Trends in HLT Research: A Survey of LDC's Data Scholarship Program

Denise DiPersio, Christopher Cieri

Linguistic Data Consortium. University of Pennsylvania

3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA

{dipersio, ccieri } AT ldc.upenn.edu

Linguistic Data Consortium

Why this Paper?

- LDC 2 year report, ongoing communication
  - consortial model
  - progress
- help applicants succeed within program
- help potential applicants succeed elsewhere
  - significant number of applicants mention
    - difficulty being published
    - time spent building and annotating their own corpus
- preview of trends in the field
- perspectives on gaps in training researchers

Data Scholarship Program

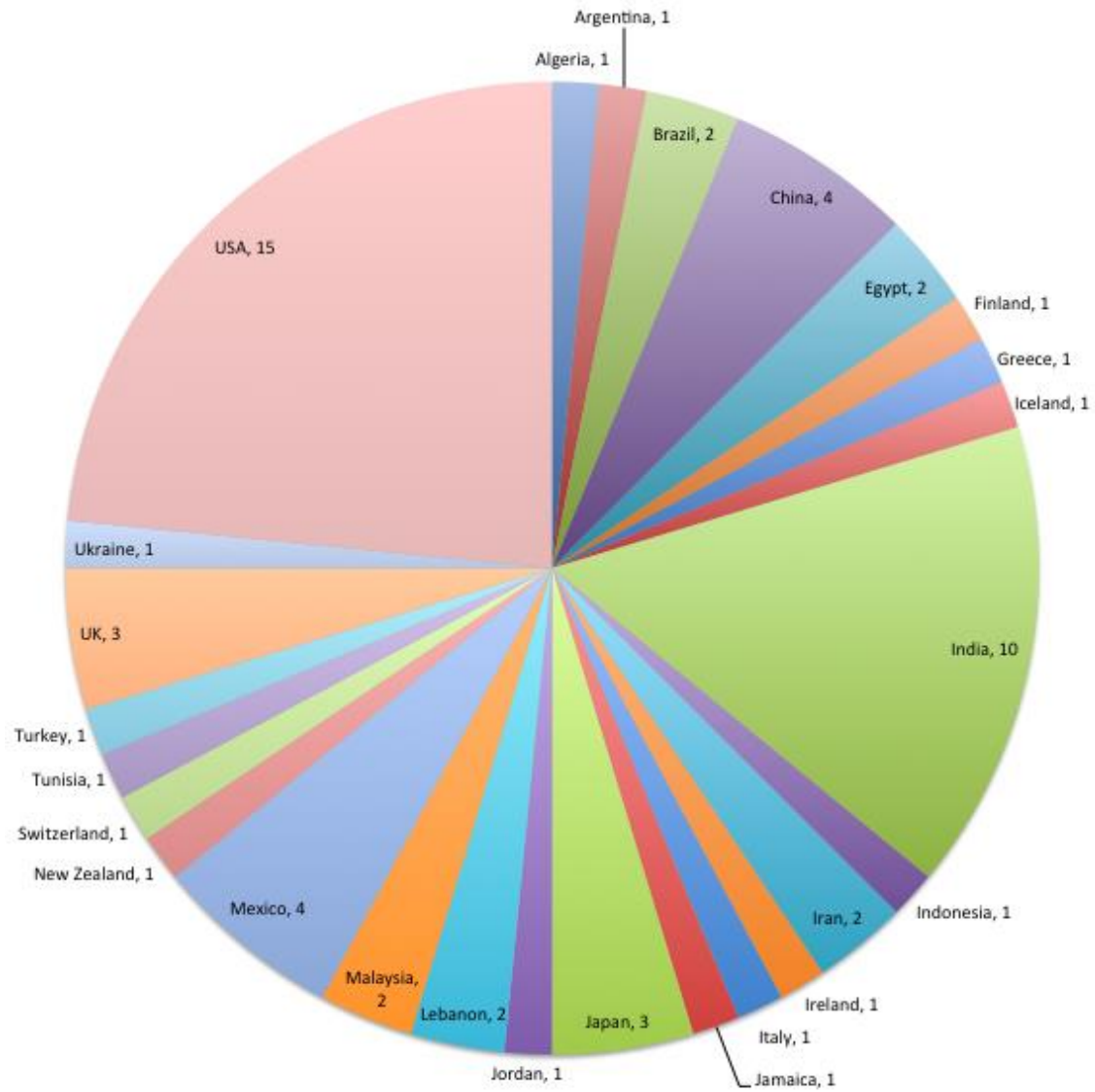
- LDC Principle: no one with a bona fide research agenda will go without data for a genuine inability to contribute
- regular student data requests to support dissertation work at institutions lacking financial resources to support Consortium
- program formalized to guarantee equal opportunity for assistance
- application
  - data use statement: research plan, data use, evaluation method
  - advisor letter: asserts high probability of success, inability to contribute
- advertised on LDC's web pages, social media platforms, monthly newsletter, conferences, LDC networks
- Benchmark:
  - since 2010 64 recipients from 26 countries
  - 110 corpora, license value >$175,000, 64% acceptance rate

