Language resources, intelligence, and community

Second International Workshop on Language Resources and Intelligence
Beijing Language and Culture University
12/16/2018

Mark Liberman

Linguistic Data Consortium
University of Pennsylvania

Beginning in the 1980s, research in Human Language Technology played a central role in the transformation of Artificial Intelligence from a field based on human-created logical rules into a field based on machine learning applied to large datasets.

It's less widely recognized that a critical part of this transformation was the creation of large and effective research communities, across organizational and national boundaries, based on well-defined "common tasks" that included only shared training data but also shared quantitative evaluations and frequent meetings to describe and discuss progress.

I will briefly survey this history, and open a discussion about the future.

Why language resources?

1. One obvious answer: Document languages

2. Another obvious answer: Provide input to machine learning.

3. A less obvious answer: Create effective research communities.

4. And goal #3 requires not just any resources, but rather:

   a) tasks designed to advance the technology
   b) metrics designed to support the tasks
   c) resources designed to support a) and b)

How do we know that?

“Classical” Artificial Intelligence (1950-1985++) was based on:

- physics
  (models of physical states, events and signals)
- facts
  (knowledge about things and relations)
- logic
  (inferences from physics, facts, and ?)

This sensible idea was the dominant approach to Human Language Technology –

- speech recognition and synthesis,
- text analysis and synthesis,
- machine translation,
- human-computer dialogue,
- …etc…

and also to computer vision, robotics, and so on.

Some quotations from the era of classical AI:

John McCarthy, “Situations, Actions, and Causal Laws”, Stanford Research Report **1963**:

Our approach to the artificial intelligence problem requires a formal theory. Namely, we believe that human intelligence depends essentially on the fact that we can represent in language facts about our situation, our goals, and the effects of the various actions we can perform. Moreover, we can draw conclusions from the facts to the effect that certain sequences of actions are likely to achieve our goals.

Bob Moore, “Automatic Deduction from Commonsense Reasoning”, SRI Research Report, **1981**:

How to enable computers to draw conclusions automatically from bodies of facts has long been recognized as a central problem in artificial-intelligence (AI) research. Any attempt to address this problem requires choosing an application (or type of application), a representation for bodies of facts, and methods for deriving conclusions. […]

We discuss the relationship of automatic deduction to the new field of “logic programming”, and we survey some of the issues that arise in extending automatic-deduction techniques to non-standard logics.

Fernando Pereira, "Logic for Natural Language Analysis", Edinburgh University PhD thesis, **1982**:

This work investigates the use of formal logic as a practical tool for describing the syntax and semantics of a subset of English, and building a computer program to answer data base queries expressed in that subset.

To achieve an intimate connection between logical descriptions and computer programs, all the descriptions given are in the definite clause subset of the predicate calculus, which is the basis of the programming language Prolog. The logical descriptions run directly as efficient Prolog programs.

Three aspects of the use of logic in natural language analysis are covered: formal representation of syntactic rules by means of a grammar formalism based on logic, extraposition grammars;. Formal semantics for the chosen English subset, appropriate for data base queries; informal semantic and pragmatic rules to translate analysed sentences into their formal semantics.

But things have changed.

Fast forward 34 years…

Amir Globerson, …, and **Fernando Pereira**, “Collective Entity Resolution with Multi-Focal Attention”, (Google) **2016**

Entity resolution is the task of linking each mention of an entity in text to the corresponding record in a knowledge base (KB). Coherence models for entity resolution encourage all referring expressions in a document to resolve to entities that are related in the KB. We explore attentionlike mechanisms for coherence, where the evidence for each candidate is based on a small set of strong relations, rather than relations to all other entities in the document. The rationale is that documentwide support may simply not exist for non-salient entities, or entities not densely connected in the KB. Our proposed system outperforms state-of-the-art systems on the CoNLL 2003, TAC KBP 2010, 2011 and 2012 tasks.

Three relevant changes:

1. Different conceptual and technological foundations
   - **1982**: logical inference and Prolog
   - **2016**: “neural nets” and TensorFlow

2. Reference to “common tasks” is new (CoNLL 2003, TAC KBP 2010, 2011 and 2012)

3. Many real-world AI successes in the background (though challenging problems remain)

Where did these changes come from?

