Word and Sentence Representations

Ondřej Bojar

May 9, 2019

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

Course Outline

— Part I: MT as a Practical Application.

1. Metrics of MT Quality.
2. Approaches to MT. SMT, PBMT, NMT, NP-hardness.
3. NMT (Seq2seq, Attention. Transformer). Neural Monkey.
4. Parallel texts. Sentence and word alignment. hunalign, GIZA++.
5. PBMT: Phrase Extraction, Decoding, MERT. Moses.
6. Morphology in MT. Factors or segmenting, data or linguistics.
7. Syntax in SMT (constituency, dependency, deep).
8. Syntax in NMT (soft constraints/multitask, network structure).
9. More on Search in SMT (weighted deduction, future cost, cube pruning).

— Part II: MT as a Step Towards Understanding.

10. **Towards Understanding: Word and Sentence Representations.**

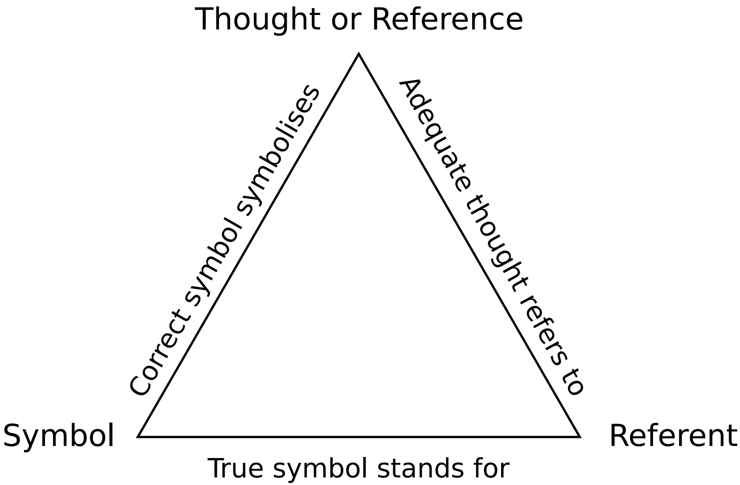
— Part III: Advanced Topics.

11. Advanced: Multi-Lingual MT. Chef's Tricks.
12. **Project presentations: <u>May 23, 2019</u>**

Outline

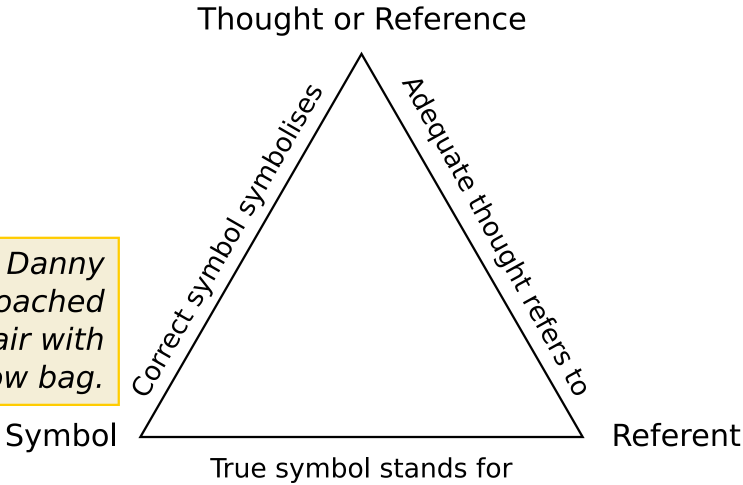
- Semiotic Triangle: Towards Understanding.
- Continuous Word Representations.
- Continuous Phrase Representations.
- Continuous Sentence Representations.
- Relating Human and NN Meaning Representations.

Semiotic Triangle



Semiotic Triangle by Ogden and Richards (1923).

Semiotic Triangle



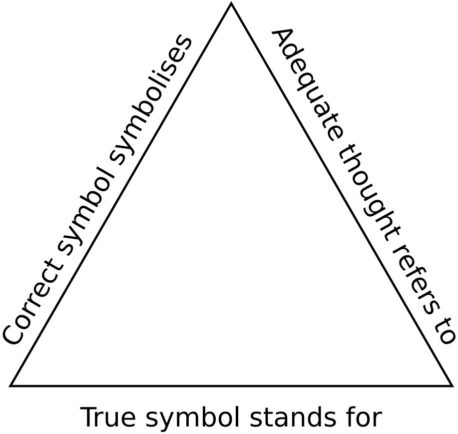
Danny approached the chair with a yellow bag.

Ambiguous sentence…

Semiotic Triangle

Thought or Reference

Danny approached the chair with a yellow bag.

