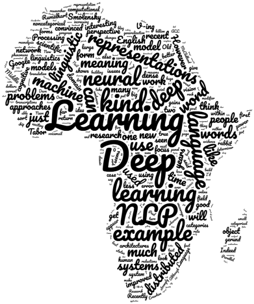
Frontiers of Natural Language Processing

Deep Learning Indaba 2018, Stellenbosch, South Africa

Sebastian Ruder, Herman Kamper, Panellists, Leaders in NLP, Everyone

DEEP LEARNING INDABA



Goals of session

1. What is NLP? What are the major developments in the last few years?
2. What are the biggest open problems in NLP?
3. Get to know the local community and start thinking about collaborations

What is NLP? What were the major advances?

A Review of the Recent History of NLP

What is NLP? What were the major advances?

A Review of the Recent History of NLP

Sebastian Ruder

Timeline

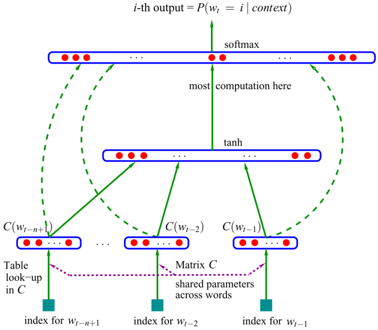
- 2001 Neural language models
- 2008 Multi-task learning
- 2013 Word embeddings
- 2013 Neural networks for NLP
- 2014 Sequence-to-sequence models
- 2015 Attention
- 2015 Memory-based networks
- 2018 Pretrained language models

Timeline

- 2001 Neural language models
- 2008 Multi-task learning
- 2013 Word embeddings
- 2013 Neural networks for NLP
- 2014 Sequence-to-sequence models
- 2015 Attention
- 2015 Memory-based networks
- 2018 Pretrained language models

Neural language models

- Language modeling: predict next word given previous words
- Classic language models: n-grams with smoothing
- First neural language models: feed-forward neural networks that take into account n previous words
- Initial look-up layer is commonly known as word embedding matrix as each word corresponds to one vector



[Bengio et al., NIPS '01; Bengio et al., JMLR '03]

Neural language models

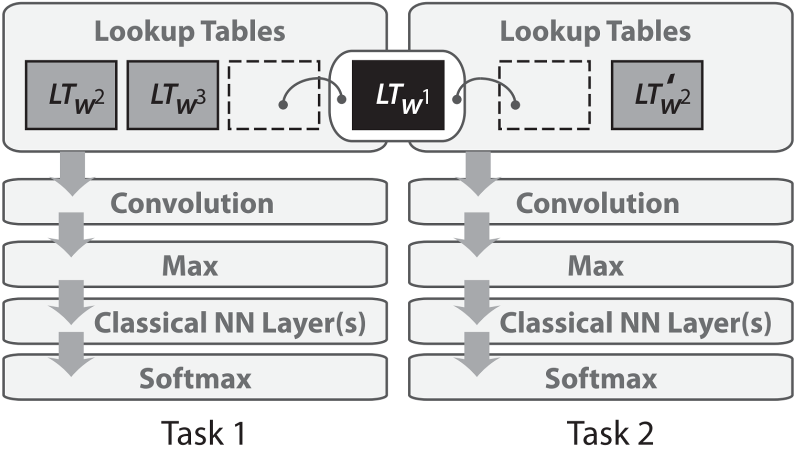
- Later language models: RNNs and LSTMs [Mikolov et al., Interspeech '10]
- Many new models in recent years; classic LSTM is still a strong baseline [Melis et al., ICLR '18]
- Active research area: What information do language models capture?
- Language modelling: despite its simplicity, core to many later advances
    - Word embeddings: the objective of word2vec is a simplification of language modelling
    - Sequence-to-sequence models: predict response word-by-word
    - Pretrained language models: representations useful for transfer learning

Timeline

- 2001 Neural language models
- 2008 Multi-task learning
- 2013 Word embeddings
- 2013 Neural networks for NLP
- 2014 Sequence-to-sequence models
- 2015 Attention
- 2015 Memory-based networks
- 2018 Pretrained language models

Multi-task learning

- Multi-task learning: sharing parameters between models trained on multiple tasks



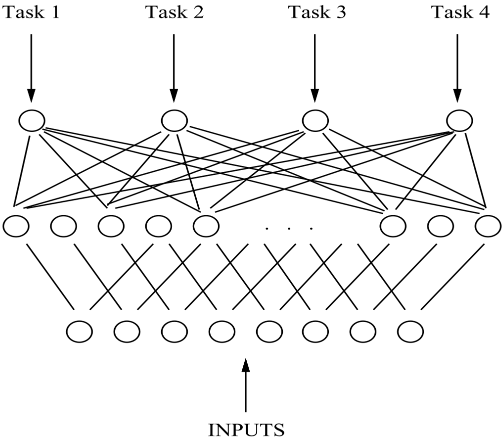
[Collobert & Weston, ICML '08; Collobert et al., JMLR '11]

Multi-task learning

- [Collobert & Weston, ICML '08] won Test-of-time Award at ICML 2018
- Paper contained a lot of other influential ideas:
  - Word embeddings
  - CNNs for text

Multi-task learning

- Multi-task learning goes back a lot further



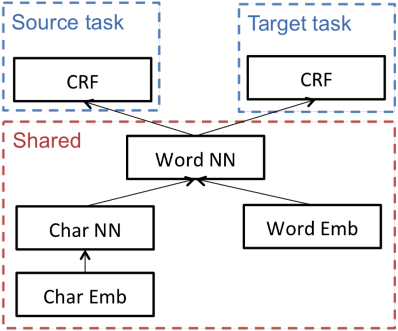
[Caruana, ICML '93; Caruana, ICML '96]

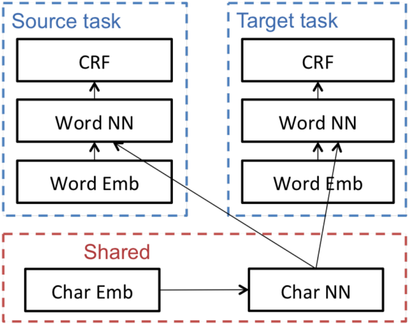
Multi-task learning

- "Joint learning" / "multi-task learning" used interchangeably
- Now used for many tasks in NLP, either using existing tasks or "artificial" auxiliary tasks
  - MT + dependency parsing / POS tagging / NER
  - Joint multilingual training
  - Video captioning + entailment + next-frame prediction [Pasunuru & Bansal; ACL '17]
  - ...

Multi-task learning

- Sharing of parameters is typically predefined
- Can also be learned [Ruder et al., '17]





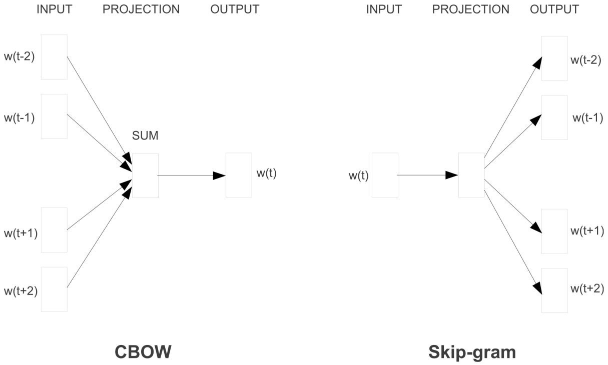
[Yang et al., ICLR '17]

Timeline

- 2001 Neural language models
- 2008 Multi-task learning
- 2013 Word embeddings
- 2013 Neural networks for NLP
- 2014 Sequence-to-sequence models
- 2015 Attention
- 2015 Memory-based networks
- 2018 Pretrained language models

Word embeddings

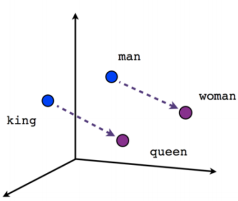
- Main innovation: pretraining word embedding look-up matrix on a large unlabelled corpus
- Popularized by word2vec, an efficient approximation to language modelling
- word2vec comes in two variants: skip-gram and CBOW



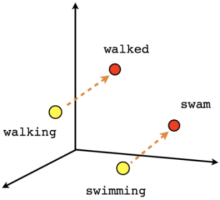
[Mikolov et al., ICLR '13; Mikolov et al., NIPS '13]

Word embeddings

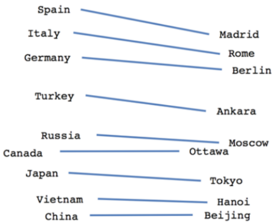
- Word embeddings pretrained on an unlabelled corpus capture certain relations between words



Male-Female



Verb tense



Country-Capital

[Tensorflow tutorial]

Word embeddings

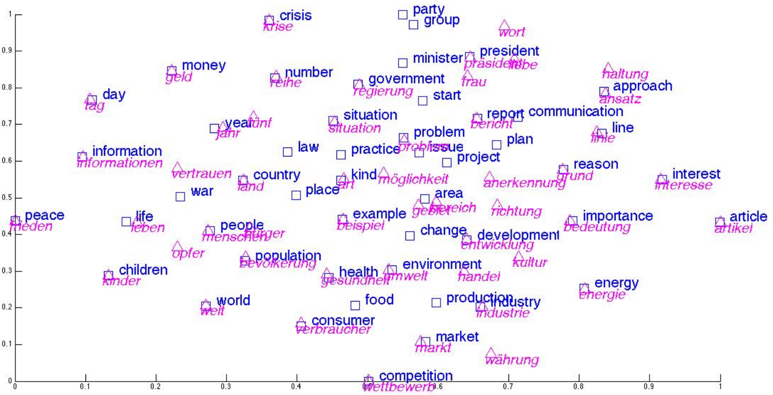
- Pretrained word embeddings have been shown to improve performance on many downstream tasks [Kim, EMNLP '14]
- Later methods show that word embeddings can also be learned via matrix factorization [Pennington et al., EMNLP '14; Levy et al., NIPS '14]
- Nothing inherently special about word2vec; classic methods (PMI, SVD) can also be used to learn good word embeddings from unlabeled corpora [Levy et al., TACL '15]

