Existing basic NLP tools and resources in KU

Vee Satayamas

Specialty Research Unit in Natural Language Processing and Intelligent Information System Technology, Kasertsart University

vee.sa@ku.th

August 16, 2017

Overview

1. Language resources
   - Word boundary and part-of-speech corpus
   - Treebank
2. Tools
   - Word segmentation
   - Part-of-speech tagging
   - Syntactic parser
   - Treebank Editor
   - Word boundary editor
   - Treebank annotation consistency checker
   - DEMO

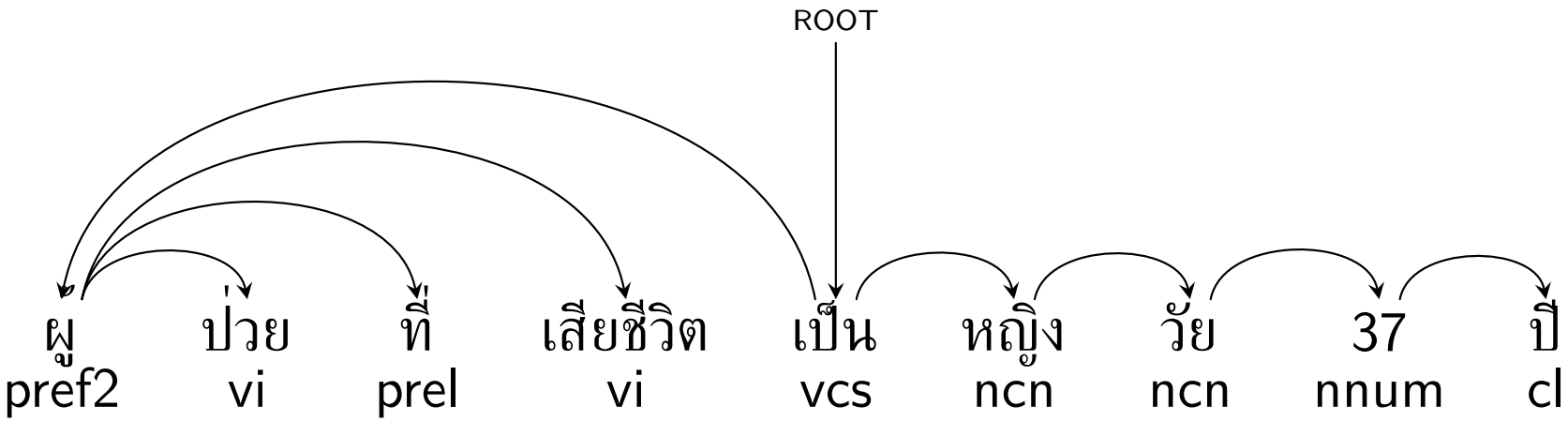
Word boundary and part-of-speech annotated corpus

new1-2-v.1
[TMB/npn _/blk แจ้ง/vt การ/pref1 # ชำระ/vt บัญชี/ncn ของ/prep บริษัท/ncn ที่/prel ธนาคาร/ncn เข้า/vi ลงทุน/vi]
[ตามที่/conj ธนาคาร/ncn ทหารไทย/npn _/blk จำกัด/adj _/blk /*punc*/*ncn*/punc _/blk ได้/prev ลงทุน/vi ใน/prep บริษัท/ncn _/blk สินพหล/npn _/blk จำกัด/adj _/blk]

