Human Language Technology and its Applications

Ko, Youngjoong

Dept. of Computer Engineering,
Dong-A University

February 7, 2014

Contents

1. Human Language Technology (HLT)

2. Spoken Language Understanding (SLU)
- Spoken Dialogue User Interface for Intelligent Robots and Smart Phones

3. Opinion Mining
- Sentiment Analysis
- Comparison Mining

4. Others
- AAC, Text Mining & Information Retrieval
- Big Data Analysis

Goals of the HLT

Computers would be a lot more useful if they could handle our email, do our library research, talk to us …

But they are fazed by natural human language.

How can we make computers have abilities to handle human language? (Or help them learn it as kids do?)

Levels of the HLT

- **Phonetics/phonology**
- **Morphology**
- **Syntax**
- **Semantics**
- **Pragmatics**
- **Discourse**

Phonetics (음성학) & Phonology (음운론)

- **The study of language sounds, how they are physically formed and systems of discrete sounds**
  - disconnect => dis-k&-'nekt
  - "It is easy to recognize speech."
  - "It is easy to wreck a nice beach."

Morphology (형태론)

- **The study of the sub-word units of meaning**
  - Disconnect
    - Dis: "not", connect: "to attach"
- **Even more necessary in some other language**
  - e.g.) Turkish
    - *uygarlastiramadiklarimizdanmissinizcasina*

    => uygar las tir ama dik lar imiz dan mis siniz casina
  - e.g.) Korean
    - *감기는*

    ⇒감기(명사) + 는(조사)

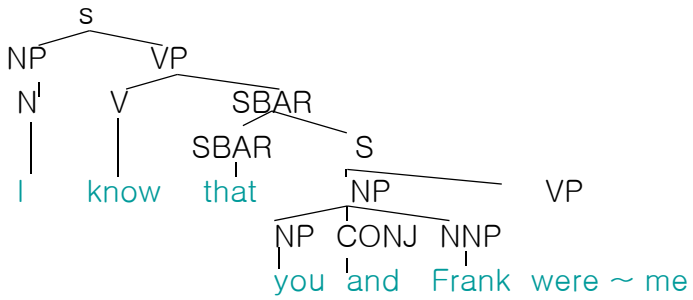
    ⇒감(동사) + 기(명사형어미) + 는(조사)

    ⇒감기(동사) + 는(어말어미)

    ⇒감(동사) + 기는(어말어미)

Syntax (구문론)

- **The study of the structural relationships between words**
  - I know that you and Frank were planning to disconnect me.



Semantics (의미론)

- **The Study of the literal meaning**
  - I know that you and Frank were planning to disconnect me.

ACTION = disconnect

ACTOR = you and Frank

OBJECT = me

Pragmatics (화용론)

- **The Study of how language is used to accomplish goals.**
  - What should you conclude from the fact I said something?
  - How should you react?
    - I'm sorry Dave, I'm afraid I can't do that.
      - ✓ Includes notions of polite and indirect styles

Discourse (담화론)

- **The study of linguistic units larger than a single utterance**
  - The structure of conversations: turn taking, thread of meaning

    Bowman: Open the pod bay doors, Hal.

    Hal: I'm sorry, Dave, I'm afraid I can't do that.

    Bowman: What are you talking about Hal?

    Hal: I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.
  - Story discourse
    - 철수는 어항을 떨어뜨렸다
    - 그는 울고 말았다.

Linguistic Rules

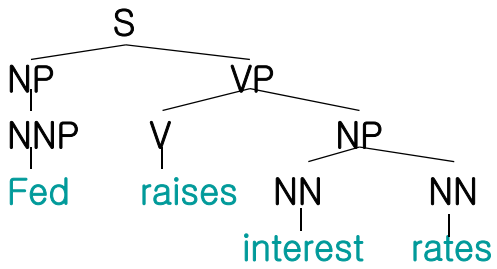
- **e.g) Morphology**
- **To make a word plural, add "s"**
  - dog -> dogs
  - baby -> babies
  - dish -> dishes
  - goose -> geese
  - child -> children
  - fish -> fish

Where are the ambiguities?

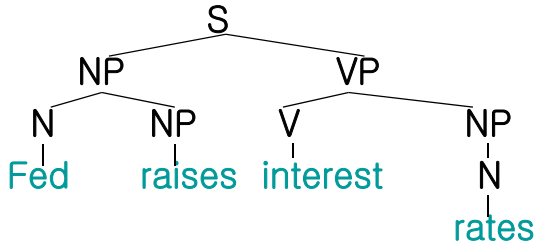
Part-Of-Speech ambiguities						Syntactic Attachment Ambiguities
		VB				
	VBZ	VBZ	VBZ			
NNP	NNS	NNS	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	in effort to control inflation

- Word sense ambiguities:
  - Fed -> "federal agent"
  - Interest -> "a feeling of wanting to know" or "learn more"
- Semantic interpretation:
  - Ambiguities above the word level.

