Does MT Understand?

Word and Sentence Representations

Ondřej Bojar

📅 April 23, 2020

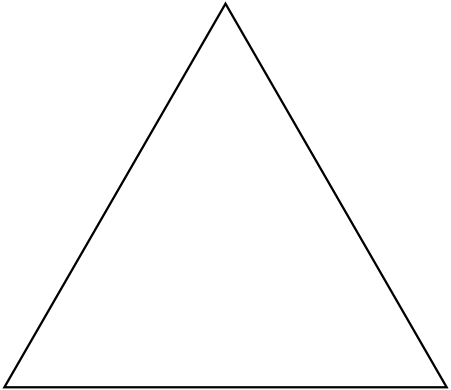
EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

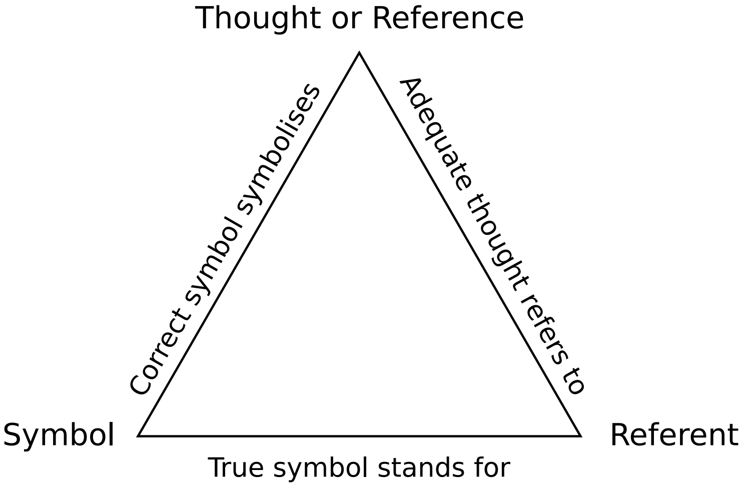
CC BY SA
unless otherwise stated

Outline

- Introducing Semiotics.
- Do Current MT Systems Understand?
- Continuous Representations.
  - What are Good Representations?
  - Continuous Word Representations.
  - Continuous Sentence Representations.
- Aspects of Meaning.
- Evaluating Sentence Representations
  - How Meaningful is Seq2Seq Representation?

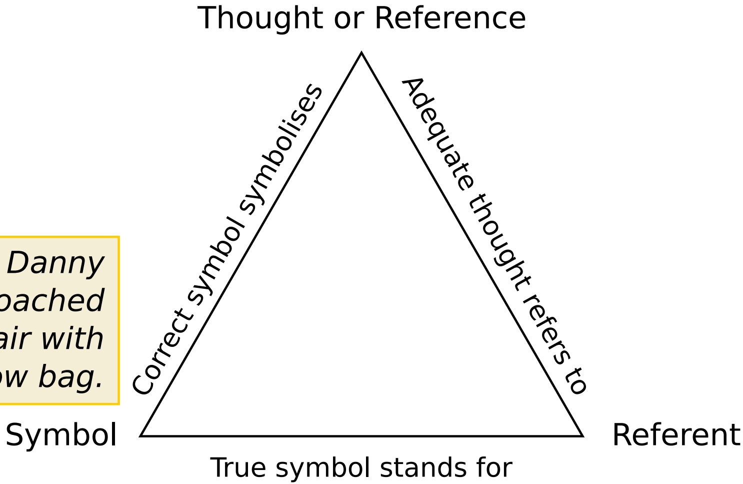


Semiotic Triangle



Semiotic Triangle by Ogden and Richards (1923).

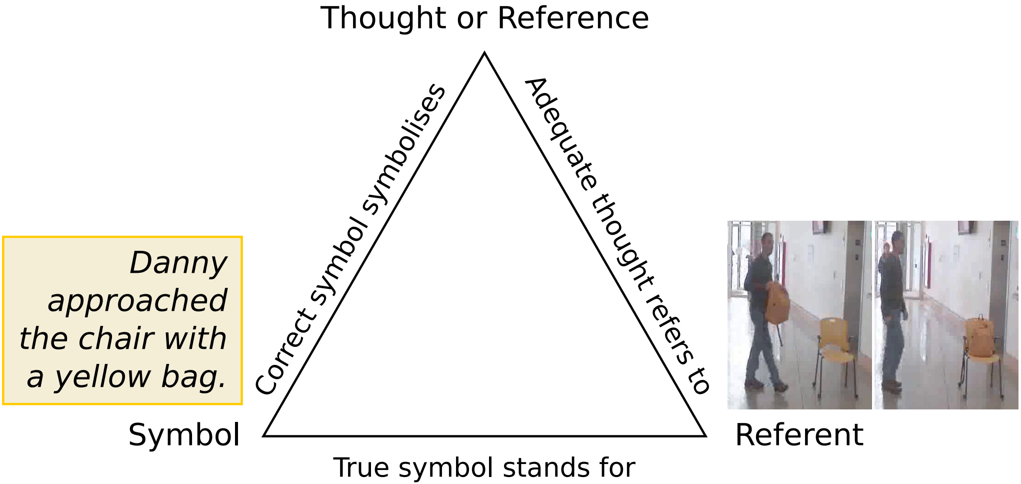
Semiotic Triangle



Danny approached the chair with a yellow bag.

Ambiguous sentence...

Semiotic Triangle



Thought or Reference

Correct symbol symbolises

Adequate thought refers to

Danny approached the chair with a yellow bag.

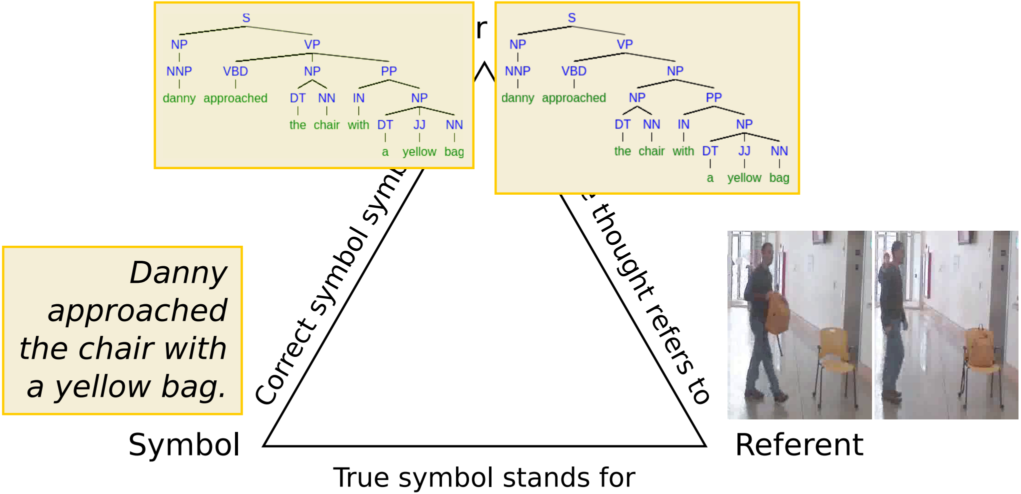
Symbol

Referent

True symbol stands for

Ambiguous sentence correspond to two situations. LavaCorpus (Berzak et al., 2015)

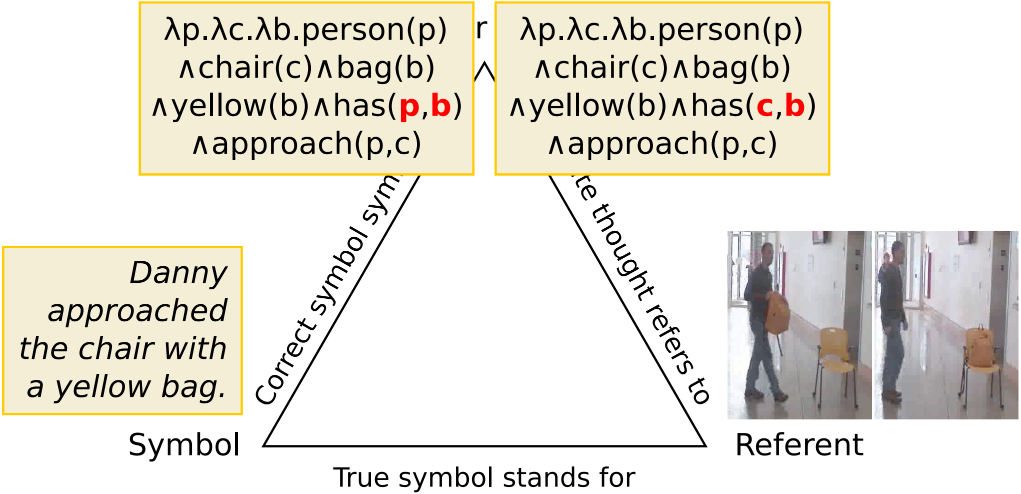
Semiotic Triangle



Danny approached the chair with a yellow bag.

Syntactic "meaning" distinguishes this already.

Semiotic Triangle



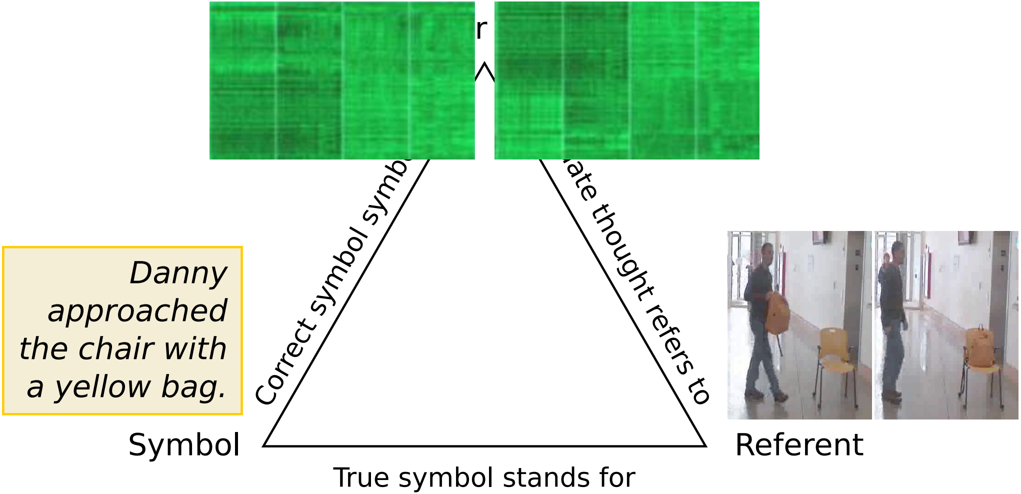
λp.λc.λb.person(p)
∧chair(c)∧bag(b)
∧yellow(b)∧has(**p**,**b**)
∧approach(p,c)

λp.λc.λb.person(p)
∧chair(c)∧bag(b)
∧yellow(b)∧has(**c**,**b**)
∧approach(p,c)

Danny approached the chair with a yellow bag.

