Ambiguity Resolution for Vt-N Structures in Chinese

Yu-Ming Hsieh[1,2] Jason S. Chang[2] Keh-Jiann Chen[1]

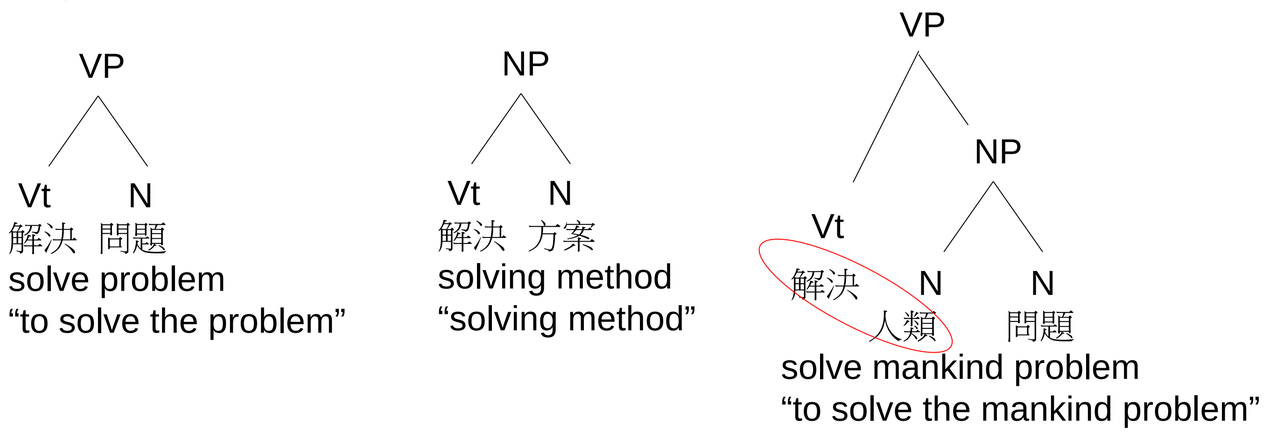
[1] Institute of Information Science, Academia Sinica, Taipei

[2] Department of Computer Science, National Tsing-Hua University, Taiwan

morris@iis.sinica.edu.tw jason.jschang@gmail.com kchen@iis.sinica.edu.tw

Problem Statement and Analysis

- In Chinese, the structure of a transitive verb (Vt) immediately followed by a noun (N) may form a verb phrase (VP), a noun phrase (NP), or there may not be a dependent relation, as shown below.

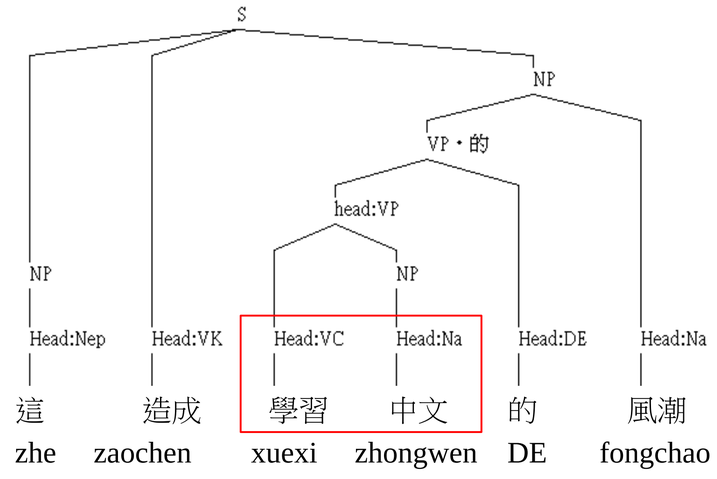


- Parsers generally prefer to the VP reading due to the statistical majority.

VP	NP	Other Relation
58%	16%	26%

Table 1. Statistical data from the Sinica Treebank.

