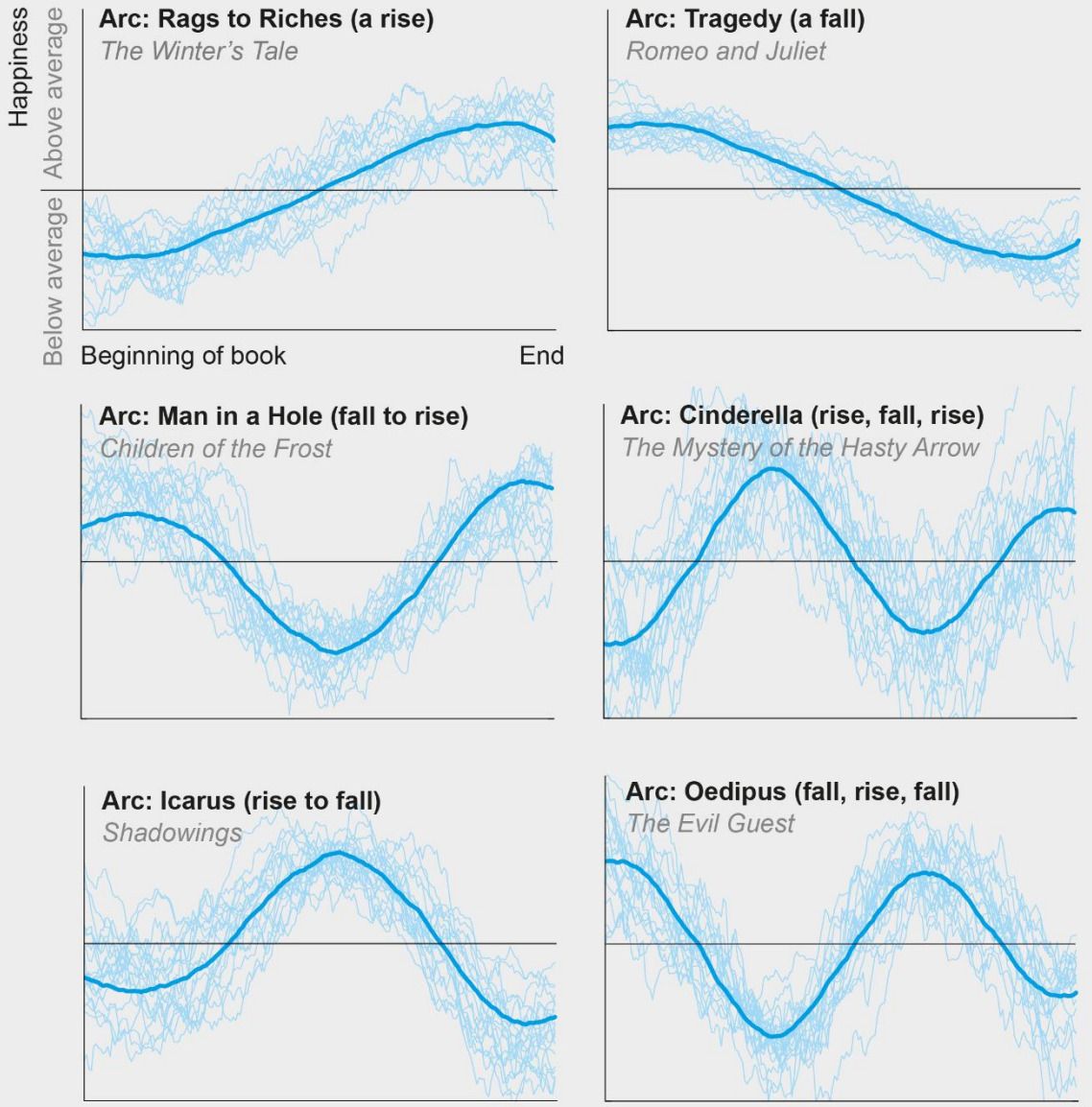
SA tasks

- sentiment classification (binary (polarity), ternary, n-ary
- subjectivity classification (vs. objectivity)
- review usefulness classification
- opinion spam classification
- emotion analysis
- hate speech, offensive speech, socially unacceptable speech
- stance detection

Emotional states in English fiction

Emotional Arcs

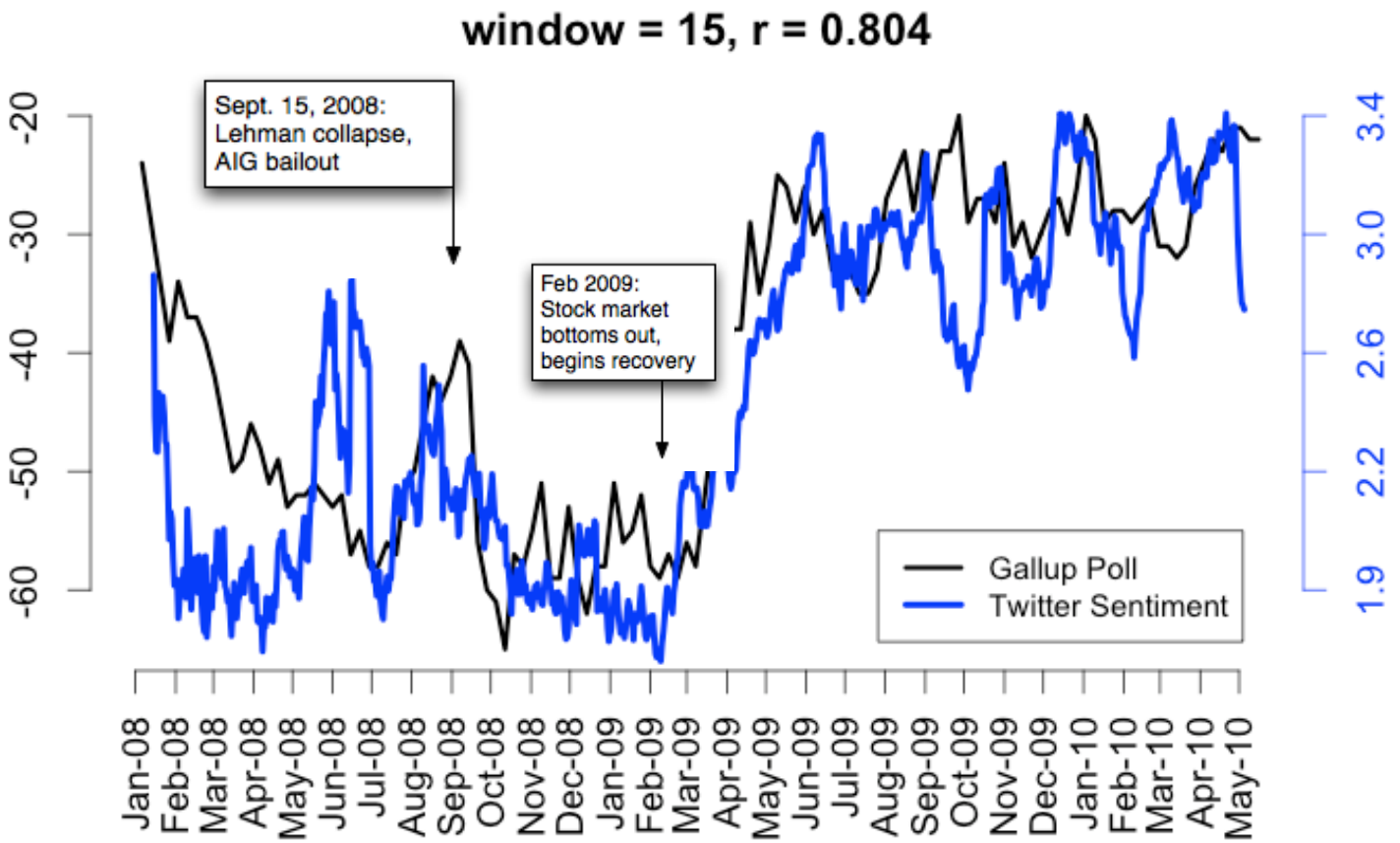
About 85 percent of 1,327 fiction stories in the digitized Project Gutenberg collection follow one of six emotional arcs—a pattern of highs and lows from beginning to end (*dark curves*). The arcs are defined by the happiness or sadness of words in the running text (*jagged plots*). All books were in English and less than 100,000 words; examples are noted.



