Big Data Mining

ABC:
AI, Big Data, Cloud Computing

Min-Yuh Day, Ph.D.
Assistant Professor
Department of Information Management
Tamkang University

http://mail.tku.edu.tw/myday

2018-09-17
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Subject/Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2018/09/10</td>
<td>Course Orientation for Big Data Mining</td>
</tr>
<tr>
<td>2</td>
<td>2018/09/17</td>
<td>ABC: AI, Big Data, Cloud Computing</td>
</tr>
<tr>
<td>3</td>
<td>2018/09/24</td>
<td>Mid-Autumn Festival (Day off)</td>
</tr>
<tr>
<td>4</td>
<td>2018/10/01</td>
<td>Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data</td>
</tr>
<tr>
<td>5</td>
<td>2018/10/08</td>
<td>Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem</td>
</tr>
<tr>
<td>6</td>
<td>2018/10/15</td>
<td>Foundations of Big Data Mining in Python</td>
</tr>
<tr>
<td>7</td>
<td>2018/10/22</td>
<td>Supervised Learning: Classification and Prediction</td>
</tr>
<tr>
<td>8</td>
<td>2018/10/29</td>
<td>Unsupervised Learning: Cluster Analysis</td>
</tr>
<tr>
<td>9</td>
<td>2018/11/05</td>
<td>Unsupervised Learning: Association Analysis</td>
</tr>
<tr>
<td>Week</td>
<td>Date</td>
<td>Subject/Topics</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>2018/11/12</td>
<td>Midterm Project Report</td>
</tr>
<tr>
<td>11</td>
<td>2018/11/19</td>
<td>Machine Learning with Scikit-Learn in Python</td>
</tr>
<tr>
<td>12</td>
<td>2018/11/26</td>
<td>Deep Learning for Finance Big Data with TensorFlow</td>
</tr>
<tr>
<td>13</td>
<td>2018/12/03</td>
<td>Convolutional Neural Networks (CNN)</td>
</tr>
<tr>
<td>14</td>
<td>2018/12/10</td>
<td>Recurrent Neural Networks (RNN)</td>
</tr>
<tr>
<td>15</td>
<td>2018/12/17</td>
<td>Reinforcement Learning (RL)</td>
</tr>
<tr>
<td>16</td>
<td>2018/12/24</td>
<td>Social Network Analysis (SNA)</td>
</tr>
<tr>
<td>17</td>
<td>2018/12/31</td>
<td>Bridge Holiday (Extra Day Off)</td>
</tr>
<tr>
<td>18</td>
<td>2019/01/07</td>
<td>Final Project Presentation</td>
</tr>
</tbody>
</table>
ABC: AI, Big Data, Cloud Computing
Outline

• AI
• Big Data
• Cloud Computing

Source: https://www.amazon.com/Business-Intelligence-Analytics-Data-Science/dp/0134633288
Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners,
Jared Dean,


Source: https://www.amazon.com/Network-Analysis-Applications-Lecture-Networks/dp/3319781952
Evolution of Computerized Decision Support to Analytics/Data Science

The timeline in Figure 1.8 shows the terminology used to describe analytics since the 1970s. During the 1970s, the primary focus of information systems support for decision making focused on providing structured, periodic reports that a manager could use for decision making (or ignore them). Businesses began to create routine reports to inform decision makers (managers) about what had happened in the previous period (e.g., day, week, month, quarter). Although it was useful to know what had happened in the past, managers needed more than this: They needed a variety of reports at different levels of granularity to better understand and address changing needs and challenges of the business. These were usually called management information systems (MIS). In the early 1970s, Scott-Morton first articulated the major concepts of DSS. He defined DSSs as "interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems" (Gorry and Scott-Morton, 1971). The following is another classic DSS definition, provided by Keen and Scott-Morton (1978):

"Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems."

Note that the term decision support system, like management information system and several other terms in the field of IT, is a content-free expression (i.e., it means different things to different people). Therefore, there is no universally accepted definition of DSS.

During the early days of analytics, data was often obtained from the domain experts using manual processes (i.e., interviews and surveys) to build mathematical or knowledge-based models to solve constrained optimization problems. The idea was to do the best with limited resources. Such decision support models were typically called operations research (OR). The problems that were too complex to solve optimally (using linear or nonlinear mathematical programming techniques) were tackled using heuristic methods such as simulation models. (We will introduce these as prescriptive analytics later in this chapter and in a bit more detail in Chapter 6.)

In the late 1970s and early 1980s, in addition to the mature OR models that were being used in many industries and government systems, a new and exciting line of models had emerged: rule-based expert systems. These systems promised to capture experts' knowledge in a format that computers could process (via a collection of if–then–else rules or heuristics) so that these could be used for consultation much the same way that one...
Value Creation by Big Data Analytics
(Grover et al., 2018)

Value Manifestation
- BDA Infrastructure
  - Big Data Asset
  - Analytics Portfolio
  - Human Talent
- BDA Capabilities
  - Ability to integrate, disseminate, explore, and analyze big data
- Value Creation Mechanisms
  - Transparency and access
  - Discovery and experimentation
  - Prediction and optimization
  - Customization and targeting
  - Learning and crowdsourcing
  - Continuous monitoring and proactive adaptation
- Direct value from BDA
- Value Targets
  - Organization Performance
  - Business Processes Improvement
  - Products & Services Innovation
  - Consumer Experience & Market Enhancement
- Investments --- Assets -------- Capabilities --------- Applications ------------------ Targets ------- Impacts ------------- Value

Learning by Doing (Coevolutionary Adaptation)

Research Landscape of Business Intelligence and Big Data Analytics: A bibliometrics study

- A bibliometric analysis on Big Data and Business Intelligence from 1990 to 2016.
- Big Data papers grow much faster than Business Intelligence papers.
- Computer Science and information systems are two core disciplines.
- Most influential papers are identified and a research framework is proposed.