Lambda calculus makes the difference clear.

Semiotic Triangle



Danny approached the chair with a yellow bag.

r

Correct symbol symb

ate thought refers to

Symbol

Referent

True symbol stands for

NN activations when processing the videos will somehow differ, too.

Why is Meaning Important in MT

Translation = expressing the same meaning in another language.

A meaning-aware translator (human or machine) will:

1. Use context to disambiguate as much as possible.
2. Ask around to learn about and understand the situation described.
3. Ideally warn the audience about unresolved ambiguities.

Recent Performance of NMT on News

Doc-Aware English→German 2019

Ave.	Ave. z	System
90.3	0.347	Facebook-FAIR
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	**Professional Translation**
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation
86.8	0.119	dfki-nmt
84.2	0.094	online-B
	... 10 more systems here ...	
76.3	−0.400	online-X
43.3	−1.769	en-de-task

Seg-Level English→Czech 2018

	Ave. %	Ave. z	System
1	84.4	0.667	**CUNI-Transformer**
2	79.8	0.521	UEDIN
	78.6	0.483	**Professional Translation**
4	68.1	0.128	ONLINE-B
5	59.4	−0.178	ONLINE-A
6	54.1	−0.354	ONLINE-G

See lecture #1 for all caveats of MT evaluation.

Do Recent Best Systems Understand?

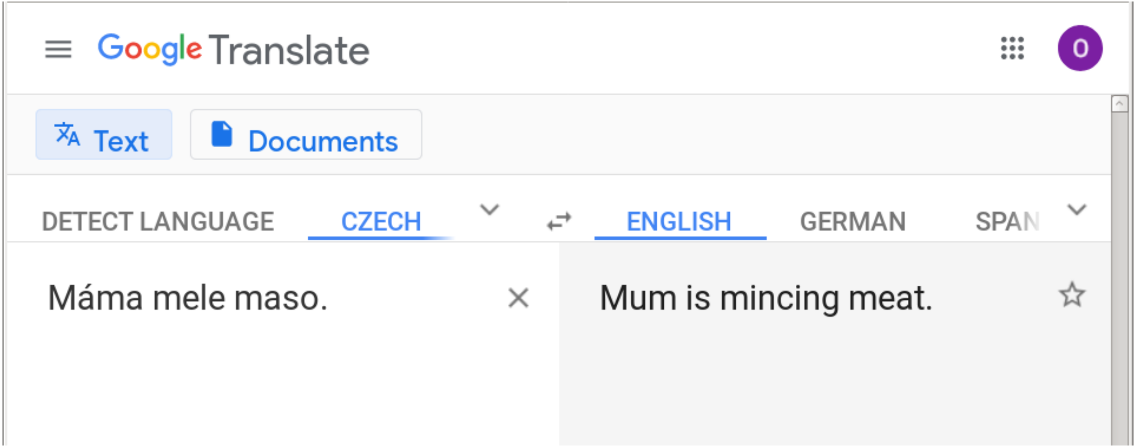
- NMT systems are trained on millions of documents.
    - To read the source and target training data of **CUNI-Transformer** you would need **50 years**, 8 hours a day, no weekends.
      (Only 40% of it was parallel.)
- NNs create internal representations.

    … So perhaps these representations are meaningful?

Test it yourself (English↔Czech):

https://lindat.mff.cuni.cz/services/transformer/

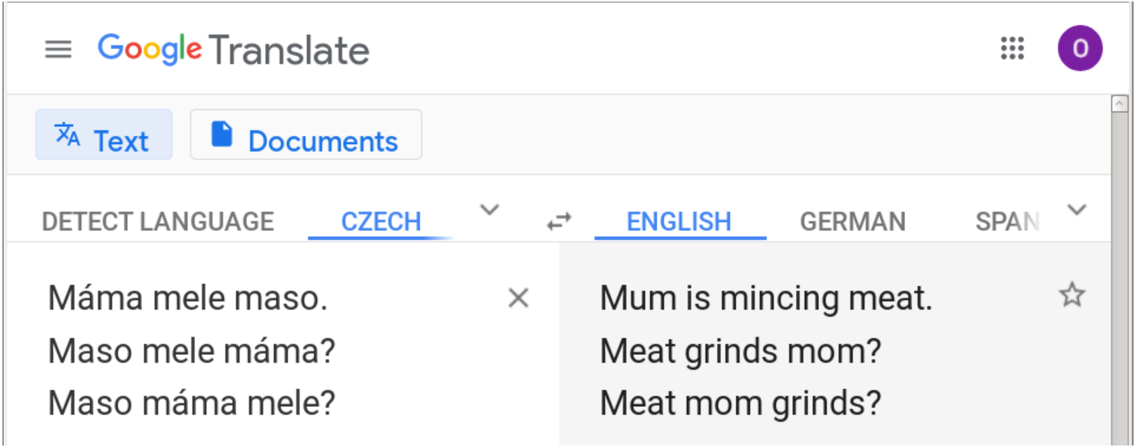
Do Recent Best Systems Understand?



Máma mele maso.

Mum is mincing meat.

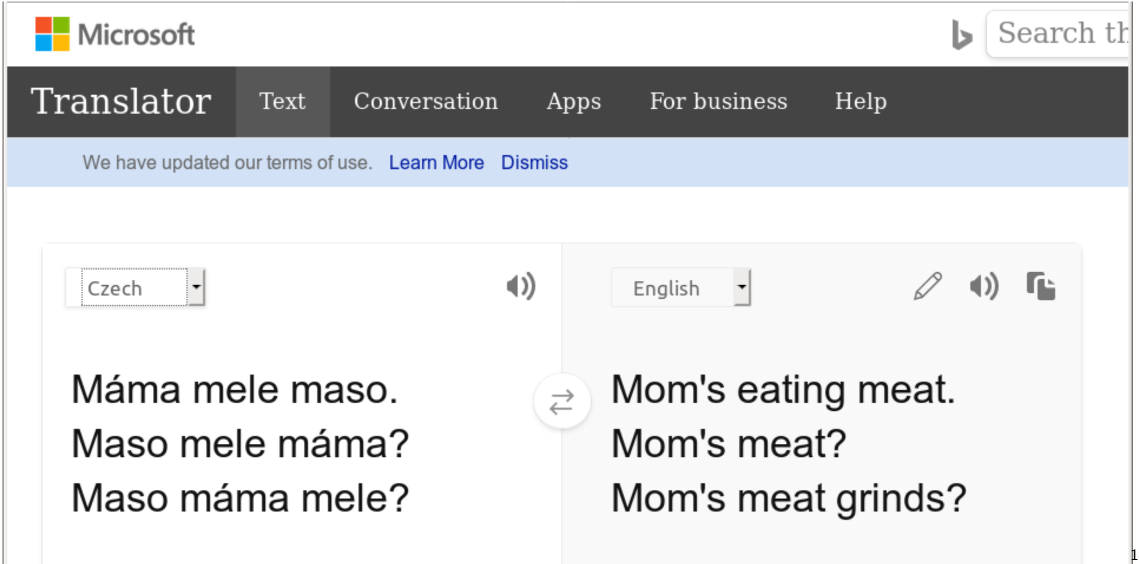
Do Recent Best Systems Understand?



Máma mele maso.
Maso mele máma?
Maso máma mele?

Mum is mincing meat.
Meat grinds mom?
Meat mom grinds?

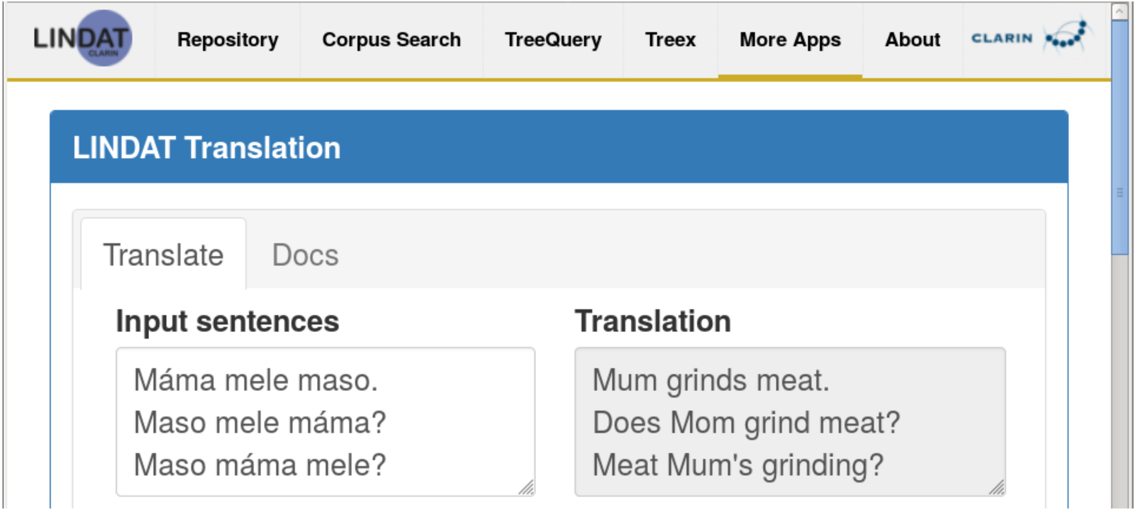
Do Recent Best Systems Understand?



We have updated our terms of use. Learn More Dismiss

Czech	English
Máma mele maso.	Mom's eating meat.
Maso mele máma?	Mom's meat?
Maso máma mele?	Mom's meat grinds?

Do Recent Best Systems Understand?



LINDAT Translation

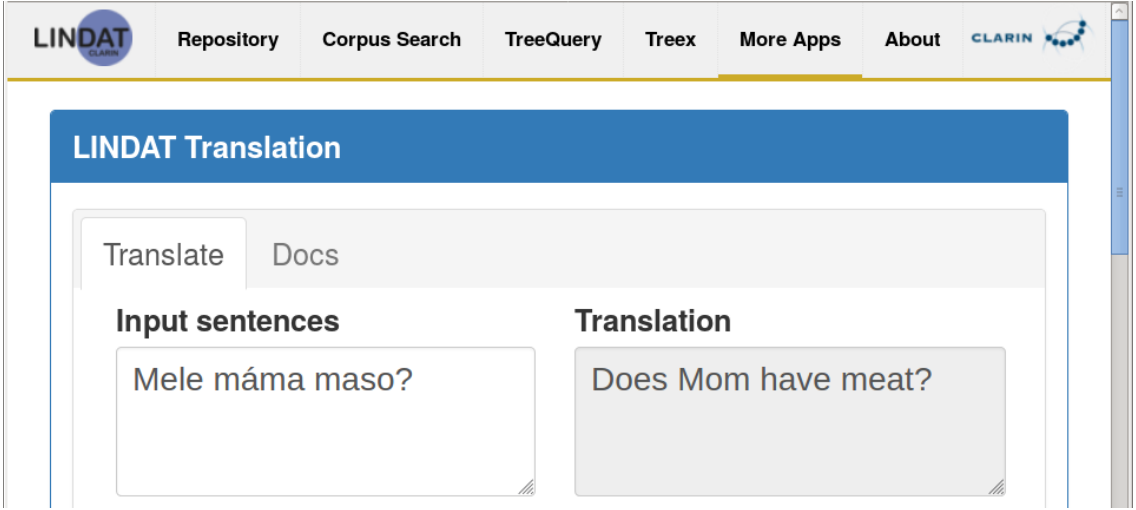
Input sentences

Máma mele maso.
Maso mele máma?
Maso máma mele?

Translation

Mum grinds meat.
Does Mom grind meat?
Meat Mum's grinding?

Do Recent Best Systems Understand?



LINDAT Translation

Input sentences

Mele máma maso?

Translation

Does Mom have meat?

Representations

Defining Representations

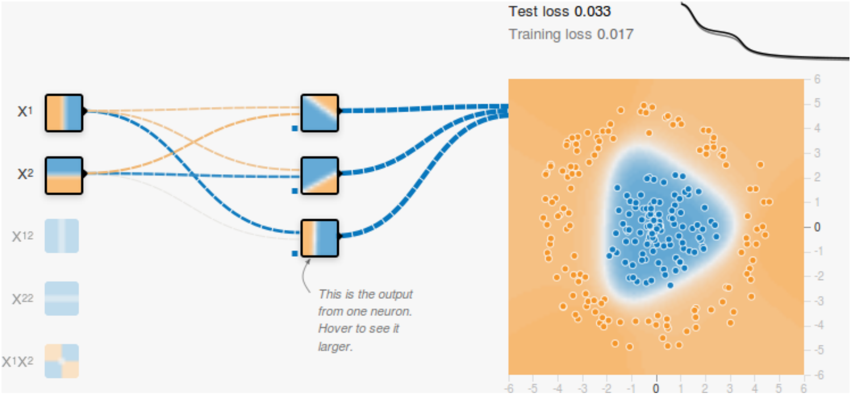
Given:

- a neural network trained to predict $\hat{y}_i \in \mathcal{Y}$ given $x_i \in \mathcal{X}$,
- and a CUT C of that network
    - (a set of neurons s.t. every path from input to output has to intersect it),

a REPRESENTATION is the mapping from $\mathcal{X}$ to $\mathcal{H}$, where

- $\mathcal{H}$ is the vector space of observed activations of neurons in C (in some arbitrary fixed order).

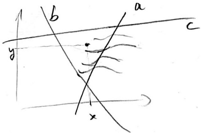
Two Cuts Here: (1) Input, (2) Hidden Layer

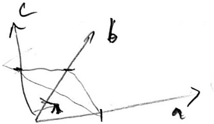


Test loss 0.033

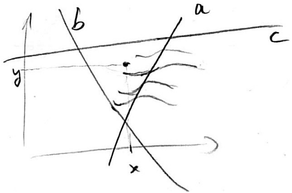
Training loss 0.017

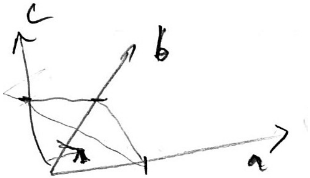
This is the output from one neuron. Hover to see it larger.





The Learned Representation





Original space allows to:

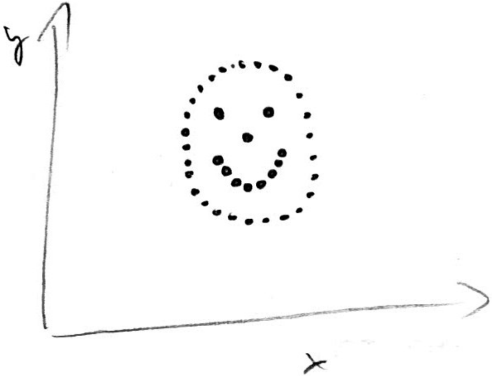
- plot input data,
- visualize separation boundaries for the first as well as subsequent layers.

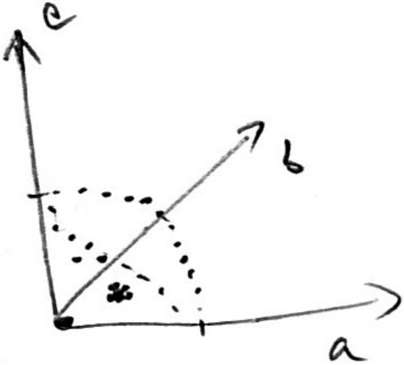
Hidden space $\mathcal{H}$ allows to:

- to linearly separate the classes.

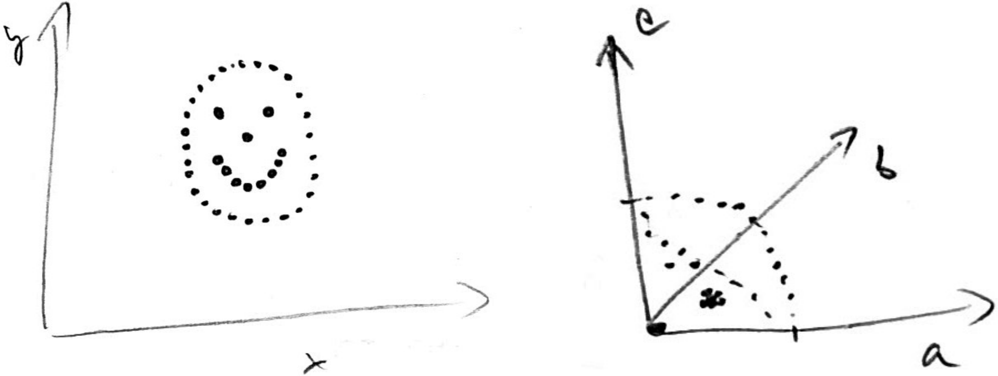
… but is it good for anything else?

Good Representations (1/2)





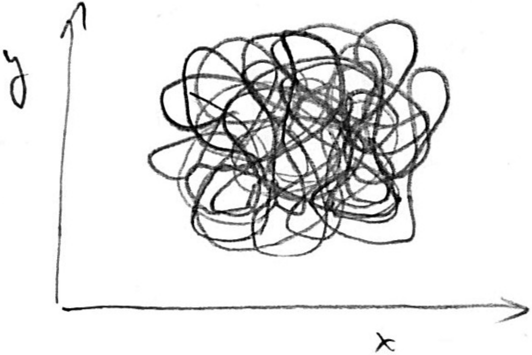
Good Representations (1/2)

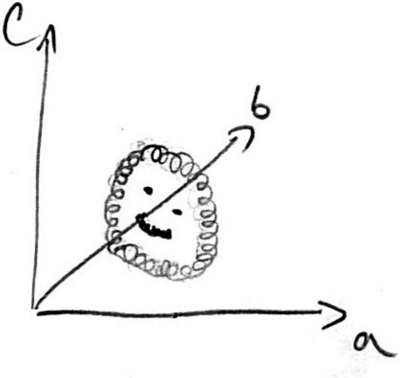


(a, b, c) is a good representation

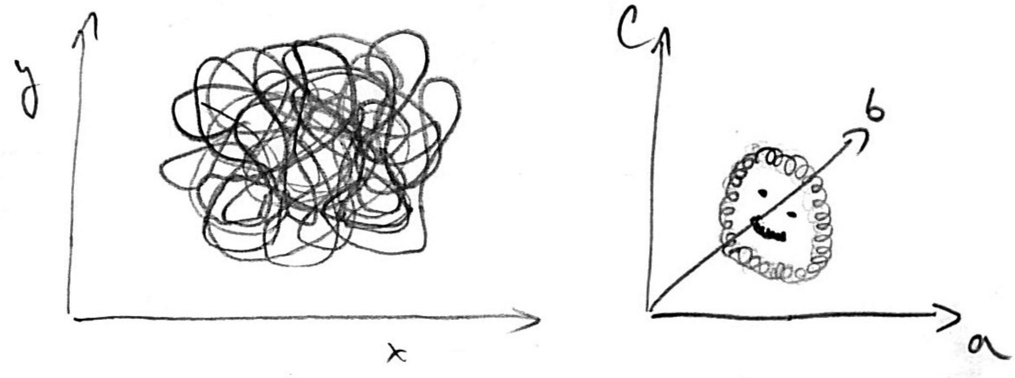
because it separates border from face features

Good Representations (2/2)





Good Representations (2/2)



(a, b, c) is a good representation
because it resembles a known picture

Which Representations Are Good?

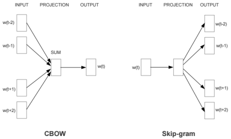
Ideas for a start:

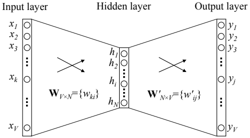
- Good representations allow to solve some main task.
  - … but for this, the NN was trained in the first place.
- Good representations allow to solve some other task.
  - Pretrained word embeddings may be useful for other tasks.
- Good representations serve well on task interfaces.
  - Divide-and-conquer vs. end-to-end training.
  - Must divide training to make use of different data sources (e.g. spoken language translation needs ASR and MT data).
- Good representations "make sense".
  - The representation of a specific TEST SET resembles something known.
  - Attaching a single layer to the representation gives a good accuracy in something, e.g. part-of-speech tags from sentence embeddings.

Word Representations

Word Embeddings

- Map each word to a dense vector.
- In practice 300–2000 dimensions are used.
    - The dimensions have no clear interpretation.
- Embeddings are trained for each particular task.
    - NNs: The matrix that maps 1-hot input to the first layer.
- The famous word2vec (Mikolov et al., 2013):
    - CBOW: Predict the word from its four neighbours.
    - Skip-gram: Predict likely neighbours given the word.



