CS5560 Knowledge Discovery and Management

Problem Set (PS-1A)

June 5, 2017

Student Name: Megha Nagabhushan

ClassID: 15

We are supposed to build a knowledge graph for the following input (text data).

1. Describe your knowledge about knowledge graph.
2. Why do we want to build such a knowledge graph?
3. What steps are required? Show your own workflow for this task.
4. What are the challenges?
5. Draw a knowledge graph for the given data.

> CHICAGO (AP) – Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

① Knowledge graphs are graphs of interlinked entities and their attributes. Knowledge graphs have an ontology as its schema.

② a) The knowledge graph are made of interlinked entities. It provides a straightforward approach where users can understand these entities by directly providing key facts. (For example, information about Kansas City)

b) For questions for which answers are not directly available, the users can use effecient exploitation techniques and locate the right portion of the knowledge graph to get the answer.

(For example, what were the 5 coldest years in Kansas City in the last 15 years?)

c) linking multiple knowledge graphs might help in decision making in large organization.

So by building a knowledge graph.

→ we are improving the learning curve.

→ graph is the most flexible data structure. So, quering and information retrieval will be easy.

→ minimizes redundancy.

→ improves decision-making.

③ <u>Step: 1</u> Natural Language Processing.

<u>Annotation</u> → <u>Lemmatization</u> → <u>POS Tagging</u> → <u>Named Entity Recognition</u>.

CHICAGO → CHICAGO → NNP → LOCATION

-LRB- → -lrb- → -LRB-

AP → AP → NER → ORGANIZATION.

-RRB- → -rrb- → -RRB-

— → - → :

Citing → cite → VBG

high → high → JJ

fuel → fuel → NN

prices → price → NNS

, → , → ,

United → United → NNP → ORGANIZATION

Annotation → Lemmatization → POS Tagging → Named Entity Recognition

Airlines → Airlines → NNPS → ORGANIZATION

said → say → VBD

Friday → friday → NNP → DATE

it → it → PRP

has → have → VBZ

increased → increase → VBN

fares → fare → NNS

by → by → IN

$ → $ → $ → MONEY

6 → 6 → CD → MONEY

per → per → IN

round → round → NN

trip → trip → NN

on → on → IN

flights → flight → NNS

to → to → TO

some → some → DT

cities → city → NNS

also → also → RB

served → serve → VBD

by → by → IN

lower-cost → lower-cost → JJ

carriers → carrier → NNS

. → . → .

American → American → NNP → ORGANIZATION

Airlines → Airlines → NNPS → ORGANIZATION

, → , → ,

a → a → DT

unit → unit → NN

AMR → AMR → NNP

, → , → ,

immediately → immediately → RB

matched → match → VBD

the → the → DT

move → move → NN

, → , → ,

spokesman → spokesman → NN

Tim → Tim → NNP → PERSON

Wagner → Wagner → NNP → PERSON

said → say → VBD

. → . → .

United → United → NNP → ORGANIZATION

, → , → ,

a → a → DT

unit → unit → NN

of → of → IN

UAL → UAL → NNP → ORGANIZATION

, → , → ,

said → say → VBD

the → the → DT

increase → increase → NN

took → take → VBD

effect → effect → NN

Thursday → Thursday → NNP → DATE.

night → night → NN → TIME

and → and → CC

applies → apply → VBZ

to → to → TO

most → most → JJS.

routes → route → NNS

where → where → WRB

it → it → PRP

competes → compete → VBZ

against → against → IN

discount → discount → NN

carriers → carrier → NNS

, → , → ,

such → such → JJ

as → as → IN

Chicago → Chicago → NNP → LOCATION.

to → to → TO

Dallas → Dallas → NNP → LOCATION

and → and → CC

Atlanta → Atlanta → NNP → LOCATION

and → and → CC

Denver → Denver → NNP → LOCATION

to → to → TO

San → San → NNP → LOCATION

Francisco → Francisco → NNP → LOCATION

, → , → ,

Los → Los → NNP → LOCATION

Angeles → Angeles → NNP → LOCATION

and → and → CC

New → New → NNP → LOCATION

. → . → .

Natural Language processing refers to the use and ability of systems to process sentences in a natural language such as English, rather than in a specialized artificial computer language such as C++.

It involves the following steps:

① Tokenization: tokenzing ~~words~~ text into words.

② Lemmatization/Stemming: ~~St~~

Stemming is the process of reducing words to their stem, base or root form - generally a written word

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single term.

c) <u>POS Tagging / Chunking</u>

Parts-of-Speech Tagging. Used a pre-defined Tagset for this process.

d) <u>Parsing</u>.

It is the process of analyzing a sentence by taking each word and determining its structure from its constituent parts. It uses two components: a parser and a grammar.

e) <u>Name Entity Recognition</u>.

It is a subtask of ~~Named-Entity-Recognition~~ information extraction that seeks to locate and classify named entities in text into pre-defined categories such as names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages etc.

f) Co-Reference Resolution

It is the task of finding all expressions that refer to the same entity in a text : It is important step for higher NLP tasks like document summarization, question answering and information extraction.

