Developing reliability metrics and validation tools for datasets with deep linguistic information

Sérgio Castro
Supervisor: Prof. Dr. António Branco
(presentation by João Silva)

Faculty of Sciences, University of Lisbon
NLX — Natural Language and Speech Group

DELPH-IN Summit
Suquamish, WA
June 2011

Presentation outline

Presentation outline

Introduction

Motivation

- Corpora with increasingly complex linguistic information
- Automated annotation with manual correction
- Ensuring reliability of annotated corpora
  - Make use of multiple annotators
  - Quantified through inter-annotator agreement (ITA) metrics

- Compare the output of the annotators:
  e.g. exact match, Parseval
  (too coarse for our purposes)
- Need to account for chance agreement

Agreement with chance correction

Observed, discounting expected

- Observed agreement (A_o)
- Expected agreement (A_e)

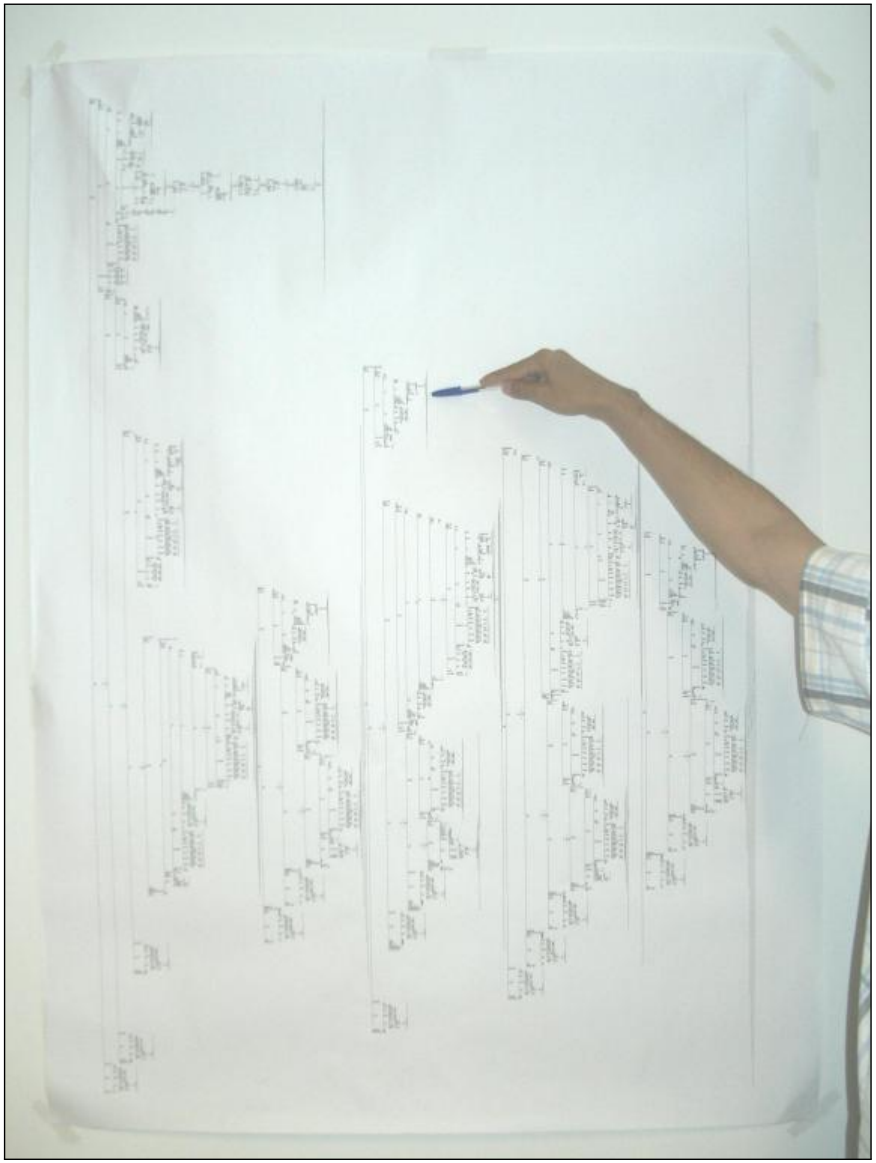
$$\frac{A_o - A_e}{1 - A_e}$$

Introduction

LX-DeepGramBank

Overview

- Double-blind annotation with adjudication
- Analyses by LX-Gram
- Manual disambiguation via semantic discriminants (with LinGO)