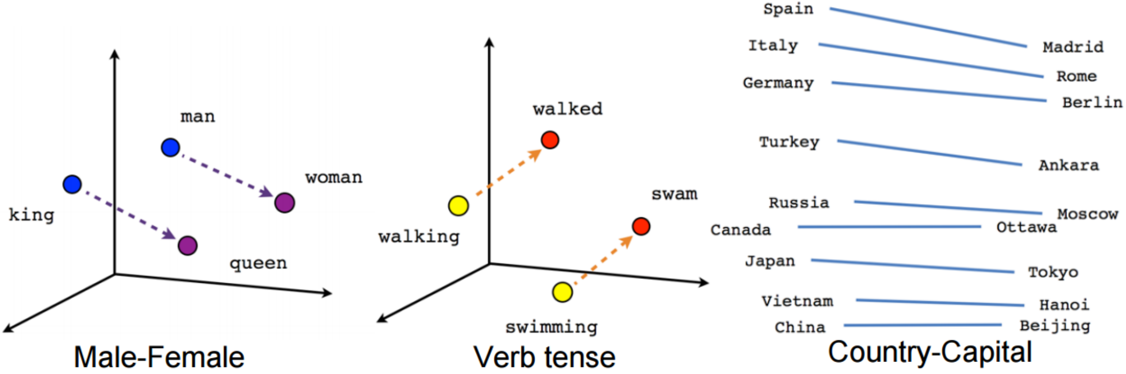


Right: CBOW with just a single-word context (`http://www-personal.umich.edu/~ronxin/pdf/w2vexp.pdf`)

Emergent Continuous Space of Words

Word2vec embeddings show interesting properties:

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen}) \tag{1}$$



Illustrations from `https://www.tensorflow.org/tutorials/word2vec`

Testset by Mikolov et al. (2013)

Question Type	Sample Pair
capital-countries	Athens – Greece
capital-world	Abuja – Nigeria
currency	Algeria – dinar
city-in-state	Houston – Texas
family	boy – girl
adjective-to-adverb	calm – calmly
opposite	aware – unaware
comparative	bad – worse
superlative	bad – worst
present-participle	code – coding
nationality-adjective	Albania – Albanian
past-tense	dancing – danced
plural	banana – bananas
plural-verbs	decrease – decreases

Caveat on Evaluation (1/2)

Consider word2vec "comprehensive" test set (Mikolov et al., 2013):

- 8.8k "semantic" and 10.6k "syntactic" questions,
- w2v "accuracy is quite good" (eyeballing)
  - The authors do mention that exact-match is "only about 60%").

Kocmi and Bojar (2016) carefully examined the test set:

- "Semantic" questions cover only 3 question types:
  - country→city, country→currency, masculine family member→ feminine
  - Vylomova et al. (2016) test many other relations, e.g. walk-run, dog-puppy, bark-dog, cook-eat.
- "Syntactic" questions constructed by combinations:
  - starting from only 313 distinct word pairs,
  - (leading to only 35 different pairs per question on average),
  - And of the 313 pairs, 286 are formed regularly.

Caveat on Evaluation (2/2)

	Test Set by	
Accuracy on “Synt Questions”	Mikolov et al.	Kocmi et al.
word2vec as released	62.5%	43.5%

Caveat on Evaluation (2/2)

	Test Set by	
Accuracy on “Synt Questions”	Mikolov et al.	Kocmi et al.
word2vec as released	62.5%	43.5%
word2vec trained on our data	42.5%	9.7%
SubGram trained on our data	42.3%	22.4%

Caveat on Evaluation (2/2)

	Test Set by	
Accuracy on “Synt Questions”	Mikolov et al.	Kocmi et al.
word2vec as released	62.5%	43.5%
word2vec trained on our data	42.5%	9.7%
SubGram trained on our data	42.3%	22.4%
Nine rules	**71.9%**	**66.4%**

Caveat on Ultimate Evaluation

Kocmi and Bojar (2016):

- submitted to TSD on March 22, 2016.
- appeared in TSD in September 2016.

... cited by 7.

Bojanowski et al. (2017):

- submitted to arxiv on July 15, 2016.
- appeared in TACL 2017.

... cited by 2994.

Caveat on Ultimate Evaluation

Kocmi and Bojar (2016):

- submitted to TSD on March 22, 2016.
- appeared in TSD in September 2016.

... cited by 7.

- No code released, no fast code implemented at all.

Bojanowski et al. (2017):

- submitted to arxiv on July 15, 2016.
- appeared in TACL 2017.

... cited by 2994.

- This is the FastText paper.

Evaluating Words against Human Assessment?

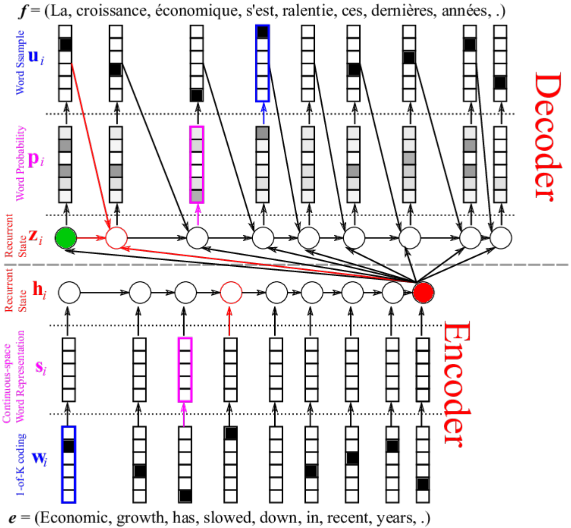
The whole idea of evaluating word vectors by relating to human judgements is risky.

- Human-produced datasets are subjective.
- Similarity vs. relatedness.
    - Relatedness: *teacher* ≈ *student*, *coffee* ≈ *cup*
    - Similarity: *teacher* ≈ *professor*, *car* ≈ *train*
    - Hill et al. (2017) observed a soft tendency:
        - Monolingual models reflect non-specific relatedness,
        - NMT models reflect conceptual similarity.
        - We saw that too for English-Czech (Abdou et al., 2017).
    - Even if we distinguish them, which should be reflected in embeddings?

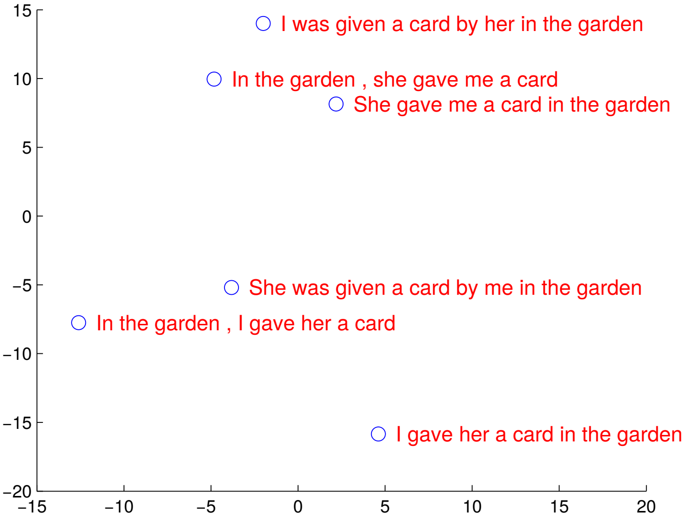
Details: Faruqui et al. (2016); Survey of eval. methods: Bakarov (2018)

Sentence Representations

Encoder-Decoder Architecture


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
u_i
Word Probability
p_i
Recurrent State
z_i
Decoder
Recurrent State
h_i
Continuous-space Word Representation
s_i
1-of-K coding
w_i
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Continuous Space of Sentences



2-D PCA projection of 8000-D space representing sentences (Sutskever et al., 2014).

Fixed-Length Representation of Sentences??

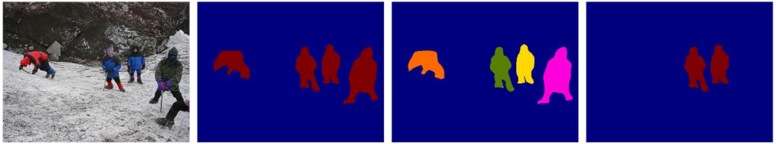
Raymond Mooney:

You can't cram the meaning of
a whole %&!$ing sentence into a single $&!*ing vector!

Aspects of Meaning

Aspects of Meaning (1/2)

- Meaning can be seen as a coarsening:
  - Pictures: Semantic segmentation.



(a) input image

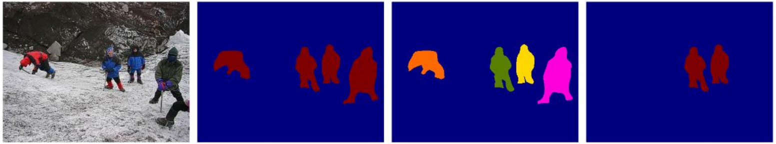
(b) object class segmentation of class ***people***

(c) object instance segmentation of class ***people***

(d) segmentation from expression *"people in blue coat"*

Aspects of Meaning (1/2)

- Meaning can be seen as a coarsening:
  - Pictures: Semantic segmentation.



(a) input image

(b) object class segmentation of class ***people***

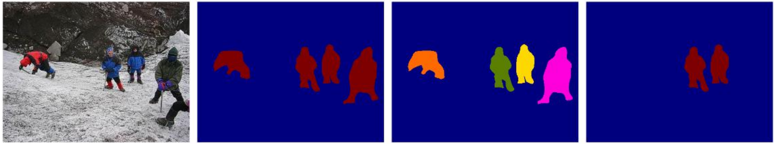
(c) object instance segmentation of class ***people***

(d) segmentation from expression *"people in blue coat"*

  - Programs: The output they give (caveat: undecidable).
  - Sentences: Reference to real world? Speaker's intention?

Aspects of Meaning (1/2)

- Meaning can be seen as a coarsening:
  - Pictures: Semantic segmentation.



(a) input image

(b) object class segmentation of class ***people***

(c) object instance segmentation of class ***people***

(d) segmentation from expression *"people in blue coat"*

  - Programs: The output they give (caveat: undecidable).
  - Sentences: Reference to real world? Speaker's intention?
- Linguistic meaning captures the structure of expressions:
  - Morphology, syntax, ...
  - Units of each layer composed into higher units (FGD, Sgall et al. (1986))

Illustration from `http://www.cs.toronto.edu/~tingwuwang/semantic_segmentation.pdf`.

Aspects of Meaning (2/2)

Aspects of sentence meaning as listed by Bojar et al. (2019).

	Symbolic	Continuous
Aspect of Meaning	Theories	Representations
Abstraction	✓	×
Compositionality	✓	$\sim$
Learnability	?	✓
Relatability (similarity, operations)	$\sim$	$\sim$
Vagueness of Meaning	×	✓
Ambiguity of Expressions	✓	×
Statefulness	$\sim$	✓

Compositionality of Meaning

Manning (2015):

> *Understanding novel and complex sentences crucially depends on being able to construct their meaning compositionally from smaller parts—words and multiword expressions—of which they are constituted.*

Compositionality of Vector Representations

Karlgren and Kanerva (2019) show “Holographic Reduced Reprs.”:

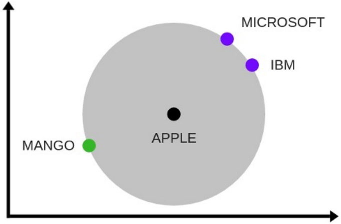
- Addition: Preserves similarity, useful to represent bag-of-...
- Hadamard product (elem-wise multiplication),
  - Invertible; product dissimilar to its operands: $A * B \not\sim A$.
  - Bipolar vectors ($\{-1, +1\}^n$) are inverse of themselves.
  - Can represent variable assignment $\{x = a, y = b, z \ = c\}$ using bipolar vectors X, Y, and Z added into a vector $(X * A) + (Y * B) + (Z * C)$. To recover the value of x, multiply by X: $X * (X * A) + X * (Y * B) + X * (Z * C)) = A + \text{noise} + \text{noise} \sim A$
- Vector elements permutation,
  - Also invertible; dissimilar; enormous number of permutations.
  - Useful to represent structures, e.g. lists: Π_1 for CAR Π_2 for CDR: (a, b) represented with $\Pi_1(a) + \Pi_2(b)$

(In highly-dimensional spaces, most vectors are dissimilar; cosine or Pearson correlation of 0.25 indicate close similarity.)

Modelling Ambiguity?

Sentence-level embeddings always produced by an encoder.

- Encoder = A deterministic mapping from expression to meaning.
- Unclear how ambiguous expressions are and should be represented.
- Same problem with word vectors already:



Ideally, an expression would correspond to
a distribution over semantic space, not a single point.

Meaning Statefulness

Stateful Meaning Representation:

- Could be modelled by decoder state:

  ≈ "State of mind after reading the input and producing a partial output."
- Better reflected in models with attention.
- Btw needed to interpret humour (Gluscevskij, 2017).

Stateless Meaning Representation:

- Encoder state.
- Points correspond to expressions.
  - Ambiguity representation unclear.

Is Sentence Meaning Continuous?

We know that one Arabic sentence can have dozens of thousands of English translation (Dreyer and Marcu, 2012):

Premiere of Iraq Nuri al-Maliki was given an excuse by President Bush, who expressed his confidence in him, and he stated that the circumstances are complicated.
President Bush said that he trusts in Nouri Maliki, head of government of Iraq, and he stated that he finds an excuse for him ”because the situation is tricky”.
Head of cabinet of Iraq Nuri al-Maliki was given an excuse by President Bush, who expressed his trust in him, and he indicated that the circumstances are difficult.
Iraq’s head of cabinet Nuri al-Maliki was given a reason by President Bush, who expressed his trust in him, and he indicated that the case is tricky.
President Bush said that he has faith in Iraqi head of cabinet Nouri al-Maliki, and he stated that he finds an excuse for him ”for the case is complicated”.

Is Sentence Meaning Continuous?

Similarly: 70k Czech translations of 1 English sentence (Bojar et al., 2013)

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi podobná.
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politického veterána.
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického veterána.

Q: Are all these paraphrases close in sent embedding spaces?

Q: How entagled are manifolds of different sents?

... work in progress with Petra Barančíková

Examining Continuous Space

Proposed strategy:

1. Propose directions of exploration.
2. Generate seed pairs of sentences for each of the directions.
3. Collect specimens along the proposed directions:
    - interpolation, a “sentence in between”,
    - extrapolation, “a sentence further in the hinted direction”.
    - Allow people to say “impossible”.
4. Validate the relations.
5. Create the partially ordered set.
6. Search for a manifold covering the ordered set.

Some first ideas explored with Chris Callison-Burch.

First dataset for Czech released (Barančíková and Bojar, 2020).

Directions of Exploration (1/2)

- Politeness
- Tense
- Verity: How much the speaker believes the message.
- Modality: Willingness/Ability of the speaker to do it.
- “Counting” / Generic Numerals, Scalar adjectives
  - I saw a handful of people there. / a big crowd / a massive crowd.
  - freezing / cold / chilly
- “Negation”, but not only reversing the main predicate
- Complexity / simplicity, Length.

Thanks to Šárka Zikánová for some of the ideas.

Directions of Exploration (2/2)

- Specificity / Generality, Vagueness.
  - Geese fly / Geese migrate / Geese migrate south / The Canadian geese flew over the pond at friendly Farms in their southward migration.
  - Hammer the hook into the wall. / Put the hook on the wall. / Do the thingy in there.
- Contextual boundness.
  - Give it to him. / Give the parcel to the man at the counter. / Give your parcel to the operator at the post office.
- High/low style/English/class.
  - Hey y'all it's a nice day ain't it?
  - Greetings! Lovely weather we are having.

First Results of Getting Pairs

Can you please give me a minute?	Could you leave me alone?
Close the door.	Close the damn door man
Can you help me find something?	I need you to help me get something.
May I talk to Mary?	Is Mary here?
I'm sorry-I don't believe we have met.	Who the hell are you?
Can you move so I can see the screen?	You aren't made of glass, you know.
Will you kindly exit?	I do not want you here!
Would you please get the mail?	Get the mail!
Can I help you?	What do you want?
Can you please help me with this?	Get over here and help me!
Can you make me breakfast?	Why are you not making me breakfast right now?
I tried to call were you busy?	You never answer your phone.

First Results of Midpointing

Can you move so I can see the screen?

Blocking the view, friend.

Move your blocking the screen

Could you move a little bit, you're blocking the screen.

Can you please move?

I can't see, can you move a little?

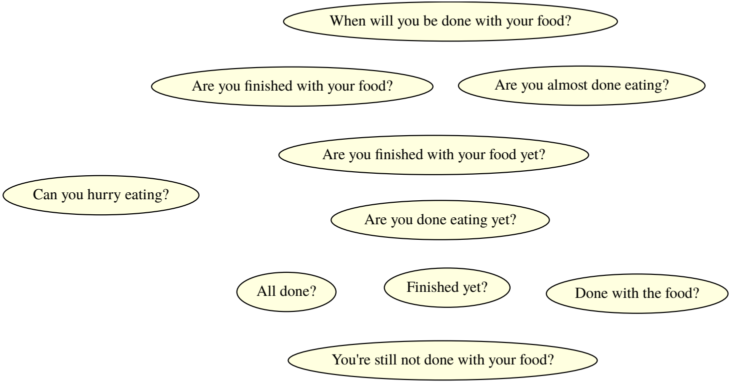
Hey can you move.

Please move.

Can you move a bit?

You aren't made of glass, you know.

After the Midpointing...



When will you be done with your food?

Are you finished with your food?

Are you almost done eating?

Are you finished with your food yet?

Can you hurry eating?

Are you done eating yet?

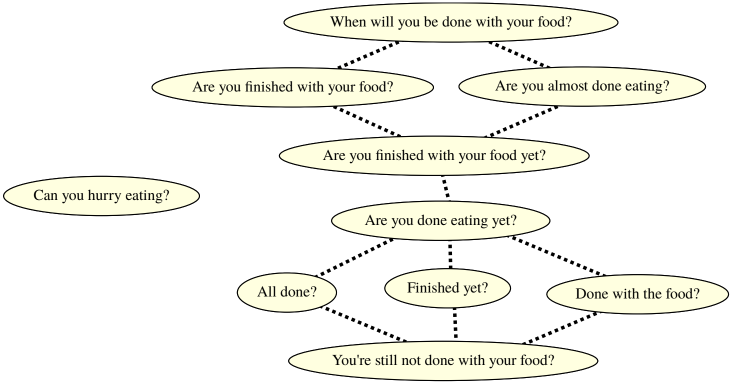
All done?

Finished yet?

Done with the food?

You're still not done with your food?

Ask Crowd to Partially Sort Them


When will you be done with your food?

Are you finished with your food?

Are you almost done eating?

Are you finished with your food yet?

Can you hurry eating?

Are you done eating yet?

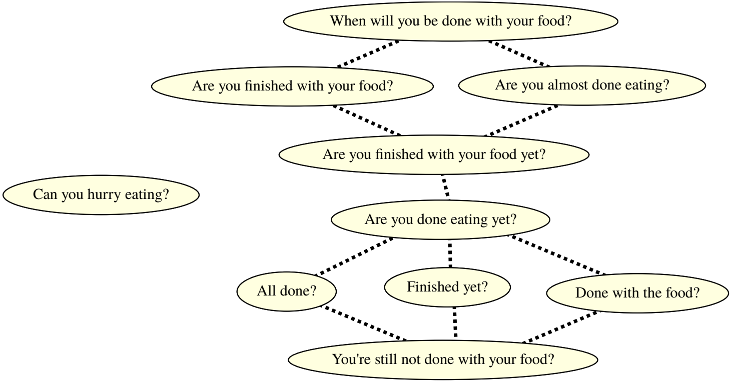
All done?

Finished yet?

Done with the food?

You're still not done with your food?

Find Methods for Manifold Learning



When will you be done with your food?

Are you finished with your food?

Are you almost done eating?

Are you finished with your food yet?

Can you hurry eating?

Are you done eating yet?

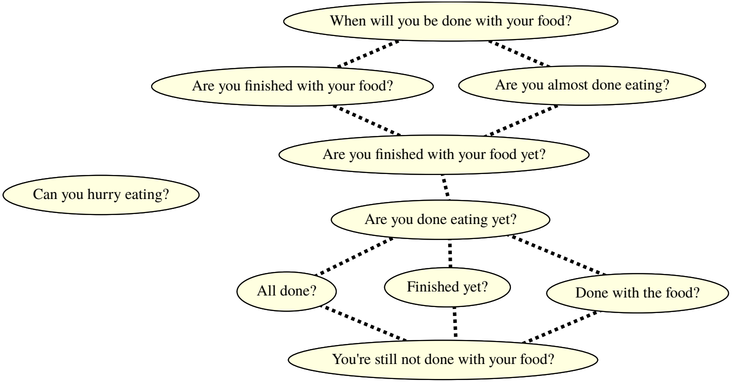
All done?

Finished yet?

Done with the food?

You're still not done with your food?

Match Posets with Learned Manifolds



When will you be done with your food?

Are you finished with your food?

Are you almost done eating?

Are you finished with your food yet?

Can you hurry eating?

Are you done eating yet?

All done?

Finished yet?

Done with the food?

You're still not done with your food?

semi-supervised.