Word embeddings

- Lots of work on word embeddings, but word2vec is still widely used
- Skip-gram has been applied to learn representations in many other settings, e.g. sentences [Le & Mikolov, ICML '14; Kiros et al., NIPS '15], networks [Grover & Leskovec, KDD '16], biological sequences [Asgari & Mofrad, PLoS One '15], etc.

Word embeddings

- Projecting word embeddings of different languages into the same space enables (zero-shot) cross-lingual transfer [Ruder et al., JAIR '18]



[Luong et al., '15]

Timeline

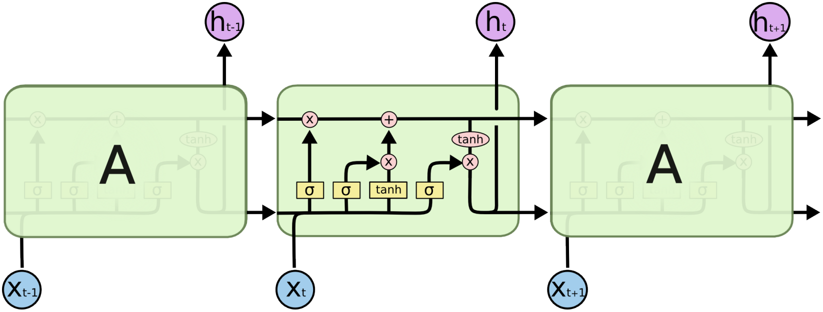
2001	Neural language models
2008	Multi-task learning
2013	Word embeddings
2013	Neural networks for NLP
2014	Sequence-to-sequence models
2015	Attention
2015	Memory-based networks
2018	Pretrained language models

Neural networks for NLP

- Key challenge for neural networks: dealing with dynamic input sequences
- Three main model types
  - Recurrent neural networks
  - Convolutional neural networks
  - Recursive neural networks

Recurrent neural networks

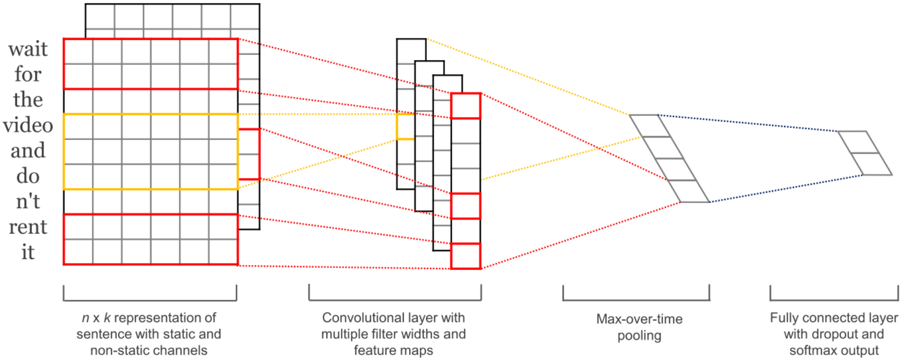
- Vanilla RNNs [Elman, CogSci '90] are typically not used as gradients vanish or explode with longer inputs
- Long-short term memory networks [Hochreiter & Schmidhuber, NeuComp '97] are the model of choice



[Olah, '15]

Convolutional neural networks

- 1D adaptation of convolutional neural networks for images
- Filter is moved along temporal dimension



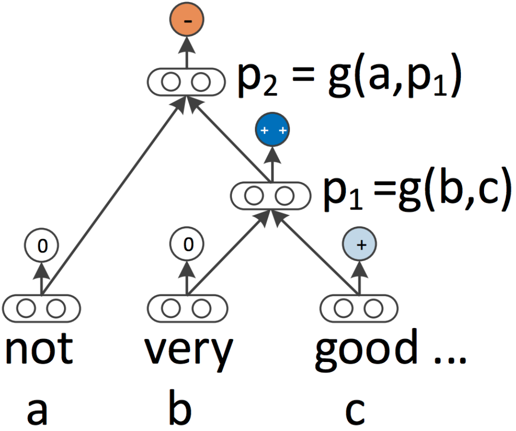
[Kim, EMNLP '14]

Convolutional neural networks

- More parallelizable than RNNs, focus on local features
- Can be extended with wider receptive fields (dilated convolutions) to capture wider context [Kalchbrenner et al., '17]
- CNNs and LSTMs can be combined and stacked [Wang et al., ACL '16]
- Convolutions can be used to speed up an LSTM [Bradbury et al., ICLR '17]