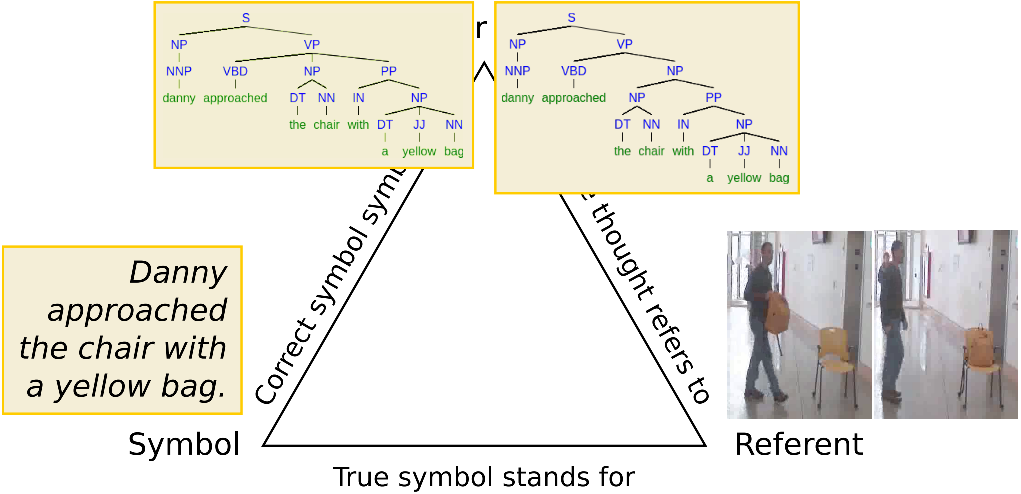


Symbol

Referent

Ambiguous sentence correspond to two situations.

Semiotic Triangle



Syntactic “meaning” distinguishes this already.

Semiotic Triangle

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

r

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.

Correct symbol sym

te thought refers to

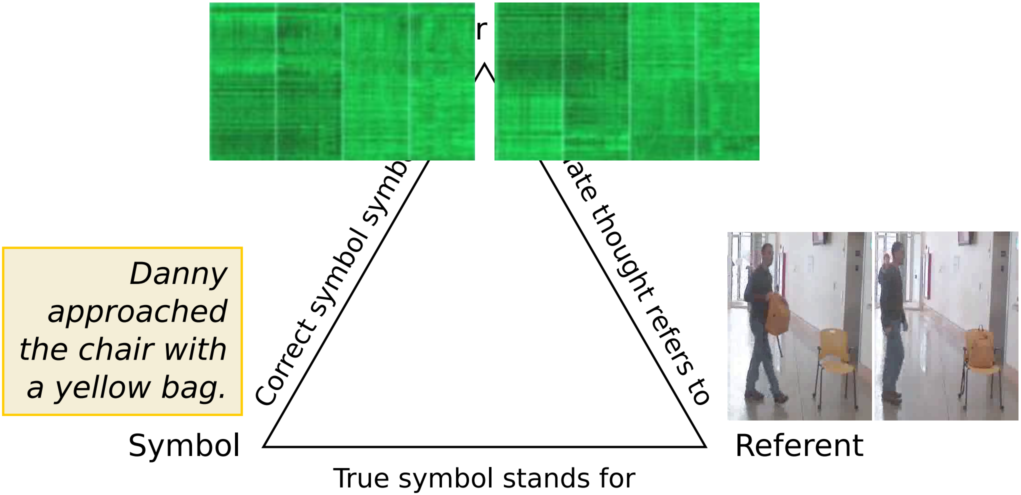
Symbol

Referent

True symbol stands for

Lambda calculus makes the difference clear.

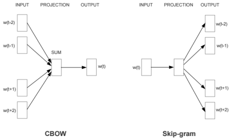
Semiotic Triangle

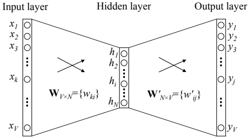


NN activations will somehow differ, too.

Word Embeddings

- Map each word to a dense vector.
- In practice 300–2000 dimensions are used, not 1–2M.
  - 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., 2013a):
  - CBOW: Predict the word from its four neighbours.
  - Skip-gram: Predict likely neighbours given the word.



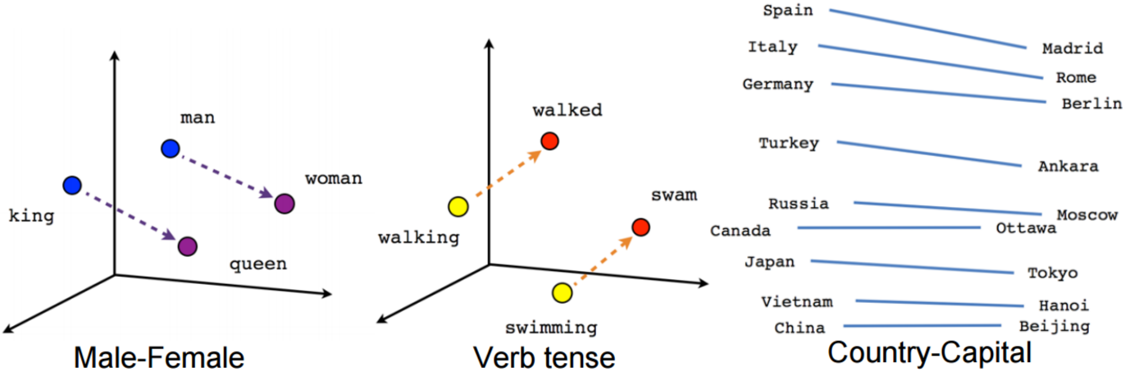


Right: CBOW with just a single-word context (`http://www-personal.umich.edu/~ronxin/pdf/w2vexp.pdf`)

Continuous Space of Words

Word2vec embeddings show interesting properties:

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



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

Testset by Mikolov et al. (2013a)

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

Problems of the Testset

- Only 3 types of "semantic" questions:
  - city-state/country, country-currency, feminine-masculine.
  - Vylomova et al. (2016) mentions many other sem. relationships:
    - e.g. walk-run, dog-puppy, bark-dog, cook-eat and others.
- "Syntactic" questions broader, but:
  - Constructed from just a few dozens of word pairs, comparing pairs with each other.
  - Overall only 313 distinct pairs throughout the whole set of 10675 questions.
  - Moreover, 268 of the 313 pairs are regularly formed:
    - e.g. by adding the suffix $+ly$ for adj$\rightarrow$adv.

