자연어 처리

Language Technology

- 사람이 사용하는 언어를 기계도 이해하고 같이 활용하고 사용할 수 있도록 하는 학문
- Mostly solved : spam detection, Part-of-speech(POS), NER
  - 스팸 탐지, 품사 찾기, 고유명사 찾기
- Making good progress : Sentiment analysis, Coreference resolution, WSD(Word sense disambiguation), Parsing, Translation, Information extraction
  - 지시대명사, 어떤 단어가 어떤 의미로 쓰였는지, 문장 구조, 번역, 정보 추출
- Still really hard : Paraphrase, Summarization task, dialog
  - 질의응답, 문장을 다른 형태로 표현, 대화

mostly solved

Spam detection

Let's go to Agra!

Buy V1AGRA …

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

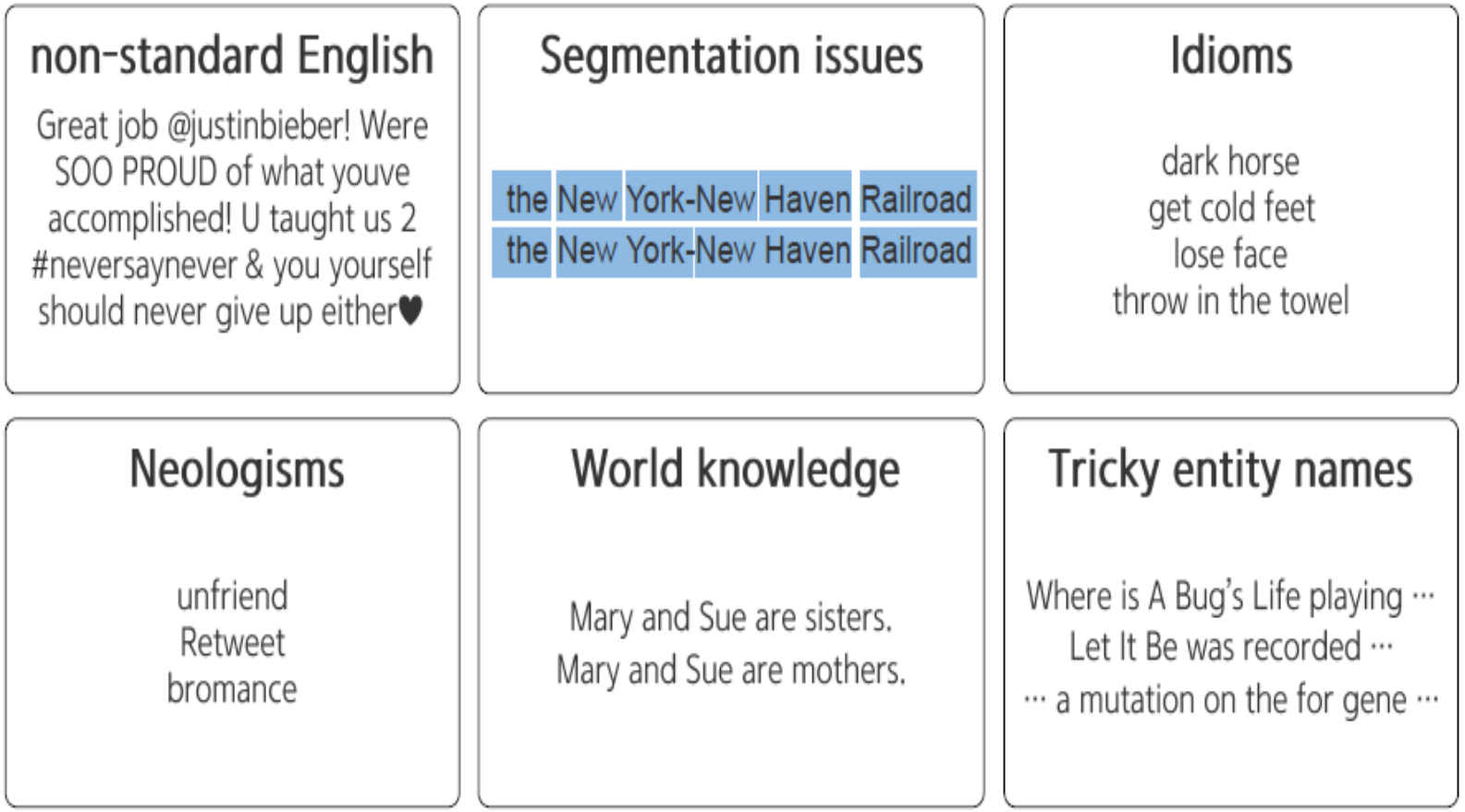
Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

NL Understanding Difficult



- 비문 사용
- 끊어 읽기
- 관용어구
- 신조어
- 지식이 필요
- Entity name (ex, Let it be : 노래 제목)

Question Answering: IBM's Watson

- **Won Jeopardy on February 16, 2011!**

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

→ Bram Stoker

질의 응답

- Factoid questions
  - 단답형
- Complex (narrative) questions
  - 서술형

Information Extraction

Subject : curriculum meeting

Date : January 15, 2012

To : Dan Jurafsky

Event : Curriculum mtg
Date : Jan-16-2012
Start : 10:00am
End : 11:30am
Where : Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

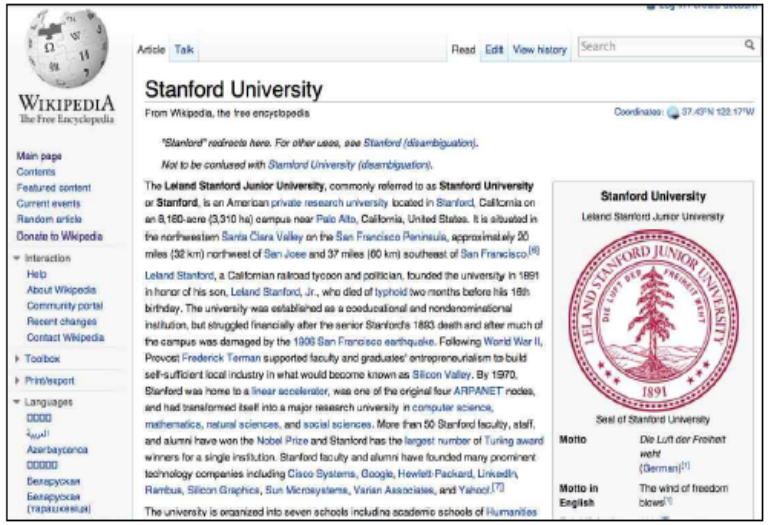
It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry

이메일 정보만 보고 문장 생성하기

Relation Extraction from Text



The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is an American private research university located in Stanford, California … near Palo Alto, California… Leland Stanford…founded the university in 1891

Stanford EQ Leland Stanford Junior University
Stanford LOC-IN California
Stanford IS-A research university
Stanford LOC-NEAR Palo Alto
Stanford FOUNDED-IN 1891
Stanford FOUNDER Leland Stanford

글로부터 자동으로 관계를 알아냄

Information Extraction & Sentiment Analysis

Attributes:

- zoom
- affordability
- size and weight
- flash
- ease of use

Size and weight

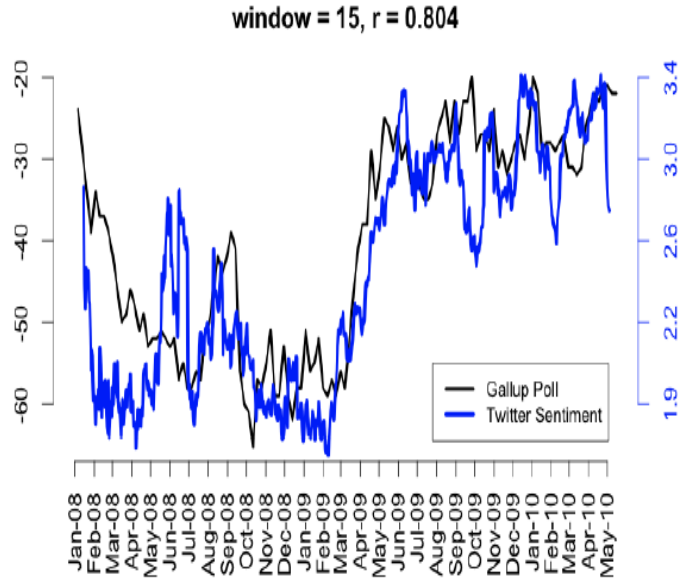
- ✓ ▶ nice and compact to carry!
- ✓ ▶ since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- ✗ ▶ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

사람이 쓴 자연어 문장으로부터 Attributes 별로 분석하고, 좋은 평인지 나쁜 평인지 찾아냄

Positive or negative movie review?

- Unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- This is the greatest screwball comedy ever filmed.
- It was pathetic. The worst part about it was the boxing scenes.

Twitter sentiment vs. Gallup poll of consumer confidence



트위터 상에 수많은 트윗들을 분석하여 설문 조사

Sentiment analysis has many other names

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

Simplest task:

- Is the attitude of this text positive or negative?

More complex:

- Rank the attitude of this text from 1 to 5

Advanced:

- Detect the target, source, or complex attitude types

- 긍정/부정
- 1점에서 5점
- 감정의 대상, 근원, 복잡한 반응을 찾아내는 것

What Makes Sentiment Analysis Difficult

- Subtly
  - 비꼬기
- Thwarted expectations
  - 좋은 내용으로 시작해서 좋지 않은 내용으로 끝남
- Ordering effects
  - 긍정적 단어와 부정적 단어의 순서

Named Entity Recognition (NER)

A very important sub-task: find and classify names in text

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organization

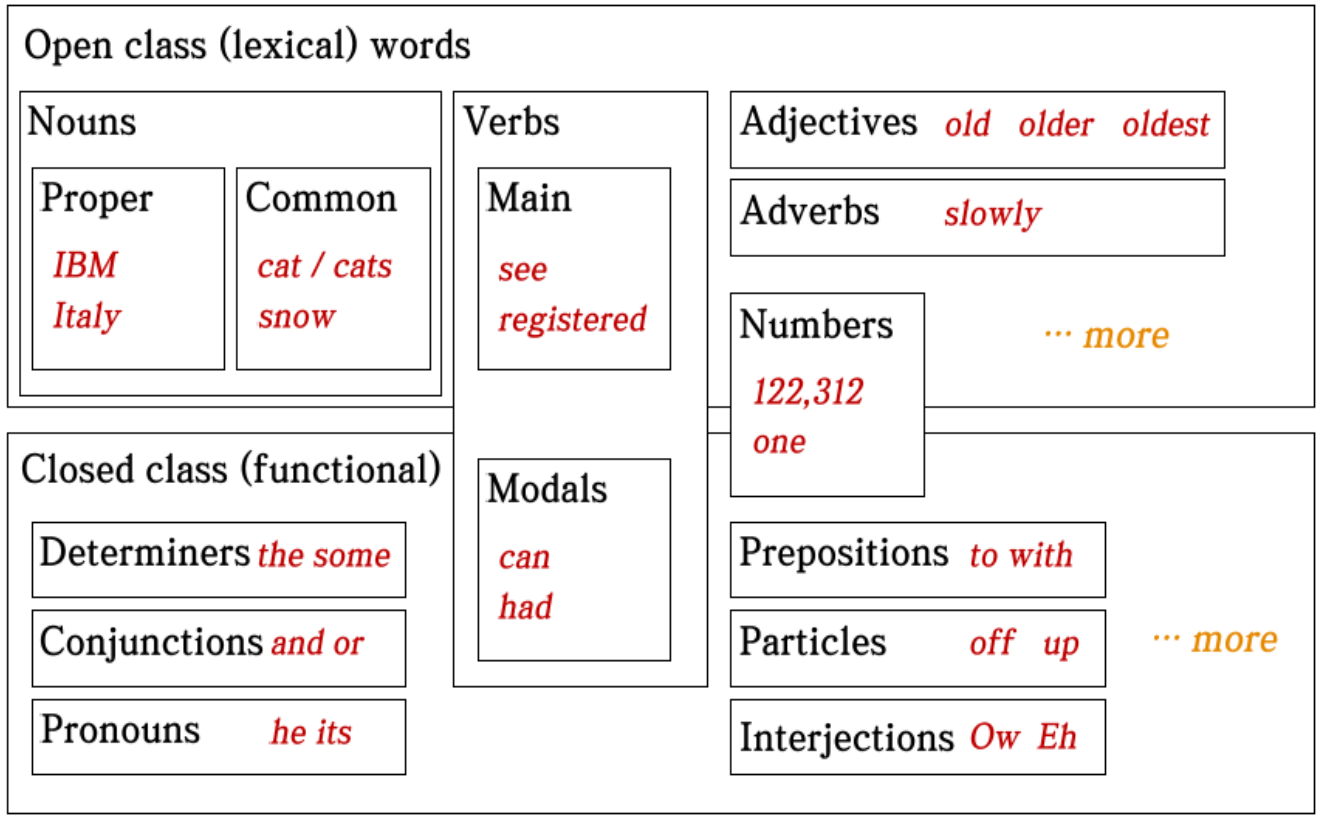
The uses

- Named entities can be indexed, linked off, etc.
- Sentiment can be attributed to companies or products
- A lot of IE relations are associations between named entities
- For question answering, answers are often named entities

주어진 문장에서 고유 명사를 찾고, 고유 명사가 사람인지, 날짜인지, 위치인지, 단체인지 판별

Part-of-Speech Tagging

- Determine the POS tag for a particular instance of a word



문장에서 품사를 찾아내기
- 하나의 단어는 여러 품사를 갖고 있어, 다른 단어들과 관계를 통해 알아내야 함

Parser works out the grammatical structure of sentences

- Parsing resolves structural ambiguity in a formal way
- e.g. I saw a girl with a telescope.

문장이 주어지면
문법적인 구조를
알아내는 과정

Word Meaning and Similarity

- Homonymy(동형이의어)
  - 형태는 똑같지만 뜻은 완전히 다름
- Polysemy
  - 한 단어가 여러 뜻을 가짐
- Synonyms
  - 동의어
- Antonyms

Machine Translation

Fully automatic!

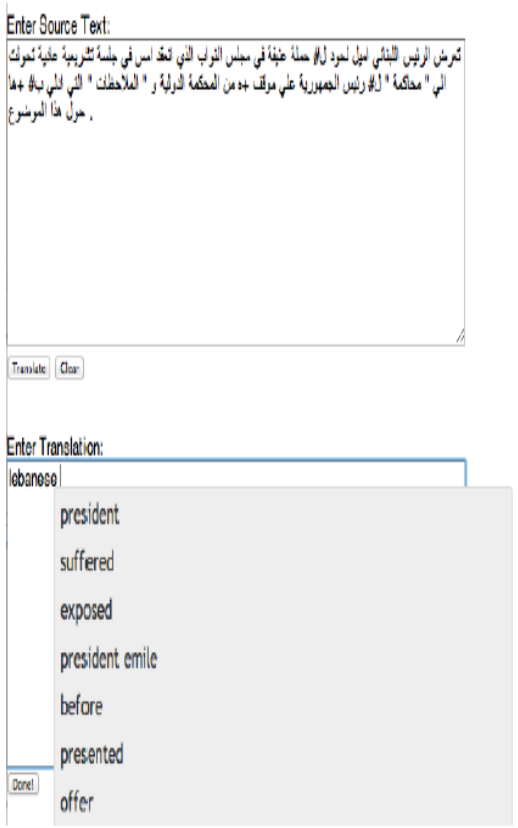
Enter Source Text:

这不过是一个时间的问题.

Translation from Stanford's Phrasal:

This is only a matter of time.

Helping human translator



Recently, machine translation is so successful!

- Google Neural Machine Translation system (GNMT) in 2016/09

Input sentence:	*Translation (PBMT):*	*Translation (GNMT):*	*Translation (human):*
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

사람보다 더 성능이 좋은 번역기

Text Summarization

- **Produce an abridged version of a text that contains information that is important or relevant to a user**

Snippets

- **Single-document summarization**
  - Given a single document, produce abstract, outline, headline

- 아주 긴 하나의 문서를 요약
- 여러 문서를 요약

- **Multiple-document summarization**
  - Given a group of documents, produce a gist of the content
    - A series of news stories on the same event
    - A set of web pages about some topic or question

Text Summarization

- **Generic summarization**
  - Summarize the content of a document
- **Query-focused summarization**
  - Summarize a document with respect to a user query
  - A kind of complex question answering (i.e. summarizing a document to construct the answer)
- **Extractive summarization**
  - Create the summary from phrases or sentences in the source document(s)
- **Abstractive summarization**
  - Express the ideas in the source documents using (at least in part) different words

- Document가 주어지면 그걸 대표하는 문장 하나 혹은 5개 미만의 문장으로 요약
- 유저가 질의를 주면 그것에 기반하여 요약하는 것
- 어떤 문서를 summarization 할 때 주어진 문서 중 중요한 특정 구 혹은 단어를 뽑아가며 요약
- 전체 글을 다 이해한 후에 그 글을 하나로 표현하는 방법