An introduction to Natural Language Processing

Pedro Balage
Pedro Balage

- Lead Data Scientist @ Farfetch - Search team
- Research and Industry experience in NLP
- Affiliated to Instituto de Telecomunicações
- Member of organization for the Lisbon Machine Learning School (LxMLS)
Why Natural Language Processing?

- Natural Language Processing (NLP) is a subfield of Artificial Intelligence that is focused on enabling computers to **understand** and **process human languages**, to get computers closer to a **human-level understanding of language**.
Why Natural Language Processing?

“The next big step for Deep Learning is natural language understanding, which aims to give machines the power to understand not just individual words but entire sentences and paragraphs.”

Yann LeCun, June 2015

https://www.kaggle.com/surveys/2017
Why should I consider NLP for my career?

Source: Forbes - 2017 AI Index Report
Why should I consider NLP for my career?

Source: Forbes - 2017 AI Index Report
The development of NLP has a very fast pace!

Source: Forbes - 2017 AI Index Report
What can I do with NLP?
Classification

- Text classification
- Spam classification
- Sentiment analysis
Sequence Labeling

- Part-of-Speech Tagging
- Chunking
- Named-Entity Recognition
Parsing

- Syntactic parsing
- Semantic parsing
Summarization

Summaries can be:

- Extractive
- Compressive
- Abstractive

The bottleneck is no longer access to information, now it’s our ability to keep up. AI can be trained on a variety of different types of texts and summary lengths. A model that can generate long, coherent, and meaningful summaries remains an open research problem.

The last few decades have witnessed a fundamental change in the challenge of taking in new information. The bottleneck is no longer access to information, now it’s our ability to keep up. We all have to read more and more to keep up-to-date with our jobs, the news, and social media. We’ve looked at how AI can improve people’s work by helping with this information deluge and one potential answer is to have algorithms automatically summarize longer texts. Training a model that can generate long, coherent, and meaningful summaries remains an open research problem. In fact, generating any kind of longer text is hard for even the most advanced deep learning algorithms. In order to make summarization successful, we introduce two separate improvements: a more contextual word generation model and a new way of training summarization models via reinforcement learning (RL). The combination of the two training methods enables the system to create relevant and highly readable multi-sentence summaries of long text, such as news articles, significantly improving on previous results. Our algorithm can be trained on a variety of different types of texts and summary lengths. In this blog post, we present the main contributions of our model and an overview of the natural language challenges specific to text summarization.
Machine Translation

I WANT FORTY KILOGRAMS OF PERSIMMONS

ICH WILL VIERZIG KILOGRAMM PERSIMONEN
Question Answering

- Question Answering
- Conversational Agents (Chatbots)
Why deep learning approaches to NLP?
A BRIEF HISTORY OF MACHINE TRANSLATION

**RBMT**
- Rule-Based Machine Translation
  - Direct Machine Translation
  - Transfer-Based RBMT
  - Interlingua Machine Translation

**EBMT**
- Example-Based Machine Translation

**SMT**
- Statistical Machine Translation
  - Word-Based
  - Syntax-Based
  - Phrase-Based SMT

**NMT**
- Neural Machine Translation
  - RNN
  - LSTM

Timeline:
- 1950
- 1966 - ALPAC Report
- 1968 - SYSTRAN
- 1980
- 1990
- 2007 - Google Translate
- 2015
Neural approaches are now state-of-the-art

![Graph showing BLEU scores over years for different translation methods: Phrase-based SMT, Syntax-based SMT, and Neural MT. The scores for each year are as follows: 2013: 20.3, 2014: 20.9, 2015: 20.8, 2016: 21.5, 2017: 22.1, and 2018: 24.7. The graph indicates an increasing trend in BLEU scores for Neural MT.]
Industry already adopted deep learning.

Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey,
Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser,
Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens,
George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa,
Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Abstract

Neural Machine Translation (NMT) is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems. Unfortunately, NMT systems are known to be computationally expensive both in training and in translation inference – sometimes prohibitively so in the case of very large data sets and large models. Several authors have also charged that NMT systems lack robustness, particularly when input sentences contain rare words. These issues have hindered NMT’s use in practical deployments and services, where both accuracy and
Traditional Machine Learning

- Representation
  - Representation of my data in a feature space
- Hypothesis Model
  - Machine Learning algorithm to split the space

- What do I want to do?
  - Regression, Classification, Clustering
- Do I have data?
  - Supervised, Unsupervised, Semi-supervised
Deep Learning vs Traditional Machine Learning

- **Traditional Machine Learning (TML)**
  - Focus on feature engineering

- **Deep Learning (DL)**
  - Focus on automatic learning word representations

### Representations of Language

<table>
<thead>
<tr>
<th>Element</th>
<th>TML</th>
<th>DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>All phonemes</td>
<td>Vector</td>
</tr>
<tr>
<td>Morphology</td>
<td>All morphemes</td>
<td>Vector</td>
</tr>
<tr>
<td>Words</td>
<td>One-hot encoding</td>
<td>Vector</td>
</tr>
<tr>
<td>Syntax</td>
<td>Phrase rules</td>
<td>Vector</td>
</tr>
<tr>
<td>Semantics</td>
<td>Lambda calculus</td>
<td>Vector</td>
</tr>
</tbody>
</table>
Representation Learning
Words
How to represent my words?

- Local representations
- Problems with this representation?
  - Sparsity
  - Vectors don't capture similarity properties.
How to represent my words?

- **Local representations**
  - Problems with this representation?
    - Sparsity
    - Vectors don't capture similarity properties.

- **Distributed representations (embeddings)**
  - Advantages of this representation?
    - More compact vectors
    - Capable of capturing similarities
Word2Vec (2013)

NLP is simply awesome

what I want to predict

CBOVW

Skip-gram
Word2Vec (2013)

- Final result is an embedding matrix that represents N words
- Is able to capture semantic relations between words
- Out of vocabulary (OOV) problem: How do I represent words that I haven’t seen during training?
Embedding representations

Male-Female

Verb tense

Country-Capital

- Spain
- Italy
- Germany
- Turkey
- Russia
- Canada
- Japan
- Vietnam
- China
- Madrid
- Rome
- Berlin
- Ankara
- Moscow
- Ottawa
- Tokyo
- Hanoi
- Beijing
Other types of representations

- Character embeddings
- Subword embeddings
NLP Libraries
NLP Python Libraries

NLTK

spaCy

GENSIM
Advanced Neural Architectures
Recurrent Neural Networks

\[ h_t \]

\[ x_t \]

\[ h_0 \]

\[ h_1 \]

\[ h_2 \]

\[ x_0 \]

\[ x_1 \]

\[ x_2 \]

\[ \ldots \]

\[ x_t \]
Recurrent Neural Networks

Detailed comparison:
https://www.slideshare.net/YanKang/rnn-explore-71268007
Language Modeling

I thought I would arrive on time, but ended up 5 minutes ____.

Language modeling -- a “fill in the blank”-style next word prediction objective which allows models to learn generic sequence representations that generalize well to new tasks.
Seq2seq

How are you?
Advanced topics
Transformer

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research niki@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez† University of Toronto aidan0cs.toronto.edu
Lukasz Kaiser* Google Brain lukasz.kaiser@google.com
Illia Polosukhin† illia.polosukhin@gmail.com
BERT

1 - Semi-supervised training on large amounts of text (books, wikipedia, etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

2 - Supervised training on a specific task with a labeled dataset.

**Semi-supervised Learning Step**

**Model:**

BERT

**Dataset:**

Predict the masked word (language modeling)

---

**Supervised Learning Step**

**Model:**

BERT (pre-trained in step #1)

**Classifier**

- 75% Spam
- 25% Not Spam

**Dataset:**

<table>
<thead>
<tr>
<th>Email message</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buy these pills</td>
<td>Spam</td>
</tr>
<tr>
<td>Win cash prizes</td>
<td>Spam</td>
</tr>
<tr>
<td>Dear Mr. Altreides, please find attached…</td>
<td>Not Spam</td>
</tr>
</tbody>
</table>
Additional Pointers
Additional Pointers

- https://www.coursera.org/learn/language-processing