Test suites, [incr tsdb()]

Ling 567
January 11, 2018

Overview

- Questions from Lab 1

- Evaluation and computational linguistics

- Evaluation and precision grammars

- Test suites and precision grammars

- Our test suites

- Features of [incr tsdb()]

- Look at Lab 2 instructions

Ask more questions!

- This class is not designed so that you can complete the work on your own with the information provided.

- I'm *relying* on you to ask questions, and not spend lots of time stuck. The 10 minute rule is for real!

  - For Canvas, I try to answer very quickly --- but that means I can sometimes miss things if you put two questions in the same thread...

Overview

- Questions from Lab 1

- Evaluation and computational linguistics

- Evaluation and precision grammars

- Test suites and precision grammars

- Our test suites

- Features of [incr tsdb()]

- Look at Lab 2 instructions

Evaluation and Computational Linguistics

- Why is evaluation so prominent in computational linguistics?

- Why is it not so prominent in other subfields of linguistics?

- What about CS?

Intrinsic v. extrinsic evaluation

- Intrinsic: How well does this system perform its own task, including generalizing to new data?

- Extrinsic: To what extent does this system contribute to the solution of some problem?

- Examples of intrinsic and extrinsic evaluation of parsers?

Test data

- Test suites

  - Hand constructed examples

  - Positive and negative examples

  - Controlled vocabulary

  - Controlled ambiguity
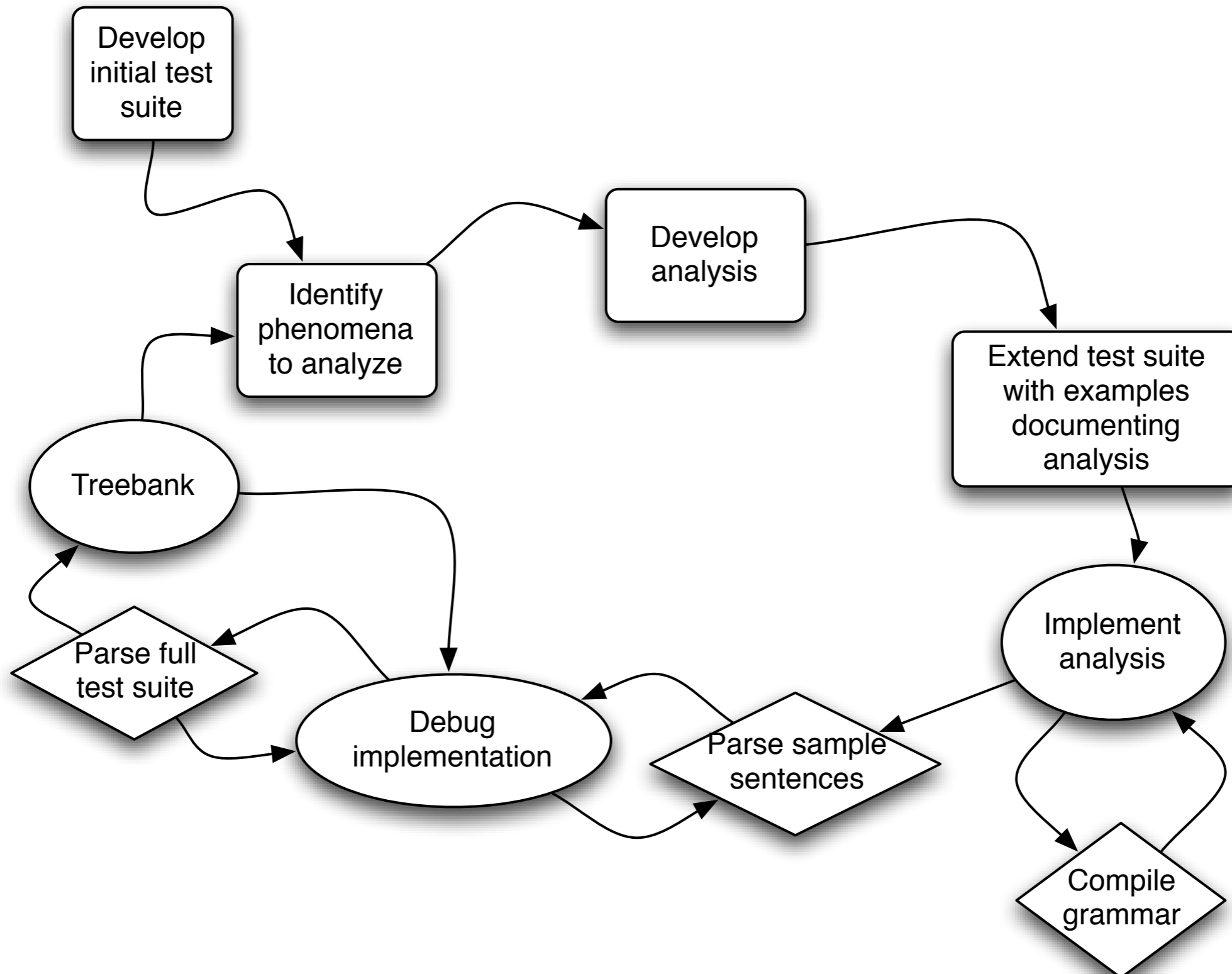
  - Careful grammatical coverage