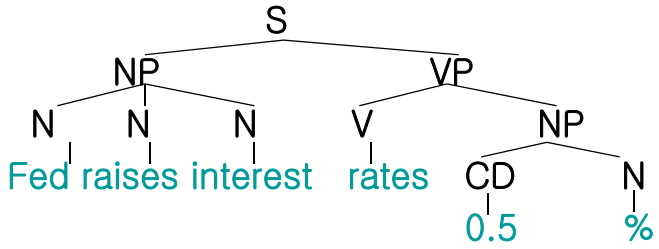
Effects of V/N Ambiguity (1)



Effects of V/N Ambiguity (2)



Effects of V/N Ambiguity (3)



Language Evolves

- **Morphology**
  - We learn new words all the time:
    bioterrorism, cyberstalker, infortainment, thumb candy, energy bar and so on.
- **Part-of-speech**
  - Historically: “kind” and “sort” were always nouns:
    - “I know that sort of men well.”
  - Now also used as degree modifiers
    - “I’m sort of hungry.”

Natural Language Computing is hard because

- **Natural language is:**
  - Highly ambiguous at all levels
  - Complex and subtle
  - Fuzzy, probabilistic
  - Involves reasoning about the world
  - Embedded a social system of people interacting
    - Persuading, insulting and amusing them
    - Changing over time

Probabilistic Models of Language

- **To handle this ambiguity and to integrate evidence from multiple levels we turn to:**
  - Bayesian Classifiers (not rules)
  - Hidden Markov Models
  - Probabilistic Context Free Grammars
  - Maximum Entropy Models
  - …other tools of Machine Learning, AI, Statistics

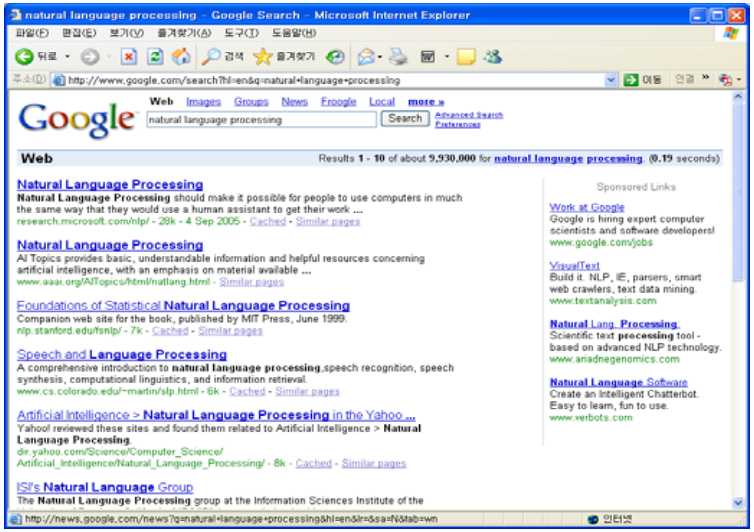
Natural Language Processing

- **NLP is the study of the computational treatment of natural languages:**
  - Most commonly Natural Language Understanding.
  - The complementary task is Natural Language Generation.
- **NLP draws on research in Linguistics, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, etc.**

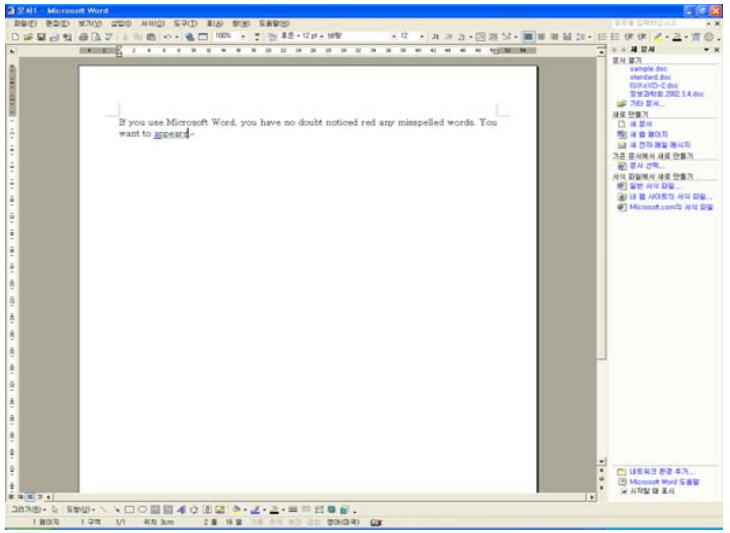
What & Where is NLP

- **Goals can be very far-reaching**
  - True text understanding
  - Reasoning and decision-making from text
  - Real-time spoken dialog
- **Or very down-to-earth**
  - Searching the Web
  - Context-sensitive spelling correction
  - Analyzing reading-level or authorship statistically
  - Extracting company names and locations from news articles

Example Applications of NLP: Information Retrieval (IR)



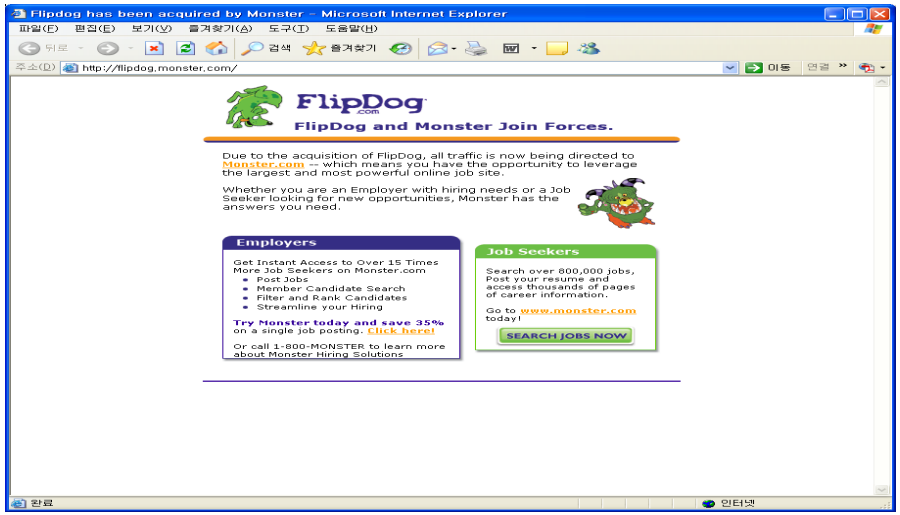
Applications of NLP: spelling correction, grammar checking



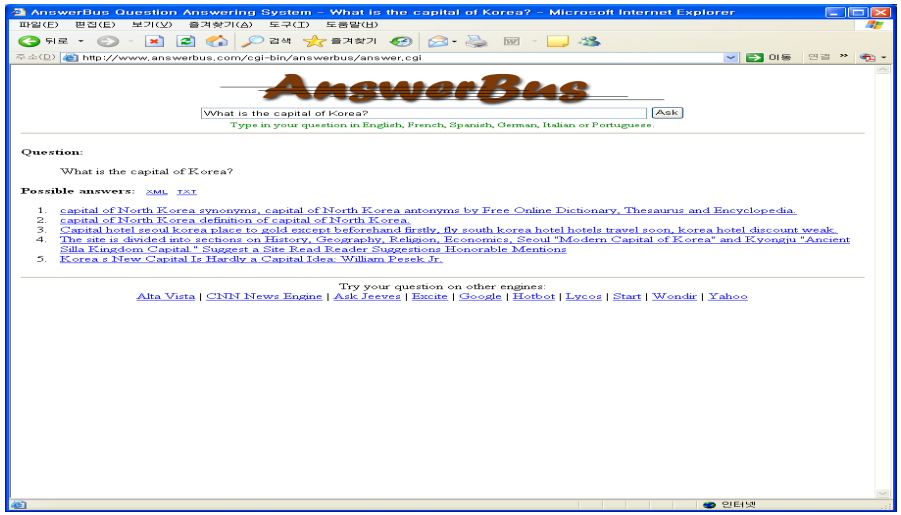
Applications of NLP: News categorization and summarization

Applications of NLP: Information Extraction: Find experts, employees

Applications of NLP: Information Extraction: Job Openings

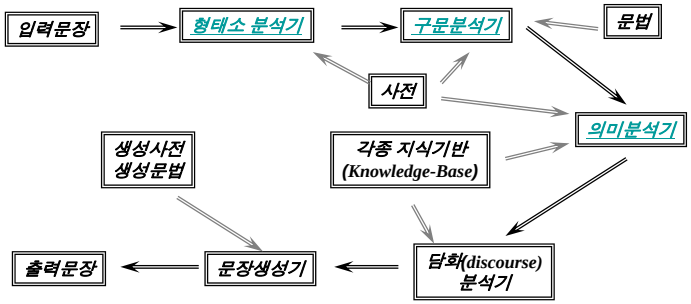


Applications of NLP: Question Answering



Natural Language Processing

❖ 자연언어처리 시스템의 구성도



Contents

1. Human Language Technology (HLT)

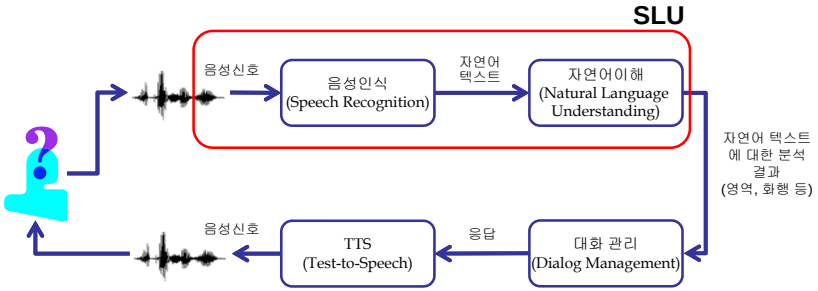
2. Spoken Language Understanding (SLU)
- Spoken Dialogue User Interface for Intelligent Robots and Smart Phones

3. Opinion Mining
- Sentiment Analysis
- Comparison Mining

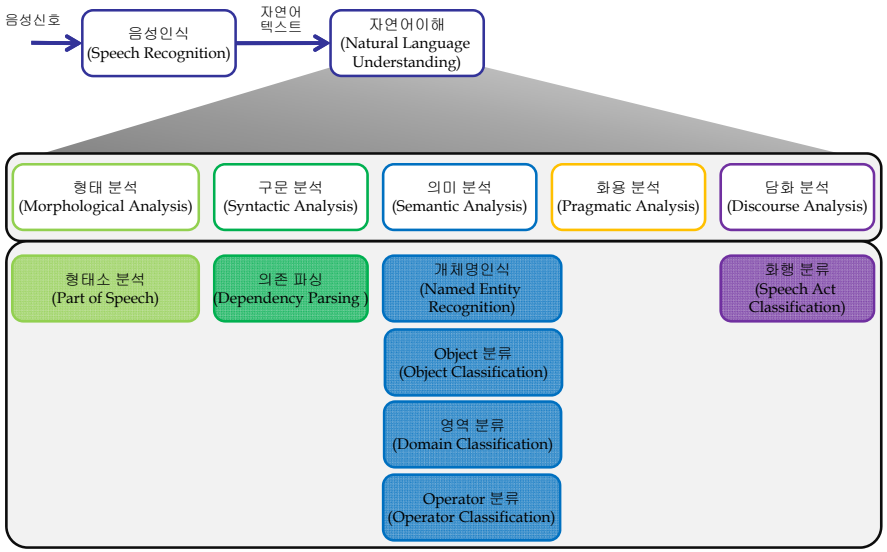
4. Others
- AAC, Text Mining & Information Retrieval
- Big Data Analysis,

Spoken Language Understanding

- 사용자의 음성을 이용한 자연어 기반의 인터페이스 사용의 증가
- 사용자가 입력한 음성을 인식하여 인식된 자연어를 분석하고 이해하는 것이 필요
  - **SLU(Spoken Language Understanding)**
    - **음성인식 + NLU(Natural Language Understanding)**



Spoken Language Understanding



형태소 분석

- **형태소**
  - 의미가 있는 최소 단위
- **형태소 분석**
  - 단어(또는 어절)를 구성하는 각 형태소를 분리
  - 분리된 형태소의 기본형 및 품사 정보를 추출

이번 주말 강원도 날씨는 어떻니

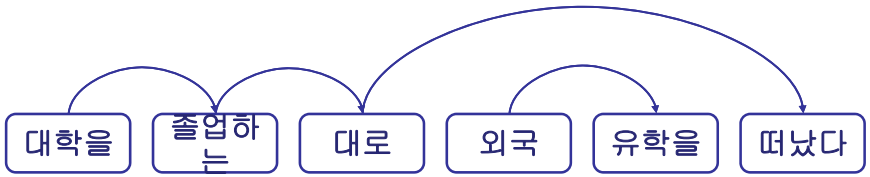
↓

이번/NNG 주말/NNG 강원도/NNP 날씨/NNG+는/JX 어떻/VA+니/EC

의존 파싱

- **의존 문법**
  - 문장을 구성하는 언어요소(형태소 혹은 어절)와 또 하나의 언어요소 사이의 의존 관계를 파악함으로써 문장을 분석
- **지배소**
  - 의미의 중심이 되는 요소
- **의존소**
  - 지배소가 갖는 의미를 보완해 주는 요소

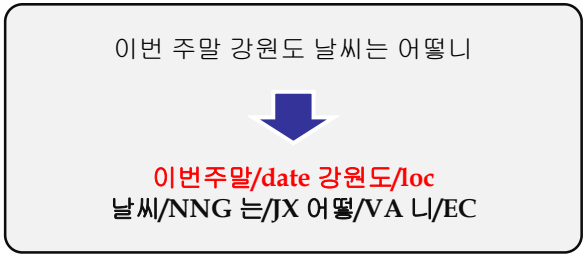
-한국어는 지배소 후위의 원칙
-지배소는 의존소보다 문장 내에서 뒤에 위치한다.



언어요소 - 어절

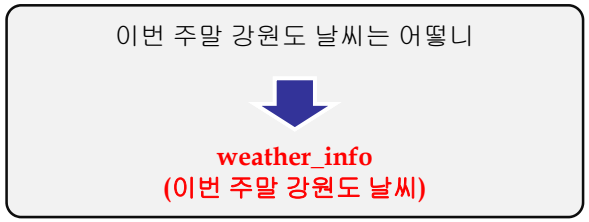
개체명 인식

- **개체명(Named Entity)**
  - 인명(Person), 지명(Location), 기관명(Organization) 등과 같은 고유명사
- **개체명 인식**
  - 한국어 문장에 개체명을 인식하여 해당 개체명 태그를 달아주는 것



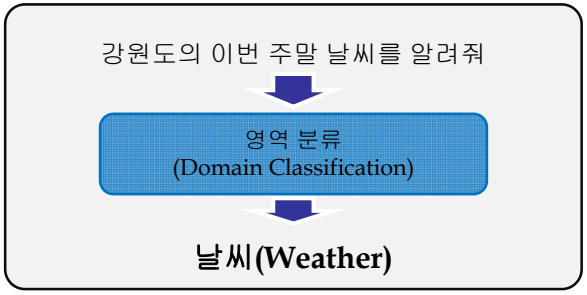
Object 분류

- **Object**
  - 사용자가 얻고자 하는 정보
  - weather_info(날씨정보), humidity_info(습도정보), bus_number(버스번호) 등
- **Object 인식**
  - 발화에서 사용자가 원하는 결과를 판단할 수 있는 구간을 인식하여 사용자가 얻고자 하는 정보를 분류



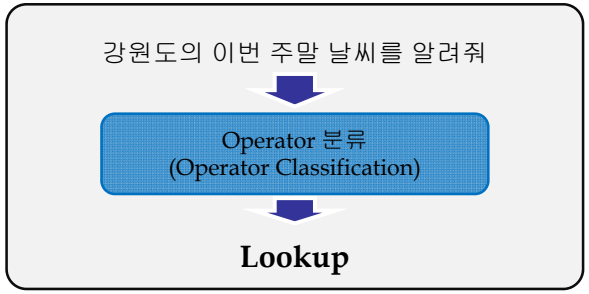
영역 분류

- **영역(Domain)**
  - 자연어 문장이 포함되는 범주(Category) 또는 주제(Topic)
- **영역 분류**
  - 자연어 문장을 분석하여 적절한 범주로 분류하는 것



Operator 분류

- **Operator**
  - 사용자가 요구하는 행동
  - 설정(set), 수정(mod), 삭제(del), 찾기(lookup) 등
- **Operator 분류**
  - 발화 속에 포함된 사용자가 요구하는 행동을 분류하는 것



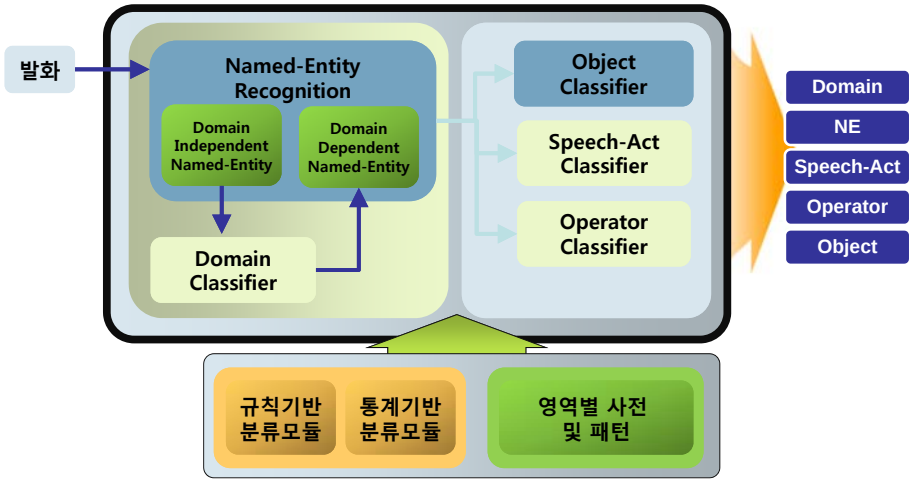
화행 분류

❖ **화행(Speech Act)**

➢ 발화 속에 포함된 대화 목적을 수행하기 위한 화자의 의도된 행위

ask_ref	정보 요구	화자가 청자에게 어떤 변수의 값을 요구(WH-question)
ask_if	정보 요구	화자가 청자에게Yes/No의 답을 요구(Y/N-question)
inform	정보 제공	화자가 청자에게 정보를 제공
response	응답	ask_ref, ask_if에 대한 대답
request	행위 요구	화자가 청자에게 어떠한 행위를 요구
accept	호응	화자가 청자의 발화에 호응
confirm	확인	확인을 요구하는 발화에 대한 응답
reject	거절	대화를 계속 진행할 수 없는 상황

SLU 구조의 예

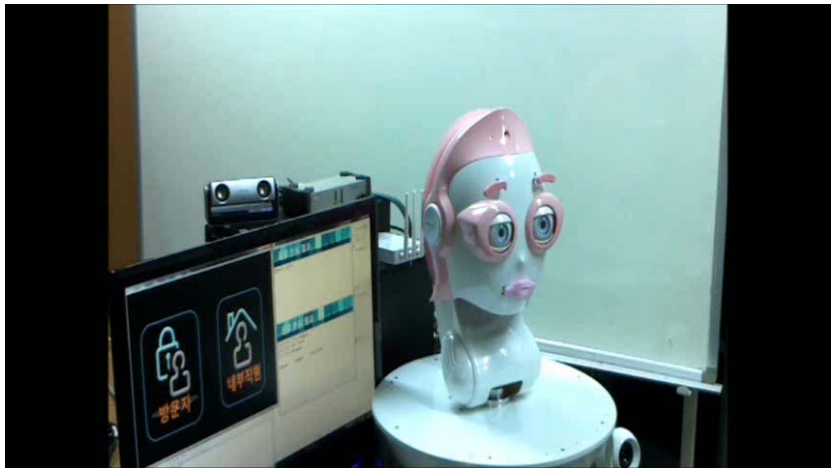


Spoken Dialogue Interface for Smart Phones

Demonstration

Spoken Dialogue Interface for Intelligent Robots

Demonstration



Contents

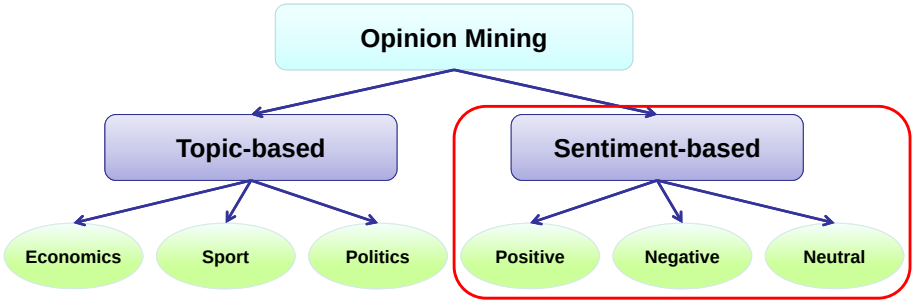
1. Human Language Technology (HLT)

2. Spoken Language Understanding (SLU)
   - Spoken Dialogue User Interface for Intelligent Robots and Smart Phones

3. Opinion Mining
   - Sentiment Analysis
   - Comparison Mining

4. Others
   - AAC, Text Mining & Information Retrieval
   - Big Data Analysis

Opinion Mining – Sentiment Analysis



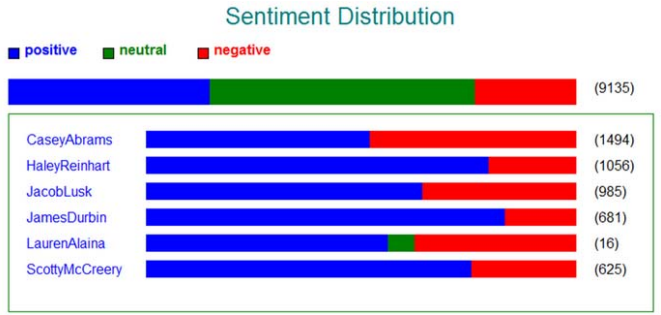
- **감정 분석(sentiment analysis) : 문서의 주관적(긍정, 부정 또는 중립) 정보를 추출하거나 식별하는 작업**
  - 자연어 처리(natural language processing), 텍스트 분석(text analysis), 컴퓨터 언어학(computational linguistics)을 이용

Levels of Sentiment Analysis

- **Document Level**
  - 문서내의 하나의 Entity
- **Sentence Level**
  - 문장내의 하나의 Entity
  - Subjectivity classification
- **Aspect / Feature Level**
  - Opinion-based classification
    - Sentiment
    - Opinion target

Practical Example

- **트위터 감정 분석을 통한 아메리칸 아이돌 결과 예측(2011)**



Methods and Features

- **기계 학습(machine learning)**
  - Latent Semantic Analysis
  - Support Vector Machine
  - Bag of Words
  - Semantic Orientation – Pointwise Mutual Information

- **전통적인 Text Categorization 에 적용되어 왔던 기계 학습 방법을 Sentiment Analysis 에 적용**
  - SVM(Support Vector Machine) ,Naïve Bayes, Maximum Entropy Model, Neural Net 등

Methods and Features

- **Bag of Words**
  - 2개의 간단한 문서:
    - 문서1 : John likes to watch movies, Mary likes too
    - 문서2 : John also like to watch football games
  - 10개의 구분되는 단어로 사전을 구성하고 사전의 인덱스를 사용해 문서를 벡터로 표현
  - 장점
    - 간단하고 효율적이며 높은 성능
  - 단점
    - 문맥 정보 및 위치정보가 손실되어 성능 저하

Dictionary

John : 1
likes : 2
to : 3
watch : 4
movies : 5
also : 6
football : 7
games : 8
Mary : 9
too : 10

Index	John	likes	to	watch	movies	also	football	games	mary	too
문서1	1	2	1	1	1	0	0	0	1	1
문서2	1	1	1	1	0	1	1	1	0	0

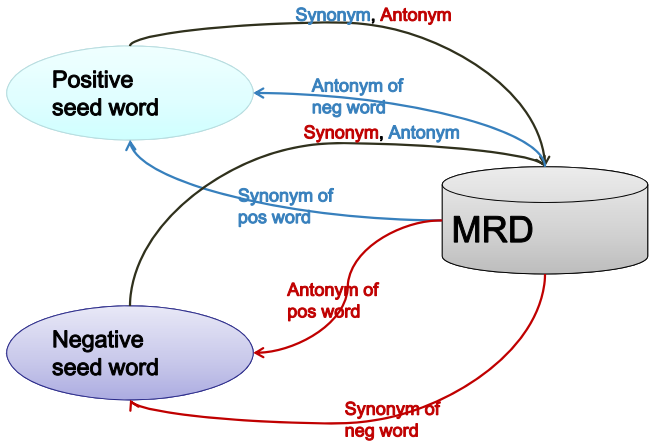
Methods and Features

- **종자 어휘(seed word)**
  - 명확히 감정을 표현하는 단어 (영어권)
    - good, excellent, nice, positive, fortunate, correct, superior
    - bad, nasty, poor, negative, unfortunate, wrong, inferior
  - 문서 및 문장의 감정 인식을 위한 중요한 자질
  - 각 7개씩의 단어로는 부족하기 때문에 확장 필요

- **한국어 감정 단어 집합 구축**
  - 영어권 연구 적용
    - 영어 시소러스를 이용하여 각 종자 어휘의 동의어와 반의어 정보를 이용하여 확장
    - 확장된 어휘 집합을 영한 사전을 이용하여 번역 후 구축

Methods and Features

- **한국어 감정 단어 집합 구축**



Methods and Features

❖ **Semantic Orientation – Pointwise Mutual Information**

➢ 구축된 감정 단어들의 감정의 강도(strength)를 추정하는 방법

$$PMI(t, t_i) = log \frac{\Pr(t, t_i)}{\Pr(t)\Pr(t_i)}$$

t	Target word
t_i	Seed word

➢ 긍정(부정) 종자 어휘와 높은 확률로 같이 등장하는 단어는 긍정(부정)일 것이다라는 가정으로부터 출발

$$O(t) = \sum_{t_i \in S_p} PMI(t, t_i) - \sum_{t_i \in S_n} PMI(t, t_i)$$

S_p	Positive seed word set
S_n	Negative seed word set

➢ 양수 값 : positve orientation, 음수 값 : negative orientation

Methods and Features

❖ **Sentiment Linguistic Resources for English**

➢ SentiWordNet

- 영어권 단어의 positive / negative / neutral 의 각 Value 정보가 들어 있는 어휘 자원
- SentiWordNet 에서 제공하는 Value 또는 PMI 를 통해서 계산된 Semantic Orientation Value 를 기계 학습 기법에 적용하여 Sentiment Analysis 를 수행하는 연구들이 활발히 진행 중

➢ 그외 영어권에는 많은 어휘 자원들이 존재

- SentiWordNet
  - ✓ http://sentiwordnet.isti.cnr.it
- General Inquirer Lexicon
  - ✓ http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
- Sentiment Lexicon
  - ✓ http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
- MPQA Subjectivity Lexicon
  - ✓ http://mpqa.cs.pitt.edu/subj_lexicon.html
- Emotion Lexicon
  - ✓ http://www.saifmohammad.com/WebPages/ResearchInterests.html

Methods and Features

❖ **최근 연구 동향**

➢ Semi-Supervised 또는 Unsupervised Machine Learning 기법을 적용

- Latent Dirichlet Allocation(LDA)

➢ 의존 파서 등 문맥 정보를 반영할 수 있는 기법을 기계 학습에 적용

- 부정표현(negation) 정보 인식 후 반영

➢ Aspect 및 Opinion Holder 를 인식하는 연구

❖ **한국어 Sentiment Analysis**

➢ 감정 단어는 감정 분석에 매우 중요한 자질

➢ 영어권 어휘 자원에 비해 타언어권 어휘 자원은 매우 부실

➢ 활발한 어휘 자원 구축 및 연구 필요

Comparison Mining

Overview

iPhone vs. Galaxy-S

Which one is better?

Property	Comparative Type	%
Sound quality	Equality	30 %
	iPone is better.	40 %
	Galaxy-S is better.	30 %
Design	Equality	25 %
	…..	?? %

Comparison Mining

System Developing Flow

- **4 Stages**
  - Stage 1: Extracting Comparative Sentences from Texts
  - Stage 2: Classifying Comparative Sentences into Different Types
  - Stage 3: Mining Comparative Entities and Predicates
  - Stage 4: Analysis and Summary
- **Evaluation**

	Stage 1 : SVM	Stage 2: TBL
Accuracy (%)	90.30	89.57

Stage 3	Accuracy	Example("X파이가 Y파이보다 싸고 맛있다")
Subject Entity (SE)	86.87	"X파이"
Object Entity (OE)	86.24	"Y파이"
Predicate (PR)	88.20	"싸고 맛있다"

Contents

1. Human Language Technology (HLT)

2. Spoken Language Understanding (SLU)
   - Spoken Dialogue User Interface for Intelligent Robots and Smart Phones

3. Opinion Mining
   - Sentiment Analysis
   - Comparison Mining

4. Others
   - AAC, Text Mining & Information Retrieval
   - Big Data Analysis

Others-AAC Software & UI for Mobile Devices

- **AAC (Augmentative and Alternative Communication) Software**
  - Software for communication that are used to express thoughts, needs, wants, and ideas for the disabled
  - Results
    - MyTalkie (마이토키), Software for Conversation by Writing (필담소프트웨어)
- **Effective UIs on Mobile Devices for Web Browsing**
  - Automatic Keyword Extraction, Text Summarization & IR system

Others-Text Summarization

- **문서 요약의 개요**
  - 생성요약(Abstraction) : 전체 문서의 내용을 압축하여 새롭게 작성
  - 추출요약(Extraction) : 문서의 중요문장을 그대로 추출하여 요약 생성
- **자동 문서 요약 절차**
  - 핵심어 추출 : 제목, 첫 문장 활용
  - 적합/부적합 문장 판정
    - Pseudo Relevance Feedback 기법 사용
  - 핵심어 확장

$$w_t = TSV_t = \log \frac{p(1-q)}{q(1-p)} = \log \frac{(r+0.5)(S-s+0.5)}{(R-r+0.5)(s+0.5)}$$

  - 문장의 중요도 계산 및 요약문 생성

$$TSVscore(S_i) = \sum_{w_j \in S_i} w_j \quad Score(S_i) = \alpha \left(\frac{TSVscore(S_i)}{TSVscoreMax} \right) + (1-\alpha) \left(1 - \frac{i-1}{N} \right)$$

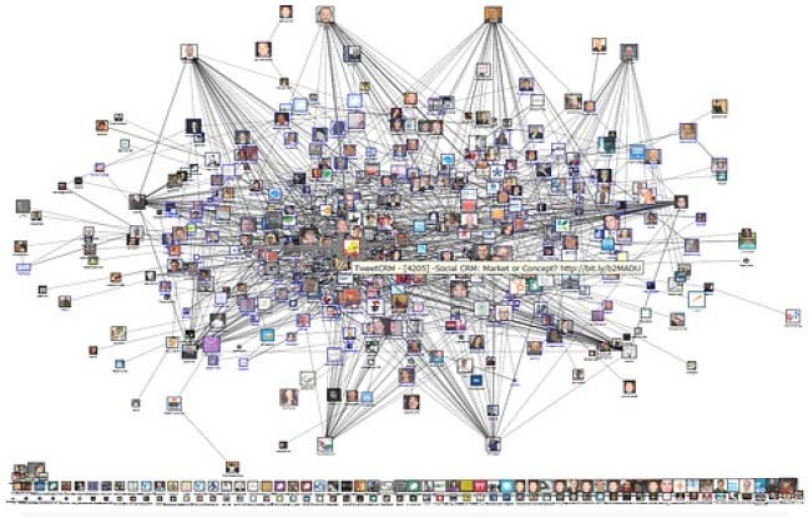
Others-AAC Software

Demonstration

Big Data Analysis

- **Big Data Analysis Techniques**
  - Related existing techniques: Data Mining, Machine Learning, Natural Language Processing, Pattern Recognition
  - 특징
    - 대용량 데이터, 비정형 데이터
  - 중요 분석 기법
    - Text Mining: NLP based information extraction
    - Opinion Mining
    - Social Network Analysis
    - Clustering Analysis
    - Visualization
  - 분석 인프라
    - Hadoop: Map-reduce
    - 통계 프로그래밍 언어: R
    - 토픽모델링: LDA (Latent Dirichlet Allocation)

Big Data Analysis



Thank you for your attention!

http://web.donga.ac.kr/yjko/

고 영 중