

Improving Identifier Informativeness Using Part of Speech Information

Dave Binkley
Matthew Hearn
Dawn Lawrie

Natural Language in Source Code

- Wealth of information 😊
- Language and Vocabulary mismatch 😞

Language Based tools need HELP

- Exploiting natural language in source code
- **Part-Of-Speech (POS)** information.

Example POS uses in SE

- Shepherd's query expansion tool
- Based on the AOIG (action-oriented identifier graph), which includes
 - a **verb** node for each distinct verb in the program,
 - a **direct object** (DO) node for each unique direct object in the program

Using natural language program analysis to locate and understand action-oriented concerns. [Shepherd, et al.]

Example POS uses in SE

An Exploratory Study of Identifier Renamings MSR 2011

Laleh M. Eshkevari

Venera Arnaoudova

Massimiliano Di Penta

Rocco Oliveto

Yann-Gaël Guéhéneuc

Giuliano Antoniol

Table 6: Classification of term grammar changes.

Renaming	Eclipse-JDT	Tomcat	Example ⁶
noun to verb	4	0	<i>editor</i> → <i>edit</i> (E)
noun to adjective	7	0	<i>qualificationPattern</i> → <i>qualifiedPattern</i> (E)
verb to noun	4	2	<i>preparedAuthenticate</i> → <i>preparedCredentials</i> (T)
verb to adjective	5	0	<i>fReconcileListeners</i> → <i>fReconcilingListeners</i> (E)
adjective to noun	5	0	<i>fLayoutHierarchicalAction</i> → <i>fShowTestHierarchyAction</i> (E)
adjective to verb	2	0	<i>isValidClassFile</i> → <i>validateClassFile</i> (E)
adverb to adjective	0	0	
Other changes	347	27	<i>filterStatic</i> (n;a) → <i>filterStatics</i> (n)
No change	230	27	

POS Taggers

- Most taggers expect (English) sentences.
- For example the Stanford Log-linear POS Tagger

Tagging Structure Fields

```
struct audio_file  
{  
    int length  
    int track_count  
    container create_mp4  
    void *data  
}
```

names split at
word markers

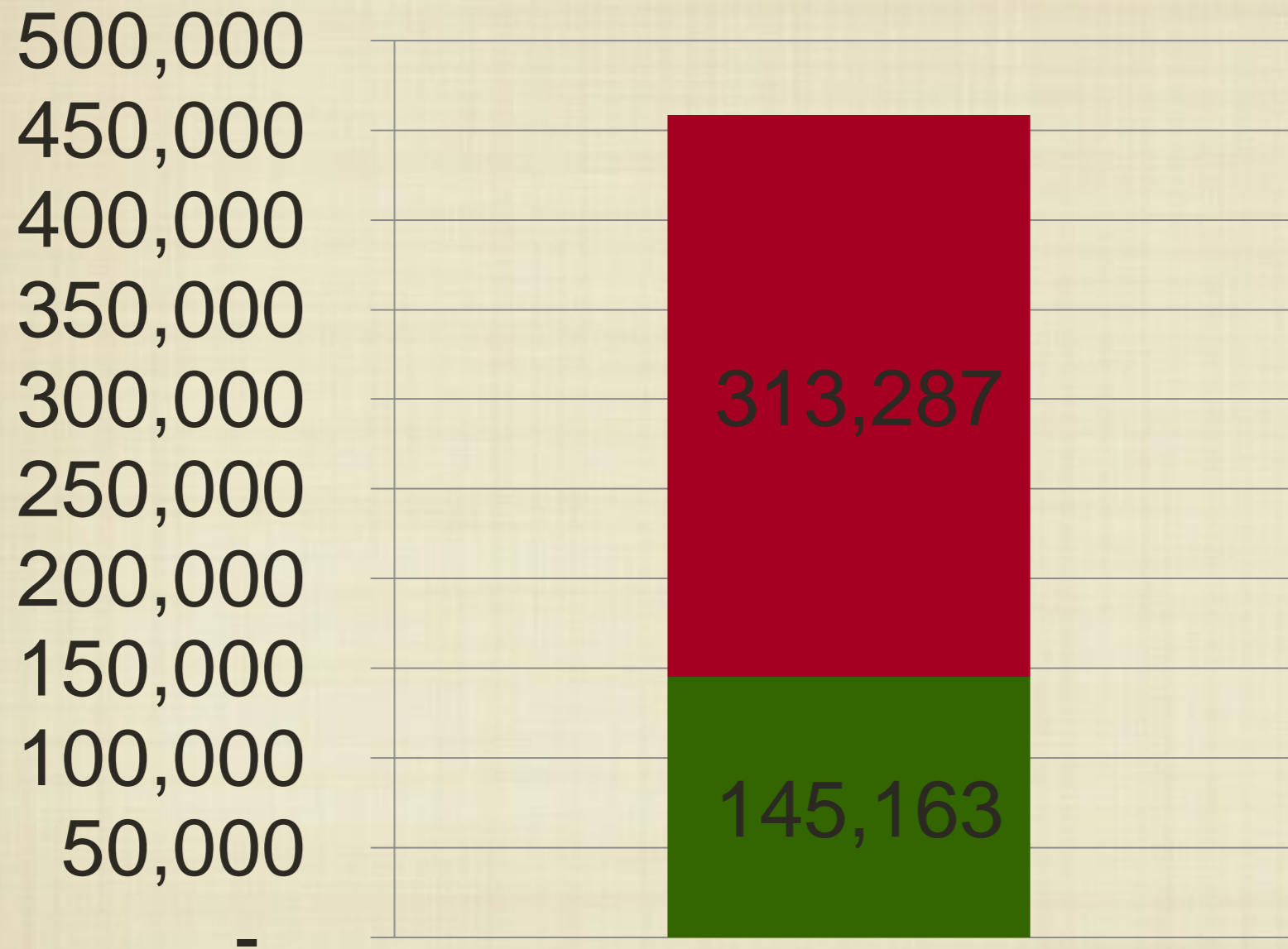
Tagger Needs Guidance

- We considered four *Templates*
 - Sentence: <split field name> “.”
 - List Item: “-” <split field name>
 - Verb: “Please,” <split field name> “.”
 - Noun: <split field name> “is a thing .”

Empirical Results

Template	Sentence	List Item	Verb	Noun
Sentence	79%	77%	72%	77%
List Item	77%	82%	71%	76%
Verb	72%	71%	76%	71%
Noun	77%	76%	71%	79%

Empirical Truth



■ Include Non-words ■ Words Only

Vocabulary Normalization

- Consider the identifiers
 - `featurelocation`
 - `floc`

Vocabulary Normalization

- Consider the identifiers

- `feature location`

- `floc`

Splitting Problem

Vocabulary Normalization

- Consider the identifiers

- `feature location`

- `f loc`

Splitting Problem

Splitting Problem

Vocabulary Normalization

- Consider the identifiers

- `feature location`

- `feature location`

Splitting Problem

Splitting and
Expansion Problem

An Application

- Rules for structure field naming
- R1: Avoid present tense verbs **
- Example
 - `create_mp4`
 - `created_mp4_container_type`

** Booleans are an exception

Summary

- POS tagging for field names
- Harder than other syntactic entities
(e.g., function and class names)

Questions?

