Cognitive Computation Group
University of Illinois at Urbana-Champaign

Tools to Support Textual Inference

Mark Sammons

University of Illinois, Urbana-Champaign

mssammon@illinois.edu

http://cogcomp.cs.illinois.edu

Multimodal Information Access & Synthesis
mias

What is at the heart of “big” NLP apps?

- Information Retrieval
- Question Answering
- Translation
- Information Extraction
- Summarization
- Recognizing Textual Entailment
- ...All require comparison of spans of text to determine whether they “match” in some way

Recognizing Textual Entailment**

- ** "Local Textual Inference" (Zaenen et al., Manning)
- Operational definition for Text Understanding:

  Given two text fragments (a Text T and a Hypothesis H),
  T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people

- Can frame many NLP tasks as RTE:
  - IE: Formulate relation as short sentence with generic placeholders, e.g. "Work-For" becomes "An organization employs a person." -- the Hypothesis; Document paragraphs become Texts
  - QA: many questions can be rephrased as statements with generic placeholders: "Something is the fastest car in the world."
  - Summarization: Detect novelty of new text span by determining whether current summary entails it or not.

OPERATOR 1: Phrasal Verb

Replace phrasal verbs
with an equivalent single word verb

T: Hurricane Katrina petroleum-supply outlook improved somewhat, yesterday, as U.S. and European governments finally reached a consensus. They finally made up their minds to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Offers by individual European governments involved supplies of crude or refined oil products.

T: Hurricane Katrina petroleum-supply outlook improved somewhat, yesterday, as U.S. and European governments finally reached a consensus. They finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Offers by individual European governments involved supplies of crude or refined oil products.

T: Hurricane Katrina petroleum-supply outlook improved somewhat, yesterday, as U.S. and European governments finally reached a consensus.
U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Offers by individual European governments involved supplies of crude or refined oil products.

OPERATOR 2: Coreference Resolution

Replace pronouns/possessive pronouns with the entity to which they refer

T: Hurricane Katrina petroleum-supply outlook improved somewhat, yesterday, as U.S. and European governments finally reached a consensus.
They finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Offers by individual European governments involved supplies of crude or refined oil products.

T: ~~Hurricane Katrina petroleum-supply outlook improved somewhat yesterday, as U.S. and European governments finally reached a consensus.~~
U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Offers by individual European governments involved supplies of crude or refined oil products.

OPERATOR 3: Focus of Attention

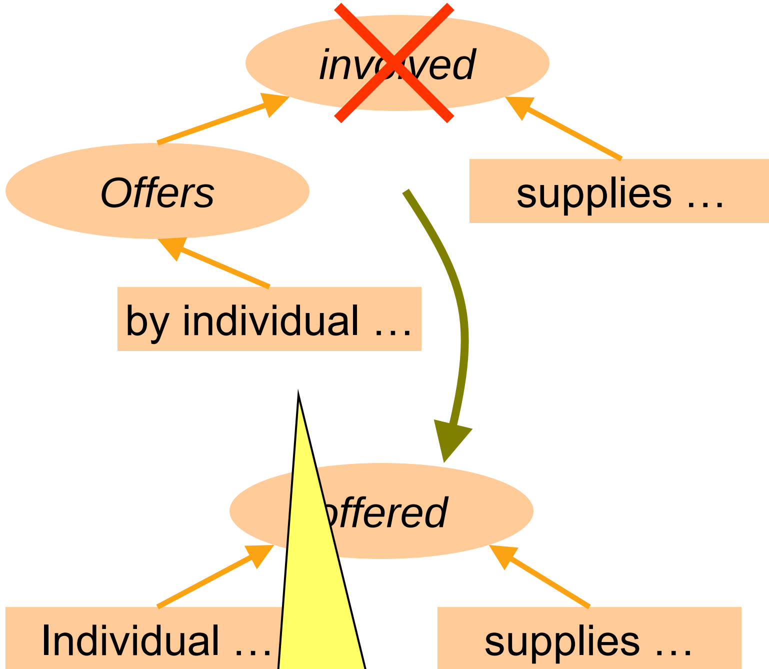
Remove segments of a sentence that do not appear to be necessary; may allow more accurate annotation of remaining words

T: U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Offers by individual European governments involved supplies of crude or refined oil products.

OPERATOR 4: Nominalization Promotion

Replace a verb that does not express a useful/meaningful relationship with a nominalization in one of its arguments



Requires semantic role labeling (for noun predicates)

