Introduction to NLP and Text Mining

Tutor: Rahmad Mahendra

Natural Language Processing & Text Mining
Short Course

Pusat Ilmu Komputer UI
22 – 26 Agustus 2016

References

- Jurafsky and Martin, Speech and Language Processing 2nd ed, Prentice-Hall, 2008.
- Manning and Schutze, Foundation of Statistical Natural Language Processing, 1999.
- Natural Language Processing course materials: Stanford University, Edinburgh University, Illinois University, University of California at Berkeley, University of Texas at Austin, ETH Zurich, National University of Singapore, Universitas Indonesia

pusilkom ui
quality results

References

- Feldman and Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2007
- Indurkhya and Damerau (ed), Handbook of Natural Language Processing 2nd ed, CRC Press, 2010

pusilkom ui
quality results

Text Mining

Text Mining

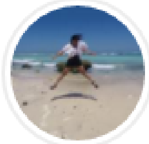
System that analyzes large quantities of natural language text dan detects lexical or linguistic patterns in an attempt to extract probably useful information. (Sebastiani, 2002)

Mining useful information from unstructured text…

pusilkom ui
quality results

Unstructured...

Free text,
Grammatical Error,
Ambiguity,
Complex,
Slank Words, ...



anggia_rurinda

Kontributor Tingkat 3

11 ulasan

3 ulasan hotel

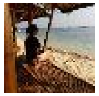
"Serasa Punya Pulau pribadi"

Diulas pada Oktober 6, 2015

Saya baru sekali pergi ke pulau umang, untuk menuju kesana dibutuhkan sekitar 6 jam perjalanan darat dan sekitar15 menit naik speed boat menuju ke pulau.Saat itu saya pergi bersama dengan teman-teman kantor. Kamar yang saya tempati cukup luas, terdapat ruang tamu dan kamar mandi di lantai bawah dan kamar tidur di lantai atas. Kebersihan kamar cukup oke, namun saya agak terganggu dengan semut-semut yang kecil yang ada disekitar kamar mandi. Breakfast di resor pulau umang cukup beragam dan rasanya cukup enak. Namun saya sangat mengagumi pemandangan di pulau ini, sangat indah dengan pasir pantai yang putih bersih..saya sangat puas bermain lari-karian, tidur-tiduran, loncat-loncat sampai main speed boat dan banana boat dan naik ke sebuah pelampung berbentuk donat (saya tidak tahu apa namanya). Over all saya cukup sengan bermain ditempat tersebut.

Tips Kamar: Semua kamar menghadap ke pantai dan laut, kamar manapun oke. Minta kamar yang ACnya dingin.

Lihat tips lainnya tentang kamar


Semi-Unstructured...

```
Report Number: 0000000W Report Status: FINAL
Type: EKG
Date: 02/02/02 13:25
Electrocardiogram Report for Accession # 00-00000W 02/02/02 13:25
VENT. RATE 76 BPM
PR INTERVAL 154 ms
QRS DURATION 88 ms
QT/QTC 422 474 ms
P-R-T AXES 55 94
NORMAL SINUS RHYTHM
NONSPECIFIC T WAVE ABNORMALITY
PROLONGED QT INTERVAL OR T-U FUSION, CONSIDER MYOCARDIAL DISEASE,
ELECTROLYTE IMBALANCE, OR DRUG EFFECTS
ABNORMAL ECG
WHEN COMPARED WITH ECG OF 01-JAN-2001 14:02,
T WAVE INVERSION NOW EVIDENT IN LATERAL LEADS
REFERRED BY: AARDVARK, M.D.,ALICE. REVIEWED BY: BEAGLE, M.D.,BOB
[report_end]
```

XML.


```
<text_report>
<text><title></title><p>Stationary ECG Study<br/>XXXX University XXXX<br/>XXXX XXXX</p></text>
<text><title>Test Date</title><p>XXXX<br/></p></text>
<text><title>XXXX Name</title><p>XXXX<br/></p></text>
<text><title>Patient XXXX</title><p>XXXX                Room:<br/></p></text>
<text><title>Gender</title><p>Male                  Technician:   JG<br/></p></text>
<text><title>DOB</title><p>XXXX<br/></p></text>
<text><title>XXXX Number</title><p>XXXX               XXXX MD:   XXXX XXXX</p>
<p>Intervals                        Axis<br/></p></text>
<text><title>XXXX</title><p>46                  P:            32<br/></p></text>
<text><title>PR</title><p>132                 QRS:          -19<br/></p></text>
<text><title>QRSD</title><p>112                 T:            -33<br/></p></text>
<text><title>QT</title><p>472<br/></p></text>
<text><title>QTc</title><p>448</p>
<p>Interpretive Statements<br/>
Sinus bradycardia<br/>T XXXX inversions in leads #, aVF are not XXXX, but are more prominent<br/>
compared to prior studies<br/>Subtle ST-T changes elsewhere are nonspecific<br/>
Low QRS voltages in precordial leads<br/>
Abnormal ECG</p>
<p>Electronically Signed On XXXX XXXX by XXXX XXXX</p></text>
</text></text_report>
```


Example: **ECG Reports**

(Angelino, 2012)

pusilkom ui
quality results

Structured...

Database

Potential Customer Table

Person	Age	Sex	Income	Customer
Ann Smith	32	F	10 000	yes
Joan Gray	53	F	1 000 000	yes
Mary Blythe	27	F	20 000	no
Jane Brown	55	F	20 000	yes
Bob Smith	50	M	100 000	yes
Jack Brown	50	M	200 000	yes

Married-To Table

Husband	Wife
Bob Smith	Ann Smith
Jack Brown	Jane Brown

(Dzerovski, 1996)

pusilkom ui
quality results

Data Mining vs Text Mining

- "Data Mining is essentially concerned with information extraction from structured databases."

- In reality, a large portion of the available information appears in textual and unstructured form. Text mining operates on textual data to extract information from a collections of texts.

(Rajman & Besancon, 1997)

pusilkom ui
quality results

Text Mining

INPUT: raw and unstructured text

This past Saturday, I bought a **Nokia** phone
and my friend bought a **Motorola** phone
with Bluetooth. We called each other when we
got home. Basically I **like** the screen.
But the **voice** on **my phone** was **not so clear**, **worse than my previous Samsung phone**. **The battery life was short too**. **My friend was quite happy with her phone**. **I wanted a phone with good sound quality just like his phone**. **So my purchase was a real disappointment**. I returned the phone yesterday."

OUTPUT:

Nokia
Screen: **good**
Battery life : **bad**
Sound quality : **bad**

Motorola
Sound quality : **good**

Samsung
Sound quality : **better- than** Nokia

pusilkom ui
quality results

Natural Language Processing

Natural Language Processing

- NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language.
- Also called Computational Linguistics
  - Also concerns how computational methods can aid the understanding of human language

Why Study NLP

- An enormous amount of knowledge is now available in machine readable form as natural language text.
- Conversational agents are becoming an important form of human-computer communication.
- Much of human-human communication is now mediated by computers.
- Lots of exciting stuff going on …

NLP Related Area

- Artificial Intelligence
- Formal Language (Automata) Theory
- Machine Learning
- Linguistics
- Psycholinguistics
- Cognitive Science
- Philosophy of Language

Linguistic Level of Analysis

- Word
- Syntax
  - concerns the proper ordering of words and its affect on meaning.
- Semantics
  - concerns the (literal) meaning of words, phrases, and sentences.
- Pragmatics
  - concerns the overall communicative and social context and its effect on interpretation.

pusilkom ui
quality results

Word

This is a simple sentence **WORDS**

Example is taken from Edinburgh's lecture notes

Morphology

This is a simple sentence **WORDS**

be
3sg
present

MORPHOLOGY

Example is taken from Edinburgh's lecture notes

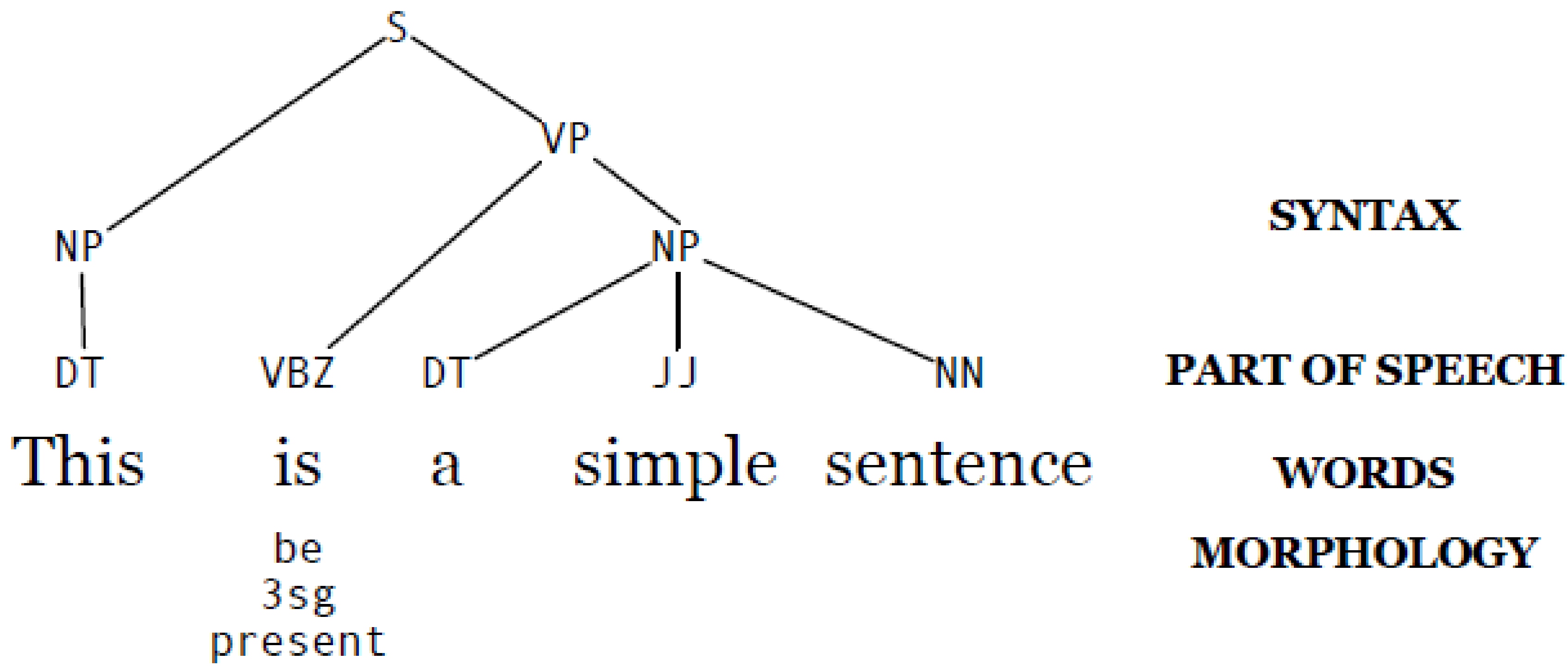
pusilkom ui
quality results

Part of Speech

DT	VBZ	DT	JJ	NN	**PART OF SPEECH**
This	is	a	simple	sentence	**WORDS**
	be 3sg present				**MORPHOLOGY**

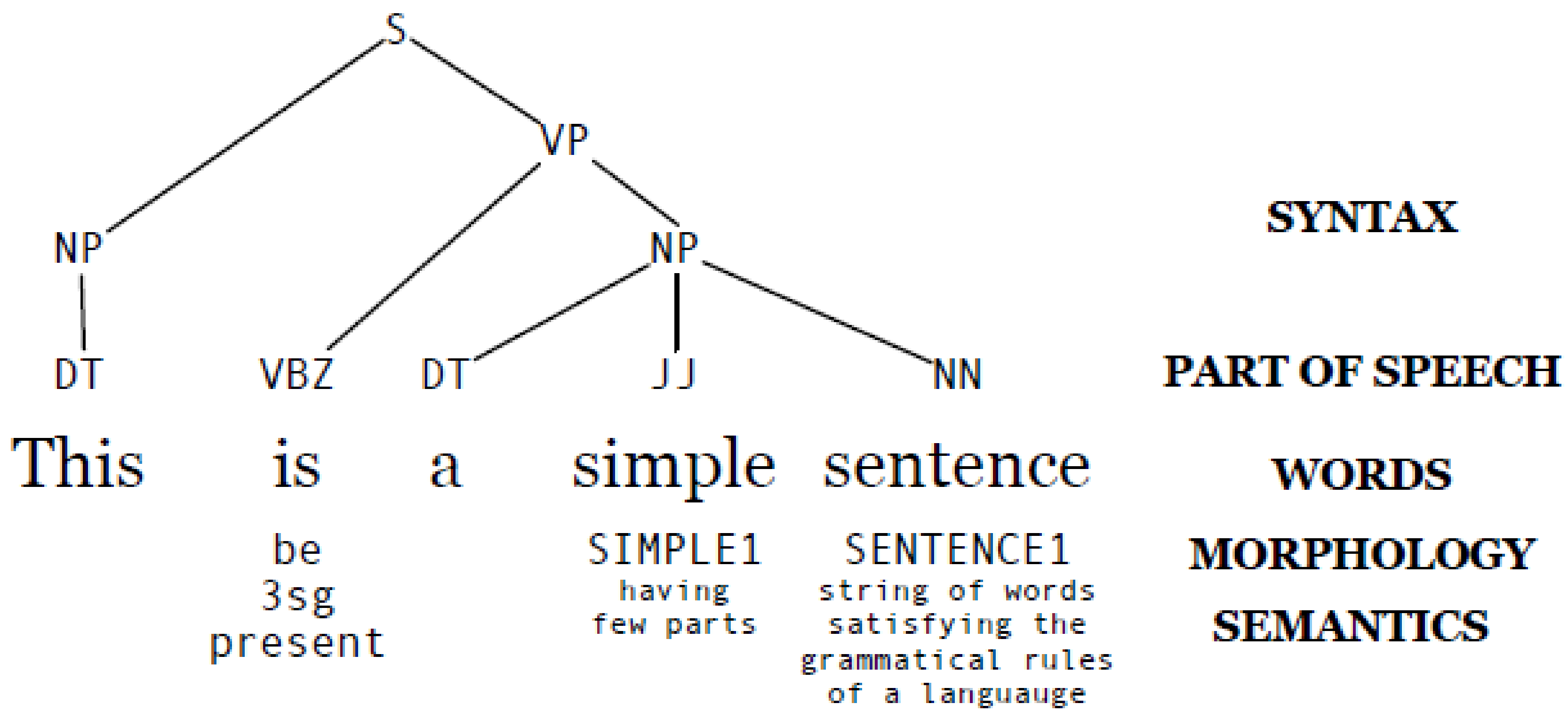
Example is taken from Edinburgh's lecture notes

Syntax



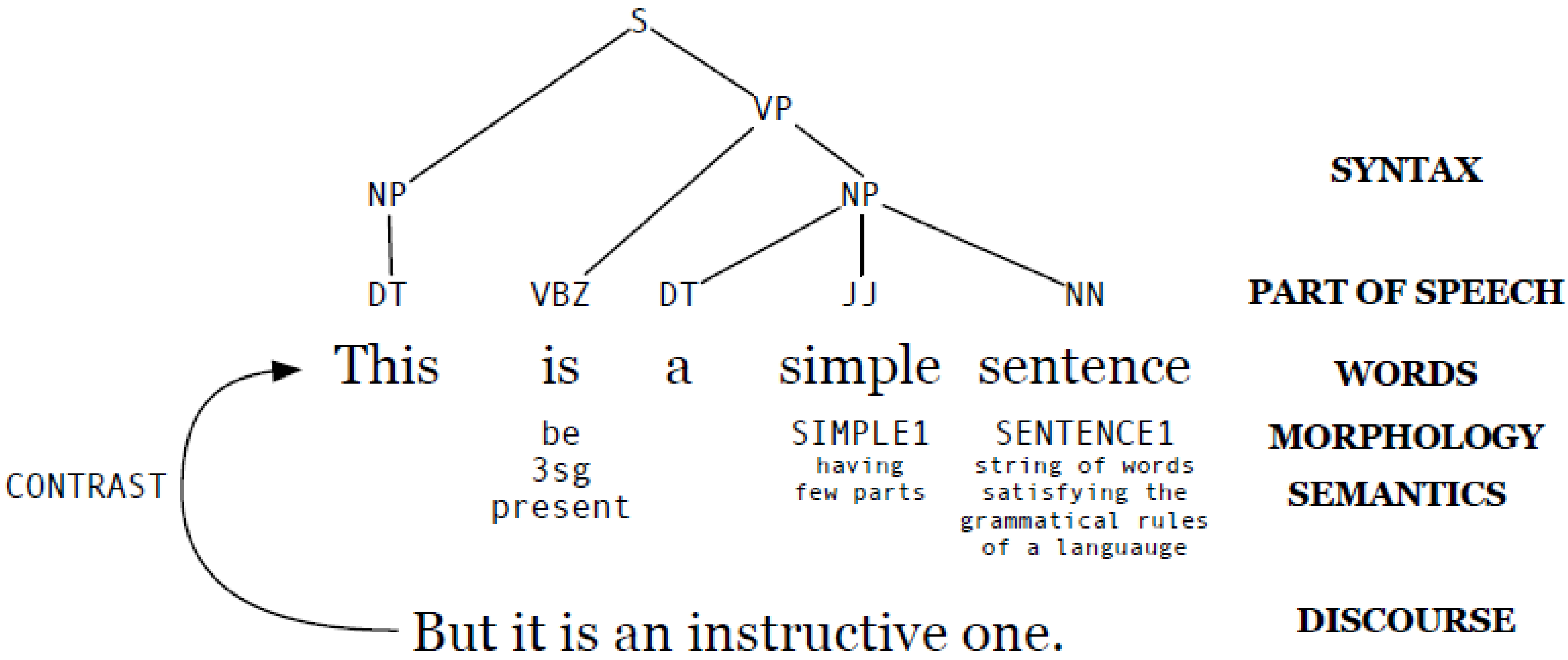
Example is taken from Edinburgh's lecture notes

Semantics



Example is taken from Edinburgh's lecture notes

Discourse



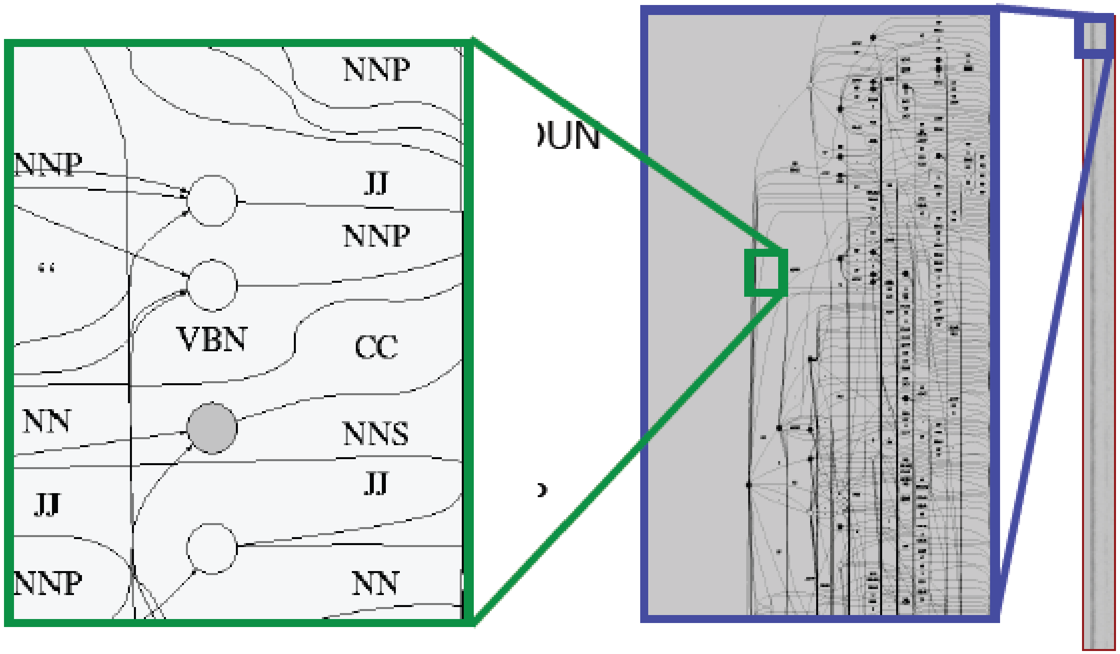
Example is taken from Edinburgh’s lecture notes

Why NLP is Hard

- Ambiguity
  - Lexical Ambiguity
  - Structural Ambiguity
  - Referential Ambiguity
- Sparsity
- Scale
- Unmodeled Variable

Ambiguity

- Time flies like an arrow
- Fruit flies like an arrow
- The boy saw the man with telescope
- Rahmad makan bakso dengan mie
- Rahmad makan pangsit dengan sumpit
- Rahmad makan soto dengan Alfan
- Kakak mengusili adik. Dia menangis sesenggukan.
- Kakak mengembalikan kunci motor adik. Dia berterima kasih.



- Language is produced with the intent of being understood. There may be relevant knowledge source related to language.

NLP Core Tasks

- Morphological Analysis
- Part-of-Speech Tagging
- Named-Entity Recognition
- Syntactic Parsing
- Semantic Parsing
- Word Sense Disambiguation
- Textual Entailment
- Coreference Resolution

Textual Entailment

TEXT	HYPOTHESIS	ENTAILMENT
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	*Yahoo bought Overture.*	TRUE
Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.	*Microsoft bought Star Office.*	FALSE
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.	*Israel was established in May 1971.*	FALSE
Since its formation in 1948, Israel fought many wars with neighboring Arab countries.	*Israel was established in 1948.*	TRUE

Examples are taken from PASCAL challenge

Coreference Resolution

- Determine which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

NLP Applications

- Spelling and Grammar Correction
- Information Retrieval
- Text Summarization
  http://autosummarizer.com/
- Text Classification

NLP Applications

- Machine Translation
  http://translate.google.com
- Question Answering
  http://start.csail.mit.edu
- Sentiment Analysis

pusilkom ui
quality results

Approach to Solve NLP Problem

- Rule Based (Symbolic)
  - Developed hand coded rules
- Statistics Based (Empirical)
  - Annotate data based on standard tagsets, then machine learn a model
- Hybrid systems
  - Often blend rule-based pre- and post-processing with ML core

(Effective) NLP Cycle

- Pick a problem (usually some disambiguation).
- Get a lot of data (hopefully labeled, but often unlabeled).
- Build the simplest thing that could possibly work.
- Repeat:
  - Examine the most common errors are.
  - Figure out what information a human might use to avoid them.
  - Modify the system to exploit that information
    - Feature engineering
    - Representation redesign
    - Different machine learning methods

pusilkom ui
quality results

THANK YOU