Linguistic Data Consortium

Success Factors

- understanding of requested database
  - database has necessary features and annotations
  - or proposal explains how they will be added
- appropriate evaluation methodology
  - in speech recognition: existing evaluation protocol & scorer
- appropriate research methodology
  - adopt accepted methodology | motivate alternative methodology > adopt new methodology without justification
- appropriate planning
  - plans to process very large corpus in very short time should mention computer resources & their deployment

Awards by Country



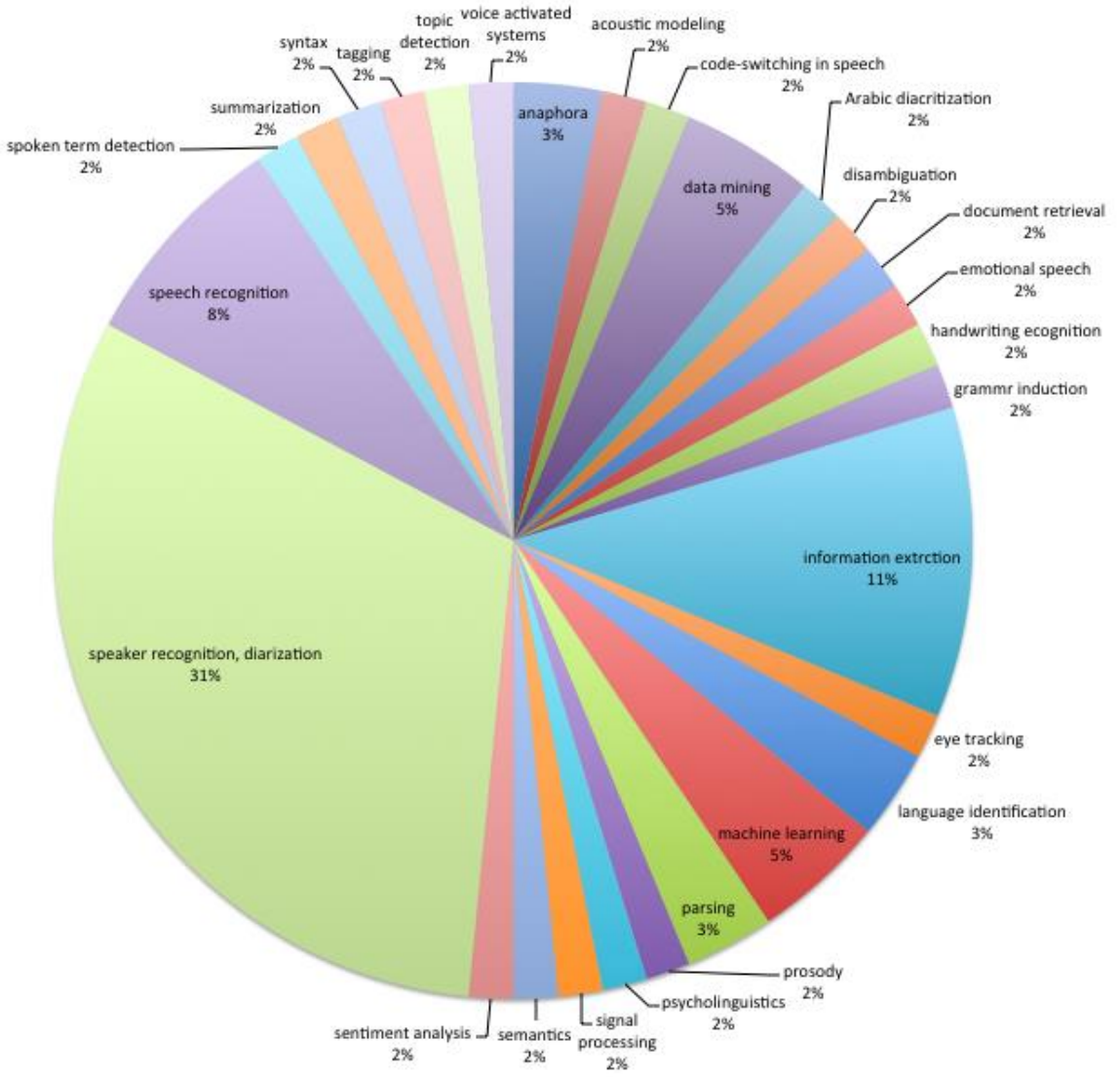
- each counts a corpus award to a person or group
- data licensed to institutions, remains after student graduates