Evaluating Sentence Representations

Evaluating Sentence Representations

Conneau and Kiela (2018) introduce SentEval:

- Given a sentence representation function, assess the fitness of the representation in multiple tasks.
  `https://github.com/facebookresearch/SentEval/`

Conneau et al. (2018) and others then compare several reprs incl.:

- SkipThough (Kiros et al., 2015):
  - Predict sentence given the surrounding sentences.
- InferSent (Conneau et al., 2017):
  - Train sentence representations on predicting entailment.

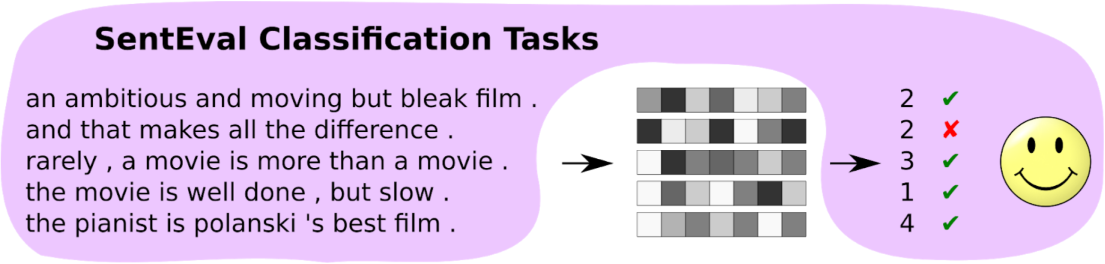
Extremely active research area, see **BlackboxNLP workshops**.

How “Semantic” are Seq2Seq Reprs?

Cífka and Bojar (2018):

- Trained several variations of Cho et al. (2014).
- Extracted sentence representations.
- Related BLEU and “semantics” of the representation:
    - Evaluation through classification.
    - Evaluation through similarity.
    - Evaluation using paraphrases.

Evaluation through classification



SentEval Classification Tasks

an ambitious and moving but bleak film . → 2 ✓

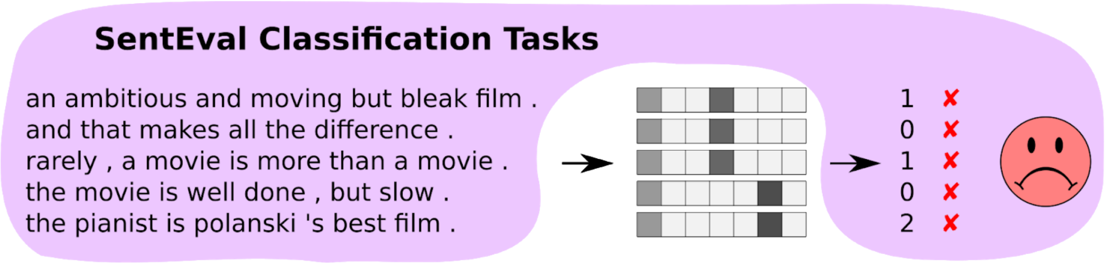
and that makes all the difference . → 2 ✗

rarely , a movie is more than a movie . → 3 ✓

the movie is well done , but slow . → 1 ✓

the pianist is polanski 's best film . → 4 ✓

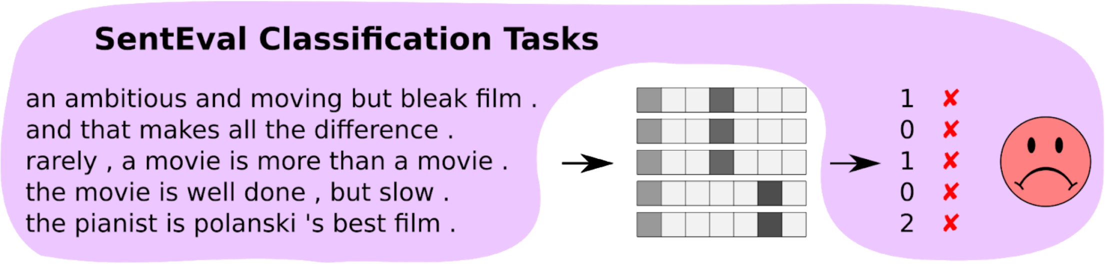
Evaluation through classification



SentEval Classification Tasks

an ambitious and moving but bleak film .
and that makes all the difference .
rarely , a movie is more than a movie .
the movie is well done , but slow .
the pianist is polanski 's best film .

Evaluation through classification

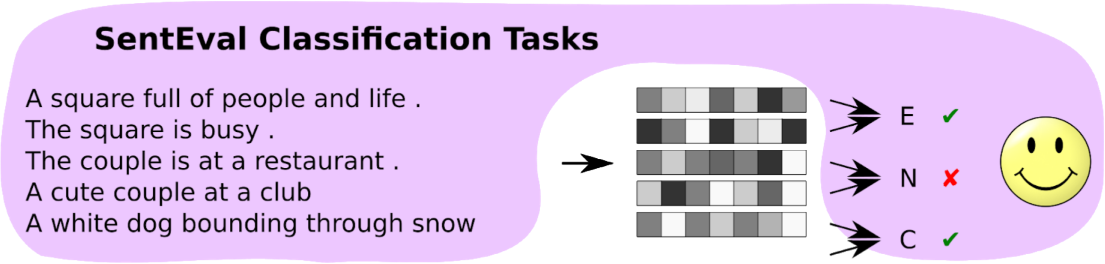


SentEval Classification Tasks

an ambitious and moving but bleak film .
and that makes all the difference .
rarely , a movie is more than a movie .
the movie is well done , but slow .
the pianist is polanski 's best film .

- Solo: movies sentiment, product review polarity, question type...

Evaluation through classification



SentEval Classification Tasks

A square full of people and life .
The square is busy .
The couple is at a restaurant .
A cute couple at a club
A white dog bounding through snow

- Solo: movies sentiment, product review polarity, question type...
- Paired: natural language inference, semantic equivalence

Evaluation through similarity

- 7 similarity tasks: pairs of sentences + human judgement

I think it probably depends on your money.	It depends on your country.	0
Yes, you should mention your experience.	Yes, you should make a resume	2
Hope this is what you are looking for.	Is this the kind of thing you're looking for?	4

  - with training set, sent. similarity predicted by regression,
  - without training set, cosine similarity used as sent. sim.,
  - ultimately, the predicted sent. similarity is correlated with the golden truth.
- In sum, we report them as “AvgSim”.

Evaluation using paraphrases: the data

- HyTER: ~200 sentences, 500 translations each
- COCO: 5k images, 5 captions each

低胸露背的黄金泳衣重五百公克,售价一千万日币。

the deep cut and halter golden swimwear weighs half kilogram selling at ten million JPY.

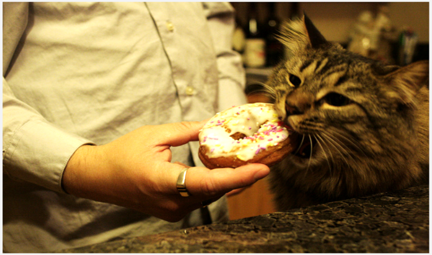
¥10,000,000 is the retail value for the low-cut gold bathing suit with a low back, and the weight is 5 hundred g.

at the weight of five hundred grams, the low cut, halter swimsuit made up of gold will sell at ten million Japanese Yen (JPY).

(Dreyer and Marcu, 2014)

Evaluation using paraphrases: the data

- HyTER: ~200 sentences, 500 translations each
- COCO: 5k images, 5 captions each



http://cocodataset.org/#explore?id=78026

(Lin et al., 2014)

a person is feeding a donut to the cat.

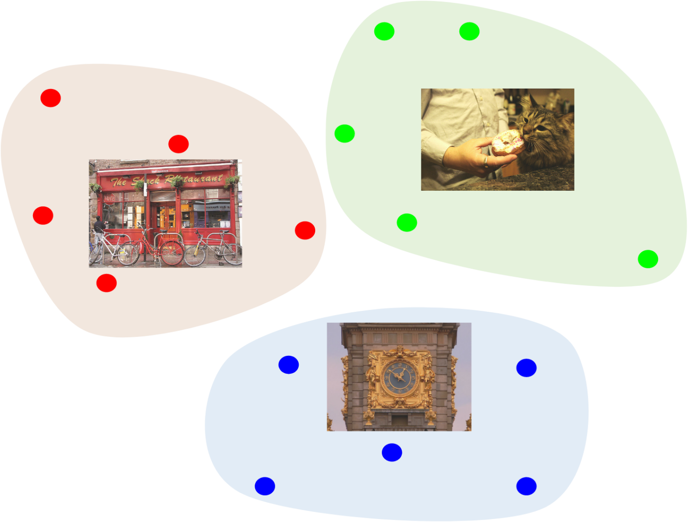
a cat being fed a donut by someone in a grey shirt.

a cat nibbles on a sprinkled donut that is being fed by the owner.

a grey cat biting into a frosted donuts

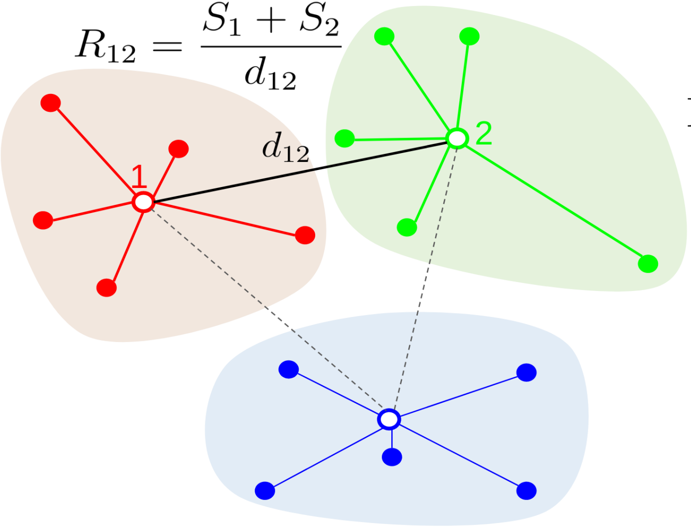
a cat is eating a donut from a person's hand.