- A better test set for Czech morphosyntax released by Kocmi and Bojar (2016)

Evaluating Words by Similarity?

The whole idea of evaluating word vector by similarity is risky.

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

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

Abdou et al. (2017)

- English-to-Czech MT, English embeddings optionally pre-trained.
  - (No improvement for NMT; Kocmi and Bojar (2017) saw a quicker start of training.)
- Evaluated embeddings from monolingual and parallel training:

Embeddings from	Monolingual Training		NMT Training	
	CBOW (no BPE)	**CBOW (BPE)**	**Baseline**	**Pretrained**
Vocabulary	Full	Common subset (265 words)		
WordSim-353 (ρ)	0.320	0.610	0.571	**0.621**	0.527
MEN (ρ)	0.300	0.610	**0.621**	0.583	0.591
SimLex-999 (ρ)	0.064	0.173	0.171	**0.519**	0.267

Pairwise cosine similarity between embeddings and standard human judgments for the common subset of the vocabularies. Best result in each row in bold.

- "Baseline" = learned by NMT only, "Pretrained" = init. by CBOW (BPE).
- Parallel ⇒ best for Similarity, Monolingual ⇒ Relatedness (MEN).

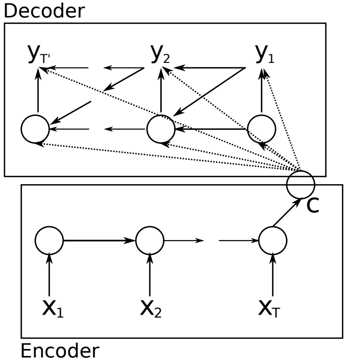
Continuous Phrase Representations

Mikolov et al. (2013b) extend SkipGram to non-compositional phrases:

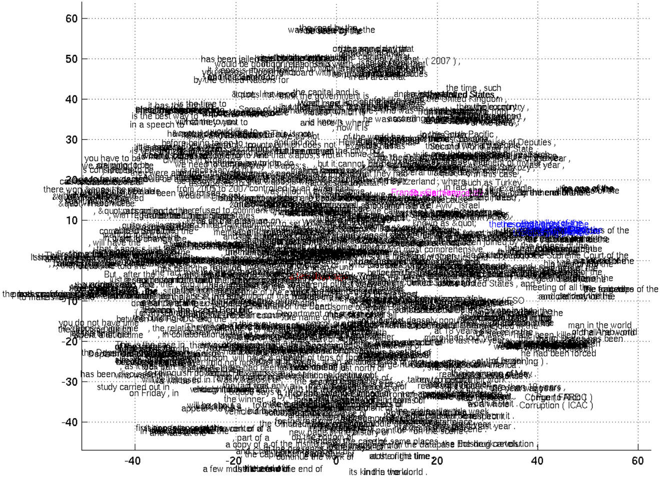
- Phrases identified in a pre-processing step used as atomic tokens.
- Vector compositionality: $v(\text{Czech}) + v(\text{currency}) \approx v(\text{koruna})$

Cho et al. (2014) propose:

- encoder-decoder architecture and
- GRU unit (name given later by Chung et al. (2014))
- to score variable-length phrase pairs in PBMT.