T: U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

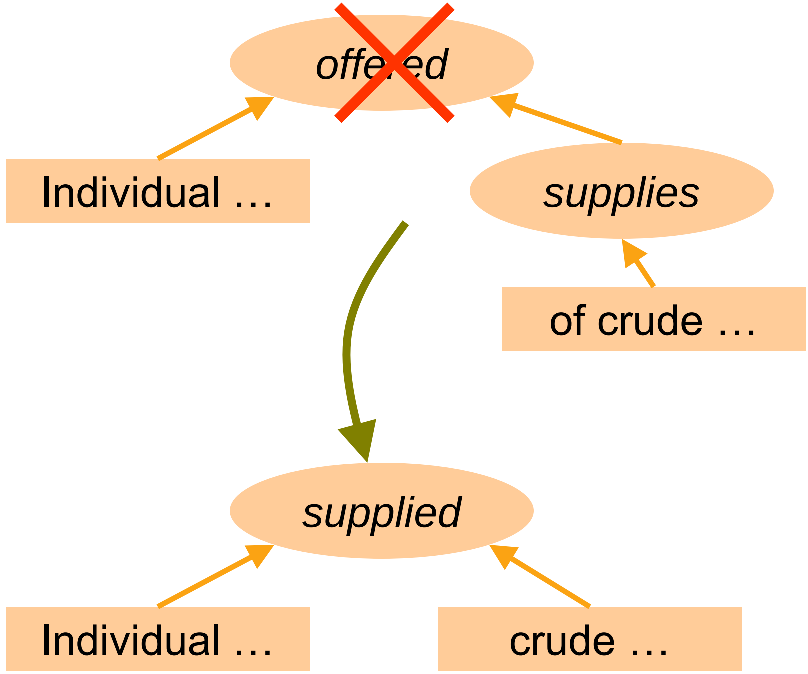
H: Individual European governments offered supplies of crude or refined oil products.

T: U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Offers by individual European governments involved supplies of crude or refined oil products.

OPERATOR 4: Nominalization Promotion

Replace a verb that does not express a useful/meaningful relationship with a nominalization in one of its arguments



T: U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

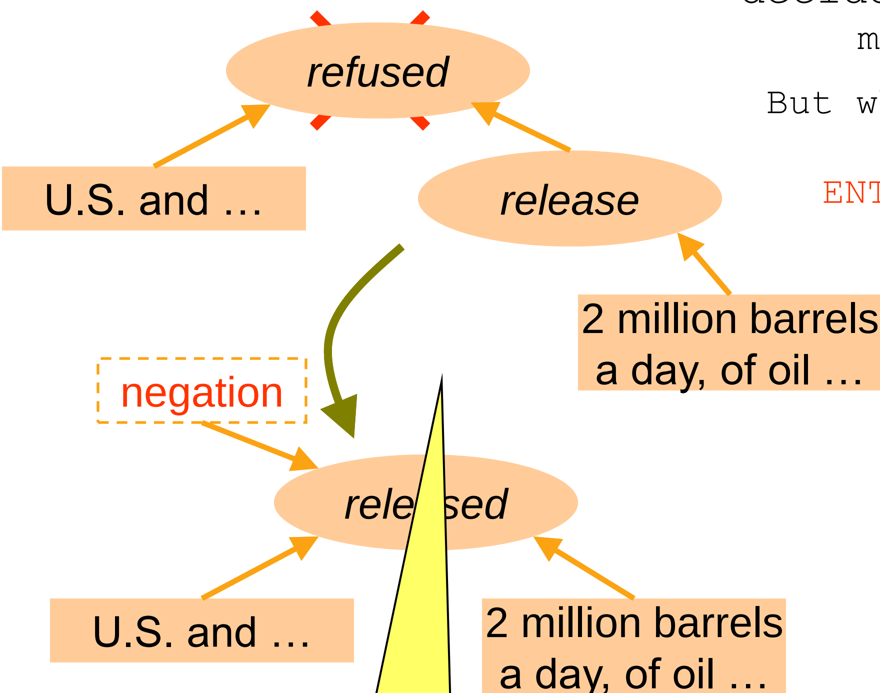
H: Individual European governments offered supplies of crude or refined oil products.

T: U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Individual European governments supplied crude or refined oil products.

OPERATOR 5: Predicate Embedding Resolution

Replace a verb compound where the first verb may indicate modality or negation with a single verb, marked with negation/modality attribute



Requires semantic role labeling (for noun predicates)

'decided' (almost) does not change the meaning of the embedded verb

But what if the embedding verb had been 'refused'?

ENTAILMENT SHOULD NOT SUCCEED

T: U.S. and European governments finally released 2 million barrels a day, of oil and refined products, from their reserves.

H: Individual European governments supplied crude or refined oil products.

T: U.S. and European governments finally decided to release 2 million barrels a day, of oil and refined products, from their reserves.

H: Individual European governments supplied crude or refined oil products.

OPERATOR 6: Predicate Matching

System matches PREDICATES and their ARGUMENTS
-- accounts for monotonicity, modality, negation, and quantifiers

ENTAILMENT SUCCEEDS

T: U.S. and European governments finally released 2 million barrels a day, of oil and refined products, from their reserves.

H: Individual European governments supplied crude or refined oil products.

Requires lexical abstraction

T: U.S. and European governments finally released 2 million barrels a day, of oil and refined products, from their reserves.

H: Individual European governments supplied crude or refined oil products.

Overview

- Common Sub-tasks in Textual Inference
- Recognizing Concepts
- Recognizing Structure Connecting Concepts
- Recognizing Relations between Concepts
- An exercise in Applied Textual Inference: Recognizing Textual Entailment

Overview

- Common Sub-tasks in Textual Inference
- Recognizing Concepts
- Recognizing Structure Connecting Concepts
- Recognizing Relations between Concepts
- An exercise in Applied Textual Inference: Recognizing Textual Entailment

Recognizing Concepts

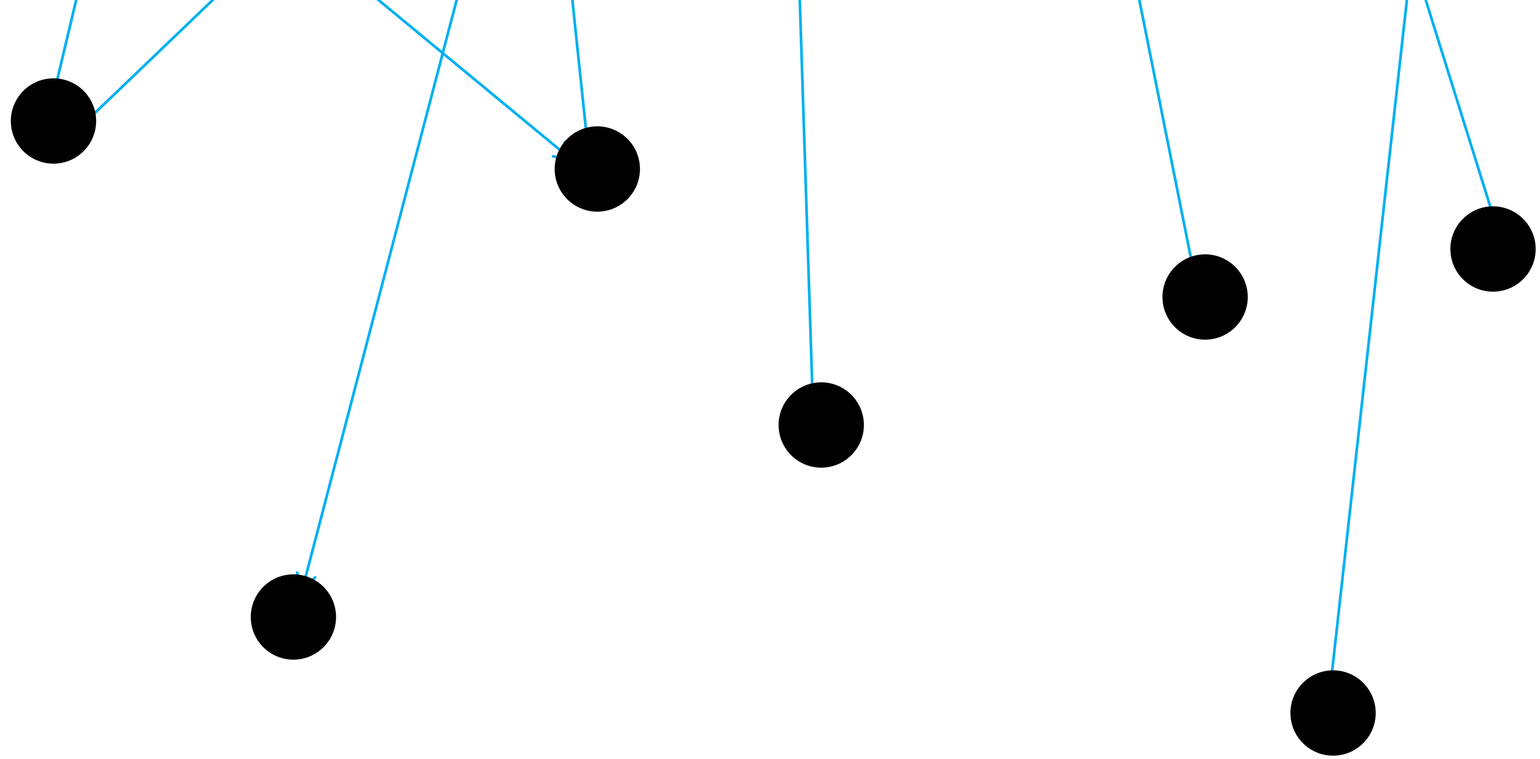
- Standard unsupervised approaches:
  - TFIDF
  - Multi-Word Expression recognition via co-occurrence statistics
  - Give boundaries, but not types
  - Moderate precision, good coverage
- Supervised approaches
  - Shallow parsing
  - Named Entity Recognition
  - Focused type information at the cost of coverage; annotation expense
- Given some kind of structured reference collection, can we learn a good concept recognizer?

"Wikification": Organizing knowledge

It's a version of ***Chicago*** – the standard classic Macintosh menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

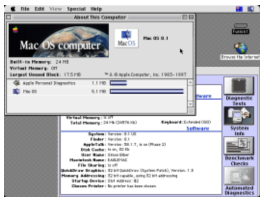
Cross-document co-reference resolution

It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.



Reference(disambiguation to Wikipedia)

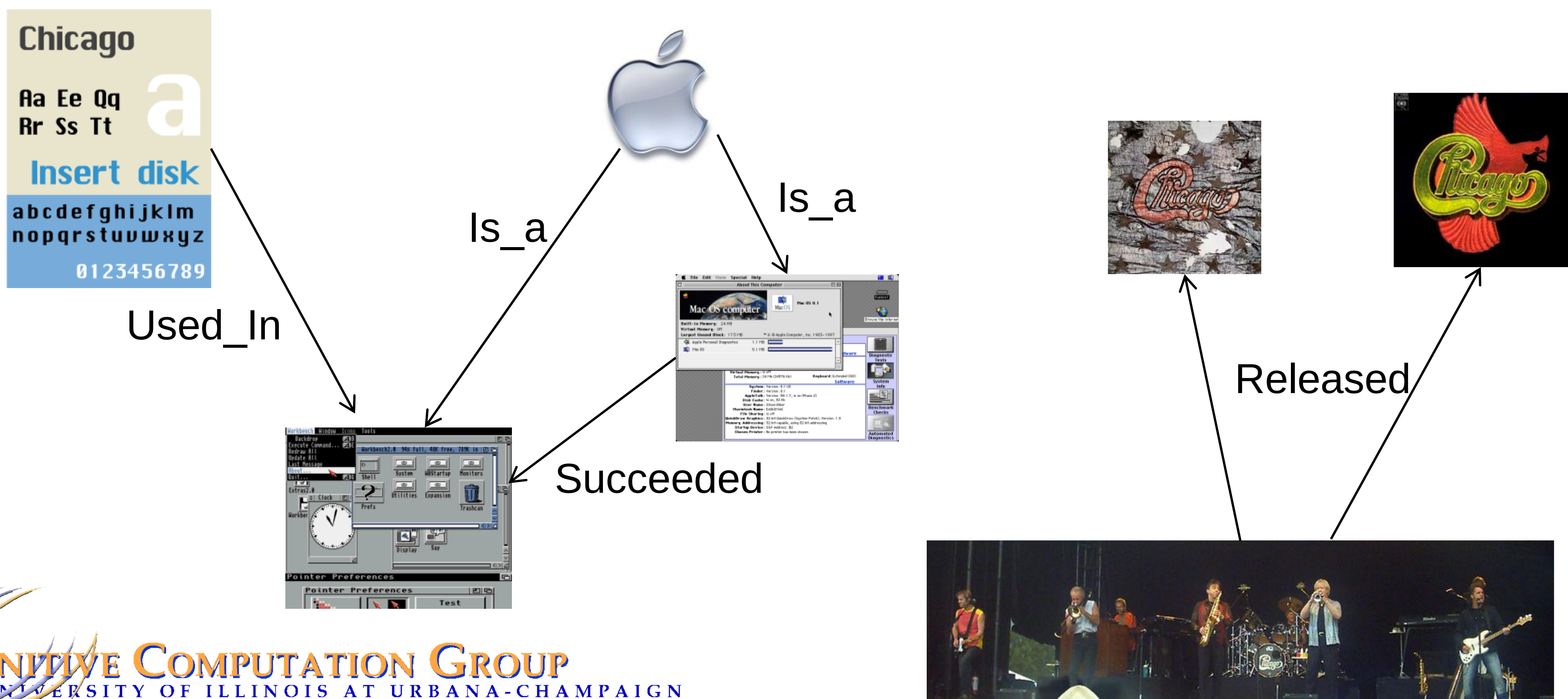
It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.



COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

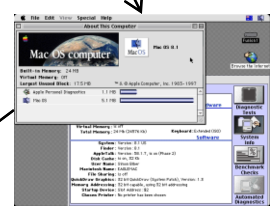
The “reference” collection has structure

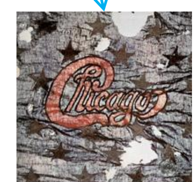
It’s a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the ”N”.	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

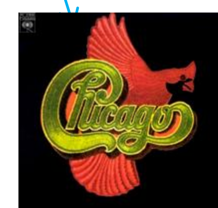


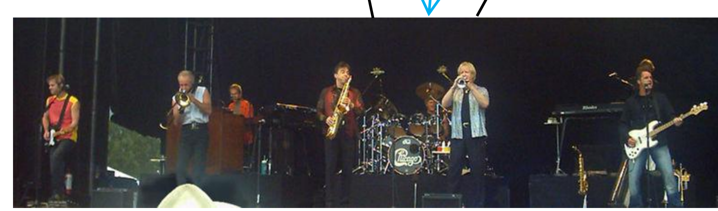
Analysis of Information Networks

It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.









Performance

Dataset	Baseline	Baseline+ Lexical	Baseline+ Lexical+ Global
ACE	94.05	96.21	97.83
MSN News	81.91	85.10	87.02
AQUAINT	93.19	95.57	94.38
Wikipedia Test	85.88	93.59	94.18

Wikifier Summary

- Broad spectrum “concept” recognizer
  - Complements NER
  - good anecdotal performance on unseen data
  - …without the annotation overhead
- Context sensitive mutual disambiguation
  - First-cut non-anaphoric co-reference capability – in a very broad domain
- A good start for bootstrapping NLP in a new domain
  - E.g. recognizing “mentions” of concepts that are/should be(?) in some ontology
- Real-time web demo:

  http://cogcomp.cs.illinois.edu/demo/wikify/

Overview

- Common Sub-tasks in Textual Inference
- Recognizing Concepts
- Recognizing Structure Connecting Concepts
- Recognizing Relations between Concepts
- An exercise in Applied Textual Inference: Recognizing Textual Entailment

Recognizing Structure Linking Concepts

- Goal: broad coverage tools giving coarse sentence structure with some semantic annotation
  - Intra-sentence: Semantic Role Labeling
  - Inter- and intra-sentence: Co-reference
- Philosophy: integrate statistical models with domain-specific constraints
  - Local decisions made by machine-learned classifiers
  - Global decision reached by optimizing local decisions with respect to constraints
  - Chosen formalism: Integer Linear Programming

Semantic Role Labeling

- Real-time web demo:

  http://cogcomp.cs.illinois.edu/demo/srl/

Co-reference

- Real-time web demo:

  http://cogcomp.cs.illinois.edu/demo/coref/

Overview

- Common Sub-tasks in Textual Inference
- Recognizing Concepts
- Recognizing Structure Connecting Concepts
- Recognizing Relations between Concepts
- An exercise in Applied Textual Inference: Recognizing Textual Entailment

Required Capabilities

- In applications requiring textual inference, we often need to know when two terms are substitutable in some way:

T: John Smith met ***Mel Gibson*** yesterday.

H: John Smith met an ***actor*** yesterday.

T: An earthquake strikes ***Taiwan***.

H: An earthquake strikes ***Japan***.

Similarity vs. Substitutability

- Similarity measures, e.g. distributional similarity metrics, identify relatedness of terms…
- …but don't tell you ***how*** the terms are related

T: An earthquake strikes ***Taiwan***.

H: An earthquake strikes ***Japan***.

T: An earthquake strikes ***Honshu***.

H: An earthquake strikes ***Japan***.

- We need specialized resources to make these finer distinctions.

So you want to compare some text….

- **How similar are two lexical expressions?**
  - Depends on what they are
  - String edit distance is usually a weak measure
  - … think about coreference resolution…

String 1	String 2	Norm. edit sim.
Shiite	Shi’ ‘ite	0.667
Mr. Smith	Mrs. Smith	0.900
Wilbur T. Gobsmack	Mr. Gobsmack	0.611
Frigid	Cold	0.167
Wealth	Wreath	0.667
Paris	France	0.167

- **Solution: specialized metrics**

NESim

- **Set of entity-type-specific measures**
  - Acronyms, Prefix/Title rules, distance metric
- **Score reflects similarity based on type information**
- **Score is asymmetric**

String 1	String 2	Norm. edit distance
Shiite	Shi’ ‘ite	0.922
Joan Smith	John Smith	0
Wilbur T. Gobsmack	Mr. Gobsmack	0.95
Frigid	Cold	0
Wealth	Wreath	0.900
Paris	France	0.411

Broad-spectrum ontologies exist!

- Simple approach: determine relations between concepts using static resources
  - WordNet, VerbNet
  - Some clever integration of e.g. WordNet + Wikipedia (YAGO)
  - Some clever "growth" of resources, e.g. Extended WordNet (Snow et al. 06, …)
- …but there are problems:
  - **Noisy** (low precision)
  - Limited **coverage** (low recall)
  - Ontology/occurrence **mismatch** (e.g. Camry Vs. Toyota Camry)

WNSim

- **Generate table mapping terms linked in WordNet ontology**
  - Synonymy, Hypernymy, Meronymy
- **Score reflects distance (up to 3 edges, undirected – e.g. via lowest common subsumer)**
- **Score is symmetric**

String 1	String 2	WNSim distance
Shiite	Shi' 'ite	0
Mr. Smith	Mrs. Smith	0
Wilbur T. Gobsmack	Mr. Gobsmack	0
Frigid	Cold	1
Wealth	Wreath	0
Paris	France	0

Taxonomic Relation Classifier (TAREC): On-demand Ontological Relations

- In textual inference, ontologies are useful to identify relations between concepts – typically, to determine whether two concepts are substitutable
- The functionality we need is, given two candidate concepts X and Y, to determine whether
  - X is substitutable for Y
  - X is definitely not substitutable for Y (direct evidence *against* a match)
  - X is not related to Y (but no direct evidence against a match)

Basic Relations

Relation	Meaning	*x*	*y*
$x \leftarrow y$	ancestor	actor	Mel Gibson
$x \rightarrow y$	child	Makalu	mountain
$x \leftrightarrow y$	sibling	copper	oxygen
$x \not\leftrightarrow y$	none	egg	C++

