Paraphrase and Entailment: semantics for applications

Bill Dolan
Microsoft Research, Redmond
NLP Summer School, Bangalore
May 2007

Motivation

- Encourage you to think about problems in computational semantics
- In particular, how can we create applications that appear to understand natural language? Without:
  - Anticipating all inputs in advance
  - Hand-coding pattern for e.g. question answering
  - Tagging massive amounts of data for supervised learning

Motivation

- Encourage you to think about problems in computational semantics
- In particular, how can we create applications that appear to understand natural language? Without:
  - Anticipating all inputs in advance
  - Hand-coding pattern for e.g. question answering
  - Tagging massive amounts of data for supervised learning

Q: *Who is John Lennon's widow?*

A: *Yoko Ono unveiled a bronze statue of her late husband, John Lennon, to complete the official renaming of England's Liverpool Airport as Liverpool John Lennon Airport.*

Motivation

- Encourage you to think about problems in computational semantics

- In particular, how can we create applications that appear to understand natural language? Without:
  - Anticipating all inputs in advance
  - Hand-coding pattern for e.g. question answering
  - Tagging massive amounts of data for supervised learning

Q: *Who is John Lennon's widow?*

A: *Yoko Ono unveiled a bronze statue of her late husband, John Lennon, to complete the official renaming of England's Liverpool Airport as Liverpool John Lennon Airport.*

Traditional approach to NLP

- Build discrete components for different analysis levels:
  - part of speech tagging
  - named entity recognition
  - parsing
  - dependency analysis
  - word sense disambiguation
  - anaphora resolution
  - discourse

- Working assumption:
  - success in all of these areas → applications that appear to understand language

Traditional approach to NLP

- Build discrete components for different analysis levels:
  - part of speech tagging
  - named entity recognition
  - parsing
  - dependency analysis
  - word sense disambiguation
  - anaphora resolution
  - discourse
- Working assumption:
  - success in all of these areas → applications that appear to understand language
- But none of this work has had much impact on applications!

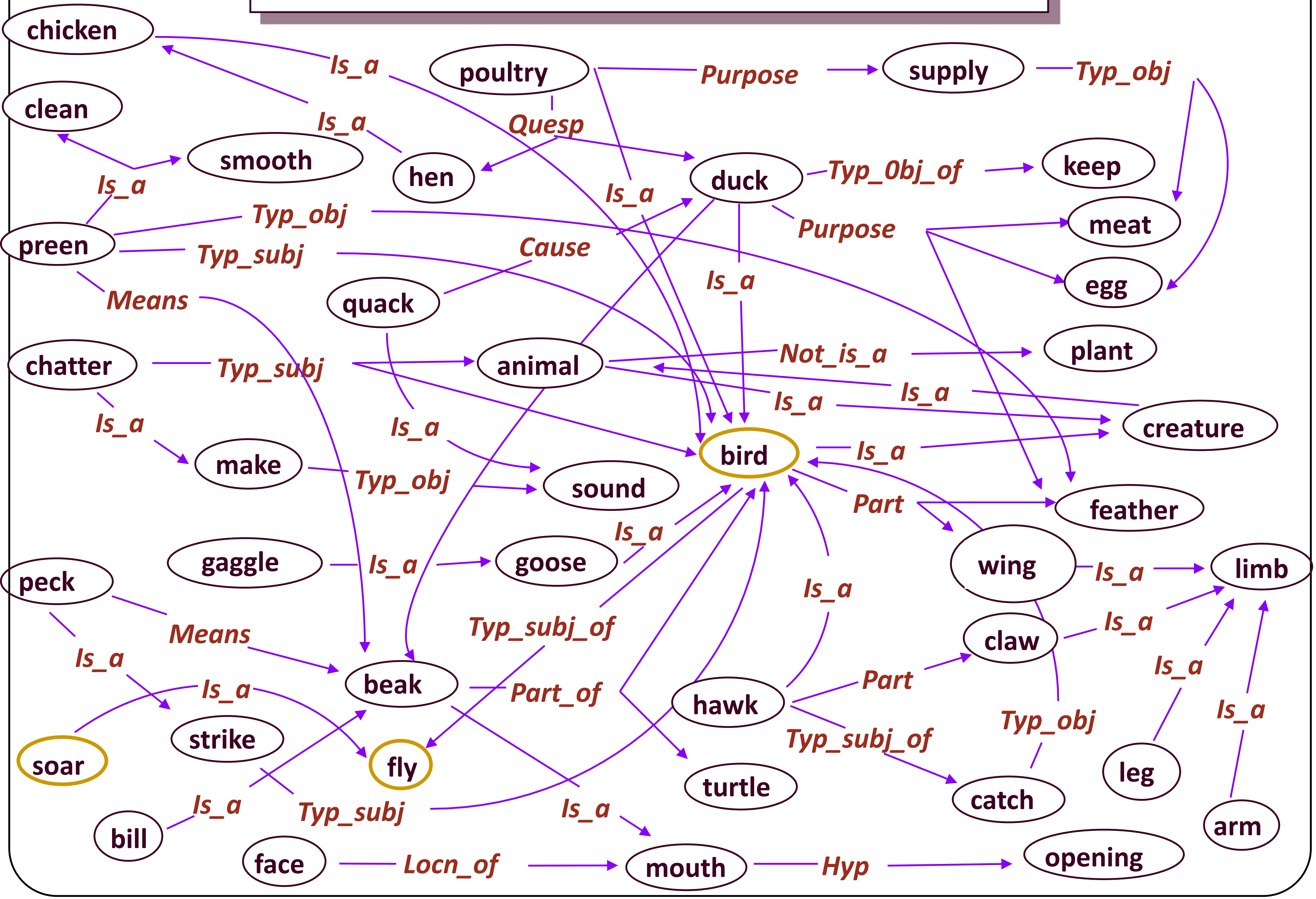
Building Applications that Appear to Understand Language

- What do applications actually *need*?
  - Information Retrieval/Question Answering
  - Summarization
  - Proofing tools: grammar, style rewrites
  - Dialog / command-and-control
- Not a word sense, syntactic structure, a cluster of related words, discourse structure, etc. Instead, the ability to:
  - Identify when two strings "mean the same thing" in a particular context, regardless of superficial similarity
  - Identify when one string "entails" the meaning of another
  - Generate string a meaning-preserving rewrite
- Empirical models of semantic overlap

A personal anecdote: what I learned from MindNet, circa 1999

- Maybe I can save you a few years' worth of trouble!
- MindNet: an automatically-constructed knowledge base (Dolan et al, 1993; Richardson et al 1998)
  - Built from free text (e.g. dictionaries, Encarta)
  - 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

Fragment of lexical space surrounding "bird"



Finding paths through the network: "bird" and "soar"

soar

Source Definitions (*Longman Dictionary of Contemporary English***):**

wing*, n 1:*

One of the two feathered limbs by which a bird flies, or a transparent limb on an insect

soar*, v 1:*

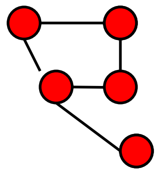
to fly; go fast or high as on wings; sail in the air

Question Answering with MindNet

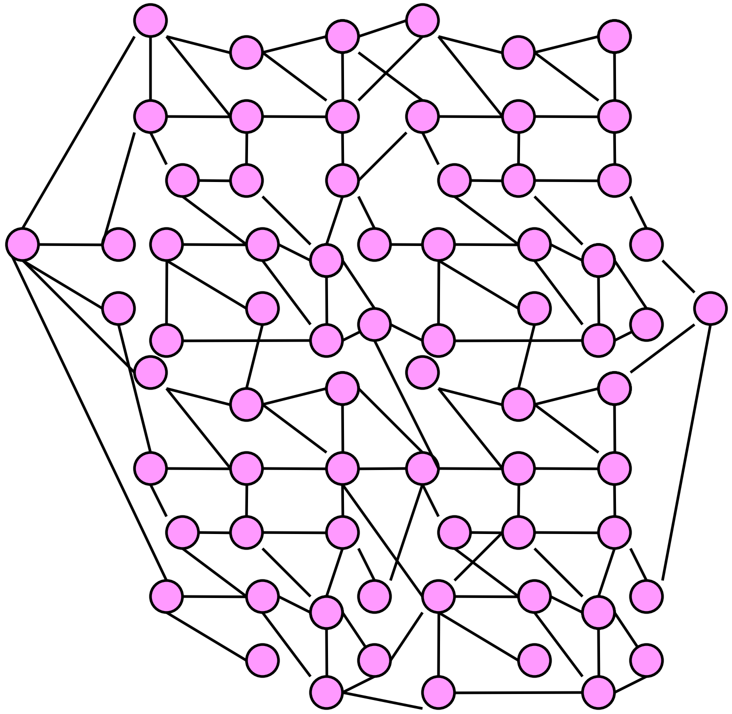
- Build a MindNet graph from:
  - Text of dictionaries
  - Target corpus, e.g. Microsoft Encarta
- Build a dependency graph from query
- Model QA as a graph matching procedure
  - Heuristic fuzzy matching for synonyms, named entities, wh-words, etc.
  - Some common sense reasoning (e.g. dates, math)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

Logical Form Matching (2)

Input LF:

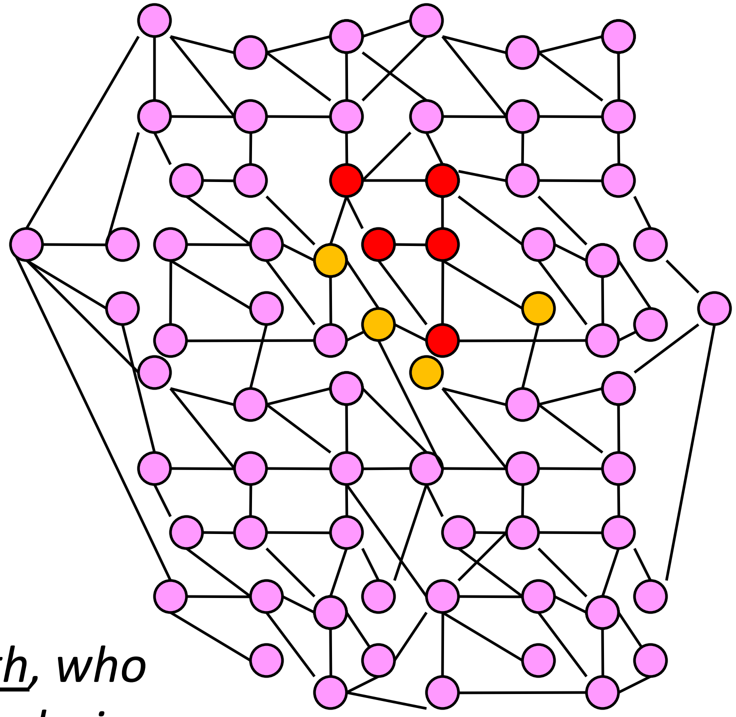


MindNet



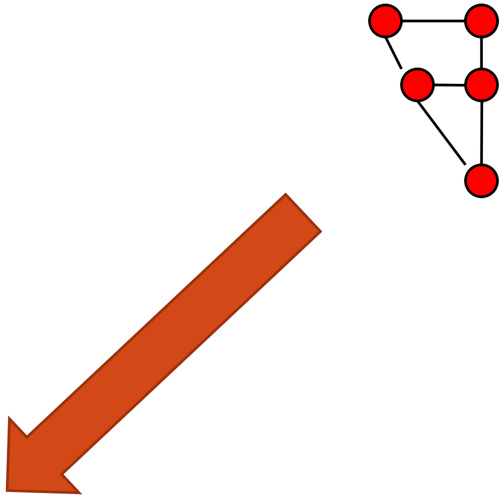
Who assassinated Abraham Lincoln?

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"

Worked beautifully!

- Just not very often…
- Most of the time, the approach failed to produce any answer at all, even though
  - Answer was present in the target corpus
  - Dependency analysis for query/target strings was correct
- Approach required close isomorphism between query and answer dependency graphs
  - Often, the data doesn’t work this way
  - Query and answer phrased in fundamentally different ways

Q: *Who is John Lennon's widow?*

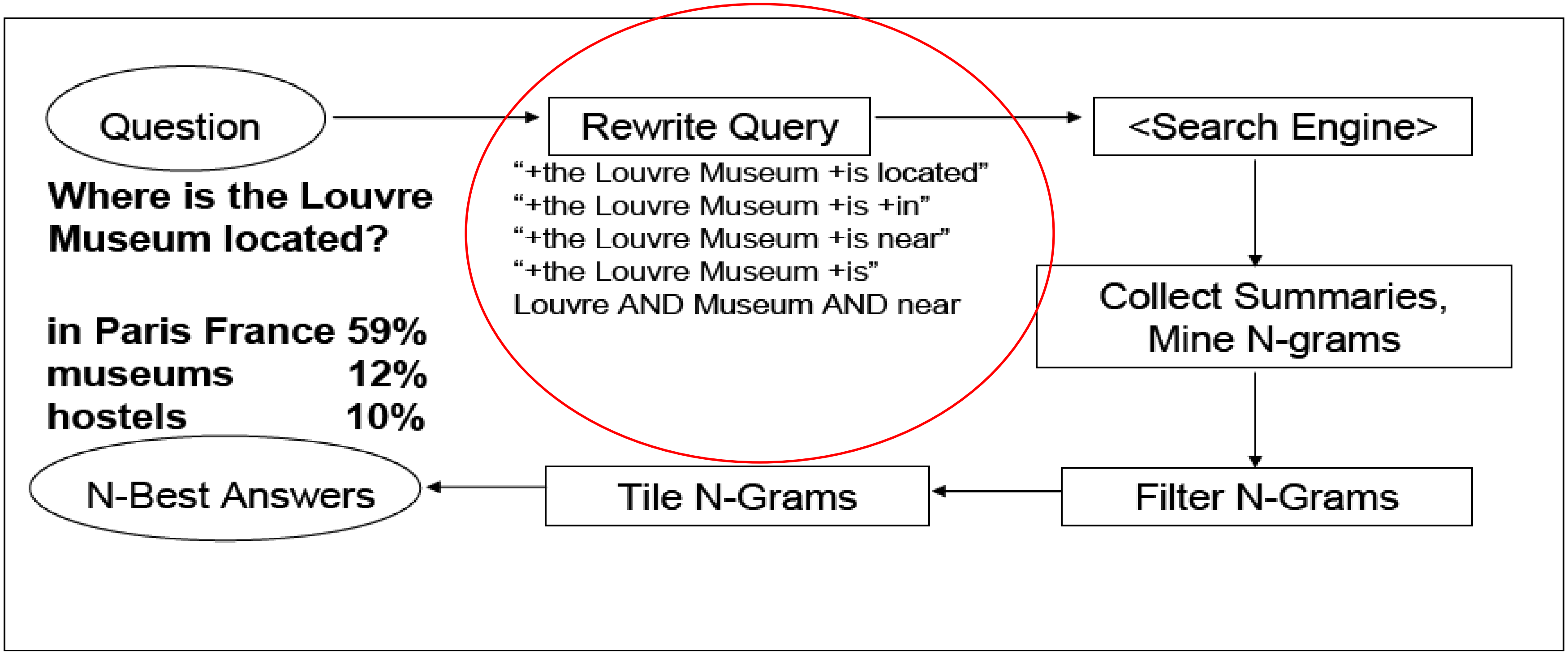
A: *Yoko Ono unveiled a bronze statue of her late husband, John Lennon, to complete the official renaming of England's Liverpool Airport as Liverpool John Lennon Airport.*

Traditional levels of linguistic analysis – at least on their own – don't "solve" this problem

Any suggestions for solutions?

Regular Expression Rewrites for QA

- Enumerate anticipated variants in advance
  - A common approach to the paraphrase problem
  - Can work well for limited query types
    - e.g. Brill et al. 2002: AskMSR “location” queries



Temptation

- Deal with match failures by heuristically relaxing matching procedures
  - e.g. explicitly allow matches of graph fragments like

Y's late husband, X = X's widow is Y

But this is a long, slippery slope!

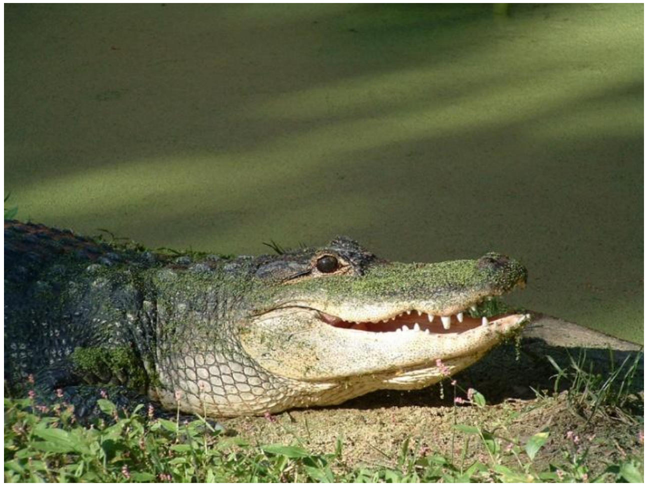
We need to find a principled solution

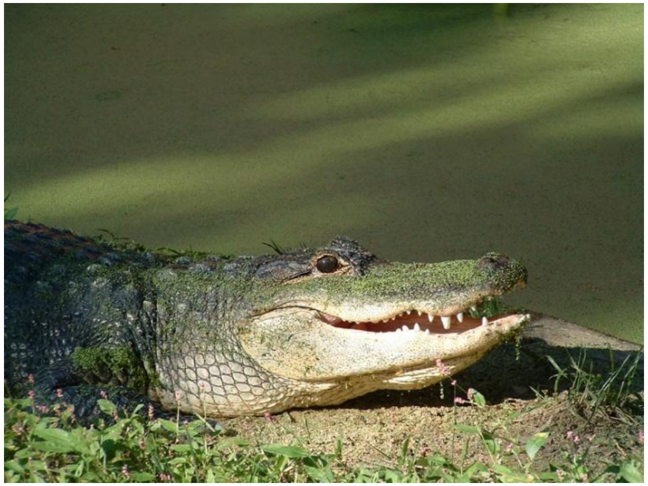
Building Applications that Appear to Understand Language

- What do applications actually *need*?
  - Empirical techniques for modeling paraphrase and entailment relationships between text chunks

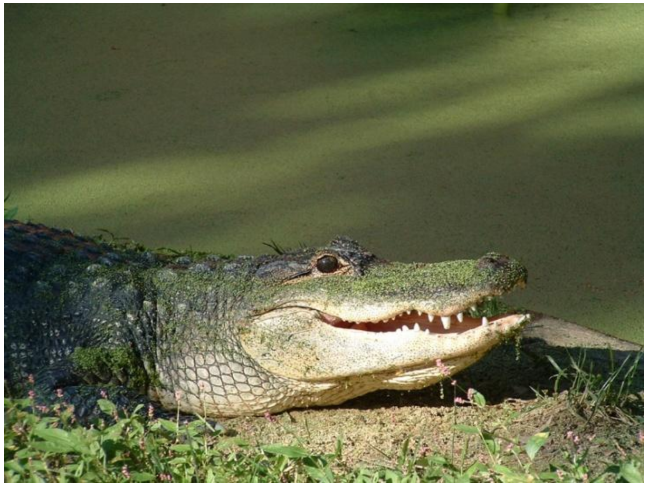
Outline

- Motivation
- Textual Entailment
  - Overview of problem, literature
  - PASCAL: Recognizing Textual Entailment Challenges
- Paraphrase Recognition and Generation
  - Overview of problem, literature
  - MSR approach
- Corpus issues
  - (time permitting): a work in progress





+



+

=

Exploding Snakes

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

= really big snake

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

= exploded

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

= in Florida

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

= swallowed, no chewing

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

= fairly big alligator

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

= not sure what happened

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

= scientists

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

So how is this relevant for applications?

Application: Question Answering

What causes childhood leukemia?

- 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.

Factoring out redundancy

Apps: Multi-document Summarization, Proofing

- 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.