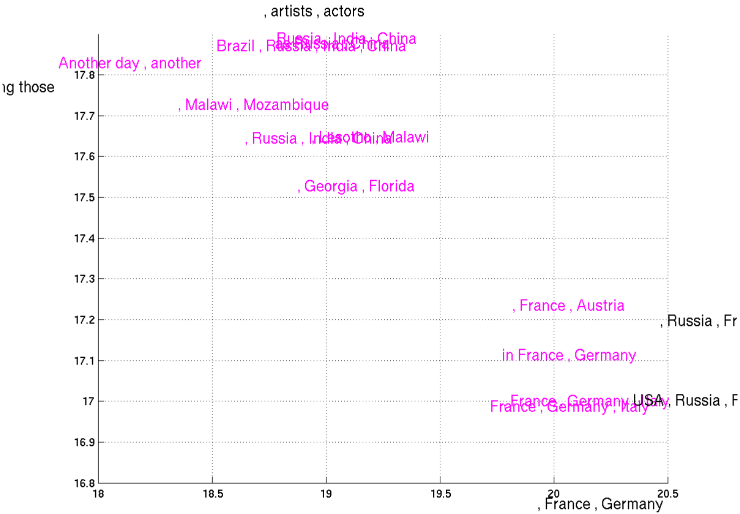


⇒ Embeddings of Phrases

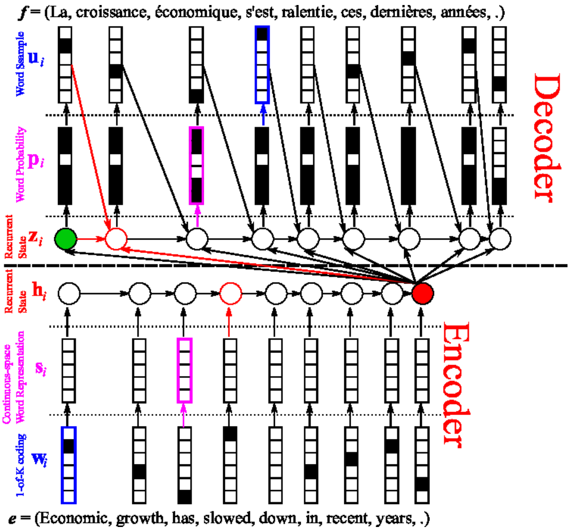


… Reveal Syntactic Similarity

… and Semantic Similarity

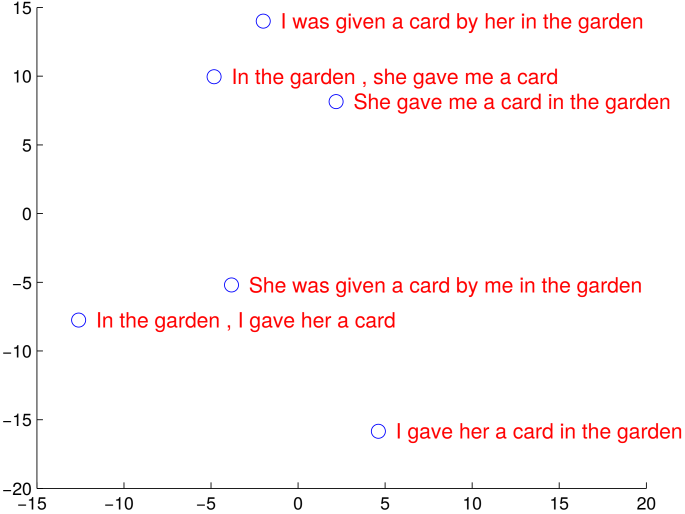


Encoder-Decoder Architecture



https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/

Continuous Space of Sentences



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

What can you cram into a single vector?

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

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.

Cífka and Bojar (2018)

- Trained several variations of Cho et al. (2014).
  - Multiple heads, to emulate attention while having a fixed-size sentence representation.
- Evaluated the models in terms of NMT (BLEU) and meaning representation evaluations:
  - SentEval,
  - Similarity of vectors corresponding to paraphrases.
- All similarity measures correlate with each other.
- BLEU is negatively correlated with them.

  The better the translation in terms of BLEU, the worse the sentence representation serves in tasks like sentiment analysis etc.

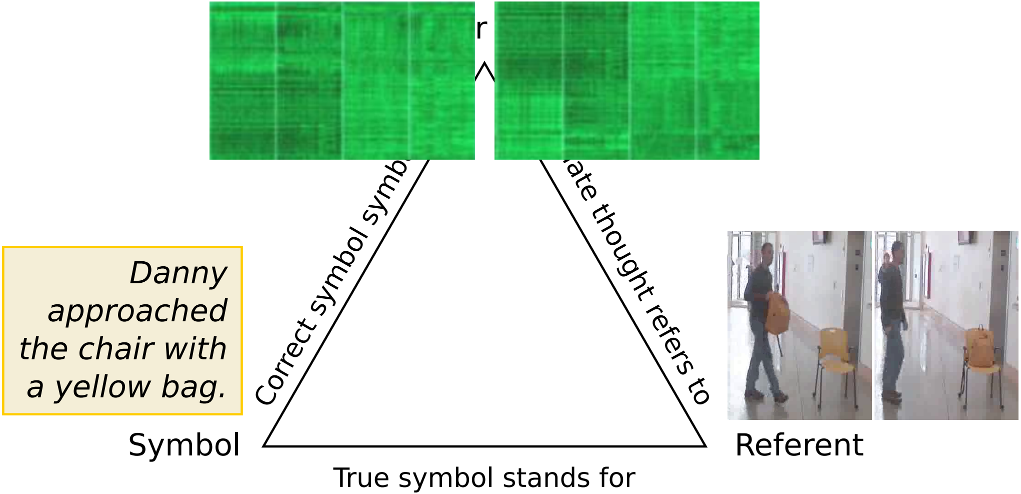
When could vectors encode?

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

- Addition: Preserves similarity, useful to represent bag-of-...
- Hadamard product (elem-wise multiplication),
  - Invertible; product dissimilar to its operands: $A * B \nsim A$.
  - 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 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 $\Pi_1(a) + \Pi_2(b)$

