A DATASET FOR EVALUATING IDENTIFIER SPLITTERS

Presented by: Dawn Lawrie, Loyola University Maryland
In Collaboration with: Dave Binkley (Loyola), Lori Pollock (U. Delaware), Emily Hill (Montclair State), and K. Vijay-Shanker (U. Delaware)
THE DATASET

For testing splitting algorithms

- 2,733 identifiers
- from C, C++, and Java
- 8,522 human splitting judgements
CREATING THE DATASET

- Randomly select identifiers
- Gathering splitting judgements
- Curating the data
Given an identifier conservatively split

Split into concepts by inserting spaces

Rated confidence of split (Scale: 0-2)
Given an identifier conservatively split

BCCell Data

Split into concepts by inserting spaces

Rated confidence of split (Scale: 0-2)
Given an identifier conservatively split

Split into concepts by inserting spaces

Rated confidence of split (Scale: 0-2)
Given an identifier conservatively split

BCCell Data

Split into concepts by inserting spaces

BC Cell Data

Rated confidence of split (Scale: 0-2)
Given an identifier conservatively split

BCCell Data

Split into concepts by inserting spaces

BC Cell Data

Rated confidence of split (Scale: 0-2)

1
Given an identifier conservatively split

Split into concepts by inserting spaces

Rated confidence of split (Scale: 0-2)

1

Also collected annotator experience
Desire a single split to facilitate comparisons among splitting algorithms

What to do with identifiers split multiple ways?

Confidence weighted majority score

- Each annotator's vote weighted by confidence

Gold set contains 2,663 identifiers
LESSONS LEARNED

- Pros and cons of random sampling
- Facilitating annotations
- Annotator Confidence
- Establishing canonical splits
SUMMARY

- 2,663 Identifier Gold Set
  - Aimed at comparing the splitting algorithms
- 2,733 Identifier Raw Set with 8,522 Judgements
  - Enables investigation of other subsets