Outline

- Motivation
- Textual Entailment
  - Overview of problem, literature
  - PASCAL: Recognizing Textual Entailment Challenges
- Paraphrase Recognition and Generation
  - Overview of problem, literature
  - MSR approach
- Corpus issues
  - (time permitting): a work in progress

Text applications require *semantic* inference

- Do two text chunks overlap semantically?
- What is the nature of the overlap?
- A systematic assault on these problems would have implications for virtually all applications that involve NL understanding

- But how do we approach the problem?

A common framework for applied semantics is needed

- Two related research threads aimed at filling this gap
  - Modeling Semantic Entailment
    - Recognizing entailment relationship between two texts
  - Recognizing and Generating Paraphrases
    - Recognizing paraphrase alternations
    - Generating meaning-preserving rewrites of an input

Desiderata for Modeling Framework

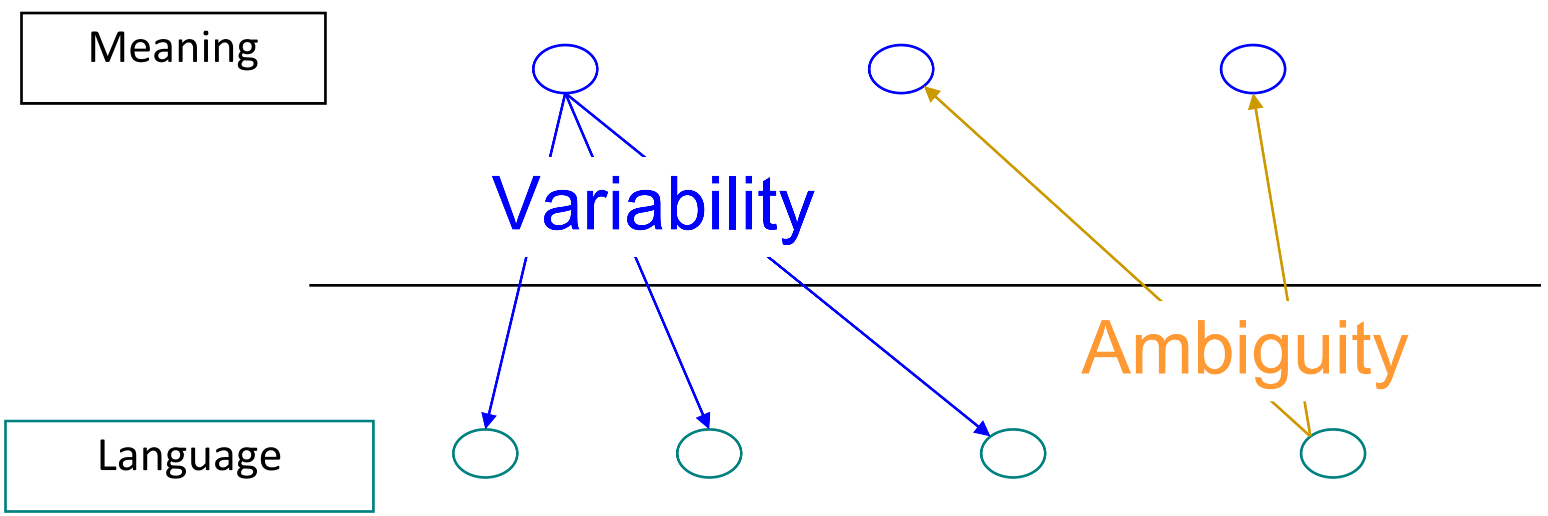
- A framework for a target level of language processing should provide:
  - *Generic module for different applications (e.g. parser)*
  - *Unified paradigm for investigating language phenomena*
  - *Unified knowledge representation*
- Most semantic work does not meet these desiderata
  - Research is instead scattered, no shared definition of what to model – much less a unified framework for doing so
  - Disjoint/isolated subfields, each with own evaluation criteria and relationships (or not) to applications
    - Word Sense Disambiguation
    - Named Entity Recognition
    - Semantic Role Labeling
    - Lexical semantics relations
- Dominating theme – interpretation
  - Semantics is about meaning, so representation needs to be invented

PASCAL RTE Challenges

- Recognizing Textual Entailment
- The textual entailment task – what and why?
- Evaluation

- *An alternative framework for investigating semantics*
- *(Thanks to Ido Dagan for many of these slides)*

Natural Language and Meaning



Variability of Semantic Expression

The Dow Jones Industrial Average closed up 255

Dow ends up

Dow gains 255 points

Dow climbs 255

Stock market hits a record high

Model variability as relations between text expressions:

- **Equivalence**: *text1* $\Leftrightarrow$ *text2* (paraphrasing)
- **Entailment:** *text1* $\Rightarrow$ *text2*

Entailment vs. Semantic Equivalence

Unidirectional Entailment

I bought a science fiction novel.

I bought a book.

Bidirectional (more specific case = paraphrase/equivalence)

On its way to an extended mission at Saturn, the Cassini probe on Friday makes its closest rendezvous with Saturn's dark moon Phoebe.

The Cassini spacecraft, which is en route to Saturn, is about to make a close pass of the ringed planet's mysterious moon Phoebe

Entailment vs. Semantic Equivalence

- Closely related notions
  - I'll be addressing the two separately, but the line between them is often blurry
- Recognition vs. Generation
  - Most work on paraphrase/equivalence has focused on recognition, but some recent work on generating paraphrases
  - For Entailment, focus solely on recognition

Example

Typical Application Inference

Question
Who bought Overture? >>

Expected answer form
X bought Overture

Overture's acquisition by Yahoo

text

entails

Yahoo bought Overture

hypothesized answer

- Same inferencing capability that we need for other applications:
  - Information Extraction: *X buy Y*
  - Summarization (multi-document)
  - MT evaluation
  - Educational applications:

KRAQ'05 Workshop - KNOWLEDGE and REASONING for ANSWERING QUESTIONS (IJCAI-05)

Call For Papers:

- Reasoning aspects:
  * information fusion,
  * search criteria expansion models
  * summarization and intensional answers,
  * reasoning under uncertainty or with incomplete knowledge
- Knowledge representation and integration:
  * levels of knowledge involved (e.g. ontologies, domain knowledge),
  * knowledge extraction models and techniques to optimize response accuracy

But other applications have very similar needs from a semantic component!

• Can notions of semantic overlap/equivalence provide a **common** empirical task?

SIGLEX ACL Special Interest Group

SemEval-2007

4th International Workshop on Semantic Evaluations

Task	Download Data	Task Summary	Further Details	Task Website (if available)
1	Data	Evaluating WSD on Cross Language Information Retrieval	View	Webpage
2	Data	Evaluating Word Sense Induction and Discrimination Systems	View	Webpage
3		Task Cancelled: ~~Pronominal Anaphora Resolution in the Prague Dependency Treebank 2.0~~		
4	Data	Classification of Semantic Relations between Nominals	View	Webpage
5	Data	Multilingual Chinese-English Lexical Sample Task	View	
6	Data	Word-Sense Disambiguation of Prepositions		Webpage
7	Data	Coarse-grained English all-words	View	Webpage
8	Data	Metonymy Resolution at Semeval-2007	View	Webpage
9	Data	Multilevel Semantic Annotation of Catalan and Spanish	View	Webpage
10	Data	English Lexical Substitution Task for SemEval-2007	View	Webpage
11	Data	English Lexical Sample Task via English-Chinese Parallel Text		Webpage
12	Data	Turkish Lexical Sample Task	View	Webpage
13	Data	Web People Search		Webpage
14	Data	Affective Text	View	Webpage
15	Data	TempEval: A proposal for Evaluating Time-Event Temporal Relation Identification	View	Webpage
16	Data	Evaluation of wide coverage knowledge resources		Webpage
17	Data	English Lexical Sample, English SRL and English All-Words Tasks		Webpage
18	Data	Arabic Semantic Labeling	LDC License	
19	Data	Frame Semantic Structure Extraction	View	Webpage

Classical Entailment Definition

- Chierchia & McConnell-Ginet (2001):
  A text t entails a hypothesis h if h is true in <u>every</u> circumstance (possible world) in which t is true

- Strict entailment - doesn't account for some uncertainty allowed in applications
  - Cottage industry in linguistic s: inventing bizarre circumstances in which an entailment doesn’t hold
  - To produce a dataset that is relevant for applications, we have to use a weaker definition

“Almost certain” Entailments

T: *The technological triumph known as GPS … was incubated in the mind of Ivan Getting.*

H: *Ivan Getting invented the GPS.*

T: *About two weeks before the trial started, I was in Shapiro’s office in Century City.*

H: *Shapiro works in Century City.*

Applied Textual Entailment

- Directional relation between two text fragments: *Text (t)* and *Hypothesis (h):*

t **entails** *h* ***(t⇒h)*** if **humans** reading *t* will infer that *h* is **most likely** true

- Operational (applied) definition:
  - Human gold standard - as in NLP applications
  - Assumes common background knowledge – both of the language and world
    - But then this is what applications expect/need
  - Assumption is that educated native speakers share enough background to allow meaningful evaluations

Probabilistic Interpretation

<u>Definition</u>:

- *t probabilistically entails h* if:
  - P(*h is true* | *t*) > P(*h is true*)
    - *t* increases the likelihood of *h* being true
- P(*h is true* | *t*): *entailment confidence*
  - The relevant entailment score for applications
  - In practice: "most likely" entailment expected

The Role of Knowledge

- For textual entailment to hold we require:
  - *text* AND *knowledge* $\Rightarrow$ *h*

    but
  - *knowledge* should not entail *h* alone

- Systems are not supposed to validate *h*'s truth without utilizing *t*

PASCAL Recognizing Textual Entailment (RTE) Challenges

2004-7

RTE-1 (2004-05): *Southampton, England*
RTE-2 (2005-06): *Venice, Italy*
RTE-3 (2006-07): *Prague, the Czech Republic*

http://www.pascal-network.org/Challenges/RTE/

Organizers:

Bar-Ilan University
MITRE
ITC-irst and CELCT, Trento
Microsoft Research

Data drawn from different applications

- Most data created from actual application output
- 7 application settings in RTE-1, 4 in RTE-2/3
  - QA
  - IE
  - “Semantic” IR
  - Comparable documents / multi-doc summarization
  - MT evaluation
  - Reading comprehension
  - Paraphrase acquisition
- RTE-2: 800 examples each in development and test sets
  - Example creation a manual process, grammar/spelling errors fixed
  - 50-50% YES/NO split
- Upper bound (human performance) for RTE2 test
  - Kappa agreement on test items: K = 0.94 (good agreement)
  - Fewer controversial examples than in RTE1

Some Examples

	TEXT	HYPOTHESIS	TASK	ENTAIL-MENT
1	*Reagan attended a ceremony in Washington to commemorate the landings in Normandy.*	*Washington is located in Normandy.*	IE	False
2	*Google files for its long awaited IPO.*	*Google goes public.*	IR	True
3	*…: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.*	*Cardinal Juan Jesus Posadas Ocampo died in 1993.*	QA	True
4	*The SPD got just 21.5% of the vote in the European Parliament elections, while the conservative opposition parties polled 44.5%.*	*The SPD is defeated by the opposition parties.*	IE	True

More Examples

	TEXT	HYPOTHESIS		ENTAIL-MENT
1	*Tibone estimated diamond production at four mines operated by Debswana – Botswana's 50-50 joint venture with DeBeers – could reach 33 million carats this year.*	*Botswana is a business partner of DeBeers.*		True
2	*The EZLN differs from most revolutionary groups by having stopped military action after the initial uprising in the first two weeks of 1994.*	*EZLN is a revolutionary group.*		True
3	*Two persons were injured in dynamite attacks perpetrated this evening against two bank branches in this Northwestern Colombian city.*	*Two persons perpetrated dynamite attacks in a Northwestern Colombian city.*		False
4	*Such a margin of victory would give Abbas a clear mandate to renew peace talks with Israel, rein in militants and reform the corruption-riddled Palestinian Authority.*	*The new Palestinian president combated corruption and revived the Palestinian economy.*		False

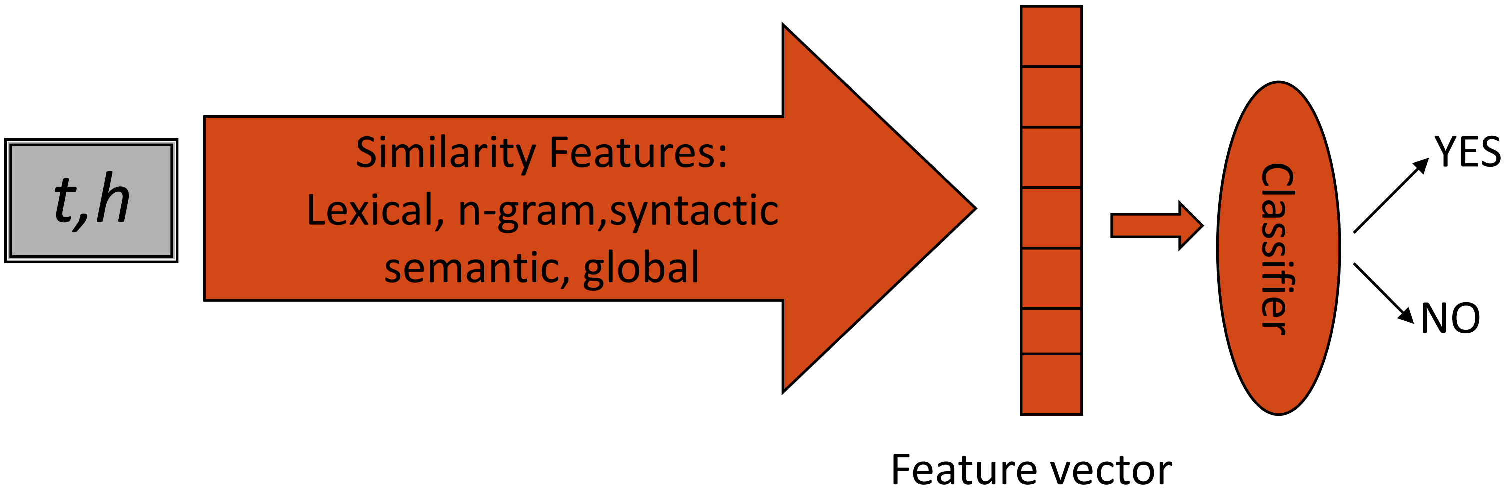
Participation and Impact

- Very successful challenges, worldwide:
  - RTE-1 – 17 groups
  - RTE-2 – 23 groups
    - 30 groups in total (but none from India!)
    - >150 downloads of training set
  - RTE-3 underway – 25 groups
    - Joint RTE/Paraphrase workshop at ACL-07 in Prague
- Great interest in the research community
  - Papers, conference sessions and areas, PhD's, influence on funded projects
  - Upcoming special issue on Textual Entailment for Journal of Natural Language Engineering (JNLE)
  - ACL-07 tutorial

Methods and Approaches (RTE-2)

- Measure similarity match between *t* and *h* (*coverage* of *h* by *t)*:
  - Lexical overlap (unigram, N-gram, subsequence)
  - Lexical substitution (WordNet, statistical)
  - Syntactic matching/transformations
  - Lexical-syntactic variations ("paraphrases")
  - Semantic role labeling and matching
  - Global similarity parameters (e.g. negation, modality)
- Cross-pair similarity
- Detect mismatch (for non-entailment)
- Logical interpretation and inference (vs. matching)

Dominant approach: Supervised Learning



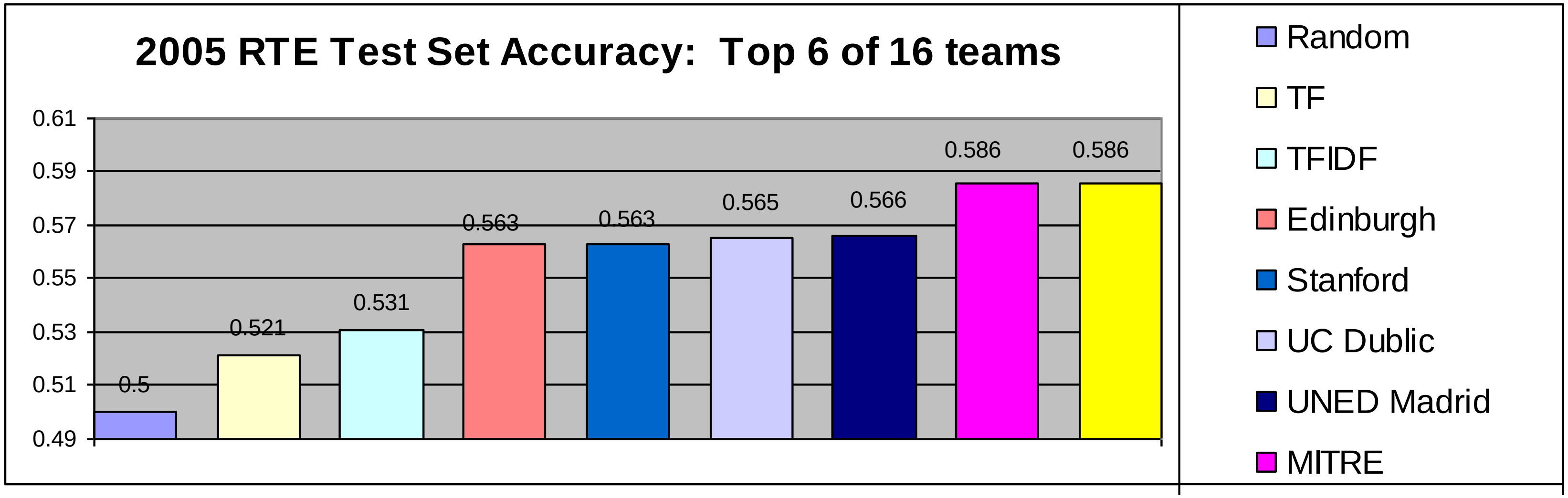
- Features model similarity and mismatch
- Classifier determines relative weights of information sources
- Train on development set and auxiliary *t-h* corpora

RTE 2: Results

First Author (Group)	Accuracy	Average Precision
Hickl (LCC)	75.4%	80.8%
Tatu (LCC)	73.8%	71.3%
Zanzotto (Milan & Rome)	63.9%	64.4%
Adams (Dallas)	62.6%	62.8%
Bos (Rome & Leeds)	61.6%	66.9%
11 groups	58.1%-60.5%	
7 groups	52.9%-55.6%	

Average: 60%
Median: 59%

Compare to results of RTE 1



- RTE1: overall, poor performance
  - On most data sets, average performance at or below baseline (47.7-51.9)
  - Only on "comparable documents" did systems perform well (avg. 73.3%)
    - Easiest set because lots of ngram overlaps, amenable to string-edit distance techniques
- Little to show for use of deep NL inferencing techniques
  - No improvements with algorithms for computing lexical similarity, WordNet, thesauri, dependency parsing, theorem provers, etc.

RTE 2: analysis

- For the first time: deeper methods (semantic/ syntactic/ logical) seemed to outperform shallow methods (lexical/n-gram) ☺

Parallels situation in MT, where deeper methods are for the first time beginning to outperform string-based methods

- But most systems based on deep analysis did not score significantly better than the lexical baseline
  - Around 60% on this dataset
  - Even the winning system, while parser-based, might have done well with string-based techniques alone

Why didn't systems do better?

