ME AND MY RESEARCH

周双双
総合研究会

Contents

- Self-introduction
- Research Progress

Who I am

- Name--------周双双(Zhou Shuangshuang しゅう　そうそう)
- What does “双” mean in Chinese?

  1st– two, a pair of

  2nd--Good things should be in pairs
- Why to choose it as my name?

  Numbers related to my birth are all even number

  birthday, birth weight and so on

Where I come from

- The Most Northeastern part of China, Heilongjiang Province(黒竜江省),Qiqihar(チチハル)

Hometown 18 years

1988.8.8

Northeastern University
(Shenyang 4+2 years)

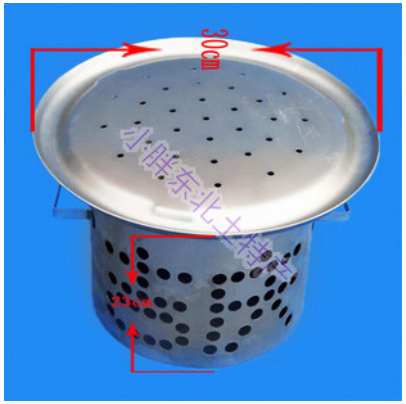
Tohoku University
Research student (6 months)
Doctoral candidate (Just two months Begin!!)

Where I come from

Ice-lantern Festival (Very Cold !!! -30° c ++)

Yakiniku （烧肉——烤肉）

Special Brazier



Hobbies

- **Travelling**
- **Sports**

Basketball, Swimming, Badminton, Mountain Climbing, Mahjong(麻雀)