Test data

- Test corpora

  - Naturally occurring

  - More open vocabulary

  - Haphazard ungrammatical examples

  - Application-focused

Uses of test data

- How far do I have left to go?

  - Internal metric

  - Objective comparison of different systems

- Where have I been?

  - Regression testing

  - Documentation

Grammar engineering workflow

Evaluating precision grammars

- Coverage over some corpus

  - Which corpus?

  - Challenges of lexical acquisition

- Coverage of phenomena

  - How does one choose phenomena?

- Comparison across languages

Levels of adequacy

- grammaticality

- "right" structure

- "right" dependencies

- "right" full semantics

- only legit parses (how can you tell?)

- some set of parses including the preferred one

- preferred parse only/within first N

Our test suites

- Map out territory we hope to cover

- Include both positive and negative examples

- Serve as an exercise in understanding the description of the language

  - IGT format

  - Creating examples where necessary

On the importance of simple examples

- Why keep examples simple?

- How simple is too simple?

- What kinds of things make an example not simple enough?

On the importance of simple examples

- Awtuw [awt] (Feldman 1986:67)

  (70) Yowmen Yawur du-k-puy-ey
        Yomen   Yawur DUR-IMPF-hit-IMPF
        'Yowmen and Yawur are hitting (someone).' [awt]

- Basque [eus] (adapted from Joppen and Wunderlich 1995:129)

  (112) Zuek   lagun-ei   opari polit-ak   ema-ten dizkiezue.
        you.PL.ERG friend-PL.DAT present nice-PL.ABS give-IMPF 3A.have.PLA.3PLD.2PLE
        'You(pl) always give nice presents to your friends.' [eus]

On the importance of simple examples

- Russian [rus] (Bender 2013:92)

a. Человек   укусил   собаку.
   Chelovek   ukusi-l   sobak-u.
   man.NOM.SG.M bite-PAST.PFV.SG.M dog-ACC.SG.F
   'The man bit the dog.' [rus]

[incr tsdb()] basics

- [incr tsdb()] stores test suite profiles as (plain text) relational databases: Each is a directory with a fixed set of files in it.

- Most files are empty.

- A profile that has not been processed has only two non-empty files: item (the items to be processed) and relations (always the same)

- Once the profile has been processed, the result of the processing is stored in some of the other files (in particular, parse and result)

[incr tsdb()] basics

- A test suite *skeleton* consists of just the item and relations files and can be used to create new test suite profiles

- [incr tsdb()] allows the user to compare two profiles to see how they differ

- It can also produce graphs plotting summary data from many profiles to visualize grammar evolution over time

- -> Demo

Overview

- Questions from Lab 1

- Evaluation and computational linguistics

- Evaluation and precision grammars

- Test suites and precision grammars

- Our test suites

- Features of [incr tsdb()]

- Look at Lab 2 instructions