Reasons to avoid Reasoning: Where does NLP stop and AI Begin?

Bill Dolan
Microsoft Research

NSF symposium on Semantic Knowledge Discovery, Organization and Use
November 15, 2008

Modeling Semantic Overlap

- Over the last few years, broadening consensus that this is the core problem in building applications that “understand” language
  - Search, QA, Summarization, Dialog, etc.
  - Bakeoffs in QA, Textual Entailment, etc.

The latest DreamWorks animation fest, "Madagascar: Escape 2 Africa," surpassed expectations, bringing in $63.5 million in its opening weekend. That put it way ahead of any competition and landed the "Madagascar" sequel the third-biggest opening weekend ever for a DreamWorks picture, behind "Shrek 2" and "Shrek the Third."

It was a zoo at the multiplex this weekend, as the animated sequel Madagascar: Escape 2 Africa easily won the box office crown.
Another Inconvenient Truth

• So far, though, unsatisfying progress toward real applications
  – Keywords still rule web search and QA
  – No obvious progress toward single-document summarization
  – Hand-coded Eliza clones dominate the dialog world
• No unified, cutting edge research agenda
  – As in e.g. Speech Recognition or Machine Translation
  – Instead, plethora of algorithms, tools, resources being used
• Still a niche field
  – Semantic overlap may be key to the “Star Trek” vision, but MT papers dominate today’s NLP conferences
• Why?
  – Is it too early in the revolution to judge results?
  – Are we using the wrong machinery?
  – Or have we mischaracterized the problem space?

Problems with the Problem

• No clear-cut definition of target phenomena
• Hand-selection of data leads to artificial emphasis on “favorites”, e.g.
  – Glaring contradictions (e.g. negation mismatch)
  – Well-studied linguistic alternations (e.g. scope ambiguities, long-distance dependencies)
• Artificial division of data into e.g. 50% True/False
  • No guarantee of match with real-world frequency
  • May greatly overstate actual utility of algorithms
• Less than ideal inter-annotator agreement (Snow et al. 2008)
What I learned from MindNet (circa 1999) and Why It’s Relevant Today

• MindNet: an automatically-constructed knowledge base (Dolan et al, 1993; Richardson et al 1998)
  – Project goal: rich, structured knowledge from free text
  – Detailed dependency analysis for each sentence, aggregated into arbitrarily large graph
  – Named Entities, morphology, temporal expressions, etc.
  – Frequency-based weights on subgraphs
  – Path exploration algorithms, learned lexical similarity function
• Built from arbitrary corpora: Encarta, web chunks, dictionaries, etc.

http://research.microsoft.com/mnex/
Question Answering with MindNet

- Build a MindNet graph from:
  - Text of dictionaries
  - Target corpus, e.g. an encyclopedia (Microsoft Encarta)
- Build a dependency graph from query
- Model QA as a graph matching procedure
  - Heuristic fuzzy matching for synonyms, named entities, wh-words, etc.
  - Some common sense reasoning (e.g. dates, math)
- Generate answer string from matched subgraph
  - Including well-formed answers that didn’t occur in original corpus

Logical Form Matching (2)

Input LF:

Who assassinated Abraham Lincoln?
Fuzzy Match against MindNet

American actor John Wilkes Booth, who was a violent backer of the South during the Civil War, shot Abraham Lincoln at Ford's Theater in Washington, D.C., on April 14, 1865.

Generate output string

“John Wilkes Booth shot Abraham Lincoln”
Evaluation

• Tested against a corpus of:
  – 1.3K naturally-collected questions
  – Full recall set (10K+ Q/A pairs) for Encarta 98, created by professional research librarian
    • Fine-grained detail on quality of linguistic/conceptual match

(More on this dataset later...)

Worked beautifully!

• Just not very often...
• Most of the time, the approach failed to produce any answer at all, even when:
  – An exact answer was present in the target corpus
  – Dependency analysis for query/target strings was correct
• What went wrong?
  – Complex linguistic alternations: paraphrase, discourse
  – AI-flavored reasoning challenges
Genre-specific Matching Issues

Q: How hot is the sun?

• Graphical Content

• Tabular Content

Genre-Specific Matching Issues (2)

• Encyclopedia article title = Antecedent for explicit/implicit subject pronoun

Q: Who killed Caesar?
A: During the spring of 44 BC, however, he joined the Roman general Gaius Cassius Longinus in a conspiracy against Caesar. Together they were the principal assassins of Caesar. (Brutus, Marcus Junius)
Simple Linguistic Alternations

Q: In what present-day country did the Protestant Reformation begin?
A: The University of Wittenberg was the scene of the beginning of the Protestant Reformation (1517), started by Martin Luther, a professor there. (*Christianity: Reformation and Counter Reformation*)

Q: Do penguins have ears?
A: Birds have highly developed hearing. Although the structure of their ears is similar to that of reptiles, birds have the added capability of distinguishing the pitch of a sound and the direction from which it comes. (*Ear*)

More Complex Linguistic Alternations

Q: What are the dangers of Radon?
A: Selenium is especially harmful to wildlife in heavily irrigated areas, and indoor radon has become a major health concern because it increases the risk of lung cancer. (*Geochemistry*)

Q: Why is grass green?
A: Plants possess, in addition to mitochondria, similar organelles called chloroplasts. Each chloroplast contains the green pigment chlorophyll, which is used to convert light energy from the sun into ATP. (*Cell Biology*)

Q: How big is our galaxy in diameter?
A: The Milky Way has been determined to be a large spiral galaxy, with several spiral arms coiling around a central bulge about 10,000 light-years thick. The diameter of the disk is about 100,000 light-years. (*Milky Way*)
# Extra-Linguistic Reasoning

## Mathematical

**Q:** How hot is the sun?

**A:** The surface temperatures of red dwarfs range from 2800 °C to 3600 °C (5100 °F to 6500 °F), which is only about 50 to 60 percent of the surface temperature of the sun. ([Flare Star](#))

## Causality

**Q:** Why do some people have freckles and other people don't?

**A:** Freckles appear in genetically predisposed individuals following exposure to sunlight or any other ultraviolet light source. ([Freckles](#))

**Q:** When was the universe created?

**A:** Some original event, a cosmic explosion called the big bang, occurred about 10 billion to 20 billion years ago, and the universe has since been expanding and cooling. ([Big Bang Theory](#))

---

# Extra-Linguistic Reasoning (2)

## Deeper Reasoning

**Q:** Do cloned animals have the same DNA make up?

**A:** While Dolly has most of the genetic characteristics of sheep A, she is not a true clone. ([Clone](#))

**Q:** Are photons particles or waves?

**A:** Radiant energy has a dual nature and obeys laws that may be explained in terms of a stream of particles, or packets of energy, called photons, or in terms of a train of transverse waves (see Photon; Radiation; Wave Motion). ([Optics](#))
How much of this is “NLP”?

- Our test corpus was as “real-world” as we could make it
  - Yet rife with seemingly unapproachable problems
  - Often, problems are simply not linguistic
    - NLP machinery irrelevant to the task
    - Require Big AI, not computational linguistics
- These problems become obvious only in recall scenario
  - None of the web’s redundancy
  - Brittleness immediately apparent
- But not unique to the QA task
  - Echoed in other applications requiring “understanding”

Search Indexing, Multi-document Summarization

A child who lives near a petrol (gas) station is four times more likely to develop leukemia than a child who lives far away from one, according to a new study. Living near to a petrol station or garage may increase the risk of acute childhood leukaemia by 400%.

Children who live in close proximity to gas stations and auto body shops have a dramatically higher rate of leukemia, according to a new study. Living near a petrol station may quadruple the risk for children of developing leukaemia, new research says. Children who live near petrol stations may be four times more susceptible to leukaemia.
Linguistic Alternations

• Spelling
  leukemia / leukaemia

• Lexical:
  petrol (gas) station
  a petrol station or garage
  gas stations and auto body shops

• Morphological
  a child / children
  a petrol station / petrol stations

• Phrasal / syntactic
  may quadruple the risk for
  may be four times more susceptible to
  is four times more likely to develop

• Ellipsis
  (…than a child who lives far away from one)

Extra-Linguistic Alternations

• Genre conventions
  ... according to a new study / new research says / null

• Mathematical Reasoning
  may quadruple the risk for
  may be four times more susceptible to
  may increase the risk ... by 400%

• In redundant context, equivalence may be clear
  — But how to answer “What quadruples the risk of Leukemia in kids?”