Public opinion surveys

Twitter sentiment vs. Gallup on consumer sentiment

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010



Statistical machine translation

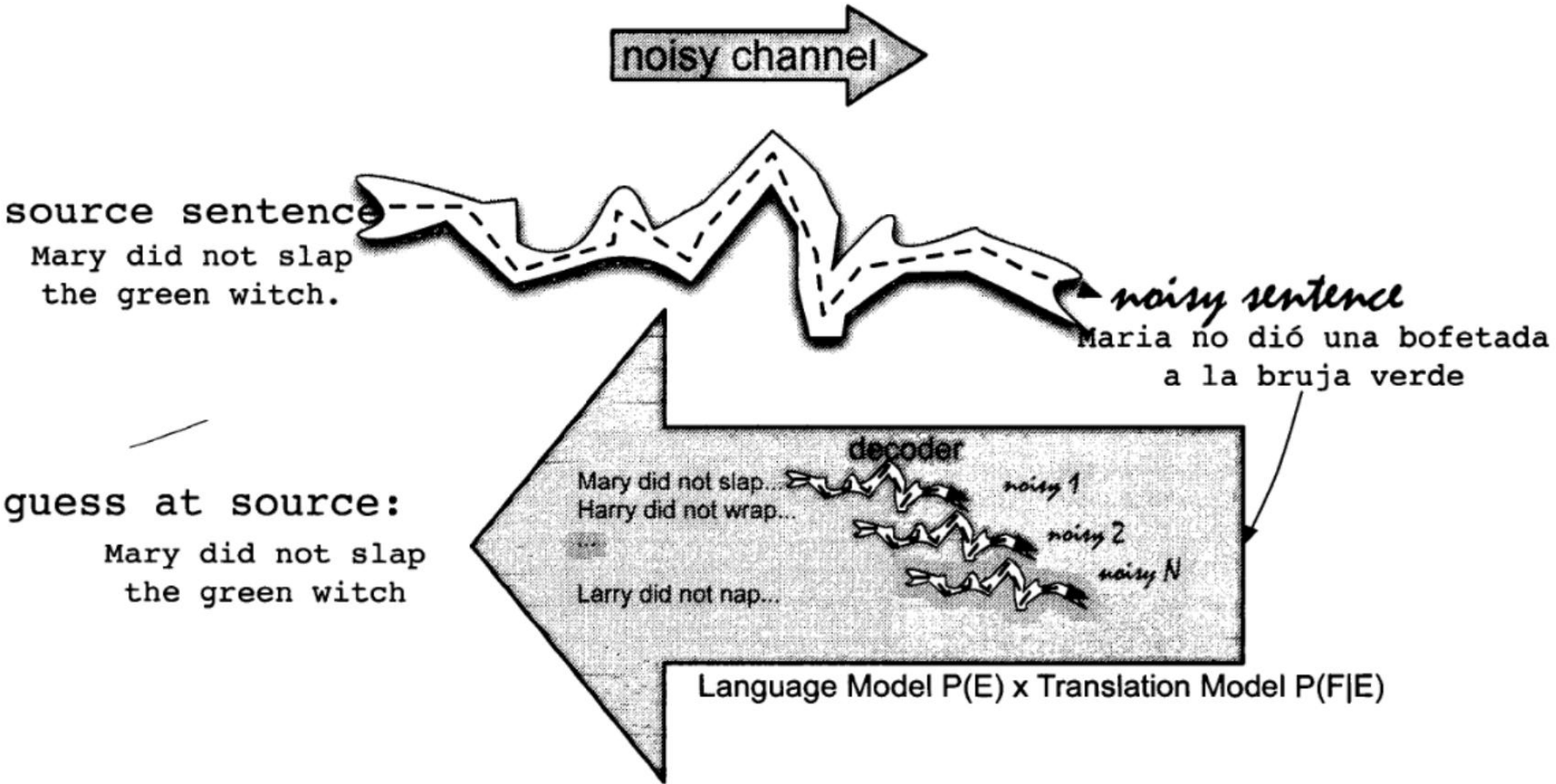
- non-neural approach: no longer used in practice but gives insight of what is needed
- idea from the theory of information
- we translate from foreign language F to English E
- a document is translated based on the probability distribution $p(e \mid f)$, i.e. the probability of the sentence e in target language based on the sentence in source language f
- Bayes rule
  $\arg\max_e p(e \mid f) = \arg\max_e p(f \mid e)\, p(e) \,/\, p(f)$
- $p(f)$ can be ignored as it is a constant for a given fixed sentence
- traditional (non-neural) approaches split the problem into subproblems
  - create a language model $p(e)$
  - a separate translation model $p(f \mid e)$
  - decoder forms the most probable e based on f

Noisy channel model

- given English sentence e
- during transmission over a noisy channel the sentence e is corrupted and we get sentence in a foreign language f, which we are able to observe
- to reconstruct the most probable sentence e we have to figure out:
  - how people speak in English (language model), p(e) and
  - how to transform a foreign language into English (translation model), p(f | e)