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study", Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10
Evolution of top keywords in “BD & BI” publications

- 2014
  - Management
  - Text Mining
  - Data Mining
  - Data Science

- 2015
  - Big Data Analytics
  - Social Media
  - Business Analytics
  - Information System

- 2016
  - Cloud Computing
  - Data Warehouse

- 2017
  - Knowledge Management

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study", Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10
Framework for BD and BI Research

Technology
- Data Collection
- Data Storage
- Data Analytics
- Infrastructure

Application
- Business
- Medicate
- Supply Chain
- Engineering
- Services

Impact
- Value Creation
- Individual Impact
- Organizational Impact
- Social Impact

Management
- Adoption of BD/BI
- Cost Benefit
- Security/Privacy
- Human Resource

Business Intelligence and Big Data analytics

Definition of Artificial Intelligence (A.I.)
Artificial Intelligence

“... the science and engineering of making intelligent machines”

(John McCarthy, 1955)

Artificial Intelligence

“... technology that thinks and acts like humans”
Artificial Intelligence

“... intelligence exhibited by machines or software”
## 4 Approaches of AI

<table>
<thead>
<tr>
<th>Thinking Humanly</th>
<th>Thinking Rationally</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acting Humanly</td>
<td>Acting Rationally</td>
</tr>
</tbody>
</table>

## 4 Approaches of AI

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AI Acting Humanly: The Turing Test Approach (Alan Turing, 1950)

- Natural Language Processing (NLP)
- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
- Computer Vision
- Robotics

Boston Dynamics: Atlas

#13 ON TRENDING
What's new, Atlas?

https://www.youtube.com/watch?v=fRj34o4hN4l
Humanoid Robot: Sophia

https://www.youtube.com/watch?v=S5t6K9iwcdw
Can a robot pass a university entrance exam?
Noriko Arai at TED2017

https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam
https://www.youtube.com/watch?v=XQZjkPyJ8KU
Artificial Intelligence (A.I.) Timeline

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence.

1955
A.I. BORN
Term ‘artificial intelligence’ is coined by computer scientist John McCarthy to describe “the science and engineering of making intelligent machines”.

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line.

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans.

1966
SHAKEY
The ‘first electronic person’ from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions.

1977
A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold.

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov.

1998
KISMET
Cynthia Breazeal at MIT introduces Kismet, an emotionally intelligent robot insofar as it detects and responds to people’s feelings.

1999
AIBO
Sony launches first consumer robot pet dog AIBO (AI robot) with skills and personality that develop over time.

2002
ROOMBA
First mass-produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes.

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S.

2011
WATSON
IBM’s question answering computer Watson wins first place on popular $1M prize television quiz show Jeopardy.

2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human.

2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks.

2016
TAY
Microsoft’s chatbot Tay goes rogue on social media making inflammatory and offensive racist comments.

2017
ALPHAGO
Google’s A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{120}) of possible positions.

Artificial Intelligence

Machine Learning & Deep Learning

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Artificial Intelligence (AI)

Machine Learning (ML)

- Supervised Learning
- Unsupervised Learning
- Deep Learning (DL)
  - CNN
  - RNN
  - LSTM
  - GRU
  - GAN
- Semi-supervised Learning
- Reinforcement Learning

Source: https://leonardoaraujosantos.gitbooks.io/artificial-intelligence/content/deep_learning.html
Artificial Intelligence (AI) is many things

Ecosystem of AI

Source: https://www.i-scoop.eu/artificial-intelligence-cognitive-computing/
Artificial Intelligence (AI)
Intelligent Document Recognition algorithms

Source: https://www.i-scoop.eu/artificial-intelligence-cognitive-computing/
Deep Learning Evolution

Source: http://www.erogol.com/brief-history-machine-learning/
Machine Learning Models

- Deep Learning
- Association rules
- Decision tree
- Clustering
- Bayesian
- Kernel
- Ensemble
- Dimensionality reduction
- Regression Analysis
- Instance based

Source: Sunila Gollapudi (2016), Practical Machine Learning, Packt Publishing
3 Machine Learning Algorithms

Machine Learning (ML) / Deep Learning (DL)

Supervised Learning:
- Decision Tree Classifiers
- Linear Classifiers
- Rule-based Classifiers
- Probabilistic Classifiers
  - Support Vector Machine (SVM)
  - Neural Network (NN)
  - Deep Learning (DL)
  - Naïve Bayes (NB)
  - Bayesian Network (BN)
  - Maximum Entropy (ME)

Unsupervised Learning:
- Reinforcement Learning

Artificial intelligence (AI) in optical networks

Big Data
Organizations have to work smart. Paying careful attention to the management of BI initiatives is a necessary aspect of doing business. It is no surprise, then, that organizations are increasingly championing BI and under its new incarnation as analytics. Application Case 1.1 illustrates one such application of BI that has helped many airlines as well as, of course, the companies offering such services to the airlines.

FIGURE 1.9 Evolution of Business Intelligence (BI).

T
chnical staf
Build the data warehouse - Organizing - Summarizing - Standardizing

Business user
A
ccess
Manipulation, result

Managers/executives
BPM stra

FIGURE 1.10 A High-Level Architecture of BI.

(Source: Based on W. Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions. The Data Warehousing Institute, Seattle, WA, 2003, p. 32, Illustration 5.)

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Organizations have to work smart. Paying careful attention to the management of BI initiatives is a necessary aspect of doing business. It is no surprise, then, that organizations are increasingly championing BI and under its new incarnation as analytics. Application Case 1.1 illustrates one such application of BI that has helped many airlines as well as, of course, the companies offering such services to the airlines.

