A Classical Chinese Corpus with Nested Part-of-Speech Tags

John S. Y. LEE
City University of Hong Kong
24 March, 2012
Presentation at LaTeCH Workshop, EACL
Outline

• Introduction
• Previous Work
• Corpus Design
• Data & Evaluation
• Conclusion
Introduction

• Chinese text corpora with linguistic annotations
  – Part-of-speech and syntactic structures
    • E.g., Lancaster Corpus [McEnery & Xiao 2004] ; Penn Chinese Treebank [Xue et al. 2005]

• Uses:
  – Teaching material of foreign language
  – Data for empirical linguistics research
  – Training data for natural language processing tasks
Introduction

• Central issue: word segmentation

  – A Chinese character may combine with its neighbor(s) to form a multi-character word
    • E.g., 他/喜歡/蘋果 ‘he/likes/apples’

  – Some cases are less clear-cut
    • E.g., should “Peking University” can be
      – one word “北京大學” ‘Peking University’
      – two words “北京/大學” ‘Peking/University’
Introduction

• General test for wordhood: “Is the meaning of the whole compositional of its parts?”

  – E.g., 沙門 ‘Buddhist monk’ is one word
    • The word bears no semantic relation with its constituent characters 沙 ‘sand’ and 門 ‘door’
Introduction

• Why word segmentation?
  – To enable part-of-speech (POS) tagging
    • Defines smallest unit for POS analysis, e.g., 他/pronoun 喜歡/verb 蘋果/noun
    • Meaningless to analyze the POS of 沙門 ‘Buddhist monk’ as 沙/noun and 門/noun
    • Critical for Chinese since it has no inflectional morphology
Introduction

• Why word segmentation? (cont’d)
  – To enable automatic word retrieval
    • E.g., search for the adjective “個人” ‘personal’ should not return the phrase “一個人為錯誤” ‘a human error’
    • Segmentation “一個人為錯誤” removes the ambiguity
Introduction

• No standard for word segmentation
  – Native speakers can agree on word boundaries in modern Chinese only about 76% of the time [Sproat et al., 1996]

• Best segmentation depends on application
  – Searching geographical terms?
    • 黃河 ‘Yellow River’ should be considered one word
  – Searching parallelisms in poems?
    • 黃河 should be two words, 黃/adj 河/noun
Introduction

• To settle on any particular word segmentation criterion is to risk of omitting useful information
  – The verb-object combination is “among the hardest cases for the word definition” [Xia, 2000]

• We propose nested POS tags
  – E.g., (黃/adj 河/noun)/proper_noun
  – Accommodate different levels of granularity
Outline

• Introduction
• Previous Work
• Corpus Design
• Data & Evaluation
• Conclusion
Previous Work

- **Corpus at Jiaotong University** [Huang et al. 2006]
  - 1000 sentences of pre-Tsin classical Chinese
  - No word segmentation is performed

- **Sheffield Corpus of Chinese** [Hu et al. 2005]
  - 109K characters of archaic Chinese, 147K characters of medieval Chinese
  - No details on word segmentation criteria
Previous Work

• Academia Sinica Ancient Chinese Corpus [Wei et al. 1997]
  – Largest word-segmented and POS-tagged corpus of classical Chinese: more than 500K characters
  – Recognizes following categories of words
    • parallel and subordinating compounds
    • proper nouns
    • reduplications
Previous Work

• Penn Chinese Treebank [Xue et al. 2005]
  – “Complex internal structures should be segmented when possible”
  – Number of syllables
    • E.g., 北京大|學 is segmented as two words since 大|學 has two bisyllabic
    • E.g., 北京市 is one word since 市 is monosyllabic
  – Bound morphemes
    • e.g., the morpheme 本 is bound with the character 人 in the word 本人 ‘oneself’
Previous Work

• Penn Chinese Treebank (cont’d)
  – Semantic compositionality
  – Insertion test
    • “can another morpheme be inserted between the two characters?”
  – XP-substitution test
    • “can the morpheme replaced by a phrase of the same type?”
Previous Work

• POS tagsets differ in complexity
  – All cover major categories
    • E.g., nouns, verbs, and adjectives
  – But there may or may not be sub-categories
    • E.g., transitive vs. intransitive verbs; common, proper or temporal nouns

<table>
<thead>
<tr>
<th>Corpus</th>
<th>POS tagset size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiaotong University corpus</td>
<td>21</td>
</tr>
<tr>
<td>Peking University corpus</td>
<td>26</td>
</tr>
<tr>
<td>Penn Chinese Treebank</td>
<td>33</td>
</tr>
<tr>
<td>Academia Sinica Balanced Corpus</td>
<td>46</td>
</tr>
<tr>
<td>Sheffield Corpus of Chinese</td>
<td>111</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• Previous Work
• Corpus Design
  – Strings without internal structure
  – Strings with internal structures
• Data & Evaluation
• Conclusion
Corpus Design

• Two-level annotation in Penn Chinese Treebank [Xia 2000]
  – Used in a limited number of constructions
    • E.g., (走/verb 上來/verb)/verb ‘walk up’

• Generalized as nested POS tags in our corpus
Outline

• Introduction
• Previous Work
• Corpus Design
  – Strings without internal structure
  – Strings with internal structures
• Data & Evaluation
• Conclusion
Strings without internal structure

- Annotated at the lowest level
- Roughly equivalent to “words” in other corpora
  - But stricter definition here

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign loanwords</td>
<td>匈奴/NR 圍酒泉</td>
</tr>
<tr>
<td>Numbers</td>
<td>少年十五/CD 十六/CD時</td>
</tr>
<tr>
<td>Reduplications</td>
<td>車馬去騏騏/AD</td>
</tr>
<tr>
<td>Bound morphemes</td>
<td>天油然/AD作雲</td>
</tr>
</tbody>
</table>
Strings without internal structure

• POS tagset of the Penn Chinese Treebank adapted for Classical Chinese
  – Most commonly used tagset
  – Prospect of leveraging existing modern Chinese annotations as training data
  – Tags are assigned not according to the meaning of the word, but the syntactic distribution [Xia, 2000]
Strings without internal structure

• Tagging according to syntactic distribution
  – E.g., 深山何處 鐘/verb
    • ‘where in the deep mountain [is it] ringing’
    • Noun ‘bell’ tagged as verb
  – E.g., 倖忽 雲/adv 散
    • ‘quickly disperses like clouds’
    • Noun ‘cloud’ tagged as adverb
  – E.g., 送君 盡/adv 惆悵
    • ‘saying farewell to you, I am utterly sad’
    • Verb ‘exhaust’ tagged as adverb
Strings without internal structure

• Most POS tags have counterparts in Classical Chinese
  – Some have roughly equivalent word classes
    • E.g., P (preposition) tag is retained for words known as coverbs [Pulleyblank, 1995]
  – Tags specific to modern Chinese are discarded
    • E.g., DER (resultative 得 de), DEV (manner 地 de), and FW (words in foreign alphabet)
Outline

• Introduction
• Previous Work
• Corpus Design
  – Strings without internal structure
  – Strings with internal structures
• Data & Evaluation
• Conclusion
Strings with internal structures

• Annotated at higher levels
• Three types:
  – Parallel compounds
  – Subordinating compounds
  – Proper names
Strings with internal structures

• Parallel compound
  – two-character noun, verb and adjective “in which neither member dominates the other” [Packard, 1998]

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar meaning</td>
<td>君王 ‘king’ = 君 ‘ruler’ + 王 ‘king’ (君/NN 王/NN)/NN</td>
</tr>
<tr>
<td>Related meaning</td>
<td>骨肉 ‘kin’ = 骨 ‘bone’ + 肉 ‘flesh’ (骨/NN 肉/NN)/NN</td>
</tr>
<tr>
<td>Opposite meaning</td>
<td>是非 ‘rumors’ = 是 ‘right’ + 非 ‘wrong’ (是/JJ 非/JJ)/NN</td>
</tr>
</tbody>
</table>
Strings with internal structures

- **Subordinating compounds**
  - Strings where “one member (the modifier) is subordinate to and modifies the other (the head)” [Packard, 1998]

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb-object</td>
<td>識事 ‘experience’ = 識 ‘understand’ + 事 ‘affairs’ (識/VV 事/NN)/NN</td>
</tr>
<tr>
<td>Subject-verb</td>
<td>日落 ‘sunset’ = 日 ‘sun’ + 落 ‘descend’ (日/NN 落/VV)/NN</td>
</tr>
<tr>
<td>Adjectival modifier</td>
<td>少年 ‘youth’ = 少 ‘few’ + 年 ‘year’ (少/JJ 年/NN)/NN</td>
</tr>
<tr>
<td>Noun modifier</td>
<td>家食 ‘household food’ = 家 ‘house’ + 食 ‘food’ (家/NN 食/NN)/NN</td>
</tr>
</tbody>
</table>
Strings with internal structures

• Proper names
  – geographical names
    • E.g., 黃河 huang he ‘Yellow River’
    • adjective modifying a noun
  – personal names with titles
    • E.g., 始興公 shi xing gong ‘Duke Shixing’
    • noun modifying a noun
Strings with internal structures

• Definition of ‘strings with internal structures’ is deliberately broad
  – Nested annotations provided for all such strings
  – User to decide which level of tags is suitable for the research objective at hand
Strings with internal structures

• Benefits (1): Provides information for alternate syntactic analyses:
  – E.g., 天氣晚來秋
    • ‘the weather turns chilly as the evening comes’
      – if 晚來 is considered two words forming a subordinate clause: 晚/noun 來/verb
    • ‘the weather turns chilly at evening’
      – if 晚來 is considered one word as a temporal noun ‘evening’: 晚來/NT

天氣/noun (晚/noun 來/verb)/temporal_noun 秋/adj
Strings with internal structures

• Benefits (2): Indicates parallelism
  – Tagging 黃河 ‘Yellow River’ as one word (proper noun) would have obscured the parallelism in this couplet

<table>
<thead>
<tr>
<th>獨</th>
<th>樹</th>
<th>臨</th>
<th>關</th>
<th>門</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘lone tree’</td>
<td>‘watches the entrance of the pass’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JJ</td>
<td>NN</td>
<td>VV</td>
<td>NN</td>
<td>NN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>黃</th>
<th>河</th>
<th>向</th>
<th>天</th>
<th>外</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Yellow River’</td>
<td>‘faces the outer sky’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JJ</td>
<td>NN</td>
<td>VV</td>
<td>NN</td>
<td>LC</td>
</tr>
<tr>
<td>NR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Strings with internal structures

• Benefits (3): Nesting at arbitrary depth
  – E.g., three-level annotation for 細柳營 ‘Camp of little willows’ reveals its parallelism at all levels with its counterpart, 新豊市 ‘City of new abundance’

<table>
<thead>
<tr>
<th>細</th>
<th>柳</th>
<th>營</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td>NR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

‘Little Willow camp’
Outline

• Introduction
• Previous Work
• Corpus Design
• Data & Evaluation
• Conclusion
Data

- Proposed word segmentation and POS tagging scheme applied to 8th-century classical Chinese poems
  - Complete works of Wang Wei and Meng Haoran
    - 521 poems
    - ~32,000 characters
Evaluation

• Word segmentation
  – i.e. segmentation of ‘strings without internal structure’
  – Performed on 1,057 character by two annotators
  – Disagreement on this task was 1.7%
Evaluation

• POS tagging
  – For strings without internal structures, disagreement was 4.9%
  – Confusion between verbs and adverbs
    • E.g., 伏檻紓三顧
      – 紓 ‘bow’ can be seen as a verb or an adverb ‘respectfully’
  – Confusion between adjective and noun
    • E.g., 命服
      – 命 ‘profession’ may be a noun modifier or adjectival modifier of 服 ‘clothes’
Evaluation

• POS tagging (cont’d)
  – For ‘strings with internal structures’, both precision and recall were 83.5%
  – Disagreement on proper names
  – Most of the discrepancies are direct results of differences in POS tagging of ‘strings without internal structures’
Outline

• Introduction
• Previous Work
• Corpus Design
• Data & Evaluation
• Conclusion
Conclusion

• We described a novel method of word segmentation and POS tagging tailored for the classical Chinese language
  – Applied on more than 32,000 characters
  – Inter-annotator agreement compares favorably with past annotation projects
  – Dependency treebank have been created [to appear in NAACL-HLT 2012]