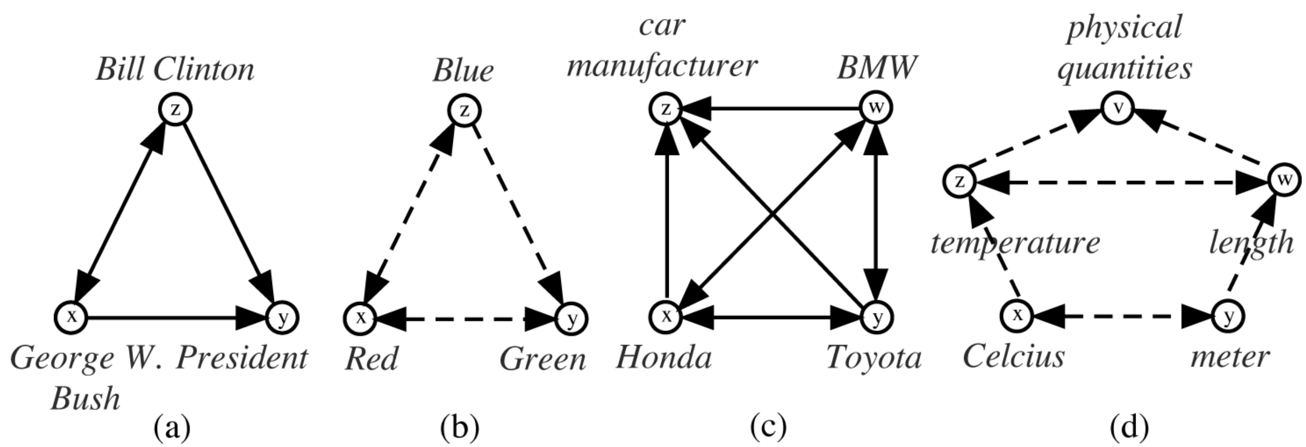
Taxonomic Relation Classifier (TAREC)

- Normalize query terms to reference collection
  - Use pattern-based extraction + web search to identify alternative terms (e.g., delimiter-based list extraction)
- Train a local classifier to compare query terms
  - Mine Wikipedia for related terms: article titles, content, and categories
- PMI: pmi(x,y) = log [Nf(x,y)/f(x)f(y)]
  where f(.) counts the # of its argument; N is the total # of Wikipedia pages.

Bag of words - Degree of similarity	
texts(x) vs. categories(y)	*texts(x) vs. texts(y)*
categories(x) vs. texts(y)	*categories(x) vs. categories(y)*

Improving Decisions with Constraints

- Improve local classifier by using concepts related to query terms X, Y to constrain them
  - Extract related terms from static ontology (YAGO)
  - Use local classifier to determine relations between them
  - Select best set of relation labels linking X, Y and other concepts that does not match a pre-specified violation pattern (e.g. b, d)



Performance

System	Wiki	WordNet	non-Wiki
Strube07	24.59	24.13	21.18
Snow06	41.23	46.91	34.46
Yago07	69.95	70.42	34.26
TAREC (local)	89.37	89.72	31.22
TAREC	**91.03**	**91.2**	**45.21**

- Limitations: Useful for Things rather than Relations
  - Majority of Wikipedia pages are about entity-like concepts
  - Need to supplement with additional knowledge for textual inference

TAREC summary

- Broad spectrum ontology-like resource
- Functional interface matched to typical inference need
- Leverages Wikipedia as reference collection
  - Dynamic resource – regular updates
- Normalizes input terms to reference “ontology”
- Uses local classification plus constrained optimization to incorporate common-sense constraints

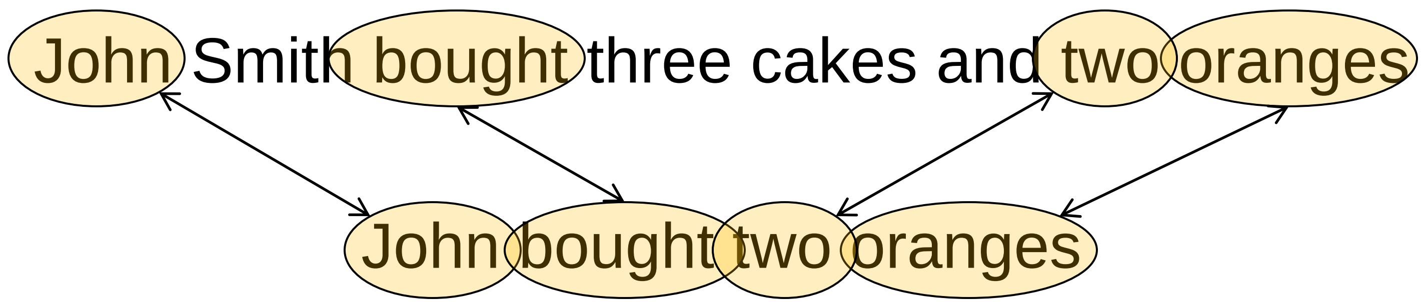
- Real-time web demo:

  http://cogcomp.cs.illinois.edu/demo/relation/

TEXTUAL ENTAILMENT SYSTEM

Alignment in RTE: Lexical Level

- **Alignment:** **a mapping from elements in the Hypothesis to elements in the Text**