- System reports point to problems with:
  - Lack of knowledge (syntactic transformation rules, phrasal paraphrase alternations, lexical relations, etc.)
  - Lack of training data

- Without knowledge of what might entail what, inference mechanisms are helpless

- Systems that coped better with these issues performed best:
  - Hickl et al.(2006) - acquisition of large entailment corpora for training
  - Tatu et al. (2006)– large knowledge bases (linguistic and world knowledge)

Some Potential Issues with RTE

- Artificial distribution of phenomena
  - Problematic cases for agreement discarded
  - Only in retrospect can you tell that e.g. polarity markers correctly cue judgments
  - Working backward from application data may bias set toward simpler examples
- 50/50 True/False split may overstate usefulness of resulting metric
  - In real world, tough to find pairs of sentences that exhibit clean entailment relationships
  - Will it massively over-flag on real data?
  - Test is gameable: do your best, then flip a coin

Groundhog: positive examples

- Burger and Ferro (2005), RTE1
  - Assume that a news article headline is a paraphrase or synopsis of the first line of the article
- Groundhog: 100K pairs
  - Filtered by discarding pairs with no shared entity/NP, pairs in stock price/sports domains
  - 2296/2500 (91.8%) judged correct in sample

Judgment	Example
YES	Text: Sydney newspapers made a secret deal not to report on the fawning and spending during the city's successful bid for the 2000 Olympics, former Olympics Minister Bruce Baird said today.
	Hypothesis: Papers Said To Protect Sydney Bid
YES	Text: An IOC member expelled in the Olympic bribery scandal was consistently drunk as he checked out Stockholm's bid for the 2004 Games and got so offensive that he was thrown out of a dinner party, Swedish offi cials said.
	Hypothesis: Offi cials Say IOC Member Was Drunk

Groundhog: negative examples

- 100K pairs for negative training:
  - Sequential sentences with shared named entities (2438/2500 or 97% correct)
  - Sentences linked by discourse connectives like “even though”, “although”, “but”, “in contrast” (2942/1000 or 94% correct)

Judgment	Example
NO	**Text:** One player losing a close friend is Japanese pitcher Hideki Irabu, who was befriended by Wells during spring training last year.
	Hypothesis: Irabu said he would take Wells out to dinner when the Yankees visit Toronto.
NO	**Text:** According to the professor, present methods of cleaning up oil slicks are extremely costly and are never completely effi cient.
	Hypothesis: *In contrast*, he stressed, Clean Mag has a 100 percent pollution retrieval rate, is low cost and can be recycled.

Some Potential Issues with RTE

- No clear distinction between paraphrase / entailment examples
  - Many examples are bidirectional

    Satomi Mitarai died of blood loss.

    Satomi Mitarai bled to death.
  - (But this may just reflect application reality)

Some Potential Issues with RTE (2)

- No explicit separation of cases requiring world knowledge vs. linguistic knowledge
  - Often difficult to decide
  - But result is that "straightforward" paraphrase examples are intermingled with complex AI-style reasoning problems

T *Meanwhile, in an exclusive interview with a TIME journalist, the first one-on-one session given to a Western print publication <u>since his election as president of Iran earlier this year, Ahmandinejad</u> attacked the "threat" to bring the issue of Iran's nuclear activity to the UN Security Council by the US, France, Britain and Germany.*

H *Ahmandidnejad is a citizen of Iran.*

Where does linguistic inferencing stop and “deeper” reasoning begin?

- Meaning equivalence encompasses much more than just linguistic equivalence or entailment

  Three adults and two children boarded the bus.
  Five people got on the bus.

  - How do we add 2+3? Is this really an NLP problem?

  I was there on Tuesday night
  I was there last night

  The grass outside is wet.
  It rained last night.

  - Where does “linguistic” knowledge end?

  I like eggplant
  I like the fruit of the plant Solanum melongena.

- Task must be bounded to ensure:
  - Clear evaluations
  - Reasonably good agreement on datatset annotation

Some suggested research directions

- Knowledge acquisition
  - Unsupervised acquisition of linguistic and world knowledge from general corpora and web
  - Acquiring larger entailment corpora
- Failure analysis of inferences
  - What types of inferences are feasible/infeasible given current approaches?
  - What types are necessary? (e.g. is linguistic information sufficient for most cases?)
  - Is deep syntactic information useful? (In principle, yes: Vanderwende et al 2005). But why is this difficult to show in practice?

Outline

- Motivation
- Textual Entailment
  - Overview of problem, literature
  - PASCAL: Recognizing Textual Entailment Challenges
- Paraphrase Recognition and Generation
  - Overview of problem, literature
  - MSR approach
- Corpus issues
  - (time permitting): a work in progress

Some previous work on Semantic Similarity

- Many efforts in a variety of NL applications to hand-code paraphrase relationships
  - Command and control, QA, etc.
  - Feasible in a small domain, but overwhelming for non-toy applications
- Recent research has treated paraphrase acquisition and generation as a machine learning problem
  - Richardson, 1997; Barzilay & McKeown, 2001; Lin & Pantel, 2002; Shinyama et al, 2002, Barzilay & Lee, 2003, Pang et al., 2003

Distributional Hypothesis

- Basis for much of the empirical work in applied lexical semantics
- Words that occur in the same contexts tend to have similar meanings (Harris, 1954)
  - “[A] word is characterized by the company it keeps”, Firth (1957)

Path Similarity in MindNet

- Richardson, 1997
- Empirical metric for lexical similarity
  - Based on dependency paths in MindNet, a semantic network of interlinked parse structures
  - Similarity function trained on known thesaurus pairs (similar by definition)

sandwich <-Object_of- eat -> Object -> apple
eagle – Is_A -> bird <- Subject_of – fly – Subject -> plane

- Result: a set of path shapes correlated with lexical similarity

* <-Object_of- * -> Object -> *
* – Is_A -> * <- Subject_of – * – Subject -> *

- Similarity of novel word pairs now measured based on shapes of paths linking them in semantic network

ball <-Object_of- throw -> Object -> rock

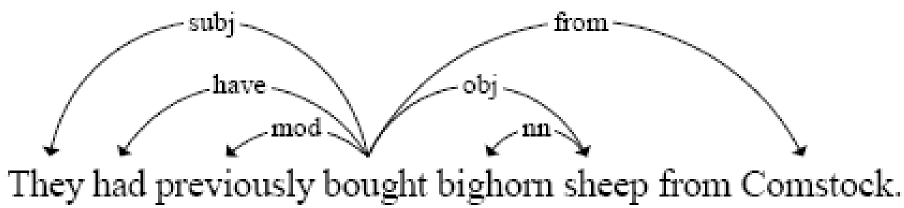
DIRT: Discovery of Inference Rules from Text

- Lin & Pantel 2001
- Automatically learns paraphrase expressions from text
  - Corpus annotated with dependency parses
  - Generalization of Distributional Hypothesis: not just similar words, but similar dependency subgraphs
- Focuses on (binary) dependency paths linking 2 nouns
  - If two paths tend to link the same sets of words, DIRT hypothesizes that the meanings of the two paths are similar

Dickens <u>is the author of</u> "Great Expectations"
"Great Expectations" <u>was written by</u> Dickens
Charles Dickens <u>wrote</u> "Great Expectations"

- Result is a set of "inference rules" or paraphrase alternation patterns

DIRT: extracting paths



(a) N:subj:V←buy→V:from:N
≡ *X* buys something from *Y*

(b) N:subj:V←buy→V:obj:N
≡ *X* buys *Y*

(c) N:subj:V←buy→V:obj:N→sheep→N:nn:N
≡ *X* buys *Y* sheep

(d) N:nn:N←sheep←N:obj:V←buy→V:from:N
≡ *X* sheep is bought from *Y*

(e) N:obj:V←buy→V:from:N
≡ *X* is bought from *Y*

DIRT: finding similar paths

N:subj:V <- find -> V:obj:N->solution->N:to:N — *"X finds a solution to Y"*

N:subj:V<-solve-> V:obj:N — *"X solves Y"*

"X finds a solution to Y"		*"X solves Y"*	
SlotX	SlotY	SlotX	SlotY
commission	strike	committee	problem
committee	civil war	clout	crisis
committee	crisis	government	problem
government	crisis	he	mystery
government	problem	she	problem
he	problem	petition	woe
legislator	budget deficit	researcher	mystery
sheriff	dispute	sheriff	murder

Overlapping slot fillers for two paths extracted from newspaper corpus

DIRT paraphrase resource

- Available on request
- Output of DIRT over 1GB of news text
  - 7 million paths extracted (231K unique)
  - Paraphrases generated from these
- Top 20 paraphrases for "X solves Y"

Y is solved by X	*X rectifies Y*
X resolves Y	*X copes with Y*
X finds a solution to Y	*X overcomes Y*
X tries to solve Y	*X eases Y*
X deals with Y, Y is resolved by X	*X tackles Y*
X addresses Y	*X alleviates Y*
X seeks a solution to Y	*X corrects Y*
X does something about Y	*X is a solution to Y*
X solution to Y	*X makes Y worse*
Y is resolved in X	*X irons out Y*
Y is solved through X	

Emerging Consensus: empirical approach trained on parallel data

- Broad assumption that this is the right direction
  - But no consensus on either data source or specific techniques
- A particular agenda: paraphrase as translation
  - MSR Redmond approach (Quirk et al 2004; Dolan et al 2004, Brockett et al 2005)
  - A key difference from most other work: generation as well as recognition

What's working in NLP?
Learned Pattern Matching

- 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
- A departure from earlier, AI-oriented approaches
  - Data-driven MT is "just" pattern matching, yet achieves much better results than "deep"/AI approaches
  - A priori, you might expect high-quality translation to require a thorough understanding of NL semantics
- Maybe we can think about monolingual understanding in the same way

Paraphrase *is* Translation

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

Paraphrase *is* Translation

A four-metre python exploded when it tried to swallow a two-metre alligator whole in Florida.

Animal experts in the US are baffled after a huge snake exploded when trying to snack on a giant alligator.

Scientists in Florida are puzzling over a Burmese python that scarfed down a six-foot alligator before its stomach ruptured.

A python in Florida's Everglades clashed with a 6-foot-long alligator. And won. But then tried to swallow him whole. And exploded.

Scientists in Florida still aren't quite sure how a 13-foot Burmese python managed to devour a six-foot alligator in the Everglades.

A 13-foot Burmese python recently burst after it apparently tried to swallow a live, six-foot alligator whole, authorities said.

4mの大蛇が1.8mの生きたワニを丸飲みして破裂！

一条身长13英尺的缅甸巨蟒将一头6英尺长的美洲鳄整个活吞了下去，结果被撑破了肚皮，双双俱亡。

미국 플로리다주 에버글레이즈 국립공원에서 길이 4m의 미얀마 비단뱀이 몸 길이가 1.8m나 되는 악어를 통째로 삼키다 배가 터져 죽은 사상 초유의 현장이 발견돼 공포 영화를 무색케 하는 전율을 자아내고 있다.

Multilingual QA: What causes childhood leukemia?

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.

Paraphrase as Monolingual MT

- Model paraphrase normalization/generation as a translation problem
  - Rely on standard string-based statistical machine translation techniques (Dolan et al, 2004; Quirk et al. 2004; Brockett et al 2005)
- Statistical Machine Translation (MT)
  - Brown et al (1993)
  - Compute optimal alignment of words in source/target sentence pairs
  - Use resulting models for passive recognition or generation (decoding)
- SMT already has some solutions
  - Not dependent on string identity
  - Can handle moderately broad domains
  - Full sentence generation implemented
  - Identifies contextually appropriate substitutions
- Follow same process as bilingual application
  - Acquire a corpus of parallel sentences
  - Train and decode using SMT techniques

"If you're a carpenter, everything looks like a nail"

- Common for a field with a successful paradigm – like SMT – to re-apply that paradigm to other problems
  - If you have a hammer, you look for nails
  - That's why SMT modeling techniques were borrowed from Speech Recognition in the first place
- But in this case, it's really the nails that look alike!
  - Whatever the paradigm, paraphrase and bilingual translation problems should be handled by the same machinery
- While we're at it: piggyback on SMT's rapid progress/energy/funding
  - Established evaluation metrics
  - Improved modeling techniques
  - Improved parallel comparable data collection techniques.
  - We'd like to piggyback on its success and proven evaluation metrics

Monolingual Translation: How hard can it be?

- Is an MT approach to equivalence feasible?
- Statistical MT varies in effectiveness depending on
  - Amount of training data: more is better
  - Character of training data:
    - Narrow vs. broad domain
    - closely parallel vs. loose translations
  - Language pair
    - The more closely related, the better (Koehn 2005)
- What could be more closely related than English-English?
  - Should be easy

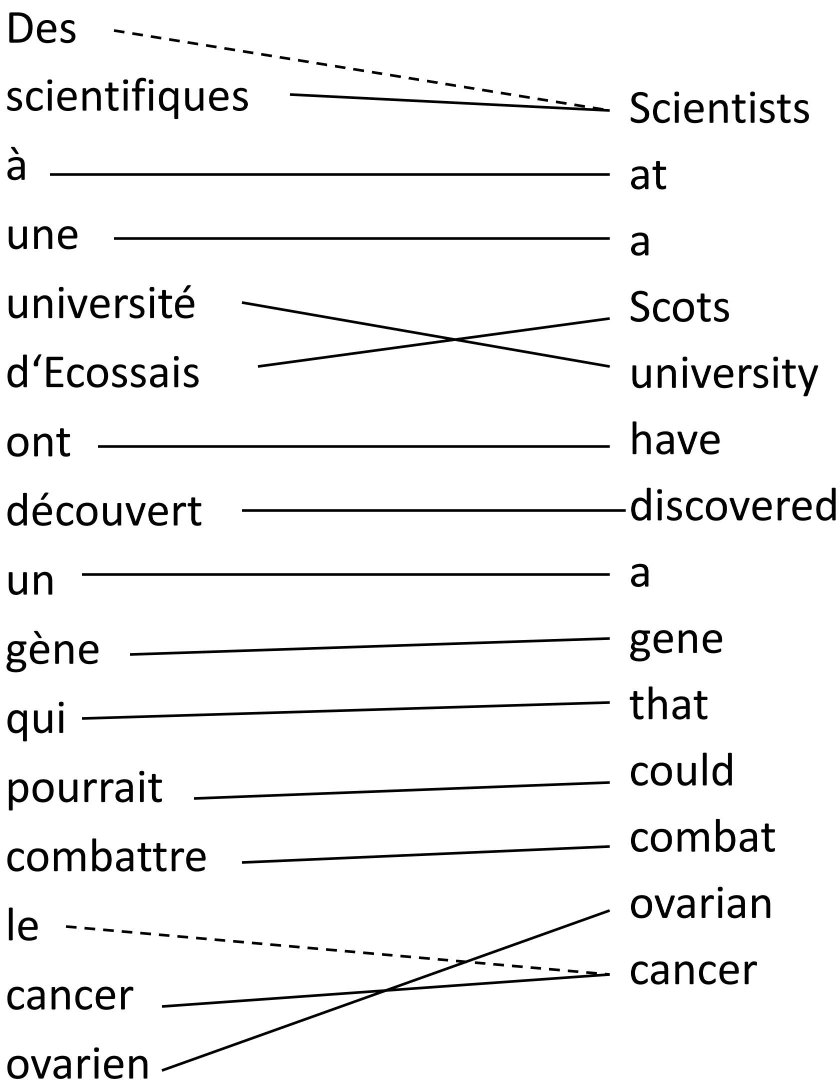
How hard can it be to align English-English?

Scientists
at
a
Scots
university
have
discovered
a
gene
that
could
combat
ovarian
cancer

How hard can it be to align English-English?

Des	
scientifiques	Scientists
à	at
une	a
université	Scots
d'Ecossais	university
ont	have
découvert	discovered
un	a
gène	gene
qui	that
pourrait	could
combattre	combat
le	ovarian
cancer	cancer
ovarien	

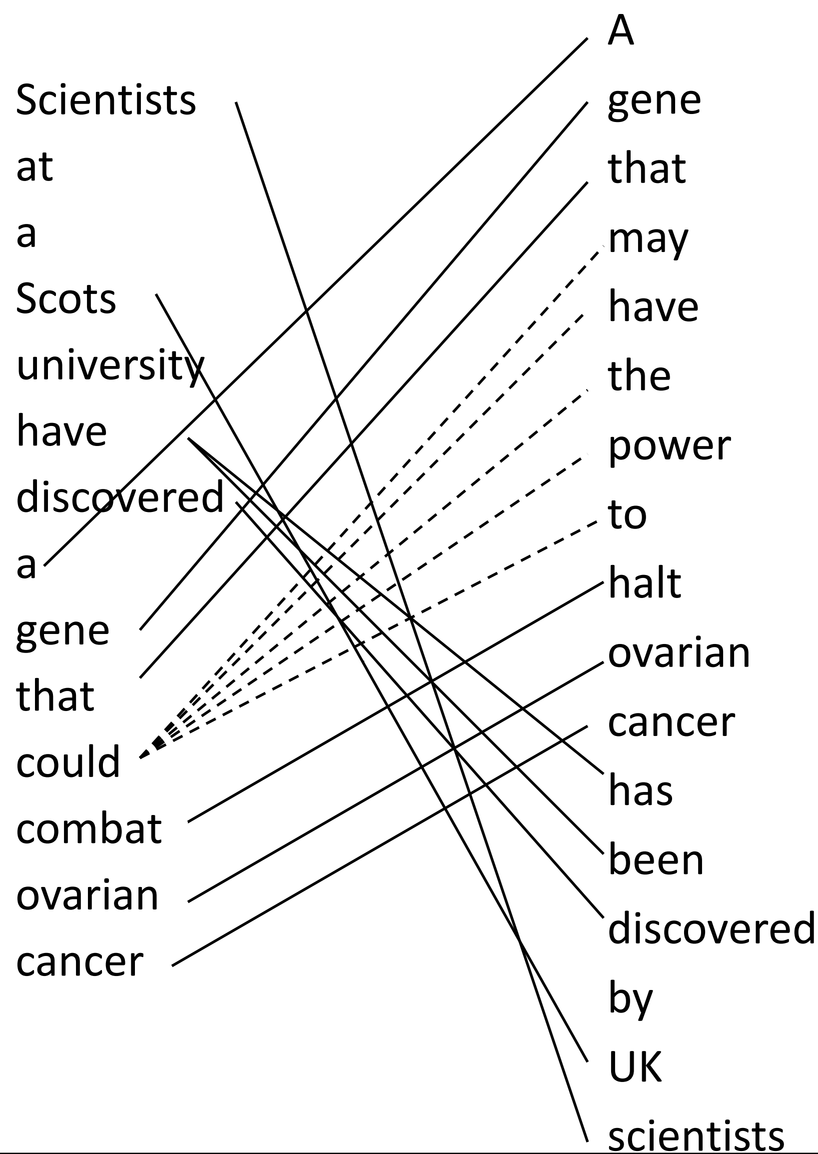
How hard can it be to align English-English?



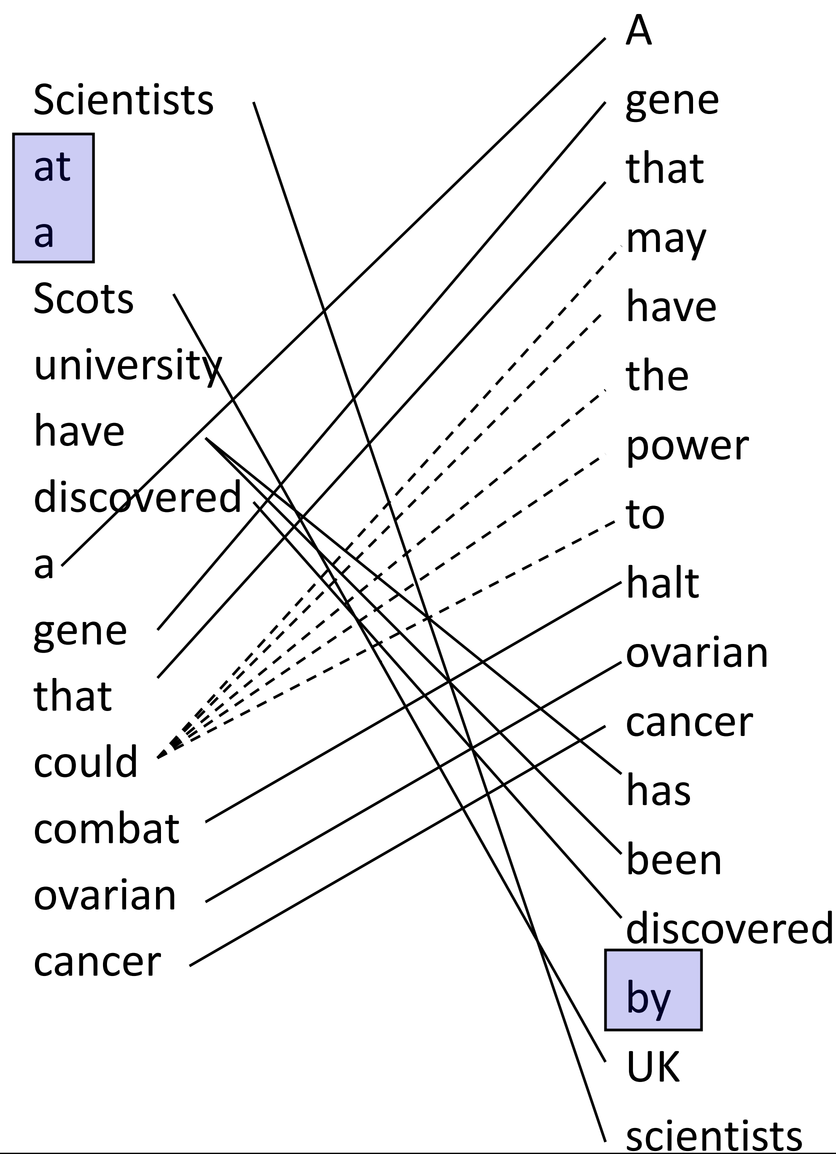
How hard can it be to align English-English?

	A
Scientists	gene
at	that
a	may
Scots	have
university	the
have	power
discovered	to
a	halt
gene	ovarian
that	cancer
could	has
combat	been
ovarian	discovered
cancer	by
	UK
	scientists

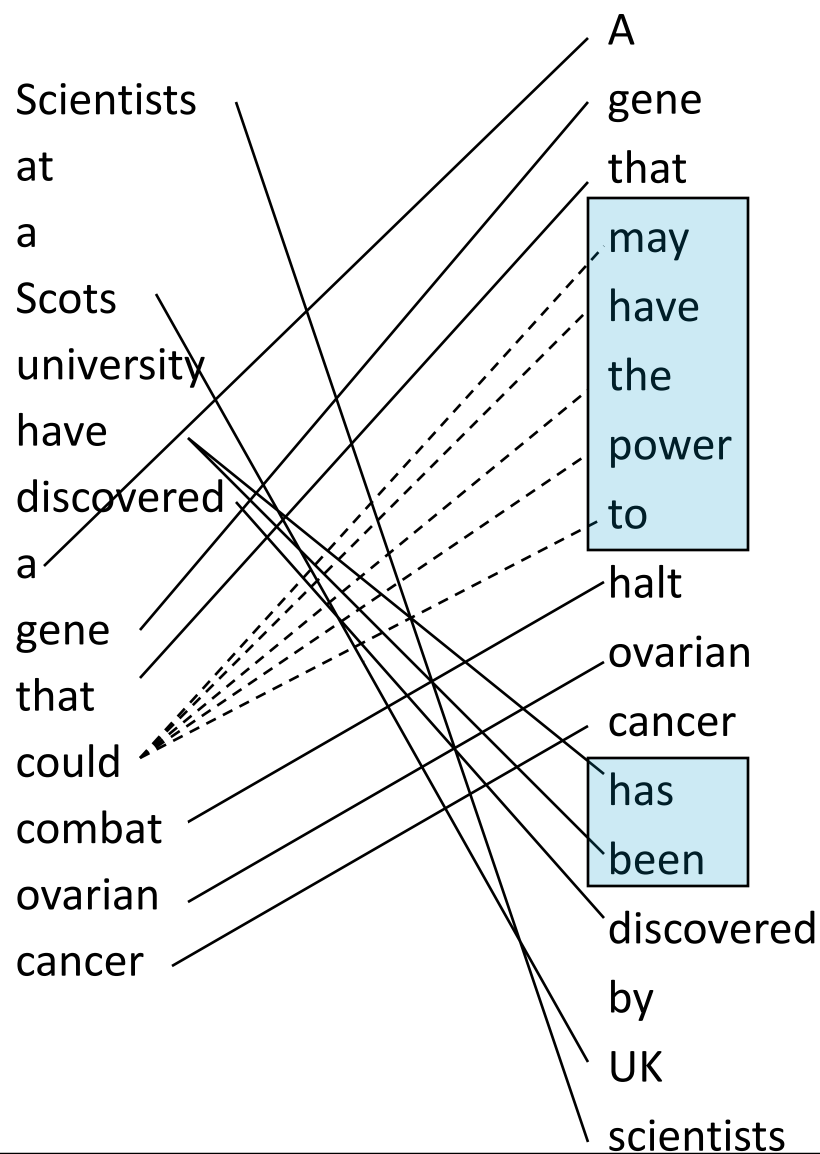
How hard can it be to align English-English?



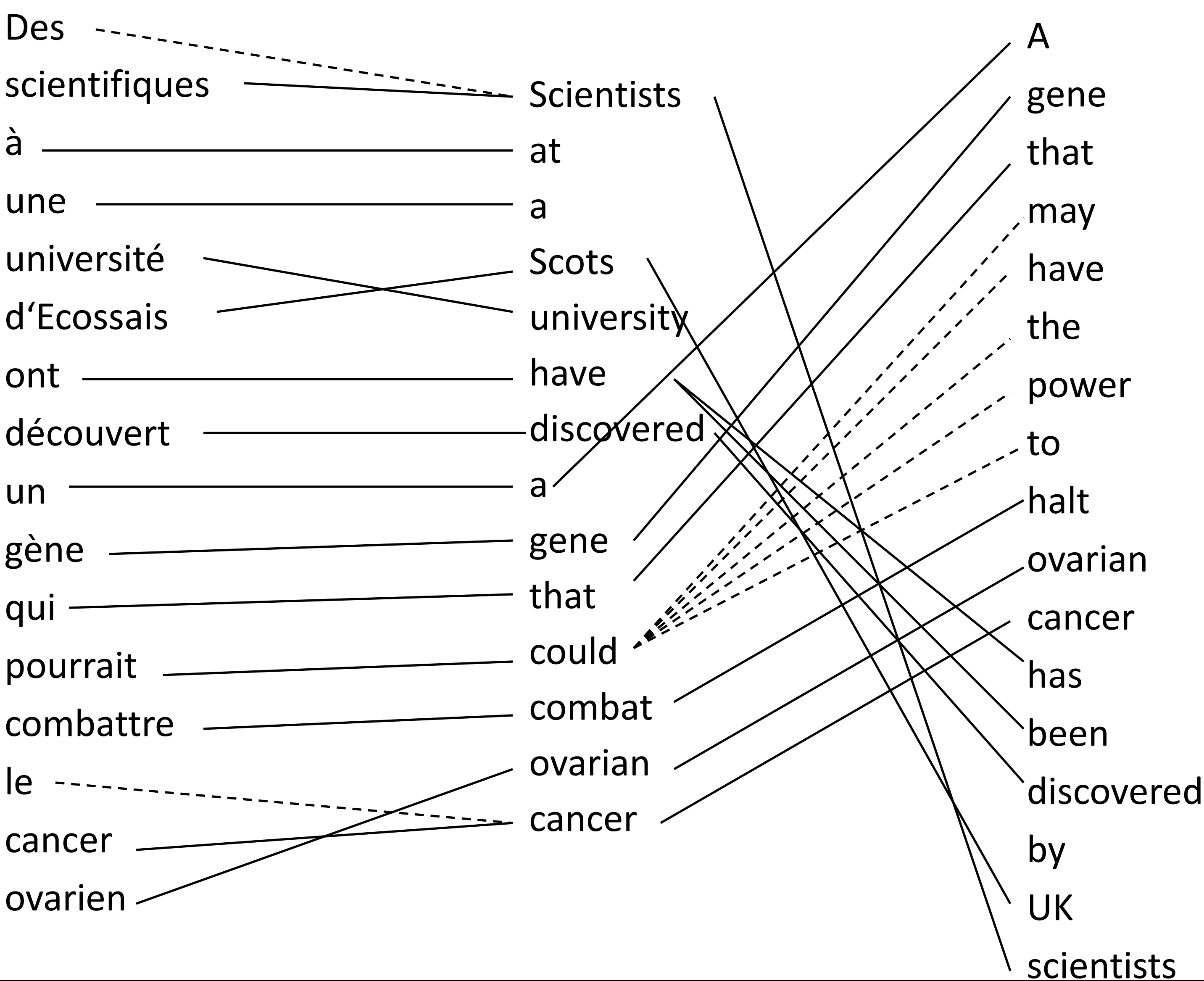
How hard can it be to align English-English?



How hard can it be to align English-English?



How hard can it be to align English-English?



Alternations are Translation-like

Elaboration:
- the NASDAQ / the tech-heavy NASDAQ

Phrasal:
- has pulled the plug on / is dropping plans for
- more than a million people / a massive crowd

Synonymy:
- charges / accusations

Anaphora:
- Prime Minister Blair / He

Reordering:
- topicalization, voice alternations
- adverb or prepositional phrase placement

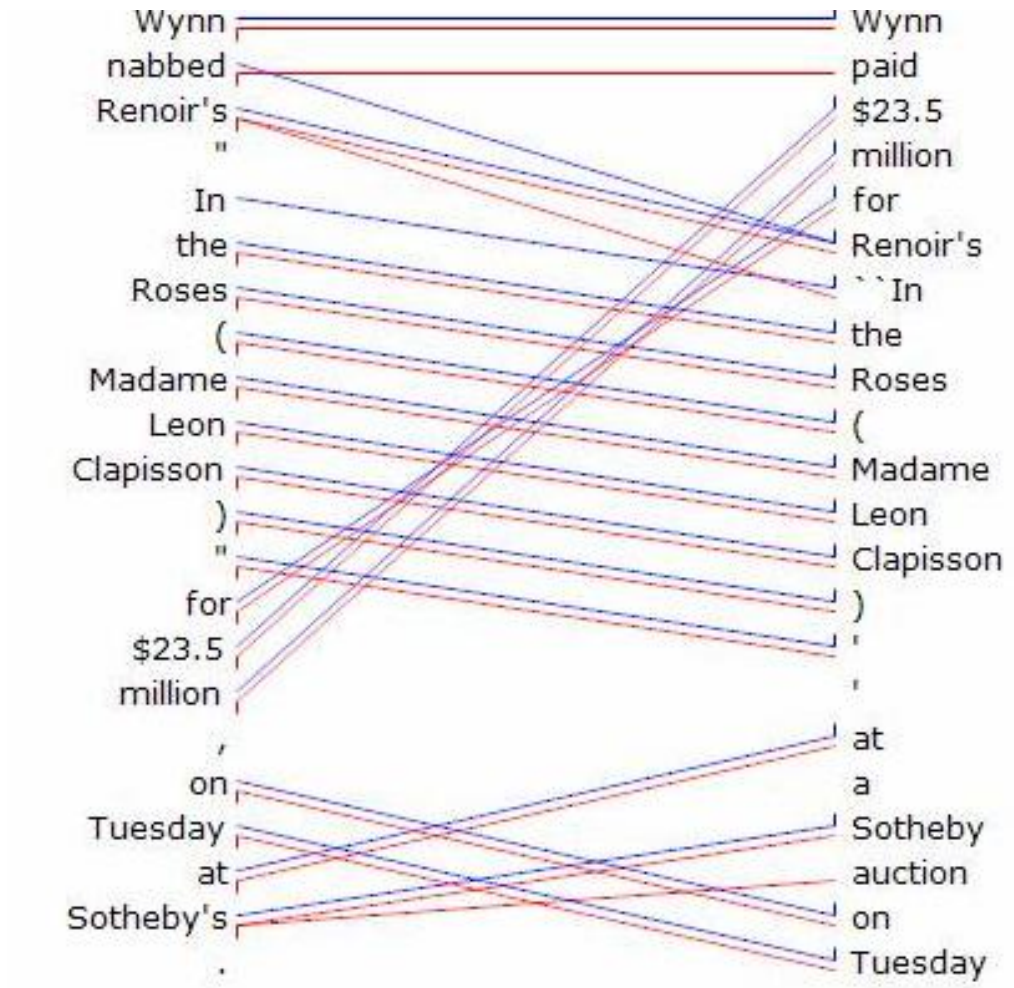
Quantifying how hard it is

- Rely on established metrics from MT
- Word align monolingual data using Giza++
- Compare using standard word alignment metric
  - Word Alignment Error Rate (Och & Ney, 2003)
  - English-French (Hansards): <5%
  - English-English sentences: 11%+ AER
    - Even though pairs selected by Levenshtein distance <12
    - AER for non-identical words ranges from 20.9%-47.5%
- Poor alignments = models less useful for
  - Measuring similarity between strings
  - Generating alternative phrasings (decoding)

Building an SMT paraphrase system (Quirk et al, 2004)

- Gather parallel, monolingual corpus of sentences
  - Currently 150K sentence pairs
  - (more on this later)
- Train models using standard statistical MT engine
  - Word alignment of parallel corpus
  - Parameters for translation models
- Apply models:
  - Align new sentence pairs to recognize paraphrase
  - Use MT decoder as statistical rewriting component

Example Giza++ Alignment



Example: paraphrase generation

Input string	the	white	house announced	a new	policy
-32.381603 [tgt:-20.721785, trans:-44.041422]	the	white	house announced	[the new]	policy
-33.183348 [tgt:-19.335800, trans:-47.030896]	the	white	house said	[the new]	policy
-33.514222 [tgt:-24.324001, trans:-42.704443]	the	[bush white]	house announced	[the new]	policy
-33.541390 [tgt:-26.930034, trans:-40.152746]	[that the]	white	[house said]	[the new]	policy
-33.897980 [tgt:-24.951244, trans:-42.844716]	the	[bush white]	house [unveiled a]	new	policy
-34.214541 [tgt:-22.851248, trans:-45.577834]	the	white	house [revealed a]	new	policy
-34.314196 [tgt:-22.938017, trans:-45.690375]	the	[bush white]	house said	[the new]	policy
-34.388251 [tgt:-29.998310, trans:-38.778191]	[that the]	[bush white]	[house said]	[the new]	policy
-34.665697 [tgt:-21.851742, trans:-47.479653]	[the bush administration]		announced	[the new]	policy
-34.684553 [tgt:-24.365965, trans:-45.003140]	his	white	[house said]	[the new]	policy
-34.813016 [tgt:-27.914257, trans:-41.711774]	a	[bush white]	[house said]	[the new]	policy
-34.916360 [tgt:-25.849709, trans:-43.983011]	this	white	[house said]	[the new]	policy
-34.930093 [tgt:-23.155255, trans:-46.704931]	the	white	house unveiled	[the new]	policy
-34.950866 [tgt:-21.547445, trans:-48.354287]	the	white	house announced	a new	policy
-34.999698 [tgt:-19.530269, trans:-50.469127]	[the bush administration]		said	[the new]	policy
-35.010633 [tgt:-22.823760, trans:-47.197506]	the	white	house [announced its]	new	policy
-35.074486 [tgt:-22.969377, trans:-47.179594]	the	white	house [offered a]	new	policy
-35.313121 [tgt:-23.706620, trans:-46.919622]	a	white	house announced	[the new]	policy
-35.339204 [tgt:-21.823518, trans:-48.854890]	the	white	house announced	[the new]	regulations

- Monotone Decoder (i.e., no attempt to reorder words/phrases)
- Trained on a small corpus of parallel sentences from news articles

Decoder Demo

Recent Related Work

Barzilay & Lee (2003)

- Learn sentence-level templatic patterns for paraphrase generation
  - Method exploits meta-information implicit in dual collections of newswire articles
  - Multi-sequence alignment identifies sentences sharing formal (and presumably semantic) properties
    - Result: word lattices capturing n-gram-based structural similarities between sentences
    - Lattices are mapped to templates that can be used to produce novel transforms of input sentences

Barzilay & Lee (2003)

- Generation results can be striking
  - Complex reorderings, phrasal replacements

latest violence bring to NUM1 number of people killed as a direct result of the palestinian, including NUM2 palestinians and NUM3 israelis.

At least NUM1 palestinians and NUM2 israelis have been killed since palestinian uprising against israeli occupation began in September 2000 after peace talks stalled.

- But sentence-level templates mean technique has limited generality
  - Requires domain with highly repetitive sentence structures
  - Since pairings involve whole sentences, semantic information can be lost/gained randomly

bomb on X1 kills NUM1 israelis

police spokeswoman said NUM1 people were killed by bomber at X1 attack on

Shinyama & Sekine 2003

- Goal: find paraphrase alternation patterns from Japanese news articles for Information Extraction (IE)

  NUMBER people have died in LOCATION
  LOCATION reported NUMBER deaths

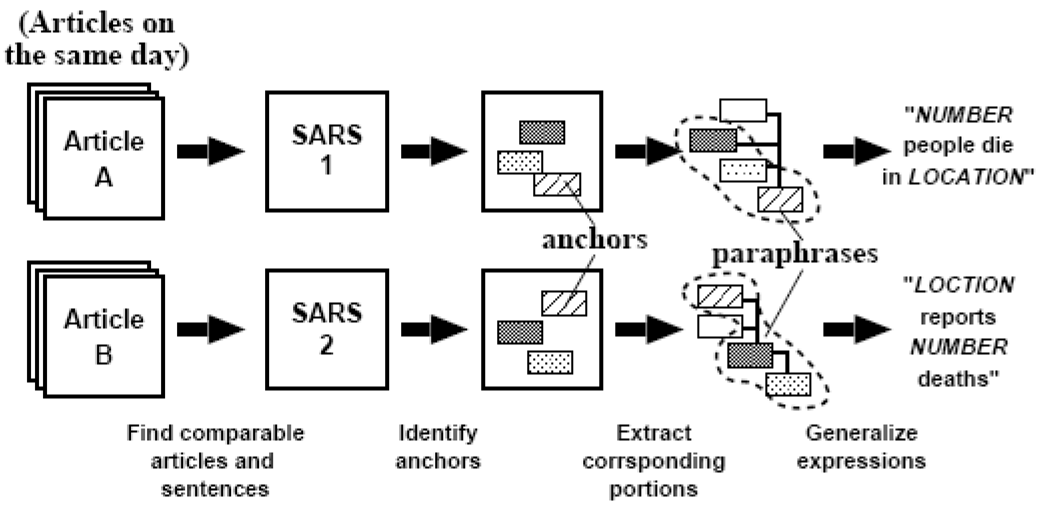
  PERSON1 killed PERSON2
  PERSON1 let PERSON2 die from loss of blood

  PERSON1 shadowed PERSON2
  PERSON1 kept his eyes on PERSON2

- Method:
  - Comparable articles from different news sources
  - Data annotated with dependency parses
  - Named entities serve as “anchors”, as in DIRT, assume material connecting them is equivalent in meaning

- Accuracy: 62% in small domain, vs. 100% human

Shinyama & Sekine (2003)



The government has announced that two more people have died in Hong Kong after contracting the SARS virus and 61 new cases of the illness have been detected.

Hong Kong reported two more deaths and 61 fresh cases of SARS Friday as governments across the world took tough steps to stop the killer virus at their orders.

Outline

- Motivation
- Textual Entailment
  - Overview of problem, literature
  - PASCAL: Recognizing Textual Entailment Challenges
- Paraphrase Recognition and Generation
  - Overview of problem, literature
  - MSR approach
- Corpus issues
- (time permitting): a work in progress

Corpus Issues

- Where will the data come from?
  - This is the major problem holding back the field
- Work in SMT has been fueled by availability of easily-aligned data
  - Professional translators
  - Typically closely related languages
- No data like this for monolingual alternations
  - What’s available is comparable, not strictly parallel
  - Equivalent news articles, translations of novels, etc.
  - Not enough of it!

An Assertion

- If we had a parallel monolingual corpus on the scale of the Hansards or EuroParl….
  - SMT would be the default approach to semantic modeling
  - For both recognition and generation
- Instead, we're seeing…
  - Templatic generation strategies
  - Task-specific heuristics
  - Purpose-built lexicons
- Remember the 1980's! (and 1990's…)

From MT, we know that these methods do not scale

Where's the monolingual Hansards?

- There isn't one
- No monolingual role comparable to a professional translator
  - Editors and teachers edit sentences, but before/after typically not saved, mixture of linguistics/content edits

Johnson, who has an extensive criminal record, lied repeatedly during his testimony to the congressional panel

Mr. Johnson was not entirely forthcoming in his comments to the panel

- People rewrite text to fix a problem
- For translation, the "problem" is obvious: the string is in the wrong language
- For paraphrasing, the "problem" is less clear
  - Native speakers have trouble with the task unless they can see a problem with the original sentence
  - Can't just pay an army to create the dataset

Even worse!

