Natural Language Processing Introduction and Practise

Wei Heng

Amazon China

hengwei@amazon.com

Oct 21, 2013

Agenda

NIp Overview

- History of Natural Language Processing?
- What is Natural Language Processing?
- Why is NLP hard?

Basic NLP Problems

- Tokenization & Tagging
- Parsing
- Common Work flow for general Chinese Processing

Future Reading

- Common Tools and Corpus
- Reference

History of Natural Language Processing

NIp began in the 1950s as the intersection of artificial intelligence and linguistics. There mainly existed two important stages.
Before 1980s, rule-based natural language processing:

1. Early simplistic approaches, for example, word-for-word Russian-to-English machine translation.
2. Chomsky's 1956 theoretical analysis of language grammar. influencing the creation(1963) of Backus-Naur Form(*BNF*) notaiton. and *BNF* is used to specify *CFG*, and is commonly used to represent programming-language syntax.
3. 1970s, lexical-analyzer(lexer) generators and parser generators such as the *lex/yacc* combination.

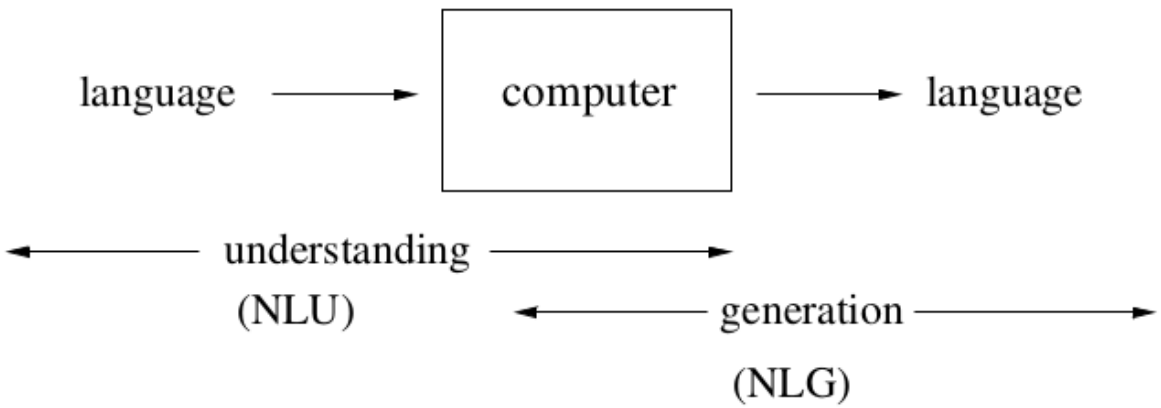
History of Natural Language Processing

After 1980s, it resulted in a fundamental reorientation, duing to the limitations of hand-written rules:

1. Simple, robust approximations replaced deep analysis
2. Evaluation became more rigorous
3. Machine learning method that used probabilities became prominent
4. Large, annotated bodies of text (corpora) were employed to train machine-learning algorithms—the annotation contains the correct answers—and provided gold standards for evaluation.

What is Natural Language Processing?

Computers using natural language as input and/or output



Modern NIp algorithms are baded on machine learning, especially statistical machine learning.

Major Tasks in Natural Language Processing

Low-level NIp tasks:

- Sentence boundary detection: abbreviations and tile (m.g. and Dr.)
- Tokenization: identify individual tokens(word, punctuation) within a sentence. A lexer plays a core role.
- Part of speech assignment to individual words.
- Morphological decomposition of compound words.
- Shallow parsing (chunking): identifying phrases from constituent part-of-speech tagged tokens.

Major Tasks in Natural Language Processing

High-level Nlp tasks:

- Spelling/grammatical error identification and recovery
- Named entity recognition
- Automatic summarization: Produce a readable summary of a chunk of text.
- Machine translation: Automatically translate text from one human language to another.
- Question answering: Given a human-language question, determine its answer.

Why is NLP hard?

[Example for Collins COMS W4705]
"At last, a computer that understand you like your mother"

Why is NLP hard?

Ambiguity:

At last, a computer that understands you like your mother"

1. It understands you as well as your mother understands you
2. It understands (that) you like your mother
3. It understands you as well as it understands your mother

1 and 3: Does this mean well, or poorly?

Why is NLP hard?

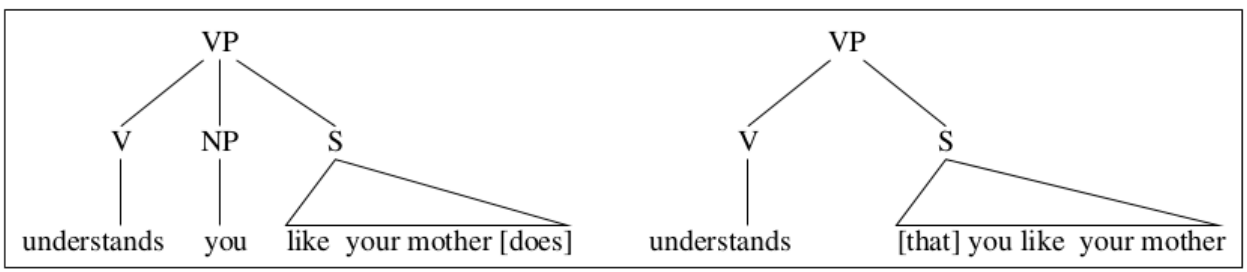
Ambiguity at Many Levels

At the acoustic level(speech recognition)

1. “ . . . a computer that understands you like your mother”
2. “ . . . a computer that understands you lie cured mother”

At the syntactic level:

Different structures lead to different interpretations.



Why is NLP hard?

At the semantic (meaning) level:

Two definitions of "mother"

1. a woman who has given birth to a child
2. a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

More Example at the semantic level:

- They put money in the bank = buried in mud?
- I saw her duck with a telescope

Tokenization & Tagging

Mainly Three type of mathods:

- Simple string comparision combined with dictionary, such as forward, backward match method.
- hard-written syntax or semantic rules, pCFG based
- statistical based learning, such Maximium Entropy, Conditional Random Field.

Tokenization and Tagging

Table: Detailed Comparison

Method	**String Match**	**Rule-based**	**Statistical**
Ambiguilty	bad	good	good
new words	bad	good	good
dictionary	yes	no	no
corpus	no	no	yes
rules	no	yes	no
complexity	easy	hard	general
accuracy	soso	accurate	good
effecify	fast	slow	general

Tokenization and Tagging

Example 1: Part-of-Speech tagging

Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ./.

Example 2: Named Entity Recognition

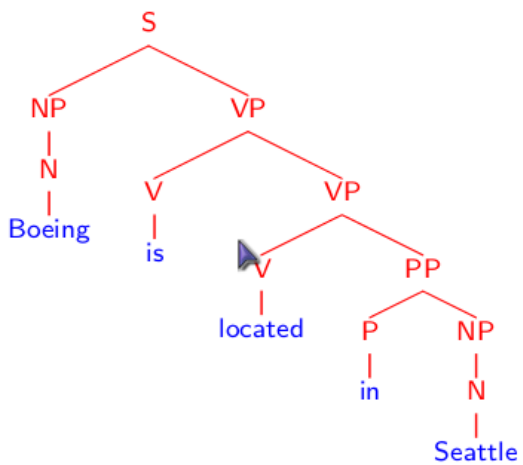
Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ./.

Parsing

INPUT:

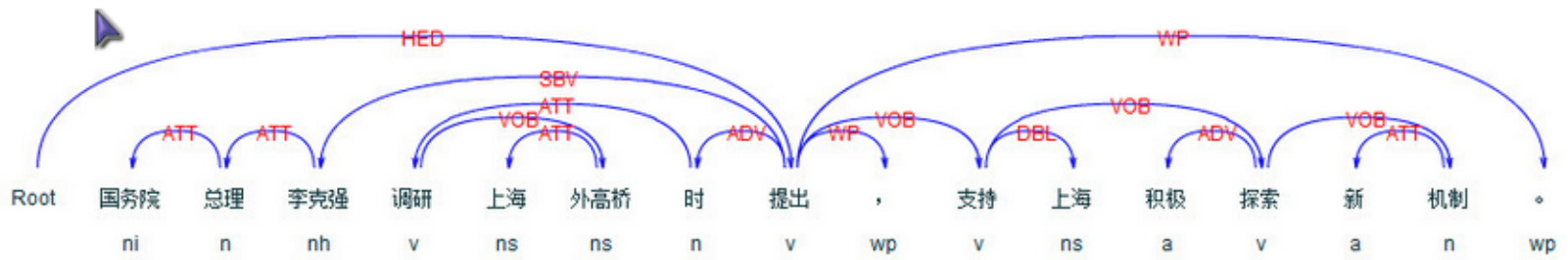
Boeing is located in Seattle.

OUTPUT:



Parsing

Chinese Parsing Example



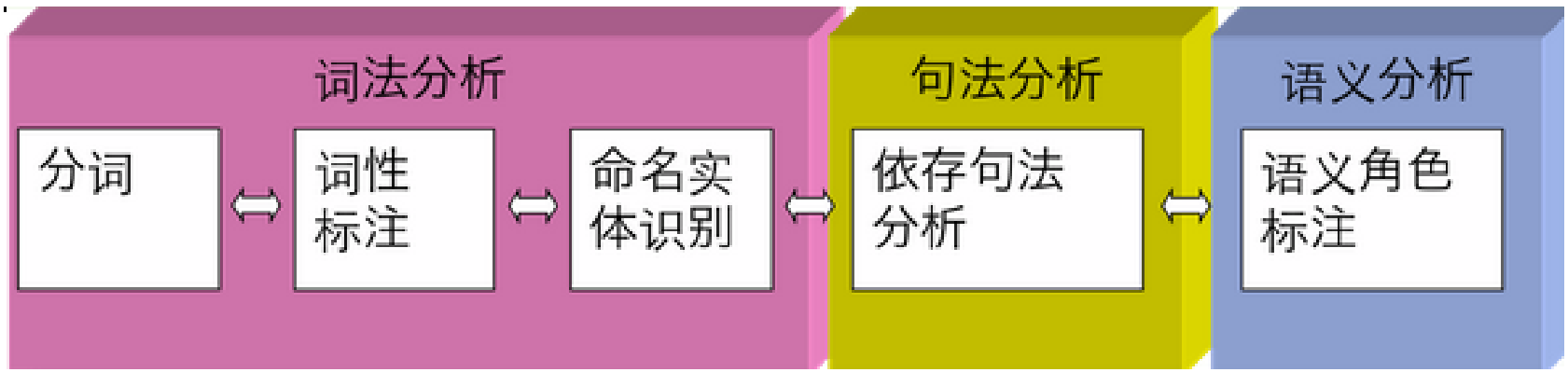
Paring

Depandency Relationship

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我送她一束花 (我 <-- 送)
动宾关系	VOB	直接宾语，verb-object	我送她一束花 (送 --> 花)
间宾关系	IOB	间接宾语，indirect-object	我送她一束花 (送 --> 她)
前置宾语	FOB	前置宾语，fronting-object	他什么书都读 (书 <-- 读)
兼语	DBL	double	他请我吃饭 (请 --> 我)
定中关系	ATT	attribute	红苹果 (红 <-- 苹果)
状中结构	ADV	adverbial	非常美丽 (非常 <-- 美丽)
动补结构	CMP	complement	做完了作业 (做 --> 完)
并列关系	COO	coordinate	大山和大海 (大山 --> 大海)
介宾关系	POB	preposition-object	在贸易区内 (在 --> 内)
左附加关系	LAD	left adjunct	大山和大海 (和 <-- 大海)
右附加关系	RAD	right adjunct	孩子们 (孩子 --> 们)
独立结构	IS	independent structure	两个单句在结构上彼此独立
核心关系	HED	head	指整个句子的核心

Common Work Flow

From LTP Cloud Platform



Common Tools and Corpus

CWS tools

- ICTCLAS Segmentation, Pos, NER
- 傻瓜分词
- Standform Segmentation and Parser Tools

Corpus

- Penn Tree for Chinese corpus
- Peking Training Set
- People's Daily(1998)

Reference

John Smith (2012)

History of Natural Language Processing

Association of Compuationall Linguistics 12(3), 45 – 678.

The Beginning