Natural Language Processing

Yue Zhang
Westlake University
Chapter 1

Introduction
Contents

• 1.1 What is Natural Language Processing (NLP)?
• 1.2 NLP tasks
  • 1.2.1 Fundamental NLP tasks
  • 1.2.2 Information Extraction tasks
  • 1.2.3 Text generation Tasks
  • 1.2.4 Other Applications
• 1.3 NLP from a Machine Learning Perspective
1.1 What is Natural Language Processing (NLP)?

1.2 NLP tasks

1.2.1 Fundamental NLP tasks

1.2.2 Information Extraction tasks

1.2.3 Text generation Tasks

1.2.4 Other Applications

1.3 NLP from a Machine Learning Perspective
What is NLP?

In the broadest sense, NLP refers to any program that automatically processes human languages.
Main approaches

Rule-based (symbolic) approach (1950s-1980s)

• The oldest approaches to NLP
• Based on human-developed rules and lexicons
• Challenges in resolving ambiguities
  "The spirit is strong, but the flesh is weak"
  "The Vodka is good, but the meat is bad"
Main approaches

Statistical approach (traditional machine learning) (1980s-2000s)

• Gradually adopted by both the academia and the industry
• Using probabilistic modeling
  • training data (corpus with markup)
  • feature engineering
  • training a model on parameters
  • applying model to test data
Main approaches

Connectionist approach (Neural networks)
(2000s-now)

• Deep learning surpasses statistical methods as the domain approach
  • free from linguistic features
  • very large neural models
  • pre-training over large raw text
Contents

• 1.1 What is Natural Language Processing (NLP)?
• 1.2 NLP tasks
  • 1.2.1 Fundamental NLP tasks
  • 1.2.2 Information Extraction tasks
  • 1.2.3 Text generation Tasks
  • 1.2.4 Other Applications
• 1.3 NLP from a Machine Learning Perspective
Fundamental Tasks

• Computational Linguistics
• Phonology
• Morphology
• Syntax
• Semantics
• Discourse
• Pragmatics
Syntactic tasks: Word level

• Morphological analysis

(English) walking
(walk + ing)

(Arabic) wktAbnA
(w + k+tAb + n + A)

(German) Wochenarbeitszeit
(Wochen + arbeits + zeit)

• Word segmentation

中国外企业务
(China foreign business in China)

• Tokenization

Mr. Smith visited
Mr. Smith visited

Wendy's new house.
Wendy 's new house.

• POS Tagging

I can open this can
(PRPR MD VB DT NN)
Syntactic tasks: Word level

• Part-of-speech (POS)

Basic syntactic role that words play in a sentence
Syntactic tasks: Sentence level

- Grammar formalisms for syntactic parsing:
Constituent parsing

Constituent parsers assign phrase labels to constituent, also referred to as phrase-structure grammars.
Dependency parsing

Dependency parsers analyze a sentence in head words and dependent words.
CCG parsing

- Lexical categories (e.g. NP, N, S\NP)
- Composition rules

*example* when the phrase \( \frac{\text{bought}}{\text{(S\NP)/NP}} \) and \( \frac{\text{a book}}{\text{NP}} \) are combined, the resulting categories (S\NP)/NP and NP are combined into S\NP, resulting in \( \frac{\text{bought a book}}{\text{S\NP}} \)
Supertagging

Also called shallow parsing, a pre-processing step before parsing.

- CCG supertagging
- Syntactic chunking

identify basic syntactic phrases from a given sentence.

He made a request for cutting down the operation budget

[NP He]    [VP made]    [NP a request]    [PP for]
[VP cutting down]    [NP the operation budget]
Semantic tasks: Word level

• Word sense disambiguation (WSD)
  Never trouble troubles till trouble troubles you.
  I saw a man saw a saw with a saw.

• Metaphor
  Love is a battlefield.
  Bob is a couch potato.
Semantic tasks: Word level

- Sense relations between words
  - Synonyms
    - quick - fast
    - bad - poor
    - big - large
  - Antonyms
    - big - small
    - bad - good
    - easy - difficult
  - Hyponyms
    - car - vehicle
    - apple - fruit
    - cat - animal
  - Meronyms
    - leaf - tree
    - nose - face
    - roof - house

- Analogy
  - king – queen / man – woman / boy – girl
Semantic tasks: Sentence level

- Predicate-argument relations
  (semantic role labeling)

Tim bought this book for $1.
Semantic graphs

Abstract Meaning Representation

Tim bought this book for $1.
Logic

“Everyone who bought this book loves it.”