- SMT approach requires a LOT of data
- In MT data, every sentence pair is information-rich
  - Mappings to different lexical items
  - Translation probabilities in different contexts
- But paraphrase alternations may contain little information
  - Most often, words map to themselves, regardless of context

    Migratory birds returned to the lake for the first time in 4 years.

    Migratory birds came back to the lake for the first time in 4 years.
- Information-poor mappings = need for even more data
  - Must acquire data from comparable sources
  - A challenge for SMT, must be faced earlier in monolingual domain
- Unlike e.g. English-French MT, focus is on the hard cases

So where will we get a monolingual Hansards?

- We’re going to have to build it
  - Scraped from the Web
    - Improved extraction from comparable corpora (Munteanu & Marcu, 2005)?
    - Better use of parallel news data?
    - Some new breakthrough??
- This may be the most important research agenda for empirical modeling of semantic overlap
  - Same problem faces broad-domain MT
  - We shouldn’t let the lack of a corpus push us into short-term solutions

One promising source: clustered news data

- Explosion of online news sources:
  - Over 6,000 sites indexed by MSN NewsBot
  - Over 4,500 sites indexed by Google News
- Mountains of paraphrase data:
  - Multiple journalists writing on the same events
  - Multiple editors editing the same copy
  - Automatic topic clustering by time and event
- MSR has released a paraphrase corpus gathered from clusters of online news articles

HOME
Newsbot
Popular Articles
World
US
Business
Sports
Entertainment
Science/Health
Technology
News
Business
Sports
Tech / Science
Entertainment
Health
Travel
Multimedia
Opinions
Weather
Local News
Newsweek
Today Show
Nightly News
Dateline NBC
Meet the Press
MSNBC TV
News Video
MSNBC Shopping
Newsbot

Microchips implanted in Mexican officials

MSNBC - 20 hours ago • Popularity Rank: 2725 • Similar Stories: 5

MEXICO CITY - Security has reached the subcutaneous level for Mexico's attorney general and at least 160 people in his office -- they have been implanted with microchips that get them...

Reuters/ DANIEL AGUILAR

MEXICAN ATTORNEY GENERAL MACEDO SPEAKS AT A PRESS CONFERENCE
Mexican attorney general Rafael Macedo de la Concha speaks at a news conference on June 24, 2004 in Mexico City, where Federal police display weapons confiscated during a raid on suspected drug traffickers. Macedo said that in a Thursday morning raid in Tijuana, police netted several suspected members of the Arellano Felix cartel, including alleged chief enforcer Mario Alberto Rivera. REUTERS/Daniel Aguilar

Mexican attorney general personally goes high-tech for security

CNEWS - 18 hours ago

MEXICO CITY (AP) - Security has reached the subcutaneous level for Mexico's attorney general and at least 160 people in his office. They have been implanted with microchips that get them...

Chip Implanted in Arm of Mexican Judicial Workers for Access to Secure Areas of Headquarters

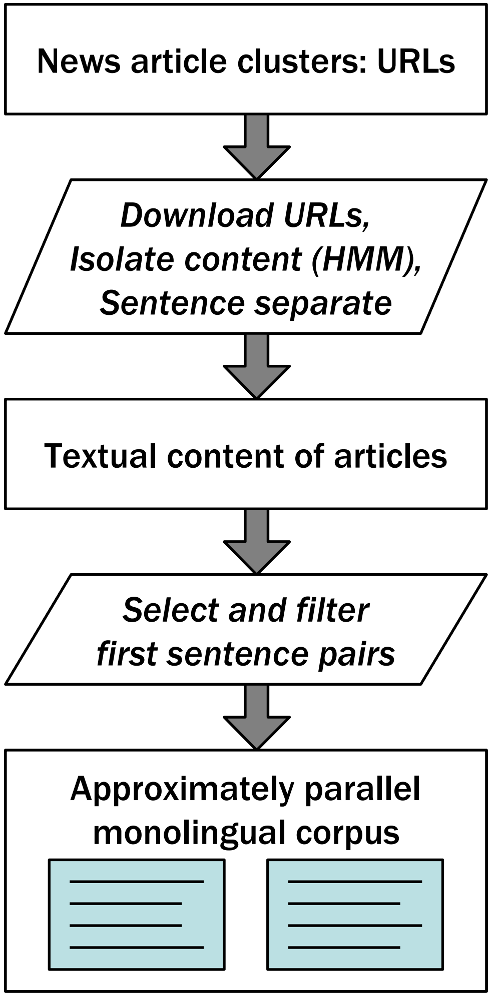
Xposed - 21 hours ago

Security has reached the subcutaneous level for Mexico's attorney general and at least 160 people in his office - they have been implanted with microchips that get them access to secure...

Chip Implanted in Mexico Judicial Workers

Data extraction

Data Collection



Over 8 months:

- 11,162 clusters
- 177,095 articles
- 15.8 articles per cluster
- 3.5m sentences

MSR Paraphrase Corpus

- Annotated corpus of naturally-occurring paraphrase and non-paraphrase general domain examples (Dolan et al. 2004)

  http://research.microsoft.com/research/nlp/msr_paraphrase.htm

  - 5801 pairs of sentences from news articles
    - Distilled from 13 million sentence pairs in 32K clustered articles
    - Word-based Levenshtein Distance ≥ 8 edits (bag of words) over sentences from news clusters
  - Annotated by multiple human judges
    - 3900 (67%) judged "semantically equivalent"
    - 1901 (33%) partial or non-equivalent
  - In "yes" cases, semantic content mostly – but not entirely -- parallel

El Watan, an Algerian newspaper, reported that the kidnappers fiercely resisted the army assault this morning, firing Kalashnikov rifles.
El Watan, an Algerian newspaper, reported that the kidnappers put up fierce resistance during the army assault firing Kalashnikov rifles.

Mr. Concannon had been doused in petrol, set himself alight and jumped onto a bike to leap eight metres onto a mattress below.

A Sydney man suffered serious burns after setting himself alight before attempting to jump a BMX bike off a toilet block into a pile of mattresses , police said.

Another Filtering Technique

- Exploit a discourse convention of news genre
  - First sentence or two of a news article typically summarize its content

<u>**Policeman rips hole in royal oil painting**</u>
Independent, UK - 3 hours ago
By Chris Bunting. A policeman who was determined to protect a royal oil painting from thieves accidentally tore a gaping hole in its canvas. **...**

Assumption: within the same topic cluster, initial sentences contain the same content

Second Corpus: first 2 Sentences

- Consider all sentence pairs from the first 2 sentences in each article within a given cluster; filter with the following criteria:
  - Max length ratio: 1/2
  - Must share at least 3 words of 4+ characters

First 2 Sentences: very rich data

The genome of the fungal pathogen that causes Sudden Oak Death has been sequenced by US scientists

Scientists have figured out the complete genetic code of a virulent pathogen that has killed tens of thousands of California native oaks

Researchers announced Thursday they've completed the genetic blueprint of the blight-causing culprit responsible for sudden oak death

The East Bay-based Joint Genome Institute said Thursday it has unraveled the genetic blueprint for the diseases that cause the sudden death of oak trees

Mr. Concannon had been doused in petrol, set himself alight and jumped onto a bike to leap eight metres onto a mattress below.

A Sydney man suffered serious burns after setting himself alight before attempting to jump a BMX bike off a toilet block into a pile of mattresses , police said.

But "rich" is not always more useful

- Pairs gathered with this technique can exhibit much more interesting alternations than those gathered with edit distance strategies
  - Little lexical overlap
  - Complex syntactic alternations
- Ironically, though, this presents problems for SMT techniques
  - Word and phrase alignment work best with simple word-for-word alternations (*English-French*, not *English-Japanese*)
  - In our experiments, pairs derived through string-edit distance metric produced more useful data for SMT models
  - Need to wait for SMT techniques to catch up!
- Interesting to speculate on where SMT would be now if the only existing parallel corpora were in English-Japanese!

Another promising use of parallel news data

- LCC's Groundhog System (Hickl et al, 2006)
  - Winner (by a wide margin) at RTE2: 75% accuracy
- Decision tree classifier, with features from:
  - Alignments (e.g. longest common substring)
  - Dependency matches based on semantic parser
  - Heuristic features (e.g. polarity mismatch between predicates)
  - Paraphrase mappings
- Main contributor to success: paraphrase information
  - Input to classifier: model trained on large corpus of MT-style parallel mappings
  - Mappings learned from text corpora
  - Paraphrase model, used alone, correct 66% of the time
    - (cf. 60 % average for systems at RTE2, 60% baseline for simple lexical overlap method)
  - (Indirect) evidence that RTE dataset is heavily skewed toward bidirectional entailment examples

Paraphrase Acquisition in Groundhog

- Paraphrase model trained on 200K additional entailment pairs scraped from the web
  - Data drawn from newswire articles on the web
  - 10K hand-labeled examples used to bootstrap data collection
  - Positive and negative examples used to train classifier
  - Sentence pairs parsed with a chunk parser
    - Chunks aligned with using a classifier trained on hand-annotated good/bad pairs
    - Result: set of aligned chunk pairs

Previous work on data collection: Tradeoff: Quality vs. Quantity

- Structured corpora (e.g. parallel news articles, multiple translations of same original)
  - Barzilay & Lee (2001)
  - Identify anchor points in comparable texts, assume that what intervenes is similar in meaning (Shinyama et al 2002)
  - Generate whole sentence paraphrases (Barzilay & Lee 2003)
  - Paraphrase as phrasal translation (Quirk et al 2004, Dolan et al 2004; Burger & Ferro, 2005; Hickl et al , 2006)
- Unstructured corpora
  - Clustering similar words (e.g. Pereira 1993, Lin 1998)
  - Anchor-based approach using free text (Lin & Pantel, 2001)

Existing Paraphrase Corpora

- Multiple translations in a second language
  - ATR English-Chinese paraphrase corpus (Zhang & Yamamoto, 2002),
  - ATR Japanese-English corpus (Shirai, et al., 2002)
  - LDC Multiple-Translation Chinese Corpus (Huang et al., 2002)
  - Multiple translations of novels (e.g., Barzilay & McKeown, 2001)
- Barzilay & Lee, 2003:

http://www.cs.cornell.edu/Info/Projects/NLP/statpar.html

Encarta QA Dataset

- If you could talk to an encyclopedia, what would you ask?
  - Gathered hundreds of questions from school children

Q: *If clouds are made of water how come they don't fall?*

- Target corpus: Encarta 98
  - Relatively small, closed corpus
  - Exhaustive recall set annotated by hand
  - Professional research librarian
  - 10K QA pairs

- Paraphrase/Entailment problems everywhere

Q: *If clouds are made of water how come they don't fall?*

A: *Cloud particles range in size from about 5 to 75 micrometers (0.0005 to 0.008 cm/0.0002 to 0.003 in). The particles are so small that light, vertical currents easily sustain them in the air.*

To conclude…

What applications would paraphrase/entailment modules enable?

- Very few!
  - Recognition: redundancy detection for proofing
  - Paraphrase generation: “Rewrite this” in word processor
- But recognition capability would play a key role for many applications, e.g.
  - key feature in a classifier deciding whether a candidate text is a good answer in a QA system
  - echo-back capability for dialog system

- Important to distinguish between “solving” these hard, long-standing application problems vs. informing a higher-level program with crucial information

To summarize…

- What NL functionality do applications need?
  - Not a word sense, a cluster of related words, a parse tree, etc.
  - Instead, the ability to:
    - Empirically measure when two strings overlap in a particular context
    - Decide whether overlap is unidirectional vs. bidirectional
    - Generate a meaning-preserving paraphrase of an input string
- I’ve tried to convince you that these abilities are at the heart of many – all? -- tasks requiring NLU
  - The RTE Challenges are a first attempt to focus the research community on a common, empirical task
  - Need to better understand the nature of the inferences needed and how traditional levels of linguistic analysis can be used to help make these inferences
  - Need much more data!

Outline

- Motivation
- Textual Entailment
  - Overview of problem, literature
  - PASCAL: Recognizing Textual Entailment Challenges
- Paraphrase Recognition and Generation
  - Overview of problem, literature
  - MSR approach
- Corpus issues
  - (time permitting): a work in progress

Semantic Structure as an Emergent Property of the Web

Parallel Monoligual Data is Scarce

- Maybe SMT will develop better techniques for mining parallel chunks from comparable corpora
  - Ideally, we'll be able to adopt the same techniques
- But work on this problem …
  - Is still at a very early stage
  - Tends to discover simple, word-for-word translations (reflects seeding with bilingual dictionary)
    - Not so useful for monolingual data, since words = themselves, as do extracted phrases and sentences
- Burger & Ferro's technique is interesting, but still limited to structured data
  - Only news domain
- We have to keep looking for other sources

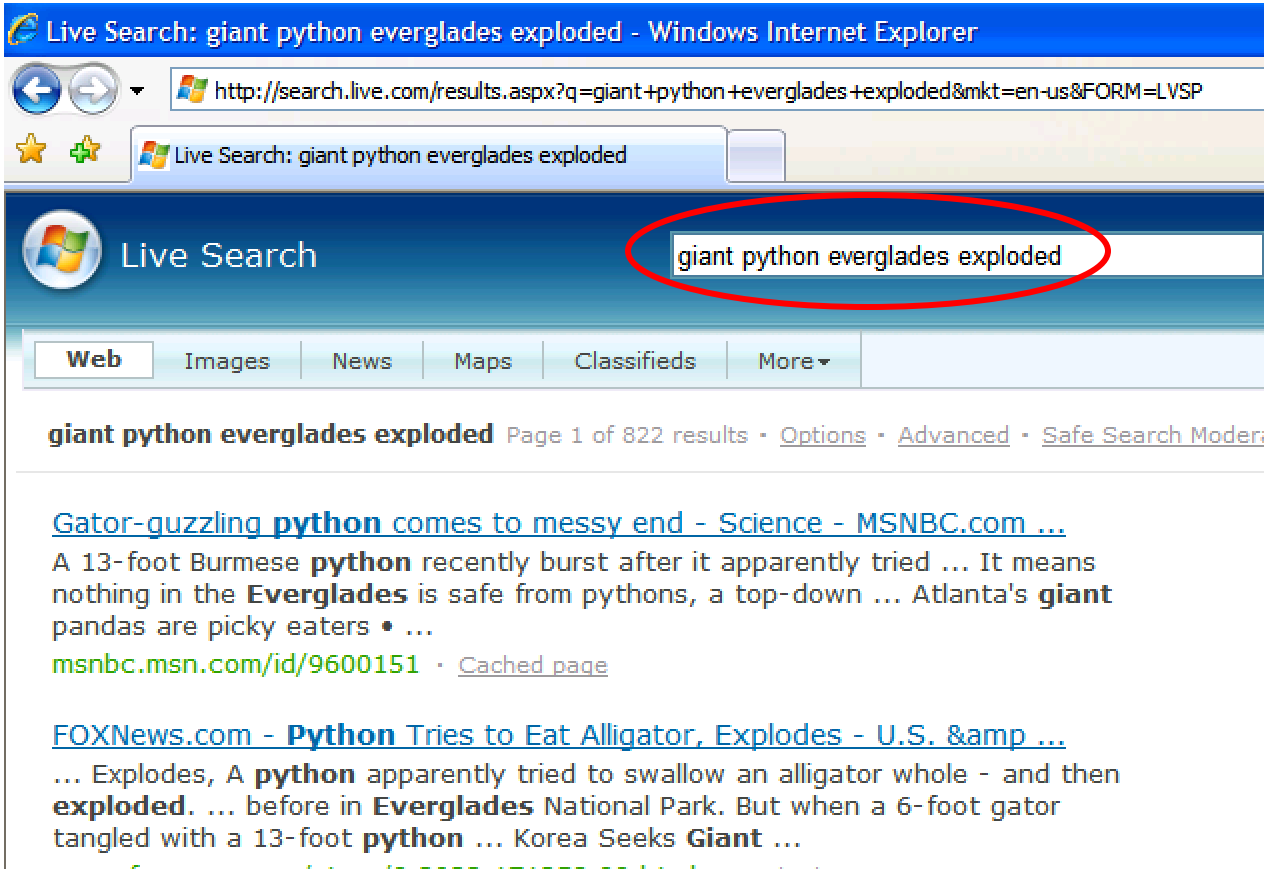
An Impasse

- Statistical MT approach is well-motivated, but…
  - Not enough parallel monolingual data in the world
  - Paraphrases are more difficult to elicit than translations, so no data fire hose
- We need an approach to learning semantic similarity that doesn't require specialized data
  - Ideally, ordinary, flat, monolingual text

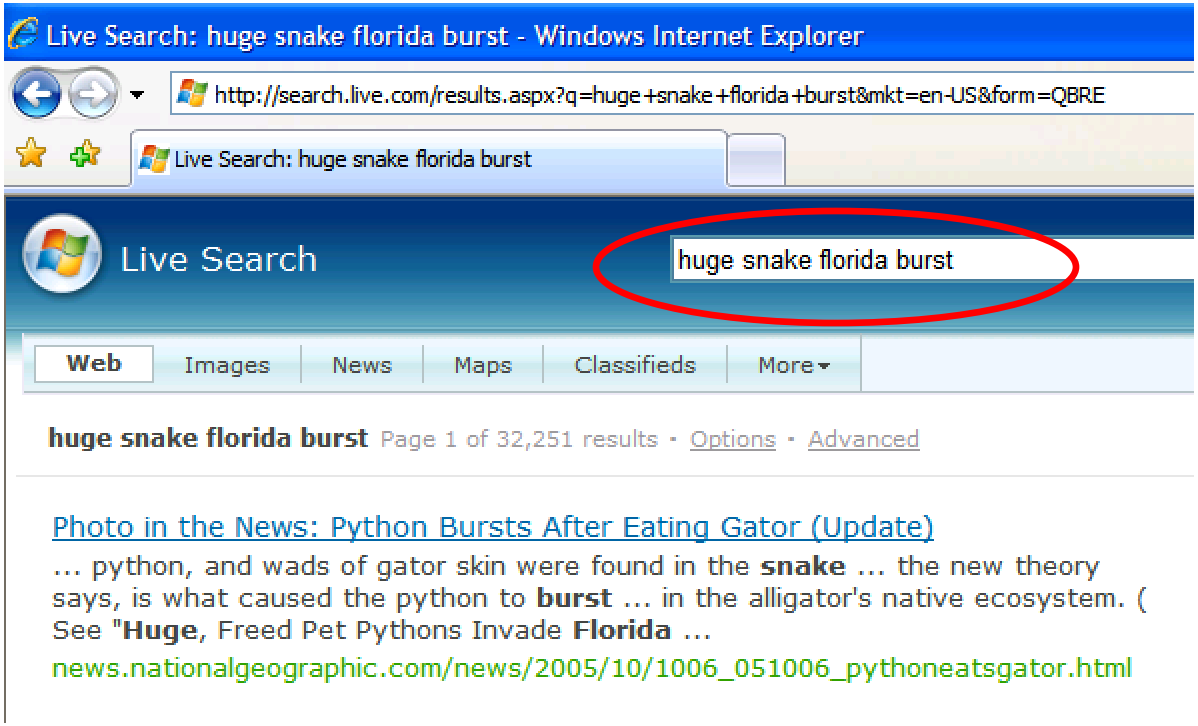
Turn to the Web

- Because that's where the paraphrases are
  - (paraphrasing Willie Sutton)
- Scale of the web changes the rules
  - The information we need is out there
  - We just have to figure out how to mine it
- Search engines work because the paraphrase problem has been solved by brute force
  - The sheer volume of text out there means that the same information is replicated in different forms
  - Any plausible choice of keywords is likely to yield something relevant

Paraphrases on the Web



Paraphrases on the Web (2)



