Toward Active Learning in Data Selection: Automatic Discovery of Language Features During Elicitation

Jonathan Clark
Robert Frederking
Lori Levin

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
Feature Detection

• **Grammatemes** - Language features that express grammatical meanings (such as number, person, tense)

• Given a set of **grammatemes** and a **structured corpus**, can we determine if these grammatemes are expressed in a particular language?

• e.g. Answers “Does this language distinguish singular nouns from plural nouns?” (“And if so, how?”)

* Source: Alena Böhmová, Silvie Cinková, Eva Hajičová. Annotation on the tectogrammatical layer in the Prague Dependency Treebank. 2005.
Feature Detection

The dog sleeps
((num sg)…)
The dogs sleep
((num dl)…)  
The dogs sleep
((num pl)...)
Feature Detection

The dog sleeps
((num sg)…)
The dogs sleep
((num dl)…)      
The dogs sleep
((num pl)...)    

犬が寝る
犬が寝る
犬が寝る
Feature Detection

The dog sleeps (num sg)...
The dogs sleep (num dl)...
The dogs sleep (num pl)...

犬が寝る
犬が寝る
犬が寝る

Marks Plural? NO
犬が寝る
犬が寝る

Marks Dual? NO
犬が寝る
犬が寝る

Bilingual Person

Feature Detection
Data Selection

• Given many potential training examples, select the ones that will help the target system most

• Many Uses - Seen in Speech Recognition, Speech Synthesis, and Machine Translation

• Corpus Navigation: Not all data is relevant for all languages

• Helps when money or time is limited

• e.g. Small Domains, MT Emergencies, and Minority Languages
Data Selection

The dog sleeps ((num sg)...)
The dogs sleep ((num dl)...)
The dogs sleep ((num pl)...)
Data Selection

Bilingual Person

The dog sleeps ((num sg)...)
The dogs sleep ((num dl)...)
The dogs sleep ((num pl)...)

犬が寝る

犬が寝る

犬が寝る

Marks Plural? NO
犬が寝る

Marks Dual? NO
犬が寝る

Feature Detection

Data Selection (Corpus Navigation)
Data Selection

The dog sleeps ((num sg)…)  
The dogs sleep ((num dl)…)  
The dogs sleep ((num pl)…)

Implicational Universal:  
No Plural Marking --> No Dual Marking

Marks Plural? NO  
犬が寝る

Marks Dual? NO  
犬が寝る

Bilingual Person

Data Selection (Corpus Navigation)
Data Selection

The dog sleeps ((num sg)…)
The dogs sleep ((num dl)…)
The dogs sleep ((num pl)…)

Bilingual Person

犬が寝る
犬が寝る

Marks Plural? NO
犬が寝る

Marks Dual? NO
犬が寝る

Feature Detection

Implicational Universal:
No Plural Marking --> No Dual Marking

Data Selection (Corpus Navigation)
Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies.
context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies.
context: Maria bakes cookies regularly or habitually.
srcsent: Maria \textbf{bakes} cookies.
tgtsent: Maria \textcolor{red}{hornea} galletas.
aligned: \((1,1),(2,2),(3,3),(4,4)\)
context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies.
tgtsent: Maria hornea galletas.
aligned: ((1,1),(2,2),(3,3),(4,4))

fstruct: [f1]( [f2](actor ((gender f)(anim human)(num sg)))
[f3](undergoer ((person 3) (num dl))) (tense pres))
cstruct: [n1](S1 [n2](S [n3](NP [n4](NNP Maria))
[n5](VP [n6](VBZ bakes) [n7](NP [n8](NNS cookies))))))

phimap: phi(n1)=f1; phi(n3)=f2; phi(n7)=f3;
headmap: h(n1)=n2; h(n2)=n5; h(n3)=n4; h(n4)=n4;
h(n5)=n6; h(n6)=n6; h(n7)=n8; h(n8)=n8;
Example Deduction Rule

