The Value Versus Volume Continuum

Training data for speech recognition

Tom Dibert
Appen

Quick Overview

- No single company approach.
- What are we talking about?
- Volume versus value.
- Government R&D trends.
- Consortia projects.

No Single Approaches

- We don’t discuss client specific programs unless it’s publically released data.
- None of the approaches we’re discussing today are “single company” processes.

What Are We Talking About?

- Our perspective
  - Means to an end.
  - Very accurate/small volume?
  - Huge volumes?

Volume Versus Value

- Many clients have enormous volumes of data – not much use without meta data.
- Few clients have small amounts, heavily annotated.
- Truth is somewhere in between.

Mainstream Commercial

- Where are most commercial clients?
- Hypothesis confirmation/correction.
- Full DC, transcription (typically domain specific).
- “Crowd style” remote data collection.

Government R&D

- Programs like BABEL
  - R&D for ASR
  - What can I learn from language A to help me with language B, and A+B for a new language?
    - Ultimate goal is reduction in time/data volumes to for rapid response in low-density/low resource languages.

Consortia Projects

- Appen and others have participated in numerous data collection consortia.
- Pro
  - Cost effective; many databases for price of one.
- Con
  - Consortium must agree on style/specs.

QUESTIONS?