“Classical AI” (physics, facts, and logic) made a lot of sense.

But unfortunately, it didn’t work.

After three decades or so, most government and industry leaders came to the conclusion that the whole enterprise was a waste of money.

Result: The “AI Winter”.

Some eminent observers expressed negative opinions very harshly.

John Pierce (manager of the teams that invented the transistor and designed the first communications satellite) wrote:

"The typical [AI engineer]... builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment."

"To sell suckers, one uses deceit and offers glamour."

"It is clear that glamour and any deceit … blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect."

So when U.S. support for Human Language Technology research began again in the mid-1980s, the programs were carefully crafted to

- protect against “glamour and deceit”
  because there was a well-defined, objective evaluation metric
  applied by a neutral agent (NIST)
  on shared data sets; and
- ensure that “simple, clear, sure knowledge is gained”
  because participants must reveal their methods
  to the sponsor and to one another
  at the time that the evaluation results are presented

The “Common Task” structure:

- A detailed task definition and “evaluation plan”
  developed in consultation with researchers
  and published as the first step in the project.
- Automatic evaluation software
  written and maintained by NIST
  and published at the start of the project.
- **Shared data:**
  Training and “dev(elopment) test” data
  is published at start of project;
  “eval(uation) test” data is withheld
  for periodic public evaluations

Not everyone liked it.

Many Piercians were skeptical:

“You can’t turn water into gasoline,
no matter what you measure.”

Many researchers were disgruntled:

“It’s like being in first grade again --
you’re told exactly what to do,
and then you’re tested over and over .”

But it worked.

Why did it work?

1. The obvious: it allowed funding to start
   (because the projects were glamour-and-deceit-proof)

   and to continue
   (because funders could measure progress over time)

Why did it work?

2. Less obvious: it allowed project-internal hill climbing
   - because the evaluation metrics were automatic
   - and the evaluation code was public

This obvious way of working was a new idea to many!

… and researchers who had objected to be tested twice a year

began testing themselves every hour…

Why did it work?

3. Even less obvious: it created a culture
(because researchers shared methods and results
on shared data with a common metric)

**Participation in this culture became so valuable
that many research groups joined without funding**

What else it did

The *common task method* created a positive feedback loop.

When everyone’s program has to interpret the same ambiguous evidence,
ambiguity resolution becomes a sort of gambling game,
which rewards the use of statistical methods,
and has led to the flowering of “machine learning”.

Given the nature of speech and language,
statistical methods need the largest possible training set,
which reinforces the value of shared data.

Iterated train-and-test cycles on this gambling game are addictive;
they create “simple, clear, sure knowledge”,
which motivates participation in the common-task culture.

The "Common Task Method"

. . . has become the standard research paradigm
in experimental computational science:

- Published training and testing data
- Well-defined evaluation metrics
- Techniques to avoid over-fitting
  (managerial as well as statistical)

Domain: ***Algorithmic analysis of the natural world.***

Over the past 30 years, variants of this method have been applied to many other problems:

machine translation, speaker identification, language identification, parsing, sense disambiguation, information retrieval, information extraction, summarization, question answering, OCR, sentiment analysis, image analysis, video analysis, … , etc.

The general experience:

1. Error rates decline by a fixed percentage each year,
   to an asymptote depending on task and data quality
2. Progress usually comes from many small improvements;
   improvement by 1% is a reason to break out the champagne.
3. Shared data plays a crucial role – and is re-used in unexpected ways.
4. Glamour and deceit have mostly been avoided.

Thus a central goal of language resource creation is:

to create effective research communities,
focused on a series of hard problems,
whose solutions bring real benefits to society.

A (partial) list of important problems that are still unsolved:

- The “Cocktail party” problem(s):
  sound enhancement, localization and separation
- Lifelike co-constructed human-machine dialog
- Diarization of real-world recordings
- Spatial language understanding in the real world
- Reference resolution (e.g. Winograd Schema problems)
- Intelligent tutoring
- Clinical diagnosis and monitoring
- Efficient adaptation of HLT to new varieties and new languages

For each problem, we should ask:

1. What is the right series of tasks to lead to eventual success?

2. What are the right evaluation metrics for each task?

3. What are the right datasets for each step of the process?

Thank you!

?