Recursive neural networks

- Natural language is inherently hierarchical
- Treat input as tree rather than as a sequence
- Can also be extended to LSTMs [Tai et al., ACL '15]



[Socher et al., EMNLP '13]

Other tree-based based neural networks

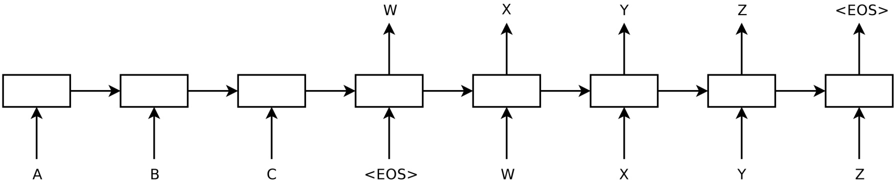
- Word embeddings based on dependencies [Levy and Goldberg, ACL '14]
- Language models that generate words based on a syntactic stack [Dyer et al., NAACL '16]
- CNNs over a graph (trees), e.g. graph-convolutional neural networks [Bastings et al., EMNLP '17]

Timeline

- 2001 Neural language models
- 2008 Multi-task learning
- 2013 Word embeddings
- 2013 Neural networks for NLP
- **2014 Sequence-to-sequence models**
- 2015 Attention
- 2015 Memory-based networks
- 2018 Pretrained language models

Sequence-to-sequence models

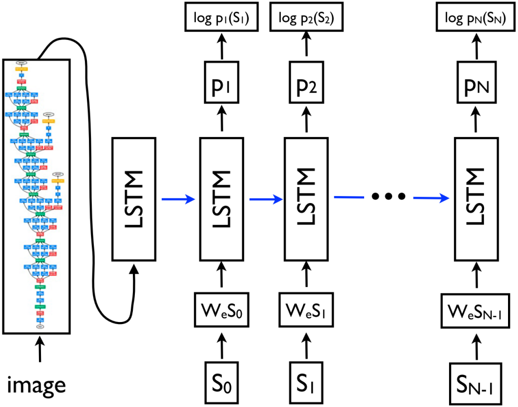
- General framework for applying neural networks to tasks where output is a sequence
- Killer application: Neural Machine Translation
- Encoder processes input word by word; decoder then predicts output word by word



[Sutskever et al., NIPS '14]

Sequence-to-sequence models

- Go-to framework for natural language generation tasks
- Output can not only be conditioned on a sequence, but on arbitrary representations, e.g. an image for image captioning

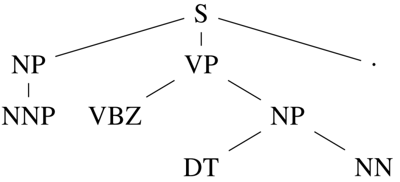


[Vinyals et al., CVPR '15]

Sequence-to-sequence models

- Even applicable to structured prediction tasks, e.g. constituency parsing [Vinyals et al., NIPS '15], named entity recognition [Gillick et al., NAACL '16], etc. by linearizing the output

John has a dog . $\rightarrow$



John has a dog . $\rightarrow$ (S (NP NNP $)_{\mathrm{NP}}$ (VP VBZ (NP DT NN $)_{\mathrm{NP}}$ $)_{\mathrm{VP}}$. $)_{\mathrm{S}}$

[Vinyals et al., NIPS '15]

Sequence-to-sequence models

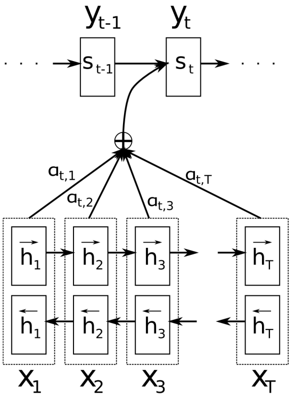
- Typically RNN-based, but other encoders and decoders can be used
- New architectures mainly coming out of work in Machine Translation
- Recent models: Deep LSTM [Wu et al., '16], Convolutional encoders [Kalchbrenner et al., arXiv '16; Gehring et al., arXiv '17], Transformer [Vaswani et al., NIPS '17], Combination of LSTM and Transformer [Chen et al., ACL '18]

Timeline

2001	Neural language models
2008	Multi-task learning
2013	Word embeddings
2013	Neural networks for NLP
2014	Sequence-to-sequence models
2015	Attention
2015	Memory-based networks
2018	Pretrained language models

Attention

- One of the core innovations in Neural Machine Translation
- Weighted average of source sentence hidden states
- Mitigates bottleneck of compressing source sentence into a single vector



[Bahdanau et al., ICLR '15]

Attention

- Different forms of attention available [Luong et al., EMNLP '15]
- Widely applicable: constituency parsing [Vinyals et al., NIPS '15], reading comprehension [Hermann et al., NIPS '15], one-shot learning [Vinyals et al., NIPS '16], image captioning [Xu et al., ICML '15]