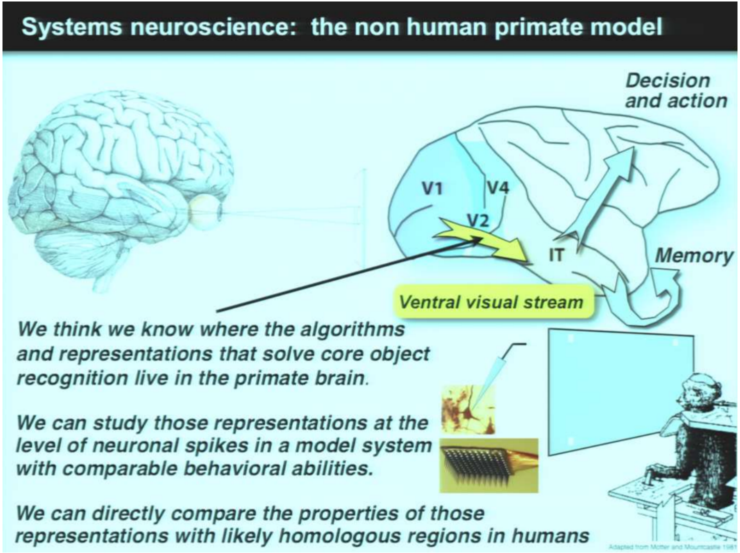
(In highly-dimensional spaces, most vectors are dissimilar; cosine or Pearson correlation of 0.25 indicate close similarity.)

Relating Human and NN Representations

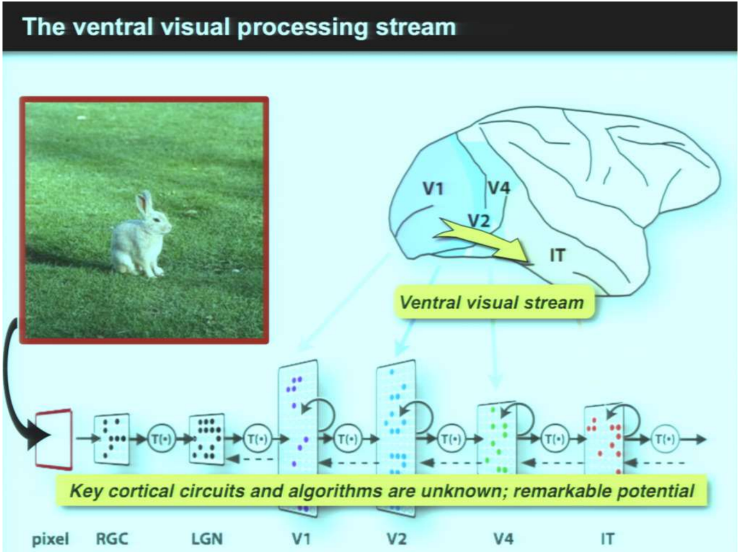


… let's take inspiration in ape and human vision first.