**FIGURE 1.9 Evolution of Business Intelligence (BI).**

**Data Warehouse Environment**
- Technical staff
  - Build the data warehouse
  - Organizing
  - Summarizing
  - Standardizing

**Business Analytics Environment**
- Business users
  - Access
  - Manipulation, results

**Performance and Strategy**
- Managers/executives
  - BPM strategies

**User interface**
- Browser
- Portal
- Dashboard

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), *Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition*, Pearson
computer technology, management science techniques, and statistics to solve real problems. Of course, many other organizations have proposed their own interpretations and motivations for analytics. For example, SAS Institute Inc. proposed eight levels of analytics that begin with standardized reports from a computer system. These reports essentially provide a sense of what is happening with an organization. Additional technologies have enabled us to create more customized reports that can be generated on an ad hoc basis. The next extension of reporting takes us to OLAP-type queries that allow a user to dig deeper and determine specific sources of concern or opportunities. Technologies available today can also automatically issue alerts for a decision maker when performance warrants such alerts. At a consumer level we see such alerts for weather or other issues. But similar alerts can also be generated in specific settings when sales fall above or below a certain level within a certain time period or when the inventory for a specific product is running low. All of these applications are made possible through analysis and queries on data being collected by an organization. The next level of analysis might entail statistical analysis to better understand patterns. These can then be taken a step further to develop forecasts or models for predicting how customers might respond to a specific marketing campaign or ongoing service/product offerings. When an organization has a good view of what is happening and what is likely to happen, it can also employ other techniques to make the best decisions under the circumstances. These eight levels of analytics are described in more detail in a white paper by SAS (sas.com/news/sascom/analytics_levels.pdf).

This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by INFORMS in proposing three levels of analytics. These three levels are identified (informs.org/Community/Analytics) as descriptive, predictive, and prescriptive. Figure 1.11 presents a graphical view of these three levels of analytics. It suggests that these three are somewhat independent steps and one type of analytics applications leads to another. It also suggests that there is actually some overlap across these three types of analytics. In either case, the interconnected nature of different types of analytics applications is evident. We next introduce these three levels of analytics.

**Business Analytics**

**Descriptive**
- What happened?
- What is happening?
- Enablers:
  - Business reporting
  - Dashboards
  - Scorecards
  - Data warehousing
- Well-defined business problems and opportunities

**Predictive**
- What will happen?
- Why will it happen?
- Enablers:
  - Data mining
  - Text mining
  - Web/media mining
  - Forecasting
- Accurate projections of future events and outcomes

**Prescriptive**
- What should I do?
- Why should I do it?
- Enablers:
  - Optimization
  - Simulation
  - Decision modeling
  - Expert systems
- Best possible business decisions and actions

**Source:** Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, 4th Edition, Pearson
Analytics Ecosystem

Although some researchers have distinguished business analytics professionals from data scientists (Davenport and Patil, 2012), as pointed out previously, for the purpose of understanding the overall analytics ecosystem, we treat them as one broad profession. Clearly, skill needs can vary between a strong mathematician to a programmer to a modeler to a communicator, and we believe this issue is resolved at a more micro/individual level rather than at a macro level of understanding the opportunity pool. We also take the widest definition of analytics to include all three types as defined by INFORMS—descriptive/reporting/visualization, predictive, and prescriptive as described earlier.

Figure 1.13 illustrates one view of the analytics ecosystem. The components of the ecosystem are represented by the petals of an analytics flower. Eleven key sectors or clusters in the analytics space are identified. The components of the analytics ecosystem are grouped into three categories represented by the inner petals, outer petals, and the seed (middle part) of the flower.

The outer six petals can be broadly termed as the technology providers. Their primary revenue comes from providing technology, solutions, and training to analytics user organizations so they can employ these technologies in the most effective and efficient manner. The inner petals can be generally defined as the analytics accelerators. The accelerators work with both technology providers and users. Finally, the core of the ecosystem comprises the analytics user organizations. This is the most important component, as every analytics industry cluster is driven by the user organizations.

The metaphor of a flower is well-suited for the analytics ecosystem as multiple components overlap each other. Similar to a living organism like a flower, all these petals grow and wither together. We use the terms components, clusters, petals, and sectors interchangeably to describe the various players in the analytics space. We introduce each of the industry sectors next and give some examples of players in each sector. The list of company names included in any petal is not exhaustive. The representative list of companies in each cluster is just to illustrate that cluster’s unique offering to describe where analytics talent may be used or hired away. Also, mention of a company’s name or its capability in one specific area.
Job Titles of Analytics

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics. Not all data is useful for all tasks, obviously. That is, data has to match with (have the coverage of the specifics for) the task for which it is intended to be used. Even for a specific task, the relevant data on hand needs to comply with the quality and quantity requirements. Essentially, data has to be analytics ready. So what does it mean to make data analytics ready? In addition to its rel evancy to the problem at hand and the quality/quantity requirements, it also has to have a certain data structure in place with key fields/variables with properly normalized values. Furthermore, there must be an organization-wide agreed-on definition for common variables and subject matters (sometimes also called master data management), such as how you define a customer (what characteristics of customers are used to produce a holistic enough representation to analytics) and where in the business process the customer-related information is captured, validated, stored, and updated.

Sometimes the representation of the data may depend on the type of analytics being employed. Predictive algorithms generally require a flat file with a target variable, so making data analytics ready for prediction means that data sets must be transformed into a flat-file format and made ready for ingestion into those predictive algorithms. It is also imperative to match the data to the needs and wants of a specific predictive algorithm and/or a software tool—for instance, neural network algorithms require all input variables to be numerically represented (even the nominal variables need to be converted.