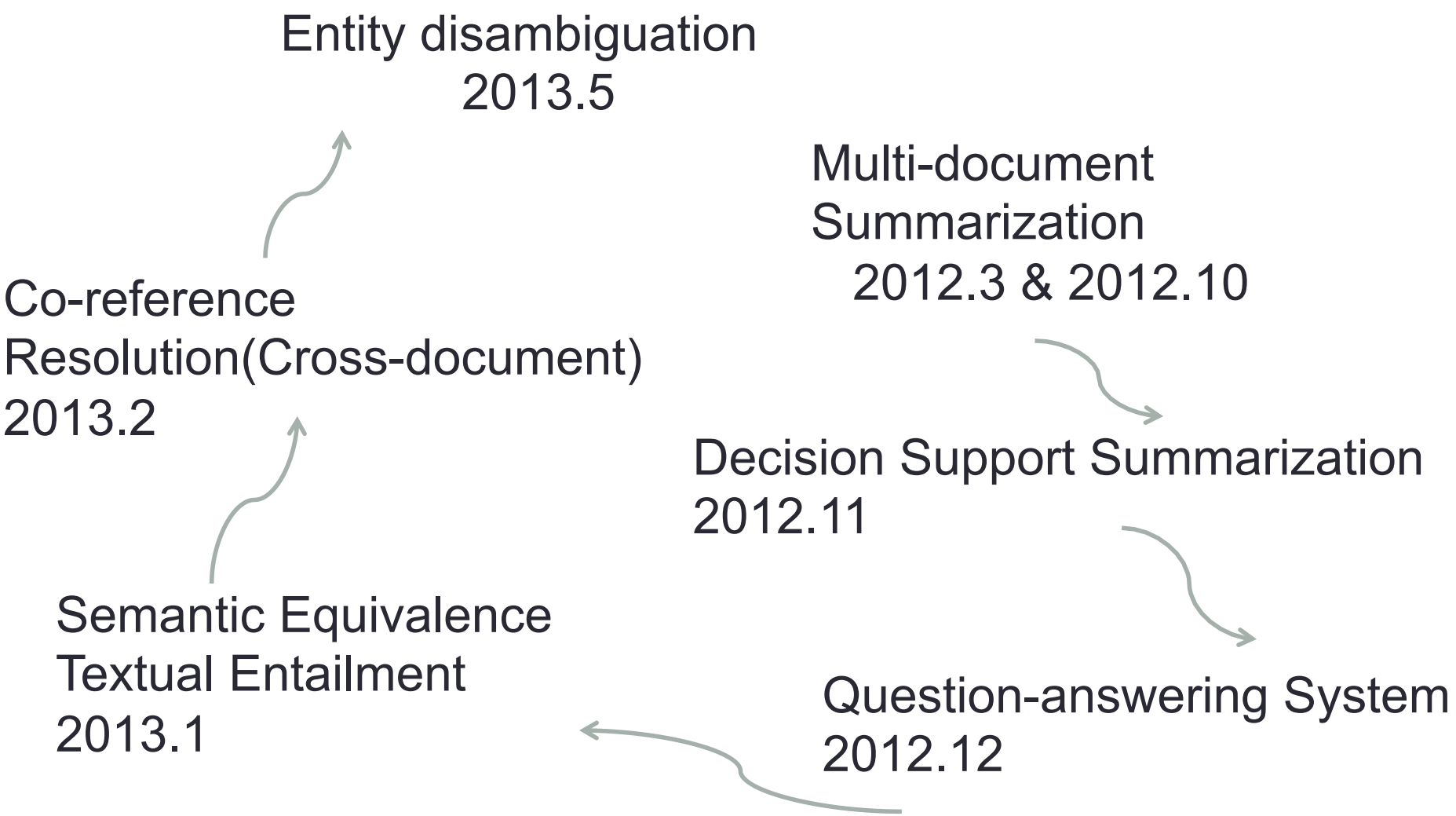
- **Music**

Listening cover wide kinds of music
Chorus two-year chorus experience in university
Karaoke → KTV