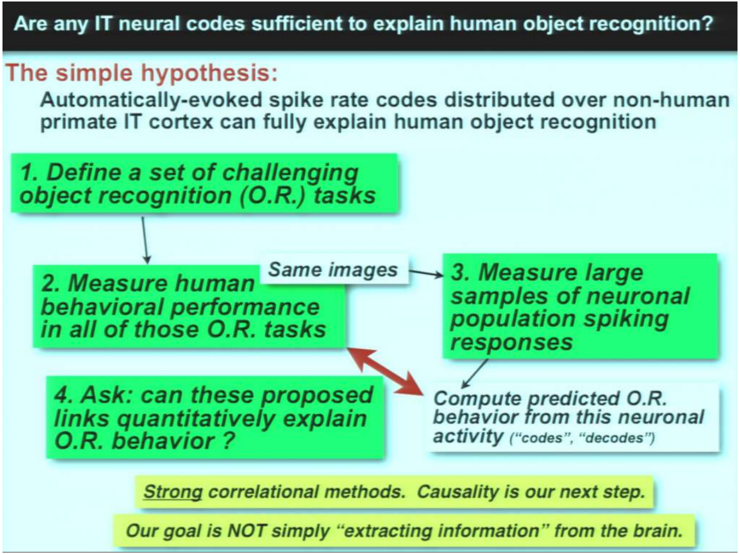
DiCarlo 2013 Tutorial on Vision



DiCarlo 2013 Tutorial on Vision



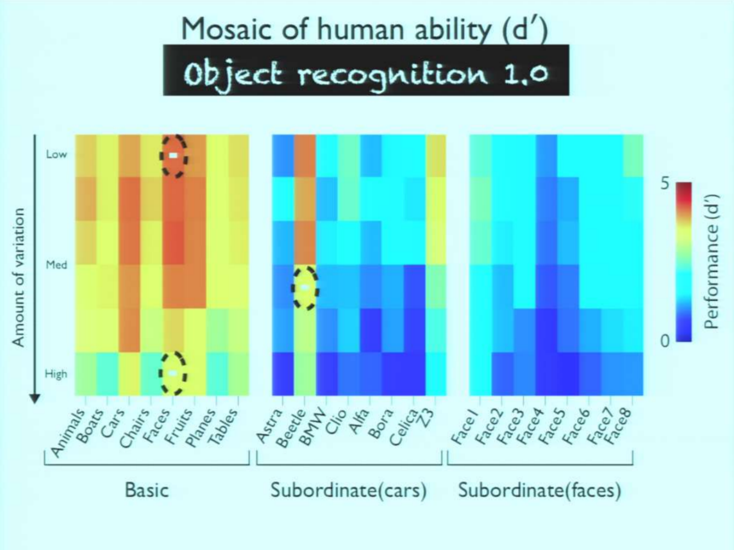
DiCarlo 2013 Tutorial on Vision



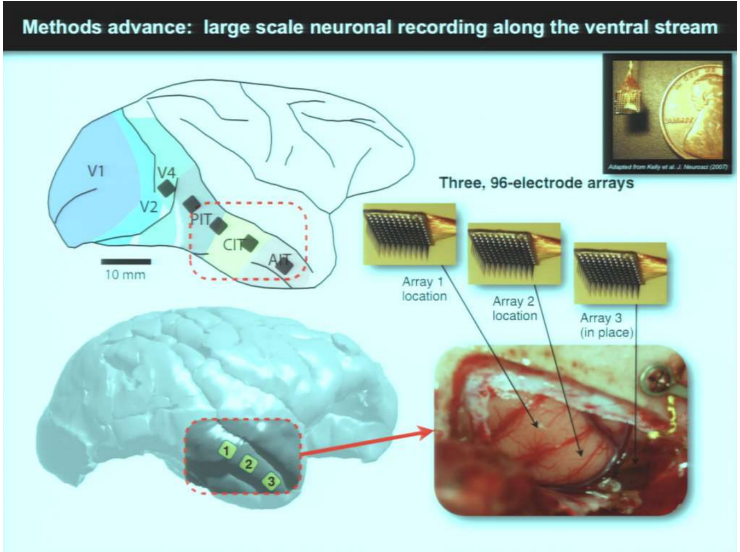
DiCarlo 2013 Tutorial on Vision

- 64 objects, can generate as many images as we like
- full parametric control
- “natural” statistics
- uncorrelated, new background every image
- not fully “natural” by design -- challenging for computer vision, doable by humans

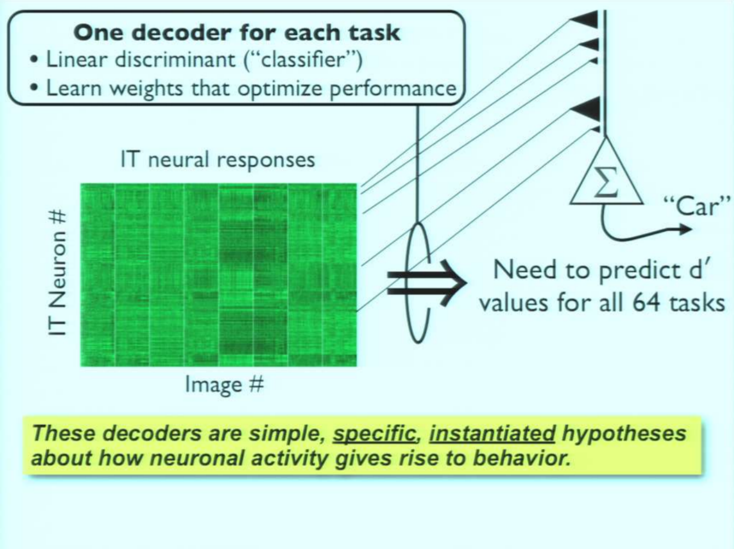
DiCarlo 2013 Tutorial on Vision



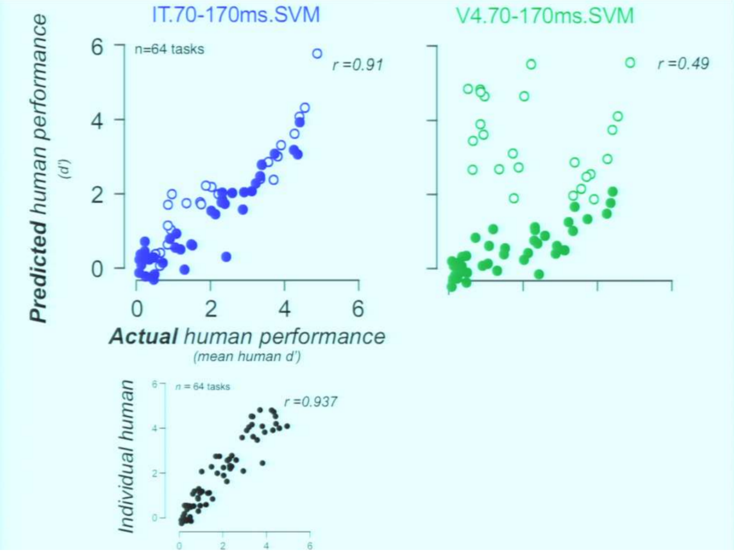
DiCarlo 2013 Tutorial on Vision



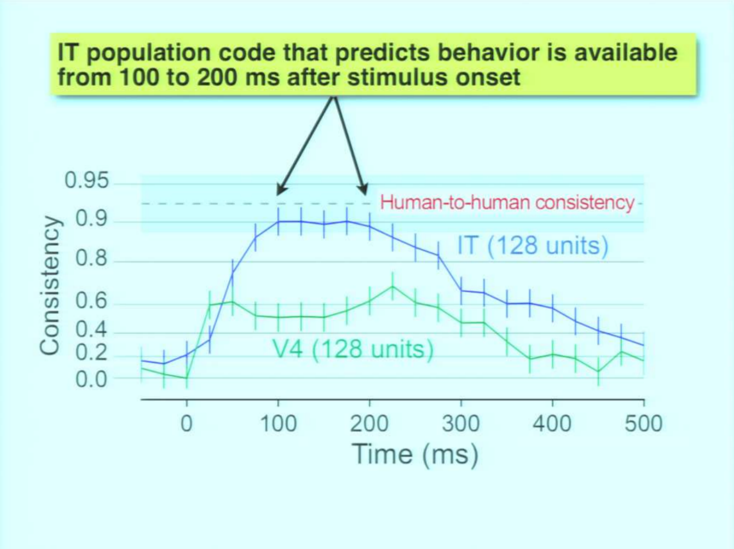
DiCarlo 2013 Tutorial on Vision



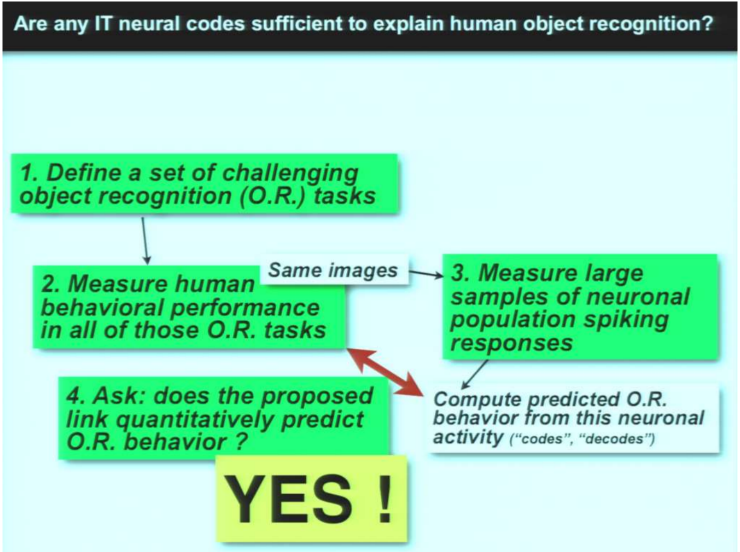
DiCarlo 2013 Tutorial on Vision



DiCarlo 2013 Tutorial on Vision



DiCarlo 2013 Tutorial on Vision



Vision: From Vision to Language

We can explain human/ape object recognition by:

- Recording apes' neuronal activity
  and attaching a single-layer NN to interpret it
- Measuring human performance

… on the same object recognition tasks.

- and relating them.

Idea:

- Record NMT behaviour (all parameters accessible)
- and human behaviour, possibly recording:
  - Objective: reading studies, eye-tracking, …
  - Subjective: introspection.

… on the same language processing tasks.

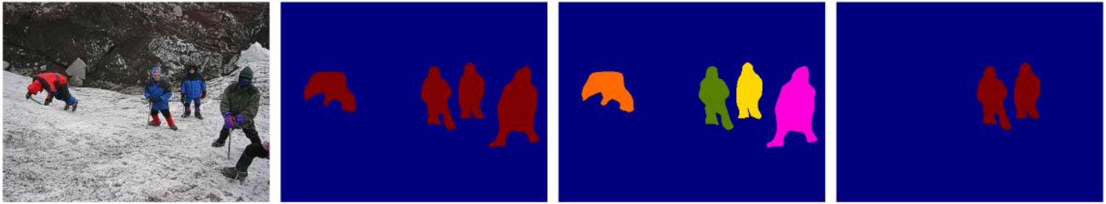
- and relate them.

Aspects of Meaning

- Meaning is a coarsening:
  - Pictures: Semantic segmentation ("reverse raytracing")
  - Programs: The output they give (caveat: undecidable).
  - Comp. Linguistics: Reference to real world? Speaker's intention?
- Meaning can be shifted, modified.
- Meanings can be compared.
- Meaning is generally compositional.
- Linguistic meaning captures the structure of expressions:
  - Morphology, syntax, ...
- Pragmatics: Named entities, numbers, anaphora...
- Expressions are ambiguous.
- Meanings are vague.
- Continuousness.

Meaning as a Coarsening

Semantic Segmentation of Pictures



(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"*

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

Compositionality of Meaning

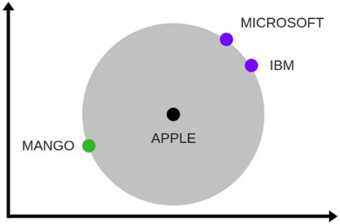
Manning (2015):

> *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.*

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.



Ideally, an expression would correspond to a distribution over semantic space.

Meaning Statefulness

Stateful Meaning Representation:

- "The state of mind after having read this and produced this output so far."
- Corresponds to models with attention.
- Btw needed to interpret humour (Gluscevskij, 2017).

Stateless Meaning Representation:

- Points correspond to expressions.
  - Ambiguity representation unclear.
- Points correspond to meanings.
  - As in models without attention.

Is Sentence Meaning Continuous?

We know that one English sentence can have 70k Czech translations (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.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
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.

Is Sentence Meaning Continuous?

Similarly for English (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".

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.

Work in progress with Chris Callison-Burch and Petra Barančíková.

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 (1/3)

Can you help me find something?
Would you help me look? Find this for me. Help me find something. Please help me find something. Will you help me? Your assistance in finding something is required.
I need you to help me get something.