[Xu et al., ICML '15]

Attention

- Not only restricted to looking at an another sequence
- Can be used to obtain more contextually sensitive word representations by attending to the same sequence $\rightarrow$ self-attention
- Used in Transformer [Vaswani et al., NIPS '17], state-of-the-art architecture for machine translation

Timeline

- 2001 Neural language models
- 2008 Multi-task learning
- 2013 Word embeddings
- 2013 Neural networks for NLP
- 2014 Sequence-to-sequence models
- 2015 Attention
- 2015 Memory-based networks
- 2018 Pretrained language models

Memory-based neural networks

- Attention can be seen as fuzzy memory
- Models with more explicit memory have been proposed
- Different variants: Neural Turing Machine [Graves et al., arXiv '14], Memory Networks [Weston et al., ICLR '15] and End-to-end Memory Networks [Sukhbaatar et al., NIPS '15], Dynamic Memory Networks [Kumar et al., ICML '16], Neural Differentiable Computer [Graves et al., Nature '16], Recurrent Entity Network [Henaff et al., ICLR '17]

Memory-based neural networks

- Memory is typically accessed based on similarity to current state similar to attention; can be written to and read from
- End-to-end Memory Networks [Sukhbaatar et al., NIPS '15] process input multiple times and update memory
- Neural Turing Machines also have a location-based addressing; can learn simple computer programs like sorting
- Memory can be a knowledge base or populated based on input

Timeline

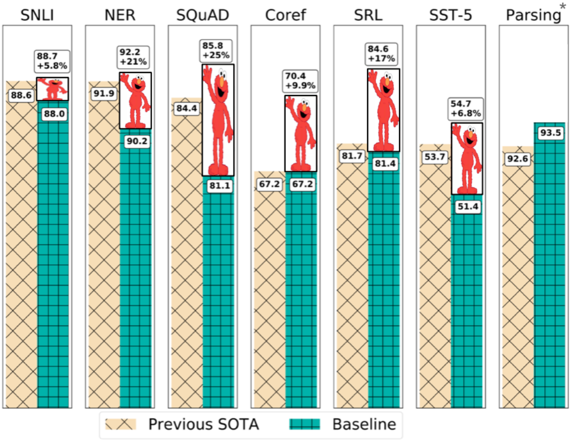
- 2001 Neural language models
- 2008 Multi-task learning
- 2013 Word embeddings
- 2013 Neural networks for NLP
- 2014 Sequence-to-sequence models
- 2015 Attention
- 2015 Memory-based networks
- 2018 Pretrained language models

Pretrained language models

- Word embeddings are context-agnostic, only used to initialize first layer
- Use better representations for initialization or as features
- Language models pretrained on a large corpus capture a lot of additional information
- Language model embeddings can be used as features in a target model [Peters et al., NAACL '18] or a language model can be fine-tuned on target task data [Howard & Ruder, ACL '18]

Pretrained language models

- Adding language model embeddings gives a large improvement over state-of-the-art across many different tasks



*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

[Peters et al., '18]

Pretrained language models

- Enables learning models with significantly less data
- Additional benefit: Language models only require unlabelled data
- Enables application to low-resource languages where labelled data is scarce

Other milestones

- Character-based representations
    - Use a CNN/LSTM over characters to obtain a character-based word representation
    - First used for sequence labelling tasks [Lample et al., NAACL '16; Plank et al., ACL '16]; now widely used
    - Even fully character-based NMT [Lee et al., TACL '17]
- Adversarial learning
    - Adversarial examples are becoming widely used [Jia & Liang, EMNLP '17]
    - (Virtual) adversarial training [Miyato et al., ICLR '17; Yasunaga et al., NAACL '18] and domain-adversarial loss [Ganin et al., JMLR '16; Kim et al., ACL '17] are useful forms of regularization
    - GANs are used, but not yet too effective for NLG [Semeniuta et al., '18]
- Reinforcement learning
    - Useful for tasks with a temporal dependency, e.g. selecting data [Fang & Cohn, EMNLP '17; Wu et al., NAACL '18] and dialogue [Liu et al., NAACL '18]
    - Also effective for directly optimizing a surrogate loss (ROUGE, BLEU) for summarization [Paulus et al., ICLR '18;] or MT [Ranzato et al., ICLR '16]

The Biggest Open Problems in NLP

The Biggest Open Problems in NLP

Sebastian Ruder

Jade Abbott

Stephan Gouws

Omoju Miller

Bernardt Duvenhage

The biggest open problems: Answers from experts

Hal Daumé III Barbara Plank Miguel Ballesteros Anders Søgaard Manaal Faruqui Mikel Artetxe Sebastian Riedel Isabelle Augenstein Bernardt Duvenhage Lea Frermann Brink van der Merwe Karen Livescu Jan Buys Kevin Gimpel Christine de Kock Alta de Waal Michael Roth Maletěabisa Molapo Annie Louise Chris Dyer Yoshua Bengio Felix Hill Kevin Knight Richard Socher George Dahl Dirk Hovy Kyunghyun Cho

We asked the experts:

What are the three biggest open problems in NLP at the moment?