• Even less approachable:
  “four times more likely to develop” = “dramatically higher risk of”
  — Interpretation of “dramatically” is determined by non-linguistic factors
Dramatically different interpretations of “Dramatically”

If you keep a gun in your home, you dramatically increase the odds that you will die of a gunshot wound... Wiebe's study found that people with a gun in their home were almost twice as likely to die in a gun-related homicide, and 16 times more likely to use a gun to commit suicide, than people without a gun in their home.

At the other end of the scale, gun owners who had committed a misdemeanor that didn’t involve violence or guns were only four times as likely to commit another crime as those with no misdemeanor convictions.

Extra-linguistic Alternations in RTE

- Some RTE examples require “extensive text comprehension” (de Marneffe et al, 2008)

T: Nike Inc. said that its profit grew 32 percent, as the company posted broad gains in sales and orders.
H: Nike said orders for footwear totaled $4.9 billion, including a 12 percent increase in U.S. orders.

- But is even “text comprehension” enough?
  - Need business knowledge to decide even whether the percentages might be mappable
- Can any technology on the horizon address problems like this?
What’s our goal?

• True Artificial Intelligence? Of course we want this! But...
  – An open-ended, ill-defined agenda
  – Impossible even to build a coherent dataset
  – Impossible, therefore, to make real progress
• So how do we proceed?

What’s working in NLP?

*Learned Pattern Matching*

• Start with a corpus that characterizes a mapping that’s of interest, e.g.
  – Parsing: string ⇔ linguistic annotation
  – Translation: L1 string ⇔ L2 string

• Exploit ML/Statistical methods to learn how to best model this relationship

• Successes in NLP over the last 15 years driven by:
  – Empirical, data-driven techniques
  – Statistical pattern-learning
    • Increasing reliance on linguistic features (e.g. syntax, semantic role played by noun phrases, etc.)
  – Well-defined metrics/test sets
What can we learn from MT?

• Big quality gains over the last 15 years. Why?
  – Paradigm shift allowing wholesale borrowing of mature technologies (e.g. speech algorithms)
  – Availability of large quantities of human-created parallel corpora
  – An automated metric (BLEU)
  – An automated, rapid training-test cycle, with large training/test sets

• But equally important has been artificially limiting the:
  1. nature of the task
  2. sophistication of the metric
  3. user experience

MT: Let’s pretend that...

• Human-quality translation isn’t a goal. Instead,
  – translate sentence-by-sentence
  – no need to summarize/restructure content as a human translator would

• Users are easily satisfied
  – Only want to translate typed input, not speech
  – Don’t mind errors and disfluencies in the output
  – Are obsessed with government policy, news, and travel
    • Everyday/colloquial language: not so much

• BLEU accurately mimics human judgments

*Each of these assumptions is ridiculous, but collectively they have allowed the field to progress rapidly*
How can we similarly attack the problem of NL Semantics?

• The problem space is overwhelming
  – Need to start somewhere!
• Isolate a piece of the problem that:
  – We can agree on
  – Has some value to users, yet is fail-soft
  – Can be approached with today’s technology
• Focus on rapid training-test cycle
  – Need large training and test datasets
  – Live with an imperfect automated metric

What problem should we be trying to address?

• **Entailment**: language + deeper reasoning
  – When can the meaning of one text fragment be inferred from the other?
  – What are the limits of this task?
    • Do we allow mathematical reasoning?
    • Do we infer that “it rained last night” given that “the grass is wet”?
• **Paraphrase**: fundamentally about *language*
  – When can one syntactic/lexical unit be substituted for a different one without affecting the meaning?
    • Harris (1954): substitutability test for constituent definition
• As a field, we should focus on Paraphrase
A child who lives near a petrol (gas) station is four times more likely to develop leukemia than a child who lives far away from one, according to a new study. Living near to a petrol station or garage may increase the risk of acute childhood leukaemia by 400%. Children who live in close proximity to gas stations and auto body shops have a dramatically higher rate of leukemia, according to a new study. Living near a petrol station may quadruple the risk for children of developing leukaemia, new research says. Children who live near petrol stations may be four times more susceptible to leukaemia.

Vivre près d’un garage ou une station d’essence pourrait quadrupler le risque de leucémie infantile, suggère une étude française.

Vivir cerca de una gasolinera puede llegar a cuadriplicar el riesgo de leucemia en niños.