# Perfective/Imperfective Aspect
(rule (sentences (A (aspect perfective))
(B (aspect progressive))})
Example Deduction Rule

# Perfective/Imperfective Aspect
(rule (sentences (A (aspect perfective))
              (B (aspect progressive)))
(overlap on)
Example Deduction Rule

# Perfective/Imperfective Aspect

(rule (sentences (A (aspect perfective))
           (B (aspect progressive)))
  (overlap on)
  (if 0.6 (different
    (target-lex (fnode (A)))
    (target-lex (fnode (B))))
  (then (WALS ”Perfective/Imperfective Aspect”
             ”Grammatical marking”))))
Example Deduction Rule

# Perfective/Imperfective Aspect
(rule (sentences (A (aspect perfective))
          (B (aspect progressive)))))

(overlap on)
(if 0.6 (different
    (target-lex (fnode (A))))
  (target-lex (fnode (B))))
(them (WALS "Perfective/Imperfective Aspect"
       "Grammatical marking"))
(if 0.4 (same
    (target-lex (fnode (A))))
  (target-lex (fnode (B))))
(them (WALS "Perfective/Imperfective Aspect"
       "No grammatical marking")))
Feature Detection Experiment

- Corpus of 60 Spanish-English sentences
- Tried to identify 21 features from the World Atlas of Language Structures

|          | Precision | Recall | F1   |
|----------|-----------|--------|------|
| Baseline | 12 / 21   | 12 / 21| 12 / 21 |
| Experimental | 19 / 21   | 19 / 21| 19 / 21 |
Toward Corpus Navigation

• Not all data is relevant for every language

• Performed while a linguistically naive bilingual person translates sentences in GUI

• After eliciting each sentence:
  * Apply feature detection
  * Choose the most valuable sentence to elicit next

• Leverages knowledge from Greenbergian Implicational Universals (from Hal Daume’s database learned from WALS)
Other Applications

• Learning feature-annotated closed-class morphemes

• Factored MT

• Selection of data for automatic grammar induction for syntactic and hybrid MT systems

• Aid for linguistics field work
Language Resources

• Result of Corpus Navigation is:

1. A resource **dense** with the “right” features

2. **Highly structured**; each language feature is linked with sentences that illustrate it

3. **Word-aligned, feature-annotated** sentences useful for studying divergences and MT
Toward Active Learning in Data Selection:
Automatic Discovery of Language Features During Elicitation

Questions?

Jonathan Clark
Robert Frederking
Lori Levin

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
| WALS Features for Experiment                  |
|---------------------------------------------|
| Gender Distinctions in Independent          |
| Personal Pronouns                          |
| Nominal and Locational Predication          |
| Occurrence of Nominal Plurality             |
| Order of Adjective and Noun                 |
| Order of Genitive and Noun                  |
| Order of Numeral and Noun                   |
| Order of Subject, Object and Verb           |
| Order of Subject and Verb                   |
| Order of Object and Verb                    |
| Perfective/Imperfective Aspect               |
| Politeness Distinctions in Pronouns         |
| Position of Interrogative Phrases in Content|
| Questions                                  |
| Position of Pronominal Possessive Affixxes  |
| Position of Tense-Aspect Affixes            |
| Inclusive/Exclusive Distinction in Independent Pronouns |
| Inclusive/Exclusive Distinction in Verbal Inflection |
| Semantic Distinctions of Evidentiality       |
| The Future Tense                            |
| Verbal Person Marking                       |
| ‘Want’ Complement Subjects                  |
| Zero Copula for Predicate Nominals          |
Production Predicates

fnode
in-order
source-lex
target-lex
*-uhead
*-ihead
same
present
not-present
Elicitation Corpus Availability

- Included in LDC’s Less Commonly Taught Languages (LCTL) Language Packs
- 13 languages have already been translated by the LDC
- Urdu language pack used in this year’s NIST MT Eval
- Bengali queued for general release this year