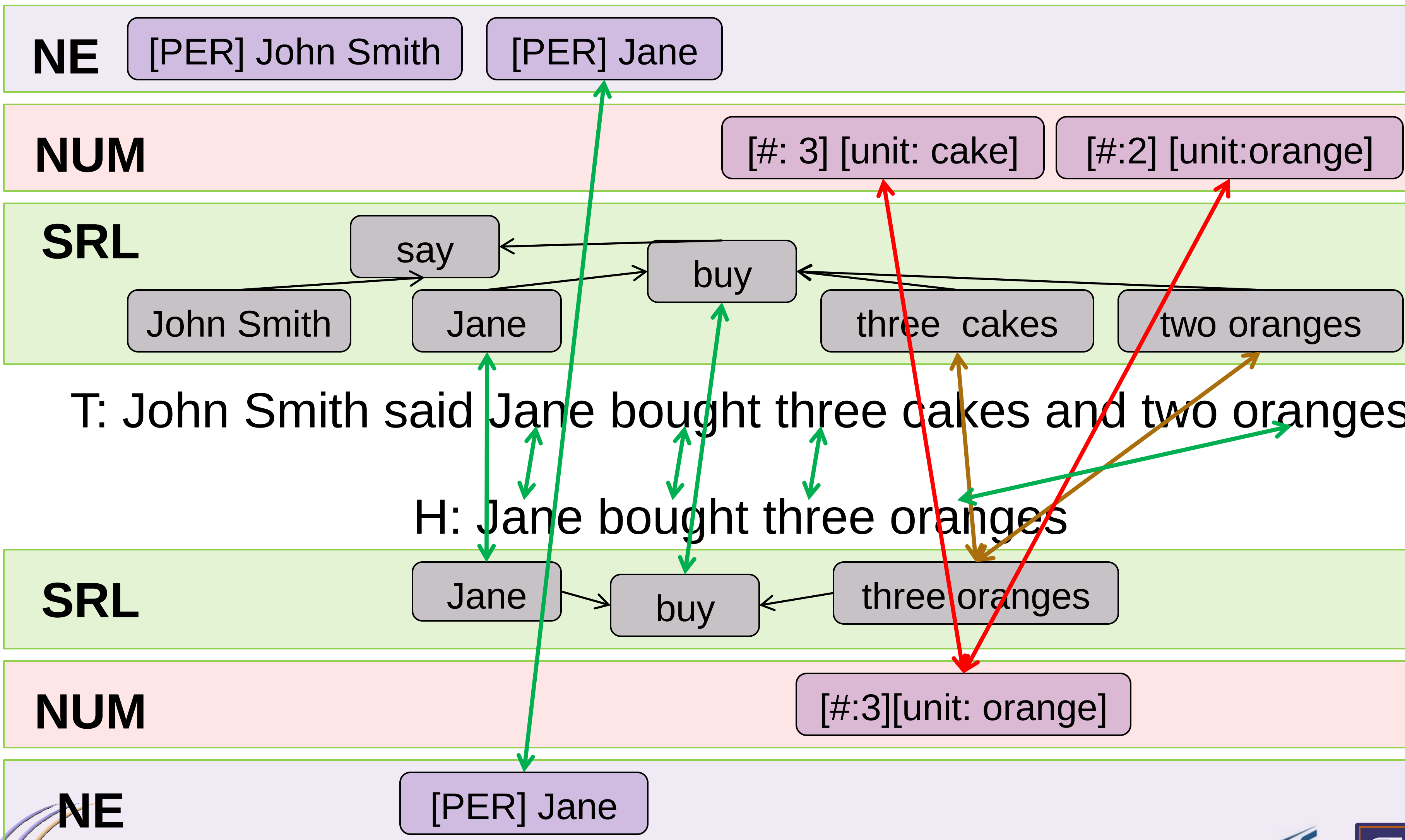
Alignment is Useful for Machine Learning in RTE

- **Machine Learning approaches provide much-needed robustness for NLP tasks**
- **RTE data sets are small, given complexity of problem**
- **Global, 2- or 3-class label on each pair**
- **We would like to resolve entailment by combining local decisions (e.g. word-level, phrase level); but *which* decisions?**
- **Alignment can be used to select a subset of the many possible comparisons, and thereby augments global label with (proxy for) finer-grained structure; can be used…**
  - **…to determine active features**
  - **…to generate labels for local classifiers**

Multiple alignments at multiple granularities

- **Intuition: exploit differences/agreements between different views of the entailment pair; avoid canonization**
- **Accommodates analysis at different granularities**
- **Resources with comparable scores can compete with each other – pick the "best"**
  - e.g. Words, Multi-word Expressions, Phrasal Verbs
- **Unscaled resources occupy different alignments (SRL, NE)**
- **Metrics can return negative numbers; use magnitude in alignments, preserve negative edge label**
  - **May be useful for contradiction features**

Multiple Alignments for RTE



Learning from Multiple Alignments

- **Extract features based on individual alignments**
  - **Can use high-precision, low-recall resources as filter features**
  - **Typical match features within alignments – e.g. proportion of tokens matched**
- **Extract features based on agreement, disagreement between different alignments**
  - **E.g. Predicate-Argument, Numerical Quantities**
- **Allows graceful degradation if some resources are unreliable; learner assigns low weights to corresponding features**

Multiple Alignments ctd.

- **Model each alignment as optimization problem**
  - **Penalize distant mappings of neighboring constituents in H, T (proxy for deep structure – favor chunk alignment)**
  - **Constraints: each token in H can be covered exactly once by an aligned constituent; edge scores must account for number of constituents covered**
  - **Solve by brute-force search**

$$\frac{1}{m}\left[\sum_i e(H_i, T_j) + \alpha . \sum_i \Delta(e(H_i, T_j), e(H_{i+1}, T_k))\right]$$

$$\sum_j I[e(H_i, T_j)] \leq 1$$

Feature Extraction

- **Main types of features:**
  - **Features assessing quality of alignment in a given view**
  - **Features assessing agreement between views**
- **Quality of Alignment features:**
  - **Proportion of constituents matched in Word, NE, SRL views**
  - **"Distortion" of match pattern**
- **Agreement features:**
  - **Proportion of token alignments agreeing with SRL constituent alignments**
  - **Negation of predicate in SRL relation match**
- **Extension: Using Coreference:**
  - **Augment SRL predicates: add arguments using Coref chains**
  - **Introduces inter-sentence structure**

Results

Corpus	System					
	Baseline	**No NE***	**Basic NE**	**No WN**	**All***	**All + Coref**
RTE5 Dev	0.628	0.640	0.623	0.647	0.648	**0.663**
RTE5 Test	0.600	0.629	0.633	0.603	0.644	**0.666**

* Submitted runs had ~60 buggy alignments in dev test; results using non-buggy alignments shown here

RTE system demo

- Note: Where are the transformations?
  - We found that chaining offered little gain while significantly complicating the architecture
  - We use transformation rules as mappings between predicate-argument structures in the SRL Comparator

- Real-time web demo:

  http://cogcomp.cs.illinois.edu/cgi-bin/rte/entailment.py

Can we do better?

- Presently, we heuristically align our representations of Text and Hypothesis to reduce the problem complexity and make learning tractable
- Even if we use machine learning for alignment, a pipeline architecture leads to error propagation
- Alternative: “indirect supervision”
  - Specify space of alignments, and a feature-based representation for it
  - Use binary RTE labels to optimize alignment that gives best performance on binary task
- A way to learn “purposefulness”?

Chang et al. 2010

- Apply indirect supervision approach to RTE and other tasks
- Use unified graph based on same input representation as fixed alignment system
- Specify match features for nodes (based on similarity score), edges, and node deletion
- Specify constraints on matching edges
  - Edge can only match if source/sink nodes are also matched
- Goal:
  - learn weights on node/edge match features such that…
  - The highest-scoring alignments for entailment pairs…
  - Yield maximum performance when used to decide binary entailment label (using threshold)

Indirect Supervision for RTE (cont'd)

- Optimization for alignment: needs a key insight
  - The *best* alignment for a negative example is "not good enough" (maximum alignment-based score should be low)
  - A positive entailment example has *at least one good alignment* (maximum alignment-based score exceeds some threshold)
- Procedure: for each example
  - Find best alignment using current hypothesis
  - Predict entailment label
  - If prediction is incorrect, update alignment feature weights
- Results: comparable to two-stage architecture

Summary

- We take a compositional approach to textual inference
  - Multi-view representation/architecture
  - Annotator/comparator pairing
- We are trying to build components that isolate specific knowledge domains, but are easy to use
  - Simple functional interface (metrics)
  - Goal: consistent API
- We are using Wikipedia as a broad-coverage general knowledge resource
  - Developed Wikifier, TAREC
  - Currently, trying to integrate them with NLP tools like Co-reference Resolver and RTE system
- Many of our tools have live demos; many are available…

References

- Annie Zaenen, Lauri Kartunnen, and Richard Crouch, “Local textual inference: can it be defined or circumscribed?”, EMSEE (2005)
- Christopher D. Manning, “Local Textual Inference: It's hard to circumscribe, but you know it when you see it - and NLP needs it.”, MS, Stanford University (2006)
- Quang Do and Dan Roth, “Constraints based Taxonomic Relation Classification”, *EMNLP* (2010)
- Lev Ratinov, Doug Downey, Mike Anderson, Dan Roth, “Local and global algorithms for disambiguation to Wikipedia”, *ACL* (2011)
- V. Punyakanok, D. Roth and W. Yih, “The importance of syntactic parsing and inference in semantic role labeling”, *Computational Linguistics* (2008)

References (cont'd)

- E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution", *EMNLP* (2008)
- Q. Do, D. Roth, M. Sammons, Y. Tu, V. Vydiswaran, "Robust, light-weight approaches to compute lexical similarity", *Computer Science Research and Technical Reports, University of Illinois* (2009)
- M. Sammons, V. Vydiswaran, T. Vieira, N. Johri, M. Chang, D. Goldwasser, V. Srikumar, G. Kundu, Y. Tu, K. Small, J. Rule, Q. Do and D. Roth, "Relation alignment for textual entailment recognition", *Text Analysis Conference (TAC)* (2009)