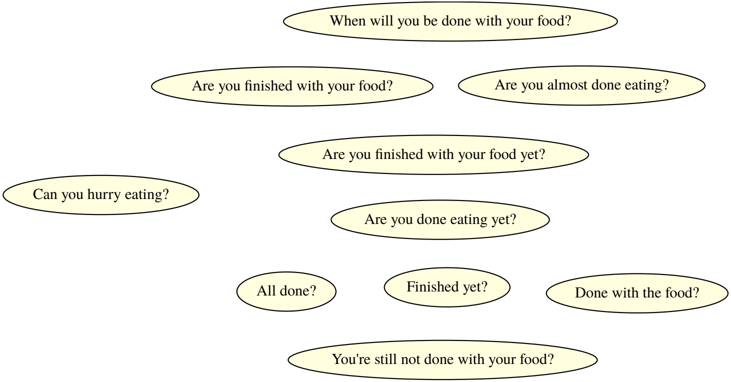
First Results of Midpointing (2/3)

Can you please give me a minute?
I'd like a minute alone.
Please wait.
Give me a minute.
One moment.
I need more time.
Come back later
Hey give me a minute.
One minute.
I need a minute to myself.
Could you leave me alone?

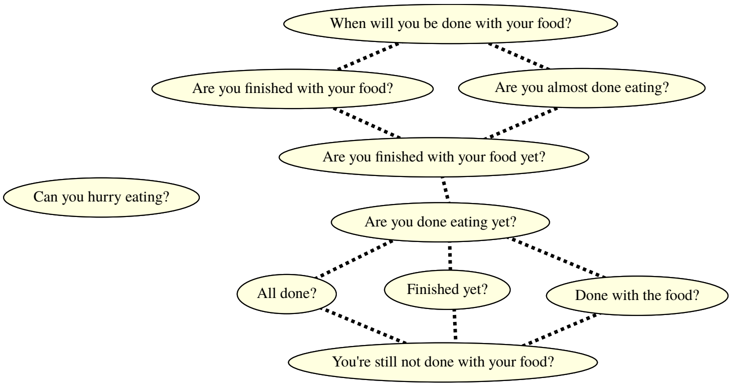
First Results of Midpointing (3/3)

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.

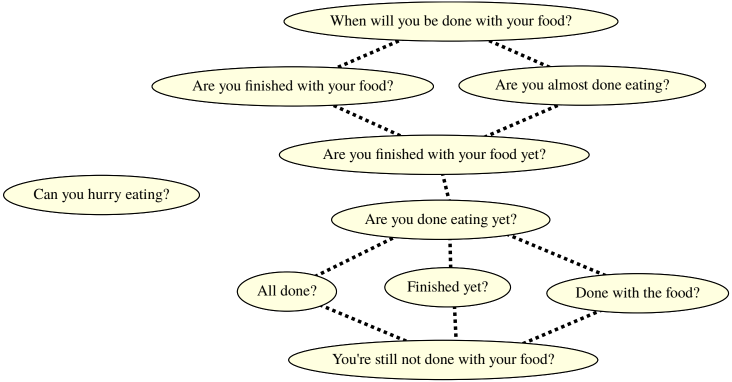
After the Midpointing...



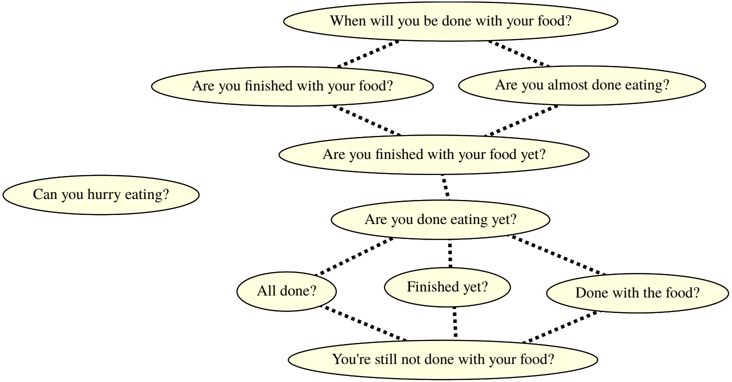
Ask Crowd to Partially Sort Them



Find Methods for Manifold Learning



Match Posets with Learned Manifolds



semi-supervised.

Some Techniques of NN Inspection

- MicroNNs, e.g. Shi et al. (2016) learning length.
- Lobotomy.
- Exploring representation space.
  - t-SNE and PCA for sentence pairs
  - Translation by search = similarity in meaning reflected in space
  - Attaching an NN to see if it can infer:
    - POS or morphology from NMT
    - Subject-Verb agreement (Linzen et al. TACL/EACL 2017)
- Linguistic exploration:
  - Various test suites (Burlot 2017, Burchhardt MQM).
  - Stanford Natural Language Inference (SNLI)
    `https://nlp.stanford.edu/projects/snli/`
  - Paraphrases (see above).

Summary

- Word vectors common and heavily used.
  - With NNs or without, esp. for fallback/robustness.
  - Usually evaluated by similarity/relatedness (somewhat dubious).
- Phrase/sentence representations very actively studied.
  - As with words, sentence representations can capture many things.
  - Representations good for NMT so far not good for meaning.
- NMT/DL very attractive for studying human language.
  - Aspects of meaning discussed.
  - NNs fit very closely to the given task (BLEU vs. SentEval).
    $\Rightarrow$ Multitask setups needed (still waiting for positive results).

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.

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.

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.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555.

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, Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on