Linguistic Data Consortium

Awards by Country

- volumes probably reflect
  - penetration of communications
  - areas of need
- underfunded research groups worldwide
  - even in computer science and engineering
- acute need for language resources in some regions
- may also reflect the availability of resources
  - Arabic, Chinese
- American applicants from diverse research groups suggests
  - spread of HLT, big data to other disciplines

Awards by Research Area



- other categorizations conceivable
- NB: applicants are not LDC members
- NB: availability of resources probably also affects applications

Awards by Corpora Requested

- Most Requested
  - NIST Speaker Recognition Evaluation, YOHO Speaker Verification
  - ACE (Automatic Content Extraction)
  - other benchmark data e.g. HUB4 Broadcast News & Transcripts
  - CALLHOME & Switchboard - transcribed telephone conversations
  - TIMIT series
  - TIDIGITS
  - Continuous Speech Recognition (CSR) – read, broadcast news
  - Gigawords – billions of words of new text
  - Topic Detection and Tracking (TDT)
  - Treebanks
- Unique
  - emotional speech -> Emotional Prosody
  - handwriting recognition -> MADCAT

Linguistic Data Consortium

Challenges

- tension between
  - desire to support young scholars
  - need to be good stewards of Consortium funds
- diversity of applicants' scientific disciplines
  - reviewers not expert in every field
  - different expectations across communities:
    - metrics-driven evaluation expected in some disciplines
    - metrics, gold standard data, scorer, **concept** absent in others
- international applicant pool
  - different approaches to completing applications
  - review committee experienced with international panels
  - however, prior knowledge or researcher/mentor often absent
- revising application process to
  - maximize success
  - maximize efficiency

Outcomes: based on awardee survey

- contributions to multiple language-related disciplines
- positive reactions from recipients
  - most described data as vital to their work
  - most reported using data, finding results as expected
- graduates: 3 already, 2 expected in 2016
- 6 published papers based on program data
- program data used in AMRITA-TCS system submitted to SRI Speakers in the Wild (SITW) Speaker Recognition Challenge
- Negative Reports
  - expected data to contain something it did not
    - failure of the vetting process
  - data set was too small
  - dissertation topic changed

Comparison to Other Programs

- unaware of programs very similar to LDC's data scholarships
  - student support
  - focused on data
  - recurring, without restriction as to corpus
- Nearest
  - ELRA offers some LRs at no cost, internships
  - GSK apparently has student pricing
  - LDC-IL occasional applications for short term projects which may attract student candidates
  - Of course, LDC also offers some data to non-members at no cost; all data to members at no cost beyond membership fee
  - CLARIN ERIC Mobility Grants data and mentoring opportunties
  - Many funding bodies support student travel, research which my include data costs.

Related and Future Work

- Future Work on Data Scholarship Program
  - seek funding to support and expand program
  - external reviewers

- Related Work Benefitting the Program
  - business system
  - delivery via direct download, cloud, grid
  - cost reduction
    - benefitting from Moore's Law for storage, computing and networking
    - not for human resources

Linguistic Data Consortium

Acknowledgements

Thanks are due to the **members** of the Linguistic Data Consortium who, through their membership fees, subsidize the Data Scholarship program as well as the many activities that sustain the Consortium and make the Data Scholarship program possible.