Noisy channel

- reasoning goes back



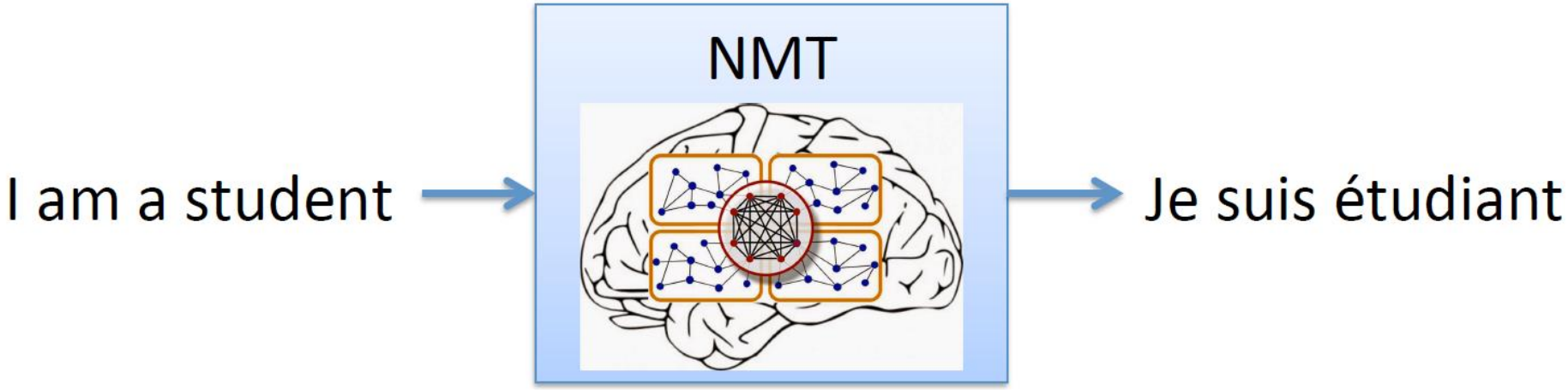
Language model

- each target (English) sentence e is assigned a probability p(e)
- estimation of probabilities for the whole sentences is not possible (why?), therefore we use language models, e.g., 3-gram models or neural language models

Translation model

- we have to assign a probability of p(f | e), which is a probability of a foreign language sentence f, given target sentence e.
- we search the e which maximizes p(e) * p(f | e)
- traditional MT approach: using translation corpus we determine which translation of a given word is the most probable
- we take into account the position of a word and how many words are needed to translate a given word

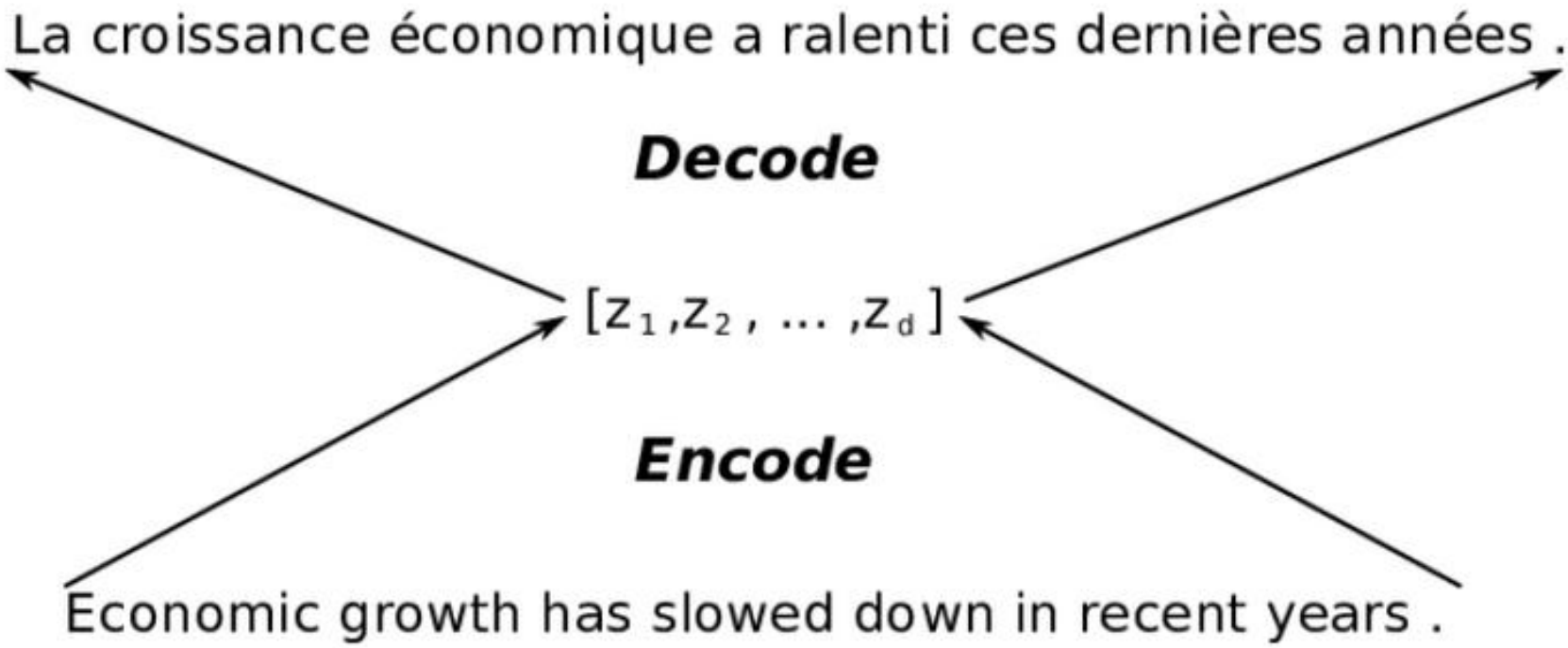
Neural machine translation (NMT)