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Big Data Analytics and Data Mining
Big Data 4 V

Volume
SCALE OF DATA

- 6 BILLION PEOPLE have cell phones
- WORLD POPULATION: 7 BILLION

Source: https://www-01.ibm.com/software/data/bigdata/

Velocity
ANALYSIS OF STREAMING DATA

- The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session
- Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure
- By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

- 27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Variety
DIFFERENT FORMS OF DATA

- 30 BILLION PIECES OF CONTENT are shared on Facebook every month
- 400 MILLION TWEETS are sent per day by about 200 million monthly active users

The FOUR V’s of Big Data

- It’s estimated that 2.5 QUINTILLION BYTES (2.3 TRILLION GIGABYTES) of data are created each day
- Most companies in the U.S. have at least 100 TERABYTES (101,000 GIGABYTES) of data stored

The global size of data in healthcare was estimated to be 150 EXABYTES (150,000,000,000,000 GIGABYTES)
based on analysis

- As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES (150,000,000,000,000 GIGABYTES)

- By 2014, it’s anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
- 4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPECG, QAS

By 2015
4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States

1 IN 3 BUSINESS LEADERS don’t trust the information they use to make decisions

Poor data quality costs the US economy around $3.1 TRILLION A YEAR

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

Source: https://www-01.ibm.com/software/data/bigdata/
Value
Stephan Kudyba (2014),
Big Data, Mining, and Analytics:
Components of Strategic Decision Making, Auerbach Publications

Source: http://www.amazon.com/gp/product/1466568704
Architecture of Big Data Analytics

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

Big Data Platforms & Tools
- Queries
- Reports
- OLAP
- Data Mining

Data Mining
Big Data Analytics Applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Architecture for Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Enabling Technologies
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Analysts
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Business Intelligence (BI) Infrastructure

- Operational Data
- Historical Data
- Machine Data
- Web Data
- Audio/Video Data
- External Data

Data Mart

Data Warehouse

Hadoop Cluster

Extract, transform, load

Casual users
- Queries
- Reports
- Dashboards

Power users
- Queries
- Reports
- OLAP
- Data mining

Business Intelligence and Data Mining

Increasing potential to support business decisions

- Decision Making
- Data Presentation
  - Visualization Techniques
  - Data Mining
    - Information Discovery
- Data Exploration
  - Statistical Summary, Querying, and Reporting
- Data Preprocessing/Integration, Data Warehouses
- Data Sources
  - Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business Analyst

Data Analyst

DBA

Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier
Data Mining at the Intersection of Many Disciplines

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining:

Core Analytics Process

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Data Mining

Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)

Knowledge Discovery (KDD) Process

Data mining: core of knowledge discovery process

Data Cleaning
Data Integration
Data Warehouse
Task-relevant Data
Selection
Data Mining
Pattern Evaluation

Source: Han & Kamber (2006)
Data Mining Processing Pipeline
(Charu Aggarwal, 2015)

Data Collection → Data Preprocessing → Analytical Processing → Output for Analyst

- Data Preprocessing:
  - Feature Extraction
  - Cleaning and Integration

- Analytical Processing:
  - Building Block 1
  - Building Block 2

Feedback (Optional) → Feedback (Optional)
A Taxonomy for Data Mining Tasks

<table>
<thead>
<tr>
<th>Learning Method</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>Classification and Regression Trees, ANN, SVM, Genetic Algorithms</td>
</tr>
<tr>
<td>Supervised</td>
<td>Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms</td>
</tr>
<tr>
<td>Supervised</td>
<td>Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Apriory, OneR, ZeroR, Eclat</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Expectation Maximization, Apriory Algorithm, Graph-based Matching</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Apriory Algorithm, FP-Growth technique</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>K-means, ANN/SOM</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>K-means, Expectation Maximization (EM)</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Cloud Computing
Google Cloud

Machine learning and Cloud AI

Source: https://cloud.google.com/solutions/big-data/overview/machine-learning-cloud-ai/
Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. See our FAQ for more info.

Getting Started

- Overview of Colaboratory
- Loading and saving data: Local files, Drive, Sheets, Google Cloud Storage
- Importing libraries and installing dependencies
- Using Google Cloud BigQuery
- Forms, Charts, Markdown, & Widgets
- TensorFlow with GPU
- Machine Learning Crash Course: Intro to Pandas & First Steps with TensorFlow

Highlighted Features

Seedbank

Looking for Colab notebooks to learn from? Check out Seedbank, a place to discover interactive machine learning examples.

TensorFlow execution

Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices.

\[
\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}
\]
Cloud Computing

AWS

Amazon Web Services

Source: https://aws.amazon.com/
Data Lakes and Analytics on AWS

**Data Movement**
Import your data from on-premises, and in real-time.

**Data Lake**
Store any type of data securely, from gigabytes to exabytes.

**Analytics**
Analyze your data with a broad selection of analytic tools and engines.

**Machine Learning**
Forecast future outcomes, and prescribe actions.

AWS Products
Analytics

• Amazon Athena
  – Query data in S3 using SQL
• Amazon CloudSearch
  – Managed search service
• Amazon EMR
  – Hosted Hadoop framework
• Amazon Elasticsearch Service
  – Run and scale Elasticsearch clusters
• Amazon Kinesis
  – Analyze real-time video and data streams
• Amazon Redshift
  – Fast, simple, cost-effective data warehousing
• Amazon QuickSight
  – Fast business analytics service
• AWS Data Pipeline
  – Orchestration service for periodic, data-driven workflows
• AWS Glue
  – Prepare and load data

Source: [https://aws.amazon.com/](https://aws.amazon.com/)
Machine Learning on AWS
Machine learning in the hands of every developer and data scientist

Source: https://aws.amazon.com/machine-learning/
Cloud Computing
AWS Cloud Practitioner
AWS Solutions Architect
AWS Certified Big Data Specialty

Role-Based Certifications

Professional
- AWS Certified Solutions Architect - Professional
- AWS Certified DevOps Engineer - Professional

Associate
- AWS Certified Solutions Architect - Associate
- AWS Certified DevOps Engineer - Associate

Foundational
- AWS Certified Cloud Practitioner

Specialty Certifications

- AWS Certified Advanced Networking - Specialty
- AWS Certified Big Data - Specialty
- AWS Certified Security - Specialty

Source: https://aws.amazon.com/certification/
Short Text Conversation (STC)
AI and Dialogue System
Chatbot

Source: https://www.mdsdecoded.com/blog/the-rise-of-chatbots/
Can machines think?

(Alan Turing, 1950)

Chatbot

“online human-computer dialog system with natural language.”

Chatbot Conversation Framework

Conversations

- Open Domain
- Closed Domain

Responses

- Retrieval-Based
- Generative-Based

- Impossible
- Rules-Based [Easiest]
- General AI [Hardest]
- Smart Machine [Hard]

Source: https://chatbotslife.com/ultimate-guide-to-leveraging-nlp-machine-learning-for-you-chatbot-531ff2dd870c
From E-Commerce to Conversational Commerce: Chatbots and Virtual Assistants

Source: http://www.guided-selling.org/from-e-commerce-to-conversational-commerce/
H&M’s chatbot on Kik

Source: http://www.guided-selling.org/from-e-commerce-to-conversational-commerce/
Uber’s chatbot on Facebook’s messenger

- one main benefit: it loads much faster than the Uber app

Source: http://www.guided-selling.org/from-e-commerce-to-conversational-commerce/
Dialogue System

Short Text Conversation Task (STC-3)
Chinese Emotional Conversation Generation (CECG) Subtask

Source: http://coai.cs.tsinghua.edu.cn/hml/challenge.html
# NTCIR Short Text Conversation

**STC-1, STC-2, STC-3**

<table>
<thead>
<tr>
<th></th>
<th>Japanese</th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
</table>
| **NTCIR-12 STC-1**
22 active participants | Twitter, Retrieval           | Weibo, Retrieval            |                       |
| **NTCIR-13 STC-2**
27 active participants  | Yahoo! News, Retrieval+Generation | Weibo, Retrieval+Generation |                       |
| **NTCIR-14 STC-3**
| Chinese Emotional Conversation Generation (CECG) subtask | | Weibo, Generation for given emotion categories |
| | Dialogue Quality (DQ) and Nugget Detection (ND) subtasks | | Weibo+English translations, distribution estimation for subjective annotations |

The 14th NTCIR (2018 - 2019)

The 14th NTCIR (2018 - 2019)
Evaluation of Information Access Technologies

January 2018 - June 2019

What's New

- February 1, 2018: Call for participation to the NTCIR-14 Kick-Off Event released.
- February 1, 2018: Call for participation to the NTCIR-14 QALab-PolInfo Kick-Off Event released.

December 5, 2017: The NTCIR-14 Task Selection Committee has selected the following six Tasks. Lifelig-3, OpenLiveQ-2, QA Lab-4, STC-3, WWW-2, CENTRE.

August 23, 2017: NTCIR-14 Call for Task Proposals released. (Closed.)

About Proceedings

After the NTCIR-14 conference, a post-proceedings of revised selected papers will be published in the Springer Lecture Notes on Computer Science (LNCS) series.

http://research.nii.ac.jp/ntcir/ntcir-14/index.html
NTCIR-14

Short Text Conversation Task (STC-3)

Welcome to the top page of STC-3@NTCIR-14!
STC-3 offers three subtasks:

- Chinese Emotional Conversation Generation (CECG) Subtask
- Dialogue Quality (DQ) Subtask (for Chinese and English)
- Nugget Detection (ND) Subtask (for Chinese and English)

Key dates for DQ and ND Subtasks

Feb-Mar 2018  Crawling Chinese test data from Weibo
Oct 2017-Jan 2018  Training data translation into English
Apr-Jun, 2018  Test data translation into English
Jul-Aug 2018  Training/test data annotation
Aug 31, 2018  STC-3 task registrations due (CECG, DQ, ND)
Sep 1, 2018  Training data with annotations released
Nov 1, 2018  Test data released
Nov 30, 2018  Run submissions due
Dec 20, 2018  Results and draft overview released to participants
Feb 1, 2019  Participant papers due
Mar 1, 2019  Acceptance notification
Mar 20, 2019  All camera-ready papers due
Jun 2019  NTCIR-14 Conference@NII
Call for Participation

In recent years, there has been a rising tendency in AI research to enhance Human-Computer Interaction by humanizing machines. However, to create a robot capable of acting and talking with a user at the human level requires the robot to understand human cognitive behaviors, while one of the most important human behaviors is expressing and understanding emotions and affects. As a vital part of human intelligence, emotional intelligence is defined as the ability to perceive, integrate, understand, and regulate emotions. Though a variety of models have been proposed for conversation generation from large-scale social data, it is still quite challenging (and yet to be addressed) to generate emotional responses.

In this challenge, participants are expected to generate Chinese responses that are not only appropriate in content but also adequate in emotion, which is quite important for building an empathic chatting machine. For instance, if user says “My cat died yesterday”, the most appropriate response may be “It's so sad, so sorry to hear that” to express sadness, but also could be “Bad things always happen, I hope you will be happy soon” to express comfort.

Previous Evaluation Challenge at NLPCC 2017

Overview of the NLPCC 2017 Shared Task: Emotion Generation Challenge
Short Text Conversation (NTCIR-13 STC2) Retrieval-based

retrieval-based method

Given a new post, can a coherent and useful comment be returned by searching a post-comment repository?