Word boundary and part-of-speech corpus

- #sentences 40750
- #words 695536

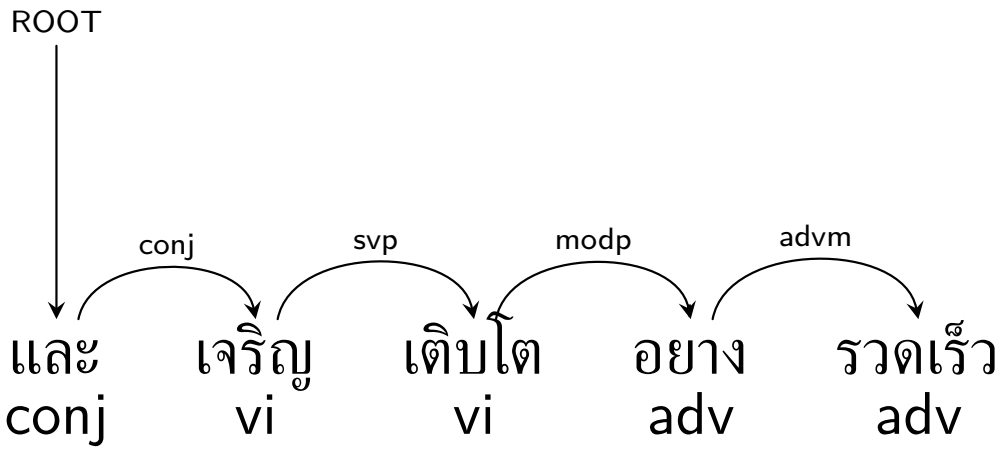
Treebank


ROOT
ผู
pref2
ป่วย
vi
ที่
prel
เสียชีวิต
vi
เป็น
vcs
หญิง
ncn
วัย
ncn
37
nnum
ปี
cl

Treebank: Internal

```
<annot rev="4" user="da">
  <node dep_attr="" pos="vcs" type="dependency">
   <snode>
      <span s="19" e="23">เป็น</span>
   </snode>
   <node dep_attr="" pos="ncn" type="dependency">
    <snode>
       <span s="23" e="27">หญิง</span>
    </snode>
    <node dep_attr="" pos="ncn" type="dependency">
     <snode>
        <span s="27" e="30">วัย</span>
     </snode>
     <node dep_attr="" pos="nnum" type="dependency">
      <snode>
         <span s="30" e="32">37</span>
      </snode>
```

Treebank: dependency label



Treebank: size

12031 sentences

KUCut

- An unsupervised learning-based word breaker[6]
- Segmenting words that are not in word list by ML
- https://bitbucket.org/veer66/kucut/