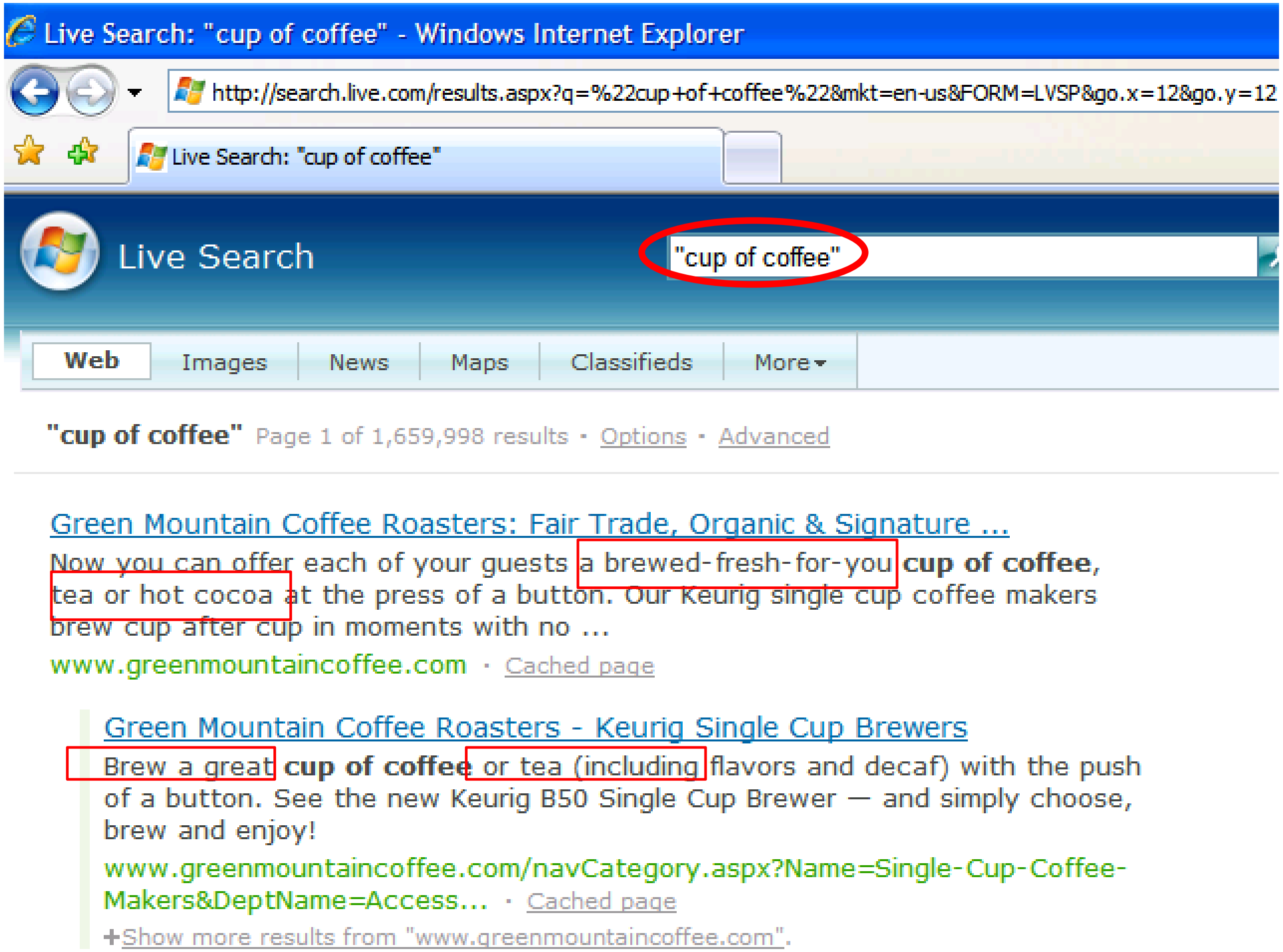
The alternations are out there

- But a keyword-based index doesn't reveal them
- How do we cluster sets of words/phrases with the same meaning?

Can Dumb^n= Intelligence?

- A very dumb approach, implemented in an inefficient way
  - Many thousands of simple Web queries
- Algorithm
  - Query the web with a quoted, arbitrarily-chosen ngram
  - Examine each snippet in top 30
    - Collect sets of left- and right-contexts of length *n*
    - Launch a new query for each context
    - Rinse, lather, repeat… endlessly
  - Post-process the assembled contexts to identify
    - Constituent boundaries
    - Semantic relationships

Query: "cup of coffee"



Collected Contexts

- cup_of_coffee.out
  - E.g. “contains a”, “as outlined by”, “starts with fresh”, “was to measure”, “is to start”, “and freshly baked”, “is easy if you”, “instructions”

- bubble_gum_flavor.out

See any patterns?

Now, start looking for generalizations

- gum.docx
- SyrupInPlaceOf.txt

Constituent Bracketing Emerges

- Simple heuristic: look for shared contexts, use to posit syntactic boundaries
  - bubble_gum.docx (left edge)
  - bubble_gum_flavor.docx (right edge)
- Counts suggestive, though not true probabilities
- Function words, light verbs, coordination, etc. quickly pop
- Contexts provide a wealth of features for clustering techniques
- As the same ngrams are examined in different contexts, corroborating evidence aggregates

Problems with the Distributional Hypothesis

- Harris, 1954
  - Words (and phrases) that occur in the same context tend to have similar meanings
  - Standard approach to calculating lexical similarity
- But distributional methods confound various dimensions
  - Syntactic class membership
    - e.g. adverbs “in a mean way, greedily, accidentally”
  - Semantic class membership
    - “New York”, “Boston”, “Los Angeles”
  - Semantic similarity
    - “without meaning to”, “by accident”
    - “Elizabeth II”, “the queen of England”
- Result
  - Noisy data that isn’t very useful for applications

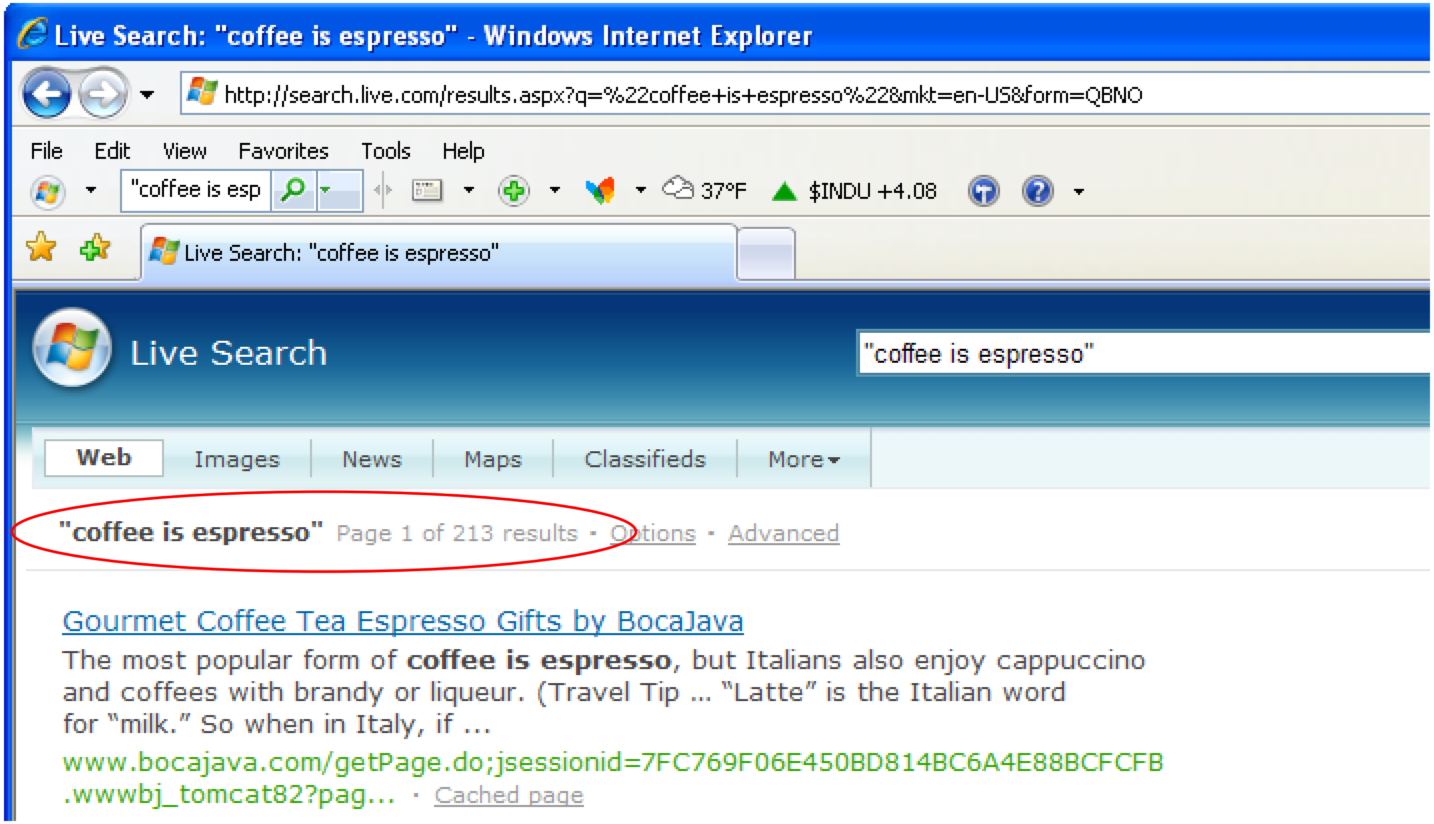
An ideal algorithm would....

- Rely on the Distributional Hypothesis
  - Avoid rare/specialized data
- Isolate cases that capture true semantic similarity
  - Not just distributional similarity
- Be unsupervised and knowledge-poor

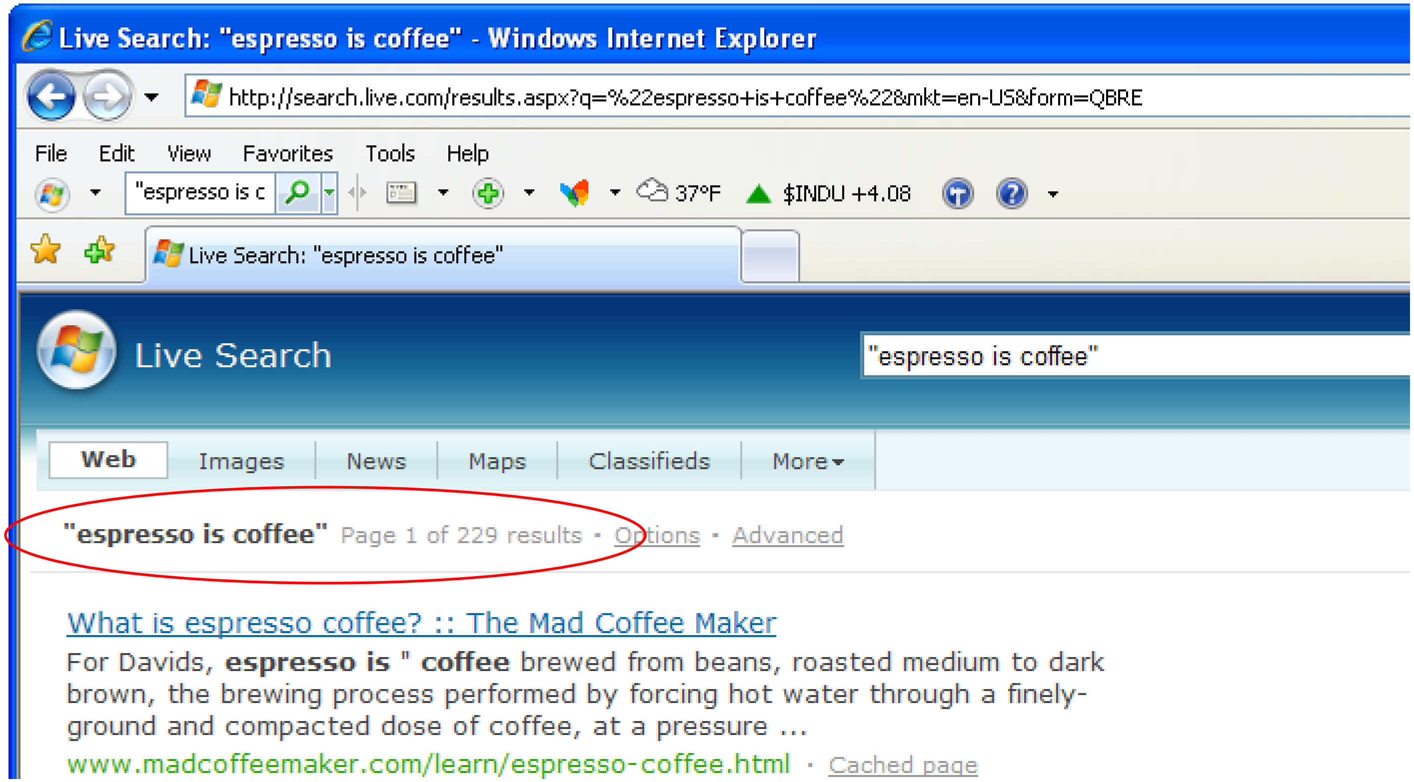
Isolating True Paraphrases

- How can we isolate alternations that reflect semantic similarity?
  - Ask the web to confirm or deny
  - More often than not, the exact answer is there
  - If there’s no evidence anywhere, then it’s probably false
- Templatic queries
  - “X is a Y”, “Y is a X”, “X and Y are both”, “X’s and Y’s are both”

Examples


Live Search: "coffee is espresso" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22coffee+is+espresso%22&mkt=en-US&form=QBNO
File Edit View Favorites Tools Help
"coffee is esp
37°F
$INDU +4.08
Live Search: "coffee is espresso"
Live Search
"coffee is espresso"
Web Images News Maps Classifieds More
"coffee is espresso" Page 1 of 213 results · Options · Advanced
Gourmet Coffee Tea Espresso Gifts by BocaJava
The most popular form of coffee is espresso, but Italians also enjoy cappuccino and coffees with brandy or liqueur. (Travel Tip ... "Latte" is the Italian word for "milk." So when in Italy, if ...
www.bocajava.com/getPage.do;jsessionid=7FC769F06E450BD814BC6A4E88BCFCFB
.wwwbj_tomcat82?pag... · Cached page

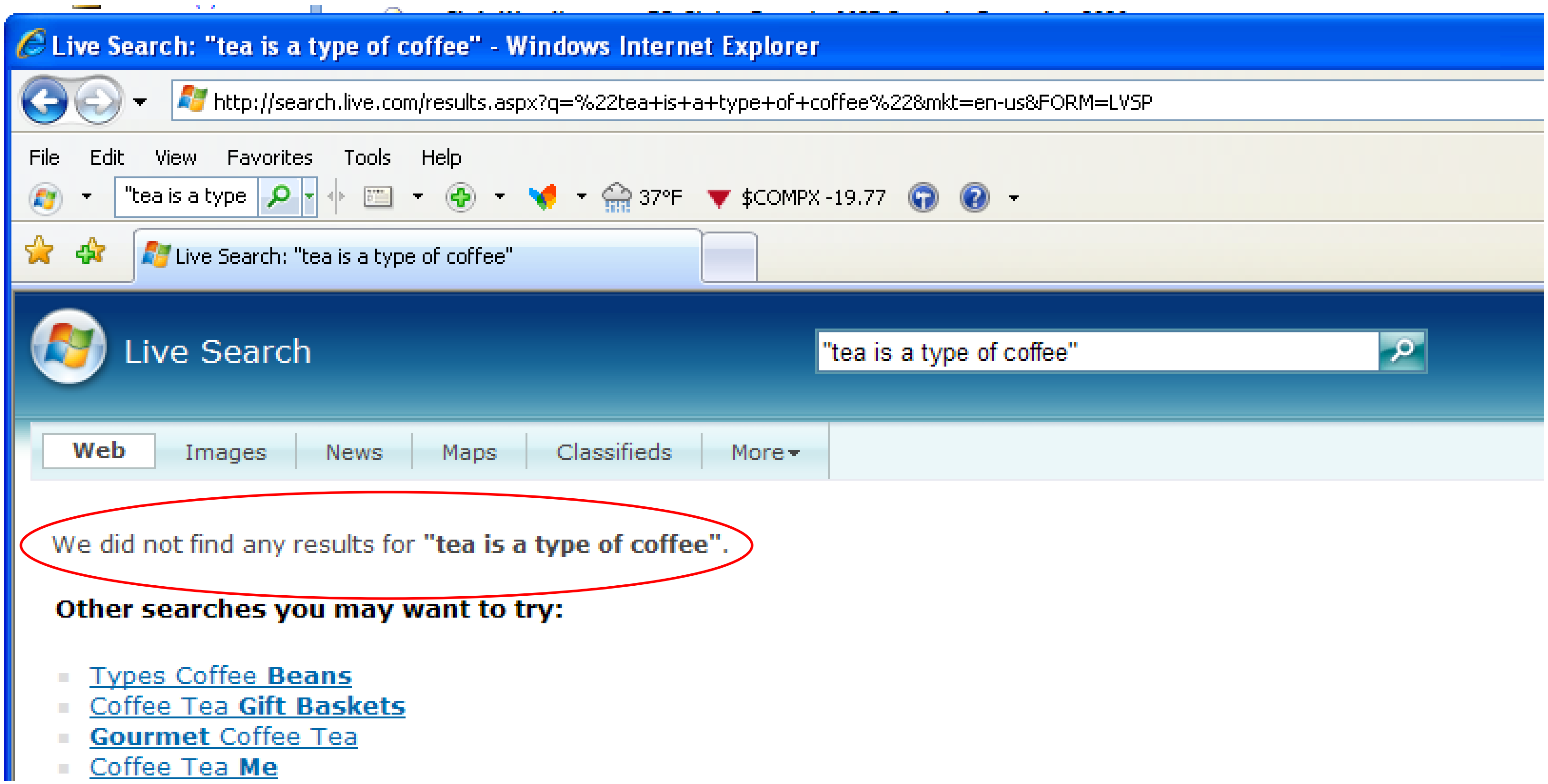
Examples


Live Search: "espresso is coffee" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22espresso+is+coffee%22&mkt=en-US&form=QBRE
File Edit View Favorites Tools Help
"espresso is c
37°F
$INDU +4.08
Live Search: "espresso is coffee"
Live Search
"espresso is coffee"
Web Images News Maps Classifieds More
"espresso is coffee" Page 1 of 229 results · Options · Advanced
What is espresso coffee? :: The Mad Coffee Maker
For Davids, espresso is " coffee brewed from beans, roasted medium to dark brown, the brewing process performed by forcing hot water through a finely-ground and compacted dose of coffee, at a pressure ...
www.madcoffeemaker.com/learn/espresso-coffee.html · Cached page