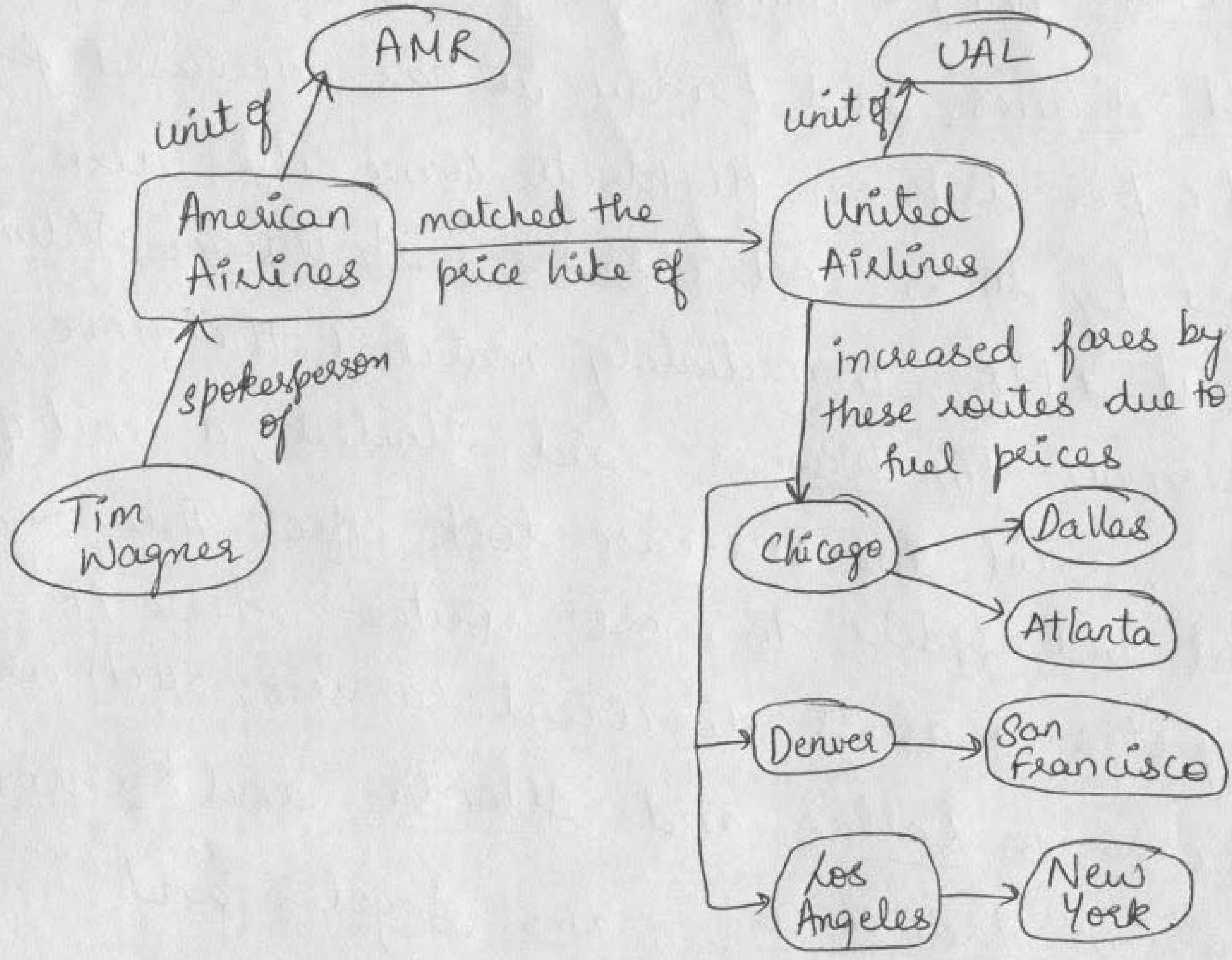
Step 2: Information Retrieval and Extraction.

CHICAGO (AP) - Citing high fuel prices,
United Airlines said Friday it has increased fares
by $6 per trip on flights to some cities also
served by lower-cost carriers. American Airlines,
a unit AMR, immediately matched the move,
spokesman Tim Wagner said. United, a unit of
UAL, said the increase took effect Thursday
night and applies to most routes where it
competes against discount carriers, such as
Chicago to Dallas and Atlanta and Denver
to San Francisco, Los Angeles and
New York

Step 3: Topic discovery

Organization	People	Places
United Airlines American Airlines AMR UAL	Tim Nagner.	Chicago Dallas Atlanta Denver San Francisco Los Angeles New York

Step: 4 Knowledge Graph Constuction.



4) Challenges of Knowledge Graph.

There is no universal algorithm to generate these knowledge graphs. As a result the knowledge graphs from different sources will adhere to different formats.

Completeness of the knowledge Graph. There can be a lot of missing information.

Correctness of the knowledge Graph. The knowledge Graph may not have the accurate data.

Public knowledge bases may not have the accurate data.

⑤ Draw knowledge Graph.

P.T.O.