Deep Learning for Natural Language Understanding

GU JIATAO

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING

THE UNIVERSITY OF HONG KONG

JIATAOGU@EEE.HKU.HK

Natural Language Understanding (NLU)

Understanding human language is difficult to represent:

- Syntax
- Semantics
- Pragmatics

Applications:

- Topic Modelling
- Information Extraction
- Machine Translation
- Text Summarization
- Dialogue Systems
- …

Representation Learning for NLU

Representation Learning:

- is concerned with questions surrounding how we can best **learn** meaningful and useful representations of data.
- The performance of machine learning methods is heavily dependent on the choice of data representation on which they are applied.

Deep Learning:

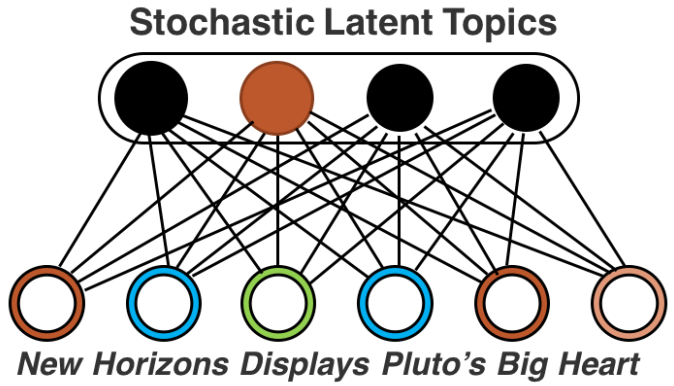
- Use deep learning (deep neural network) methods to generate proper data representation
- Typically, **vectorial/distributed representation**.

Undirected Topic Models

Modelling a document using a Restrict Boltzmann Machine (RBM).

- **Traditional Learning Algorithm:**
  - Contrastive Divergence (CD):
    - slow and unstable

- **Our methods:**
  - Noise Contrastive Estimation
    - Fast but cannot directly applied to documents.
  - **α-NCE:**
    - **modifying the traditional NCE and applicable to natural language**
    - **much faster than CD.**
    - **performance comparable**



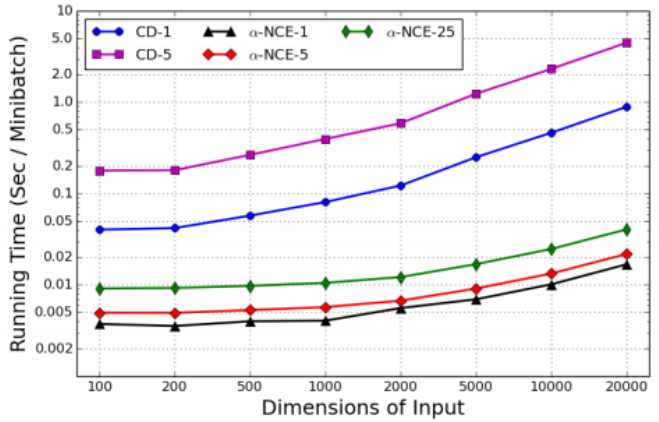


Figure 1: Comparison of running time

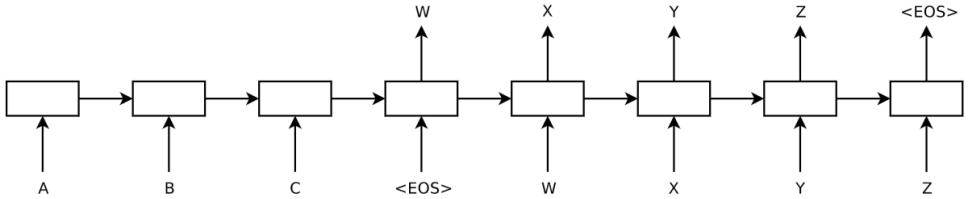
This work is accepted as a short paper in ACL-IJCAI 2015

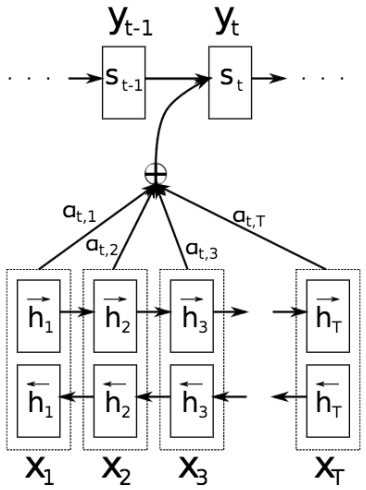
Sequence-to-Sequence Learning

Different from Topic Modelling, this part focus more on word orders and has more applications on real-life, such as machine translation.