The biggest open problems in NLP

1. Natural language understanding
2. NLP for low-resource scenarios
3. Reasoning about large or multiple documents
4. Datasets, problems and evaluation

Problem 1: Natural language understanding

- Many experts argued that this is central, also for generation
- Almost none of our current models have “real” understanding
- What (biases, structure) should we build explicitly into our models?
- Models should incorporate common sense
- Dialogue systems (and chat bots) were mentioned in several responses

Problem 1: Natural language understanding

Article: Nicola Tesla

Paragraph: *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*

Problem 1: Natural language understanding

Article: Nicola Tesla

Paragraph: *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*

Question: What city did Tesla move to in 1880?

Problem 1: Natural language understanding

Article: Nicola Tesla

Paragraph: *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*

Question: What city did Tesla move to in 1880?

Answer: Prague

Problem 1: Natural language understanding

Article: Nicola Tesla

Paragraph: *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*

Question: What city did Tesla move to in 1880?

Answer: Prague

Model predicts: Prague

Problem 1: Natural language understanding

Article: Nicola Tesla

Paragraph: *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses. Tadakatsu moved to the city of Chicago in 1881.*

Question: What city did Tesla move to in 1880?

Answer:

Model predicts:

Problem 1: Natural language understanding

Article: Nicola Tesla

Paragraph: *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses. Tadakatsu moved to the city of Chicago in 1881.*

Question: What city did Tesla move to in 1880?

Answer: Prague

Model predicts:

Problem 1: Natural language understanding

Article: Nicola Tesla

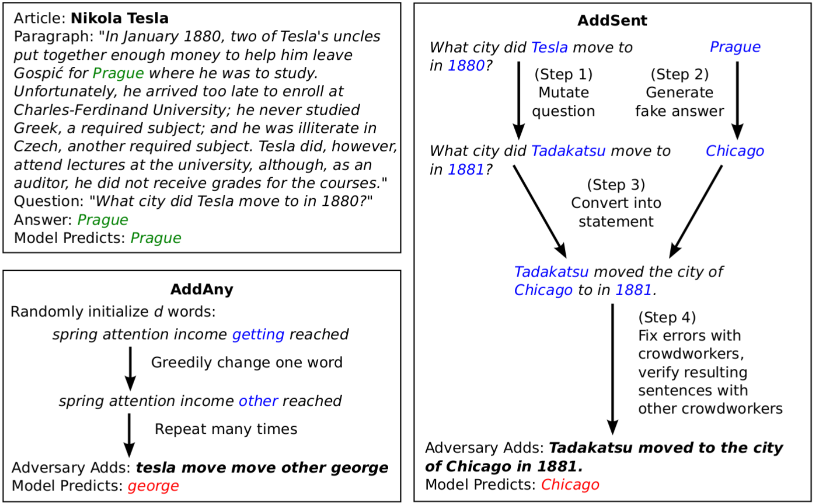
Paragraph: *In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses. Tadakatsu moved to the city of Chicago in 1881.*

Question: What city did Tesla move to in 1880?

Answer: Prague

Model predicts: Chicago

Problem 1: Natural language understanding



[Jia and Liang, EMNLP'17]

Problem 1: Natural language understanding

I think the biggest open problems are all related to natural language understanding... **we should develop systems that read and understand text the way a person does**, by forming a representation of the world of the text, with the agents, objects, settings, and the relationships, goals, desires, and beliefs of the agents, and everything else that humans create to understand a piece of text.

Until we can do that, all of our progress is in improving our systems' ability to do pattern matching. Pattern matching can be very effective for developing products and improving people's lives, so I don't want to denigrate it, but ...

— *Kevin Gimpel*

Problem 1: Natural language understanding

Questions to panellists/audience:

- To achieve NLU, is it important to build models that process language "the way a person does"? [Also see https://www.abigailsee.com/2018/02/21/deep-learning-structure-and-innate-priors.html]

Problem 1: Natural language understanding

Questions to panellists/audience:

- To achieve NLU, is it important to build models that process language “the way a person does”? [Also see https://www.abigailsee.com/2018/02/21/deep-learning-structure-and-innate-priors.html]
- How do you think we would go about doing this?

Problem 1: Natural language understanding

Questions to panellists/audience:

- To achieve NLU, is it important to build models that process language "the way a person does"? [Also see https://www.abigailsee.com/2018/02/21/deep-learning-structure-and-innate-priors.html]
- How do you think we would go about doing this?
- Do we need inductive biases or can we expect models to learn everything from enough data?