- **Dramas**

Korean, Japan, American, Thai…

The way of Research Theme

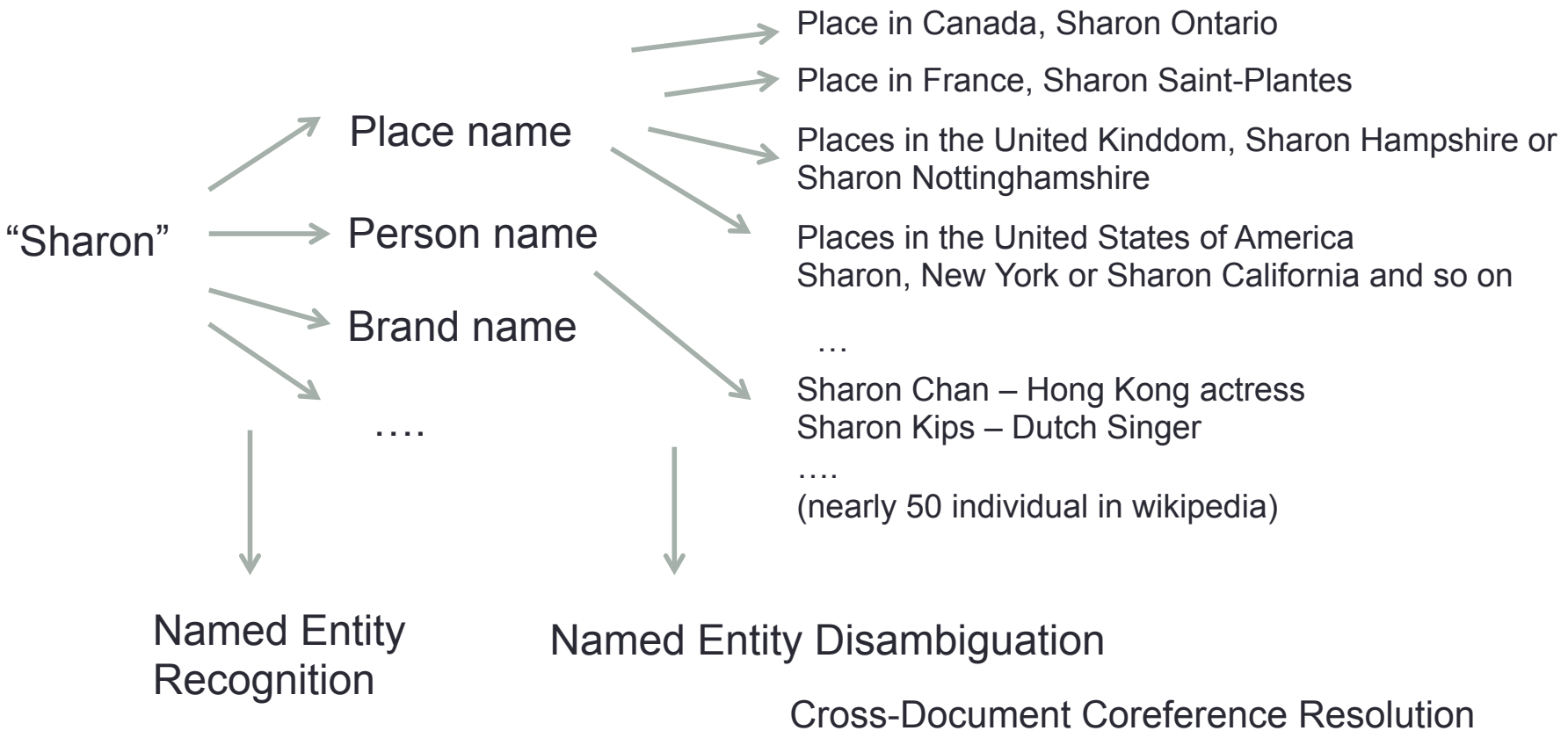


Research Progress

- Entity disambiguation vs. Word sense / concept disambiguation
- Examples:

Research Progress

- Entity disambiguation vs. Word sense disambiguation
- Examples:



Entity Disambiguation

Example:

John Williams

Richard Kaufman goes a long way back with **John Williams**. Trained as a classical violinist, Californian Kaufman started doing session work in the Hollywood studios in the 1970s. One of his movies was Jaws, with **Williams** conducting his score in recording sessions in 1975...

John Williams	author	1922-1994
J. Lloyd Williams	botanist	1854-1945
John Williams	politician	1955-
John J. Williams	US Senator	1904-1988
John Williams	Archbishop	1582-1650
John Williams	composer	1932-
Jonathan Williams	poet	1929-

Processing

Source document

Knowledge base

personal context

Represent entity with features based on its context

Ranking methods

Candidates list → Final answer

GeoNLP : Geo entity disambiguation and tagging in Social Media

- Motivation
- Social Media

-- Huge User Group
-- Potential information acquisition data source
-- Can be grouped or tagged with similar content
-- Bias temporal and spatial information, user interest related, real-time

- Geo entity disambiguation and tagging

-- geolocate event
-- Inform location information
-- Complement other NLP applications in multiple ways

- Motivation --- Application aspects
- Travel Guiding

“Geotagging Tweets Using Their Content” , Sharon Paradesi,2011

- Weather Observing

“Typonym-based Geotagging for Observing Precipitation from Social and Scientific Data Streams” , Asanobu Kitamoto et al. 2012

- Disaster Map

"GeoNLP: Toward Intelligent Geo-Tagging for Natural Language Text”, Asanobu KITAMOTO, Takeshi SAGARA, Masatoshi ARIKAWA, 2011

GeoNLP : Geo entity disambiguation and tagging in social media

Task definition

Tweets stream

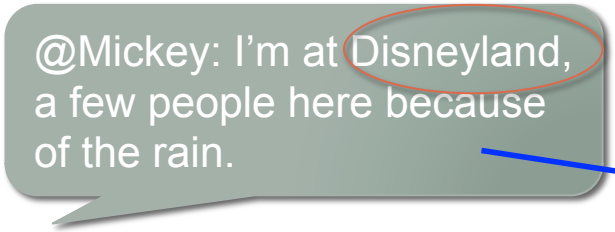
@Mickey: I'm at Disneyland, a few people here because of the rain.

@食事姫: 本町のラーメン太郎は最高。

…

GeoNLP : Geo entity disambiguation and tagging in social media

Task definition



Weather: rain → real-time weather inform
A few people → easy to play now

Tweets stream

restaurant recommend

Valuable Information

GeoNLP : Geo entity disambiguation and tagging in social media

Task definition

Tweets stream

@Mickey: I'm at Disneyland, a few people here because of the rain.

Which Disneyland? HongKong? Tyoko?

Weather: rain → real-time weather inform

A few people → easy to play now

@食事姫: 本町のラメン太郎は最高。

Honmachi in Sendai? Or in Aomori? Or?

restaurant recommend

…

Tasks need to do

Valuable Information

Geo entity disambiguation and tagging in social media

Task definition

@Mickey: I'm at Disneyland, few people here because of the rain.

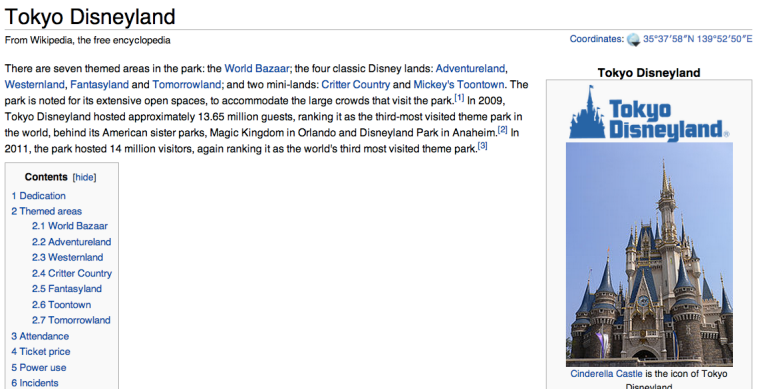
Tweets stream

@食事姫: 本町のラメン太郎は最高。

…

Tokyo Disneyland	Lat/Lng: 35.6328/139.8805
Disneyland Park(Paris)	Lat/Lng: 48.873/2.777
Hong Kong Disneyland	Lat/Lng: 22.5/114.05
Shanghai Disneyland Park	Lat/Lng: 31.144/121.657

Geo entity Disambiguation


Tokyo Disneyland

From Wikipedia, the free encyclopedia

Coordinates: 35°37′58″N 139°52′50″E

There are seven themed areas in the park: the World Bazaar; the four classic Disney lands: Adventureland, Westernland, Fantasyland and Tomorrowland; and two mini-lands: Critter Country and Mickey's Toontown. The park is noted for its extensive open spaces, to accommodate the large crowds that visit the park.[1] In 2009, Tokyo Disneyland hosted approximately 13.65 million guests, ranking it as the third-most visited theme park in the world, behind its American sister parks, Magic Kingdom in Orlando and Disneyland Park in Anaheim.[2] In 2011, the park hosted 14 million visitors, again ranking it as the world's third most visited theme park.[3]

Contents [hide]

1 Dedication
2 Themed areas
2.1 World Bazaar
2.2 Adventureland
2.3 Westernland
2.4 Critter Country
2.5 Fantasyland
2.6 Toontown
2.7 Tomorrowland
3 Attendance
4 Ticket price
5 Power use
6 Incidents

Tokyo Disneyland

Cinderella Castle is the icon of Tokyo

Geo entity tagging

@Mickey: I'm at Disneyland (Tokyo Disneyland Lat/lng: 35.6328/139.8805), few people here because of the rain.

GeoNLP : Geo entity disambiguation and tagging in social media

- Work need to do

- Automatically Geo entity recognition

Characteristic & Challenges :

- Geo/Non-Geo entities distinct

-- Places named after people
(e.g. Sharon)
-- Common noun phrases(Paradesi 2011)
(e.g. Love, Need)
-- Small location name(Ji 2011)
(e.g. “Del Rio” is part of “Gruene, Texas”)

- Solution candidates:

Start-of-the-art information extraction tools, such as Wikipedia Miner, DBPedia Spotlight, which can exact entities and their description

- Problems remain:

-- Potential false negative
-- Lack of database contains small location name , especially store names and so on.

GeoNLP : Geo entity disambiguation and tagging in social media

- Tasks need to do

- Automatically Geo entity recognition

- Geo entity disambiguation and tagging

Characteristic & Challenges:
-- Short Text(e.g. Twitter, less than 140 words)
-- Local lexical context is sparse
-- informal write style
-- Personal utterance, containing imprecise, subjective and ambiguous expressions
-- Difficult to both human and machine
Solution candidate:
Leverage user interest(Elizabeth L.Murnane et al 2013)
- Document context → Personal knowledge context
- An ambiguous entity has the mostly concept related to user's interest
- Building a model of user interest from external structured semantic data(Wikipedia user's edit log)
Problems remain:
How about users who don't have edit log?

Current Studies Survey

- Travel Guiding

"Geotagging Tweets Using Their Content" , Sharon Paradesi,2011

- Identify the locations referenced in a tweet and show relevant tweets to a user based on that user's location
- TwitterTagger, geotags tweets in near real-time and shows tweets related to surrounding areas
- Two step disambiguation: Geo/Non-Geo disambiguation, Geo/Geo disambiguation
- USGS(United States Geological Survey)location database as external data source, provides location information
- Evaluation : Split into true positives and false positives manually from a random sample of geotagged tweets

Defects:

- Unconsidered false negatives
- Tag locations only in the U.S.

Current Studies Survey

- Weather Observing

"Typonym-based Geotagging for Observing Precipitation from Social and Scientific Data Streams" , Asanobu Kitamoto et al. 2012

- Use social data stream to observe weather, because weather is a typical daily conversation topic, observing such as precipitation, wind and so on
- Based on toponym-based geotagging of weather events, collect information about real-time and long-term weather of one place
- GeoNLP API (limit to open) disambiguate and tag location names in tweets(in Japanese)
- Social data streams can be used as complementary data source to scientific data streams

Defects :
- Lack of text understanding method to handle case like this, " Place A is raining, but place B is snowing".
- Distinguish case like this " it is raining" or "it is not raining".
- Simple evaluation(comparing with radar imagery), lack of training dataset.

Geo entity disambiguation and tagging in social media

- Others

Data source input

-- Raw data, derived from Twitter streaming API or Twitter search API

-- Corpus, looking for…

External data source:

-- Specific geo database, Geonames, GeoWordNet, USGS…

-- General knowledge base, Wikipedia, DBPedia…

-- Others knowledge base, YAGO2…

- Other Difficulties

Evaluation

- Lack of labeled data
- Manual evaluation?
  Huge labor cost
- Small place names

Geo entity disambiguation and tagging in social networking

- Next to do

-- Proper Corpus finding, find methods from unlabeled data

"Learning from positive and unlabeled examples", F. Letouzey, F. Denis, and R. Gilleron. 2000.

"Building text classifiers using positive and unlabeled examples". B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. 2003.

"Entity Disambiguation with Type taxonomy", Zhicheng Zheng et al. 2013.(solve problem when lacking of labeled training data)

-- More similar Work investigation

"Toponym Resolution in Text, Annotation, Evaluation and Applications of Spatial Grounding of Place Names", Jochen Lothar Leidner, 2008

"Toponym Resolution in Social Media", Neil Ireson et al. 2010

-- Formalize task definition

-- Simple experiment demo starting

THANKS FOR YOUR ATTENTION