Brief Introduction to NLP

Natural Language and Knowledge Processing Group

Research Fellow: Keh-Yih Su
Table of Content

• A Simple Example: Syntactic Machine Translation
  – Process Flow

• Main Problems in NLP
  – Ambiguity & Ill-formedness

• Main Tasks in NLP
  – Tasks required in NLP, Bottlenecks

• Detailed Steps Overview
  – Analysis, Transfer and Generation
A Simple Example: Syntactic Machine Translation (Transfer Approach)

Analysis is target independent and generation is source independent.
Example: Miss Smith put two books on this dining table.

- Analysis
  - (1) Morphological Analysis

  Miss
  Smith
  put (+ed)
  two
  book+s
  on
  this
dining table.
Example: Miss Smith put two books on this dining table. Analysis (Cont.)

(2) Syntactic Analysis

```
S
  /\  
 NP  VP
  /\   
 V   NP  PP
   /\     
 Miss Smith put two books on this dining table
```
Transfer:

- (1) Lexical Transfer

  Miss $\Rightarrow$ 小姐
  Smith $\Rightarrow$ 史密斯
  put (+ed) $\Rightarrow$ 放
  two $\Rightarrow$ 兩
  book+s $\Rightarrow$ 書
  on $\Rightarrow$ 在…上面
  this $\Rightarrow$ 這
  dining table $\Rightarrow$ 餐桌

  - Literal Translation:
    小姐史密斯放兩書在上面這餐桌
Transfer (Cont.):

(2) Syntactic Transfer

史密斯小姐 把 兩書 放 在這餐桌上面
Generation:

- Generation

Transfer: 史密斯小姐把兩本書放在這張餐桌上面

⇒ 史密斯小姐把兩(本)書放在這(張)餐桌上面

中文翻譯： 史密斯小姐把兩本書放在這張餐桌上面
position \(\text{:= initial + rate \times 60}\)

\[\begin{array}{|c|c|}
\hline
\text{1} & \text{position} & \ldots \\
\hline
\text{2} & \text{initial} & \ldots \\
\hline
\text{3} & \text{rate} & \ldots \\
\hline
\end{array}\]

\[\begin{array}{c}
\text{id}_1 \text{:= id}_2 + \text{id}_3 \times 60
\end{array}\]

\[\begin{array}{c}
\text{id}_1 \text{:= id}_2 + \text{id}_3 \times 60
\end{array}\]
Intermediate code generator

temp1 := inttoreal (60)
temp2 := id3 * temp1
temp3 := id2 + temp2
id1 := temp3

code optimizer

temp1 := id3 * 60.0
id1 := id2 + temp1

code generator

Binary Code
source program

lexical analyzer

syntax analyzer

semantic analyzer

symbol-table manager

intermediate code generator

code optimizer

error handler

code generator

target program

[Aho 86]
Table of Content

• A Simple Example: Syntactic Machine Translation
  – Process Flow

• Main Problems in NLP
  – Ambiguity & Ill-formedness

• Main Tasks in NLP
  – Tasks required in NLP, Bottlenecks

• Detailed Steps Overview
  – Analysis, Transfer and Generation
Natural Language is more complicated

- **Ambiguity (岐義) --- For decoder**
  - Many different ways to interpret a given sentence.
  - #1: “[I saw [the boy] [with a telescope]]"
  - #2: “[I saw [the boy [with a telescope]]]"

- **Ill-Formedness (不合設定, 非正規) --- For decoder**
  - Language is involving, and no rule-set could completely cover various cases to appear tomorrow
  - New Words: Singlish, bioinformatics, etc.
  - New Usage: missing verb in "Which one?" …
Various Kinds of Ambiguities (1)

• Sentence Segmentation:
  - An English “Period” might not be the Sentence-Delimiter (e.g., Eq. 3, Mr. King, etc.)
  - Should Chinese “Comma” be regarded as the Sentence-Delimiter?