- Unlike inflectional languages, Chinese verbs modify nouns without morphological inflection, e.g., 養殖/farming 池/pond.
- Linguistically motivated features to build a Vt-N classifier include lexical words, semantic knowledge, the morphological structure of verbs, neighboring parts-of-speech, and the syllabic length of words.



“This causes the trend of learning Chinese.”

Figure 1. An example of a tree with a Vt-N structure.

Feature Extraction

Feature Type	x	y
Word	w_1=學習	w_2=中文
PoS	t_1=VC	t_2=Na
Semantic	st_1=study|學習	st_2=language|語言
Context	t_{-2}=Nep; t_{-1}=VK; t_3=DE; t_4=Na	
Additional Information	Vmorph=VV	Nlen=2
Relation Type	rt = H/X	

Table 3. The features labels of Vt-N pair in Figure 1.

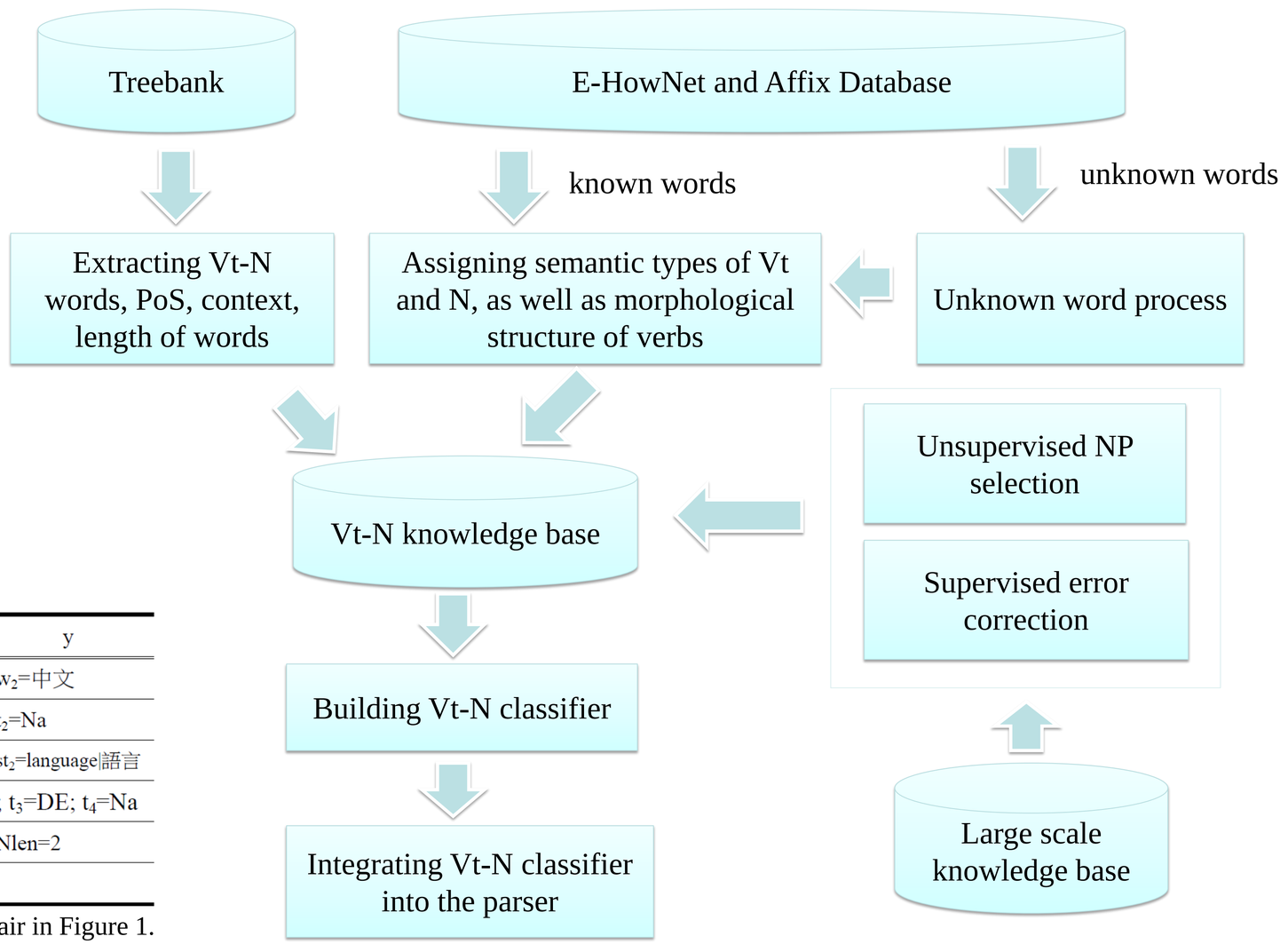
- Analysis of NP(Vt-N) structures in the Sinica Treebank reveals the following four types of semantic structures.