POS tagging

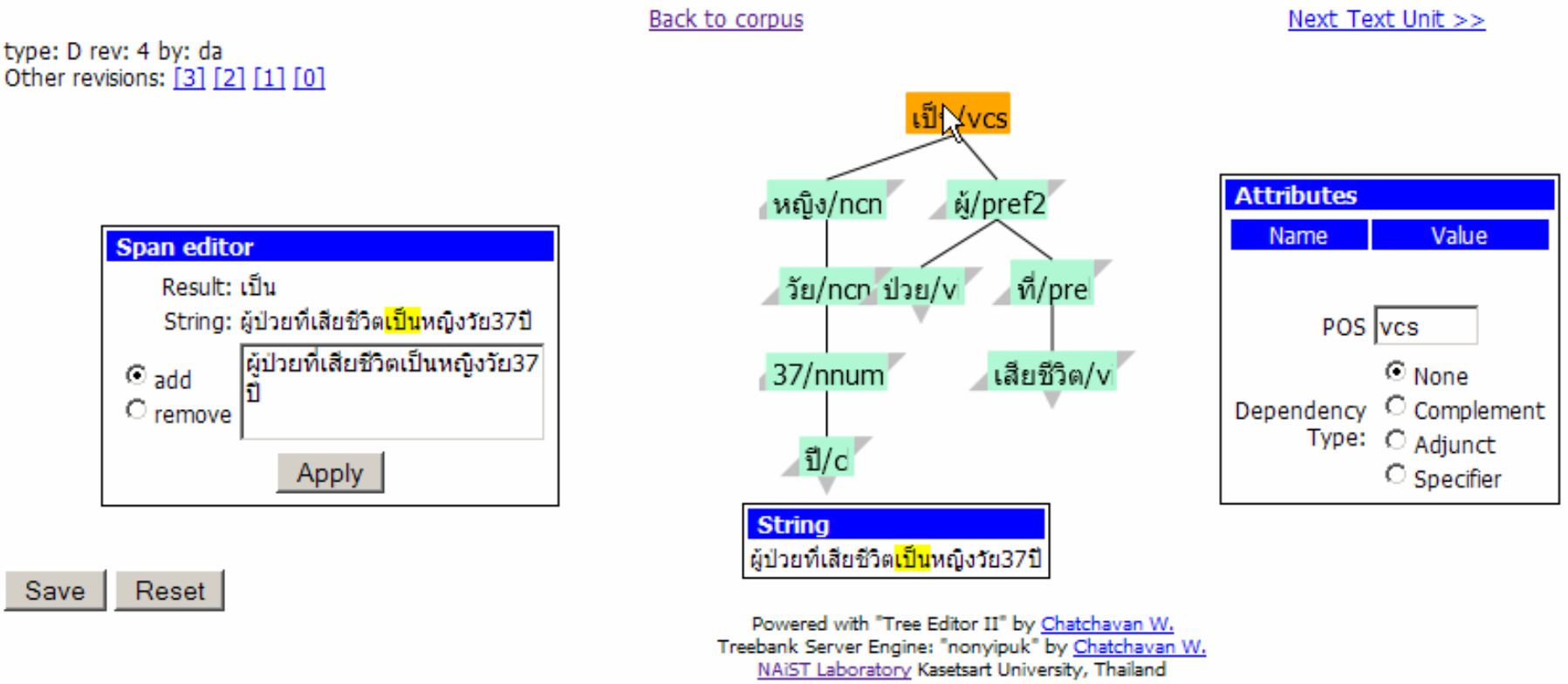
- Use existing HMM-based tagging engine[1]
- Trained using our part-of-speech annotated corpus

Syntactic parser

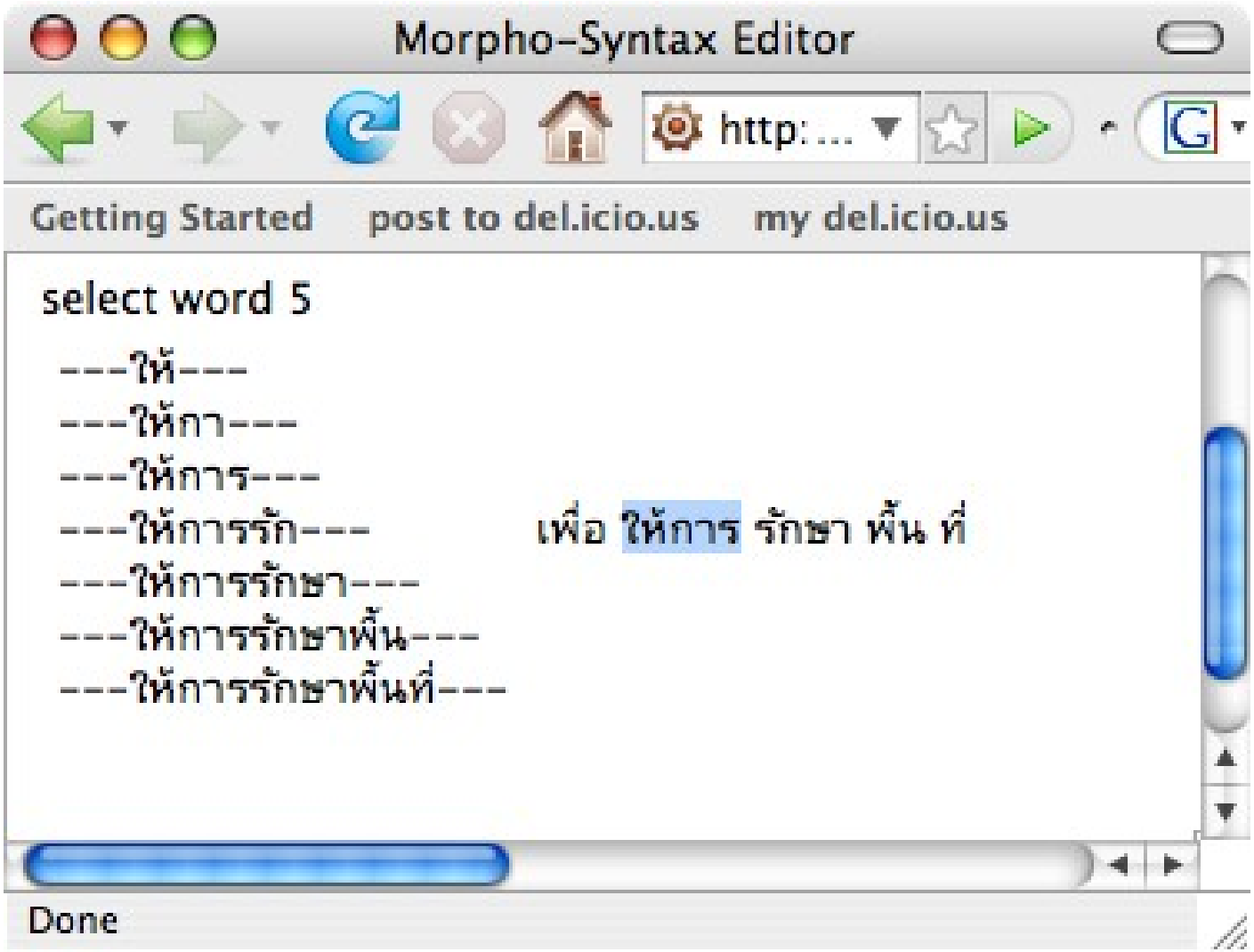
- Dependency parsing based on word segmentation with part-of-speech lattace[5]
- Use dependency treebank as training set
- Accuracy: 78.02%

Treebank Editor

Wiki-like treebank editor[7]



Word boundary editor[4]



Treebank annotation consistency checker

Searching for syntactic tree annotations, which are annotated differently in the same of similar circumstances[3]

Demo: Word segmentation and part-of-speech tagging

- Yaito[2]
- Jitar[1]
- Our word boundary and part-of-speech annotated corpus

References

[1] Daniël de Kok. Jitar HMM part of speech tagger. `https://github.com/danieldk/jitar`, 2014. [Online; accessed 15-August-2017].

[2] Vee Satayamas. A word tokenizer for ASEAN languages written in Kotlin. `https://gitlab.com/veer66/yaito`, 2017. [Online; accessed 15-August-2017].

[3] Vee Satayamas and Asanee Kawtrakul. Wide-coverage grammar extraction from thai treebank. In *Proceedings of Papillon 2004 Workshops on Multilingual Lexical Databases*, 2004.

[4] Vee Satayamas, Asanee Kawtrakul, and Christian Boitet. Annoted-w, a specialized editor for annotating word boundaries collaboratively. In *The seventh international Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Chonburi, Thailand, December 2007.

[5] Sutee Sudprasert, Asanee Kawtrakul, Christian Boitet, and Vincent Berment. Dependency parsing with lattice structures for resource-poor languages. *IEICE TRANSACTIONS on Information and Systems*, 92(10):2122–2136, 2009.

[6] สุธี สุดประเสริฐ. การตัดคำภาษาไทยโดยการเรียน รู้จากคลังเอกสาร และเอกสารนำเข้าแบบไม่มีผลเฉลย. 2003.

[7] Chatchavan Wacharamanotham, Mukda Suktarachan, and Asanee Kawtrakul. The development of web-based annotation system for thai treebank. In *The seventh international Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Chonburi, Thailand, December 2007.