Evaluation using paraphrases: the metrics



Cluster separation: Davies-Bouldin index

$$R_{12} = \frac{S_1 + S_2}{d_{12}}$$

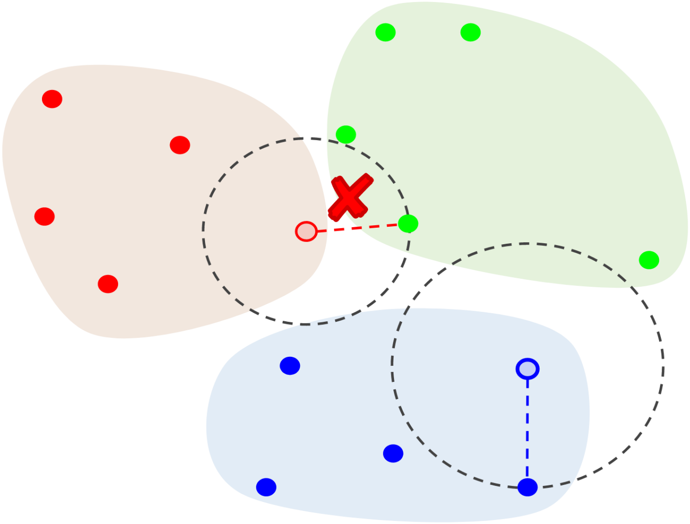


$$\mathrm{DB} = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i} R_{ij}$$

For each cluster, find the **least well-separated** one

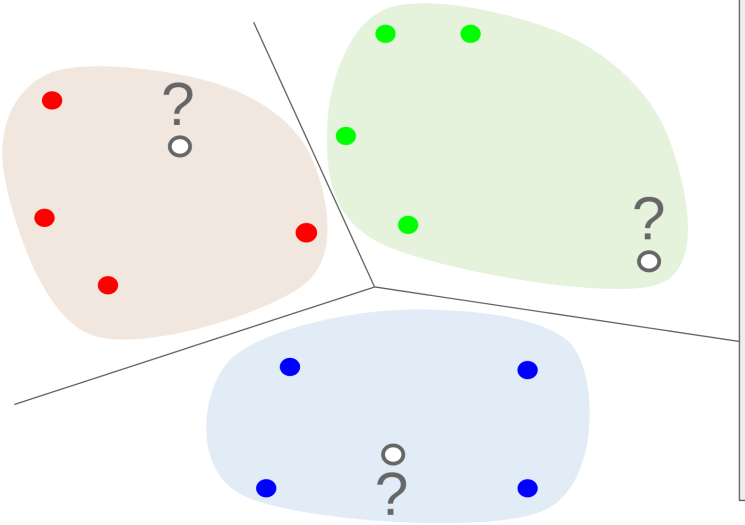
(Davies and Bouldin, 1979)

Paraphrase retrieval task (NN)



Retrieve the **nearest neighbor** and check whether it lies in the same cluster

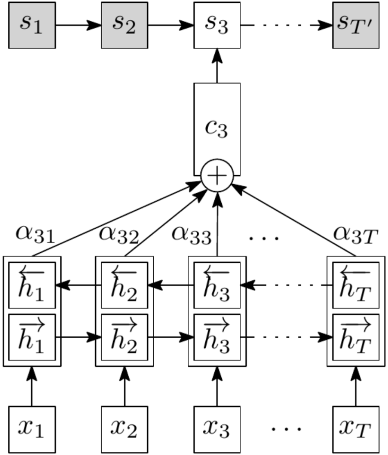
Classification task



1. Remove some points from the clusters.
2. Train an LDA classifier with the remaining points.
3. Classify the removed points back.

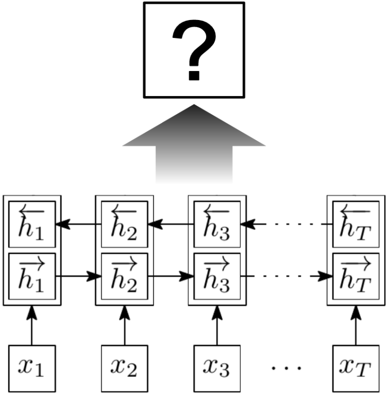
Sequence-to-sequence with attention

- Bahdanau et al. (2014)
- α_{ij}: weight of the j^{th} encoder state for the i^{th} decoder state
- no sentence embedding



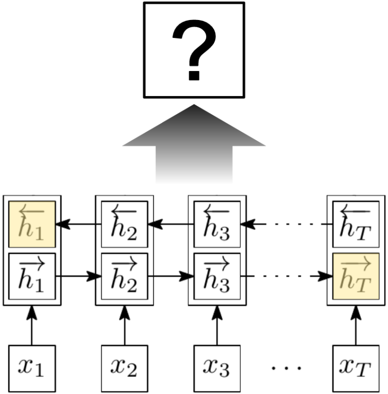
Ways of getting sentence embeddings

- final state
- max/average pooling
- inner attention



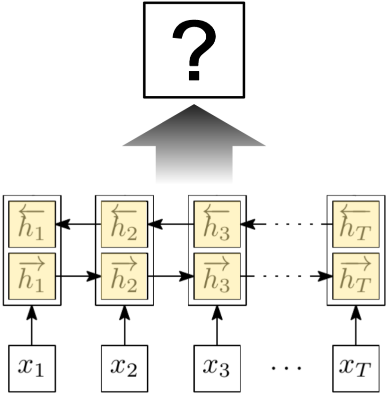
Ways of getting sentence embeddings

- final state
- max/average pooling
- inner attention



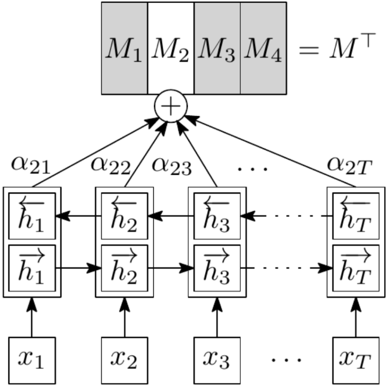
Ways of getting sentence embeddings

- final state
- max/average pooling
- inner attention

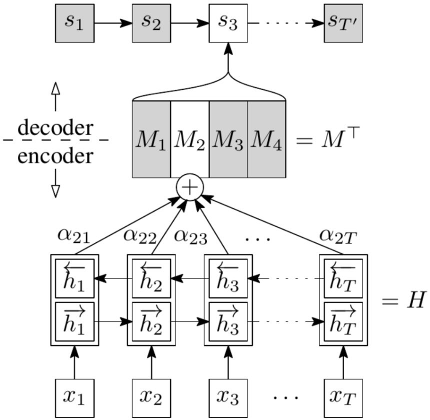


Multi-head inner attention

- Liu et al. (2016), Li et al. (2016), Lin et al. (2017)
- α_{ij}: weight of the jth encoder state for the ith column of $M^\top$
- concatenate columns of $M^\top$ → sentence embedding
- linear projection of columns to control embedding size



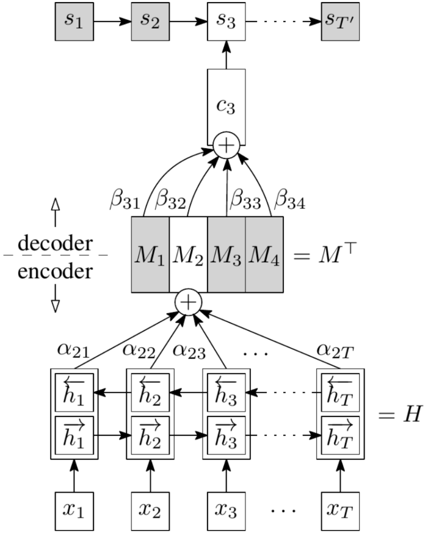
Proposed NMT architectures



s_1
s_2
s_3
$s_{T'}$
decoder
encoder
M_1 M_2 M_3 M_4 $= M^\top$
α_{21}
α_{22}
α_{23}
$\dots$
α_{2T}
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$ $\overleftarrow{h_T}$
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$ $\overrightarrow{h_T}$
$= H$
x_1 x_2 x_3 $\dots$ x_T

ATTN-CTX

decoder operates on entire embedding

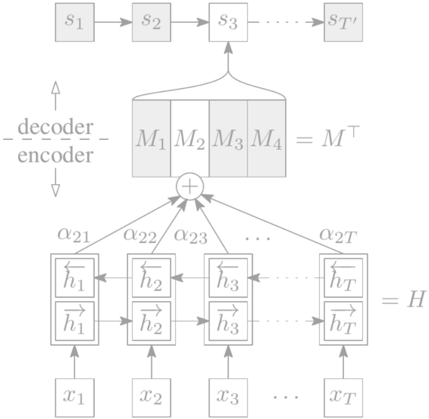


s_1
s_2
s_3
$s_{T'}$
c_3
β_{31}
β_{32}
β_{33}
β_{34}
decoder
encoder
M_1 M_2 M_3 M_4 $= M^\top$
α_{21}
α_{22}
α_{23}
$\dots$
α_{2T}
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$ $\overleftarrow{h_T}$
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$ $\overrightarrow{h_T}$
$= H$
x_1 x_2 x_3 $\dots$ x_T

ATTN-ATTN (*compound att.)*

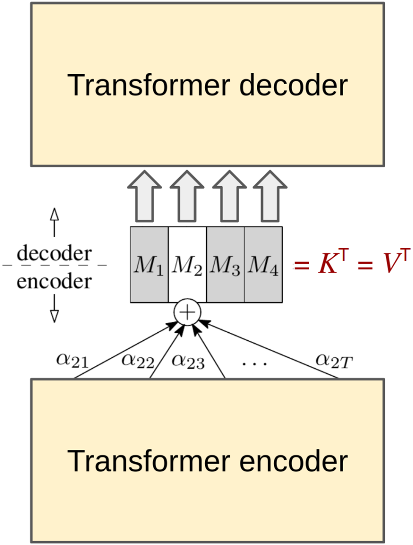
decoder „selects“ components of embedding

Proposed NMT architectures


s_1
s_2
s_3
$s_{T'}$
decoder
encoder
M_1 M_2 M_3 M_4 $= M^\top$
α_{21} α_{22} α_{23} $\cdots$ α_{2T}
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_T$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_T$
$= H$
x_1 x_2 x_3 $\cdots$ x_T

ATTN-CTX

decoder operates on entire embedding


Transformer decoder
decoder
encoder
M_1 M_2 M_3 M_4 $= K^\top = V^\top$
α_{21} α_{22} α_{23} $\cdots$ α_{2T}
Transformer encoder

TRF-ATTN-ATTN

Transformer (Vaswani et al., 2017) with inner attention

Evaluated NMT models

- model architectures:
  - **FINAL**, **FINAL-CTX**: no attention
  - **AVGPOOL**, **MAXPOOL**: pooling instead of attention
  - **ATTN-CTX**: inner attention, constant context vector
  - **ATTN-ATTN**: inner attention, decoder attention
  - **TRF-ATTN-ATTN**: Transformer with inner attention
- translation from English (to Czech or German), evaluating embeddings of English (source) sentences
  - en→cs: CzEng 1.7 (Bojar et al., 2016)
  - en→de: Multi30K (Elliott et al., 2016; Helcl and Libovický, 2017)

Sample Results – translation quality en → cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	18.4	42.5	88.6
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – translation quality en → cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	18.4	42.5	88.6
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

BLEU is consistent with human evaluation.