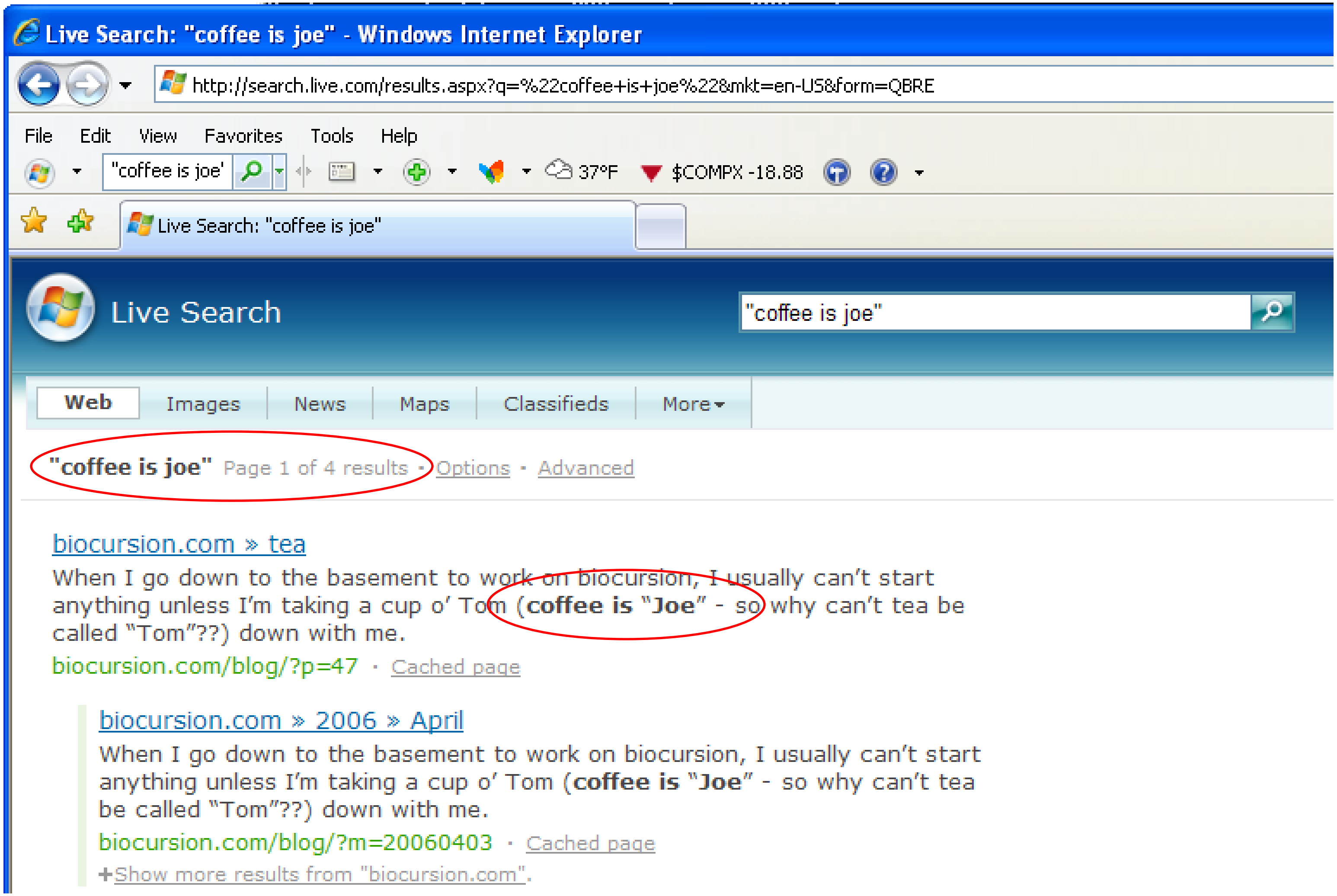
Templatic Web Queries

- Goal is to keep precision up
  - But use web, not specialized corpora, to guarantee association
- Start with a set of templatic queries
  - Fill slots in templates with words/phrases from context sets
  - Launch a quoted-string query
  - Record presence/absence of hits, counts
  - Goal is to confirm or deny semantic relationship
  - Results take on significance in aggregate…
- Enormous numbers of queries
  - Most will have a null result
  - Any single result untrustworthy, but meaningful in aggregate…
- Requires Web scale: even then, information is sparse
- SyrupInPlaceOf.txt

Examples

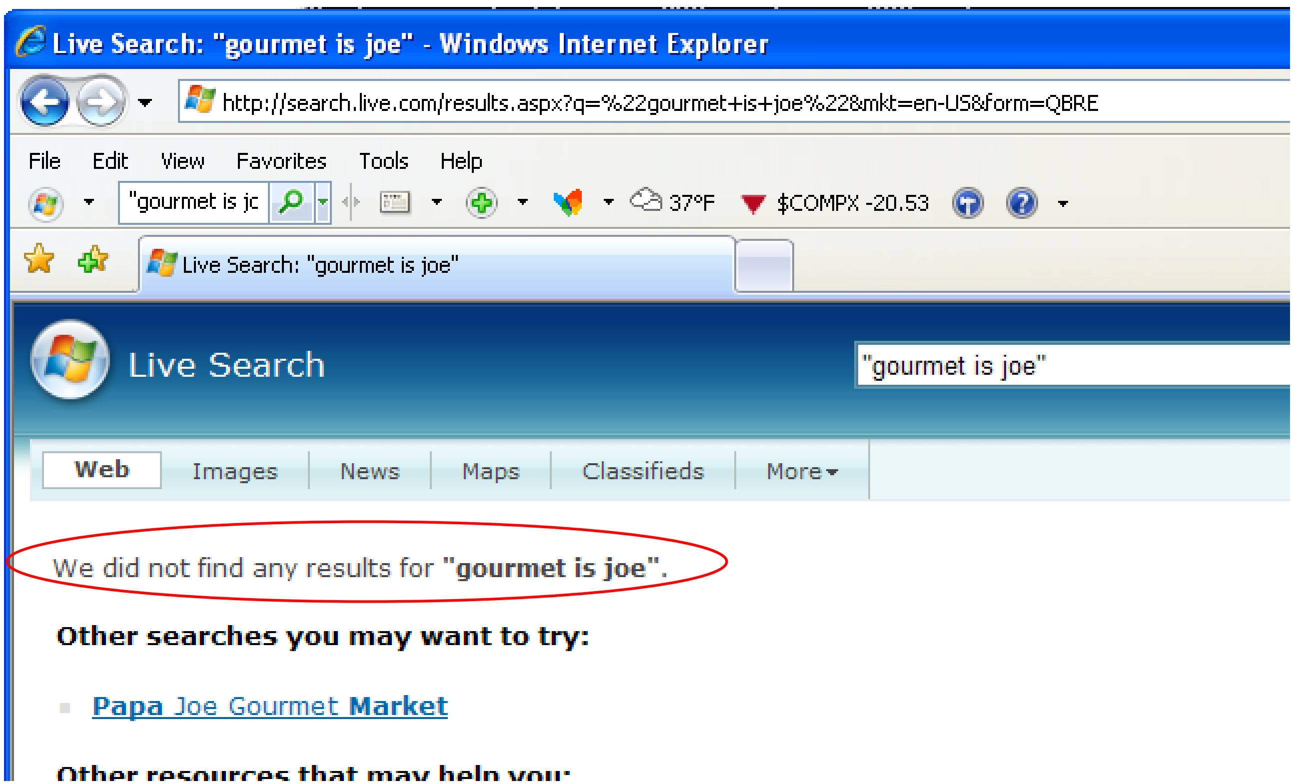

Live Search: "tea is a type of coffee" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22tea+is+a+type+of+coffee%22&mkt=en-us&FORM=LVSP
File Edit View Favorites Tools Help
"tea is a type
37°F
$COMPX -19.77
Live Search: "tea is a type of coffee"
Live Search
"tea is a type of coffee"
Web Images News Maps Classifieds More
We did not find any results for "tea is a type of coffee".
Other searches you may want to try:
Types Coffee Beans
Coffee Tea Gift Baskets
Gourmet Coffee Tea
Coffee Tea Me

Examples



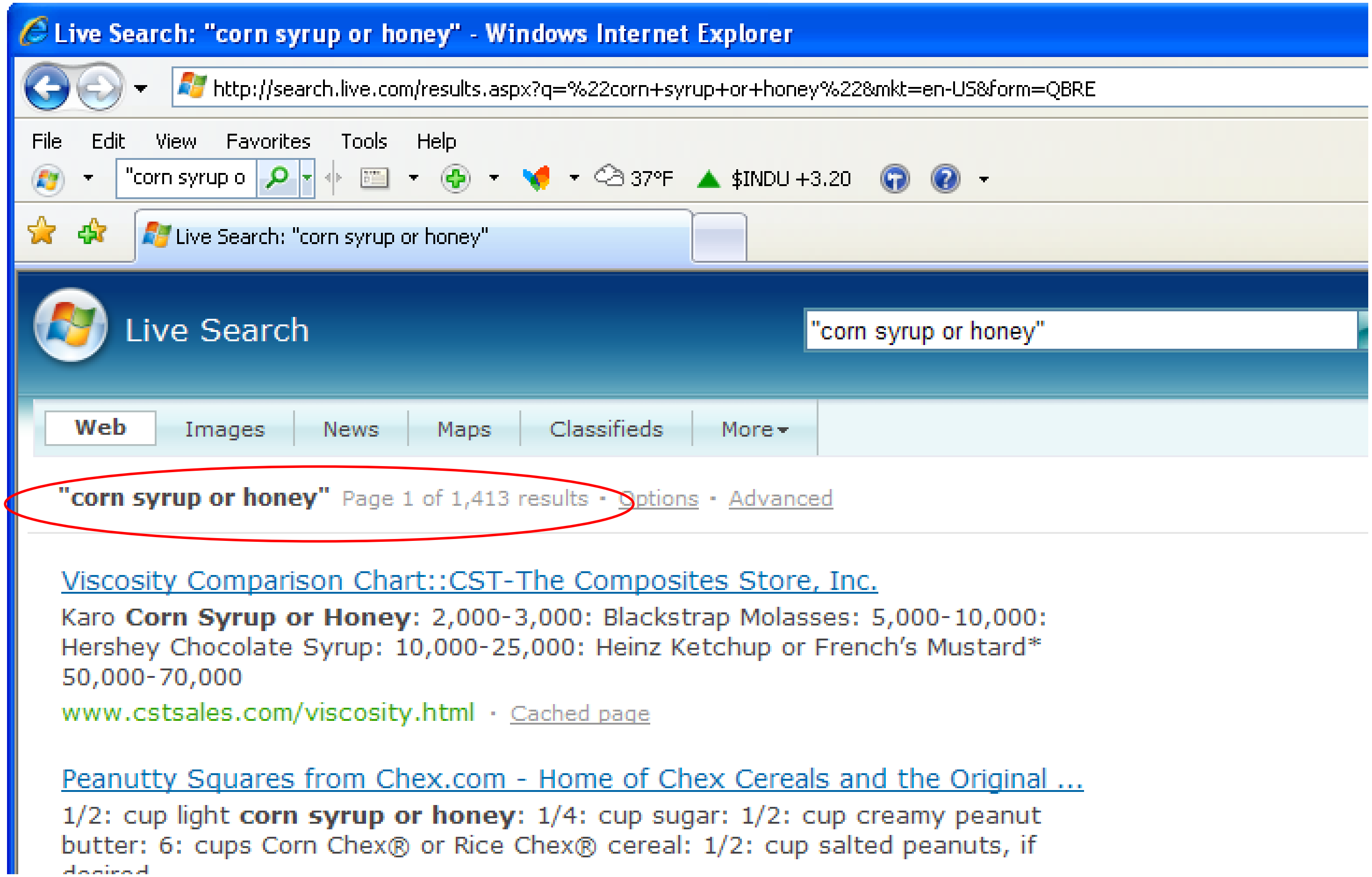
Live Search: "coffee is joe" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22coffee+is+joe%22&mkt=en-US&form=QBRE
File Edit View Favorites Tools Help
"coffee is joe'
37°F
$COMPX -18.88
Live Search: "coffee is joe"
Live Search
"coffee is joe"
Web Images News Maps Classifieds More
"coffee is joe" Page 1 of 4 results · Options · Advanced
biocursion.com » tea
When I go down to the basement to work on biocursion, I usually can't start anything unless I'm taking a cup o' Tom (coffee is "Joe" - so why can't tea be called "Tom"??) down with me.
biocursion.com/blog/?p=47 · Cached page
biocursion.com » 2006 » April
When I go down to the basement to work on biocursion, I usually can't start anything unless I'm taking a cup o' Tom (coffee is "Joe" - so why can't tea be called "Tom"??) down with me.
biocursion.com/blog/?m=20060403 · Cached page
+Show more results from "biocursion.com".

Examples

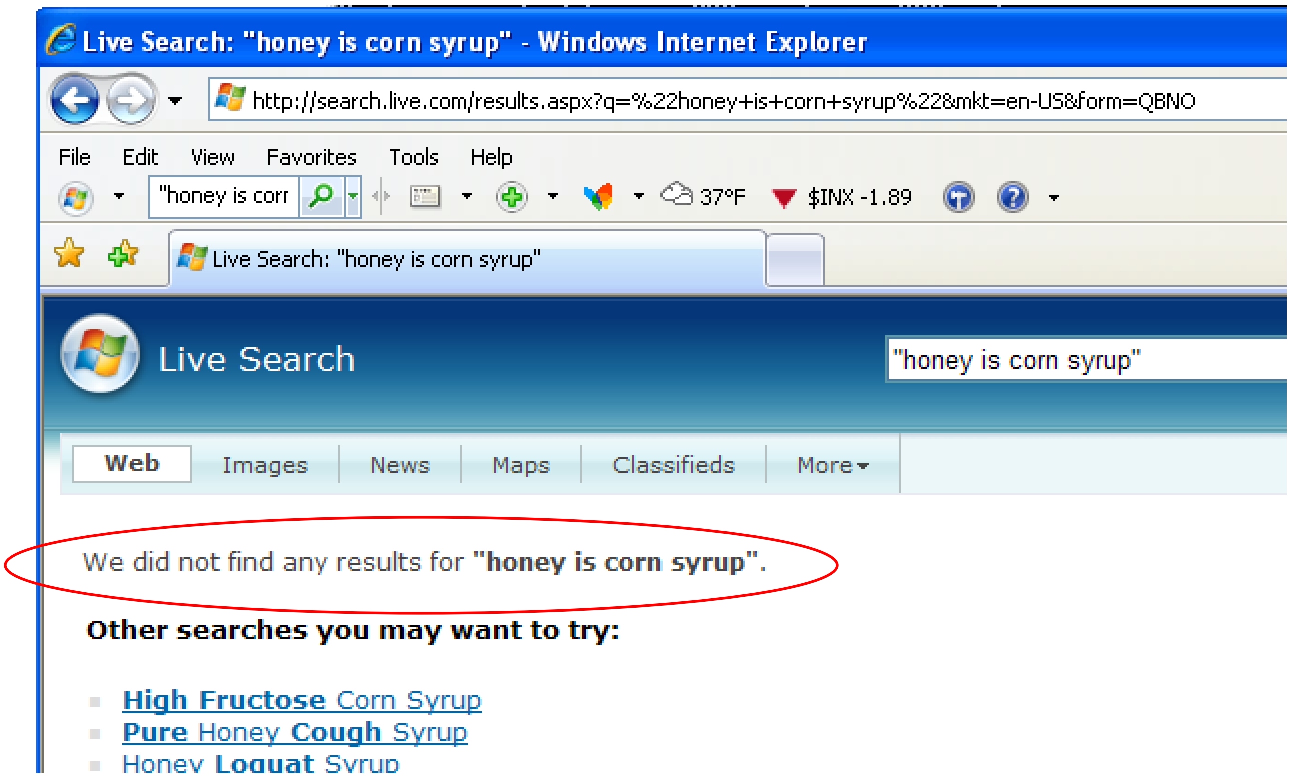


Live Search: "gourmet is joe" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22gourmet+is+joe%22&mkt=en-US&form=QBRE
File Edit View Favorites Tools Help
"gourmet is jc
37°F
$COMPX -20.53
Live Search: "gourmet is joe"
Live Search
"gourmet is joe"
Web
Images
News
Maps
Classifieds
More
We did not find any results for "gourmet is joe".
Other searches you may want to try:
Papa Joe Gourmet Market

Examples

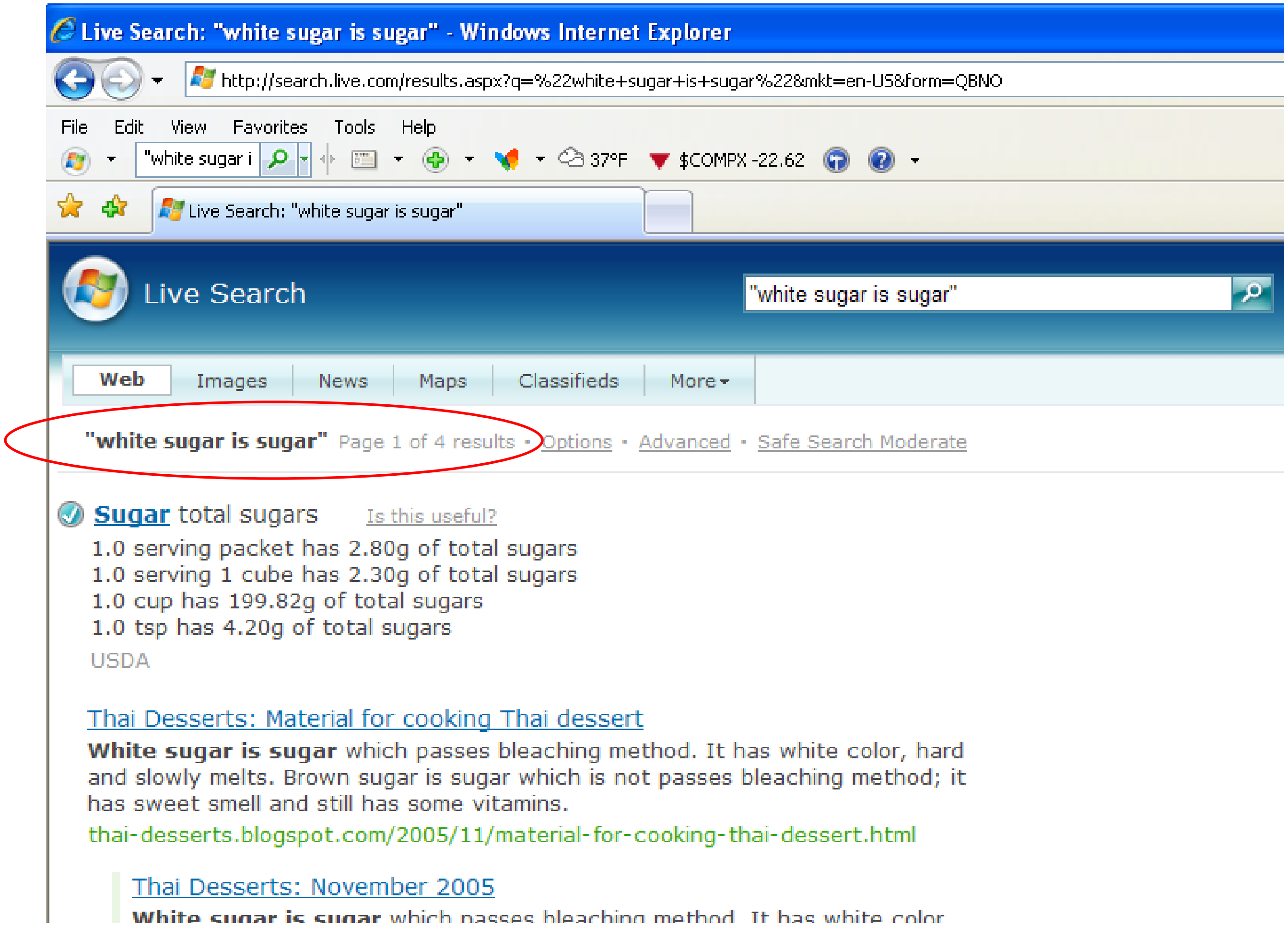

Live Search: "corn syrup or honey" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22corn+syrup+or+honey%22&mkt=en-US&form=QBRE
File Edit View Favorites Tools Help
"corn syrup o
37°F
$INDU +3.20
Live Search: "corn syrup or honey"
Live Search
"corn syrup or honey"
Web Images News Maps Classifieds More
"corn syrup or honey" Page 1 of 1,413 results · Options · Advanced
Viscosity Comparison Chart::CST-The Composites Store, Inc.
Karo Corn Syrup or Honey: 2,000-3,000: Blackstrap Molasses: 5,000-10,000: Hershey Chocolate Syrup: 10,000-25,000: Heinz Ketchup or French's Mustard* 50,000-70,000
www.cstsales.com/viscosity.html · Cached page
Peanutty Squares from Chex.com - Home of Chex Cereals and the Original ...
1/2: cup light corn syrup or honey: 1/4: cup sugar: 1/2: cup creamy peanut butter: 6: cups Corn Chex® or Rice Chex® cereal: 1/2: cup salted peanuts, if