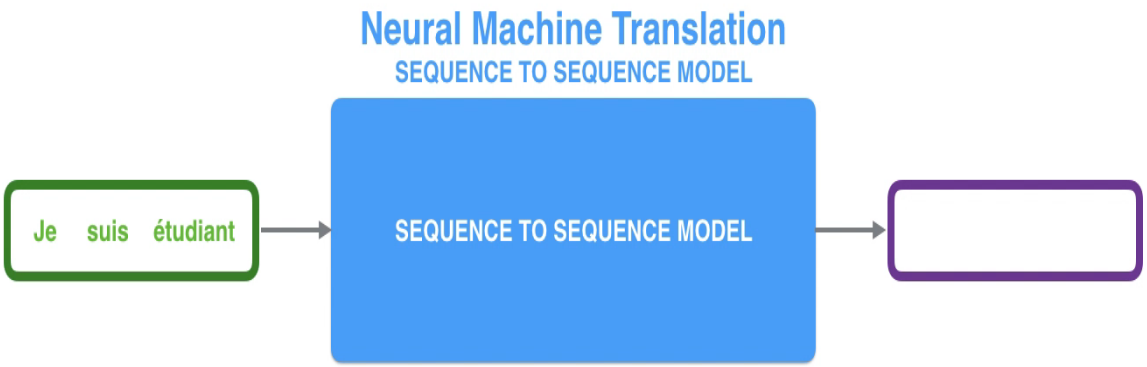
(Sutskever et al., 2014; Cho et al., 2014)