• Tokenization:
  - English Split-Idiom and Compound-Noun matching
  - Word Segmentation (no space between Chinese chars)
    - [[土地公有] 政策] or [[土地公] 有政策]?
    - [這個[討論][會]很熱烈] or [這個[討論會]很熱烈]?
  - Order: tens of candidates per sentence
Various Kinds of Ambiguities (2)

• Lexical: “The current design ....”
  – "current": noun vs. adjective
  – “design": noun vs. verb

Ø Order: hundreds of candidates per sentence
Various Kinds of Ambiguities (3)

• Syntactic:
  – “[I saw [the boy [in the park]] [with a telescope]]"
  – “[I saw [the boy [in the park [with a telescope]]]]"
  ● Order: several hundreds to thousands (even millions)

↔ Analogy in artificial language: dangling-else problem [Aho 86]
  • " [ IF (...) then [ IF (...) then (...) ELSE (...) ] ] "
  • " [ IF (...) then [ IF (...) then (...) ] ELSE (...) ] “
  • Choose the nearest “IF”, if not particularly specified
Various Kinds of Ambiguities (4)

• Semantic:
  – Lexicon-Sense:
    • Eng: Bank (money vs. river), Take (90+ senses)
    • Chn: 方便, 生產, 打 (打人, 打電話, 打蛋, 打電玩, ...)
  – Case: Agent vs. Patient
    --- "[The police] were ordered [to stop drinking] by midnight”
    --- “他們答應 張三 可以 參加 會議.”

• Pragmatic:
  – Example: “你好厲害哦” (different intentions)
Various Kinds of Ill-formededness (1)

• Unknown Words (not found in dictionaries)
  – Due to *vocabulary size, proper noun, typing error, breeding words* (e.g., Singlish: Singapore English; Linsanity), *new technical terms* (e.g., bioinformatics), *abbreviation* (NLP, ML, 海陸兩會, 客服, 活儲, 新流感)

• Incomplete Lexicon Information (Known Words)
  – Known usage, but missing desired information (e.g., part-of-speech)
  – New usage (e.g., “Please xerox a copy for me.”)
Various Kinds of Ill-formedness (2)

- Un-grammatical sentences (cannot be parsed by the given grammar)
  - Example: "Which one?", “這部電影不錯看”

- Semantic constrain violation
  - Example: “My car drinks gasoline like water.” (subject-verb agreement), “他很機車”

- Ontology violation
  - Example: There is a plastic bird on the desk. Can this bird fly? (Sowa 2000).
Translation is hard (even by human)

• Sometimes the “same” phrase means different things in different geographical areas:
  – Example: Knock somebody up (Margaret King)
    • Wake them in the morning
    • Get them pregnant

• Sometimes contradictory phrases might mean the same thing in different geographical areas:
  – Example: Valid Ticket and Invalid Ticket (Martin Kay)
Machine Translation is harder

• The computer system has to make choices even when the human isn’t (normally) aware that a choice exists.

– Example from Margaret King:
  • The farmer’s wife sold the cow because she needed money.
  • The farmer’s wife sold the cow because she wasn’t giving enough milk.

– Another example:
  • The mother with babies under four....
  • The mother with babies under forty....
Table of Content

• A Simple Example: Syntactic Machine Translation
  – Process Flow

• Main Problems in NLP
  – Ambiguity & Ill-formedness

• Main Tasks in NLP
  – Tasks required in NLP, Bottlenecks

• Detailed Steps Overview
  – Analysis, Transfer and Generation
Main Tasks for Building NLP Systems (1)

• Knowledge Representation
  – How to organize and describe intra-linguistic, inter-linguistic, and extra-linguistic knowledge.
  – How to make knowledge known to the computer?

• Knowledge Control Strategies
  – How to efficiently use knowledge for ambiguity resolution and ill-formedness recovery
Main Tasks for Building NLP Systems (2)

• Knowledge Integration
  – How to jointly consider the information from different stages (e.g., syntactic score, semantic score, etc.): Natural language contains redundant information in different levels, they will enhance each other if they can be jointly considered
  – How to jointly consider knowledge from various sources (e.g., WordNet, Hownet, various dictionaries, translation-memory, etc.)

• Knowledge Acquisition
  – How to systematically and cost-effectively set up knowledge bases
  – How to maintain the consistency of knowledge base
Main Bottleneck: Knowledge Acquisition (1)

- Knowledge Acquisition is usually the bottleneck
  - Language usage is complex (not governed by any simple and elegant model), and dynamic (which changes with different groups, locations, and time)
  - Required knowledge is huge, messy and fine-grained
  - Inducing rules by human is usually very expensive, and time-consuming
  - Consistency is difficult to maintain when the system scales up
Main Bottleneck: Knowledge Acquisition (2)

• Seesaw phenomenon is generally observed
  – Traditional rule-based approaches are very hard to ensure global improvement, even if it is possible. (Human can only jointly consider 5-9 objects at the same time.)

• Need cheap and systematic ways to acquire knowledge
  – Complex problems need a large amount of knowledge, which is very difficult and expensive to build and maintain
  – Machine Learning seems to be the only way to go
Acquisition and Representation (1)

• Knowledge can be represented in different forms (e.g., “the/det, design/v/n”)
  – Knowledge can be represented either explicitly (such as rules) or implicitly (such as parameters).
    • Example 1: IF \[ C_{i-1} \text{ is } Det \], then \[ C_i \text{ cannot be a } Verb \]
    • Example 2: \[ P(C_i = Verb \mid C_{i-1} = Det) = 0 \]
    • Example 3: weighting coefficients in neural-network

– We usually classify various approaches by their associated Knowledge Representation Form (e.g., Rule-Based, Example-Based, Statistics-Based, etc.)
Acquisition and Representation (2)

• The Task of Knowledge Acquisition is closely coupled with its Knowledge Representation Form
  – Changing the Knowledge Representation Form usually also changes the way to acquire knowledge (Rules <= human, Parameters <= computer)

• We should consider the Knowledge Representation Form from the Knowledge Acquisition point of view
  – Since Knowledge Acquisition is the bottleneck, we should consider it first
  – First select the suitable knowledge acquisition mode, then decide the corresponding appropriate knowledge representation form
Table of Content

• A Simple Example: Syntactic Machine Translation
  – Process Flow

• Main Problems in NLP
  – Ambiguity & Ill-formedness

• Main Tasks in NLP
  – Tasks required in NLP, Bottlenecks

• Detailed Steps Overview
  – Analysis, Transfer and Generation
Detailed Steps: Analysis (1)

- **Text Pre-processing (separating texts from tags)**
  - Clean up garbage patterns (usually introduced during file conversion)
  - Recover sentences and words (e.g., `<B>C</B> omputer`)
  - Separate Processing-Regions from Non-Processing-Regions (e.g., File-Header-Sections, Equations, etc.)
  - Extract and mark strings that need special treatment (e.g., Topics, Keywords, etc.)
  - Identify and convert markup tags into internal tags (de-markup; however, markup tags also provide information)
Detailed Steps: Analysis (2)

• Discourse and Sentence Segmentation
  – Divide text into various primary processing units (e.g., sentences)
  – Discourse: Cue Phrases
  – Sentence: mainly classify the type of “Period” and “Carriage Return” in English ("sentence stops" vs. “abbreviations/titles”)

• Stemming
  – English: perform morphological analysis (e.g., -ed, -ing, -s, -ly, re-, pre-, etc.) and Identify root form (e.g., got <get>, lay <lie/lay>, etc.)
  – Chinese: mainly detect suffix lexemes (e.g., 孩子們, etc.)
  – Text normalization: Capitalization, Hyphenation, …
Detailed Steps: Analysis (3)

