Inductive Detection of Language Features via Clustering Minimal Pairs

Toward Feature-Rich Grammars in Machine Translation

Jonathan Clark
Robert Frederking
Lori Levin

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA

ACL SSST-2 - June 20, 2008
Outline

- Overview of Feature Detection
- Example Application: Feature-Rich Grammars
- The Process of Feature Detection
- Results
- Conclusions
Feature Detection

((num sg)…)
The dog eats
El perro come

((num dl)…)
The dogs eat
Los perros comen

((num pl)…)
The dogs eat
Los perros comen
Feature Detection

context: There are two dogs

((num sg)...) The dog eats
El perro come

((num dl)...) The dogs eat
Los perros comen

((num pl)...) The dogs eat
Los perros comen
Feature Detection

((num sg)...) The dog eats El perro come
((num dl)...) The dogs eat Los perros comen
((num pl)...) The dogs eat Los perros comen
Feature Detection

Does this language distinguish singular, dual, or plural agents? If so, how?
Feature Detection

Does this language distinguish singular, dual, or plural agents? If so, how?

| Feature       | Candidate Lexical Items             |
|---------------|-------------------------------------|
| (num sg)      | come, el, perro                     |
| (num dl, pl)  | comen, los, perros                  |

((num sg)…) 
The dog eats El perro come

((num dl)…) 
The dogs eat Los perros comen

((num pl)…) 
The dogs eat Los perros comen
Outline

✓ Overview of Feature Detection
- Example Application: Feature-Rich Grammars
- The Process of Feature Detection
- Results
- Conclusions
Problem: Long-Distance Interactions

English → Urdu

“A student named Irshad who was throwing flour in the water for the fish . . . ”
Example Application: Feature-Rich Grammars

«ek talb a.SG student» ← 12 words → «raha tha ... PROG.SG.M be.PAST.SG.M»
Example Application: Feature-Rich Grammars

NP_A

ek   talb   a.SG   student

12 words

VP

raha   tha . . .
PROG.SG.M   be.PAST.SG.M
Example Application: Feature-Rich Grammars

(np sg)

(np sg)

12 words

→

ektalba.a.SG student

→

rakah.a.PRG.SG.M be.PAST.SG.M
Example Application: Feature-Rich Grammars

(NP<sub>A</sub> num) = (VP num)

1. (num sg)
2. (num sg)
3. (NP<sub>A</sub> num)
4. (VP num)

12 words

← 12 words →

ek talb
a.SG student

raha tha...

PROG.SG.M be.PAST.SG.M
Example Application: Feature-Rich Grammars

\[(\text{NP}_A \text{ num}) = (\text{VP num})\]

\[\text{(num sg)}\]

\[\text{S}\]

\[\text{(num pl)}\]

\[\text{NP}_A\]

\[\text{VP}\]

\[\text{ek talb a.SG student} \leftrightarrow 12 \text{ words} \rightarrow \text{raka txa... PROG.SG.M be.PAST.SG.M}\]
Overview of Feature Detection

Example Application: Feature-Rich Grammars

- The Process of Feature Detection
- Results
- Conclusions
Inductive Feature Detection

1. Elicitation Corpus
2. Minimal Pair Clustering
3. Feature Value Clustering
4. Morpheme-Feature Pairing
5. Morpheme-Feature Pairs
Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies.
Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies.
Elicitation Corpus Entry

context: Maria bakes cookies regularly or habitually.
srcsent: Maria bakes cookies .
tgtsent: Maria hornea galletas .
aligned: ((1,1),(2,2),(3,3),(4,4))
Elicitation Corpus Entry

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))
Elicitation Corpus Entry

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))

Distributed in this year’s NIST Urdu MT Evaluation via the LDC’s LCTL Language Pack
Inductive Feature Detection

- Elicitation Corpus
- Minimal Pair Clustering
- Feature Value Clustering
- Morpheme-Feature Pairing
- Morpheme-Feature Pairs
Minimal Pair Clustering

Wildcard Feature Structure:

((actor (num *) (gender n))
 (undergoer (num sg) (gender n))
 (tense pres))

The dog sees the cat
El perro ve el gato
((actor (num sg) (gender n))
 (undergoer (num sg) (gender n))
 (tense pres))

The dogs see the cat
Los perros ven el gato
((actor (num pl) (gender n))
 (undergoer (num sg) (gender n))
 (tense pres))
Minimal Pair Clustering

Wildcard Feature Structure:

\[((actor \text{ num } *)\text{ gender n)})
(undergoer \text{ num } sg\text{ gender n)})
(tense \text{ pres}))\]

The dog sees the cat
El perro ve el gato
(((actor \text{ num } sg)\text{ gender n)})
(undergoer \text{ num } sg\text{ gender n)})
(tense \text{ pres}))

The dogs see the cat
Los perros ven el gato
(((actor \text{ num } pl)\text{ gender n)})
(undergoer \text{ num } sg\text{ gender n)})
(tense \text{ pres}))
Inductive Feature Detection

- Elicitation Corpus
- Minimal Pair Clustering
- Feature Value Clustering
- Morpheme-Feature Pairing
- Morpheme-Feature Pairs
Feature Value Clustering

**s1:** The dog sees the cat
El perro ve el gato
((actor (num sg)…))
Feature Value Clustering

**s1:** The dog sees the cat
El perro ve el gato
((actor (num sg)…))

{ (s1, s2, NEQ) }

**s2:**
The dogs see the cat
Los perros ven el gato
((actor (num dl)…))
Feature Value Clustering

\(s_1:\) The dog sees the cat
El perro ve el gato
((actor (num sg)…))