- sequence to sequence machine translation (seq2seq)
- end-to-end optimization

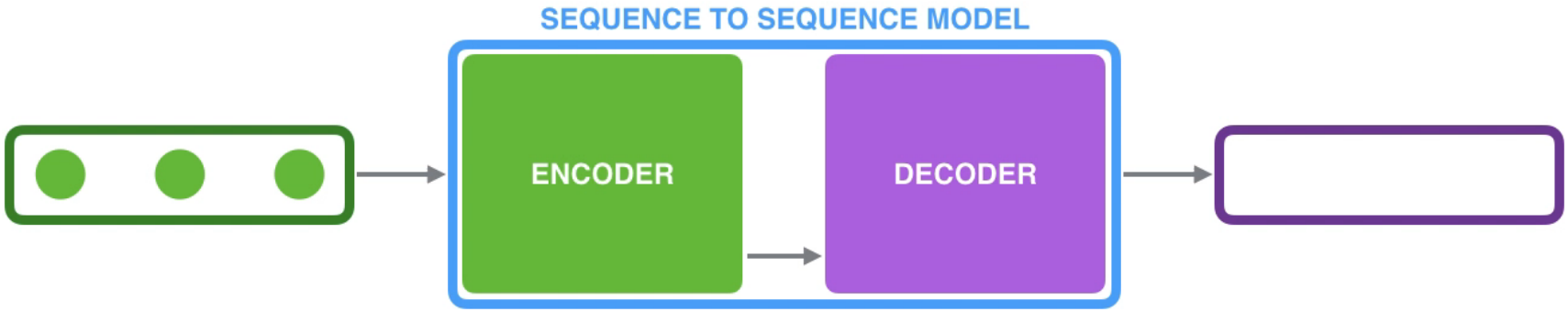
Encoder-Decoder model



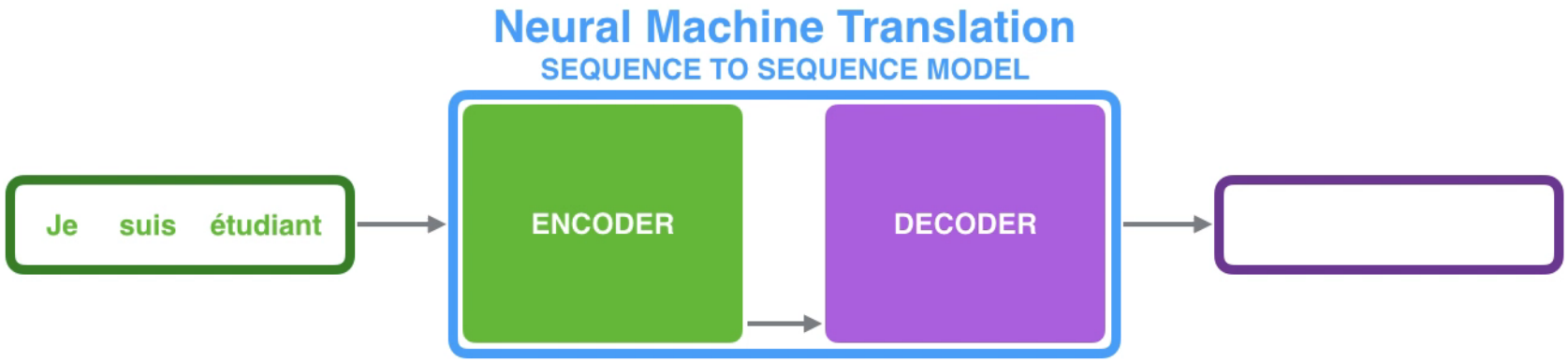
Seq2Seq for NMT

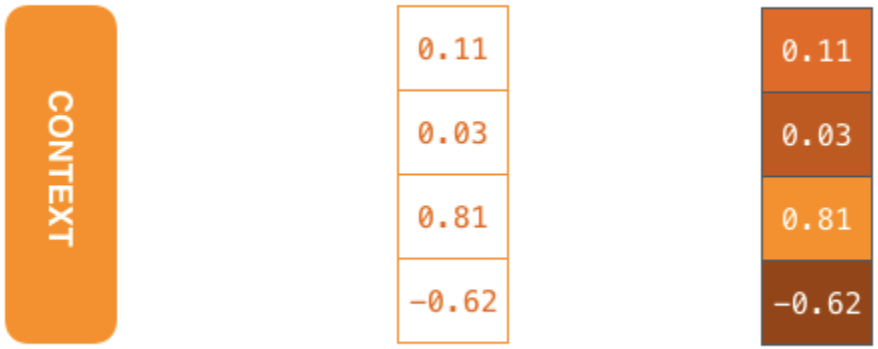


Encoder-decoder for sequences



Encoder-decoder for NMT

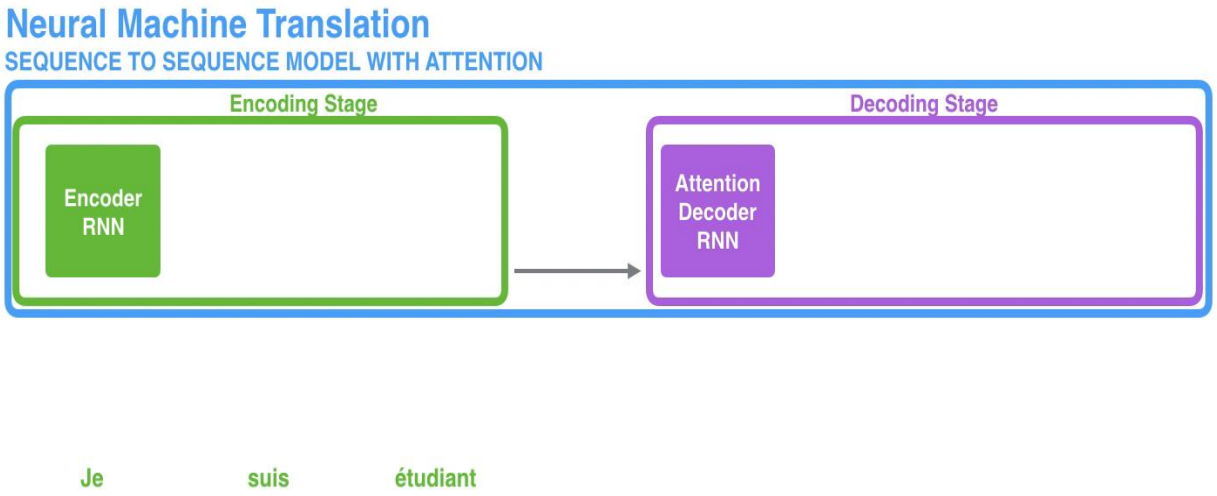




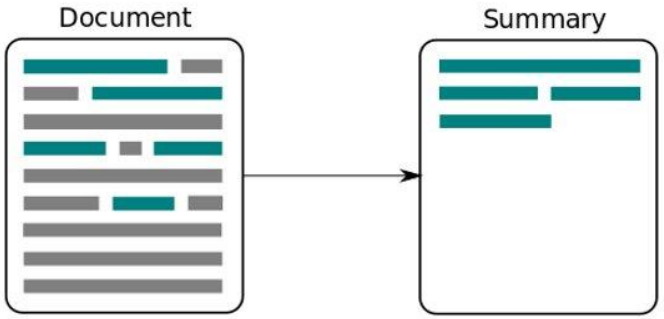
Training NMT

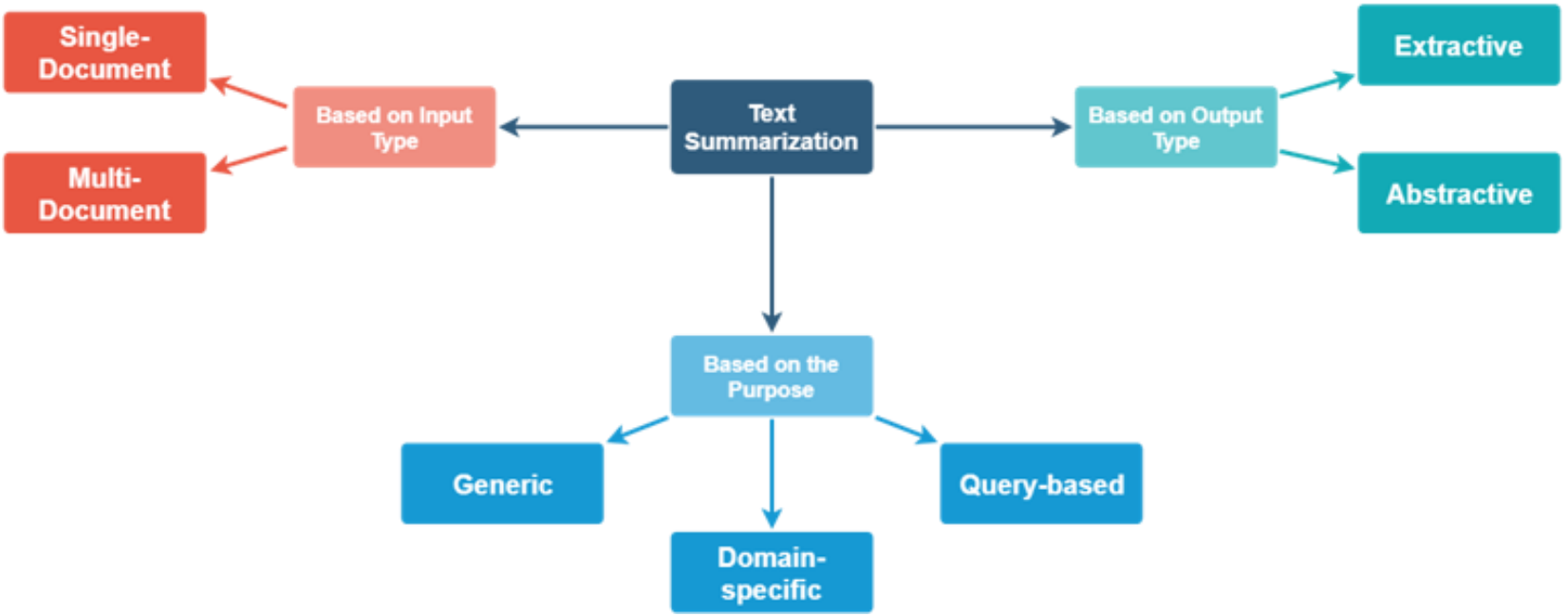
- using transformers
- softmax for output
- we maximize