Representation of "Todos os computadores têm um disco"

Introduction

Objectives

- Develop a granular ITA metric
  - Accept/reject and Parseval are too coarse
  - Look at individual disambiguation options
  - Account for chance agreement
- Implement a tool
  - Analyse LinGO logs
  - Produce reports

Presentation outline

Agreement metric

The Y-Option Kappa metric (K_Y^s)

Observed The proportion of discriminants on which the annotators agree from the total set of discriminants

$$A_o^s = \frac{agr_s}{|D_s|}$$

Expected Assume random (uniform) choice

Agreement Observed, discounting expected

$$K_Y^s = \frac{A_o^s - A_e^s}{1 - A_e^s}$$

Difficulties 1/2

- Some discriminants are implicitly marked
  - Marking a discriminant discards at least one parse
    Discriminants belonging to that parse are automatically marked
  - Accepting a parse marks its discriminants
- Several markings may result from a single manual choice
  - The tool must group these discriminants together
    (a choice and its consequences are still one choice)

Difficulties 2/2

- LinGO logs do not store every discriminant
  - No information is stored when there is only one parse
  - Upon rejection, only already marked options are stored (if nothing was marked, nothing is stored)
  - Mismatch between the set of discriminants of each annotator
- LinGO bugs/crashes

Adapting the metric

Overview

In theory, theory and practice are the same.
In practice they are not.

The set of sentences is divided into three subsets:

1. Sentences accepted by both annotators
2. Sentences rejected by at least one annotator
3. Sentences without stored information

Adapting the metric

❶ Sentences accepted by both annotators, S_{both}

The "well-behaved" case

- All discriminants are available, proceed as expected
- Calculate the proportion of divergence, P_D
  (to be used in the other cases)

Example: Proportion of divergence, P_D

$\lvert O_s \rvert$	$\lvert O_s^{eq} \rvert$	$\lvert O_s^{dif} \rvert$	P_D^s
12	8	4	0.33
7	6	1	0.14
9	0	9	1.00
10	7	3	0.30
17	12	5	0.29

$$P_D^{S_{both}} = \frac{4+1+9+3+5}{12+7+9+10+17} = 0.40$$

Adapting the metric

❷ Sentences rejected by at least one annotator, S_{R1}

The "incomplete information" case

- Divide options into two subsets:
  - O_{com} are options common to both annotators
  - O_{uniq} are options only present for one annotator

The O_{com} set is well-behaved. For O_{uniq} use an estimation:

$$\frac{|O_{uniq}| \cdot \left(1 - P_D^{S_{both}}\right)}{|O_{uniq}|}$$

Adapting the metric

❸ Sentences without stored information, S_{noop}

The "no information at all" case

- Estimate number of options per sentence

$$O_{avg} = \frac{|O_{S_{both}}| + |O_{S_{R1}}|}{|S_{both}| + |S_{R1}|}$$

- Estimate "observed" agreement

$$\frac{O_{avg} \cdot \left(1 - P_D^{S_{both}}\right)}{O_{avg}}$$

Experiment

Some (approximate) numbers:

- 50,000 sentences
- 15,000 are parsed (30% coverage)
- 12,300 remain (due to LinGO bugs/crashes)

Agreement

- Y-Option Kappa of 0.91 (over the 12,300 sentences)
  (traditionally, the acceptable threshold is at 0.80)

Presentation outline

Closing remarks

- Y-Option Kappa metric
  - Observed, discounting expected
  - Granular ITA metric
    (each semantic discriminant is a choice)
- Tool for LinGO log analysis
  - Handles incomplete information in log files
    (with estimates from "well-behaved" cases)
  - Handles automatic markings
    (by grouping a choice and its consequences)
- DeepGramBank is reliable

Thank you.