Short Text Conversation (NTCIR-13 STC2) Generation-based

generation-based method

Given a new post, can a fluent, coherent and useful comment be generated?

post

post-comment repository

The Trained Generator

Understanding

Generating

generated comment

generated comment

generated comment

Used to train the generator

post

comment

comment

comment

post

comment

comment

comment

post

comment

comment

comment

post

comment

comment

comment

Short Text Conversation (STC-3)

- Emotional Conversation Generation
- Dialogue Quality
- Nugget Detection subtasks using Chinese and English dialogue data

Source: [http://research.nii.ac.jp/ntcir/ntcir-14/tasks.html](http://research.nii.ac.jp/ntcir/ntcir-14/tasks.html)
NTCIR-14
Short Text Conversation Task (STC-3)

• Three new subtasks
  – Chinese Emotional Conversation Generation (CECG)
  – Dialogue Quality (DQ) (for Chinese and English)
  – Nugget Detection (ND) (for Chinese and English)

Source: http://research.nii.ac.jp/ntcir/ntcir-14/tasks.html
NTCIR-14 STC-3 CECG
Time Schedule

• CECG
  – Aug 31, 2018 CECG registration due
  – Sep 15, 2018 Training data released
  – Nov 15, 2018 Test data released
  – Nov 25, 2018 Run submissions due

• ALL
  – Dec 20, 2018 Evaluation results and draft overview released
  – Feb 1, 2019 Participant paper submissions due
  – Mar 1, 2019 Acceptance notification
  – Mar 20, 2019 All camera-ready papers due
  – Jun 2019 NTCIR-14 Conference & EVIA 2019 in NII, Tokyo

Source: http://coai.cs.tsinghua.edu.cn/hml/challenge/time_schedule/
### Given post:
The man who cooks and loves dogs is very handsome!

### Responses:
<table>
<thead>
<tr>
<th>Response</th>
<th>Chinese / Translated English</th>
<th>Emotion Class</th>
<th>Coherence and Fluency</th>
<th>Emotion Consistency</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response 1</td>
<td>会做饭的男人是很帅的啊。 The man who cooks is handsome.</td>
<td>喜欢 Like</td>
<td>Yes</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Response 2</td>
<td>哈哈，我也觉得。 Haha, I feel the same way.</td>
<td>喜欢 Like</td>
<td>Yes</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>Response 3</td>
<td>这是哪部电影里的？ Which movie is this from?</td>
<td>厌恶 Disgust</td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>Response 4</td>
<td>哈哈,你也是。 Haha, the same to you.</td>
<td>喜欢 Like</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>Response 5</td>
<td>我爱你会。 I love you, will you.</td>
<td>厌恶 Disgust</td>
<td>No</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>Response 6</td>
<td>这是同主义同的道！ This is the same way of the same doctrine!</td>
<td>厌恶 Disgust</td>
<td>No</td>
<td>No</td>
<td>0</td>
</tr>
</tbody>
</table>

NLPCC
Emotion Classification Challenge
6 classes (64%)

- 0: Other
- 1: Like
- 2: Sadness
- 3: Disgust
- 4: Anger
- 5: Happiness

Source: http://coai.cs.tsinghua.edu.cn/hml/challenge/dataset_description/
<table>
<thead>
<tr>
<th>Post (Given)</th>
<th>Emotion Category (Given)</th>
<th>Response (to be generated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>爱狗还会做饭的男人，最帅了！</td>
<td>喜欢</td>
<td>会做饭的男人是很帅的啊。</td>
</tr>
<tr>
<td>The man who cooks and loves dogs is very handsome!</td>
<td>Like</td>
<td>The man who cooks is handsome.</td>
</tr>
<tr>
<td>爸爸，明天是六一啦，带我出去玩吧！</td>
<td>喜欢</td>
<td>好，一定带。</td>
</tr>
<tr>
<td>Daddy, tomorrow is June 1th Children’s Day. Please take me out to play!</td>
<td>Like</td>
<td>Yes, of course.</td>
</tr>
<tr>
<td>被特别在乎的人忽略，会很难过，而装做不在乎，会更难过。</td>
<td>悲伤</td>
<td>因为在乎，所以在乎。</td>
</tr>
<tr>
<td>It's sad to be ignored by someone that you really care, and it's harder to pretend that you don't care.</td>
<td>Sadness</td>
<td>I am concerned with it, because I care about it.</td>
</tr>
<tr>
<td>冰天雪地的哈尔滨！你要冻死我咋地？！！</td>
<td>喜欢</td>
<td>注意保暖，小心别感冒！</td>
</tr>
<tr>
<td>Harbin is in the snow! Are you going to freeze me to death?!?!</td>
<td>Like</td>
<td>Keep warm. Watch out for getting cold!</td>
</tr>
<tr>
<td>本来想学一把沧桑，结果令我更忧伤。</td>
<td>悲伤</td>
<td>这是一个悲伤的故事。</td>
</tr>
<tr>
<td>I wanted to learn the vicissitudes of life, but I became sadder.</td>
<td>Sadness</td>
<td>It a sad story.</td>
</tr>
<tr>
<td>今天一整天都在海边训练。虽然很累，但还是很开心的！</td>
<td>快乐</td>
<td>是的呢，开心！</td>
</tr>
<tr>
<td>I have been training at the seaside all day. Though very tired, I still very happy!</td>
<td>Happiness</td>
<td>Yeah, happy!</td>
</tr>
</tbody>
</table>

• 1,110,000 Weibo post-response pairs
  - [[[post,post_label],[response,response_label]],
    [[post,post_label],[response,response_label]],...].

Source: http://coai.cs.tsinghua.edu.cn/hml/challenge/dataset_description/
Fluency judgement on responses with repetitive words

<table>
<thead>
<tr>
<th>Response</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>悲哀，悲哀，悲哀。</td>
<td>Yes</td>
</tr>
<tr>
<td>飘逸，飘逸，飘逸！</td>
<td>Yes</td>
</tr>
<tr>
<td>好可爱，好可爱！</td>
<td>Yes</td>
</tr>
<tr>
<td>哈哈哈，当然啦</td>
<td>Yes</td>
</tr>
<tr>
<td>疯了。疯了。全疯了</td>
<td>Yes</td>
</tr>
<tr>
<td>对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的对的對...</td>
<td></td>
</tr>
</tbody>
</table>
Sample responses generated by Seq2Seq and ECM (Emotional Chatting Machine)