In the real world, as in translation, non-linguistic alternations occur

Text messaging spiked on election night, service providers say, with AT&T reporting a 44% surge in traffic.

There was a huge surge in SMS traffic during the 10 minutes after Barack Obama was officially named the president elect of the U.S. says Sybase 365.

Mobile messaging traffic surged Tuesday night immediately following the official confirmation of Barack Obama's election as U.S. President, according to mobile services provider Sybase 365.

Ten minutes after Obama was officially elected as the country’s 44th president at about midnight on the East Coast, the volume of messages surged more than three times the normal amount for that time of day, according to the mobile messaging service provider.

Mobile messaging traffic experienced an unprecedented surge Tuesday evening in the 10 minutes immediately following the official confirmation of Barack Obama's election as U.S. president, according to mobile messaging and content management firm Sybase 365.
But Key Advantages

• Prospect of amassing a significant training/test corpus
  – News data offers a significant source of sentence pairs with overlapping content
    • Over time, millions of pairs can be automatically collected and structured
    • Never enough data, but there is a lot out there, more every day
  – Naturally occurring corpus, so no issues with artificial skewing

• Can we eliminate the need for human annotators?
  – Maybe... but even if not, annotation task can be restricted until cognitive complexity is low/inter-rater agreement high
  – Candidate pairs can be automatically filtered, provisionally aligned
  – Annotators highlight only relevant portions of strings
    • Progressively enforce harsher standard for what counts as “paraphrase” until interannotator agreement is acceptable
    • This will tend to exclude many of the most interesting paraphrases, but that’s ok for now
    • Ideal task for Mechanical Turkers: cheap, large volume of data

Limiting Annotation Complexity

Text messaging spiked on election night, service providers say, with AT&T reporting a 44% surge in traffic.

There was a huge surge in SMS traffic during the 10 minutes after Barack Obama was officially named the president elect of the U.S. says Sybase 365.

Mobile messaging traffic surged Tuesday night immediately following the official confirmation of Barack Obama's election as U.S. President, according to mobile services provider Sybase 365.

Ten minutes after Obama was officially elected as the country’s 44th president at about midnight on the East Coast, the volume of messages surged more than three times the normal amount for that time of day, according to the mobile messaging service provider.

Mobile messaging traffic experienced an unprecedented surge Tuesday evening in the 10 minutes immediately following the official confirmation of Barack Obama's election as U.S. president, according to mobile messaging and content management firm Sybase 365.
The Metric?

• BLEU: a fully automated MT evaluation metric (Papineni et al, 2001)
  – Modified N-gram precision (comparing a test sentence to reference sentence(s))
    • Weighted geometric mean of 1-4 grams
    • Brevity penalty computed at corpus level

• Problem is reducible to MT, so use the same metric
  – News clusters yield multiple paraphrases tests sentence, increasing the reliability of an ngram-overlap metric

• But BLEU is totally inadequate for assessing semantics!
  – Of course it is! But so is BLEU for MT
  – “E pur si muove”, Galileo Galilei 1633

But Paraphrasing a Useful Application?

• Not as useful as translation, but there are some:
  – Editorial suggestions
  – Dialog echo-back
  – PlagiarizeThis!™ feature?

• A useable measure of paraphrase overlap is likely a key feature for ML’d approaches to:
  – Search / QA, MT, Summarization, etc.

• Will allow the field to make rapid, measurable progress
  – Rapid training-test cycle
  – Many proven techniques can be borrowed from MT

• Not the “Star Trek” vision, but a key piece of it
  – And much more approachable
MSR QA Dataset

• Will shortly be released for research purposes
  – Questions, detailed answer annotations, full text of Encarta 98
  – Check Microsoft Research web page for download information
• Approximates a real-world QA scenario
  – School-age children (10-13) asked “If you could talk to an encyclopedia, what would you ask?”
  – Questions were written on paper, keyed in with corrected spellings
  – A professional research librarian spent months searching target corpus to create an exhaustive recall set
    • Detailed information about linguistic similarity of question to answer
    • Assessment of whether the match coincides with likely query intent
  – 1.3K questions, 10K+ Q/A pairs
• Web keyword search works well for many questions
  – But goal was to force a hard look at recall in a finite corpus

Example of QA pair
Thank you!

(And please use the MSR QA Corpus!)