\{ (s_1, s_2, \text{NEQ}) \} \quad \{ (s_1, s_3, \text{NEQ}) \}

\(s_2:\)
The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

\{ (s_1, s_3, \text{EQ}) \}

\(s_3:\)
The dogs see the cat
Los perros ven el gato
((actor (num pl)…)))
Feature Value Clustering

**s1:** The dog sees the cat
El perro ve el gato
((actor (num sg)…))

\{ (s1, s2, NEQ) \} \{ (s1, s3, NEQ) \}

**s2:**
The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

**s3:**
The dogs see the cat
Los perros ven el gato
((actor (num pl)…))
**Feature Value Clustering**

**s1:** The dog sees the cat
El perro ve el gato
((actor (num sg)…))

**s2:** The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

**s3:** The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

}\{ (s1, s2, NEQ),
  (s1, s3, NEQ) \}

\{ (s1, s3, EQ) \}

\(sg\) \(\rightarrow\) \(dl-pl\)
Inductive Feature Detection

1. Elicitation Corpus
2. Minimal Pair Clustering
3. Feature Value Clustering
4. Morpheme-Feature Pairing
5. Morpheme-Feature Pairs
Morpheme-Feature Pairing

**s1:** The dog sees the cat
El perro ve el gato
((actor (num sg)...))

**s2:** The dogs see the cat
Los perros ven el gato
((actor (num dl)...))

**s3:** The dogs see the cat
Los perros ven el gato
((actor (num dl)...))

\{ (s1, s2, \textbf{NEQ}), (s1, s3, \textbf{NEQ}) \}

\{ (s1, s3, \textbf{EQ}) \}
**Morpheme-Feature Pairing**

**s1:** The dog sees the cat
El perro ve el gato
((actor (num sg)…))

**s2:** The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

**s3:** The dogs see the cat
Los perros ven el gato
((actor (num dl)…))

\{ (s1, s2, NEQ),
  (s1, s3, NEQ) \}

\{ el, perro, ve \}

\{ los, perros, ven \}

\{ (s1, s3, EQ) \}
Outline

✓ Overview of Feature Detection
✓ Example Application: Feature-Rich Grammars
✓ The Process of Feature Detection
  • Results
  • Conclusions
Evaluation

Elicitation Corpus

Minimal Pair Clustering

Feature Value Clustering

Morpheme-Feature Pairing

Morpheme-Feature Pairs
Experiment

- Analyzed LCTL Urdu-English Elicitation Corpus (~3000 sentences)
- Evaluated by Urdu native speaker knowledgeable in linguistics

| Judgement       | Morpheme-Feature Pairings | Example Output       |
|-----------------|---------------------------|----------------------|
| Correct         | 68                        | “hai” → (tense pres) |
| Ambiguous       | 29                        | “raha” → (tense pres)|
| Incorrect       | 109                       | “nahin” → (tense pres)|
| TOTAL           | 206                       |                      |
Experiment

- Found **68 Correct Morpheme-Feature Pairs**
- = **53 Word Types**
- In an Urdu corpus of **17M tokens** and **200k types**, the 53 types selected by feature detection cover:
  - **0.02% of Types (Vocabulary)**
  - **38.6% of Tokens**

Tokens in Blind Test Set

- Annotated: 39%
- Remaining: 61%
Outline

✓ Overview of Feature Detection
✓ Example Application: Feature-Rich Grammars
✓ The Process of Feature Detection
✓ Results
• Conclusions
Future Work

Morphological Segmentation

- Morpheme-Feature Pairing
  - Feature Value Clustering
    - Minimal Pair Clustering
      - Elicitation Corpus
Elicitation Corpus

Minimal Pair Clustering

Feature Value Clustering

Morpheme-Feature Pairing

Morpheme-Feature Pairs

Presented Method

| Feature       | Candidate Lexical Items       |
|---------------|------------------------------|
| (num sg)      | come, el, perro              |
| (num dl, pl)  | comen, los, perros           |

Future Work

| Feature       | Candidate Morphemes          |
|---------------|------------------------------|
| (num sg)      | VB+Ø, el, N+o                |
| (num dl, pl)  | VB+n, los, N+os              |
Elicitation Corpus

Minimal Pair Clustering

Feature Value Clustering

Morpheme-Feature Pairing

Morpheme-Feature Pairs

Future Work

Morphological Segmentation
Elicitation Corpus

Minimal Pair Clustering

Feature Value Clustering

Morpheme-Feature Pairing

Morpheme-Feature Pairs

Future Work

Morphological Segmentation

Filtering
Minimal Pair Clustering

Feature Value Clustering

Morpheme-Feature Pairing

Elicitation Corpus

Future Work

Morphological Segmentation

Filtering

Feature Constraint Learning

Morpheme-Feature Pairs
Other Applications

- Factored MT
- **Data Selection** via active learning for synchronous grammar induction
- Aid for linguistics **field work**
Conclusion

• We now have
  • Feature-annotations for lexical items that convey grammatical meanings
  • Significant coverage
  • Structural MT systems stand to benefit by incorporating this morphosyntactic information
Inductive Detection of Language Features via Clustering Minimal Pairs: Toward Feature-Rich Grammars in Machine Translation

Questions?

Jonathan Clark
Robert Frederking
Lori Levin

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA