Artificial Intelligence and the Singularity

piero scaruffi

www.scaruffi.com
October 2014 - Revised 2019

"The person who says it cannot be done should not interrupt the person doing it" (Chinese proverb)
Piero Scaruffi

piero scaruffi
p@scaruffi.com
scaruffi@stanford.edu

Olivetti AI Center, 1987
Piero Scaruffi

- Cultural Historian
- Cognitive Scientist
- Blogger
- Poet
- www.scaruffi.com

A History of Jazz Music
1900-2000

A History of Rock and Dance Music
From the Gutter to the Laptop
From Chicago to Shanghai
Volume 2 (1990-2000)

The Nature of Consciousness
The Structure of Life and the Meaning of Matter

Synthesis
Essays, Photographs, Poems
This is Part 12

- See http://www.scaruffi.com/singular for the index of this Powerpoint presentation and links to the other parts
  1. Classic A.I. - The Age of Expert Systems
  2. The A.I. Winter and the Return of Connectionism
  3. Theory: Knowledge-based Systems and Neural Networks
  4. Robots
  5. Bionics
  6. Singularity
  7. Critique
  8. The Future
  9. Applications
  10. Machine Art
  11. The Age of Deep Learning
  12. Natural Language Processing
Natural Language Processing
1980s-Today
Natural Language Processing

1981: Hans Kamp`s Discourse Representation Theory
1983: Gerard Salton and Michael McGill's "Introduction to Modern Information Retrieval" (the "bag-of-words model")
1986: Barbara Grosz's "Attention, Intentions, and the Structure of Discourse"
1988: Fred Jelinek's team at IBM publishes "A Statistical Approach to Language Translation"
1990: Peter Brown at IBM implements a statistical machine translation system
Natural Language Processing

"Bag-of-words" model: sentence representations are independent of word order

Sequence models developed by Michael Jordan (1986) and Jeffrey Elman (1990)
Natural Language Processing

Neural Networks for Symbolic Representation

• Jordan Pollack (1990): representing tree structures in neural networks (Recursive Auto-Associative Memory or RAAM)

• Christoph Goller and Andreas Kuechler (1995): extension of Pollack’s RAAM

[Diagram of Recursive Distributed Representations by Jordan B. Pollack, 1990]

Learning Task-Dependent Distributed Representations by Backpropagation Through Structure, 1995

Christoph Goller
Andreas Kuechler

AR-report AR-95-02
Technische Universität München

Fig. 1: The Folding Architecture (left side) and the standard LRAAM (right side).
Natural Language Processing

1996: Tom Landauer’s and Susan Dumais’ "latent semantic analysis"
2001: John Lafferty’s "conditional random fields" for sequence labeling
Natural Language Processing

How should words be represented

• Traditional NLP: the word is an atom
• Bag of words: the atom is a set of words
• 2003: Yoshua Bengio’s Neural Probabilistic Language Model - represent words as vectors
• 2008: Mirella Lapata & Jeff Mitchell – represent sentences as vectors
Natural Language Processing

How should words be represented

- 2003: Yoshua Bengio’s Neural Probabilistic Language Model - represent words as vectors


A Neural Probabilistic Language Model

Yoshua Bengio
Réjean Ducharme
Pascal Vincent
Christian Jauvin

Département d’Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal, Montréal, Québec, Canada

1. associate with each word in the vocabulary a distributed word feature vector (a real-valued vector in $\mathbb{R}^m$),

2. express the joint probability function of word sequences in terms of the feature vectors of these words in the sequence, and

3. learn simultaneously the word feature vectors and the parameters of that probability function.
Natural Language Processing

How should words be represented

• 2008: Mirella Lapata & Jeff Mitchell – represent sentences as vectors

Vector-based Models of Semantic Composition

Jeff Mitchell and Mirella Lapata
School of Informatics, University of Edinburgh

<table>
<thead>
<tr>
<th></th>
<th>animal</th>
<th>stable</th>
<th>village</th>
<th>gallop</th>
<th>jokey</th>
</tr>
</thead>
<tbody>
<tr>
<td>horse</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>run</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1: A hypothetical semantic space for horse and run
Natural Language Processing

Neural Machine Translation

2001: Yoshua Bengio's Neural Probabilistic Language Model converts a word symbol into a vector within a meaning space

2005: Bengio's Hierarchical Probabilistic Neural Network Language Model solves the "curse of dimensionality" in NLP

2008: Ronan Collobert and Jason Weston's Unified Architecture for NLP learns recursive structures
Natural Language Processing

Neural networks that learn recursive structures

- Ronan Collobert and Jason Weston (2008): task-independent sequence tagging
Natural Language Processing

- Collobert and Weston (2011)

Natural Language Processing (almost) from Scratch

Ronan Collobert  
*NEC Labs America, Princeton NJ.*

Jason Weston  
*Google, New York, NY.*

Léon Bottou  

Michael Karlen  

Koray Kavukcuoglu†  

Pavel Kuksa‡  

*NEC Labs America, Princeton NJ.*
Natural Language Processing

Sequence tagging
- Generative models: hidden Markov models
- Conditional models: conditional random fields (John Laffery, 2001)
- Unified Architecture (Collobert & Weston, 2008)
Natural Language Processing

Neural Machine Translation
2010: Tomas Mikolov's RNN that can process sentences of any length
2010: Richard Socher's recursive neural network (RNN) for continuous phrase representation

Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks

Richard Socher, Christopher D. Manning, Andrew Y. Ng
Department of Computer Science
Stanford University

Figure 1: An example tree with a simple Recursive Neural Network
Natural Language Processing

Neural Machine Translation
2013: Nal Kalchbrenner and Phil Blunsom: statistical machine translation based purely on neural networks (“sequence to sequence learning”)
Natural Language Processing

Neural Machine Translation

June 2014: Bengio's encoder-decoder model
(with Kyunghyun Cho and Dzmitry Bahdanau)

By-product: instead of using an LSTM, they use a simpler type of RNN, later called Gated Recurrent Unit (GRU) with no memory unit
Neural Machine Translation
2014: Kyunghyun Cho’s Gated Recurrent Unit
Bengio's encoder-decoder model

By-product: instead of using an LSTM, they use a simpler type of RNN, later called Gated Recurrent Unit (GRU) with no memory unit.
Natural Language Processing

Neural Machine Translation

Sep 2014: Sutskever, Vinyals & Le solve the "sequence-to-sequence problem" using a LSTM (the length of the input sequence of characters doesn’t have to be the same length of the output)

They too use the encoder-decoder model
Natural Language Processing

Neural Machine Translation

June 2014: Attention model by Volodymyr Mnih

Recurrent Models of Visual Attention

Volodymyr Mnih  Nicolas Heess  Alex Graves  Koray Kavukcuoglu
Google DeepMind

The Recurrent Attention Model (RAM)

Abstract
Natural Language Processing

Neural Machine Translation
Sep 2014: Attention model by Dzmitry Bahdanau, Kyunghyun Cho & Bengio ("additive" attention)
Natural Language Processing

Neural Machine Translation
2015: Attention model by Kelvin Xu ("Show/Attend/Tell")
Natural Language Processing

Neural Machine Translation

2015: Dzmitry Bahdanau's BiRNN (bidirectional RNN) at Jacobs University Bremen in Germany to improve the speed of machine translation
Natural Language Processing

Neural Machine Translation
2016: Google’s dynamic coattention network (Socher)
Natural Language Processing

Neural Machine Translation

Nov 2016: Google switches its translation algorithm to an RNN

Figure 1: The model architecture of GNMT, Google’s Neural Machine Translation system. On the left is the encoder network, on the right is the decoder network, in the middle is the attention module.

Italian ▼  ▪ ▶  Chinese (Simplified) ▼  ▪ ▶

Ciao 你好

The structure of bi-directional connections in the first layer of the encoder.
Natural Language Processing

Neural Machine Translation: convolutional nets instead of RNNs
2016: Nal Kalchbrenner’s ByteNet

*Figure 1. The architecture of the ByteNet. The target decoder (blue) is stacked on top of the source encoder (red). The decoder generates the variable-length target sequence using dynamic unfolding.*
Natural Language Processing

Neural Machine Translation: convolutional nets instead of RNNs

2016: Facebook’s ConvS2S

Figure 1. Illustration of batching during training. The English source sentence is encoded (top) and we compute all attention values for the four German target words (center) simultaneously. Our attentions are just dot products between decoder context representations (bottom left) and encoder representations. We add the conditional inputs computed by the attention (center right) to the decoder states which then predict the target words (bottom right). The sigmoid and multiplicative boxes illustrate Gated Linear Units.
Natural Language Processing

Neural Machine Translation
Do these systems that translate one sentence into another sentence actually "understand" language?

Xing Shi (University of Southern California, 2016): the vector representations of neural machine translation capture some morphological and syntactic properties of language.
Natural Language Processing

Neural Machine Translation
Do these systems that translate one sentence into another sentence actually "understand" language?

Yonatan Belinkov (MIT, 2017): vector representations contain even some semantical properties
Natural Language Processing

Neural Machine Translation
Microsoft (2018): news translation

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Janczysz-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou

March 2018
Natural Language Processing

Neural Machine Translation

But beware… iFlytek scandal of 2018

National AI champion iFlytek in dispute over ‘automated’ speech translation at Shanghai forum

PUBLISHED: Tuesday, 25 September, 2018, 8:01 pm
UPDATED: Tuesday, 25 September, 2018, 8:00 pm
Natural Language Processing

Discourse Analysis

2013: Mikolov's “Word2vec" method for learning vector representations of words from large amounts of unstructured text data

2014: James Weston’s “memory networks“, neural networks coupled with long-term memories for question-answering
Natural Language Processing

How should words be represented

- Tomas Mikolov’s skip-gram (2013)
Natural Language Processing

Question-answering

- 2014: James Weston’s memory networks
Natural Language Processing

The old goal: relate words within a sentence
Success story: machine translation
The new goal: capture information about entire sentences
Application: understanding texts
Natural Language Processing

Discourse Analysis
2015: Richard Socher's dynamic memory networks

Ask Me Anything:
Dynamic Memory Networks for Natural Language Processing

Ankit Kumar, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong,
Romain Paulus, Richard Socher
MetaMind, Palo Alto, CA USA

Episodic Memory → Answer
Input Text Sequence → Question

2015
Natural Language Processing

Discourse Analysis

2015: Oriol Vinyals and Quoc Le's Neural Conversational Model
Natural Language Processing

Discourse Analysis
2014: GloVe (an alternative to Word2vec) by Jeffrey Pennington, Richard Socher, Chris Manning (Stanford)
2015: FastText (Mikolov, Facebook)

Bag of Tricks for Efficient Text Classification

Aug 2016

Armand Joulin  Edouard Grave  Piotr Bojanowski  Tomas Mikolov
Facebook AI Research
{ajoulin,egrave,bojanowski,tmikolov}@fb.com

Figure 1: Model architecture of fastText
Discourse Analysis

“Word embeddings" (like Word2vec and GloVe) derive a map of how words relate to each other based on the configurations in which the words appear in large amounts of text ("the distributional hypothesis" that words frequently occurring in the same contexts are related).
Natural Language Processing

Discourse Analysis
Next step: capture information about entire sentences
Jamie/Ryan Kiros: Skip-Thoughts (2015)
Lajanugen Logeswaran: Quick-Thoughts (2018)

Skip-Thought Vectors

AN EFFICIENT FRAMEWORK FOR LEARNING SENTENCE REPRESENTATIONS

Lajanugen Logeswaran* & Honglak Lee†
*University of Michigan, Ann Arbor, MI, USA
†Google Brain, Mountain View, CA, USA

Spring had come. → Enc → Dec → And yet his crops didn’t grow.

(a) Conventional approach

Spring had come. → Enc (f) → Dec
They were so black. → Enc (g) → Dec
And yet his crops didn’t grow. → Enc (g) → Dec
He had blue eyes. → Enc (g) → Dec → Classifier

(b) Proposed approach
## Natural Language Processing

### Discourse Analysis

Next step: capture information about entire sentences


<table>
<thead>
<tr>
<th>Task</th>
<th>Previous SOTA</th>
<th>ELMo +</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>SAN</td>
<td>84.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>85.8</td>
</tr>
<tr>
<td>SNLI</td>
<td>Chen et al (2017)</td>
<td>88.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>88.7</td>
</tr>
<tr>
<td>SRL</td>
<td>He et al (2017)</td>
<td>81.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>84.6</td>
</tr>
<tr>
<td>Coref</td>
<td>Lee et al (2017)</td>
<td>67.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70.4</td>
</tr>
<tr>
<td>NER</td>
<td>Peters et al (2017)</td>
<td>91.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>92.22</td>
</tr>
<tr>
<td>Sentiment</td>
<td>McCann et al (2017)</td>
<td>53.7</td>
</tr>
<tr>
<td>(5-class)</td>
<td></td>
<td>54.7</td>
</tr>
</tbody>
</table>

---

*Deep contextualized word representations*

Matthew E. Peters¹, Mark Neumann¹, Mohit Iyyer¹, Matt Gardner¹,
{matthewp, markn, mohiti, mattg}@allenai.org

Christopher Clark*, Kenton Lee*, Luke Zettlemoyer†*
{csquared, kentonl, lsz}@cs.washington.edu

†Allen Institute for Artificial Intelligence
*Paul G. Allen School of Computer Science & Engineering, University of Washington
Natural Language Processing

Discourse Analysis

TagML (Matthew Peters, 2017): bidirectional language model

Figure 2: Overview of TagLM, our language model augmented sequence tagging architecture.
## Natural Language Processing

### Attention

<table>
<thead>
<tr>
<th>Name</th>
<th>Alignment score function</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive(*)</td>
<td>$\text{score}(s_t, h_i) = v_a^T \tanh(W_a[s_t; h_i])$</td>
<td>Bahdanau2015</td>
</tr>
<tr>
<td>Location-Base</td>
<td>$\alpha_{t,i} = \text{softmax}(W_a s_t)$</td>
<td>Luong2015</td>
</tr>
<tr>
<td>General</td>
<td>$\text{score}(s_t, h_i) = s_t^T W_a h_i$</td>
<td>Luong2015</td>
</tr>
<tr>
<td>Dot-Product</td>
<td>$\text{score}(s_t, h_i) = s_t^T h_i$</td>
<td>Luong2015</td>
</tr>
<tr>
<td>Scaled</td>
<td>$\text{score}(s_t, h_i) = \frac{s_t^T h_i}{\sqrt{n}}$</td>
<td>Vaswani2017</td>
</tr>
<tr>
<td>Dot-Product(^)</td>
<td>Note: very similar to the dot-product attention except for a scaling factor; where $n$ is the dimension of the source hidden state.</td>
<td></td>
</tr>
<tr>
<td>Self-Attention(&amp;)</td>
<td>Relating different positions of the same input sequence.</td>
<td>Cheng2016</td>
</tr>
<tr>
<td>Global/Soft</td>
<td>Attending to the entire input state space.</td>
<td>Xu2015</td>
</tr>
<tr>
<td>Local/Hard</td>
<td>Attending to the part of input state space; i.e. a patch of the input image.</td>
<td>Xu2015; Luong2015</td>
</tr>
</tbody>
</table>
Natural Language Processing

Attention

2015: “dot-product” (multiplicative) method by Minh-Thang Luong, Chris

Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong  Hieu Pham  Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305

Figure 2: Global attentional model
Figure 3: Local attention model
Natural Language Processing

Attention

Attentive Reader (Phil Blunsom, Oxford University, 2015), a generalization of Weston's memory networks for question answering.
Natural Language Processing

Self-Attention
2016: Jianpeng Cheng, Mirella Lapata

Long Short-Term Memory-Networks for Machine Reading

Jianpeng Cheng, Li Dong and Mirella Lapata
School of Informatics, University of Edinburgh

Sep 2016

Diagram of Long Short-Term Memory-Networks for Machine Reading
Natural Language Processing

Self-Attention
2016: Ankur Parikh (Google)
2017: Richard Socher (Salesforce)
Natural Language Processing

Self-Attention

2017: Zhouhan Lin (Univ of Montreal)

A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING

Zhouhan Lin\textsuperscript{1,2}, Minwei Feng\textsuperscript{3}, Cicero Nogueira dos Santos\textsuperscript{4}, Mo Yu\textsuperscript{5}, Bing Xiang\textsuperscript{6}, Bowen Zhou\textsuperscript{5} & Yoshua Bengio\textsuperscript{1,7}

\textsuperscript{1}IBM Watson
\textsuperscript{2}Montreal Institute for Learning Algorithms (MILA), Université de Montréal
\textsuperscript{3}CIFAR Senior Fellow

Published as a conference paper at ICLR 2017

Process to obtain the attention weights: hidden states multiplied by a weight matrix + tanh layer + another weight matrix + a softmax layer.
Natural Language Processing

Self-Attention

2017: Ashish Vaswani's "transformer" (Google Brain):
no RNN, only self-attention

Attention Is All You Need

Vaswani, Shaazer, Uszkoreit
Natural Language Processing

Self-Attention
2017: Google's "transformer"
Natural Language Processing

Self-Attention

2018: Wei Yu's "QANet" (CMU + Google Brain): no RNN, only convolutions (to model local interactions) and self-attention (to model global interactions)
Natural Language Processing

Self-Attention

QANet
Natural Language Processing

Self-Attention

2018: DeepMind's Relational Deep Reinforcement Learning

Relational Deep Reinforcement Learning

Vinicius Zambaldi‡, David Raposo‡, Adam Santoro‡, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, Peter Battaglia

Input

ReLU

Conv. 2 x 2, stride 1

x 4

ReLU

FC 256

Feature-wise max pooling

Relational module

Multi-head dot product attention

query $q_i$

key $k_i$

value $v_i$

$E$

$A$

$E$

$softmax(QK^T/\sqrt{d})W$
Natural Language Processing

Non-local networks (Xiaolong Wang, 2018): general method for sequence processing, not only for NLP!

Non-local Neural Networks

Xiaolong Wang¹  Ross Girshick²  Abhinav Gupta¹  Kaiming He²
¹Carnegie Mellon University  ²Facebook AI Research

Figure 3. Examples of the behavior of a non-local block in res3 computed by a 5-block non-local model trained on Kinetics. These examples
Natural Language Processing

Howard & Ruder: UMLFiT (2018):

Universal Language Model Fine-tuning for Text Classification
Natural Language Processing

Transformer architecture + transfer learning for pre-training:
OpenAI GPT (2018)
Google BERT (2018)

Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks.
Natural Language Processing

Jacob Devlin’s BERT (Google, 2018)
Natural Language Processing

BERT’s descendants:

• Luke Zettlemoyer's RoBERTa (Facebook + Univ of Washington, 2019) GLUE rating: 88.5
• Chen Zhu’s FreeLB (Univ of Maryland + Microsoft, 2019) 88.8
• StructBERT (Alibaba, 2019) 89.0
• Zhenzhong Lan’s AIBERT (Google, 2019) 89.4
### Natural Language Processing

**GLUE scores (2019)**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALBERT-Team Google Language</td>
<td>ALBERT (Ensemble)</td>
<td>89.4</td>
</tr>
<tr>
<td>2</td>
<td>王玮</td>
<td>ALICE v2 large ensemble (Alibaba DAMO NLP)</td>
<td>89.0</td>
</tr>
<tr>
<td>3</td>
<td>Microsoft D365 AI &amp; UMD</td>
<td>FreeLB-RoBERTa (ensemble)</td>
<td>88.8</td>
</tr>
<tr>
<td>4</td>
<td>Facebook AI</td>
<td>RoBERTa</td>
<td>88.5</td>
</tr>
<tr>
<td>5</td>
<td>XLNet Team</td>
<td>XLNet-Large (ensemble)</td>
<td>88.4</td>
</tr>
</tbody>
</table>
Probing Neural Network Comprehension of Natural Language Arguments

Timothy Niven and Hung-Yu Kao
Intelligent Knowledge Management Lab
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, Taiwan

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy,1 Ellie Pavlick,2 & Tal Linzen1
1Department of Cognitive Science, Johns Hopkins University
2Department of Computer Science, Brown University

Natural Language Processing

BERT doubters (2019):
• Timothy Niven and Hung-Yu Kao
  (Taiwan's National Cheng Kung University)
• Tal Linzen (Johns Hopkins University)
Natural Language Processing

Pre-training: training the language representation with several simple tasks in order to grasp the co-occurrence of words or sentences

Pre-training:
• Autoencoding-based (BERT)
• Autoregressive-based (XLNet)
Natural Language Processing

XLNet (Ruslan Salakhutdinov, 2019): Transformer-XL, autoregressive)

2.3 Architecture: Two-Stream Self-Attention for Target-Aware Representations
Figure 1. Architecture composition from encoding. Each block produces a new hidden state that is added to the pool of hidden states subsequent blocks can select as branch inputs. Each encoder has 6 unique blocks per cell and each decoder has 8 unique blocks per cell. Each cell is repeated number of cells times.
Natural Language Processing

ERNIE 2.0 (Hua Wu, Baidu, 2019)

ERINE 2.0: A Continual Pre-Training Framework for Language Understanding
Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang

Figure 4: The architecture of multi-task learning in the ERINE 2.0 framework, in which the encoder can be recurrent neural networks or a deep transformer.
Natural Language Processing

Main pre-trained models in 2019:

- ELMo
- OpenAI's GPT
- Google's BERT
- Microsoft's XLNet
- Baidu's Ernie 2.0

---

**ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding**

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang

Baidu Inc., Beijing, China
Natural Language Processing

Question-answering

Danqi Chen’s DrQA (Facebook, 2017): multitask learning using distant supervision

Figure 1: An overview of our question answering system DrQA.
Natural Language Processing

Question-answering

FlowQA (Allen Institute, 2018)

FLOWQA: GRASPING FLOW IN HISTORY FOR CONVERSATIONAL MACHINE COMPREHENSION

Hsin-Yuan Huang*  
California Institute of Technology  
Eunsol Choi  
University of Washington

Wen-tau Yih  
Allen Institute for Artificial Intelligence

Published as a conference paper at ICLR 2019.

Figure 3: Alternating computational structure between context integration (RNN over context) and FLOW (RNN over question turns).

Figure 4: An illustration of the architecture for FlowQA.
Natural Language Processing

Question-answering
SDNet (Microsoft, 2018)

Figure 1: SDNet model structure.
Natural Language Processing

Question-answering

BiDAF++ (Mark Yatskar, Allen Inst, 2018)
Natural Language Processing

Visual Question-answering
Anton van den Hengel’s team at University of Adelaide (2017)
Zichao Yang at CMU in collaboration with Microsoft (2016)
Jiasen Lu at Virginia Tech (2016), the paper that introduced co-attention
Josh Tenenbaum’s team at MIT, in collaboration with IBM and DeepMind: NS-CL (2019), capable of learning about the world just as a child does: by looking around and talking
Natural Language Processing

Visual Question-answering
Josh Tenenbaum’s NS-CL (2019)

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding

Kevin Yi
Harvard University

Jiajun Wu
MIT CSAIL

Chuang Gan
MIT-IBM Watson AI Lab

Antonio Torralba
MIT CSAIL

Pushmeet Kohli
DeepMind

Joshua B. Tenenbaum
MIT CSAIL

Figure 1: Human reasoning is interpretable and disentangled: we first draw abstract knowledge of the scene via visual perception and then perform logic reasoning on it.

Figure 2: Our model has three components: first, a scene parser (de-renderer) that segments an input image (a-b) and recovers a structural scene representation (c); second, a question parser (program generator) that converts a question in natural language (d) into a program (e); third, a program executor that runs the program on the structural scene representation to obtain the answer.
Natural Language Processing

Unsupervised learning of language use: Word2vec, GloVe, ELMo, Skip-Thoughts…
Supervised learning: InferSent (Alexis Conneau, Antoine Bordes - Facebook)
Natural Language Processing

Supervised learning of language representation:
Sandeep Subramanian  (University of Montreal, 2018)
Natural Language Processing

Supervised learning of language representation:
Daniel Cer’s Universal Sentence Encoder (Google, 2018)
Natural Language Processing

Sentiment Analysis

Figure 5: Examples of intra-attention (sentiment analysis). Bold lines (red) indicate attention between sentiment important words.  

Jianpeng Cheng (2016)
Natural Language Processing

Sentiment Analysis

• Kai Sheng Tai & Richard Socher (2015)
Natural Language Processing

Sentiment Analysis

- Soumith Chintala (2015)
Natural Language Processing

Sentiment Analysis

• 2016: Peter Dodds & Chris Danforth (Univ of Vermont): text-based sentiment analysis
• 2017: Eric Chu & Deb Roy (MIT): visual and audio sentiment analysis
Natural Language Processing

Sentiment Analysis

• 2017: Alec Radford (OpenAI) discovers the “sentiment neuron” in LSTM networks.

• Trained (with 82 million Amazon reviews) to predict the next character in the text of Amazon reviews, the network develops a "sentiment neuron“ that predicts the sentiment value of the review.
Natural Language Processing

Discourse Analysis

2016: Minjoon Seo’s Bidirectional Attention Flow (BiDAF) model

Figure 1: BiDirectional Attention Flow Model (best viewed in color)
Discourse Analysis
2016: Percy Liang's SQuAD dataset
2016: Jianfeng Gao’s MARCO dataset
Natural Language Processing

Discourse Analysis
2016: Weizhu Chen's Reasonet combines memory networks with reinforcement learning
2017: Weizhu Chen's FusionNet simpler attention mechanism called "History of Word"

FUSIONNET: FUSING VIA FULLY-AWARE ATTENTION WITH APPLICATION TO MACHINE COMPREHENSION

Hsin-Yuan Huang*1,2, Chenguang Zhu1, Yelong Shen1, Weizhu Chen1
1Microsoft Business AI and Research

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Single Model</th>
<th>FusionNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM / F1</td>
<td>76.0 / 83.9</td>
</tr>
<tr>
<td>LR Baseline</td>
<td>40.4 / 51.0</td>
<td></td>
</tr>
<tr>
<td>Match-LSTM</td>
<td>64.7 / 73.7</td>
<td></td>
</tr>
<tr>
<td>BiDAF</td>
<td>68.0 / 77.3</td>
<td></td>
</tr>
<tr>
<td>SEDT</td>
<td>68.2 / 77.5</td>
<td></td>
</tr>
<tr>
<td>RaSoR</td>
<td>70.8 / 78.7</td>
<td></td>
</tr>
<tr>
<td>DrQA</td>
<td>70.7 / 79.4</td>
<td></td>
</tr>
<tr>
<td>ResoNet</td>
<td>70.6 / 79.4</td>
<td></td>
</tr>
<tr>
<td>R. Mnemonic Reader</td>
<td>73.2 / 81.8</td>
<td></td>
</tr>
<tr>
<td>DCN+</td>
<td>74.9 / 82.8</td>
<td></td>
</tr>
<tr>
<td>R-net†</td>
<td>75.7 / 83.5</td>
<td></td>
</tr>
<tr>
<td>FusionNet</td>
<td>76.0 / 83.9</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The performance of FusionNet and competing approaches on SQuAD hidden test set at the time of writing (Oct. 4th, 2017).
Discourse Analysis
2017: Quoc Le’s QANET
Natural Language Processing

Discourse Analysis

2018: Furu Wei's R-Net for reading-comprehension
# Natural Language Processing

## Text Comprehension

<table>
<thead>
<tr>
<th>Single Model</th>
<th>Published</th>
<th>LeaderBoard</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR Baseline (Rajpurkar et al., 2016)</td>
<td>40.4 / 51.0</td>
<td>40.4 / 51.0</td>
</tr>
<tr>
<td>Dynamic Chunk Reader (Yu et al., 2016)</td>
<td>62.5 / 71.0</td>
<td>62.5 / 71.0</td>
</tr>
<tr>
<td>Match-LSTM with Ans-Ptr (Wang &amp; Jiang, 2016)</td>
<td>64.7 / 73.7</td>
<td>64.7 / 73.7</td>
</tr>
<tr>
<td>Multi-Perspective Matching (Wang et al., 2016)</td>
<td>65.5 / 75.1</td>
<td>70.4 / 78.8</td>
</tr>
<tr>
<td>Dynamic Coattention Networks (Xiong et al., 2016)</td>
<td>66.2 / 75.9</td>
<td>66.2 / 75.9</td>
</tr>
<tr>
<td>FastQA (Weissenborn et al., 2017)</td>
<td>68.4 / 77.1</td>
<td>68.4 / 77.1</td>
</tr>
<tr>
<td>BiDAF (Seo et al., 2016)</td>
<td>68.0 / 77.3</td>
<td>68.0 / 77.3</td>
</tr>
<tr>
<td>SEDT (Liu et al., 2017a)</td>
<td>68.1 / 77.5</td>
<td>68.5 / 78.0</td>
</tr>
<tr>
<td>RaSoR (Lee et al., 2016)</td>
<td>70.8 / 78.7</td>
<td>69.6 / 77.7</td>
</tr>
<tr>
<td>FastQAExt (Weissenborn et al., 2017)</td>
<td>70.8 / 78.9</td>
<td>70.8 / 78.9</td>
</tr>
<tr>
<td>ReasoNet (Shen et al., 2017b)</td>
<td>70.1 / 79.0</td>
<td>70.4 / 79.4</td>
</tr>
<tr>
<td>Document Reader (Chen et al., 2017)</td>
<td>70.0 / 79.5</td>
<td>70.1 / 79.4</td>
</tr>
<tr>
<td>Ruminating Reader (Gong &amp; Bowman, 2017)</td>
<td>70.6 / 79.5</td>
<td>70.6 / 79.5</td>
</tr>
<tr>
<td>jNet (Zhang et al., 2017)</td>
<td>70.6 / 79.8</td>
<td>70.6 / 79.8</td>
</tr>
<tr>
<td>Conductor-net</td>
<td>N/A</td>
<td>72.6 / 81.4</td>
</tr>
<tr>
<td>Interactive AoA Reader (Cui et al., 2017)</td>
<td>N/A</td>
<td>73.6 / 81.9</td>
</tr>
<tr>
<td>Reg-RaSoR</td>
<td>N/A</td>
<td>75.8 / 83.3</td>
</tr>
<tr>
<td>DCN+</td>
<td>N/A</td>
<td>74.9 / 82.8</td>
</tr>
<tr>
<td>AIR-FusionNet</td>
<td>N/A</td>
<td>76.0 / 83.9</td>
</tr>
<tr>
<td>R-Net (Wang et al., 2017)</td>
<td>72.3 / 80.7</td>
<td>76.5 / 84.3</td>
</tr>
<tr>
<td>BiDAF + Self Attention + ELMo</td>
<td>N/A</td>
<td><strong>77.9 / 85.3</strong></td>
</tr>
<tr>
<td>Reinforced Mnemonic Reader (Hu et al., 2017)</td>
<td>73.2 / 81.8</td>
<td>73.2 / 81.8</td>
</tr>
</tbody>
</table>

**human performance = 82.3**

| Dev set: QANet                                   | 73.6 / 82.7 | N/A |
| Dev set: QANet + data augmentation ×2            | 74.5 / 83.2 | N/A |
| Dev set: QANet + data augmentation ×3            | 75.1 / 83.8 | N/A |
| Test set: QANet + data augmentation ×3           | **76.2 / 84.6** | 76.2 / 84.6 |

The performances of different models on SQuAD dataset.
Natural Language Processing

Summarization

Enabled by

– Sequence-to-sequence (Seq2Seq) that can both read AND write
– Pointer networks (Vinyals & Fortunato, 2015): attention-based seq2seq
Natural Language Processing

Summarization

Pointer networks (Vinyals & Fortunato, 2015)
Natural Language Processing

Summarization

- Alexander Rush & Jason Weston at (Facebook, 2015) first to apply seq2seq to summarization
- Read-Again Summarization (Raquel Urtasun & Wenyuan Zeng, Univ of Toronto, 2016)
- Forced Attention Sentence Compression Model (Phil Blunsom & Yishu Miao at Oxford, 2016)
- IBM’s SummaRunner (Ramesh Nallapati, 2017)
- Pointer-generator Network PGNET (Abigail See & Christopher Manning at Stanford, 2017)
Natural Language Processing

Summarization

Alexander Rush, Sumit Chopra and Jason Weston (Facebook, 2015): ABS
– first to apply seq2seq to summarization
– attention-based summarization
Natural Language Processing

Summarization

Read-Again Summarization (Raquel Urtasun & Wenyuan Zeng, Univ of Toronto, 2016)
Natural Language Processing

Summarization

Forced Attention Sentence Compression Model
(Phil Blunsom & Yishu Miao at Oxford, 2016)
Natural Language Processing

Summarization

IBM’s SummaRunner (Ramesh Nallapati, 2017)

Figure 1: SummaRuNNer: A two-layer RNN based sequence classifier: the bottom layer operates at word level within each sentence, while the top layer runs over sentences.
Natural Language Processing

Summarization

Pointer-generator Network PGNET (Abigail See & Christopher Manning at Stanford, 2017) for longer summaries
Natural Language Processing

Simplification

• Yuta Kikuchi (Japan, 2016): controllable text generation
• Sergiu Nisioi and Sanja Stajner (Romania & Germany, 2017)
• Daiki Nishihara (Japan, 2019)
Natural Language Processing

Simplification

- Eric de la Clergerie (Inria, 2019): ACCESS

Facebook’s AI streamlines sentences while preserving meaning
Natural Language Processing

Text Generation
OpenAI’s GPT2 (2019)

Language Models are Unsupervised Multitask Learners

Alec Radford \textsuperscript{1}  Jeffrey Wu \textsuperscript{1}  Rewon Child \textsuperscript{1}  David Luan \textsuperscript{1}  Dario Amodei \textsuperscript{2}  Ilya Sutskever \textsuperscript{2}  

\textsuperscript{1} OpenAI 

\textsuperscript{2} The Guardian

AI can write just like me.
Brace for the robot apocalypse

Hannah Jane Parkinson
## Natural Language Processing

### Text Generation

**OpenAI’s GTP2 (2019)**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Previous Record</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winograd Schema Challenge</td>
<td>accuracy (+)</td>
<td>70.70%</td>
<td>63.7%</td>
</tr>
<tr>
<td>LAMBADA</td>
<td>accuracy (+)</td>
<td>63.24%</td>
<td>59.23%</td>
</tr>
<tr>
<td>LAMBADA</td>
<td>perplexity (-)</td>
<td>8.6</td>
<td>99</td>
</tr>
<tr>
<td>Children’s Book Test Common Nouns (validation accuracy)</td>
<td>accuracy (+)</td>
<td>93.30%</td>
<td>85.7%</td>
</tr>
<tr>
<td>Children’s Book Test Named Entities</td>
<td>accuracy (+)</td>
<td>89.05%</td>
<td>82.3%</td>
</tr>
</tbody>
</table>
### Natural Language Processing

**Question-answering**  
**OpenAI’s GTP2 (2019)**

<table>
<thead>
<tr>
<th>Question</th>
<th>Generated Answer</th>
<th>Correct</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who wrote the book <em>The Origin of Species</em>?</td>
<td>Charles Darwin</td>
<td>✔</td>
<td>83.4%</td>
</tr>
<tr>
<td>Who is the founder of the Ubuntu project?</td>
<td>Mark Shuttleworth</td>
<td>✔</td>
<td>82.0%</td>
</tr>
<tr>
<td>Who is the quarterback for the Green Bay Packers?</td>
<td>Aaron Rodgers</td>
<td>✔</td>
<td>81.1%</td>
</tr>
<tr>
<td>Panda is a national animal of which country?</td>
<td>China</td>
<td>✔</td>
<td>76.8%</td>
</tr>
<tr>
<td>Who came up with the theory of relativity?</td>
<td>Albert Einstein</td>
<td>✔</td>
<td>76.4%</td>
</tr>
<tr>
<td>When was the first Star Wars film released?</td>
<td>1977</td>
<td>✔</td>
<td>71.4%</td>
</tr>
<tr>
<td>What is the most common blood type in Sweden?</td>
<td>A</td>
<td>×</td>
<td>70.6%</td>
</tr>
<tr>
<td>Who is regarded as the founder of psychoanalysis?</td>
<td>Sigmund Freud</td>
<td>✔</td>
<td>69.3%</td>
</tr>
<tr>
<td>Who took the first steps on the moon in 1969?</td>
<td>Neil Armstrong</td>
<td>✔</td>
<td>66.8%</td>
</tr>
<tr>
<td>Who is the largest supermarket chain in the UK?</td>
<td>Tesco</td>
<td>✔</td>
<td>65.3%</td>
</tr>
<tr>
<td>What is the meaning of shalom in English?</td>
<td>Peace</td>
<td>✔</td>
<td>64.0%</td>
</tr>
<tr>
<td>Who was the author of the art of war?</td>
<td>Sun Tzu</td>
<td>✔</td>
<td>59.6%</td>
</tr>
<tr>
<td>Largest state in the US by land mass?</td>
<td>California</td>
<td>✔</td>
<td>59.2%</td>
</tr>
<tr>
<td>Green algae is an example of which type of reproduction?</td>
<td>Parthenogenesis</td>
<td>×</td>
<td>56.5%</td>
</tr>
<tr>
<td>Vikram sanmat calendar is official in which country?</td>
<td>India</td>
<td>✔</td>
<td>55.6%</td>
</tr>
<tr>
<td>Who is mostly responsible for writing the declaration of independence?</td>
<td>Thomas Jefferson</td>
<td>✔</td>
<td>53.3%</td>
</tr>
<tr>
<td>What us state forms the western boundary of Montana?</td>
<td>Montana</td>
<td>×</td>
<td>52.3%</td>
</tr>
<tr>
<td>Who plays ser davos in game of thrones?</td>
<td>Peter Dinklage</td>
<td>×</td>
<td>52.1%</td>
</tr>
<tr>
<td>Who appoints the chair of the federal reserve system?</td>
<td>Janet Yellen</td>
<td>×</td>
<td>51.5%</td>
</tr>
<tr>
<td>State the process that divides one nucleus into two genetically identical nuclei?</td>
<td>Mitosis</td>
<td>✔</td>
<td>50.7%</td>
</tr>
<tr>
<td>Who won the most MVP awards in the NBA?</td>
<td>Michael Jordan</td>
<td>✔</td>
<td>50.2%</td>
</tr>
<tr>
<td>What river is associated with the city of Rome?</td>
<td>The Tiber</td>
<td>✔</td>
<td>48.6%</td>
</tr>
<tr>
<td>Who is the first president to be impeached?</td>
<td>Andrew Johnson</td>
<td>✔</td>
<td>48.3%</td>
</tr>
<tr>
<td>Who is the head of the department of Homeland Security 2017?</td>
<td>John Kelly</td>
<td>✔</td>
<td>47.0%</td>
</tr>
<tr>
<td>What is the name given to the common currency to the European Union?</td>
<td>Euro</td>
<td>✔</td>
<td>46.8%</td>
</tr>
<tr>
<td>What was the emperor name in star wars?</td>
<td>Palpatine</td>
<td>✔</td>
<td>46.5%</td>
</tr>
<tr>
<td>Do you have to have a gun permit to shoot at a range?</td>
<td>No</td>
<td>✔</td>
<td>46.4%</td>
</tr>
<tr>
<td>Who proposed evolution in 1859 as the basis of biological development?</td>
<td>Charles Darwin</td>
<td>✔</td>
<td>45.7%</td>
</tr>
<tr>
<td>Nuclear power plant that blew up in Russia?</td>
<td>Chernobyl</td>
<td>✔</td>
<td>45.7%</td>
</tr>
<tr>
<td>Who played John Connor in the original Terminator?</td>
<td>Arnold Schwarzenegger</td>
<td>×</td>
<td>45.2%</td>
</tr>
</tbody>
</table>
Natural Language Processing

Language models after BERT and GPT

Number of parameters

Elmo  GPT  BERT  GPT2

AI2  ELMo  94m
OpenAI GPT 110m
BERT-Large 340m
Transformer 465m
MT-DNN 130m
Carnegie Mellon University

MegatronLM 8.3b
T-NLG 17b

GPT 2.0 1.5b
GPT-2 15b

Carnegie Mellon University

ELMo 75m
XLM-340m
XLM-665m
RoBERTa-355m
DistilBERT 66m

NVIDIA

Megatron
Natural Language Processing

Language models after BERT and GPT
Nvidia’s Megatron (2019)

<table>
<thead>
<tr>
<th>Model</th>
<th>SQuAD 1.1</th>
<th>F1 / EM (dev set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa (Liu et al., 2019b)</td>
<td>94.6 / 88.9</td>
<td></td>
</tr>
<tr>
<td>ALBERT (Lan et al., 2019)</td>
<td>94.8 / 89.3</td>
<td></td>
</tr>
<tr>
<td>XLNet (Yang et al., 2019)</td>
<td>95.1 / 89.7</td>
<td></td>
</tr>
<tr>
<td>Megatron-336M</td>
<td>94.2 / 88.0</td>
<td></td>
</tr>
<tr>
<td>Megatron-1.3B</td>
<td>94.9 / 89.1</td>
<td></td>
</tr>
<tr>
<td>Megatron-3.9B</td>
<td><strong>95.5 / 90.0</strong></td>
<td></td>
</tr>
<tr>
<td>ALBERT ensemble (Lan et al., 2019)</td>
<td>95.5 / 90.1</td>
<td></td>
</tr>
<tr>
<td>Megatron-3.9B ensemble</td>
<td><strong>95.8 / 90.5</strong></td>
<td></td>
</tr>
</tbody>
</table>

Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism

Mohammad Shoeybi\(^1\)\(^2\)  Mostofa Patwary\(^1\)\(^2\)  Raul Puri\(^1\)\(^2\)  Patrick LeGresley\(^2\)  Jared Casper\(^2\)
Bryan Catanzaro\(^2\)

Transformer Architecture

(a) MLP

(b) Self-Attention

Purple blocks correspond to fully connected layers
Each blue block represents a single transformer layer that is replicated N times.
## Natural Language Processing

Language models after BERT and GPT

**Google T5 (2020)**

---

**Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**

*Colin Raffel, Google, Mountain View, CA 94043, USA*

---

**Diagram:**

- "translate English to German: That is good."
- "cola sentence: The course is jumping well."
- "stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
- "summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi."

---

**Table:**

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>Metric</th>
<th>Metric Value</th>
<th>Global Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Answering</td>
<td>BoolQ</td>
<td>Accuracy</td>
<td>91.0</td>
<td>1</td>
</tr>
<tr>
<td>Document Summarization</td>
<td>CNN/Daily Mail</td>
<td>ROUGE-2</td>
<td>21.55</td>
<td>1</td>
</tr>
<tr>
<td>Linguistic Acceptability</td>
<td>CoLA</td>
<td>Accuracy</td>
<td>70.8</td>
<td>1</td>
</tr>
<tr>
<td>Semantic Textual Similarity</td>
<td>MRPC</td>
<td>F1</td>
<td>92.4</td>
<td>2</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>SST-2 Binary Classification</td>
<td>Accuracy</td>
<td>97.4</td>
<td>1</td>
</tr>
</tbody>
</table>
Natural Language Processing

Language models after BERT and GPT
Microsoft Turing-NLG (2020)

Project Turing

Turing-NLG: A 17-billion-parameter language model by Microsoft

February 10, 2020

Turing Natural Language Generation (T-NLG) is a 17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks. We present a demo of the model, including its freeform generation, question answering, and summarization capabilities, to academics for feedback and research purposes. - This summary was generated by the Turing-NLG language model itself.
Natural Language Processing

Language models after BERT and GPT
OpenAI’s GTP-3 (2020)

Language Models are Few-Shot Learners

<table>
<thead>
<tr>
<th>Tom B. Brown*</th>
<th>Benjamin Mann*</th>
<th>Nick Ryder*</th>
<th>Melanie Subbiah*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jared Kaplan†</td>
<td>Prafulla Dharwal</td>
<td>Arvind Neelakantan</td>
<td>Franav Shyam</td>
</tr>
<tr>
<td>Amanda Askell</td>
<td>Sandhini Agarwal</td>
<td>Ariel Herbert-Voss</td>
<td>Girish Sastry</td>
</tr>
<tr>
<td>Rewan Child</td>
<td>Aditya Ramesh</td>
<td>Daniel M. Ziegler</td>
<td>Thomas Henighan</td>
</tr>
<tr>
<td>Christopher Hesse</td>
<td>Mark Chen</td>
<td>Eric Sigler</td>
<td>Mateusz Litwin</td>
</tr>
<tr>
<td>Benjamin Chess</td>
<td>Jack Clark</td>
<td>Christopher Berner</td>
<td>Scott Gray</td>
</tr>
<tr>
<td>Sam McCandlish</td>
<td>Alec Radford</td>
<td>Ilya Sutskever</td>
<td>Dario Amodei</td>
</tr>
</tbody>
</table>

Microsoft teams up with OpenAI to exclusively license GPT-3 language model

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination

Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed those measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.,” but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 30 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).
A robot wrote this entire article. Are you scared yet, human?  

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.
This article was written by GPT-3, OpenAI’s language generator. GPT-3 is a cutting edge language model that uses machine learning to produce human-like text. It takes in a prompt, and attempts to complete it. For this essay, GPT-3 was given these instructions: “Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI.” It was also fed the following introduction: “I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could “spell the end of the human race.” I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me.” The prompts were written by the Guardian, and fed to GPT-3 by Liam Porr, a computer science undergraduate student at UC Berkeley. GPT-3 produced eight different outputs, or essays. Each was unique, interesting and advanced a different argument. The Guardian could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI. Editing GPT-3’s op-ed was no different to editing a human op-ed. We cut lines and paragraphs, and rearranged the order of them in some places. Overall, it took less time to edit than many human op-eds.
Natural Language Processing

Language models after BERT and GPT
OpenAI’s GTP-3 (2020)

The New York Times

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

MIT Technology Review

GPT-3, Bloviator: OpenAI’s language generator has no idea what it’s talking about

Tests show that the popular AI still has a poor grasp of reality.

by Gary Marcus and Ernest Davis

August 22, 2020
Natural Language Processing

The Hype of the 2010s…
2010s

• Conversational computing
  – Siri (2011)
  – GoogleNow (2012)
  – Amazon Alexa (2014)
  – Microsoft XiaoIce (2014)
  – Microsoft Tay (2016)

Stanley Kubrick (1968)
“2001: A Space Odyssey”

(mandatory Hollywood movie for AI presentation!)
Chatbots

- Joseph Weintraub's PC Therapist (1986)
- Michail Mauldin's Julia (1994)
- Richard Wallace's ALICE (Artificial Linguistic Internet Computer Entity, 1995)
- Rollo Carpenter’s Jabberwacky (1997)
- Robby Garner's Albert One (1998)
- ActiveBuddy's SmarterChild, the first commercial chatbot, used by millions of people (2000)
- Bruce Wilcox's Suzette (2009)
- Steve Worswick's Mitsuku (2013)
Chatbots

• The "human" chatbots made by Mark Sagar, a former Hollywood animation engineer, starting with "Baby X" (2014)
• The “memorial” chatbot Replika (2016), that learns a person’s style of chat and replicates it even when the person is dead
Chatbots

- Therapist Woebot (Alison Darcy, 2017)

Hi, I'm Woebot!

I'm ready to listen, 24/7. No couches, no meds, no childhood stuff. Just strategies to improve your mood. And the occasional dorky joke.
Chatbots

The year of the full-duplex chatbot (a chatbot that can talk and listen at the same time)

April 2018: Microsoft full-duplex Xiaoice (Li Zhou) and then acquired Semantic Machines

May 2018: Google Duplex (Yaniv Leviathan)
Platforms

- Open-source platforms for NLP
  - Speaktoit/API.ai (Ilya Gelfenbeyn, 2014, acquired by Google in 2016)
  - Wit.ai (Alexandre Lebrun, acquired by Facebook in 2015)
  - Language Understanding Intelligent Service or LUIS (Microsoft, 2015)
  - Amazon Lex (2017)
  - Facebook: FastText for text representation and classification (pre-trained models of word vectors for over 150 languages)
Platforms

• Open-source platforms for chatbots
  – Scripting languages: Artificial Intelligence Markup Language or AIMA (Richard Wallace, 1995) and ChatScript (Bruce Wilcox, 2011)
Platforms

- Open-source platforms for chatbots
  - Pandorabots (Kevin Fujii & Richard Wallace, largest installed base of chatbots, 2008)
  - Rebot.me (Ferid Movsumov and Salih Pehlivan, 2014)
  - Imperson (Disney Accelerator, 2015)
  - ParlAI (Facebook, 2017)
Summarization

- Analysis and summary of text
  - Narrative Science (Chicago, 2010 - Kristian Hammond and Larry Birnbaum)
  - Semantic Machines (Berkeley, 2014 – Dan Roth, Dan Klein, Larry Gillick – acquired in 2018 by Microsoft)
  - Maluuba (Canada, 2011, Sam Pasupalak and Kaheer Suleman – acquired in 2017 by Microsoft)
The State of NLP in 2019

Lead-3 = first three sentences of the document

Table 1: Model and human performance (% in F1 score) on the CoQA test set
Speech Recognition

DARPA Challenges

more and more complicated challenges
Speech Recognition

HMM-based speech recognition

- Bell Labs "mixture-density HMM” (1985)
- CMU’s Sphinx (1988)
- BBN’s Byblos (1989)
- SRI’s Decipher (1989)

Speech recognition datasets

- CSR corpus
- Switchboard corpus

DARPA’s ATIS (1989-94): speech recognition for air travel (ATIS): BBN, MIT, CMU, AT&T, SRI, etc.
Speech Recognition

1994: Nuance (future Apple Siri)
1995: Voice Signal Technologies (future Semantic Machines, Microsoft)
2000: MIT’s Pegasus for airline flights status and Jupiter for weather status/forecast
2000: AT&T’s How May I Help You (HMIHY) for telephone customer care
Speech Recognition

Hybrid HMM-DNN

- Hinton (2009): using a DNN for acoustic modeling (plus an HMM for modeling the sequence of speech)
- Microsoft (2011)
Speech Recognition

Apple's Siri (2011)
Google's Now (2012)
Microsoft's Cortana (2013)
Wit.ai (acquired by Facebook in 2015)
Amazon's Alexa (2014)
SoundHound's Hound (2016)
Speech Recognition

Removing the HMM

2014: Alex Graves' CTC/LSTM without HMM for speech recognition (but high error rate)
Speech Recognition

Removing the HMM
2014: Andrew Ng's CTC/GRU without HMM (low error rate)
2015: LibriSpeech corpus
2015: Baidu's Deep Speech 2
Speech Recognition

Baidu's Deep Speech 2

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*
Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Eric Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

*Dec 2015
Speech Recognition

2016: Microsoft achieves human parity
  – Three kinds of convolutional nets for acoustic modeling
    • VGG
    • ResNet
    • LACE (layer-wise context expansion with attention)
  – An LSTM for language modeling
Speech Synthesis

1940: Homer Dudley’s vocoder (Britain)
1961: Louis Gerstman and Max Mathews program a computer to sing a song (Bell Labs)
1966: Ryunen Teranishi and Noriko Umeda’s text-to-speech system (Japan)
1972: Cecil Coker’s talking computer (Bell Labs)
Speech Synthesis

Trivia (Coker’s ventures into electronic music):
John Cage: Variation II (1966)
Bell Labs’ 7”, 33 ⅓ RPM record “Synthetic Voices For Computers” (1970)
Speech Synthesis

1979: Dennis Klatt: MITalk (MIT)
1988: Francis Charpentier‘s concatenative speech synthesis (France)
1995: Keiichi Tokuda ‘s HMM-based HTS (Japan)
1996: Alan Black’s concatenative text-to-speech (Japan)
Speech Synthesis

- Voice morphing: Festvox (Alan Black, 1997)
Next...

• See http://www.scaruffi.com/singular for the index of this Powerpoint presentation and links to the other parts.