“Tim bought this book.”

We can infer that “Tim loves this book.”

\[(\text{tim}(x) \land \text{book}(y) \land \text{buy}(x, y))\]
\[\forall x(\text{book}(y) \land \text{buy}(x, y) \Rightarrow \text{love}(x, y))\]
\[\rightarrow (\text{tim}(x) \land \text{book}(y) \land \text{love}(x, y))\]
More Semantic Parsing Cases

Lambda calculus

\[(\lambda x. x y (\lambda y. + y)) x\]

Text to SQL

```
SELECT
    s.name [schema], t.name [table], i.name [index],
    ips.avg_fragmentation_in_percent [fragmentation], ips.page_count [pages]
FROM sys.dm_db_index_physical_stats(DB_ID(),DEFAULT,DEFAULT,DEFAULT,DEFAULT) ips
JOIN sys.indexes i ON i.index_id = ips.index_id AND i.object_id = ips.object_id
JOIN sys.tables t ON t.object_id = ips.object_id
JOIN sys.schemas s on s.schema_id = t.schema_id
WHERE ips.page_count > 500
```
Text entailment

a directional semantic relation between two texts

*Text*: Tim went to the Riverside for dinner

*Hypotheses1*: The Riverside is an eating place  

True

*Hypotheses2*: Tim had dinner  

True

*Hypotheses3*: Tom had lunch  

False

*Hypotheses4*: Tim did not have dinner  

Contradiction
Discourse tasks

- Discourse: multiple sub-topics and coherence relations
- Discourse parsing: Analyze the coherence relations between sub-topics in a discourse.

Rhetoric structure theory

(1) The movie is interesting.  (2) Tim wants to watch it.  (3) He cannot do it this week.  (4) He has a final exam next Monday.
Discourse segmentation

(a) [The movie is interesting] and (b) [Tim wants to watch it]
(c) but [he cannot do this] because (d) [he has a final exam next Monday]

discourse markers

and  but  because  ...
Contents

• 1.1 What is Natural Language Processing (NLP)?
• 1.2 NLP tasks
  • 1.2.1 Fundamental NLP tasks
  • 1.2.2 Information Extraction tasks
  • 1.2.3 Text generation Tasks
  • 1.2.4 Other Applications
• 1.3 NLP from a Machine Learning Perspective
Information extraction (IE)

Obtain structured information from unstructured texts.
Entities

• Named entity recognition (NER)
To identify all named entity mentions from a given piece of text

Mary went to Chicago to meet her boyfriend John Smith

[PER Mary] went to [LOC Chicago] to meet [PER John Smith]
Anaphora Resolution

• resolves what a pronoun or noun phrase refers to

• Zero-pronoun resolution
detects and interprets dropped pronouns
Co-references

• Co-reference resolution

finds all expressions that refer to the same entities in a text

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tim watched eight Harry Potter movies. He found the series fascinating.</td>
<td>{Tim, he},</td>
</tr>
<tr>
<td></td>
<td>{eight Harry Potter movies, the series}</td>
</tr>
<tr>
<td>“I had a very bad dinner at The Oceanside.”, said Jennifer, “It was too salty.” She did not like the restaurant itself either, since it was very crowded.</td>
<td>{I, Jennifer, She}</td>
</tr>
<tr>
<td></td>
<td>{dinner, It}</td>
</tr>
<tr>
<td></td>
<td>{The Oceanside, the restaurant, it}</td>
</tr>
</tbody>
</table>
Relations between entities represent knowledge

- common relations
- hierarchical
- domain-specific

- Bangkok - Type-Instance - Hilton - Hotel
- Bill Gates - Affiliation - Microsoft
- Singapore - Physical - Malaysia
Relations

- Relation extraction

identify relations between entity under a set of pre-specified relation categories.

Tim_{PER} met his wife Mary_{PER} when he was working at MSRA_{ORG} in Beijing_{LOC}.
Knowledge graph

a type of databases, entities form nodes and relations form edges.
Knowledge graph

- Entity linking (entity disambiguation) determines the identity of entity mentioned from text.

  Same entity has multiple mentions
  
  USA  The US  The states  America

- Related task: Named entity normalization finds a canonical term for named entity mentions
Knowledge graph

- Link prediction (knowledge graph completion)
  Knowledge graphs allow knowledge inference.

Given “John is a singer”, “John is from Rome”, “Rome is in Italy”,

“John is from Italy”
“Italy has a singer”
Events

• Event Detection

Trigger word: “Trump visited Tokyo.”

“Trump’s Tokyo visit has finished.”

Event type classification

“DIPLOMATIC VISIT”

Argument extraction

“VISITOR=Trump”
Events

• News event detection (first story detection)
• Event factuality prediction (predict the likelihood of event)

<table>
<thead>
<tr>
<th>Event</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trump's visit to Tokyo has finished</td>
<td>1</td>
</tr>
<tr>
<td>Trump's visit to Tokyo is scheduled on June 1</td>
<td>0.96</td>
</tr>
<tr>
<td>Trump is likely to visit Tokyo in this Asia trip</td>
<td>0.7</td>
</tr>
</tbody>
</table>

• Event time extraction (e.g. temporal ordering of events)
• Causality detection
Events

• Event coreference resolution
  “I interviewed Mary yesterday. It went very smooth.”
  “it” refers to the interviewing event

• Zero-pronouns
  "Mary went to Russia to see the World Cup. Tom too."  *verb phrase ellipsis*
Events

• Script learning

Aims to extract a set of partially ordered events knowledge

In the scenario "restaurant visiting"

• "customer to be seated"
• "customer to order food"
• "waiter to serve food"
• "customer to eat food"
• "customer to pay"
## Sentiment analysis (opinion mining)

<table>
<thead>
<tr>
<th>Task</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Sentiment classification</td>
<td>This is a film well worth seeing.</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>It’s too slowly paced to be a thriller.</td>
<td>negative</td>
</tr>
<tr>
<td>(B) Targeted sentiment</td>
<td>[IOS] is much better than [Android].</td>
<td>{IOS: negative,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Android: negative}</td>
</tr>
<tr>
<td></td>
<td>Does [Amazon] support [Alipay]?</td>
<td>{Amazon: neutral,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alipay: neutral}</td>
</tr>
<tr>
<td>(C) Aspect-oriented sentiment</td>
<td>The USB receiver is small and fits inside the mouse</td>
<td>{USB receiver:</td>
</tr>
<tr>
<td></td>
<td>when not in use. Batteries are easy to install.</td>
<td>positive,</td>
</tr>
<tr>
<td></td>
<td>It is shorter than a normal mouse, which is going to</td>
<td>Battery: positive,</td>
</tr>
<tr>
<td></td>
<td>take some getting used to. I wish it were the same</td>
<td>Size: negative}</td>
</tr>
<tr>
<td></td>
<td>size as a normal mouse.</td>
<td></td>
</tr>
<tr>
<td>(D) More Fine-grained</td>
<td>Tim blamed Mary for not buying the watch.</td>
<td>{Opinion holder: Tim</td>
</tr>
<tr>
<td>classification</td>
<td></td>
<td>Opinion target: Mary</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Opinion expression: not buying the watch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sentiment polarity: negative}</td>
</tr>
</tbody>
</table>
Sentiment related tasks

• Sarcasm detection
  "Like you care!"

• Sentiment lexicon acquisition
  lexicons that contain sentiment-baring words, polarities and strengths

• Emotion detection
  "anger","disappointed","excited"
Sentiment related tasks

• Stance detection and argumentation mining
  "for", "against"

Nuclear Energy

- attack
- support

Nuclear energy may have horrific consequences if an accident occurs

Nuclear energy has an enormous capacity for energy production with no carbon emissions
1.1 What is Natural Language Processing (NLP)?

1.2 NLP tasks

1.2.1 Fundamental NLP tasks

1.2.2 Information Extraction tasks

1.2.3 Text generation Tasks

1.2.4 Other Applications

1.3 NLP from a Machine Learning Perspective
Realization

The generation of natural language text from syntactic/semantic representations

Semantic dependency graphs (logical forms) example:

Logical form for *he has a point he wants to make*, with gold standard CCG supertags for each node.
Data-to-text Generation

The generation of natural language text from syntactic/semantic representations

Example of a set of triples and the corresponding text:

A.C. Cesena manager Massimo Drago
Massimo Drago club S.S.D. Potenza Calcio
Massimo Drago club Calcio Catania

↓

Massimo Drago played for the club SSD Potenza Calcio and his own club was Calcio Catania. He is currently managing AC Cesena.
Summarization

Different types of summarization

According to input
- Single document summarization
- Multi-document summarization
- Extractive summarization
- Abstractive summarization

According to output
- Title generation and key phrase generation
- Key phrase extraction (keyword extraction)
<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>hello world</td>
<td>Bonjour le monde</td>
</tr>
</tbody>
</table>
Grammar error correction

• Grammar error detection
• Disfluency detection
• Writing quality assessment

Input (Erroneous)
She see Tom is caught by policeman in park at last night.

Output (Corrected)
She saw Tom caught by a policeman in the park last night.
Question answering (QA)

• Knowledge-base QA

A semantic graph for an example question “What was the first Taylor Swift album?”
Question answering (QA)

• Reading comprehension (machine reading) answer questions in interpretive ways

An example from the Stanford Question Answering Dataset (SQuAD):

Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory. The processing element carries out arithmetic and logic operations, and a sequencing and control unit can change the order of operations in response to stored information. Peripheral devices allow information to be retrieved from an external source, and the result of operations saved and retrieved.

• In computer terms, what does CPU stand for?
• What are the devices called that are from an external source?
• What are two things that a computer always has?
Question answering (QA)

• Community QA

An example of Question Answering from website forum showing three pairwise interactions between the original question $q$, the related question $q'$, and a comment $c$ in the related question thread.

```
q: Can I drive with an Australian driver’s license in Qatar?

q': How long can i drive in Qatar with my international driver's permit before I'm forced to change my Australian license to a Qatari one? When I do change over to a Qatari license do I actually lose my Australian license? I'd prefer to keep it if possible...

c: depends on the insurer, Qatar Insurance Company said this in email to me: “Thank you for your email! With regards to your query below, a foreigner is valid to drive in Doha with the following conditions: Foreign driver with his country valid driving license allowed driving only for one week from entry date Foreign driver with international valid driving license allowed driving for 6 months from entry date Foreign driver with GCC driving license allowed driving for 3 months from entry”. As an Aussie your driving licence should be transferable to a Qatar one with only the eyetest (temporary, then permanent once RP sorted).
```
Question answering (QA)

• Open QA

An example from the Natural Questions corpus:

**Question:** what color was john wilkes booth’s hair

**Wikipedia Page:** John_Wilkes_Booth

**Long answer:** Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astonishing memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair, and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.

**Short answer:** jet-black
Dialogue systems

- Chit-chat
- Task-oriented dialogues
Contents

• 1.1 What is Natural Language Processing (NLP)?
• 1.2 NLP tasks
  • 1.2.1 Fundamental NLP tasks
  • 1.2.2 Information Extraction tasks
  • 1.2.3 Text generation Tasks
• 1.2.4 Other Applications
• 1.3 NLP from a Machine Learning Perspective
Information retrieval

- Text classification / text clustering
  - Text topics classification
    "finance", "sports", "Tech"...
  - Spam detection
    \textit{email spam}
  - Opinion spam detection
    \textit{whether a review contains deceptive false opinions}
  - Language identification
    "French", "English"
  - Rumor detection
    \textit{false statement}
  - Humor detection
Recommendation system

leverage text reviews for recommending
Text mining and text analytics

- derive high-quality information from text
  - Clinical decision assistance
  - Stock market prediction
  - Movie revenue prediction
  - Presidential election results prediction
Contents

• 1.1 What is Natural Language Processing (NLP)?
• 1.2 NLP tasks
  • 1.2.1 Fundamental NLP tasks
  • 1.2.2 Information Extraction tasks
  • 1.2.3 Text generation Tasks
  • 1.2.4 Other Applications
• 1.3 NLP from a Machine Learning Perspective
Machine learning perspective

- The current dominant method

- The historical of research
Machine learning perspective

NLP tasks are many and dynamically evolving, but fewer according to machine learning nature

• Classification
  Output is a distinct label from a set

• Structure prediction
  Outputs are structures with inter-related sub structures

• Regression
  Output is a real valued number, e.g. predicting stock prices
Categorized by the training data

• Unsupervised learning
data without human annotation

• Supervised learning
data with human annotated gold-standard output labels

• Semi-supervised learning
both data with labels and data without annotation
Summary

• What is Natural Language Processing (NLP)
• A spectrum of NLP problems
• NLP from a machine learning perspective
Resources

• NLP toolkits

NLTK - leading platform for text processing libraries and corpora
https://www.nltk.org
AllenNLP - NLP research library built on PyTorch
https://allennlp.org/
Stanford’s Core NLP Suite
http://nlp.stanford.edu/software/corenlp.shtml
Huggingface Transformer - pretrained models ready to use
https://github.com/huggingface/transformers
Resources

• Word level syntax

POS tagging online:
https://part-of-speech.info
The Stanford log-linear POS tagger
https://nlp.stanford.edu/software/tagger.html
NLP4j - robust POS tagging using dynamic model selection
https://emorynlp.github.io/nlp4j/
Flair - with a state-of-the-art POS tagging model
https://github.com/zalandoresearch/flair/
Resources

• Syntax

spaCy - industrial-strength NLP in python, for parsing and more  [https://spacy.io/](https://spacy.io/)
phpSyntaxTree - generate graphical syntax trees  [http://ironcreek.net/phpsyntaxtree/](http://ironcreek.net/phpsyntaxtree/)
The Stanford Parser  [https://nlp.stanford.edu/software/lex-parser.html](https://nlp.stanford.edu/software/lex-parser.html)
Penn Treebank  [https://www.sketchengine.eu/penn-treebank-tagset/](https://www.sketchengine.eu/penn-treebank-tagset/)
CCGBank  [http://groups.inf.ed.ac.uk/ccg/ccgbank.html](http://groups.inf.ed.ac.uk/ccg/ccgbank.html)
Resources

• Lexical semantics

WordNet - the de-facto sense inventory for English
https://wordnet.princeton.edu/
Open Mind Word Expert sense-tagged data
http://www.cse.unt.edu/~rada/downloads.html#omwe
CuiTools - a complete word sense disambiguation system
http://sourceforge.net/projects/cuitools/
WDS Gate - a WSD toolkit using GATE and WEKA
http://sourceforge.net/projects/wsdgate/
SEMPRE - a toolkit for training semantic parsers
https://nlp.stanford.edu/software/sempre/
Resources

• Semantic roles

PropBank - the proposition bank
https://propbank.github.io/
Implied Relationships - predicate argument relationships
http://u.cs.biu.ac.il/~nlp/resources/

• Logic

GEO880
http://www.cs.utexas.edu/users/ml/nldata/geoquery.html
DeepMind logical entailment dataset
https://github.com/deepmind/logical-entailment-dataset
Resources

• AMR

AMR - abstract meaning representation
https://amr.isi.edu/
Segrada - semantic graph database
https://segrada.org/

• Text entailment

The Stanford Natural Language Inference (SNLI) Corpus
https://nlp.stanford.edu/projects/snli/
MultiNLI - the multi-genre NLI corpus
https://www.nyu.edu/projects/bowman/multinli/
Resources

• Discourse segmentation

PDTB - Penn Discourse Treebank
https://www.seas.upenn.edu/~pdtb/
Prague Discourse Treebank - annotation of discourse relations
https://ufal.mff.cuni.cz/pdit2.0
Resources

• NER

Stanford Named Entity Recognizer (NER)
https://nlp.stanford.edu/software/CRF-NER.html

OpeNER - open Polarity Enhanced Name ENtity Recognition
https://www.opener-project.eu/

CoNLL 2003 language-independent named entity recognition
http://www.cnts.ua.ac.be/conll2003/ner/

OntoNotes
https://catalog.ldc.upenn.edu/LDC2013T19

MUC-3 and MUC-4 datasets
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
Resources

• Co-reference

BART coreference system
http://www.bart-coref.org/
CherryPicker - a coreference resolution tool with cluster ranker
http://www.hlt.utdallas.edu/~altaf/cherrypicker/
Resources

• Relation extraction

The NewYorkTimes(NYT) - supervised relationship extraction
https://catalog.ldc.upenn.edu/LDC2008T19
ACE2004 - multilingual training corpus
https://catalog.ldc.upenn.edu/LDC2005T09
SemWval2010
http://semeval2.fbk.eu/
TACRED - relation extraction dataset built on newswire, web text
https://nlp.stanford.edu/projects/tacred/
RewRel - the largest supervised relation classification dataset
http://www.zhuhao.me/fewrel/
Resources

• Knowledge graph

Microsoft Text Analytics
https://labs.cognitive.microsoft.com/en-us/project-entity-linking

Dexter - a open source framework for entity linking
http://dexter.isti.cnr.it/

neleval - for named entity linking and coreference resolution
https://pypi.org/project/neleval/
Resources

• Events

ACE(KBP) automatic content extraction
https://cs.nyu.edu/grishman/jet/guide/ACEstructures.html

TimeBank 1.2
https://catalog.ldc.upenn.edu/LDC2006T08

TAC KBP 2017 - event tracking

Story Cloze Test corpora
http://cs.rochester.edu/nlp/rocstories/
Resources

• Sentiment

The Stanford Sentiment Treebank (SST) - movie reviews
https://nlp.stanford.edu/sentiment/index.html
MPQA - news articles manually annotated for opinions
http://mpqa.cs.pitt.edu/corpora/
SemEval17 - consist of 5 subtasks, both Arabic and English
http://www.aclweb.org/anthology/S17-2088
The IMDb dataset - reviews from IMDb with label
https://kaggle.com/carolzhangdc/imdb-5000-movie-dataset
MeaningCloud
Https://www.meaningcloud.com
Resources

• Machine translation

Workshop on Statistical Machine Translation (WMT)
http://www.statmt.org/wmt14/translation-task.html
International Workshop on Spoken Language Translation (IWSLT)
http://workshop2015.iwslt.org/
OpenNMT - open source neural machine translation
http://opennmt.net/
BinQE - a machine translation dataset annotated with binary quality judgements
https://ict.fbk.eu/binqe/
T2T for neural translation
https://github.com/tensorflow/tensor2tensor
Resources

• Summarization

The CNN / Daily Mail dataset - training machine reading systems
https://arxiv.org/abs/1506.03340

• Grammar error correction

CoNLL-2014 Shared Task - benchmark GEC systems
https://www.comp.nus.edu.sg/~nlp/conll14st/
Resources

• QA

CoQA - a conversational question answering dataset
https://stanfordnlp.github.io/coqa/
QBLink - sequential open-domain question answering
https://sites.google.com/view/qanta/projects/qblink
DrQA: Open Domain Question Answering
https://github.com/facebookresearch/DrQA
DocQA: Multi-Paragraph Reading Comprehension by AllenAI
https://github.com/allenai/document-qa
Resources

• Dialogue system

MultiWOZ (2018) - for goal-driven dialogue system
http://dialogue.mi.eng.cam.ac.uk/index.php/corpus/
DailyDialog Dataset (2017)
http://yanran.li/dailydialog
DeepPavlov - open-source library for dialogue systems
https://deeppavlov.ai/
KVRET - multi-turn, multi-domain, task-oriented dialogue dataset
Resources

• Recommendation system

Amazon product review
http://jmcauley.ucsd.edu/data/amazon/

Case Recommender - recommender tool
https://github.com/caserec/CaseRecommender

MyMediaLife - recommender system library
http://www.mymedialite.net/

LIBMF - a matrix-factorization library for recommender system
https://www.csie.ntu.edu.tw/~cjlin/libmf/
Resources

• Text mining and text analytics

GATE - general architecture for text engineering
https://gate.ac.uk/
OpenNLP - Apache OpenNLP library
https://opennlp.apache.org/
LingPipe - tool kit for processing text
http://alias-i.com/