  P(output sentence | input sentence)
- we sum errors on all outputs
- backpropagation
- training on correct translations
- as the translation, we return a sequence of words with the highest probability (not necessary greedily)

NMT with attention



Text summarization



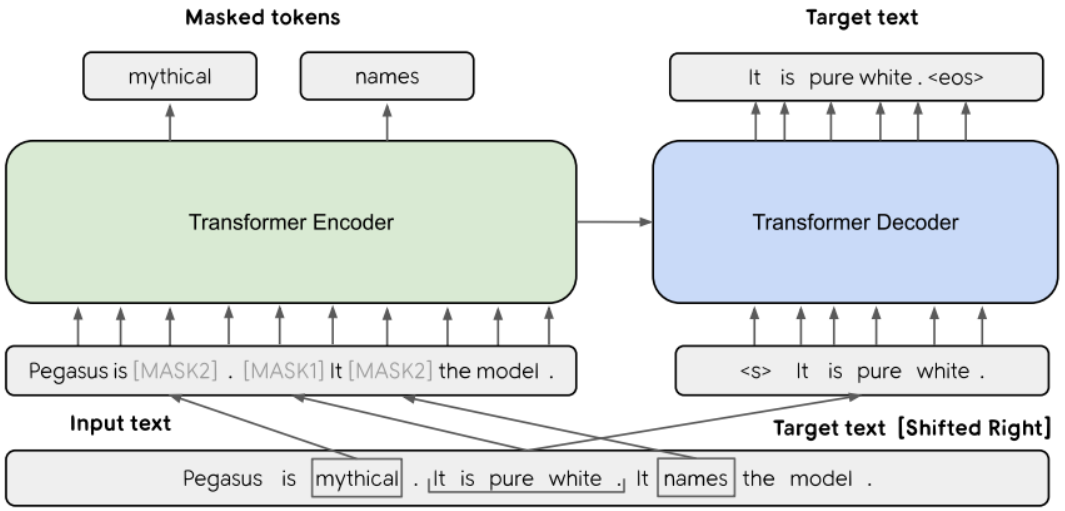


Text summarization

- Evaluation:
  - ROUGE scores,
  - BERTScore,
  - with QA:
    - question generation,
    - searching for answers in the summary
  - human
- LLMs
- Short and long texts
- cross-lingual