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – translation quality en → cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	18.4	42.5	88.6
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

Attention in the encoder helps translation quality.

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – translation quality en → cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	18.4	42.5	88.6
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

More attention heads → better translation quality.

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – representation eval. en → cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX (“Cho“)	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

Sample Results – representation eval. en→cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX (“Cho“)	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

Baselines are hard to beat.

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

Sample Results – representation eval. en → cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX ("Cho")	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

Attention harms the performance.

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

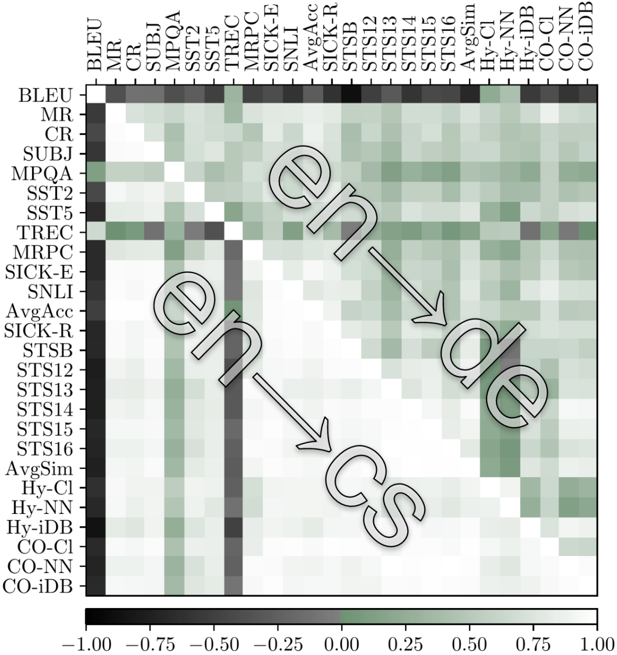
Sample Results – representation eval. en → cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX (“Cho“)	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

More heads → worse results.

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

Full Results – correlations



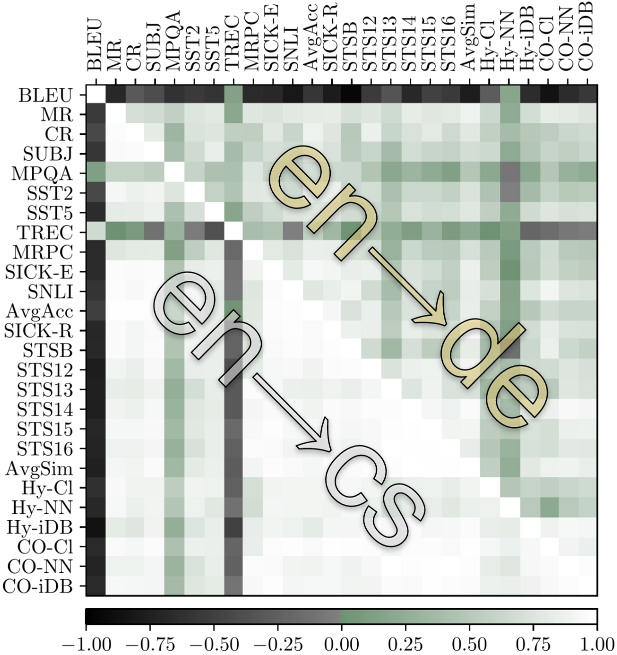
BLEU vs. other metrics:
−0.57 ± 0.31 (en → **cs**)
−0.36 ± 0.29 (en → **de**)

Pairwise average (except BLEU):
0.78 ± 0.32 (en → **cs**)
0.57 ± 0.23 (en → **de**)

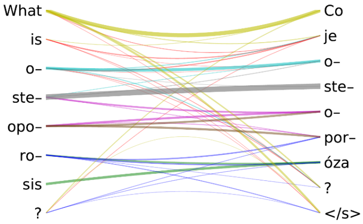
Full Results – correlations excluding Transformer

BLEU vs. other metrics:
−0.57 ± 0.31 (en → **cs**)
−0.54 ± 0.27 (en → **de**)

Pairwise average (except BLEU):
0.78 ± 0.32 (en → **cs**)
0.62 ± 0.23 (en → **de**)

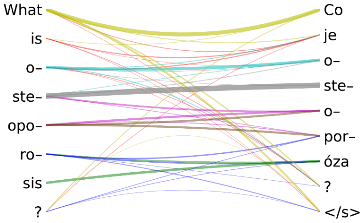


Compound attention interpretation

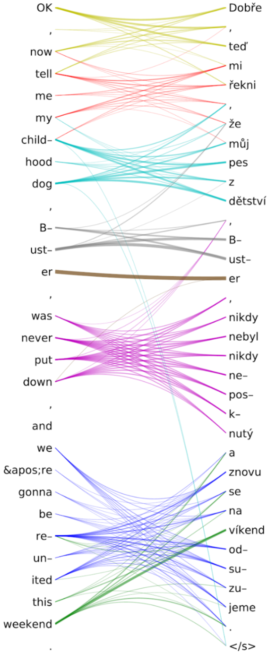


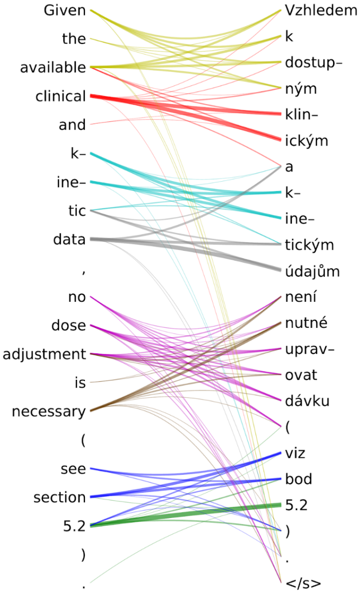
ATTN-ATTN en-cs model with 8 heads

Compound attention interpretation

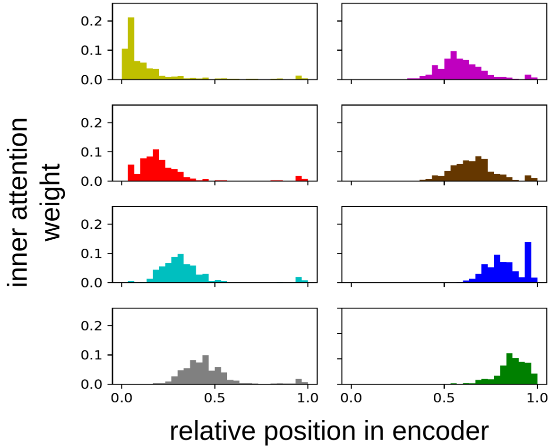


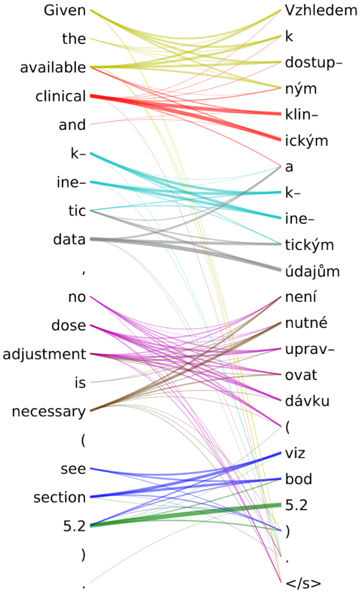
ATTN-ATTN en-cs model with 8 heads

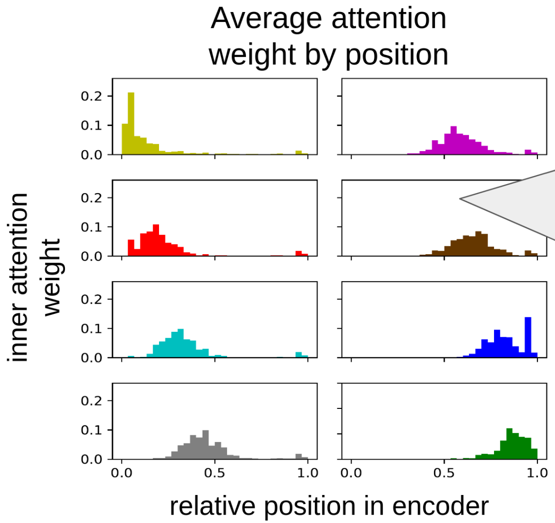


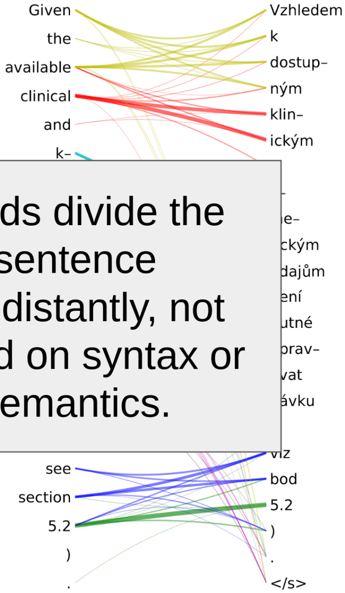


Average attention weight by position









Heads divide the sentence equidistantly, not based on syntax or semantics.

Summary

- NMT systems can surpass humans within the given domain.
- We discussed learned representations.
  - Illustrated word and sentence embeddings.
- We discussed aspects of meaning.
- Some level of "understanding" can be found in the representations.
  - Follow the BlackBoxNLP workshops:
  - POS, Syntax, Word Derivations, Compositionality...
  - Still very far from human understanding.
- Big caveats need to be taken when interpreting results.
  - The "utility" of syntax in NMT discussed last week.
  - The exact composition of the task and the test set.

References

Mostafa Abdou, Vladan Gloncak, and Ondřej Bojar. 2017. Variable mini-batch sizing and pre-trained embeddings. In Proceedings of the Second Conference on Machine Translation, pages 680–686, Copenhagen, Denmark, September. Association for Computational Linguistics.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. CoRR, abs/1801.09536.

Petra Barančíková and Ondřej Bojar. 2020. COSTRA 1.0: A Dataset of Complex Sentence Transformations. In Proceedings of the LREC 2020. ELRA.

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2015. Do you see what I mean? visual resolution of linguistic ambiguities. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1477–1487, Lisbon, Portugal, September. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. 2013. Scratching the Surface of Possible Translations. In Proc. of TSD 2013, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

Ondřej Bojar, Raffaella Bernardi, and Bonnie Webber. 2019. Representation of sentence meaning (a jnle special issue). Natural Language Engineering, 25(4).

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and Meaning Representation in Opposition? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1362–1371. Association for Computational Linguistics.