Types	Examples
Telic(Vt) + Host(N)	研究/工具 research tool 探測/機 detective machine
Host-Event(Vt) + Attribute(N)	攻擊/策略 attacking strategy 書寫/內容 written context
Agentive(Vt) + Host(N)	叫/聲 shouting sound 炸/雞 fried chicken
Apposition(Vt) + Affair(N)	追撞/事故 collision accident 破壞/運動 destructive movement

Table 2. Semantic types of NP(Vt-N) and examples.

- Difficulties:
  - Feature selection. Decision on which features to adopt and their combination is a difficult task in classification.
  - Unknown word issue. Unknown word processing has technical problems that affect the prediction of the semantic types and morph-structures of unknown words.
  - Data sparseness problem. Due to the limited size of the current Treebank, we should mine useful information from all available resources.

Proposed Models



Experiment Results

Evaluation of the Vt-N Classifier

- The results of Table 4 show that contextual information (M2) and lexical feature (M3) are the most important feature used to disambiguate VP, NP and independent structures. Results also demonstrate the benefits of using the semantic type (M4), verb morph-structure and noun length features (M5).

Models	Feature for Vt-N	P1(%)	P2(%)
M1	(t_1,t_2)	61.94	59.10
M2	+ (t_{-1},t_1) (t_2,t_3) (t_{-2},t_{-1},t_1) (t_2,t_3,t_4) (t_{-1},t_3)	76.59	72.30
M3	+ (w_1,t_1,w_2,t_2) (w_1,w_2) (w_2) (w_1)	83.55	80.20
M4	+ (st_1,t_1,st_2,t_2) (st_1,t_1) (st_2, t_2)	84.63	81.90
M5	+ (Vmorph) (Nlen)	85.01	83.00

Table 4. The results of using different feature combinations. P1(%) is the 10-fold cross validation accuracy of the training data; P2(%) is the accuracy of the test data.

Knowledge from Large-scale Unlabeled Data

- We applied the data selection method (*distance=1, with an intransitive verb followed by an object noun*) and data correction method to learn more useful knowledge.
- The results in Table 5 show that the proposed methods can improves the accuracy.

	Treebank+ ASBC+Correction	Treebank+ ASBC-Vi-N	Treebank
size of training instances	56,521	46,258	8,017
M5 - P2(%)	88.40	83.90	83.00

Table 5. Experiment results of classifiers with different training data.

Vt-N classifier and PCFG Parser Integration

- We integrate the Vt-N models into the PCFG parser. The formula of the integrated structural evaluation model is as follows:

$$Score(T,S)=\sum_{i=1}^{n}(w_1\times RP_i+w_2\times VtNP_i)$$

	PCFG + M5 (Treebank+ASBC+Correction)	PCFG + M5 (Treebank)	PCFG
P2(%)	87.88	80.68	77.09
BF(%)	84.68	83.64	82.80

Table 6. The performance of the PCFG parser with and without model M5. The BF (*bracketed f-score*) is the parsing performance metric.