Summarizers – Pegasus

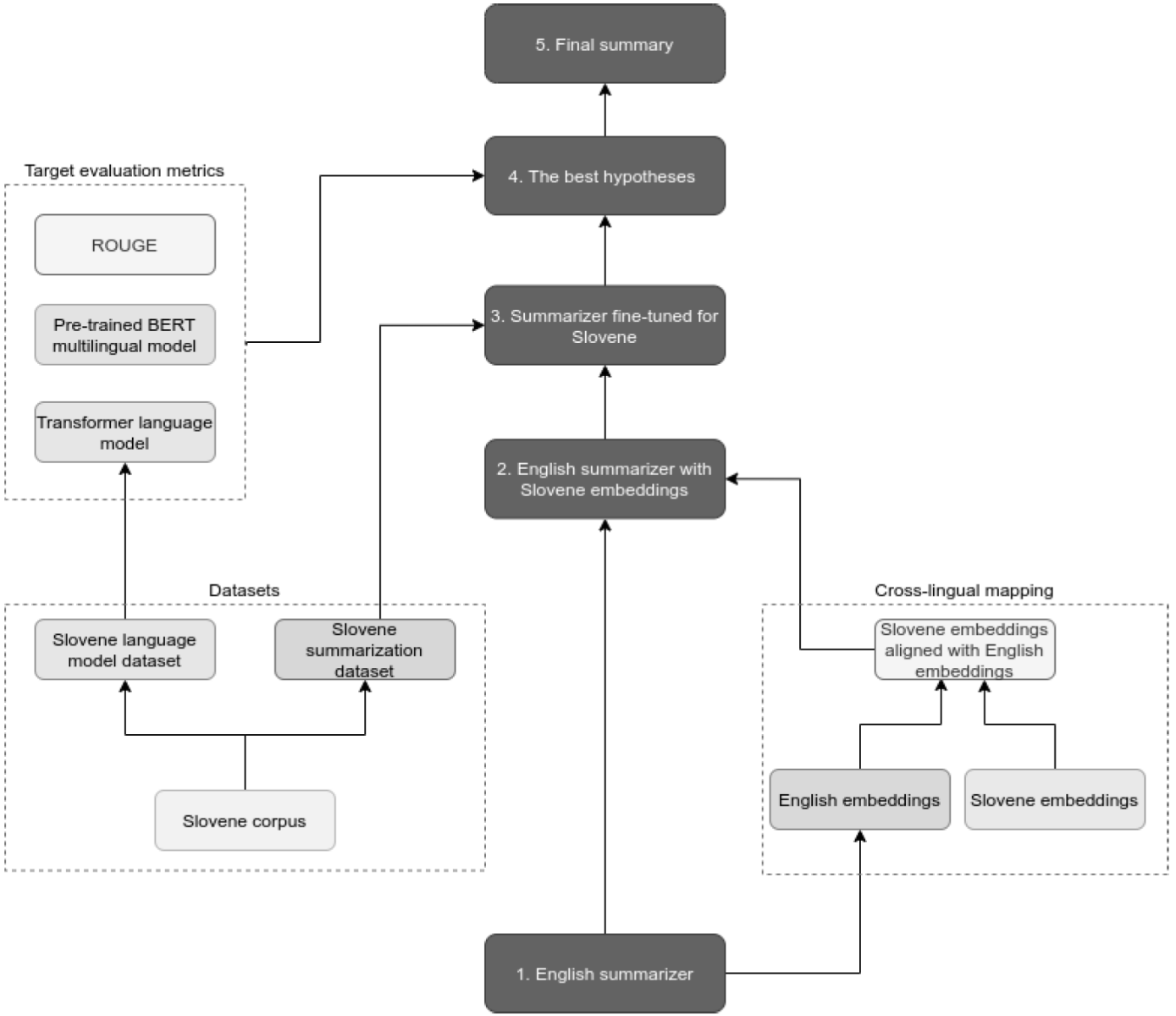
- An example of a different, successful transformer model
- Transformer BART model
- encoder-decoder architecture
- text garbling and reconstruction
- Auxiliary tasks: masked language model and missing sentence generation
- Demo: https://ai.googleblog.com/2020/06/pegasus-state-of-art-model-for.html



(Zhang et al., 2019)

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

XL summarization architecture



Žagar, A., & Robnik-Šikonja, M. (2022). Cross-lingual transfer of abstractive summarizer to less-resource language. Journal of Intelligent Information Systems, 58(1), 153-173.