- **Tokenization**
  - English: mainly identify compound/idiom split-idiom (e.g., turn on) and split-idiom (e.g., turn NP on)
  - Chinese: Word Segmentation (e.g., [土地] [公有] [政策])
  - Regular Expression: numerical strings/expressions (e.g., twenty millions), date, ... (each being associated with a specific type)

- **Named Entities Recognition**
  - Detect Person-Name (e.g., Bill Clinton), Location-Name (e.g., Taipei city, Taiwan), Organization-Name (e.g., Institute of Information Science), Time (September 16, 2014)
  - Decide Boundaries and Type
Detailed Steps: Analysis (4)

- **Tagging**
  - Assign Part-of-Speech (e.g., n, v, adj, adv, etc.)
  - Associated forms are basically independent of languages starting from this step

- **Parsing**
  - Decide suitable syntactic relationship (e.g., PP-Attachment)

- **Decide Word-Sense**
  - Decide appropriate lexicon-sense (e.g., River-Bank, Money-Bank, etc.)

- **Assign Case-Label**
  - Decide suitable semantic relationship (e.g., Patient, Agent, etc.)
Detailed Steps: Analysis (5)

- **Anaphora Resolution**
  - Pronoun reference (e.g., “he” refers to “the president”)

- **Decide Discourse Structure**
  - Decide suitable discourse segments relationship (e.g., Evidence, Concession, Justification, etc. [Marcu 2000].)
Detailed Steps: Analysis (6)

• Convert into Logical Form (Optional)
  – Co-reference resolution (e.g., “president” refers to “Bill Clinton”), scope resolution (e.g., negation), Temporal Resolution (e.g., today, last Friday), Spatial Resolution (e.g., here, next), etc.
  – Determine IS-A (also Part-of) relationship, etc.
  – Mainly used in inference related applications (e.g., Q&A, etc.)
Detailed Steps: Transfer (1)

• Decide suitable Target Discourse Structure
  – For example: Evidence, Concession, Justification, etc. [Marcu 2000].

• Decide suitable Target Lexicon Senses
  – Sense Mapping may not be one-to-one (sense resolution might be different in different languages, e.g. “snow” has more senses in Eskimo)
  – Sense-Token Mapping may not be one-to-one (lexicon representation power might be different in different languages, e.g., “DINK”, “睨”, etc). It could be 2-1, 1-2, etc.
Detailed Steps: Transfer (2)

- Decide suitable Target Sentence Structure
  - For example: verb nominalization, constitute promotion and demotion (usually occurs when Sense-Token-Mapping is not 1-1)

- Decide appropriate Target Case
  - Case Label might change after the structure has been modified
  - (Example) verb nominalization: “… that you (AGENT) invite me” $\Leftrightarrow$ “… your (POSS) invitation”
Detailed Steps: Generation (1)

• Adopt suitable Sentence Syntactic Pattern
  – Depend on Style (which is the distributions of lexicon selection and syntactic patterns adopted)

• Adopt suitable Target Lexicon
  – Select from Synonym Set (depend on style)

• Add “de” (Chinese), comma, tense, measure (Chinese), etc.
  – Morphological generation is required for target-specific tokens
Detailed Steps: Generation (2)

- Text Post-processing
  - Final string substitution (replace those markers of special strings)
  - Extract and export associated information (e.g., Glossary, Index, etc.)
  - Restore customer’s markup tags (re-markup) for saving typesetting work
Common Phenomena in NLP

• An early step might need information from later steps
  – For example, identifying split-idiom in the Tokenization step needs to verify a specified constituent (e.g., turn NP on)
  – One way to handle that is to adopt a Black-Board approach; however, it is not efficient (ref. Verbmobile report [Wahlster 00]).

• Output may not be unique
  – Zero, when a rule-based approach encounters ill-formed input
  – Usually several candidates are possible (even under Unification Grammar Formalism)