Examples

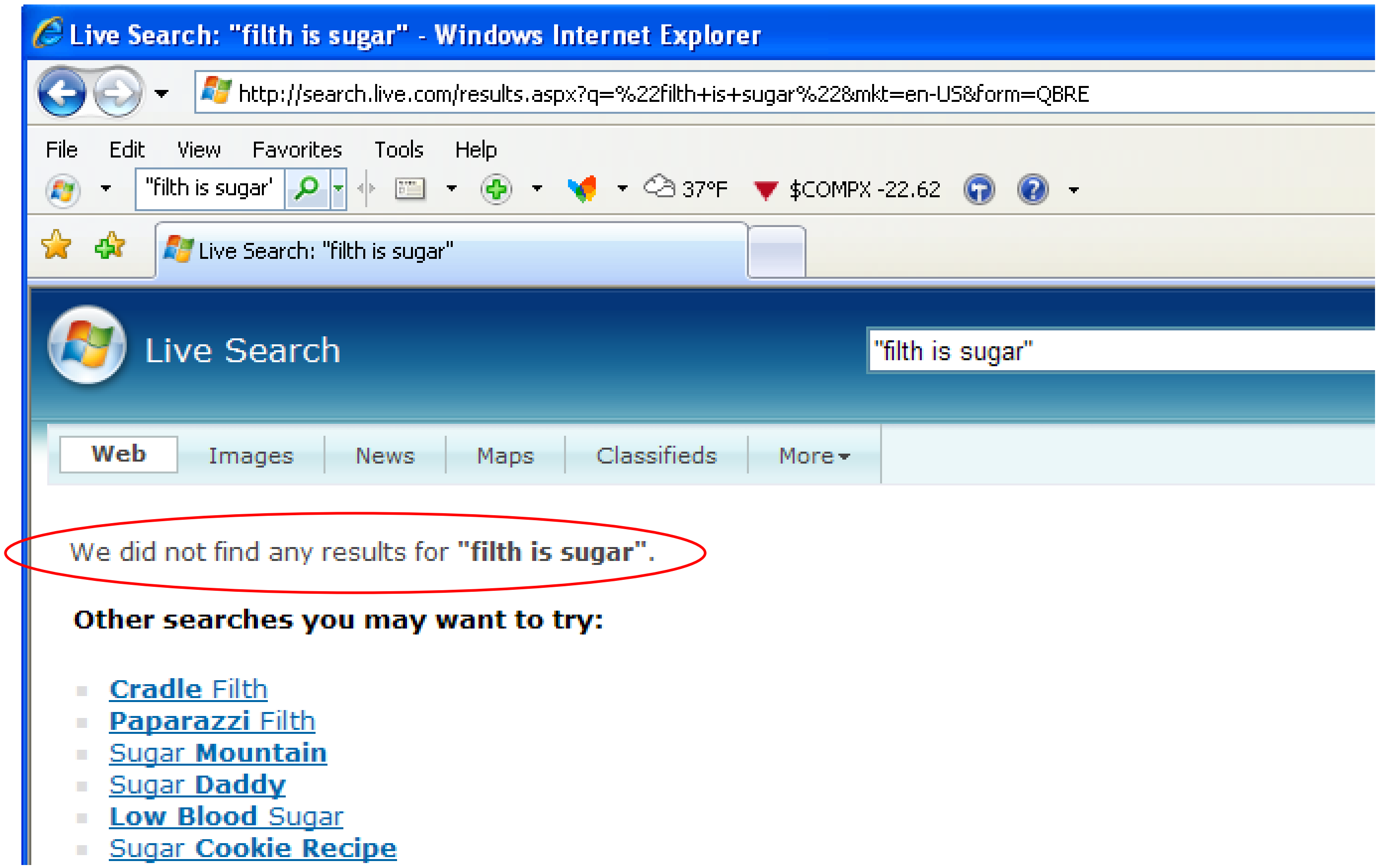


Live Search: "honey is corn syrup" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22honey+is+corn+syrup%22&mkt=en-US&form=QBNO
File Edit View Favorites Tools Help
"honey is corr
37°F
$INX -1.89
Live Search: "honey is corn syrup"
Live Search
"honey is corn syrup"
Web Images News Maps Classifieds More
We did not find any results for "honey is corn syrup".
Other searches you may want to try:
High Fructose Corn Syrup
Pure Honey Cough Syrup
Honey Loquat Syrup

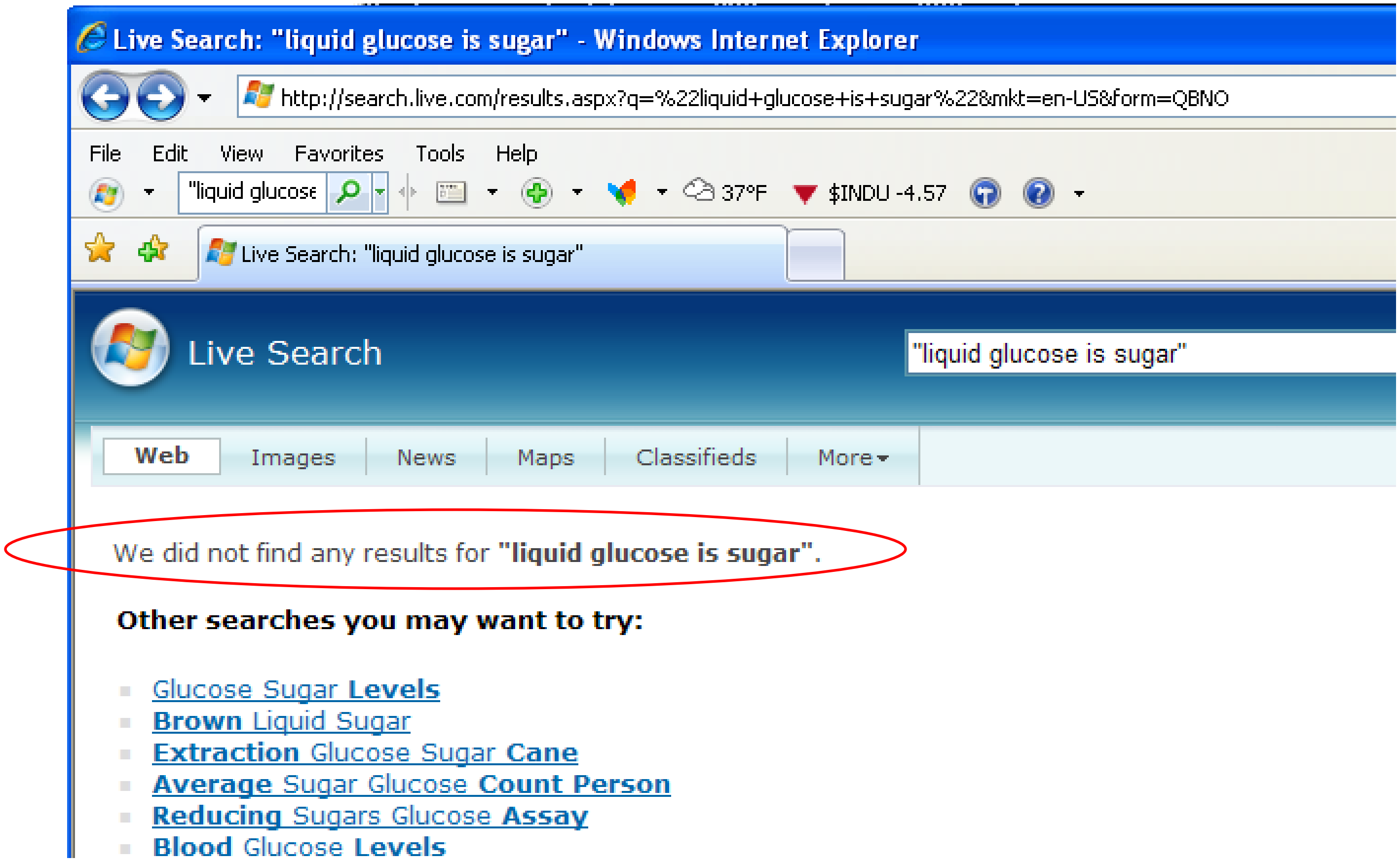
Examples


Live Search: "white sugar is sugar" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22white+sugar+is+sugar%22&mkt=en-US&form=QBNO
File Edit View Favorites Tools Help
"white sugar i
37°F
$COMPX -22.62
Live Search: "white sugar is sugar"
Live Search
"white sugar is sugar"
Web Images News Maps Classifieds More
"white sugar is sugar" Page 1 of 4 results · Options · Advanced · Safe Search Moderate
Sugar total sugars Is this useful?
1.0 serving packet has 2.80g of total sugars
1.0 serving 1 cube has 2.30g of total sugars
1.0 cup has 199.82g of total sugars
1.0 tsp has 4.20g of total sugars
USDA
Thai Desserts: Material for cooking Thai dessert
White sugar is sugar which passes bleaching method. It has white color, hard and slowly melts. Brown sugar is sugar which is not passes bleaching method; it has sweet smell and still has some vitamins.
thai-desserts.blogspot.com/2005/11/material-for-cooking-thai-dessert.html
Thai Desserts: November 2005

Examples

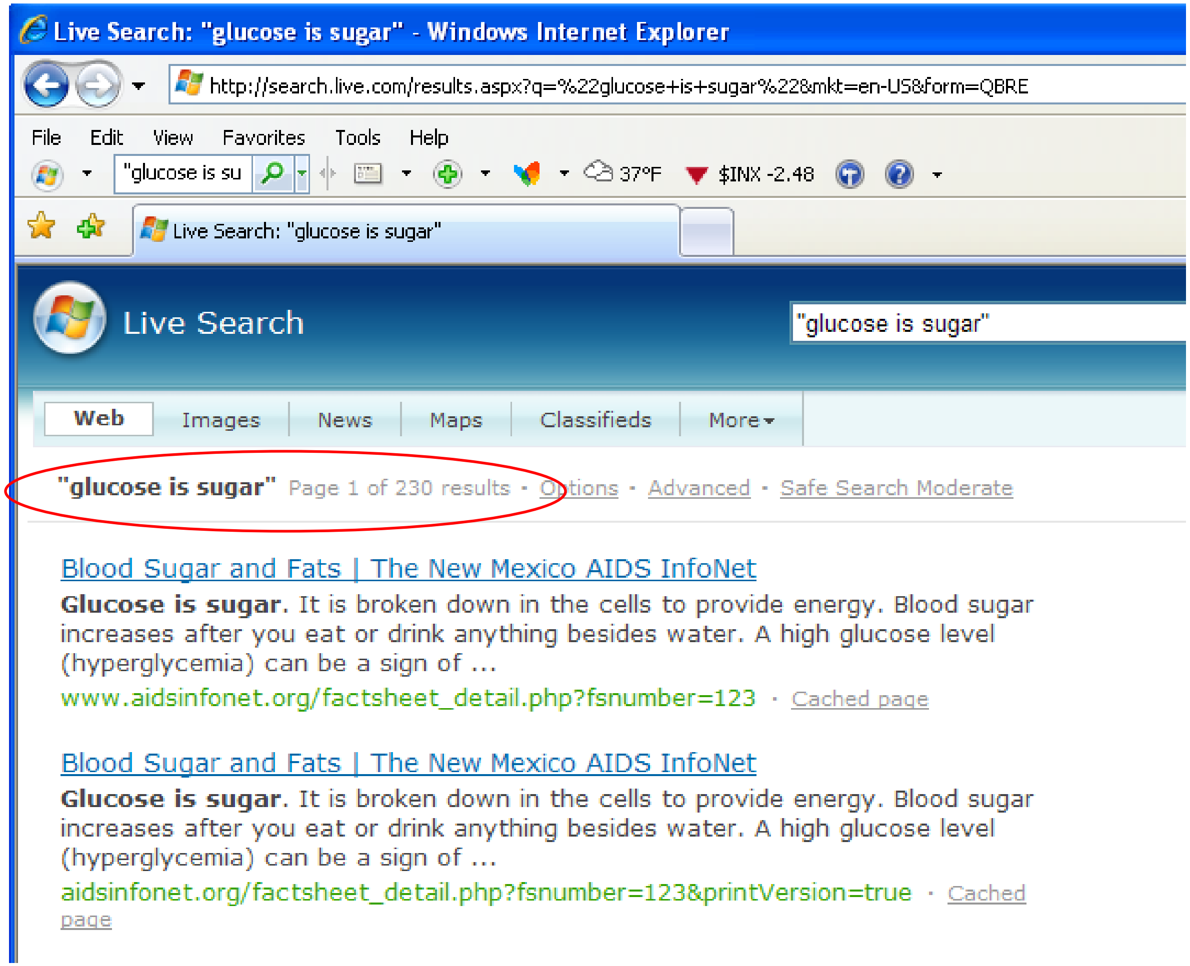

Live Search: "filth is sugar" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22filth+is+sugar%22&mkt=en-US&form=QBRE
File Edit View Favorites Tools Help
"filth is sugar"
37°F
$COMPX -22.62
Live Search: "filth is sugar"
Live Search
"filth is sugar"
Web Images News Maps Classifieds More
We did not find any results for "filth is sugar".
Other searches you may want to try:
Cradle Filth
Paparazzi Filth
Sugar Mountain
Sugar Daddy
Low Blood Sugar
Sugar Cookie Recipe

Examples

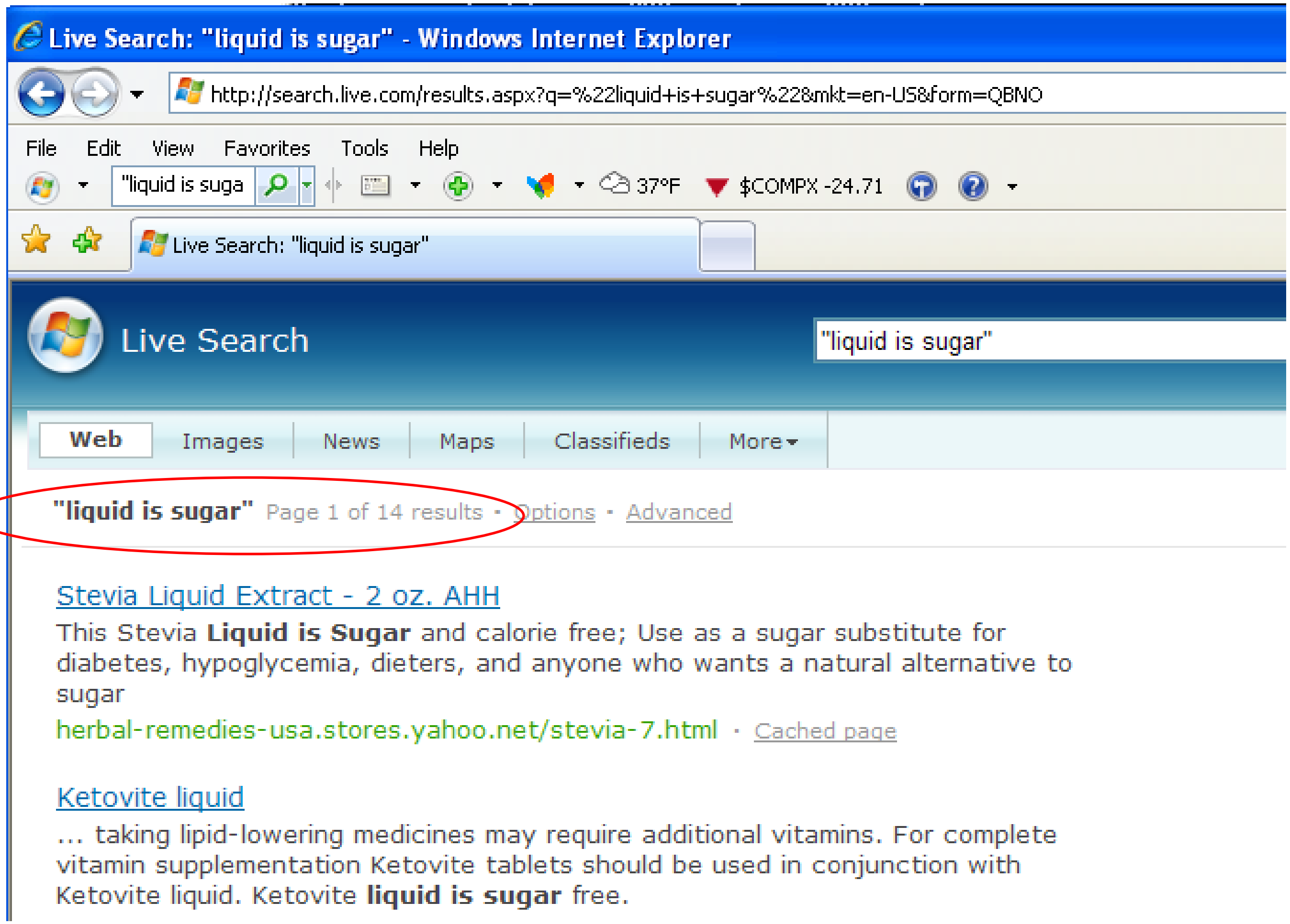


Live Search: "liquid glucose is sugar" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22liquid+glucose+is+sugar%22&mkt=en-US&form=QBNO
File Edit View Favorites Tools Help
"liquid glucose
37°F
$INDU -4.57
Live Search: "liquid glucose is sugar"
Live Search
"liquid glucose is sugar"
Web
Images
News
Maps
Classifieds
More
We did not find any results for "liquid glucose is sugar".
Other searches you may want to try:
Glucose Sugar Levels
Brown Liquid Sugar
Extraction Glucose Sugar Cane
Average Sugar Glucose Count Person
Reducing Sugars Glucose Assay
Blood Glucose Levels

Examples


Live Search: "glucose is sugar" - Windows Internet Explorer
http://search.live.com/results.aspx?q=%22glucose+is+sugar%22&mkt=en-US&form=QBRE
File Edit View Favorites Tools Help
"glucose is su
37°F
$INX -2.48
Live Search: "glucose is sugar"
Live Search
"glucose is sugar"
Web Images News Maps Classifieds More
"glucose is sugar" Page 1 of 230 results · Options · Advanced · Safe Search Moderate
Blood Sugar and Fats | The New Mexico AIDS InfoNet
Glucose is sugar. It is broken down in the cells to provide energy. Blood sugar increases after you eat or drink anything besides water. A high glucose level (hyperglycemia) can be a sign of ...
www.aidsinfonet.org/factsheet_detail.php?fsnumber=123 · Cached page
Blood Sugar and Fats | The New Mexico AIDS InfoNet
Glucose is sugar. It is broken down in the cells to provide energy. Blood sugar increases after you eat or drink anything besides water. A high glucose level (hyperglycemia) can be a sign of ...
aidsinfonet.org/factsheet_detail.php?fsnumber=123&printVersion=true · Cached page

Any single query may mislead…



Examples of String Queries

- Hypernymy
  - *X is a(n) Y / Y is a(n) X*
  - *X's are Y's / Y's are X's*
  - *X is a type of Y / Y is a type of X*
  - etc.
- Set membership
  - *"Y, X, (and / or) Z" (+ permutations)*
  - *"X or Y or Z"*
  - *"an X or a Y or a Z"*
  - *"the X but not the Y"*
  - *"the X, not the Y"*
- Complement of Hearst's (1992) string patterns
  - Hearst looked for patterns like "X and other Y" in a document
  - We use templatic patterns to verify hypotheses constructed by looking across a set of documents
- Serves as a check on the Distributional Hypothesis' tendency to lump together disparate types of alternations

Still a prototype, but promising

- Structure from nothing
  - With just a bit of heuristic prodding
- Exhaustive exploration of the web doesn’t appear necessary for interesting results
  - High frequency alternations appear early on
- Morphosyntactic structure
  - Constituent boundaries
    - With relative strength information from cumulative counts
  - Named entity classes
  - Morphological alternations
  - Phrases with variable slots
- Lexical/Phrasal semantic structure
  - Hypernymy & set membership
  - Contextually-appropriate synonyms/phrasal alternations

Limitations

- Inefficient
  - Vast numbers of queries, most yielding a null result
  - Direct access to a structured web index would be dramatically more efficient, but current approach ok for experimentation
- Can’t learn major syntactic alternations
  - e.g. active/passive
- Precision is still likely to be lower than with specialized data
  - But lots of knobs and levers to tweak: classifiers for clustering contexts, more heuristic secondary queries, etc.

Real Utility is as a tool to aid SMT

- SMT approach to paraphrase is the best-motivated
  - We need sophisticated techniques for mining free text for monolingual phrasal mappings
  - Web-crawl approach is promising strategy
- Phrasal alternations are useful input for higher-level alignment algorithms
  - Comparable corpus extraction a hot topic in SMT (e.g. Munteanu & Marcu 2006)
  - If we can do a good job at the phrase level, easier to find sentence-level alternations