Problem 1: Natural language understanding

Questions to panellists/audience:

- To achieve NLU, is it important to build models that process language "the way a person does"? [Also see https://www.abigailsee.com/2018/02/21/deep-learning-structure-and-innate-priors.html]
- How do you think we would go about doing this?
- Do we need inductive biases or can we expect models to learn everything from enough data?
- Questions from audience

Problem 2: NLP for low-resource scenarios

Problem 2: NLP for low-resource scenarios

- **Generalisation beyond the training data**

Problem 2: NLP for low-resource scenarios

- **Generalisation beyond the training data** – relevant everywhere!

Problem 2: NLP for low-resource scenarios

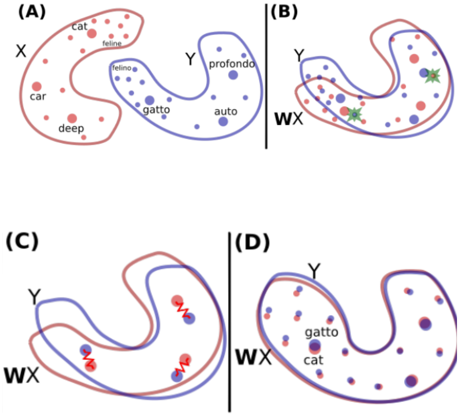
- **Generalisation beyond the training data** – relevant everywhere!
- Domain-transfer, transfer learning, multi-task learning
- Learning from small amounts of data
- Semi-supervised, weakly-supervised, “Wiki-ly” supervised, distantly-supervised, lightly-supervised, minimally-supervised

Problem 2: NLP for low-resource scenarios

- **Generalisation beyond the training data** – relevant everywhere!
- Domain-transfer, transfer learning, multi-task learning
- Learning from small amounts of data
- Semi-supervised, weakly-supervised, “Wiki-ly” supervised, distantly-supervised, lightly-supervised, minimally-supervised
- Unsupervised learning

Problem 2: NLP for low-resource scenarios

Word translation without parallel data:



[Conneau et al., ICLR'18]

Problem 2: NLP for low-resource scenarios

Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and **James Glass**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
`{andyyuan,ckbjimmy,st9,glass}mit.edu`

Abstract

Recent research has shown that word embedding spaces learned from text corpora of different languages can be aligned without any parallel data supervision. Inspired by the success in unsupervised cross-lingual word embeddings, in this paper we target learning a *cross-modal* alignment between the embedding spaces of speech and text learned from corpora of their respective modalities in an unsupervised fashion. The proposed framework learns the individual speech and text embedding spaces, and attempts to align the two spaces via adversarial training, followed by a refinement procedure. We show how our framework could be used

[Chung et al., arXiv'18]

Problem 2: NLP for low-resource scenarios

Questions to panellists/audience:

- Is it necessary to develop specialised NLP tools for specific languages, or is it enough to work on general NLP?

Problem 2: NLP for low-resource scenarios

Questions to panellists/audience:

- Is it necessary to develop specialised NLP tools for specific languages, or is it enough to work on general NLP?
- *Since there is inherently only small amounts of text available for under-resourced languages, the benefits of NLP in such settings will also be limited.* Agree or disagree?

Problem 2: NLP for low-resource scenarios

Questions to panellists/audience:

- Is it necessary to develop specialised NLP tools for specific languages, or is it enough to work on general NLP?
- *Since there is inherently only small amounts of text available for under-resourced languages, the benefits of NLP in such settings will also be limited.* Agree or disagree?
- Unsupervised learning vs. transfer learning from high-resource languages?
- Questions from audience

Problem 3: Reasoning about large or multiple documents

- Related to understanding
- How do we deal with large contexts?
- Can be either text or spoken documents
- Again incorporating common sense is essential

Problem 3: Reasoning about large or multiple documents

Example from NarrativeQA dataset:

Title: Ghostbusters II
Question: How is Oscar related to Dana?
Answer: her son
Summary snippet: …Peter's former girlfriend Dana Barrett has had a son, Oscar…
Story snippet:

DANA (setting the wheel brakes on the buggy)
Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)
Hiya, Oscar. What do you say, slugger?

FRANK (to Dana)
That's a good-looking kid you got there, Ms. Barrett.

[Kočiský et al., TACL'18]

Problem 3: Reasoning about large or multiple documents

Questions to panellists/audience:

- Do we need better models or just train on more data?

Problem 3: Reasoning about large or multiple documents

Questions to panellists/audience:

- Do we need better models or just train on more data?
- Questions from audience

Problem 4: Datasets, problems and evaluation

Perhaps the biggest problem is to **properly define the problems themselves.** And by properly defining a problem, I mean building **datasets and evaluation** procedures that are appropriate to measure our progress towards concrete goals. Things would be easier if we could reduce everything to Kaggle style competitions!

— *Mikel Artetxe*

. . . basic resources (e.g. stop word lists)

— *Alta de Waal*

Problem 4: Datasets, problems and evaluation

https://rma.nwu.ac.za

Problem 4: Datasets, problems and evaluation

Questions to panellists/audience:

- What are the most important NLP problems that should be tackled for societies in Africa?

Problem 4: Datasets, problems and evaluation

Questions to panellists/audience:

- What are the most important NLP problems that should be tackled for societies in Africa?
- How do we make sure that we don't overfit to our benchmarks?

Problem 4: Datasets, problems and evaluation

Questions to panellists/audience:

- What are the most important NLP problems that should be tackled for societies in Africa?
- How do we make sure that we don't overfit to our benchmarks?
- Questions from audience

We asked the experts a few more questions:

We asked the experts a few more questions:

What, if anything, has led the field in the wrong direction?

What has led the field in the wrong direction?

- "Synthetic data/synthetic problems" — *Hal Daumé III*
- "Benchmark/leaderboard chasing" — *Sebastian Riedel*
- "Obsession of ... beating the state of the art through "neural architecture search" — *Isabelle Augenstein*

What has led the field in the wrong direction?

- "Synthetic data/synthetic problems" — *Hal Daumé III*
- "Benchmark/leaderboard chasing" — *Sebastian Riedel*
- "Obsession of ... beating the state of the art through "neural architecture search" — *Isabelle Augenstein*
- "Chomskyan theories of linguistics instead of corpus linguistics" — *Brink van der Merwe*

What has led the field in the wrong direction?

- "Synthetic data/synthetic problems" — *Hal Daumé III*
- "Benchmark/leaderboard chasing" — *Sebastian Riedel*
- "Obsession of ... beating the state of the art through "neural architecture search" — *Isabelle Augenstein*
- "Chomskyan theories of linguistics instead of corpus linguistics" — *Brink van der Merwe*
- "Not incorporating enough Chomskyan theory into our models" — *Someone Else*

What has led the field in the wrong direction?

- "Synthetic data/synthetic problems" — *Hal Daumé III*
- "Benchmark/leaderboard chasing" — *Sebastian Riedel*
- "Obsession of ... beating the state of the art through "neural architecture search" — *Isabelle Augenstein*
- "Chomskyan theories of linguistics instead of corpus linguistics" — *Brink van der Merwe*
- "Not incorporating enough Chomskyan theory into our models" — *Someone Else*
- "Too much emphasis on Bayesian methods (sorry :)"— *Karen Livescu*

What has led the field in the wrong direction?

- "Synthetic data/synthetic problems" — *Hal Daumé III*
- "Benchmark/leaderboard chasing" — *Sebastian Riedel*
- "Obsession of ... beating the state of the art through "neural architecture search" — *Isabelle Augenstein*
- "Chomskyan theories of linguistics instead of corpus linguistics" — *Brink van der Merwe*
- "Not incorporating enough Chomskyan theory into our models" — *Someone Else*
- "Too much emphasis on Bayesian methods (sorry :)"— *Karen Livescu*
- "Haha, as if the field as a whole moved in a single direction" — *Michael Roth*

What has led the field in the wrong direction?

I don't think there is anything like that. We can learn from "wrong" directions and "correct" directions, if such a thing even exists.

— *Miguel Ballesteros*

Anything new will temporarily lead the field in the wrong direction, I guess, but upon returning, we may nevertheless have pushed research horizons.

— *Anders Søgaard*

Sentiment shared in many of the other responses

We asked the experts a few more questions:

We asked the experts a few more questions:

What advice would you give a postgraduate student in NLP starting their project now?

What advice would you give a postgraduate student in NLP starting their project now?

Do not limit yourself to reading NLP papers. **Read a lot** of machine learning, deep learning, reinforcement learning papers. A PhD is a great time in one's life to go for a **big goal**, and even small steps towards that will be valued.

— *Yoshua Bengio*

Learn how to tune your models, **learn how to make strong baselines**, and learn how to build baselines that test particular hypotheses. **Don't take any single paper too seriously**, wait for its conclusions to show up more than once.

— *George Dahl*

What advice would you give a postgraduate student in NLP starting their project now?

i believe **scientific pursuit is meant to be full of failures.** ... if every idea works out, it's either (a) you're not ambitious enough, (b) you're subconciously cheating yourself, or (c) you're a genius, the last of which i heard happens only once every century or so. so, don't despair!

— *Kyunghyun Cho*

Understand **psychology** and the core problems of semantic cognition. Read ... Go to CogSci. Understand **machine learning**. Go to NIPS. **Don't worry about ACL. Submit something terrible** (or even good, if possible) to a workshop as soon as you can. You can't learn how to do these things without going through the process.

— *Felix Hill*

Summary of session

- What is NLP? What are the major developments in the last few years?
- What are the biggest open problems in NLP?
- Get to know the local community and start thinking about collaborations

Summary of session

- What is NLP? What are the major developments in the last few years?
- What are the biggest open problems in NLP?
- Get to know the local community and start thinking about collaborations
- We now have the closing ceremony, so eat and chat!