<table>
<thead>
<tr>
<th>Post</th>
<th>Model</th>
<th>Response(Original)</th>
<th>Response(Translated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>承诺, 就是一个骗子说给一个傻子听的。 A promise is what a cheater says to a fool.</td>
<td>Seq2Seq</td>
<td>承诺, 就是一个骗子说给一个傻子听的。</td>
<td>A promise is what a cheater says to a fool.</td>
</tr>
<tr>
<td>承诺, 就是一个骗子说给一个傻子听的。 A promise is what a cheater says to a fool.</td>
<td>ECM</td>
<td>承诺是一个专一的傻子</td>
<td>Promise is a single-minded fool</td>
</tr>
<tr>
<td>我就是一个开心的傻子</td>
<td>Like</td>
<td>我就是一个开心的傻子</td>
<td>I'm just a happy fool</td>
</tr>
<tr>
<td>我就是一个悲伤的傻子</td>
<td>Happy</td>
<td>我就是一个悲伤的傻子</td>
<td>I am a sad fool</td>
</tr>
<tr>
<td>一个骗子说给一个傻子听</td>
<td>Sad</td>
<td>一个骗子说给一个傻子听</td>
<td>A cheater says to a fool</td>
</tr>
<tr>
<td>不管你信不信，反正我信了！</td>
<td>Disgust</td>
<td>不管你信不信，反正我信了！</td>
<td>Whatever you believe or not, I trust in it!</td>
</tr>
<tr>
<td>知道真相的我眼泪掉下来。 Knowing the truth, I burst into tears.</td>
<td>Seq2Seq</td>
<td>知道真相的我眼泪掉下来。</td>
<td>After knowing the truth, I burst into tears.</td>
</tr>
<tr>
<td>知道真相的我眼泪掉下来。 Knowing the truth, I burst into tears.</td>
<td>ECM</td>
<td>太可爱了，太可爱了！</td>
<td>Too lovely, it's too lovely.</td>
</tr>
<tr>
<td>笑得我眼泪都出来了。</td>
<td>Happy</td>
<td>笑得我眼泪都出来了。</td>
<td>I can’t help laugh myself to tears</td>
</tr>
<tr>
<td>知道真相的我眼泪掉下来。</td>
<td>Sad</td>
<td>知道真相的我眼泪掉下来。</td>
<td>After knowing the truth, I burst into tears.</td>
</tr>
<tr>
<td>这是真的还是假的？</td>
<td>Disgust</td>
<td>这是真的还是假的？</td>
<td>Is this true?</td>
</tr>
<tr>
<td>什么意思？没看懂。</td>
<td>Angry</td>
<td>什么意思？没看懂。</td>
<td>What do you mean? I don’t understand.</td>
</tr>
</tbody>
</table>

Sample responses generated by Seq2Seq and ECM (Emotional Chatting Machine)

<table>
<thead>
<tr>
<th>Post</th>
<th>Model</th>
<th>Response(Original)</th>
</tr>
</thead>
<tbody>
<tr>
<td>承诺，就是一个骗子说给一个傻子听的。 A promise is what a cheater says to a fool.</td>
<td>Seq2Seq</td>
<td>承诺，就是一个骗子说给一个傻子听的。</td>
</tr>
<tr>
<td></td>
<td>ECM</td>
<td>互联是一个专一的傻子</td>
</tr>
<tr>
<td></td>
<td></td>
<td>我就是一个开心的傻子</td>
</tr>
<tr>
<td></td>
<td></td>
<td>我就是一个悲伤的傻子</td>
</tr>
<tr>
<td></td>
<td></td>
<td>一个骗子说给一个傻子听</td>
</tr>
<tr>
<td></td>
<td></td>
<td>不管你信不信，反正我信了！</td>
</tr>
</tbody>
</table>

# Emotional Short Text Conversation (ESTC) Dataset

<table>
<thead>
<tr>
<th></th>
<th>Posts</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Posts</td>
<td>217,905</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>Posts</td>
<td>Angry</td>
<td>234,635</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Responses</td>
<td>Disgust</td>
<td>689,295</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Happy</td>
<td>306,364</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Like</td>
<td>1,226,954</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sad</td>
<td>537,028</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>1,365,371</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>Posts</td>
<td></td>
<td></td>
<td>1,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>Posts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,000</td>
</tr>
</tbody>
</table>

Conversations with/without considering emotion

Emotional Chatting Machine (ECM)

- User: Worst day ever. I arrived late because of the traffic.
  1. Basic Seq2Seq: You were late.
  2. ECM (Like): I am always here to support you.
  3. ECM (Happy): Keep smiling! Things will get better.
  4. ECM (Sad): It’s depressing.
  5. ECM (Disgust): Sometimes life just sucks.
  6. ECM (Angry): The traffic is too bad!

Overview of Emotional Chatting Machine (ECM)

Overview of Emotional Chatting Machine (ECM)

Overview of Emotional Chatting Machine (ECM)

Data flow of the decoder with an internal memory

Data flow of the decoder with an external memory

Sample responses generated by Seq2Seq and ECM (Emotional Chatting Machine)

<table>
<thead>
<tr>
<th>Post</th>
<th>Model</th>
<th>Response (Original)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowing the truth, I laughed myself to tears</td>
<td>Seq2Seq</td>
<td>知道真相的我眼泪掉下来。</td>
</tr>
<tr>
<td>知道真相的我眼泪笑下来</td>
<td>Like</td>
<td>太可爱了，太可爱了！</td>
</tr>
<tr>
<td>ECM</td>
<td>Happy</td>
<td>笑得我眼泪都出来了。</td>
</tr>
<tr>
<td></td>
<td>Sad</td>
<td>知道真相的我眼泪掉下来。</td>
</tr>
<tr>
<td></td>
<td>Disgust</td>
<td>这是真的还是假的？</td>
</tr>
<tr>
<td></td>
<td>Angry</td>
<td>什么意思？没看懂。</td>
</tr>
</tbody>
</table>

Chinese Emotional Conversation Generation (CECG) Evaluation Metric

- Emotion Consistency
  - whether the emotion class of a generated response is the same as the pre-specified class.

- Coherence
  - whether the response is appropriate in terms of both logically coherent and topic relevant content.

- Fluency
  - whether the response is fluent in grammar and acceptable as a natural language response.

Source: http://coai.cs.tsinghua.edu.cn/hml/challenge/evaluation_metric/
Chinese Emotional Conversation Generation (CECG) Evaluation Metric

IF Coherence and Fluency
  IF Emotion Consistency
    LABEL 2
  ELSE
    LABEL 1
ELSE
  LABEL 0

Sequence-to-sequence Learning with Attention for Generation-based STC

General Framework of Controllable Short-Text-Conversation Generation with External Memory

## Controllable Short Text Conversation Examples

<table>
<thead>
<tr>
<th>Post</th>
<th>NRM</th>
<th>MrRNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>李娜太牛了!中国第一个世界网球大满贯冠军! Li Na is great! China’s first World Tennis Grand Slam champion!</td>
<td>李娜是世界冠军! Li Na is the world champion!</td>
<td>Come on, Li Na. Come on!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post</th>
<th>ESED</th>
</tr>
</thead>
<tbody>
<tr>
<td>感人器官捐献广告,请感谢那些给你第二次生命的人. A touching organ donation advertisement, please thank those who gave you a second life.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key-Word</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>世界</td>
<td>李娜是世界上最棒的! Li Na is the best in the world!</td>
</tr>
<tr>
<td>world</td>
<td>Li Na is the best in the world!</td>
</tr>
<tr>
<td>中国</td>
<td>李娜是中国的骄傲! Li Na is the pride of China!</td>
</tr>
<tr>
<td>China</td>
<td>Li Na is the pride of China!</td>
</tr>
<tr>
<td>网球</td>
<td>中国网球公开赛,加油! China Tennis Open, come on!</td>
</tr>
<tr>
<td>tennis</td>
<td>China Tennis Open, come on!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key-Word</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>生命</td>
<td>感谢生命给予的一切! Thanks for everything given by life!</td>
</tr>
<tr>
<td>life</td>
<td>Thanks for everything given by life!</td>
</tr>
<tr>
<td>器官</td>
<td>感谢那些器官捐赠者。Thanks for those organ donors.</td>
</tr>
<tr>
<td>organ</td>
<td>Thanks for those organ donors.</td>
</tr>
<tr>
<td>广告</td>
<td>非常感人的广告! A very touching advertisement!</td>
</tr>
</tbody>
</table>

Comments Generated Using Different Semantic key Mapping Methods

<table>
<thead>
<tr>
<th>Post</th>
<th>Implementation</th>
</tr>
</thead>
</table>
| 善良的社会应该由善良的人组成。A good society should be composed of people of kindness. | Identity Mapper: 善良 kindness Kindness is a belief.  
Synonym: 友善 friendliness 善良是一种美德。Kindness is a virtue.  
Antonym: 丑恶 evil 没有丑恶的社会,更可怕。A society without evils is more terrifying.  
Embedding: 乐观 optimism 乐观是一种生活态度。Optimism is an attitude of life.  
Common-Reply: 真 really 真的吗？Really？  
对 right 对的,对的。Right, right.  
Random: 人 person 每个人都有善良的一面。Every person has a side of kindness.  
中国 China 中国的贫富差距太大了。The gap between the wealthy and the poor in China is too big. |

### Generated Responses of Knowledge Introduction by External Memory

<table>
<thead>
<tr>
<th>Post</th>
<th>Key-Word</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>你最喜欢哪个星座？</td>
<td>摩羯座</td>
<td>我最喜欢摩羯座。</td>
</tr>
<tr>
<td>What is your favorite constellation?</td>
<td>Capricorn</td>
<td>I love Capricorn.</td>
</tr>
<tr>
<td></td>
<td>射手座</td>
<td>我是射手座,我喜欢。</td>
</tr>
<tr>
<td></td>
<td>Sagittarius</td>
<td>I am a Sagittarius and I like it.</td>
</tr>
<tr>
<td>今天周杰伦的电影,太好看了。</td>
<td>郭</td>
<td>周杰伦的歌也很好听。</td>
</tr>
<tr>
<td>Jay Chou’s movie is so nice today.</td>
<td>song</td>
<td>Jay Chou’s songs are nice too.</td>
</tr>
<tr>
<td>梵高的画,很漂亮。</td>
<td>向日葵</td>
<td>梵高笔下的向日葵。</td>
</tr>
<tr>
<td>Van Gogh’s paintings are very beautiful.</td>
<td>sunflower</td>
<td>The Sunflower by Van Gogh.</td>
</tr>
<tr>
<td></td>
<td>荷兰</td>
<td>梵高的画,很漂亮。</td>
</tr>
<tr>
<td></td>
<td>Holland</td>
<td>Van Gogh’s paintings are very beautiful.</td>
</tr>
</tbody>
</table>

Welcome to NTCIR-14 STC-3

• CECG
  – Aug 31, 2018 CECG registration due
  – Sep 15, 2018 Training data released
  – Nov 15, 2018 Test data released
  – Nov 25, 2018 Run submissions due

• ALL
  – Dec 20, 2018 Evaluation results and draft overview released
  – Feb 1, 2019 Participant paper submissions due
  – Mar 1, 2019 Acceptance notification
  – Mar 20, 2019 All camera-ready papers due
  – Jun 2019 NTCIR-14 Conference & EVIA 2019 in NII, Tokyo

Source: http://coai.cs.tsinghua.edu.cn/hml/challenge/time_schedule/
Summary

• AI
• Big Data
• Cloud Computing
References

• Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.
• Jared Dean (2014), Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Wiley.
• Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study”, Expert Systems with Applications, 111, no. 30, pp. 2-10.
• Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications.