WIKIPEDIA BASED INFORMATION EXTRACTION SYSTEM

(TEMPLATE-FILLING)

CS 6320 | NATURAL LANGUAGE PROCESSING | FALL 2018

TEAM PHOENIX:
ROMI PADAM (rkp 170230)
PRAHALYA REDDY KALUVA (pxk172630)

Table of Contents -

Problem Description

The purpose of this project is to extract information from unstructured data. Wikipedia provides a huge dataset to work with as it contains diverse set of terms and complex abstraction of common knowledge. We took set of Wikipedia pages on famous people in technology for the task of proper classification of entities, assignment of entities into roles and relations and drawing inferences. We represented the information as templates consisting of fixed sets of slots.

Proposed Solution

We used template filling approach to find information from the Wikipedia corpus using python scripts and then filled the slots in the associated templates with the extracted information. The slots contain text segments extracted directly from the text and concepts that have been inferred from the text via additional processing. We divided this major task into four subtasks which are outlined below -

Task 1: Created a set of templates and slots.

For this project, we identified semantic entities and events , and created a total of 12 information templates and 40 slots.

Template 1
Key_People
- Key_Person_Name
- Birth_Place
- Birth_Date
- Nationality

Template 2
Family
- Key_Person_Name
- Ancestry
- Father
- Mother
- Children
- Relatives

Template 3
Occupation
- Key_Person_Name
- Work

- Board_Member_Of

Template 4

Education

- Key_Person_Name
- Year
- Degree
- Discipline
- University

Template 5

Career

- Key_Person_Name
- Company
- Title
- Number_Of_Years

Template 6

Health

- Key_Person_Name
- Disability_Flag
- Disability
- Illness

Template 7

Financial_Status

- Key_Person_Name
- Salary
- Net_Worth
- Possessions

Template 8

Recognition

- Key_Person_Name
- Year
- Accolade
- Patents
- Tributes

Template 9

Philanthropic_Endeavours

- Key_Person_Name
- Charitable Organization
- Amount_Donated

Template 10

Death

- Key_Person_Name
- Place_Died

Date_Died
Cause_Of_Death
Resting_Place

Template 11
Publications
Key_Person_Name
Books
Papers
Films

Template 12
Innovations
Key_Person_Name
Entity
Year
Cofounders

Task 2: Created a corpus of natural language statements.

Wikipedia, in particular, is a rich source of textual data and contains vast collection of knowledge. We built a corpus from the set of English Wikipedia articles, which are freely and conveniently available online.

In order to build the corpus, we used Wikipedia-API, a python wrapper for Wikipedia's API. We wrote a python script scraper.py to built the corpus by simply scraping the data and removing References section. Our corpus comprises of 85,000 words and is built for following famous people in technology -

- Steve Jobs
- Tim Cook
- Steve Wozniak
- Jeff Bezos
- Mark Zuckerberg
- Bill Gates
- Stephen Hawking
- Larry Page
- Marc Benioff
- Paul Allen
- Jerry Yang
- Elon Musk
- Steve Ballmer
- Tim Berners-Lee
- Kevin Systrom
- Sundar Pichai

- Evan Spiegel
- Alexander Fleming
- Bjarne Stroustrup
- Jack Dorsey
- Bob Iger

A snippet from the corpus -
“Steven Paul Jobs (; February 24, 1955 – October 5, 2011) was an American business magnate and investor. Steve Jobs was the chairman, chief executive officer (CEO), and co-founder of Apple Inc. ; chairman and majority shareholder of Pixar; a member of The Walt Disney Company's board of directors following its acquisition of Pixar; and the founder, chairman, and CEO of NeXT. …”

Task 3: Extracted NLP based features from corpus.

We implemented a deeper NLP pipeline to extract the following NLP based features from the natural language statements.

- Tokenized the corpus into sentences and words.
  Using NLTK package, we segmented the entire corpus into sentences and then tokenized the sentences into words.

- Lemmatized the words to extract lemmas as features.
  Using NLTK package, we lemmatized the corpus to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

- Part-of-speech(POS) tagged the words to extract POS tag features.
  Using NLTK package, we classified words into their part-of-speech and label them with appropriate POS tag.

- Performed dependency parsing to identify parsed tree based patterns as features.
  Using Stanford dependency parser, we analyzed the grammatical structure of the sentences and established relationships between head words and words that modify those heads.

- Used WordNet to extract hypernyms, hyponyms, meronyms and holonyms as features.

Task 4: Filled templates from corpus.

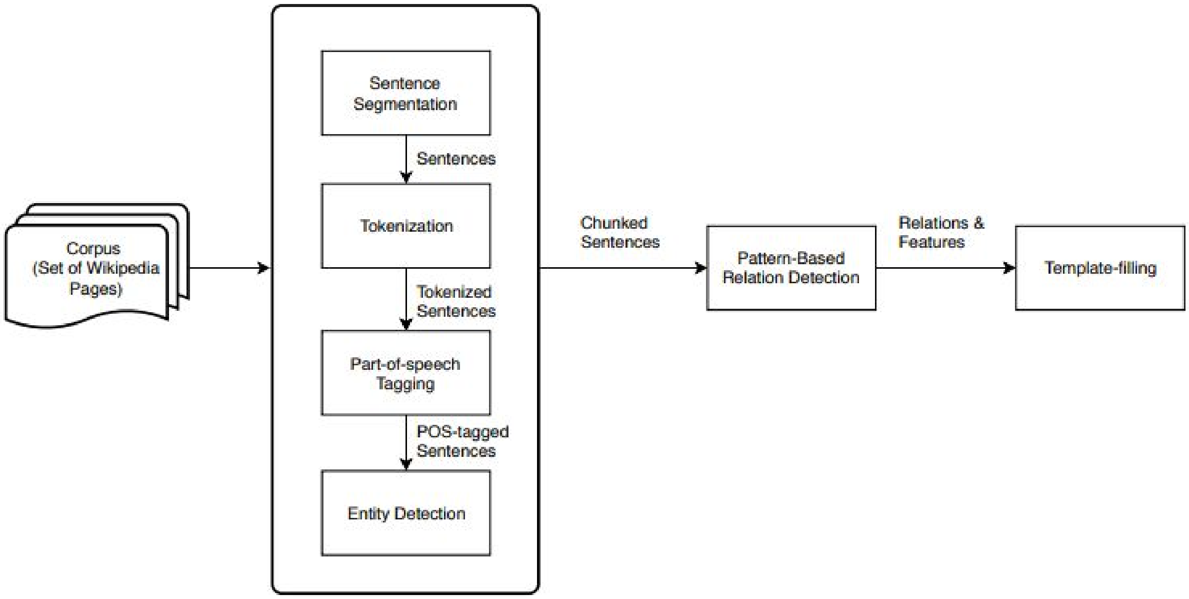
We implemented a heuristic based approach to extract filled information templates from the corpus of natural language statements.

We ran the deeper NLP pipeline (from Task 3) on the entire corpus and then, using a python script and pattern-based information extraction methods, extracted features specific to the templates.

Programming Tools Used

- Python 3.6.1.
- Wikipedia-API (Python wrapper for Wikipedia's API).
- Natural Language ToolKit (NLTK).
- Stanford NER Tagger.
- Anaphora Resolution.
- Stanford Dependency Parser.
- Geotext.

Architectural Diagram



Above architectural diagram can be defined as follows:

- Scrapped data related to some of the famous technology people from the Wikipedia pages using the Wikipedia API.

- Preprocessed the data using NLTK package features. Chunking was mainly performed for entity detection. We also used RegexpParser chunker to identify entities using patterns and GeoText to identify cities and countries.
- We used Pronoun Anaphora for reference resolution. Interrogated the corpus to extract phrases matching a particular context and a sequence of parts-of-speech tags. We used Anaphora for reference resolution. Used Wordnet to find breadth of sentences for a given pattern.
- Finally, filled the slots in templates with texts and information extracted from above steps.

Results & Error Analysis

Template 1 - Key_People

Template -

Key_People

Key_Person_Name
Birth_Place
Birth_Date
Nationality

Example -

Snippet from corpus -

“Paul Gardner Allen (January 21, 1953 – October 15, 2018) was an American business magnate, investor, software engineer, humanitarian, and philanthropist. Allen was born on January 21, 1953, in Seattle, Washington, to Kenneth Sam Allen and Edna Faye (née Gardner) Allen.”

Output -

Key_People(Allen, January 21 , 1953, Seattle,)

Template 2 - Family

Template -

Family

Key_Person_Name
Ancestry
Father
Mother
Children
Relatives

Example -
Snippet from corpus -
“Bill Gates' ancestry includes English, German, Irish, and Scots-Irish.Gates was born in Seattle, Washington, on October 28, 1955. Bill Gates is the son of William H. Gates Sr. (b. 1925) and Mary Maxwell Gates (1929–1994). Bill and Melinda Gates have said that they intend to leave their three children $10 million each as their inheritance. Gates has one older sister, Kristi (Kristianne), and a younger sister, Libby.”

Output -
Family(Bill, ['English', 'German', 'Irish'], William, Mary, three, [Kristi, Libby])

Template 3 - Occupation
Template -
Occupation

Key_Person_Name
Work
Board_Member_Of

Example -
Snippet from corpus -
“Steven Paul Jobs (; February 24, 1955 – October 5, 2011) was an American business magnate and investor. Steve Jobs was the chairman, chief executive officer (CEO), and co-founder of Apple Inc. ; chairman and majority shareholder of Pixar; a member of The Walt Disney Company's board of directors following its acquisition of Pixar; and the founder, chairman, and CEO of NeXT. Jobs was a board member at Gap Inc. from 1999 to 2002.”

Output -
Occupation(Jobs, [investor, chairman, chief executive officer, shareholder] , Gap Inc.)

Template 4 - Education
Template -
Education

Key_Person_Name
Year
Degree
Discipline
University

Example -
Snippet from corpus -
“Tim Cook earned a Bachelor of Science (B.S.) in industrial engineering from Auburn University in 1982, and his Master of Business Administration (MBA) from Duke University's Fuqua School of Business in 1988.”
Output -
Education(Tim, 1988, ['Bachelor of Science', 'Master of Business Administration'], ['industrial engineering'], [Auburn University, Duke University])

Template 5 - Career
Template -
Career

Key_Person_Name
Company
Title
Number_Of_Years

Example -
Snippet from corpus -
“After graduating from Auburn University in 1982, Cook spent 12 years in IBM's personal computer business, ultimately serving as the director of North American fulfillment.”

Output -
Career(Cook, IBM’s , director, 12)

Template 6 - Health
Template -
Health

Key_Person_Name
Disability_Flag
Disability
Illness

Example -
Snippet from corpus -
“Page explained that Larry Page has been suffering from a vocal cord issue for 14 years, and, as of his May 2013 post, doctors were unable to identify the exact cause. In October 2013, Business Insider reported that Page's paralyzed vocal cords are caused by an autoimmune disease called Hashimoto's thyroiditis, and

prevented him from undertaking Google quarterly earnings conference calls for an indefinite period.”.

Output -
“Health(Larry Page, Y, , a vocal cord issue for 14 years , and , as of his May 2013 post , doctors were unable to identify the exact cause .)”

Template 7 - Financial_Status
Template -
Financial_Status

Key_Person_Name
Salary
Net_Worth
Possessions

Example -
Snippet from corpus -
“As of 2014, Musk's annual salary is one dollar, similar to that of Steve Jobs and other CEOs; the remainder of his compensation is in the form of stock and performance-based bonuses. After realizing the site "pravda. com" is used by the Ukrainian Internet newspaper Ukrayinska Pravda, Musk bought the site pravduh. com on May 25, 2018.”

Output -
Financial_Status(Musk, one dollar, $22.8 billion ,bought the site pravduh)

Template 8 - Recognition
Template -
Recognition

Key_Person_Name
Year
Accolade
Patents
Tributes

Example -
Snippet from corpus -
“Bezos was awarded an honorary doctorate in science and technology from Carnegie Mellon University in 2008.”
“Allen held 43 patents from the United States Patent and Trademark Office.”
“The Starmus III Festival in 2016 was a tribute to Stephen Hawking.”

Output -
Recognition(Bezos, 2008, honorary doctorate, ,)
Recognition(Allen, , , 43,)
Recognition(Stephen Hawking, 2016, , ,Starmus III)

Template 9 - Philanthropic_Endeavors
Template -
Philanthropic_Endeavours
    Key_Person_Name
    Cause
    Charitable Organization

Example -
Snippet from corpus -
“In February 2007, Yang and his wife gave $75 million to Stanford University, their alma mater, $50 million of which went to building the "Jerry Yang and Akiko Yamazaki Environment and Energy Building", a multi-disciplinary research, teaching and lab building designed with sustainable architecture principles. “

Output -
Philanthropic_Endeavors(Yang, ['Stanford University'], $ 75 million)

Template 10 - Death

Template -
Death
    Key_Person_Name
    Place_Died
    Date_Died
    Cause_Of_Death
    Resting_Place

Examples -
Example 1:
Snippet from corpus:
“Steve Jobs died of respiratory arrest related to the tumor at age 56 on October 5, 2011. Steve is buried in an unmarked grave at Alta Mesa Memorial Park, the only nonsectarian cemetery in Palo Alto.”

Output:
Death(Steve, , , respiratory arrest related to the tumor, Alta Mesa Memorial Park)

Example 2:

Snippet from corpus:
“Stephen Hawking died at his home in Cambridge, England, early in the morning of 14 March 2018, at the age of 76.”

Output:
Death(Stephen, Cambridge, 14 March 2018, ,)

Template 11 - Publications

Template -
Publications

Key_Person_Name
Books
Papers
Films

Example -
Snippet from corpus -
“1999: Pirates of Silicon Valley – a TNT film directed by Martyn Burke. Wozniak is portrayed by Joey Slotnick while Jobs is played by Noah Wyle.”
“The Road Ahead, written by Gates with Microsoft executive Nathan Myhrvold and journalist Peter Rinearson, was published in November 1995.”
“Hawking continued his writings for a popular audience, publishing *The Universe in a Nutshell* in 2001.”
Output -
Publications(Martyn Burke, , , Pirates of Silicon Valley)
Publications([Gates, Nathan Myhrvold, Peter Rinearson] , Road Ahead, ,)
Publications(Hawking, , Universe in a Nutshell,)

Template 12 - Innovations

Template -
Innovations

Key_Person_Name
Entity
Year
Cofounders

Example -
Snippet from corpus -

“Berners-Lee joined the board of advisors of start-up State. com, based in London. As of May 2012, Berners-Lee is president of the Open Data Institute, which Tim Berners-Lee co-founded with Nigel Shadbolt in 2012.”

Output -
Innovations(Tim, Open Data Institute, 2012, Nigel Shadbolt)

Precision Matrix -

Template #	Template Name	Results		Precision
		True Positive	False Positive	
1	Key_People	11	5	68.75
2	Family	10	6	62.5
3	Occupation	5	8	38.46153846
4	Education	17	8	68
5	Career	6	9	40
6	Health	9	7	56.25
7	Financial_Status	7	7	50
8	Recognition	4	10	28.57142857
9	Philanthropic_Endeavors	19	10	65.51724138
10	Death	8	3	72.72727273
11	Publications	5	11	31.25
12	Innovations	4	8	33.33333333
				51.28006787

Problems Encountered & Resolution

1. While creating corpus from Wikipedia pages, our scraping process was extracting all the irrelevant sections like References, Future Reading, etc. We partitioned the corpus to remove these sections.

Pending Issues

1. NLTK doesn’t perform correct POS tagging in certain of the cases. For e.g. Hawking is identified as a verb VBN.

2. While chunking data, NLTK doesn’t classify entities correctly into PERSON, ORGANISATION and GPE.

Potential Improvements

1. Use a better POS tagger which correctly labels part-of-speech tags.
2. Identify entities (organization, person and location) with better precision.