The core idea is "**Sequence-to-sequence Learning**":

- We have:
  - Source sequence
  - Target sequence
- What we do:
  - Model **the transformation** using a **Encoder** and **Decoder**
  - With additional techniques such as **Attention Mechanism**
- **What we can do more:**
  - **Look deep into the Seq2Seq learning framework.**





Incorporating Transformation Rules

Basic Idea:

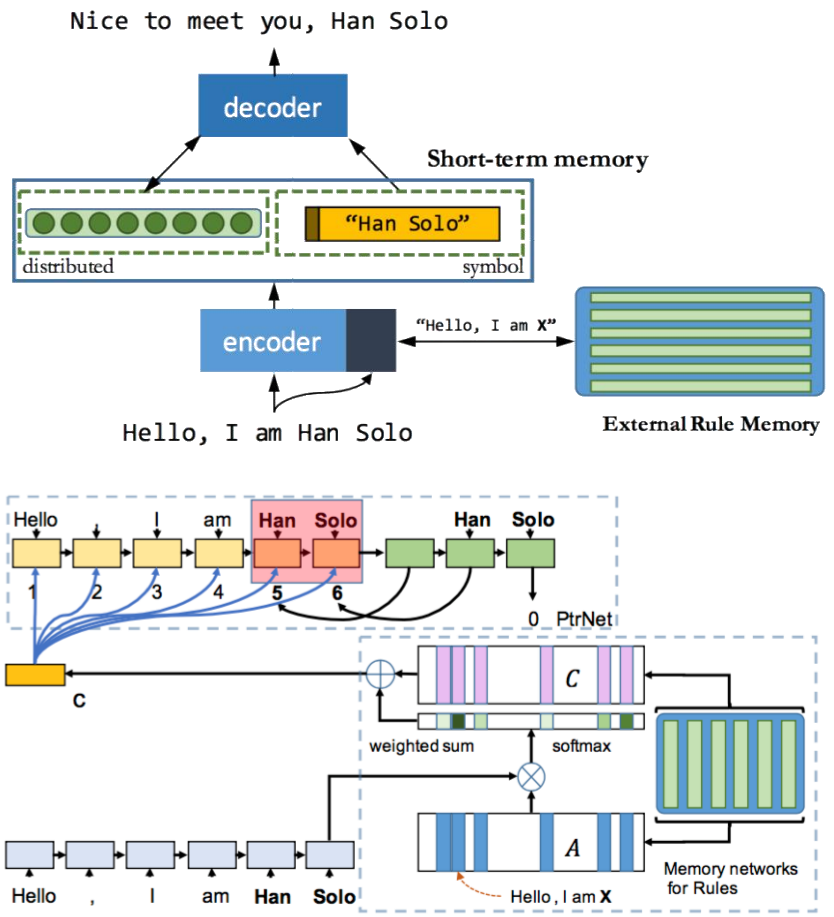
- Incorporating rules into Seq2Seq Learning

Useful in dialogue system:

- A simple dialogue system can be seen as a **Seq2Seq Learning framework**

- Transformation Rules:
  - ***Hello, I am X -> Nice to meet you, X***

- Combine existing transformation rules into a neural network based dialogue system.



This work is accepted as a workshop paper in ICLR 2016

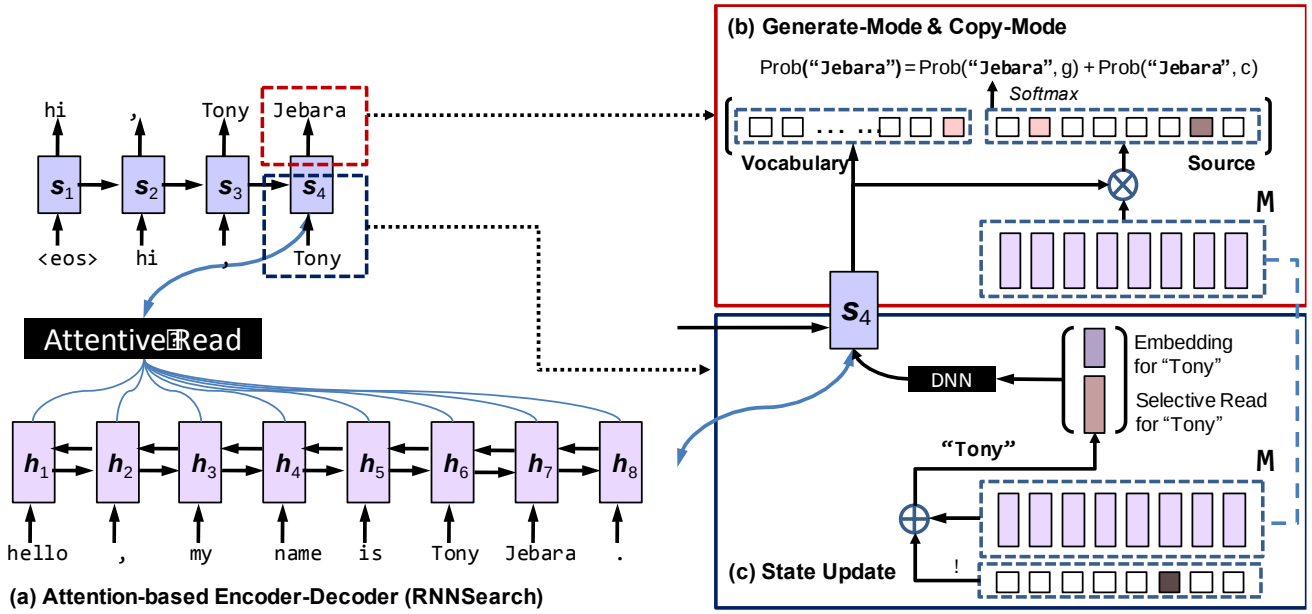
Incorporating Copying Mechanism

Basic Idea:

- Incorporating a special Copying Mechanism in traditional Seq2Seq learning framework.

Useful in **text summarization, machine translation and dialogue system**:

- Sometimes we not only need generate words, but also need copy from the source.
- **A hybrid addressing mechanism of both context and location.**



This work is accepted as a long paper in ACL 2016

Incorporating Copying Mechanism

Input(3): 今天上午 9 点半， 复旦 投毒案 将 在 上海 二中院 公开审理 。 被害 学生 黄洋 的 亲属 已 从 四川 抵达 上海， 其父 称待 刑事 部分 结束 后， 再 提 民事 赔偿， 黄洋 92 岁 的 奶奶 依然 不知情 。 今年 4 月， 在 复旦 上海医学院 读 研究生 的 黄洋 疑遭 室友 林森 浩 投毒， 不幸身亡 。 新民 网

Today 9:30, the Fudan poisoning case will be will on public trial at the Shanghai Second Intermediate Court. The relatives of the murdered student Huang Yang has arrived at Shanghai from Sichuan. His father said that they will start the lawsuit for civil compensation after the criminal section. HuangYang 92-year-old grandmother is still unaware of his death. In April, a graduate student at Fudan University Shanghai Medical College, Huang Yang is allegedly poisoned and killed by his roommate Lin Senhao. Reported by Xinmin

---

Golden: 林森 浩 投毒案 今日 开审 92 岁 奶奶 尚不知 情 (the case of Lin Senhao poisoning is on trial today, his 92-year-old grandmother is still unaware of this)

RNN context: 复旦投毒案： 黄洋疑遭室友投毒凶手已从四川飞往上海，父亲命案另有4人被通知家属不治？

CopyNet: 复旦 投毒案 今 在 沪 上 公开审理 (the Fudan poisoning case is on public trial today in Shanghai)

Input(4): 华谊 兄弟 （ 300027 ） 在 昨日 收盘 后 发布公告 称， 公司 拟 以 自有 资金 3.978 亿元 收购 浙江 永乐 影视 股份 有限公司 若干 股东 持有 的 永乐 影视 51 % 的 股权 。 对于 此项 收购， 华谊 兄弟 董秘 胡明 昨日 表示： “ 和 永乐 影视 的 合并 是 对 华谊 兄弟 电视剧 业务 的 一个 加强 。

Huayi Brothers (300027) announced that the company intends to buy with its own fund 397.8 million 51% of Zhejiang Yongle Film LTD's stake owned by a number of shareholders of Yongle Film LTD. For this acquisition, the secretary of the board, Hu Ming, said yesterday: "the merging with Yongle Film is to strengthen Huayi Brothers on TV business".

Golden: 华谊 兄弟 拟 收购 永乐 影视 5 1 % 股权 (Huayi Brothers intends to acquire 51% stake of Zhejiang Yongle Film)

RNN context: 华谊兄弟收购永乐影视51%股权：与永乐影视合并为“和唐”影视合并的“UNK”和“UNK”的区别？

CopyNet: 华谊 兄弟 拟 3.978 亿 收购 永乐 影视 董秘 称 加强 电视剧 业务 (Huayi Brothers is intended to 397.8 million acquisition of Yongle Film secretaries called to strengthen the TV business)

Input(7): 工厂， 大门 紧锁， 约 20 名 工人 散 坐在 树荫下 。 “ 我们 就是 普通工人， 在 这里 等 工资 。 ” 其中 一人 说道 。 7 月 4 日 上午， 记者 抵达 深圳 龙华区 清 湖 路上 的 深圳 愿景 光电子 有限公司 。 正如 传言 一般， 愿景 光电子 倒闭 了， 大 股东 邢毅 不知 所踪 。

The factory’s door is locked. About 20 works are scattered to sit under the shade. “We are ordinary workers, we are waiting for our salary here.” one of them said. In the morning of July 4th, reporters arrived at Shenzhen Yuanjing Photoelectron Corporation located at Qinghu Road, Longhua District, Shenzhen. Just as the rumor says, Yuanjing Photoelectron Corporation is closed down and the large shareholder Xing Yi is missing.

Golden: 深圳 亿元 级 LED 企业倒闭 烈日 下 工人 苦 等 老板 (Hundred-million CNY worth LED enterprise is closed down and workers wait for the boss under the scorching sun)

RNN context: 深圳 “<UNK>”：深圳<UNK><UNK>，<UNK>，<UNK>，<UNK>

CopyNet: 愿景 光电子 倒闭 20 名 工人 散 坐在 树荫下 (Yuanjing Photoelectron Corporation is closed down, 20 works are scattered to sit under the shade)