RECOGNIZING TEXTUAL ENTAILMENT

- a tutorial

MARK SAMMONS

IDAN SZPEKTOR

V.G. VINOD VYDISWARAN

The Holy Grail of NLP….

- Understanding Natural Language Text

British left waffles on nukes

  - **Traditional approach: map it to a canonical form**
  - **Can then (in theory) integrate multiple statements from diverse sources to derive "new" facts**
  - Question #1: How to represent its meaning?

$$\exists_x \exists_y \exists_z British(x) \wedge Waffles(y) \wedge Nukes(z) \wedge leave\_on(x, y, z)$$

  - Question #0.5: What ***is*** its meaning?
  - Question #0.1: What does ***understand*** mean?

UNDERSTANDING

British left waffles on nukes

- Canonical NLP task: search
- If I can reliably say when a document matches, it is sufficient…

UK citizens put breakfast food on weapons 

UK citizens deposited nuclear material in Belgium 

Why we think you're here

- You are "interested" in RTE
- Presumably, want to know
  - What is the task?
  - Why is it worth my attention?
  - What have people done in RTE (that is "interesting")?
  - If I want to join in, where should/could I start? (what are the interesting problems within RTE?)

TABLE OF CONTENTS

Part I:

MOTIVATION AND TASK DEFINITION

Section Outline

- Motivation
- Definition
- Task and Evaluation
- Applications

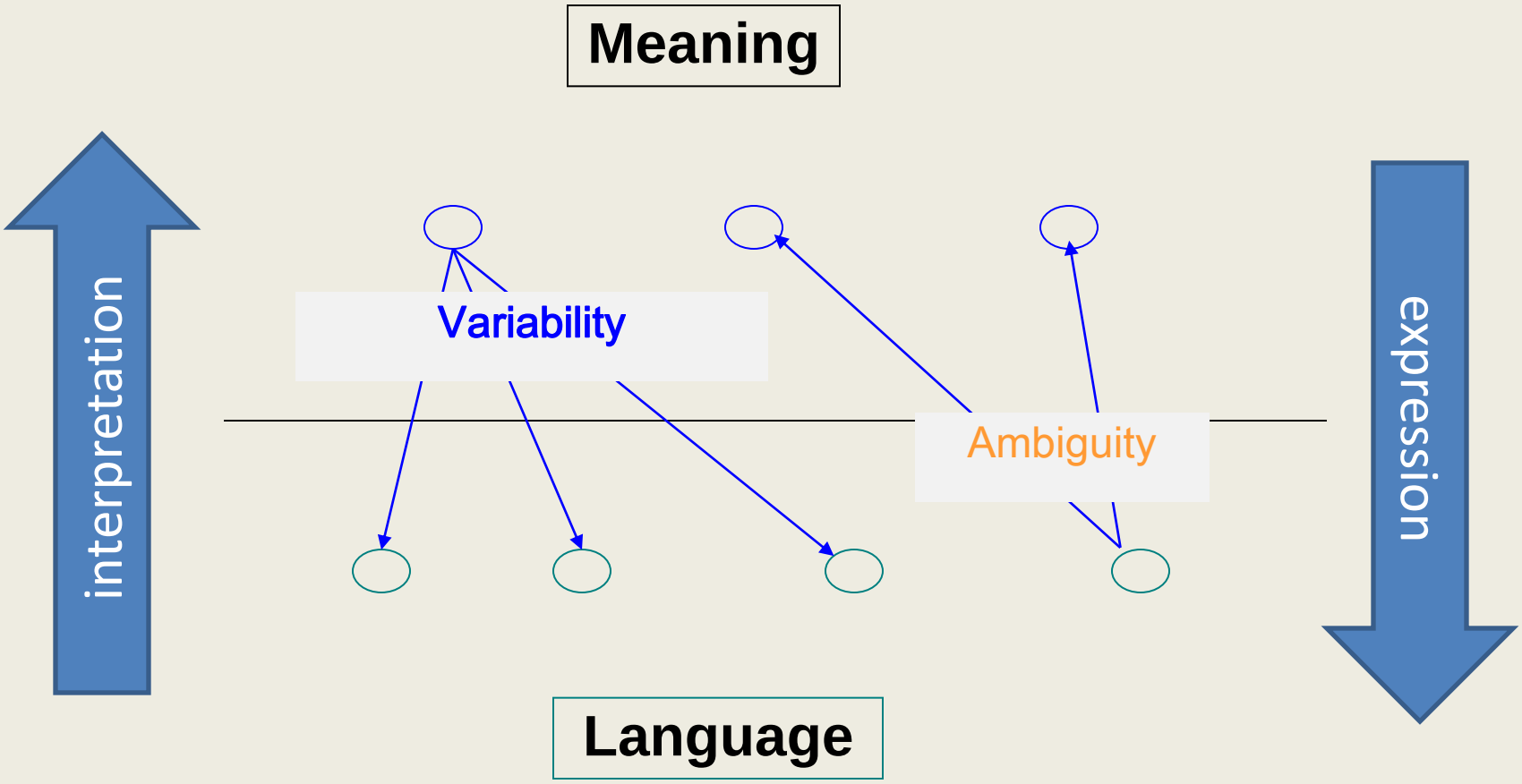
Section Outline

- Motivation
- Definition
- Task and Evaluation
- Applications

MOTIVATION

- Text applications require *semantic* inference
- A common framework for applied semantics is needed, but still missing
- Textual entailment may provide such framework

Natural Language and Meaning

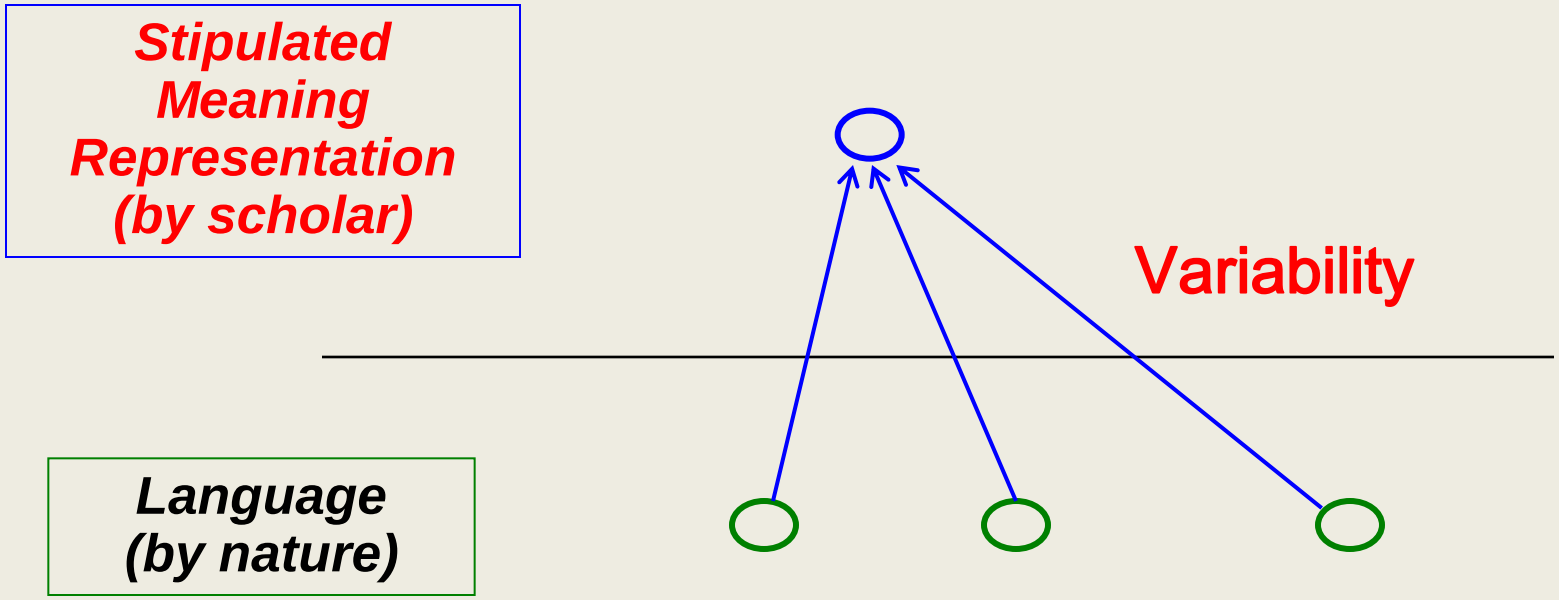


Desiderata for Modeling Framework

- A framework for a target level of language processing should provide:
  - Generic (feasible) module for applications
  - Unified (agreeable) paradigm for investigating language phenomena

- Most semantics research is scattered
  - WSD, NER, SRL, lexical semantics relations…
  - Dominant approach: interpretation
  - What else needs to be done?

Classical approach = interpretation



- Logical forms, word senses, semantic roles, named entity types, … - scattered interpretation tasks
- Feasible/suitable framework for applied semantics?

Variability of Semantic Expression

The Dow Jones Industrial Average closed up 255

Dow ends up

Dow climbs 255

Dow gains 255 points

Stock market hits a record high

Model variability as relations between text expressions:

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

Typical Application Inference: *Entailment*

Question
Who bought Overture? >> *Expected answer form*
X bought Overture

Overture's acquisition by Yahoo

text

⇨

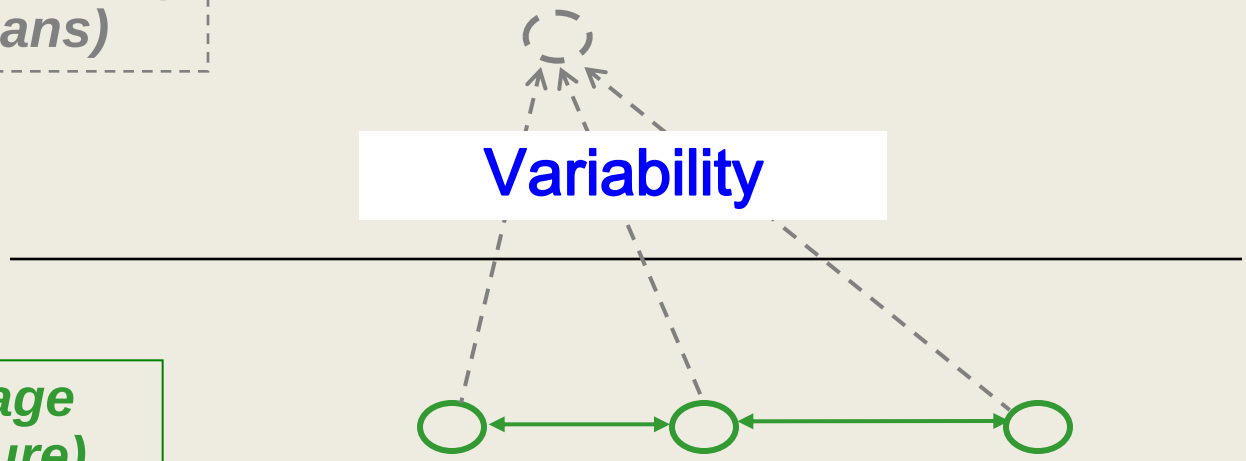
Yahoo bought Overture

hypothesized answer

- Similar for IE: X acquire Y
- Similar for "semantic" IR: t: Overture was bought for …
- Summarization (multi-document) – identify redundant info
- MT evaluation (and recent ideas for MT)

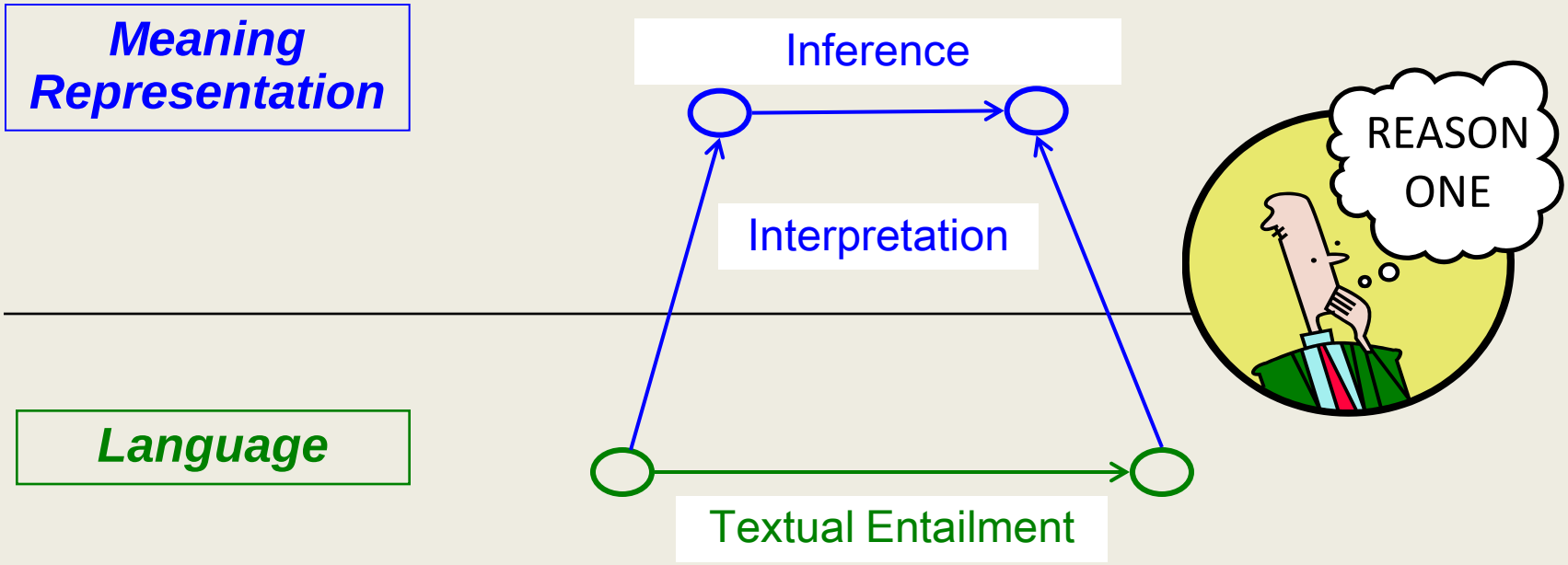
Textual Entailment = Text Mapping

Assumed Meaning (by humans)



Language (by nature)

General Case – Inference



- **Entailment mapping is the actual applied *goal***
  - ***but also a touchstone for understanding!***
- **Interpretation becomes possible *means***
  - **Various representation levels may be investigated**

Section Outline

- Motivation
- **Definition**
- Task and Evaluation
- Applications

APPLIED TEXTUAL ENTAILMENT

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

> ***t entails h*** ($t \Rightarrow h$) **if**
> **humans reading *t* will infer that *h* is most likely true**

- **Operational (applied) definition (Dagan et al., 2006):**
  - **Human gold standard - as in other NLP applications**
  - **Assuming "common background knowledge" – which is indeed expected from applications**

CONTRADICTION

- Definition 2.

  The Hypothesis H of an entailment pair
  contradicts the Text T if
  the relations/events described by H
  are highly unlikely to be true
  given the relations/events described by T.

- Justification: filtering facts from diverse/noisy sources, detecting state changes

EXAMPLE

Text:

> The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

Hyp 1: BMI acquired an American company.

Hyp 2: BMI bought employee-owned LexCorp for $3.4Bn.

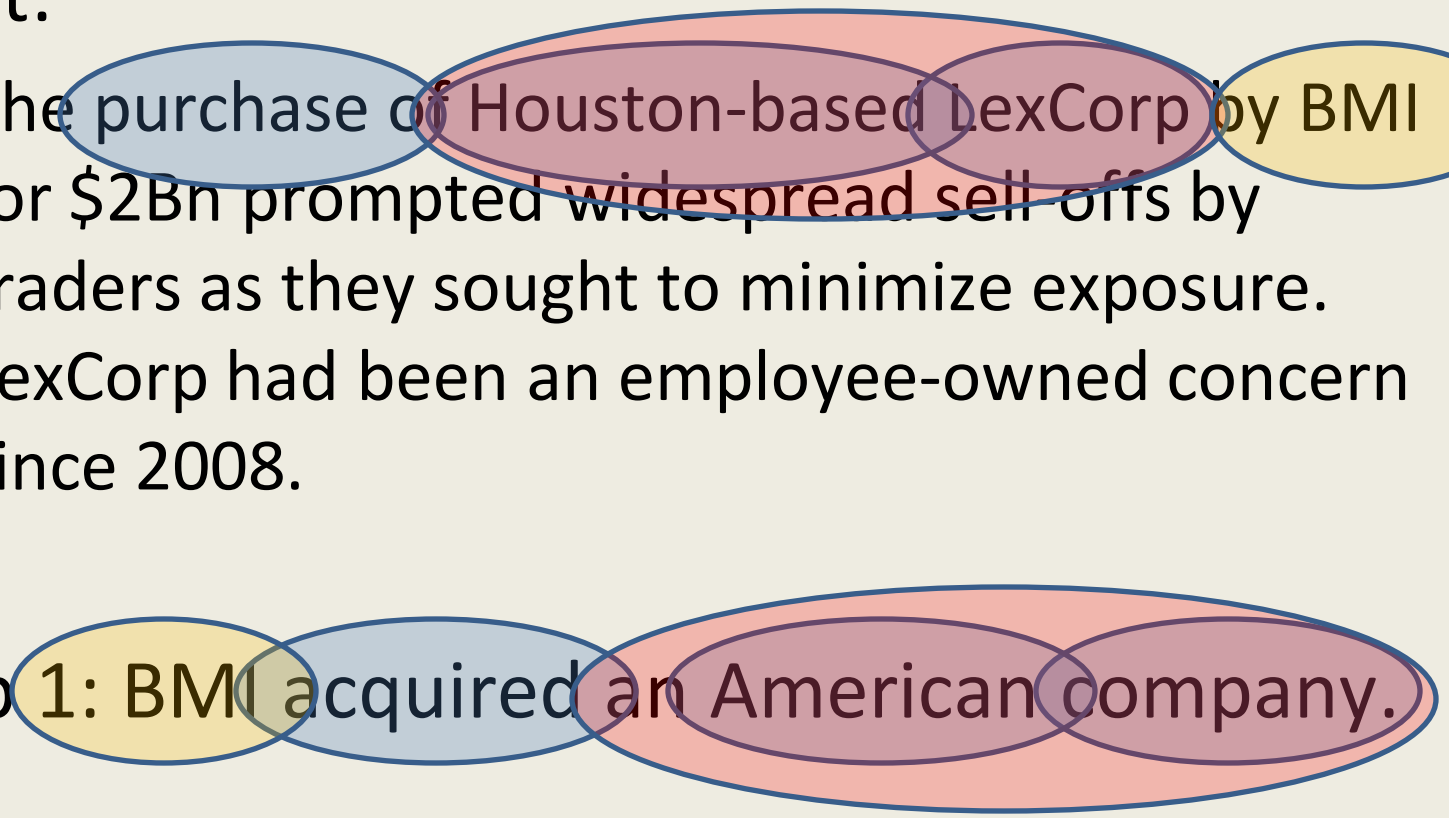
Hyp 3: BMI is an employee-owned concern.

EXAMPLE: ENTAILMENT

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 1: BMI acquired an American company.



EXAMPLE: CONTRADICTION

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 2: BMI bought employee-owned LexCorp for $3.4Bn.

EXAMPLE: UNKNOWN

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 3: BMI is an employee-owned concern.

The Role of Knowledge

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

  but
  - *knowledge* should not entail *h* alone
  - Justification: consider time-dependent information, e.g. PresidentOf(US, X)

- Systems are not supposed to validate *h*'s truth regardless of *t* (e.g. by searching *h* on the web)

[ID: 5T-39 ENTAIL: NO]

TEXT:

…While no one accuses Madonna of doing anything illegal in adopting the 4-year-old girl, reportedly named Mercy, there are questions nonetheless about how Madonna is able to navigate Malawi's 18-to-24 month vetting period in just a matter of days or weeks…

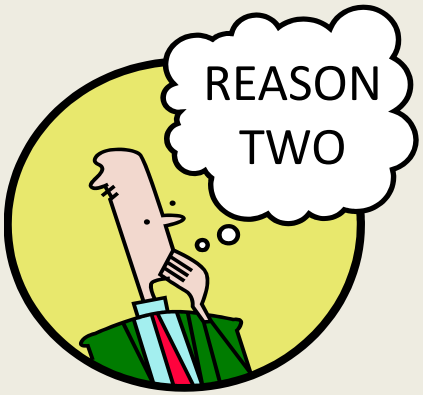
HYPOTHESIS:

Madonna is 50 years old.

Aside: "Why RTE?" Again…

In the previous examples, we needed to integrate knowledge about:

- Named Entities
- Coreference
- Semantic Roles/Syntactic Dependencies
- Nominalization
- Lexical Semantics
- Spatial Inference/Meronymy



Section Outline

- Motivation
- Definition
- Task and Evaluation
- Applications

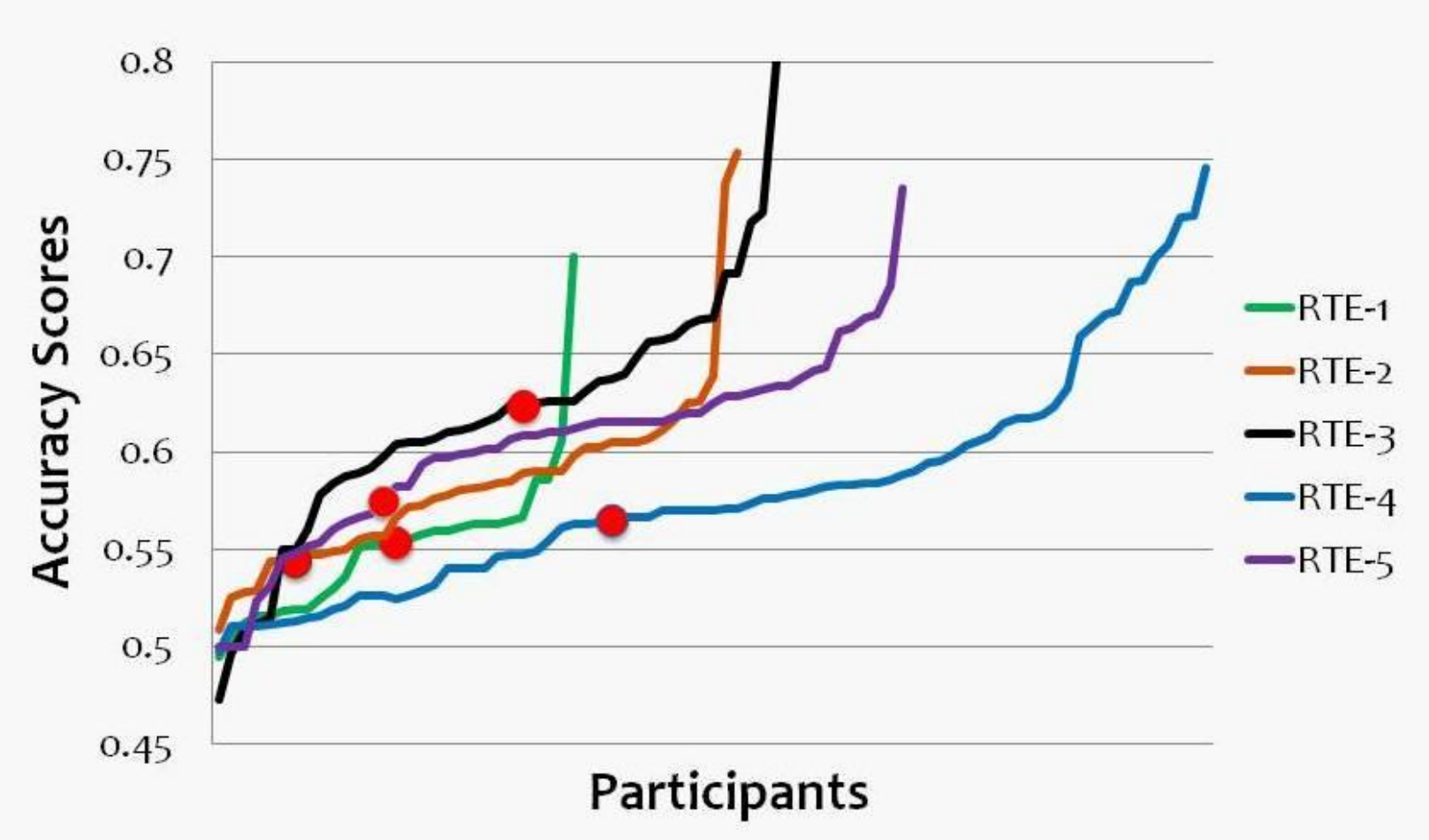
EVALUATION

- Examples drawn from NLP tasks/domains
- **~90% pairwise inter-annotator agreement**
- RTE 1-3: ~800 dev, 800 test RTE pairs each (‘05- ‘07)
  - Boolean label: “entailed” vs. “not entailed”
  - BALANCED data set
- RTE 4-5: Ave. text length = 40, 100 words (‘08, ‘09) respectively, 2-way and 3-way tasks
  - “entailed”, “contradicted”, and “unknown”
- Some pilot RTE task data sets as well
- RTE 6 (2010): shift to application focus: IR-like setting

Evaluation (cont'd)

- **Two measures currently used:**
  - **Accuracy (#correct / #total)**
  - **Confidence Weighted Score (2-way only)**
    - **Rank solutions from most confident positive to uncertain to most confident negative**
- **Typically, not much difference in system ranking between the two measures**
- **(Bergmair 2009) proposes a nice RTE evaluation metric based on Mutual Information…**
- **Relatively high lexical baseline performance is indicative of the difficulty of this task**

How well are we doing?



Section Outline

- Motivation
- Definition
- Task and Evaluation
- Applications

Applications of RTE

- IE (Spektor et al. 2008)
- QA (Hickl et al. 2006; Celikyilmaz et al. 2009)
- IR (Exhaustive applications)
  (Roth et al. 2009)
- MT task (Mirkin et al. 2009);
  MT evaluation (Pado et al. 2009)

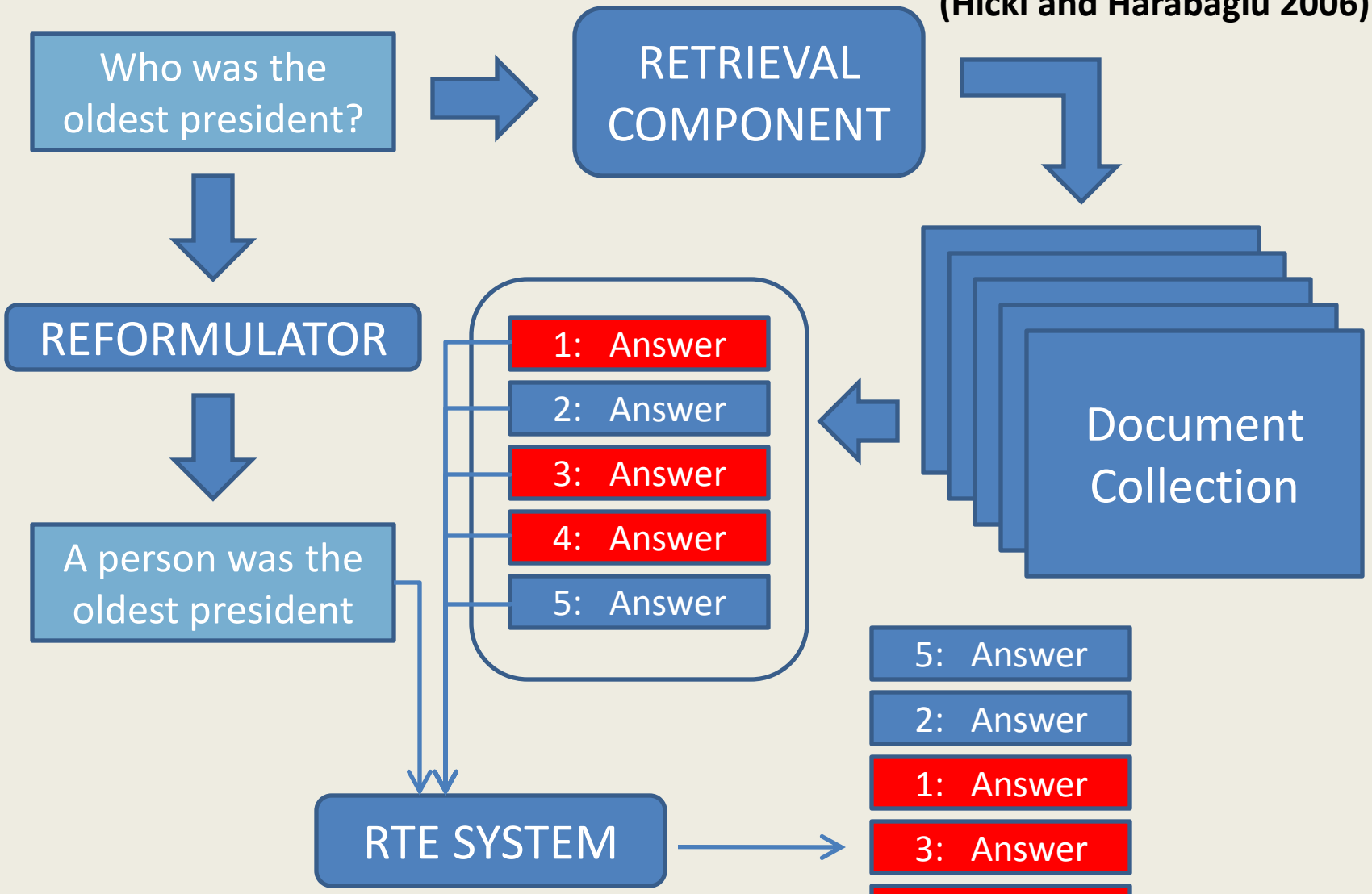
RTE for Question Answering

(Hickl and Harabagiu 2006)

- Question Answering: given a query in natural language, find answers in a document set
- For "factoid" questions, can be thought of as an Information Retrieval problem
  - Question preprocessed, analyzed to focus search
  - IR component returns a set of "answers" – sections of documents likely to contain the answer to the question.
  - QA system assigns them scores – based on e.g. keyword matches, topic match to question category, etc.

QUESTION ANSWERING

(Hickl and Harabagiu 2006)



RTE FOR QA

(Hickl and Harabagiu 2006)

- Intuition: correct answer may not be top ranked, but is often within the top K results
- Re-rank candidate answers using RTE component
- RTE approach: rephrased question is Hypothesis, each candidate answer is a Text
  - **For each Text:**
    If Text |= Hypothesis, push answer to top
- Improves system accuracy from 30.6% to 42.7%

Scalable RTE for Exhaustive Search

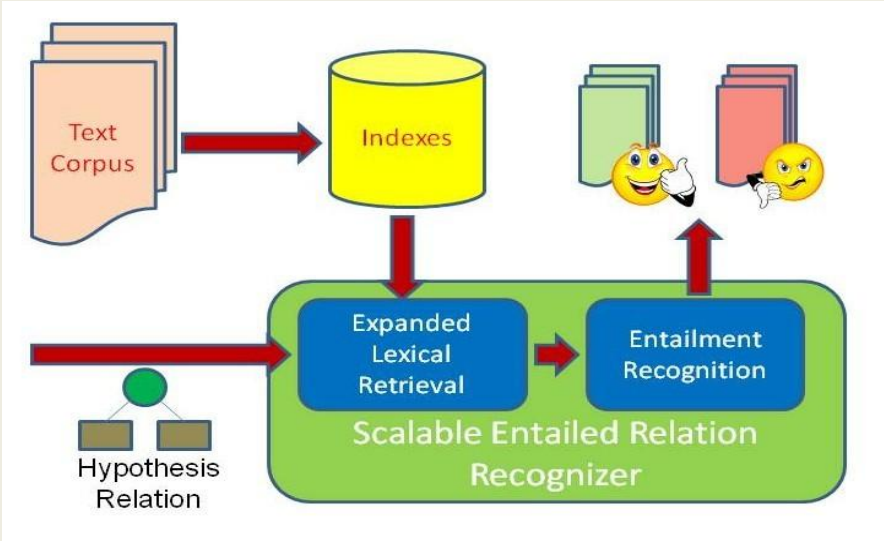
(Roth et al., 2009)

- Target applications like document downgrading (detect classified information): must retrieve ALL instances of specified query
- Artificial corpus generated from IE, IR subtasks from RTE 1-3 (cross-product of H, T from all pairs)
- Two-stage architecture:
  - Push some RTE capabilities into Retrieval step; index shallow semantic markup (NE, NQ, MWE), use similarity metrics in retrieval
  - Post-retrieval RTE step filters results using deeper structure

Scalable RTE cont'd

(Roth et al., 2009)

- "Smart" Retrieval step; index shallow semantic markup (NE, NQ, MWE), use similarity metrics in retrieval
- Post-retrieval RTE step filters results using deeper structure



- Performance on RTE 1-3 evaluation (just considering pairs from RTE 1-3) was among top 3 published results for each
- Reduction from ~3.8M RTE operations (naïve system) to ~14K

RTE FOR INFORMATION EXTRACTION

(Szpektor et al. 2008)

(emphasis: context for rules):

- Use ACE 2005 event corpus, expressed as entailment pairs
- Target relations (Hypotheses) are small set of templates, e.g "X acquire Y" PLUS manually constructed contextual preferences (== "background knowledge")
- Identify when candidate sentences entail target relations
  - Use learned rules with contextual preferences based on DIRT approach (Lin and Pantel, 2001)
- Evaluate ranking of matches, show improvement over non-contextual rule-based approach

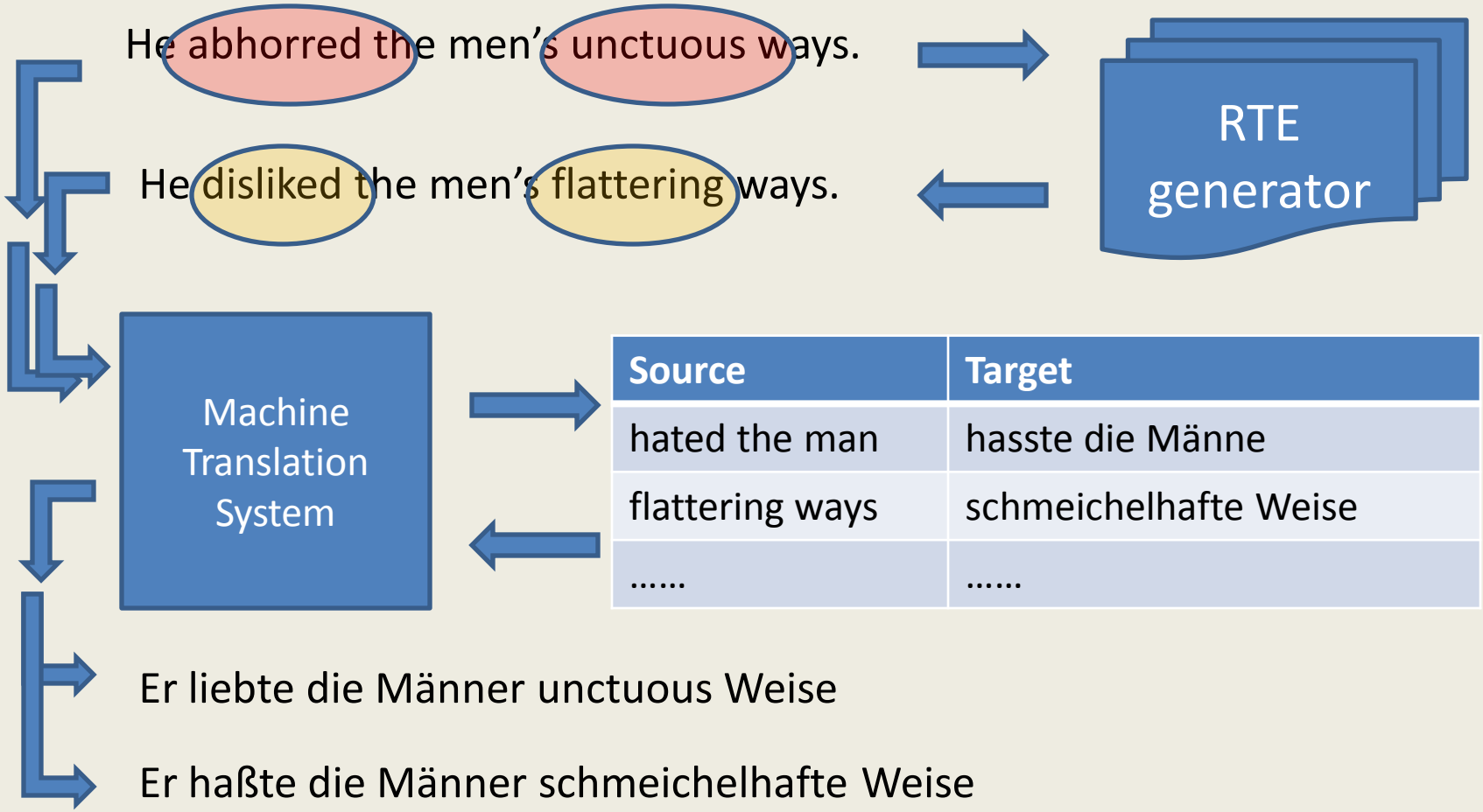
RTE for Machine Translation

(Mirkin et al., 2009)

- Problem: incomplete tables in translation source model (e.g. due to domain shift, data sparseness, scarce source language resources)
- Solution: when unknown terms encountered, generate entailed versions of source language representation using entailment-driven process
  - Intuition: may lose some information (entailed => more general), but still get some reliably translated meaning
- Show significant improvement in recall (~15% raw gain) for only modest loss in precision (~3% raw loss) for proportion of ACCEPTABLE TRANSLATIONS

RTE FOR MT CTD.

(Mirkin et al., 2009)



SUMMARY

- At least three strong reasons to work on RTE:
  - Closely reflects the real task in many NLP applications (text-to-text inference)
  - Links a broad range of existing (and yet-to-be-developed) NLP applications/resources
  - Can be productively applied to other NLP tasks.
- It is (one instantiation of) the grand NLP challenge!

Part II:

ENTAILMENT SYSTEMS IN ACTION

COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

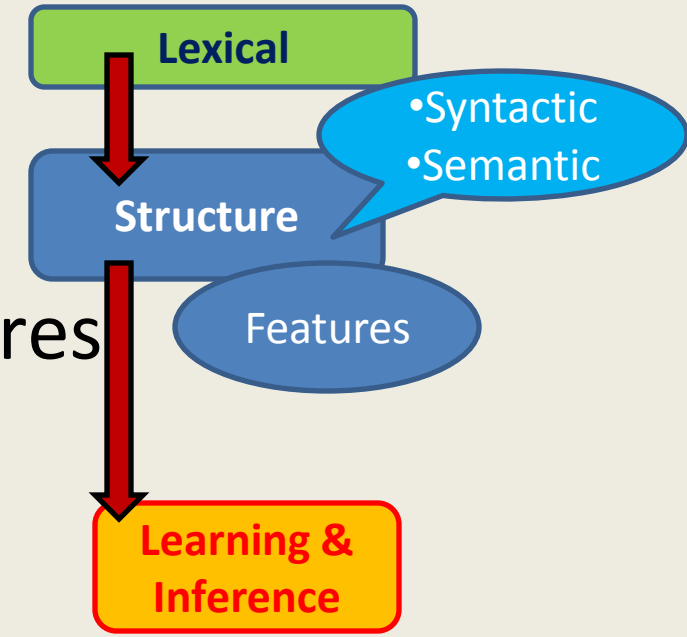
YAHOO! RESEARCH

Current state-of-the-art

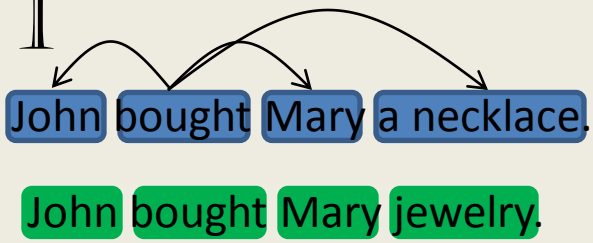
- A sample of distinct approaches
  - Motivated by intuitive RTE process
  - Strengths and weaknesses of each
- Analyze common threads
- Identify common problems
- Propose a framework for thinking about RTE

Approaches Touched Upon

1. Lexical
2. Tree-based similarity
3. Predicate-argument structures
4. Logical form
5. Cross-pair similarity
6. Learning entailment via alignment

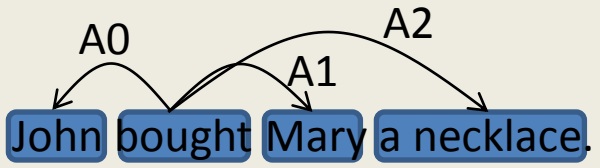


Terminology 1



- Constituent
  - Any induced structure, including (trivially) words
  - E.g. phrase, argument, relation, predicate, parse (sub)tree
  - May comprise multiple (smaller) constituents and edges
- Edge
  - Labeled arc connecting two constituents
  - E.g. “role” in predicate-argument structure

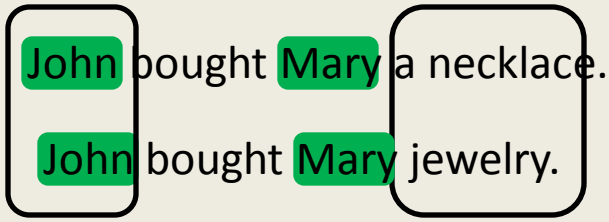
TERMINOLOGY 2



John bought Mary jewelry.

- Predicate-Argument Structure (Relation)
  - Representation of some event in the text, having arity >= 1
  - Predicate is often a verb, but may be expressed in other ways, including completely implicitly
- View
  - The structure(s) induced from a particular annotation source (or set of comparable sources)
  - E.g. Named Entity, or Shallow Parse, or Word, …

Terminology 3



- Annotator (Analytic)
  - A source of analysis for a text span
  - E.g. Named Entity Recognizer; SRL; POS
- Comparator (Metric)
  - A resource comparing two Constituents of a specified type
  - For simplicity, returns score $S \in [-1, 1]$
    - Could return other information also
  - Constituents may have complex structure

Approach 1: Lexical

- Bag-of-words model: words (and possibly NEs) form the lexical constituents
- For each word in H, find "best" word in T
- Normalize scores across sentence-pairs
- Find a threshold to distinguish the good matches from the bad matches

EXAMPLE: CONTRADICTION/ENTAILMENT

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 2: BMI bought employee-owned LexCorp for $3.4Bn.

  under

How to measure similarity…

- bought ⇔ purchase
- How to compare numeric quantities?
  - "$2 Bn" and "under $3.4 Bn"

Solution:

- Define similarity metrics between words and NEs
  - word similarity based on WordNet
  - NE similarity based on rules (acronyms, abbreviated first names, etc.)
- Similarity metrics for Numeric Quantities (NQs)
  - Tokenize, find units, and compare

Other Questions to Answer…

- bought (v) ⇔ purchase (n)
- How to compare numeric quantities?
  - "$2 Bn" and "under $3.4 Bn"

- Is "for" as important as "BMI"?
- How to tokenize?
  - "employee-owned" vs. "employee" "-" "owned"
- Which "LexCorp" to choose?
  - More importantly, which Text word to choose?
- How to threshold the similarity score?

Local Lexical Matching (LLM)

- Greedy approach to match words in Hyp to words in Text
  - The one with highest similarity score wins
  - Tie resolved randomly (does not matter)
  - A text word may be closest to multiple Hyp words
  - Score normalized by size of Hyp
- Similarity metrics for words, NEs, and NQs
  - NE and NQ similarity over multiple words
- Threshold value learnt over dev set (simple LTU)

Error Cases – (1) More Knowledge

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 1: BMI acquired an American company.

Error Cases – (2) More Structure

- Text:

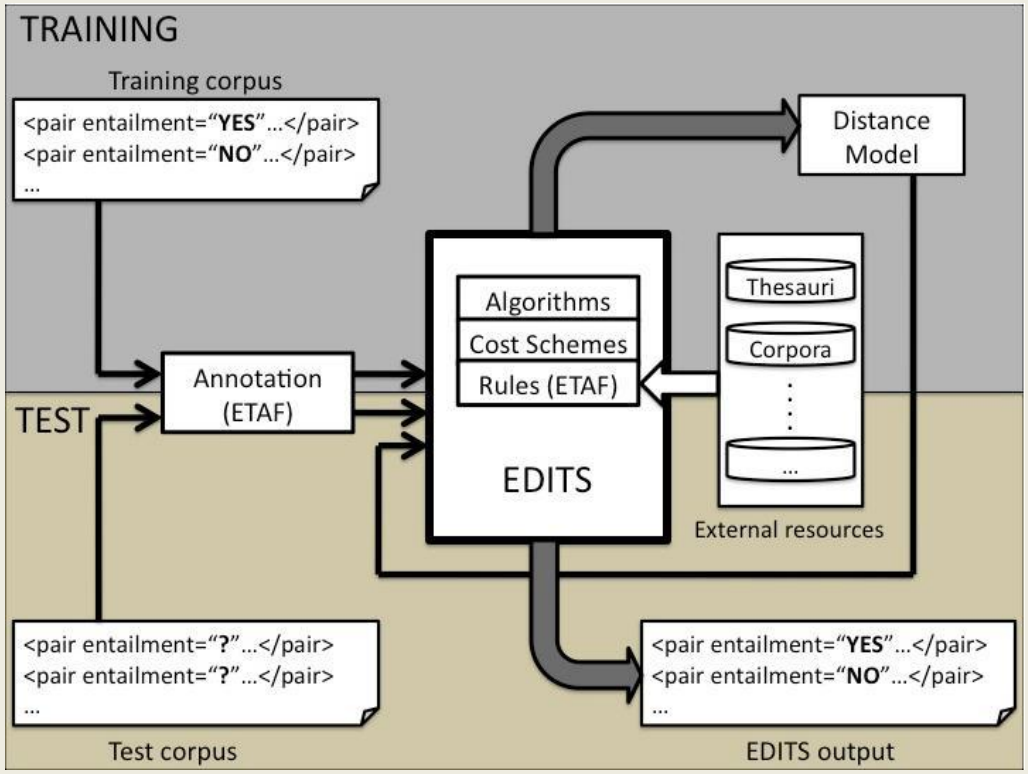
  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 3: BMI is an employee-owned concern.

Approach 2: Tree Similarity

- Basic tree edit distance – not too successful
- Needs to be combined with other token distance metrics
- FBKIrst defined a framework based on edit distances over string, tokens, and tree-level
  - Achieved accuracy of 60.2% on RTE5 2-way
  - 71% in EVALITA 2009 (Italian RTE competition)

EDITS FRAMEWORK



- EDITS: Edit Distance Textual Entailment Suite
- Three modules
  - Edit distance algorithm
  - Cost scheme
  - Rules (from Wikipedia)
- Learn threshold function and weights for edit operations (insertion, deletion, substitution)
- Inference: simple threshold unit based on lowest cost edit-distance operations

EDITS DETAILS: EDIT DISTANCE

- String Edit distance
  - Edit ops on characters
  - Levenshtein distance
- Token Edit distance
  - Edit ops over sequence of tokens in Text and Hyp
  - Levenshtein over tokens
- Tree Edit distance
  - Edit ops over nodes of syntactic representation of Text and Hyp
  - Zhang-Shasha algorithm

EDITS DETAILS: RULES

- Rules for Entailment and Contradiction
- Text units → Hyp units (with some prob.)
- Derived from
  - WordNet (hyponymy, synonymy): 2700 rules
  - VerbOcean: 18K rules on "stronger than" relation
  - Wikipedia: relatedness between words using LSI

$$Entailment(T,H) = \frac{ED(T,H)}{(\mathrm{ED(T,\ \_)} + \mathrm{ED(\ \_, H)})}$$

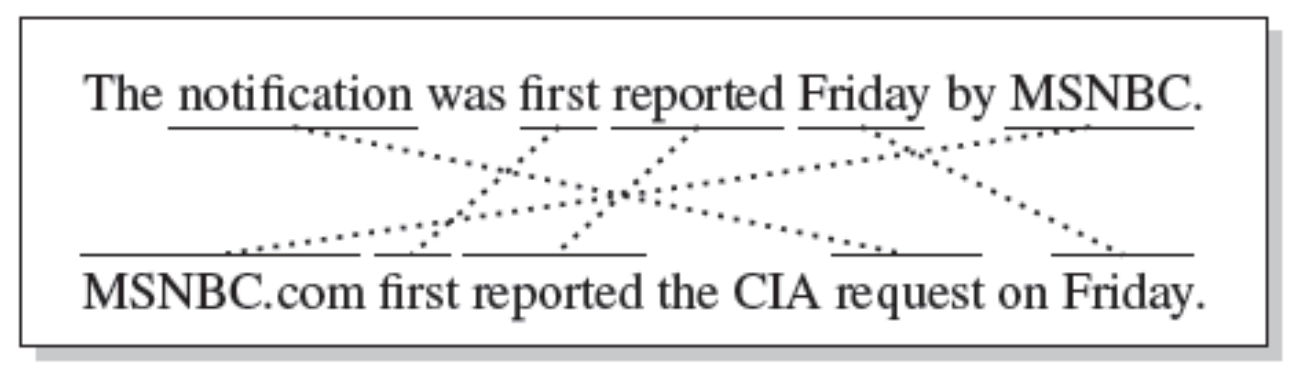
Delete T and Insert H

Approach 3: Predicate-Argument Structures

- Systems in action: DFKI
- Similarity defined over semantic structure of sentences, including arity of relation verbs, core arguments, and sentence passivization.
- Lexical similarity augmented with roles played by semantic units in the pair of sentences
- Shallow Lexical alignment may help focus predicate-argument match to relevant sub-structures

Using Tree Structure as Features

- Learning Constrained Latent Representations (LCLR) framework: Chang, et.al. NAACL 2010
  - Uses declarative Integer Linear Programming (ILP) inference formulation
- Define an intermediate representation: Alignment between Text and Hyp
  - As alignment of NEs, predicates and arguments



Learning over Graph Structure

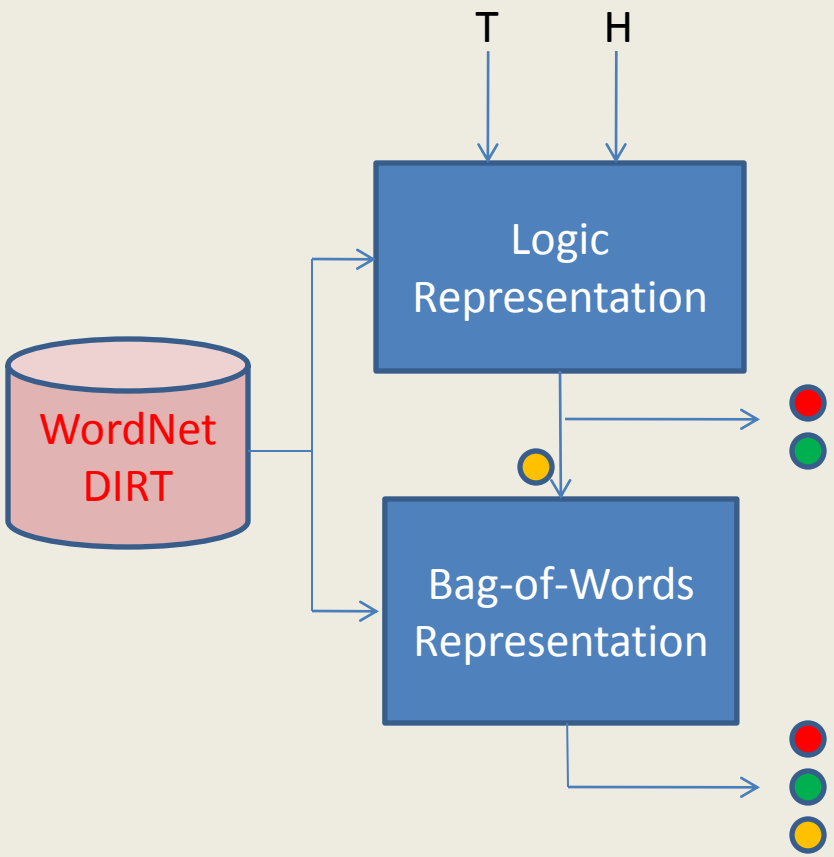
- Text and Hyp are acyclic graphs
  - Words are nodes
  - Directed edges link verbs to the head words of semantic role labeling arguments
  - Dependency edges between words
- Constraints based on word mapping, edge mapping, and word deletion
- Learn features based on hidden alignment structure
- Inference using a trained SVM classifier

Approach 4: Logical Form

- Systems in action: BLUE (Boeing), MITRE, LCC
- Transform Text to logic-based representation
- Infer Hyp using a theorem prover
- Bag-of-words alignment used to backoff
- Includes dependency parsing, POS, Coref
- Uses WordNet and DIRT rules to generate a chain of reasoning from T to H
  - Limited by errors in knowledge sources

BOEING LANGUAGE UNDERSTANDING ENGINE (BLUE)

4/6: Logical Reformulation



- Logic representation of T
  - Parsing
- Try to derive H from T
  - Using Logic, WordNet and DIRT
- If entailment/contradiction, output result and reasoning
- Back-off to BOW model, ignoring syntax structure
- 61.5% in RTE5 2-way

BLUE: LOGIC MODULE

- Parse T using a bottom-up chart parser, SAPIR
- Generates a Logic Form (LF) – a normalized tree structure with variables for NPs, constituents

  LF for *"A soldier was killed in a gun battle."*

```
(DECL
  ((VAR _X1 "a" "soldier")
  (VAR _X2 "a" "battle" (NN "gun" "battle")))
  (S (PAST) NIL "kill" _X1 (PP "in" _X2)))
```

- Includes some disambiguation (e.g. POS)
- Converted to logic representation of assertions

```
object(kill01,soldier01)
in(kill01,battle01)
modifier(battle01,gun01)
```

APPLYING BLUE TO RTE

- Subsumption (≥)
  - If representation of H subsumes (is more general than) T
    - "A person likes a person" ≥ "A man loves a woman"
- Syntactic predicate match
  - Structural/Lexical
    - active - passive, modifiers, rule-based (on and onto)
  - WordNet
    - synonyms, hypernyms across POS
    - similar, pertains, and derivational
  - DIRT rules, esp. verb paraphrases
    - (X rel_1 Y) → (X rel_2 Y)

similar-to (speedy#s2, fast#a1)
pertains-to (rapidly#r1, quick#a1)
derives (destroy#v1, destruction#n1)

ERROR ANALYSIS

- WordNet
  - T: ...Japanese capital of Tokyo...
    H: Tokyo is the capital of Japan.
  - T: Clarkson died...
    H: Actress Lana Clarkson killed...

pertains-to (Japanese#a1, Japan#n2)

killing#n2 ≥ death#n7

- DIRT
  - T: The U.S. holds about 240 men at the U.S. base in Cuba...
    H: About 240 people are detained in Guantanamo.
  - T: A man hijacked a passenger plane in the Jamaican resort of Montego Bay...
    H: A plane crashed in the Jamaican resort of Montego Bay.

IF Y is held by X THEN Y is detained by X

IF Y is hijacked in X THEN Y crashes in X

SUMMARY OF BLUE

- Logic representation helps in ~ 30% cases
  - Accuracy relatively good (~ 63.5%)
- DIRT rules have low coverage (~ 10 – 15%)
- Only ~ 50% DIRT rules sensible
- Need for additional "knowledge" resources
  - "Slumdog Millionaire" is a movie.
- Syntactic knowledge alone didn't help
  - Error-prone preprocessing
  - implicit structure in both T and H, not discriminative
- Hypotheticals not addressed yet (even if X → X)

Approach 5: Cross-pair similarity

- Zanzotto et.al. proposed approach based on syntactic tree kernels
- Define similarity between pairs of sentences using modified dependency tree repr.
  - Nodes abstract the syntactic units
  - Anchor the matching with lexical alignment
  - Generalize anchors to semantic units and learn higher-level patterns
- SVM learnt over inter-pair distance measures
- Similarity functions also incorporates Wikipedia

OVERALL IDEA

(Zanzotto, Moschitti, 2006)

Cross-pair similarity

$K_S((T',H'),(T'',H'')) \approx K_T(T',T'') + K_T(H',H'')$

How do we build it:

- Using a syntactic interpretation of sentences
- Using a similarity among trees $K_T(T',T'')$: this similarity counts the number of subtrees in common between T' and T''

This is a syntactic pair feature space

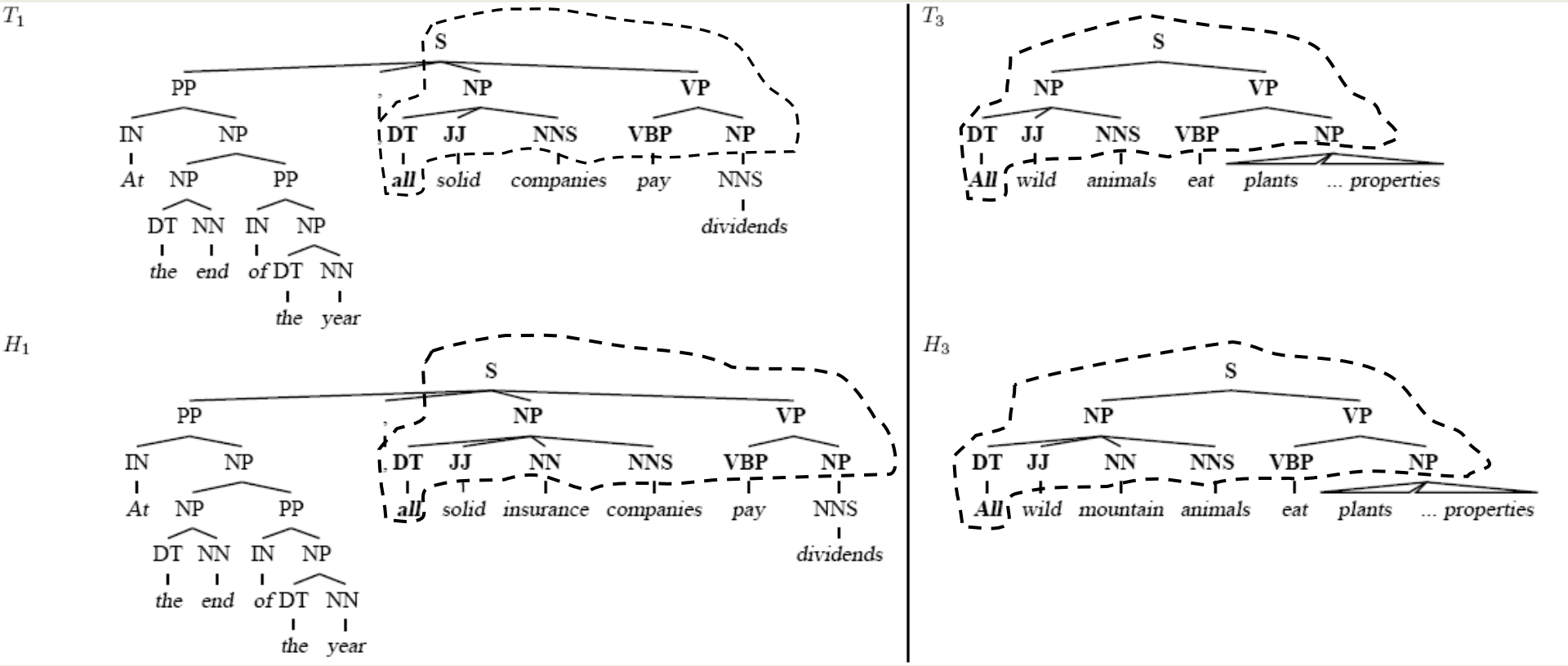
Page 69

OBSERVING THE SYNTACTIC PAIR FEATURE SPACE

(Zanzotto, Moschitti, 2006)

Slides from Fabio Zanzotto et.al.'s talk in EMNLP 2009.

Can we use syntactic tree similarity?

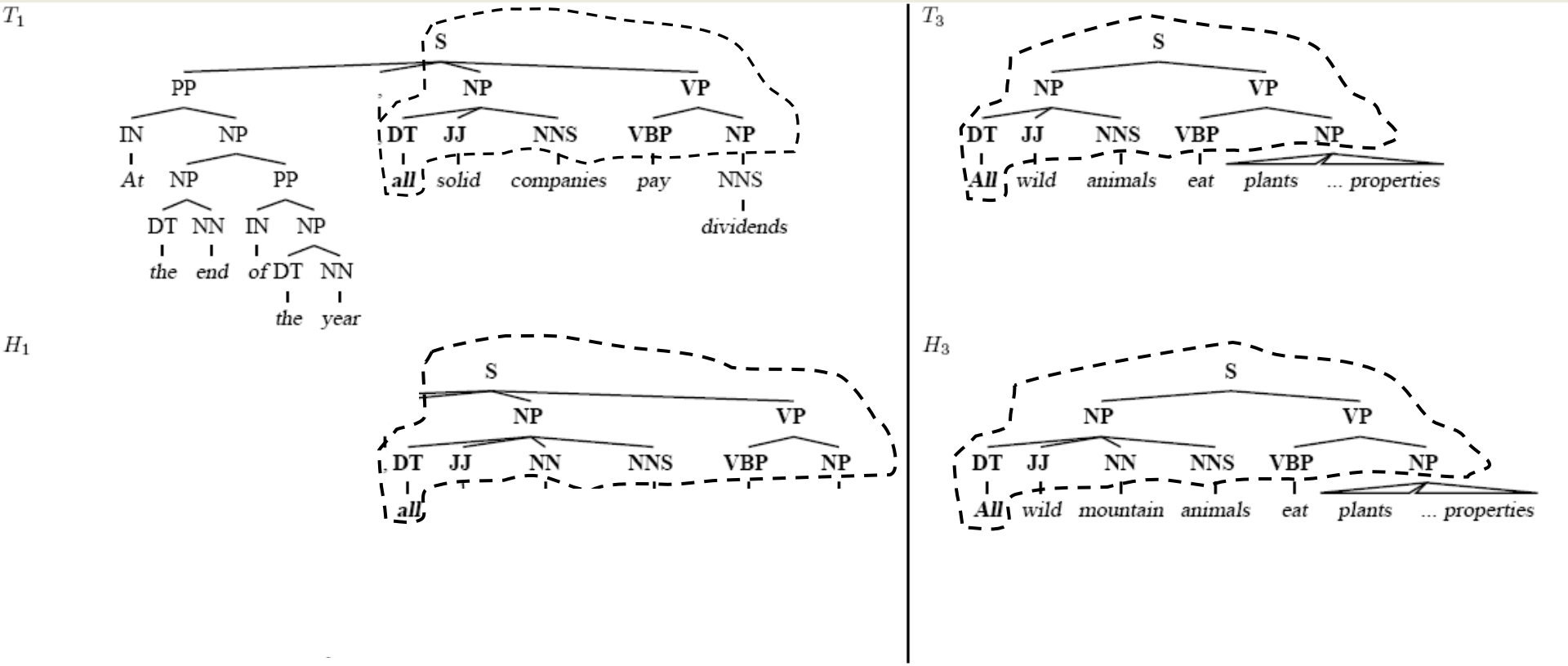


Page 70

OBSERVING THE SYNTACTIC PAIR FEATURE SPACE

(Zanzotto, Moschitti, 2006)

Can we use syntactic tree similarity?



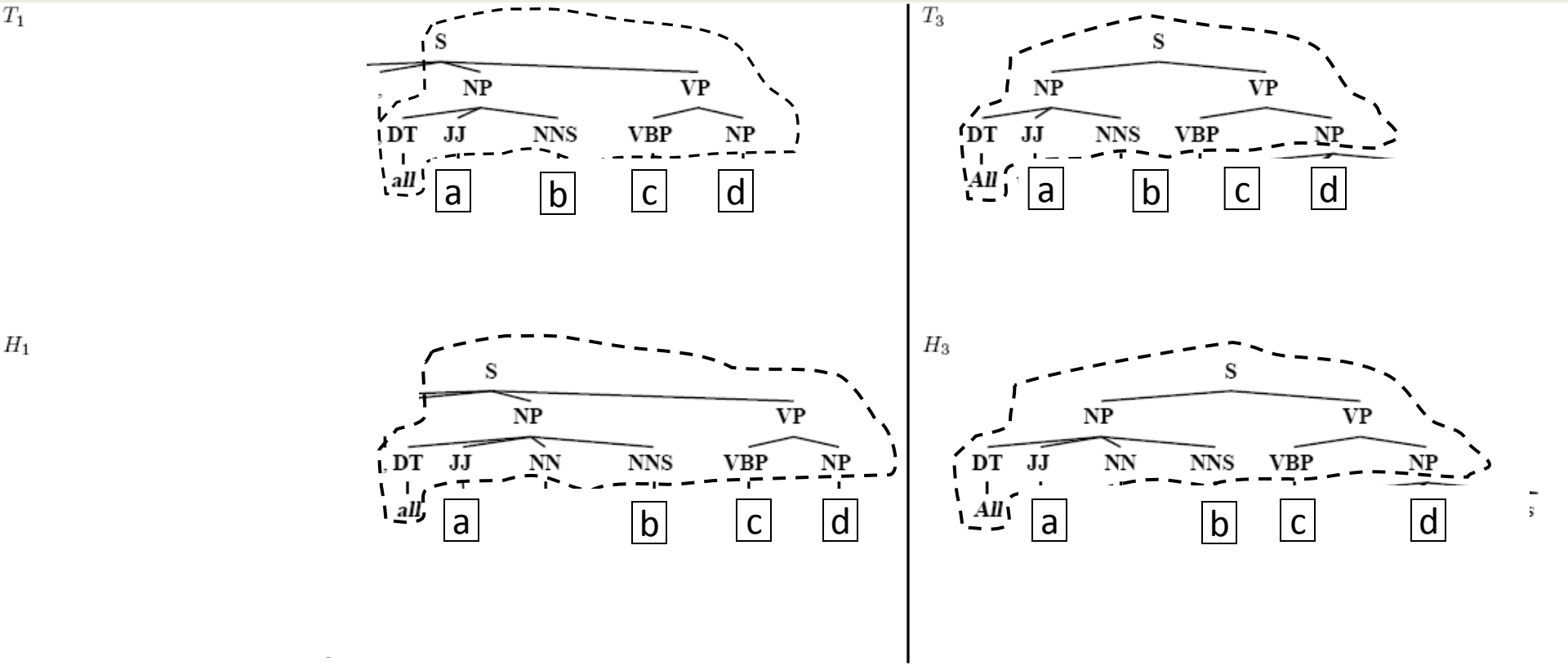
Page 71

OBSERVING THE SYNTACTIC PAIR FEATURE SPACE

(Zanzotto, Moschitti, 2006)

Can we use syntactic tree similarity? Not only!

Implied structures can lead to rewrite rules



Page 72

EXPLOITING REWRITE RULES

(Zanzotto, Moschitti, 2006)

To capture the *textual entailment recognition rule* (*rewrite rule* or *inference rule*), the *cross-pair similarity* measure should consider:

- the *structural/syntactical* similarity between, respectively, *texts* and *hypotheses*
- the *similarity among the intra-pair relations* between constituents

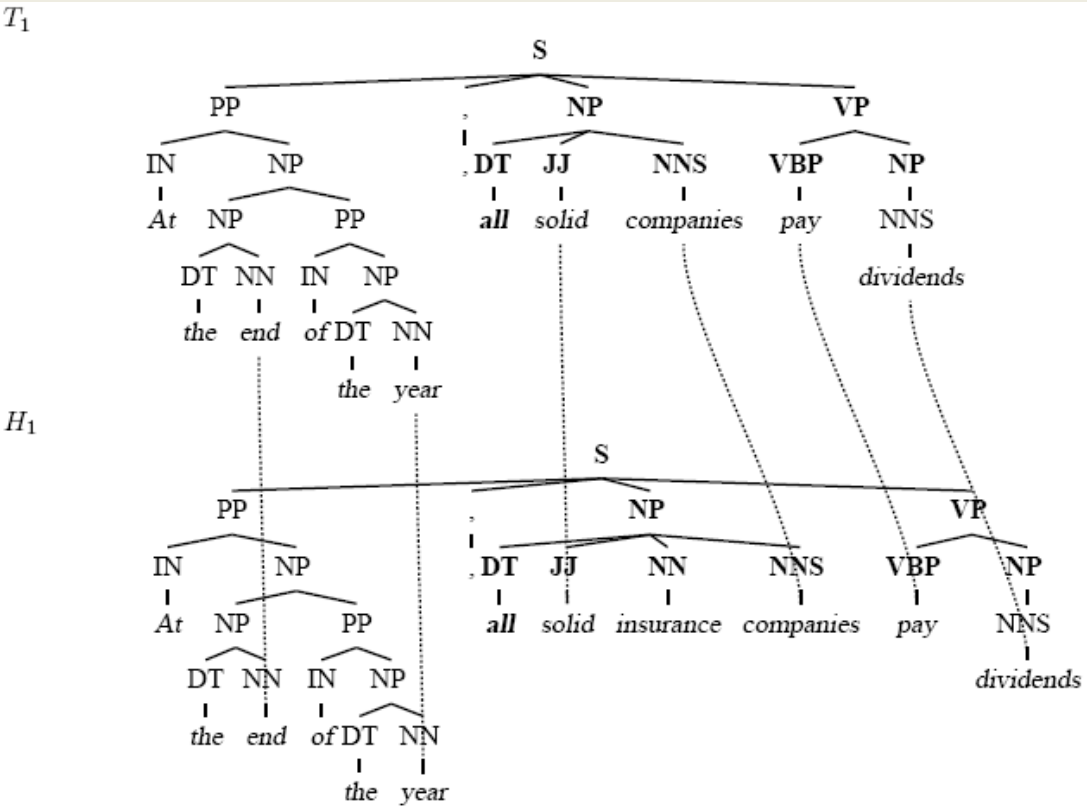
Page 73

EXPLOITING REWRITE RULES

(Zanzotto, Moschitti, 2006)

Intra-pair operations

→ Finding ***anchors***



Page 74

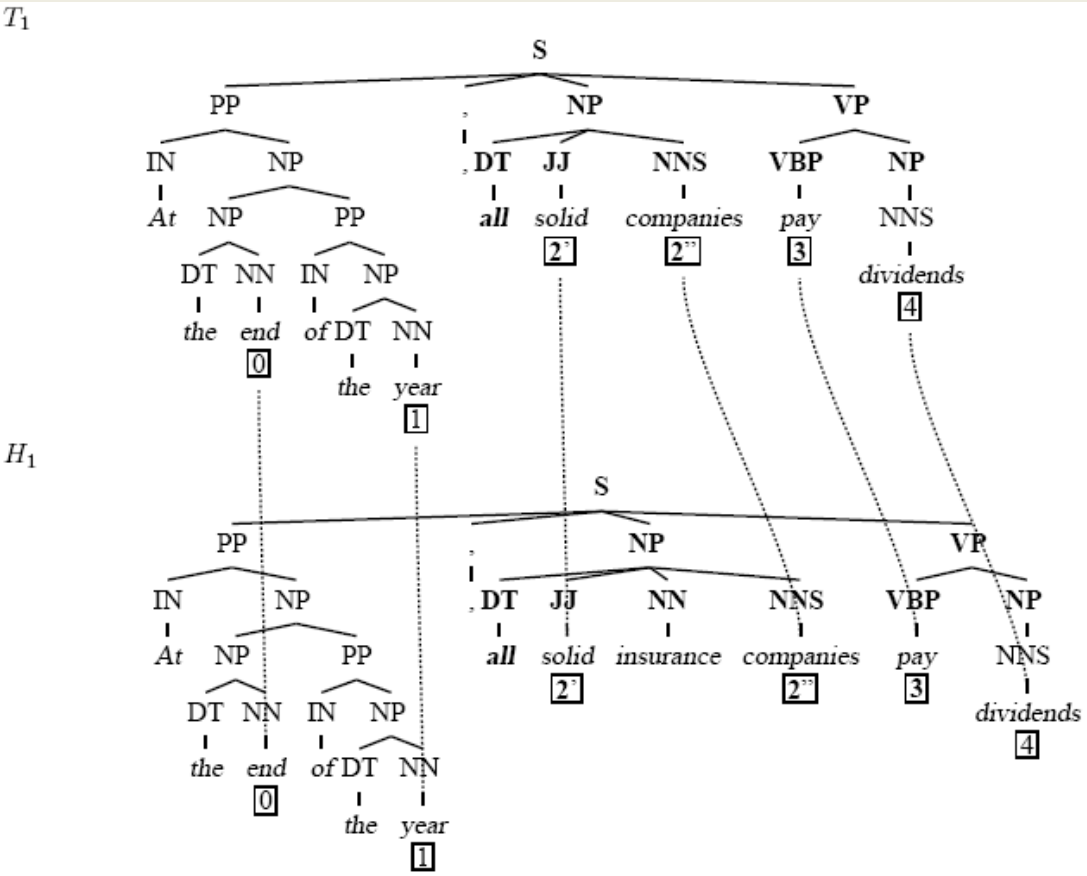
EXPLOITING REWRITE RULES

(Zanzotto, Moschitti, 2006)

Intra-pair operations

→Finding ***anchors***

→Naming anchors with ***placeholders***



Page 75

EXPLOITING REWRITE RULES

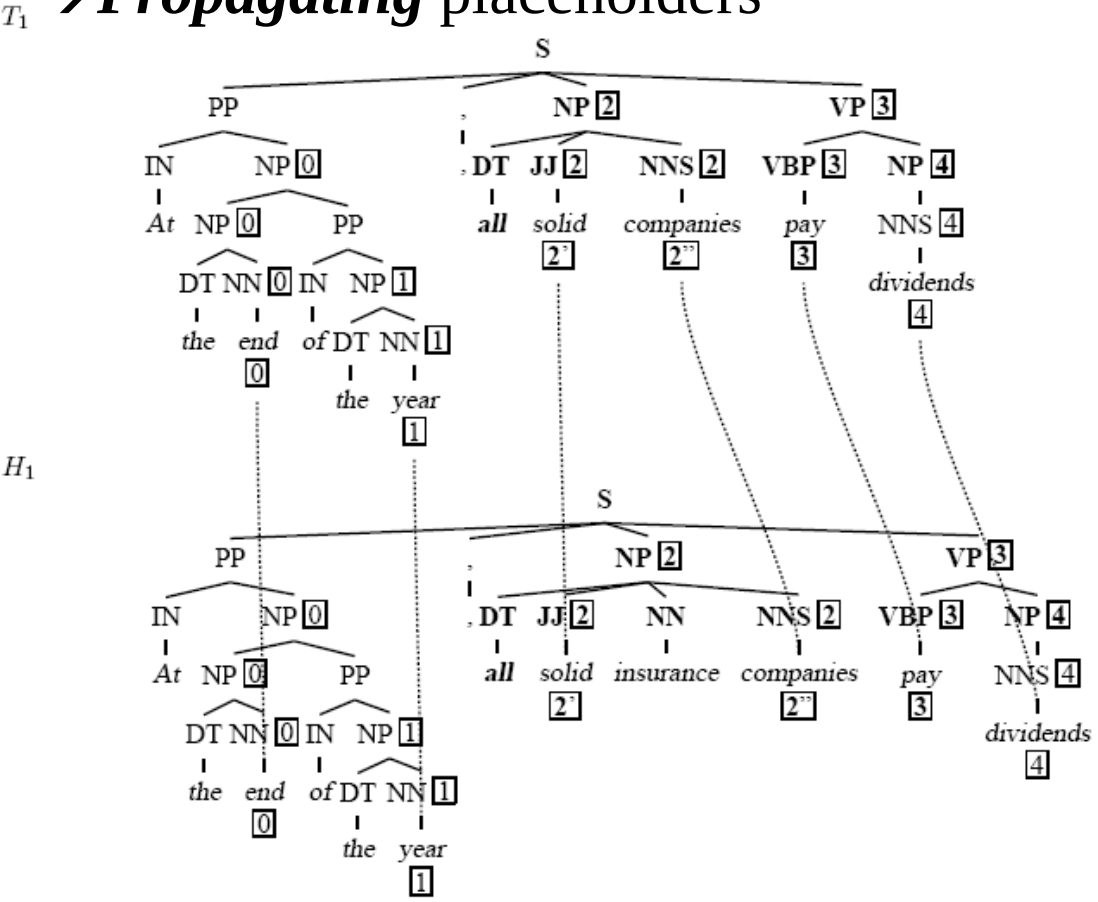
(Zanzotto, Moschitti, 2006)

Intra-pair operations

→Finding ***anchors***

→Naming anchors with ***placeholders***

→***Propagating*** placeholders



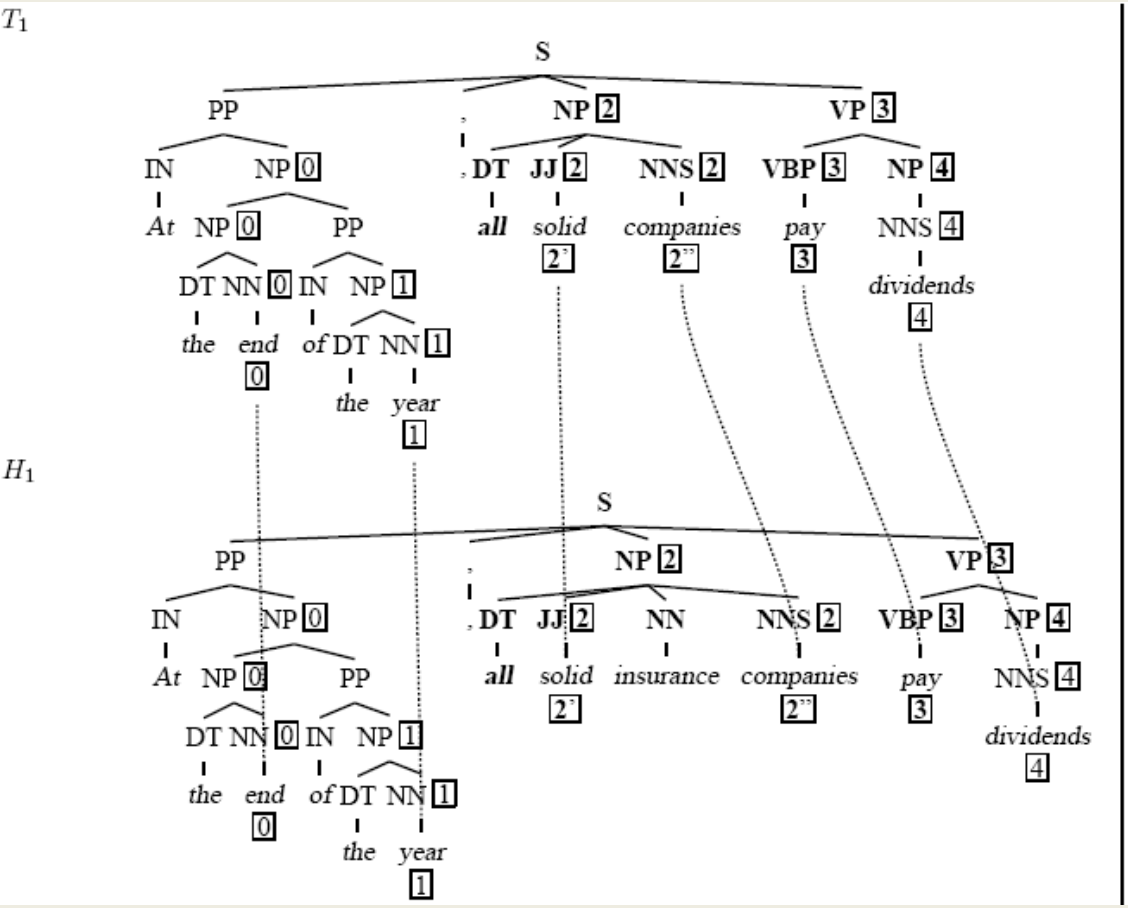
Page 76

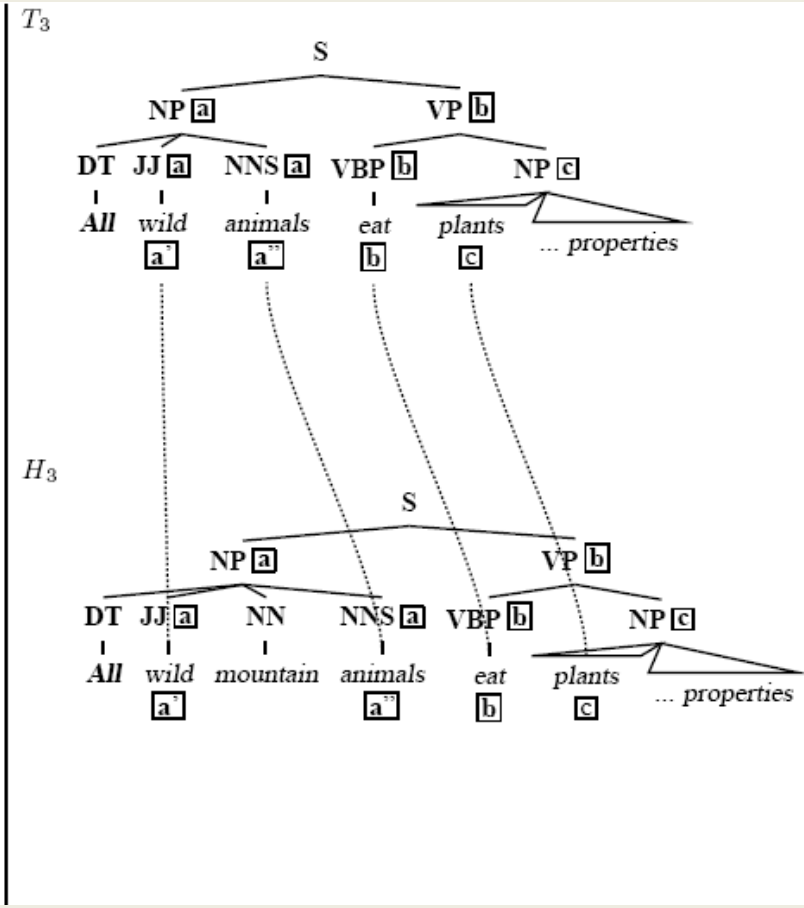
EXPLOITING REWRITE RULES

(Zanzotto, Moschitti, 2006)

Intra-pair operations *Cross-pair operations*

→Finding ***anchors***
→Naming anchors with ***placeholders***
→***Propagating*** placeholders





Page 77

EXPLOITING REWRITE RULES

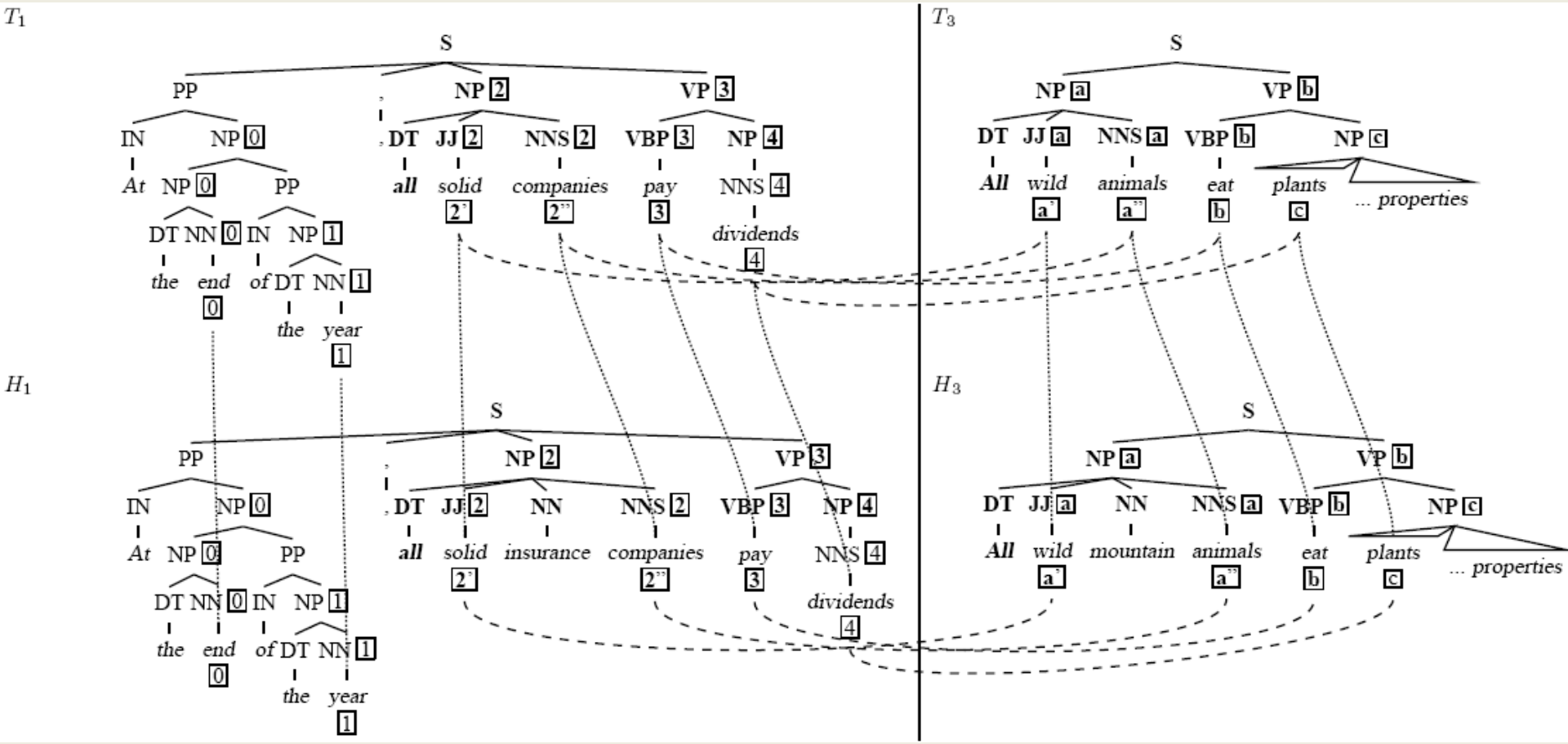
(Zanzotto, Moschitti, 2006)

Intra-pair operations

→Finding ***anchors***
→Naming anchors with ***placeholders***
→***Propagating*** placeholders

Cross-pair operations

→Matching placeholders across pairs



Page 78

EXPLOITING REWRITE RULES

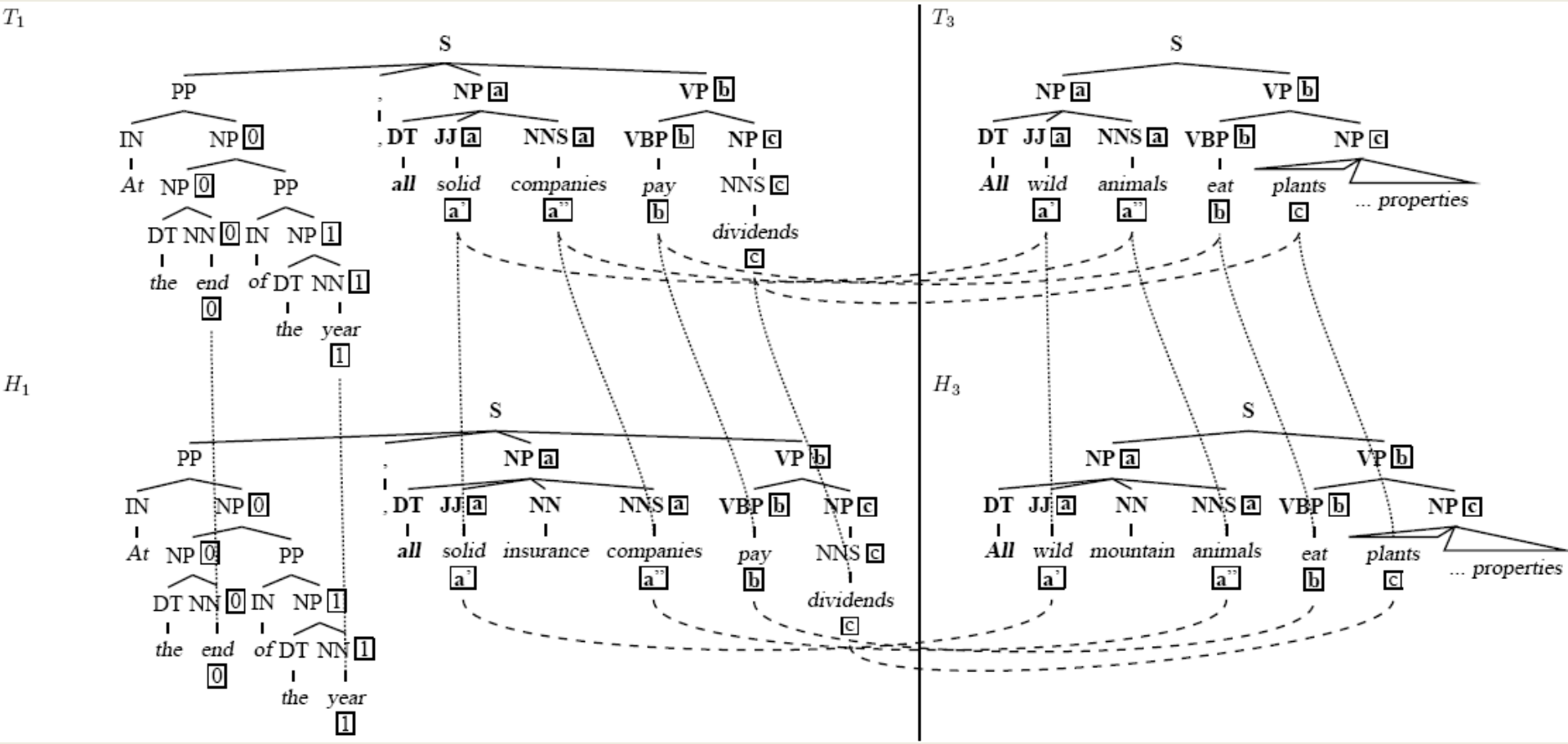
Intra-pair operations

→Finding ***anchors***
→Naming anchors with ***placeholders***
→***Propagating*** placeholders

Cross-pair operations

→Matching placeholders across pairs

→Renaming placeholders



Page 79

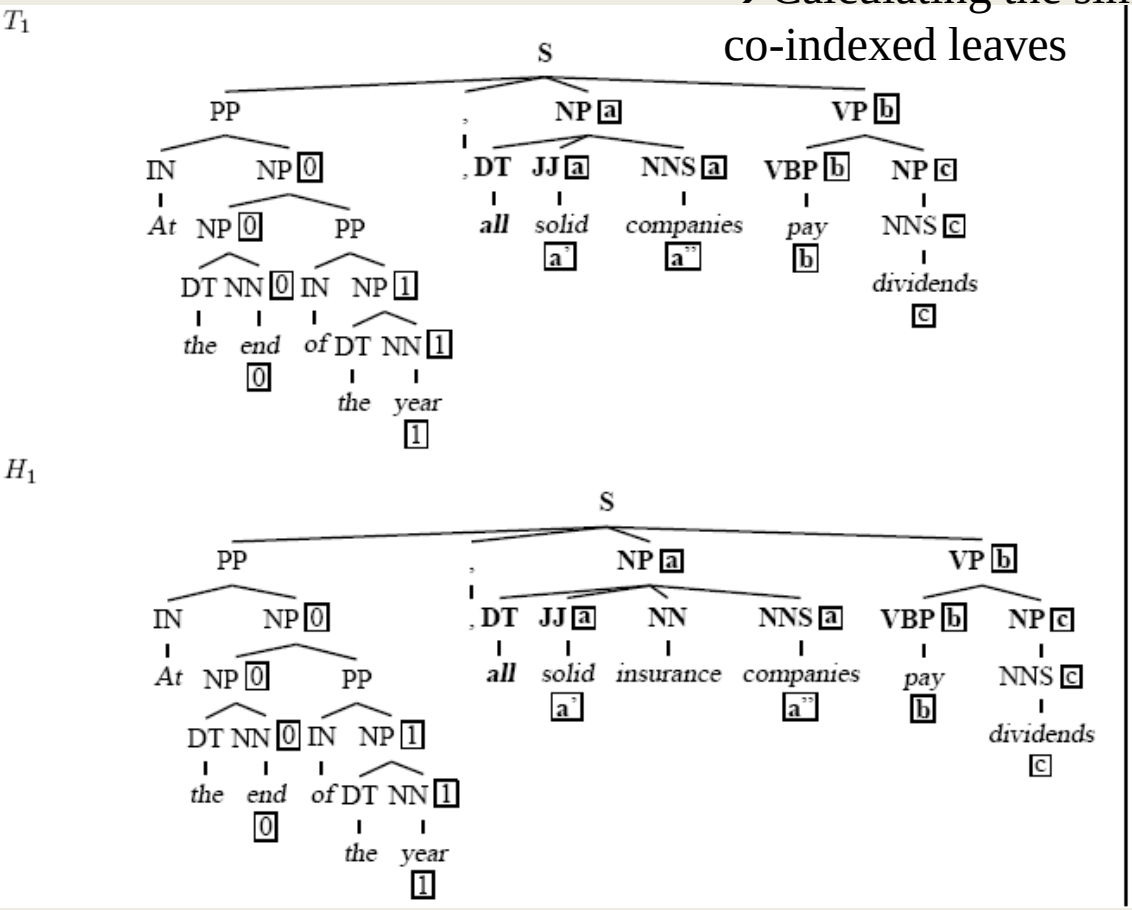
EXPLOITING REWRITE RULES

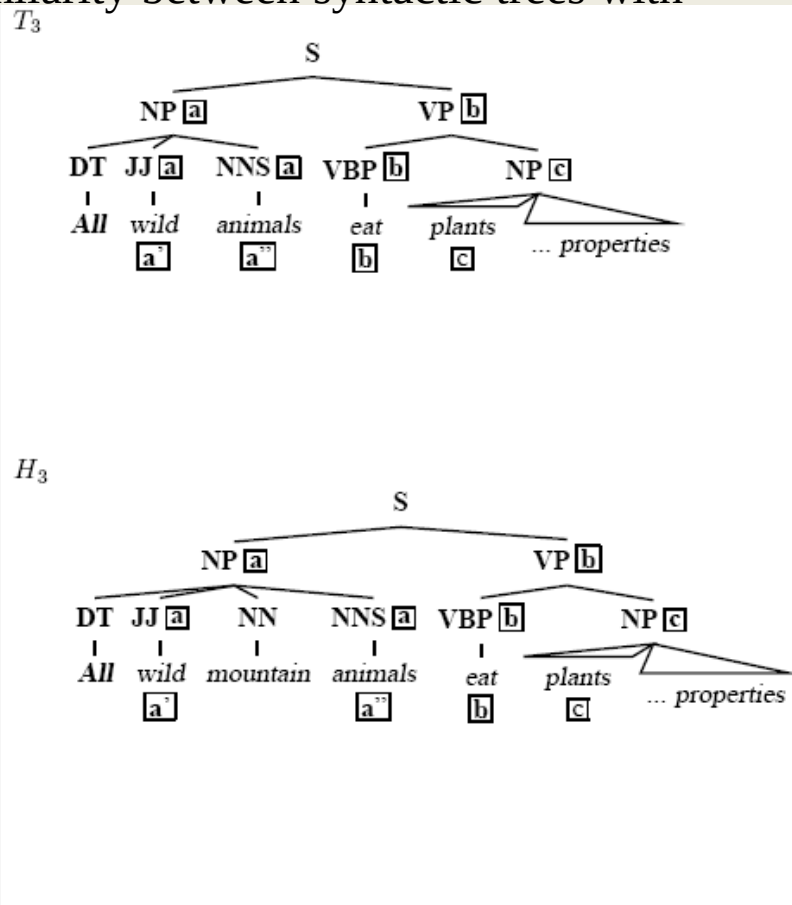
Intra-pair operations

→Finding ***anchors***
→Naming anchors with ***placeholders***
→***Propagating*** placeholders

Cross-pair operations

→Matching placeholders across pairs
→Renaming placeholders
→Calculating the similarity between syntactic trees with co-indexed leaves





Page 80

EXPLOITING REWRITE RULES

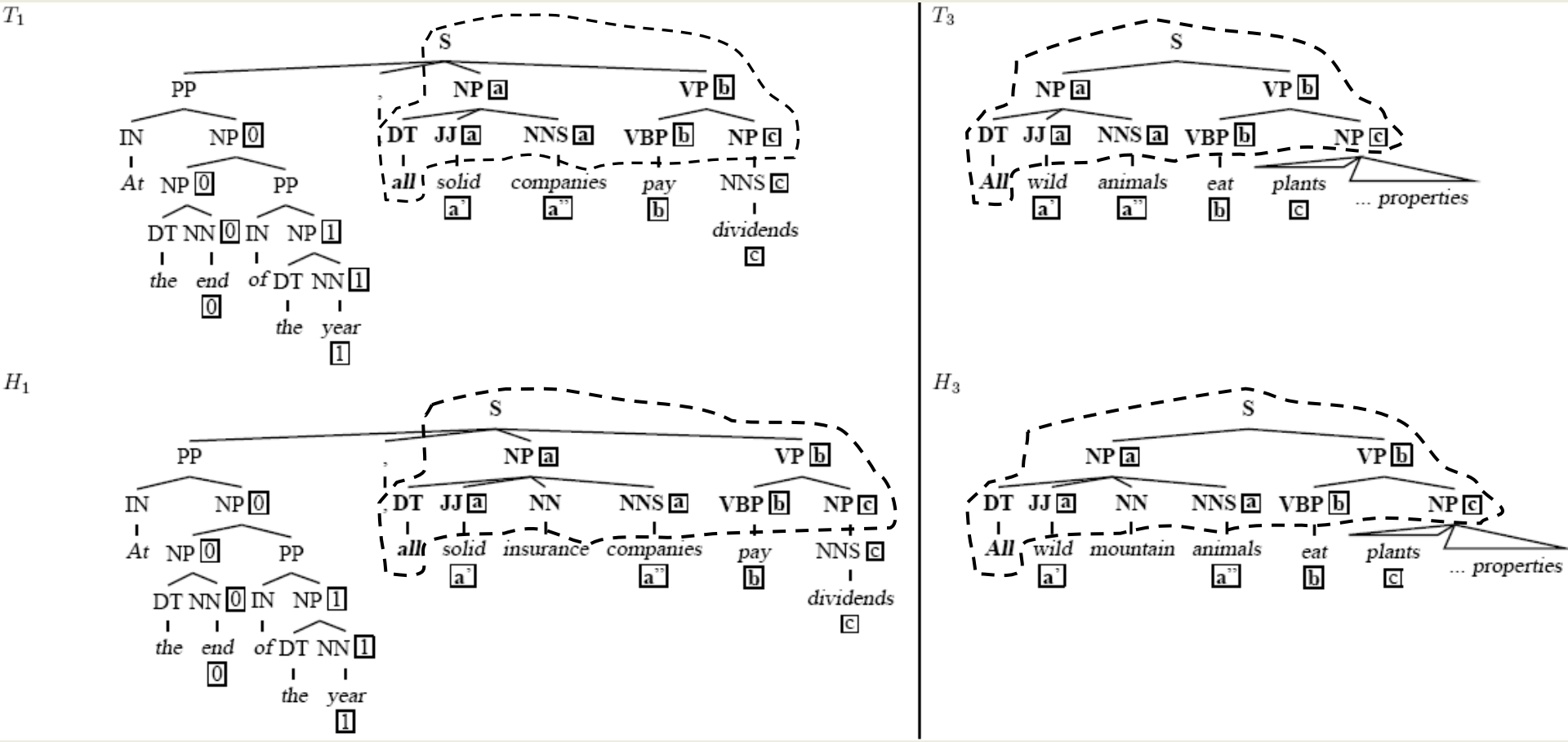
(Zanzotto, Moschitti, 2006)

Intra-pair operations

→Finding ***anchors***
→Naming anchors with ***placeholders***
→***Propagating*** placeholders

Cross-pair operations

→Matching placeholders across pairs
→Renaming placeholders
→Calculating the similarity between syntactic trees with co-indexed leaves



Page 81

Approach 6: Learning Alignment

- Idea: break entailment into smaller decisions
- Alignment as a way to recognize relevant Text portions
- Portions of text compared using closed set of operations
  - Operations include lexical similarity, structural similarity
  - Possible to define concepts such as semantic containment and semantic exclusion
  - May be extended using Knowledge bases

Learning Entailment via Alignment

- Formulated as an optimization function to align Hyp tokens to Text tokens, using lexical and dependency structure similarity
- Use learnt alignment to train global classifier
- Classifiers learnt to recognize context structures such as negation, monotonicity

INFERENCE VIA ALIGNMENT

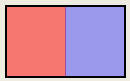
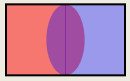
- Text represented as basic semantic premises
- Transformed to Hyp using series of edit operations
- Entailment decision predicted for each edit operation
- Decisions propagated through syntax tree
- Final label inferred using decisions over the edit sequence

NATURAL LOGIC (NATLOG)

- MacCartney and Manning, Stanford
- Use natural logic representation for TE
- Initial implementation of alignment based entailment inference
- Inference patterns built over shallow surface forms, instead of full semantic interpretation
- Can be used for post-enrichment: rules would bring structures sufficiently close for NatLog operations to become sufficient

7 BASIC ENTAILMENT RELATIONS

Slides based out of Bill MacCartney and Chris Manning's talk in COLING 2008.

Venn	symbol	name	example
	P = Q	equivalence	*couch* = *sofa*
	P ⊏ Q	forward entailment (strict)	*crow* ⊏ *bird*
	P ⊐ Q	reverse entailment (strict)	*European* ⊐ *French*
	P ^ Q	negation (exhaustive exclusion)	*human* ^ *nonhuman*
	P | Q	alternation (non-exhaustive exclusion)	*cat* | *dog*
	P _ Q	cover (exhaustive non-exclusion)	*animal* _ *nonhuman*
	P # Q	independence	*hungry* # *hippo*

Relations are defined for all semantic types: *tiny* ⊏ *small*, *hover* ⊏ *fly*, *kick* ⊏ *strike*, *this morning* ⊏ *today*, *in Beijing* ⊏ *in China*, *everyone* ⊏ *someone*, *all* ⊏ *most* ⊏ *some*

Entailment & Semantic Composition

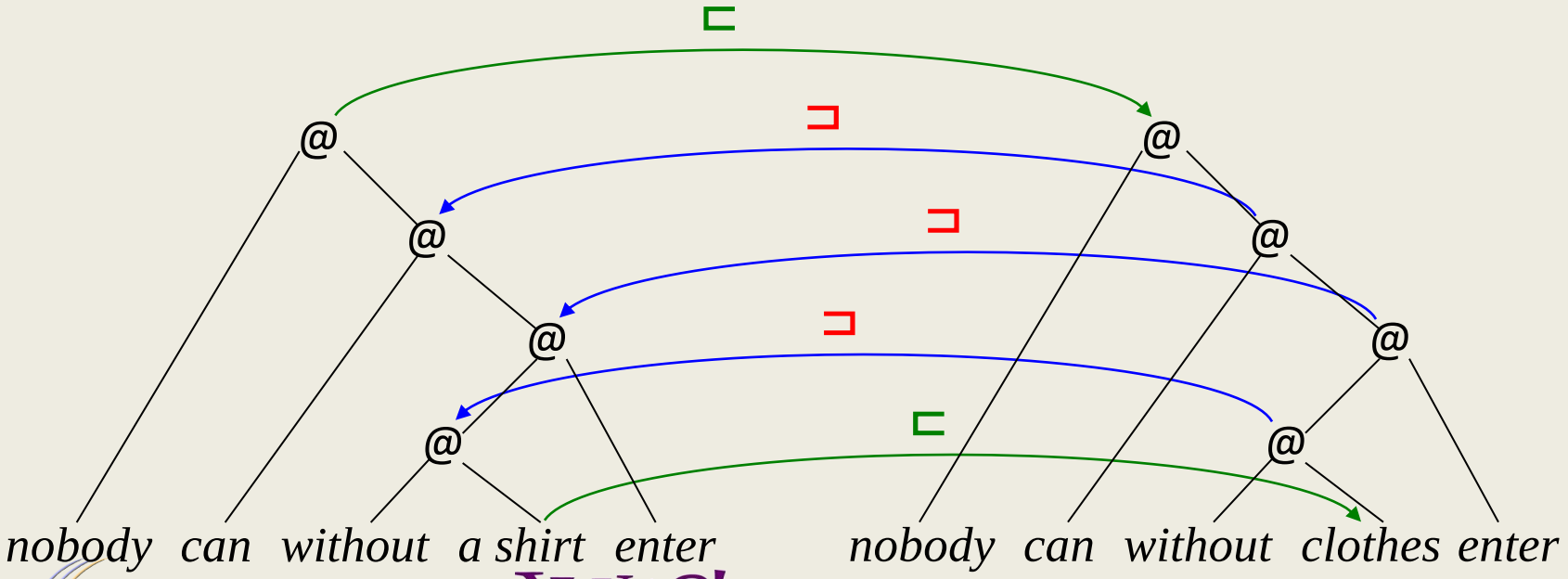
- Ordinarily, semantic composition preserves entailment relations: *eat pork* ⊏ *eat meat*, *big bird* | *big fish*
- But many semantic functions behave differently:

  tango ⊏ *dance* ⇒ *refuse to tango* ⊐ *refuse to dance*

  French | *German* ⇒ *not French* _ *not German*

Projecting Entailment Relations Upward

- If two compound expressions differ by a single atom, their entailment relation can be determined compositionally
  - Assume idealized semantic composition trees
  - Propagate entailment relation between atoms upward, according to projectivity class of each node on path to root



A (WEAK) INFERENCE PROCEDURE

1. Find sequence of edits connecting *P* and *H*
   - Insertions, deletions, substitutions, …
2. Determine lexical entailment relation for each edit
   - Substitutions: depends on meaning of substituends: *cat* | *dog*
   - Deletions: ⊏ by default: *red socks* ⊏ *socks*
   - But some deletions are special: *not ill* ^ *ill*, *refuse to go* | *go*
   - Insertions are symmetric to deletions: ⊐ by default
3. Project up to find entailment relation across each edit
4. Compose entailment relations across sequence of edits
   - à la Tarski's relation algebra

NATLOG SYSTEM

1. Linguistic analysis
2. Alignment
3. Lexical entailment classification
4. Entailment projection
5. Entailment composition

Running Example

P *Jimmy Dean refused to move without blue jeans.*

H *James Dean didn't dance without pants*

yes

RESULTS ON RTE3: NATLOG

System	Data	% Yes	Prec %	Rec %	Acc %
Stanford RTE	dev	50.2	68.7	67.0	67.2
	test	50.0	61.8	60.2	60.5
NatLog	dev	22.5	73.9	32.4	59.2
	test	26.4	70.1	36.1	59.4
Hybrid	dev	56.0	69.2	75.2	70.0
	test	54.5	64.4	68.5	64.5

4% gain (significant, $p < 0.05$)

(each data set contains 800 problems)

Part III:

THE STATE OF THE ART

Outline

- Common Threads in Existing Research
  - What current approaches can already do
  - What current approaches can't do
- Inference in Successful RTE Systems
  - Global similarity functions
  - Alignment and Machine Learning
  - Rethinking Alignment

Outline

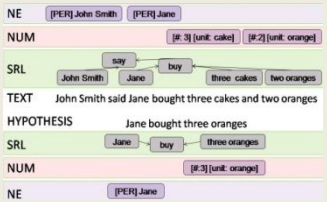
- Common Threads in Existing Research
  - What current approaches can already do
  - What current approaches can't do
- Inference in Successful RTE Systems
  - Global similarity functions
  - Alignment and Machine Learning
  - Rethinking Alignment

COMMON TE ARCHITECTURE

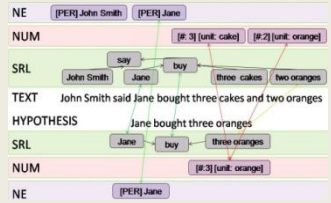
RTE DATA

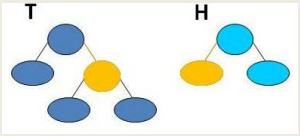
PREPROCESSING

ENRICHMENT



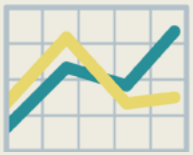
ALIGNMENT





GLOBAL SIMILARITY FUNCTION

FEATURE EXTRACTION & CLASSIFICATION



DECISION

Common Capabilities

- Robustness vs. missing knowledge
  - distance measure/machine learning
  - Combining many local decisions
- Extensive use of existing NLP resources
  - Shallow semantics (NER, syntactic parse, SRL)
  - lexical/structural knowledge resources (WordNet, VerbOcean, Wikipedia, DIRT)
- Specialized knowledge resources
  - Ad-hoc, system-specific (e.g. Numerical Reasoning in Iftene et al., 2009)
- Some notion of alignment

SEMANTIC PHENOMENA

- **A number of semantic phenomena have been identified as significant to Textual Entailment.**
  - **Very little quantification per phenomenon has been done.**
  - **See (Sammons et al. 2010) for a recent attempt.**
- **A large number of them are being handled (in a restricted way) by some RTE systems.**
  - **Transformation rules; metrics; specialized annotation/normalization**
- **Semantic implications of interpreting syntactic structures (Braz et. al'05; Bar-Haim et. al. '07)**
  - **Model-theoretic interpretation.**
  - **Each enrichment using e.g. entailment rule makes one interpretation (more) explicit.**
  - **does not CHANGE meaning – (assuming soundness of rules).**

SEMANTIC PHENOMENA (CONT.)

- **Conjunctions**
  - **Jake and Jill ran up the hill** — **Jake ran up the hill**
  - **Jake and Jill met on the hill** — ***Jake met on the hill**
- **Clausal modifiers**

  T: But celebrations were muted as many Iranians observed a Shi'ite mourning month.

  H: Many Iranians observed a Shi'ite mourning month.
- **Relative clauses**
  - **The assailants fired six bullets at the car, which carried Vladimir Skobtsov.**
  - **The car carried Vladimir Skobtsov.**

SEMANTIC PHENOMENA (CONT.)

- **Conjunctions**
  - **Jake and Jill ran up the hill** — **Jake ran up the hill**
  - **Jake and Jill met on the hill** — ***Jake met on the hill**
- **Clausal modifiers**

  T: But celebrations were muted as many Iranians observed a Shi'ite mourning month.

  H: Many Iranians observed a Shi'ite mourning month.
- **Relative clauses**
  - **The assailants fired six bullets at the car, which carried Vladimir Skobtsov.**
  - **The car carried Vladimir Skobtsov.**

SEMANTIC PHENOMENA (CONT.)

- **Appositives**
  - **Frank Robinson, a one-time manager of the Indians, has the distinction for the NL…**
  - **Frank Robinson is a one-time manager of the Indians.**
- **Passive/active**
  - **We have been approached by the investment banker.**
  - **The investment banker approached us.**
- **Genitive modifier**
  - **Malaysia's crude palm oil output has risen.**
  - **The crude palm oil output of Malaysia has risen.**

LOGICAL STRUCTURE

- **Factivity: Uncovering the context in which a verb phrase is embedded**
  - **We believe the terrorists entered the building.**
- **Polarity : negative markers or a negation-denoting verb (e.g. *deny*, *refuse, fail*)**
  - **The terrorists failed to enter the building.**
  - **Terrorists never entered the building.**
- **Modality/Negation Dealing with modal auxiliary verbs (can, must, should), that modify verbs' meanings**
  - **The terrorists might not have entered the building.**
- **Can be hard to identify the scope of the modifier.**

Logical Structure Cont'd

- Superlatives/Comparatives/Monotonicity: inflecting adjectives or adverbs.
  - Examples:

TEXT: **All companies are required to file reports at the end of the fiscal year.**

HYP 1: **All tax companies are required to file reports.** ✓

Hyp 2: **All companies are required to file tax reports.** ✗

- Quantifiers, determiners and articles

Hyp 3: **Some companies are required to file reports.** ✓

Hyp 4: **300 companies are required to file reports.** ✗

OUTLINE

- COMMON THREADS IN EXISTING RESEARCH
  - What current approaches can already do
  - What current approaches can't do
- INFERENCE IN SUCCESSFUL RTE SYSTEMS
  - Global similarity functions
  - Alignment and Machine Learning
  - Rethinking Alignment

WHAT IS A "HARD" EXAMPLE?

- NIST TAC published the outputs for all participating RTE systems (2-way and 3-way labels for RTE test sets)
- We compared the top 5 system outputs to the gold standard
- We selected examples for which all 5 made incorrect predictions

ID: 5T-11: CONTRADICTION

TEXT: A Soyuz capsule carrying a Russian cosmonaut, an American astronaut and U.S. billionaire tourist Charles Simonyi has docked at the international space station. Russian cosmonaut Gennady Padalka manually guided the capsule...

HYP: Charles Simonyi is a Russian cosmonaut.

- high lexical similarity; implicit (weak) relation; relation/argument exclusion (possibly based on numerical reasoning…)

ID: 5T-25 ENTAIL: YES

TEXT: At least 14 people have been killed in a suicide bomb attack in southern Sri Lanka, police say. The telecoms minister was among about 35 people injured in the blast at the town of Akuressa, 160km (100 miles) south of the capital, Colombo…

HYP: 49 people were hit by a suicide bomber in Akuressa.

- Lexical and concept mapping
- Numerical Reasoning/abstraction/synthesis

HARD EXAMPLES IN RTE 5

[id: 5T-79 entail: NO contradict: NO]

TEXT: …Mr. Goddard said he had hatched the idea for the Unemployment Olympics because he yearned for the chance to "battle all the unemployed people for stuff." "It's also not a bad time to be unemployed," said Mr. Goddard, who is from Rochester Hills, Mich., and lives in the East Village…

HYP: The Unemployment Olympics took place in the East Village.

5T-281 CONTRADICTION

TEXT: Pop music producer Phil Spector… has been convicted of second-degree murder in the 2003 shooting of actress Lana Clarkson... Clarkson died February 3, 2003 at Spector's mansion, the "Pyrenees Castle"… Spector was arrested after police were called to the mansion, finding Clarkson dead of a gunshot wound…

HYP: Actress Lana Clarkson killed music producer Phil Spector.

ID: 5T-437: ENTAILED

TEXT: The Japanese Nikkei 225 has recorded it's third biggest drop in history with a massive sell-off in the exchange that has resulted in USD 250 billion being knocked of the index's value. Toyota, which is the second largest carmaker in the world, fell by the largest amount in 21 years, while Elpida Memory, the world's largest manufacturer of computer memory, dropped in value to a record low.

HYP: Japan's economy is not flourishing.

Common Weaknesses

- Current approaches strongly dependent on explicit representation of semantic content
  - Lexical + local structural similarity tends to dominate
  - We do not recover errors in deep structure well
- Knowledge resources lack broad coverage
  - Which is to say, much needed knowledge is missing
- Pipelined architecture is prevalent
  - Lossy, especially in staged systems
- Back-off measures make error analysis difficult
  - Little explanatory power

Outline

- Common Threads in Existing Research
  - What current approaches can already do
  - What current approaches can't do
- **Inference in Successful RTE Systems**
  - Global similarity functions
  - Alignment and Machine Learning
  - Rethinking Alignment

Inference in RTE

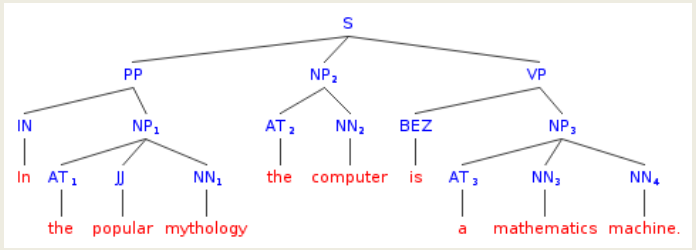
- Three general responses to RTE problem:
  - Logical Form/theorem proving
  - Machine Learning/statistical
  - Similarity function
- LF problematic:
  - How to map from NL to LF?
  - How to handle missing knowledge?
  - Previous efforts very brittle (Bayer et al. 2005)
    - use LF/TP only as component in ensemble (Tatu et al., 2006) or as source of coarse features (Bos and Makert, 2006)
    - use shallow back-off model (Clark and Harrison, 2009)

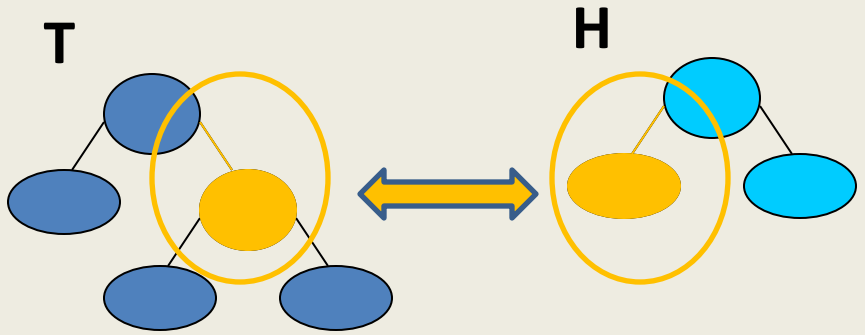
SIMILARITY FUNCTIONS

- F(T', H') -> R
  - Threshold to separate classes
  - For 3-way task, either 2 stages (Wang et al. 2009) or 2 thresholds (Iftene et al. 2009)
- Compositional
  - Combine local scores
  - Global adjustments possible
    - **Abstraction of simple modal/factive/polarizing structures**
    - **Ad-hoc filtering rules**
- (Zanzotto et al. 2006) use inter-pair similarity function, train using RTE labels (2-way)

IFTENE ET AL. 2009

- Dependency parse-based structures
- For T, H nodes, compare node lexical entry AND connecting structure (to parent)





- Aggregate multiple "fitness functions" (metrics)
- Aggregate local scores, adjust globally for filtering rules

WHY (WHEN) DO SIMILARITY FUNCTIONS WORK?

- Given a set of local comparisons (entailment decisions), it is not very likely that most will be in agreement yet be inconsistent with the global label
- Some tolerance of noise
  - e.g., "expect 1 mistake" -> lower threshold
  - Like using set of mediocre predictors, "K of M" rule
- Extends lexical model with limited context – most reliable analytic structure (dependency parse)

SIMILARITY FUNCTIONS

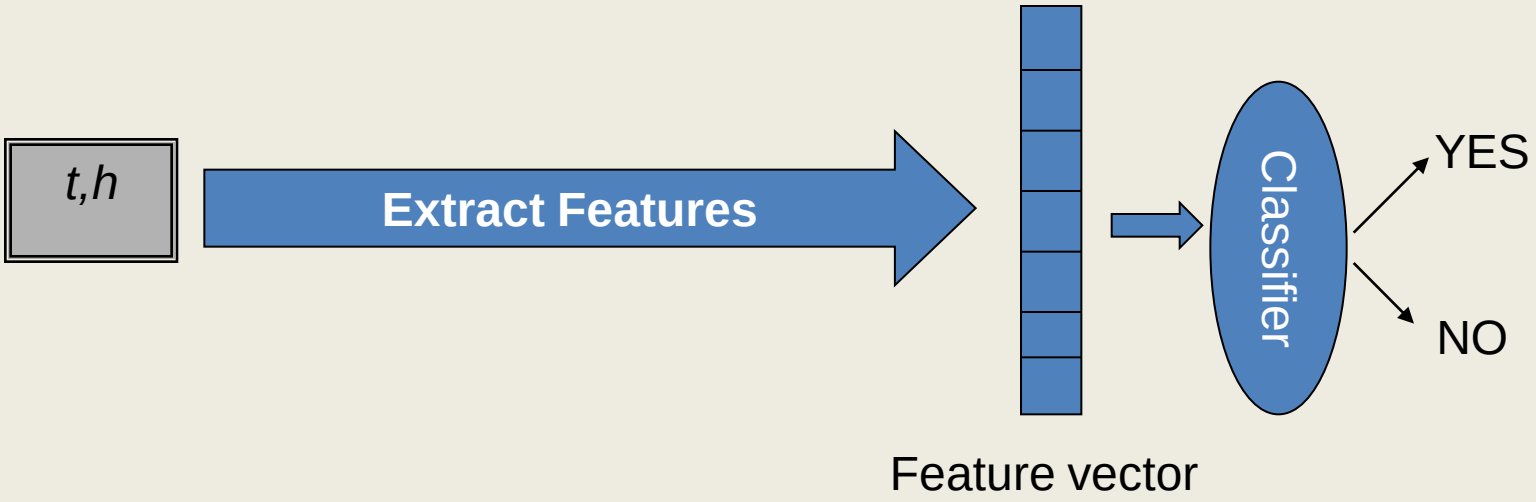
- Strongly dependent on distribution of corpus
  - Not good for precise distinctions in structure, e.g. contradiction cases
  - Encouraging that this works for RTE corpora, which were not selected with this model in mind ("unbiased" corpora…)
- Requires normalization of T, H in cases where "large" inference steps needed (e.g. T is missing explicit H content)
  - Needs background/domain knowledge
- Modular, in that they may use type-specific similarity resources
  - But scaling issues not clearly addressed

OUTLINE

- COMMON THREADS IN EXISTING RESEARCH
  - What current approaches can already do
  - What current approaches can't do
- INFERENCE IN SUCCESSFUL RTE SYSTEMS
  - Global similarity functions
  - Alignment and Machine Learning
  - Rethinking Alignment

MACHINE LEARNING IN RTE

Standard ML approach…

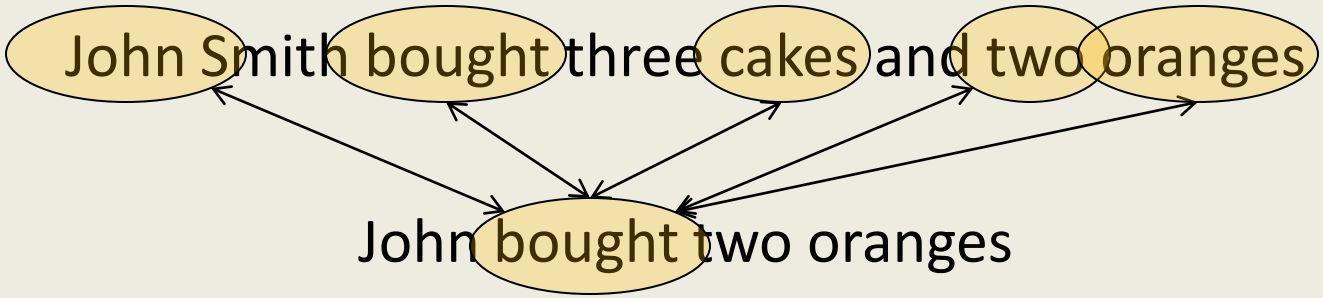


But which features?

- Naïve approach (e.g. Words and POS) yields RTE5 dev feature "If Text contains Madonna then ENTAILS"
- Not many labeled examples (esp. given problem complexity)
- Solution: similarity features

SIMILARITY FEATURES

- There are many similarity features – which ones are the correct ones?

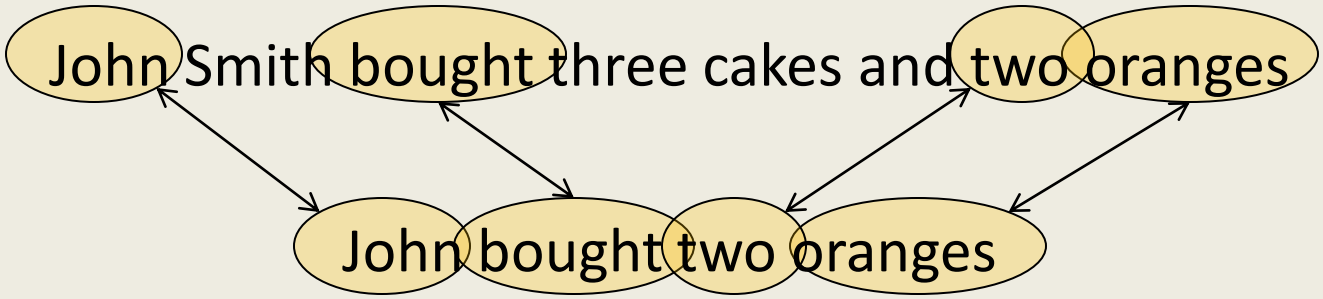


- If we use all possible comparisons for each word, can we get a signal?

Selecting Similarity Features

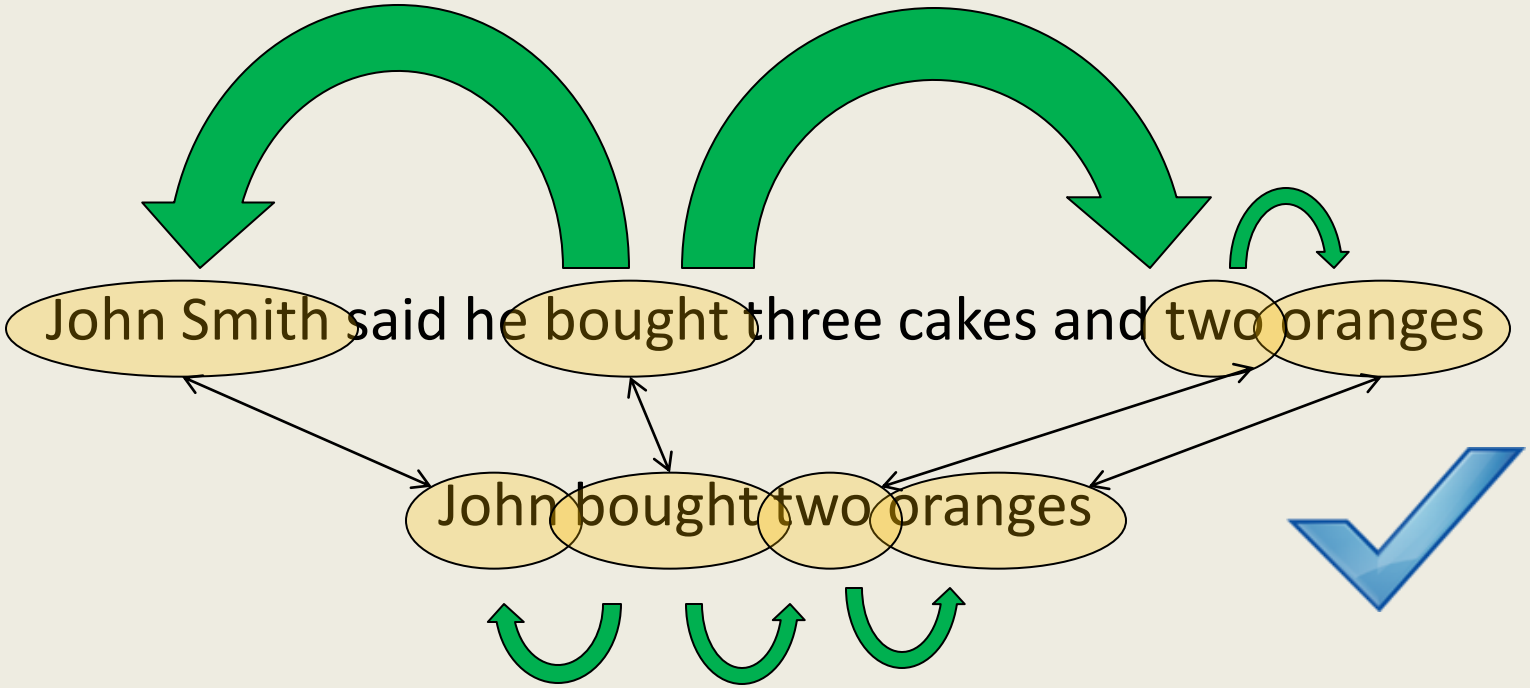
- Impose ***constraints*** on the aggregate set of comparisons we entertain
- E.g. each Hypothesis element can match at most one Text element

Alignment: a mapping from elements in the Hypothesis to elements in the Text under specified constraints



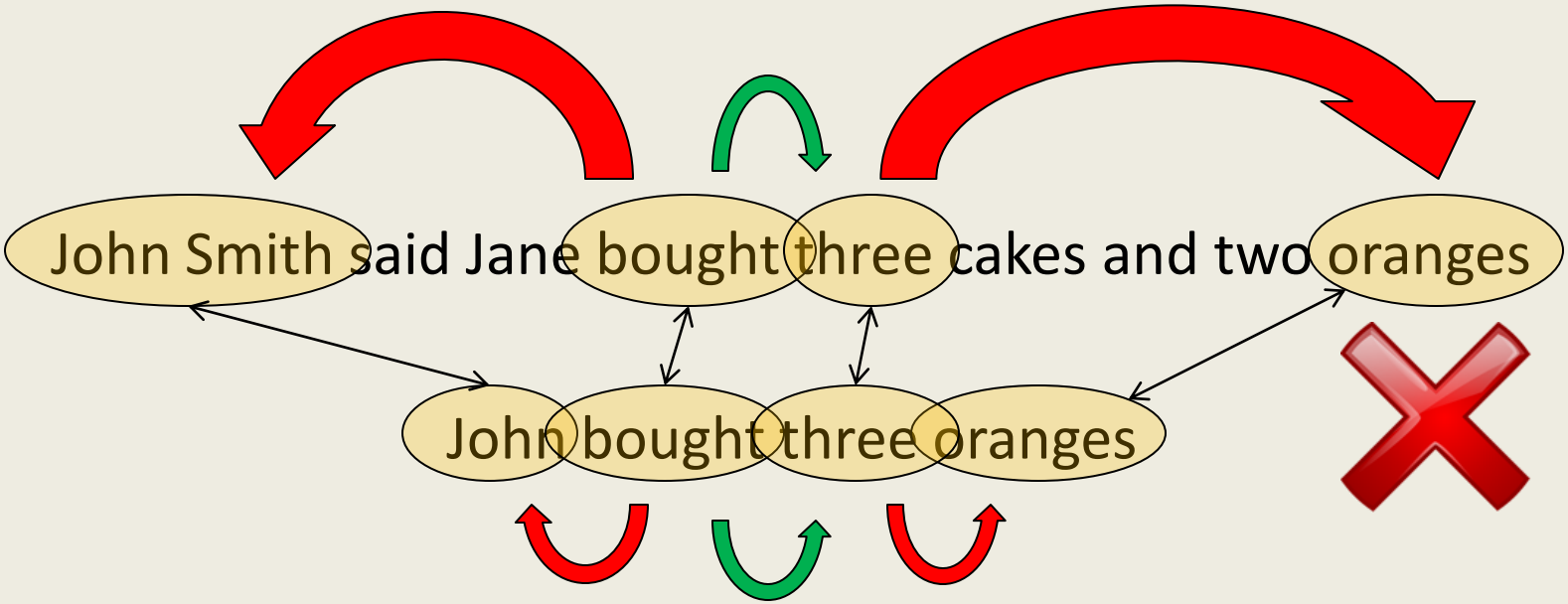
Shallow Alignment as Focus of Attention

- Pick a "good" shallow alignment
- Use this to query deeper structure/extract features



Shallow Alignment as Focus of Attention

- Pick a "good" shallow alignment
- Use this to query deeper structure/extract features



Alignment Research

Chambers et al. 2007, deMarneffe et al. 2007

- learn "alignment" from lexical-level labelings
  - Intuition: abstract away some logical structure, irrelevant Text content
  - Identify the parts of T that "support" H
- Identify "relevant" parts of T via word, edge weight vectors
- Use alignment to extract features for discerning "entailed" from "not entailed", using deeper semantic structure

Chambers et al. Alignment

- Alignment score:

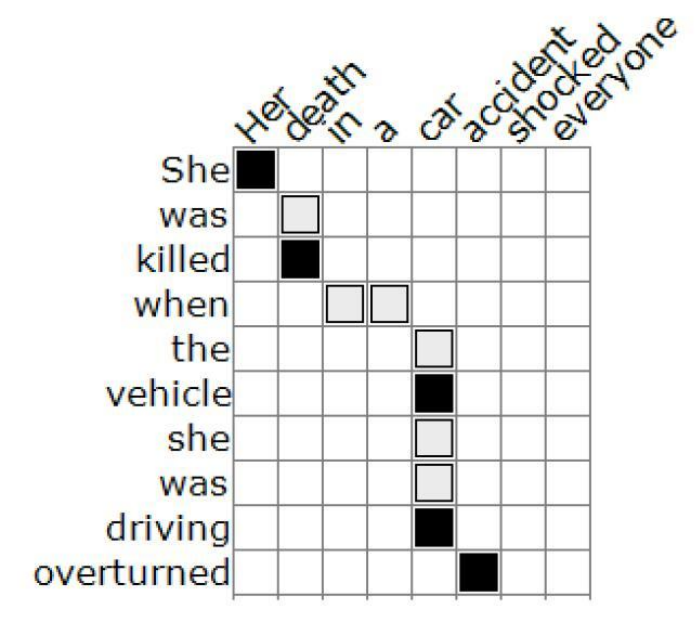
$$\text{score}(a) = \sum_{i \in h} \text{score}_w(h_i, a(h_i)) + \sum_{(i,j) \in e(h)} \text{score}_e((h_i, h_j), (a(h_i), a(h_j)))$$

- Lexical similarity of aligned H, T words PLUS:
- Given dependency relationship between two words in H, similarity of dependency relation between the mapped words in T
- Stochastic local search to explore space of alignment
  - **Initialize with greedy lexical alignment**
  - **Gibbs-like exploration of space of alignments**

Learning Alignment (Cont'd)

Limitations:

- Results are not stellar: lexical level mapping not sufficiently expressive?



- Expensive annotation effort
- Alignment annotation is difficult for negative examples, and even for some positive examples

Multiple Alignments

Sammons et al. 2009: each alignment:

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

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

e: alignment edge
α: weight of distance parameter
m: number of tokens in H

Δ: distance function between mapped constituents

- Alignment weights based on specialized similarity metrics; parameters set using heuristics
  - Avoid scaling problem: separate metrics into different alignments
- Extract features based on comparison between alignments

WHY DOES ALIGNMENT WORK?

- Comparable to similarity metric approach
  - Trying to capture deeper structure
- Supports discriminative ML by generating sufficiently coarse features
- Works best on cases where content in H is explicit in T
  - But with better deep structure/appropriate representation, expect to do better
- Better inputs => better alignments
  - Problem: pipeline effect for erroneous annotations AND for erroneous alignment

Problems with Alignment

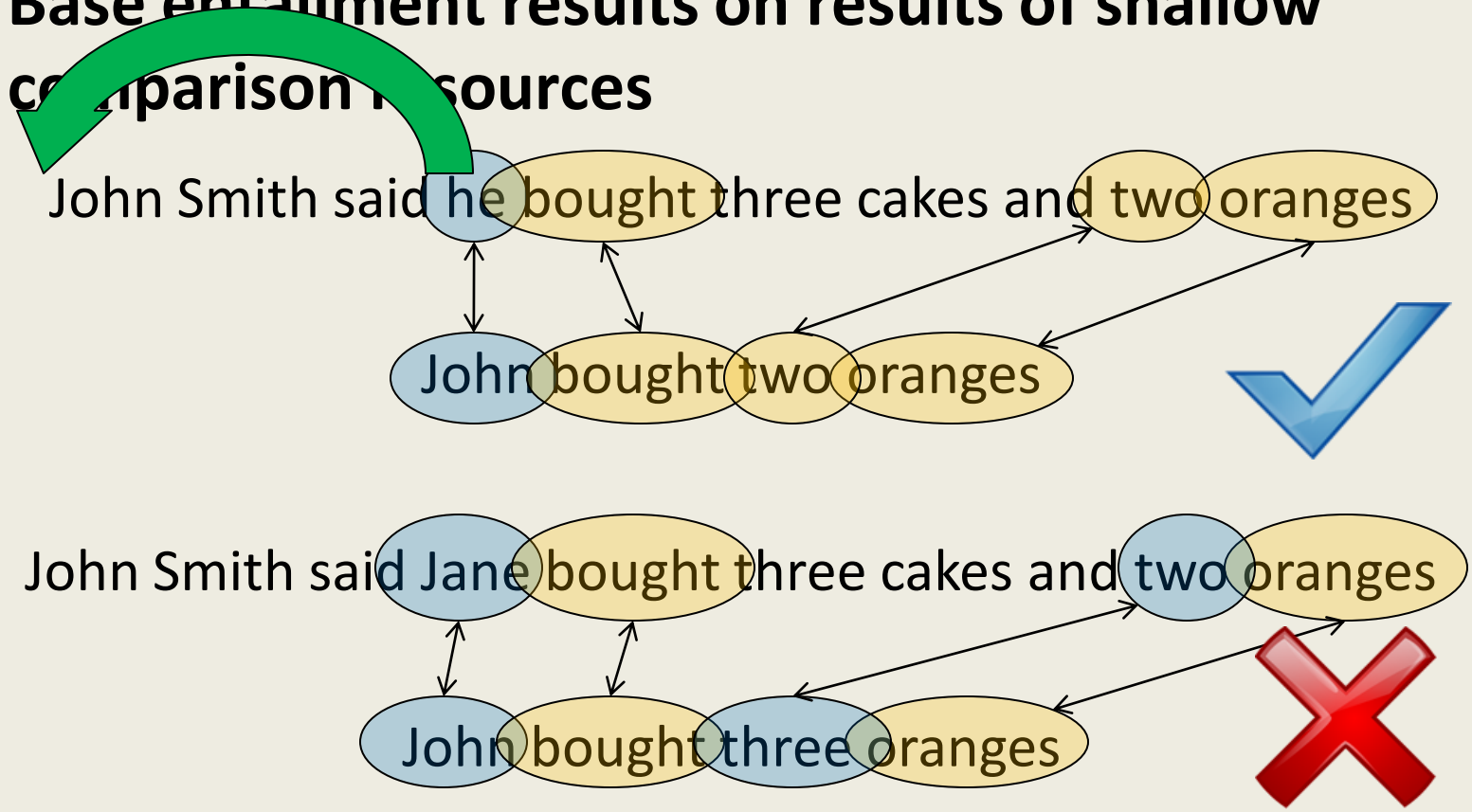
- Mapping "relevant" parts may be correct intuition, but "relevant" seems to depend on deep structure
  - Fixed heuristic/learned mapping based on shallow cues is problematic
  - Distance is not a reliable proxy for deep structure
- May be multiple match candidates for many H constituent (i.e., shallow alignment may pick the wrong one)
  - Alignment constraints introduce a problem in fixed two-stage system

OUTLINE

- COMMON THREADS IN EXISTING RESEARCH
  - What current approaches can already do
  - What current approaches can't do
- INFERENCE IN SUCCESSFUL RTE SYSTEMS
  - Global similarity functions
  - Alignment and Machine Learning
  - Rethinking Alignment

Alternative: Using Structure as Focus of Attention

- **Find best structural match**
- **Base entailment results on results of shallow comparison resources**



Deep-first approach

- Getting correct structure is HARD
  - P(all correct) = 0.9^3 per predicate-argument structure*

  *based on SRL training domain, i.e. optimistic
- Errors in deep structure → problem selecting correct local decision
- Other preprocessing errors – e.g. Coreference – will propagate in same way as shallow-first approach

A BETTER ALIGNMENT MODEL?

Zanzotto et al. 2006, 2007, 2009:

- **Learn to distinguish alignments for positive, negative TE examples**
- **Alignment is fixed, but we learn from what we have (potential to recover from some *consistent* input noise…)**
- **Stated goal is to learn FOL rules expressing structural mappings**
  - **Seems problematic: variability of language seems too great to simply learn absent constraining principles**
  - **How many RTE examples needed?**
  - **How limiting is the quality of the (fixed) input alignment?**

A BETTER ALIGNMENT MODEL?

Chang et al. 2010:

- **Bootstrap alignment and classification**
- **Semi-supervised approach (can use other data)**
  - **Indirect supervision: binary labels, characterization of *space* of alignments**
  - **Learn best model within given space, that optimizes performance on binary task**
- **Learn alignment from binary entailment labels**
- **Agnostic to specific alignment process**
  - **But if we have good inputs (metrics, enrichment via rules), we expect 'discovered' alignments to emphasize these resources**

Open Questions

Global Similarity:

- How to account for deep structure, e.g. factivity, polarity; presently, seems ad-hoc
- Not clear how enrichment resources (e.g., Relation Extractors) are currently used: packed forest approach?
- Issues of scaling different similarity resources not well addressed/explained

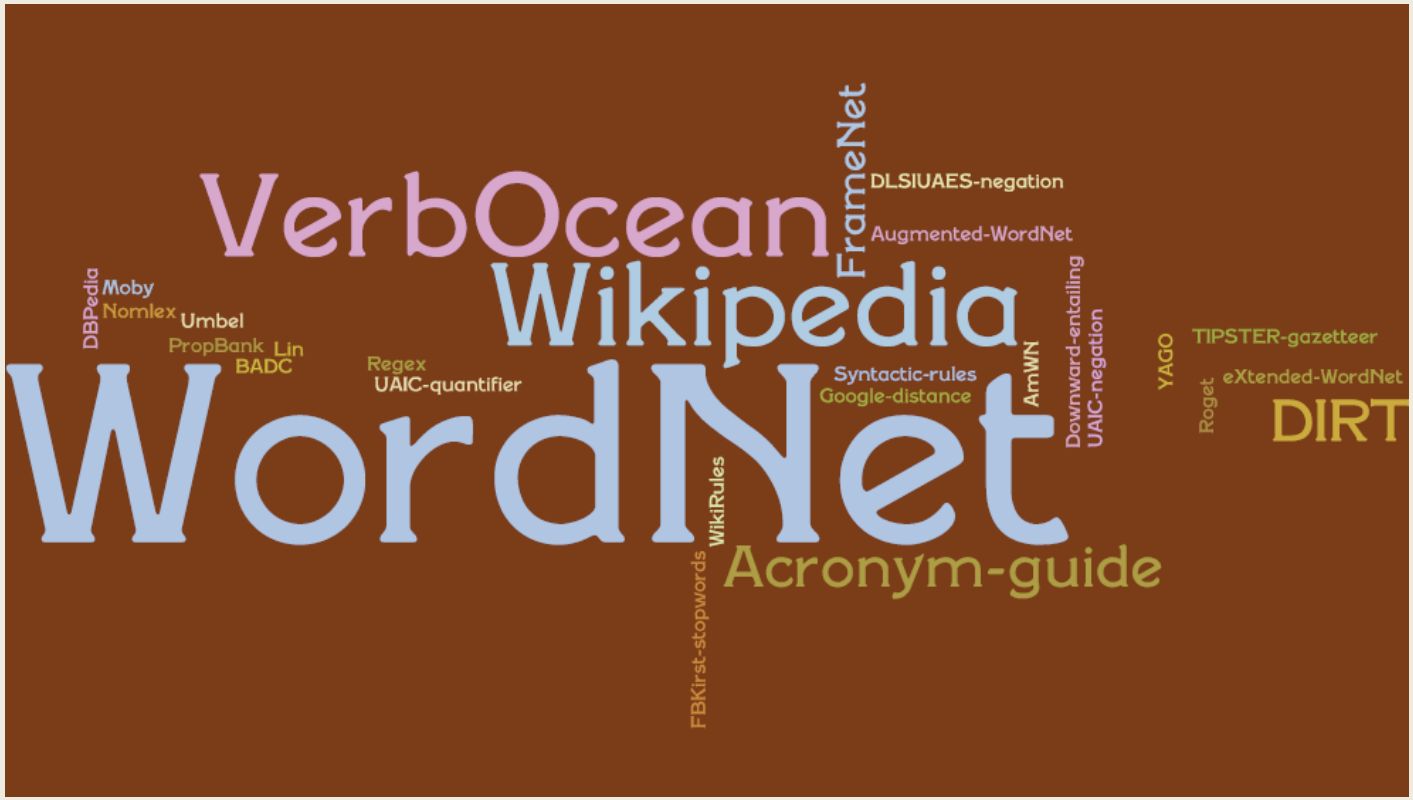
Alignment:

- Lexical-level alignment is not sufficiently informative, too expensive to generate.
- Is Deep-First alignment appropriate/feasible?
- Can we *learn* alignments beyond the lexical level, informed by entailment labels? (Chang et al. may be promising direction)

Global Sim. vs. Alignment

- Based on common intuition: structure is important!
  - Just, ***which*** structure
- Both are outperforming shallow lexical models
- Underlying models are very similar
  - **Alignment adds more constraints to application of similarity metrics; presently used mainly as *input* to RTE**
  - **Alignment explicitly oriented to application of Machine Learning techniques**
  - **Alignment models have broader application – are more agnostic regarding chosen level of representation**
  - **Both have problems with missing knowledge**

Part IV:
KNOWLEDGE ACQUISITION AND APPLICATION



The Knowledge Bottleneck

- Linguistic and world knowledge – integral part of RTE
- Missing knowledge resources – a barrier for further advances in RTE (Bar-Haim et al., 2006, Giampiccolo et al., 2007)

We need:

- Broad-coverage entailment knowledge resources
- Models for applying knowledge selectively in context
  - Even using WordNet effectively is still an open issue (WSD)

OUTLINE

- Knowledge representation by entailment rules
- Rule-base Acquisition
- Context-sensitive rule application
- Evaluation of rule-bases

KNOWLEDGE REPRESENTATION

ENTAILMENT RULES

- Most of the knowledge utilized by TE systems may be represented by entailment rules
- *Entailment rule*: entailment relation between two text fragments, possibly with variables
  - *lhs* → *rhs* (*entailing* → *entailed*)
  - Paraphrases: bidirectional entailment rules

New York → *city* (lexical rule)

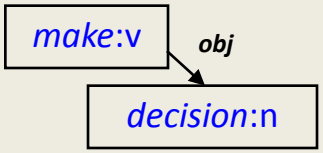
X buy Y from Z ↔ *Z sell Y to X* (template-based rule)

Y is V[ed] by X → *X V Y*

- Local inferences – combined to form complex entailments

LEXICAL RULES

- Lexical rules describe entailment relations between terms or phrases
- Substitutable rules: substituting *lhs* with *rhs* generates a valid text
  - *New York* → *city* "*I visited New York*" → "*I visited [a] city*"
  - *buy* ↔ *purchase* "*I bought a car*" → "*I purchased a car*"
- Non-substitutable rules: cannot simply substitute *lhs* with *rhs*
  - *definition* ↔ *define* "*My definition is wrong*" ↛ "*My define is wrong*"
  - *car* → *wheel*
  - *The Magical Mystery Tour* → *Beatles*
- Typically represented as surface strings or parse sub-trees
  - *make a decision* ↔ *decide*



↔ *decide*

TEMPLATE-BASED RULES

- Rules between templates with shared arguments
  - Templates are text fragments with variables
  - Highly generic representation – useful also for syntactic-based rules

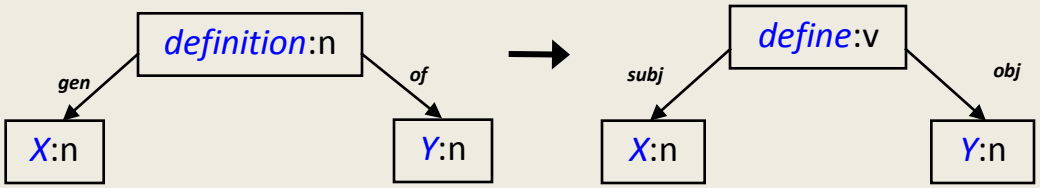
X buy Y → *X pay for Y*

X snore → *X sleep*

X's definition of Y ↔ *X define Y*

X's definition by Y ↔ *Y define X*

- Typically represented as transformations between parse sub-trees



  - Additional syntactic annotation for semantic disambiguation
    (Macleod et al., 1998; Szpektor and Dagan, 2009)

X broke$_{intransitive}$ → *X was damaged* vs. *X broke*$_{transitive}$ → *X damaged*

Higher Representation Levels

- Lexical semantic – based on further semantic annotation
  - Semantic Role Labeling: X_{Buyer} buy ↔ sell to X_{Buyer}

- First-order-logic

  excellent:JJ(x1) → *of:IN(x1, x2) highest:JJ(x1) quality:NN(x1)*

  - Large scale rule-sets are usually acquired from lower representation levels, and may then be converted to logic form

- Unpopular representations due to lack of robust parsers

Entailment Rule-Set Acquisition

Large-scale resources

- Broad coverage is a necessity – requiring huge resources
  - Calls for unsupervised approaches

- Typical approaches:
  - Rules generation from manually constructed resources
    - Very accurate
    - Limited rule coverage
  - Statistical learning from corpora
    - Good potential for broad coverage
    - Mediocre accuracy

Lexical Rule Acquisition from Manually Constructed Resources

The WordNet Lexicon (Miller, 1995)

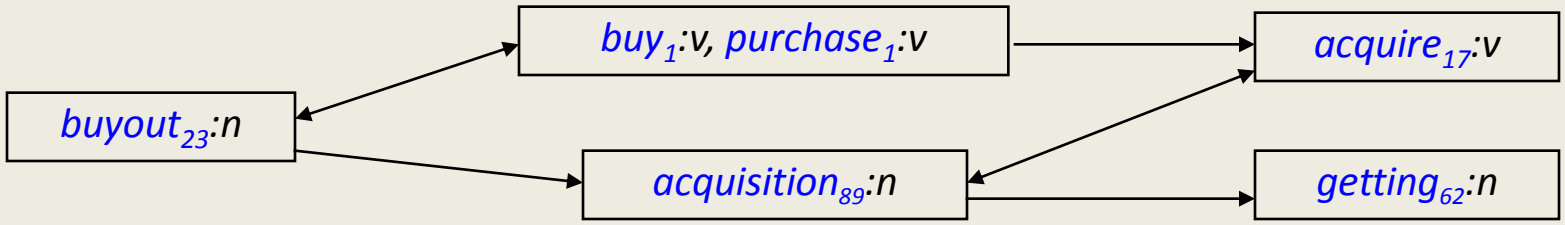
- WordNet – lexical database organized by meanings (*synsets)*

  S1: buy, purchase (obtain by purchase)

  S2: bribe, corrupt, buy, … (make illegal payments to in exchange for favors …)

- WordNet contains lexical relations – some useful for inference
  - hypernymy (*capital* → *city*), instance-of (*Paris* → *city*), derivationally-related (*acquire* ↔ *acquisition*), meronymy (*car* → *wheel*)

- Relations define a directed "entailment" graph for terms
  - Traverse the graph to generate entailment rules
  - Measure distance between terms on the graph (WordNet similarity)



WordNet Extensions

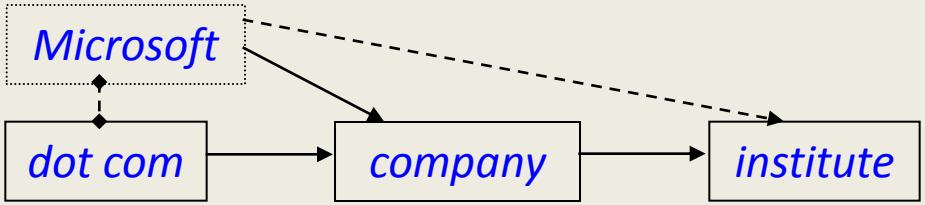
- *eXtended WordNet* (Moldovan and Rus, 2001)
  automatically generate rules from WordNet glosses

  S: excellent, first-class (of the highest quality)

  ⇓

  X is excellent → *X is of the highest quality*

- *Augmented WordNet* (Snow et al., 2006)
  automatically add new terms to the WordNet graph
  1. Extract hyponym candidates using a *hypernymy* classifier
  2. Greedily add the candidate that best meets the transitivity constraints in the graph



WIKIPEDIA

- Wikipedia – a free, web-based multilingual encyclopedia

> **Encyclopedia**
> An **encyclopedia** (also spelled **encyclopaedia** or **encyclopaedia**) is a type of reference work, a compendium holding information from either all branches of knowledge or a particular branch of knowledge.

- Term similarity based on LSA (Mehdad et al., 2009)
  Apple :: *Macintosh*

- Pattern-based rule extraction
  - Terms in first sentence entailed by the title (Shnarch et al., 2009)
    encyclopedia → *reference work* ; *encyclopedia* → *compendium*
  - Relation extraction anywhere in a page (Iftene and Balahur-Dubrescu, 2008)
    Dalmatia → *Italy* ; *Berlusconi* → *Italy*

WIKIPEDIA-RELATED KBS

- DBpedia (Auer et al., 2007): structured information from Wikipedia
  - Contains properties and relations for a topic

Microsoft

Products	*Microsoft Windows, Microsoft Office, Bing, Zune*
Key persons	*Steve Ballmer, Bill Gates, Ray Ozzie*

- Yago (Suchanek et al., 2007): link between Wikipedia's category hierarchy and WordNet's ontology

Dr. Dre → *[wiki] G-funk musicians* → *[wiki] hip hop musicians* → *[wn] musician*

Karl Marx → *[wiki] political philosophers* → *[wn] philosopher*

Other Manually Constructed Resources

- Thesauri – synonyms and related terms
  - *Moby thesaurus*

    tree: …, *alder*, …, *apple*, …, *block*, …, *corner*, …, *family tree*, …, *genealogy*, …, *peach*, …, *stick*, …, *timber*,…

- Gazetteers – geographical dictionaries
  - *Tipster gazetteer*

    Sao Paulo (city) → *Sao Paulo* (province) → *Brazil* (country)

    Sao Paulo (island) → *Brazil* (country)

- Acronym and Abbreviation Lists
  - *BADC*

    AIS – *Airborne Imaging Spectrometer*

    EGS – *European Geopysical Society*

Corpus-based statistical lexical rule acquisition

DISTRIBUTIONAL SIMILARITY

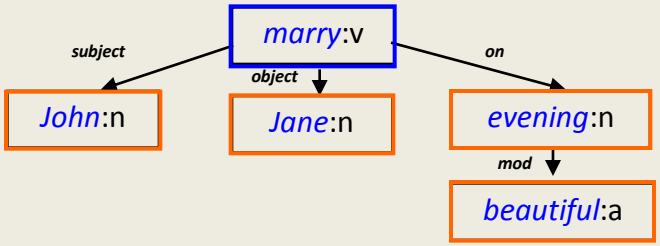
- Unsupervised learning of rules based on distributional similarity between terms
  - Assumption (*Harris, 1954*): terms that appear in similar contexts have similar meanings

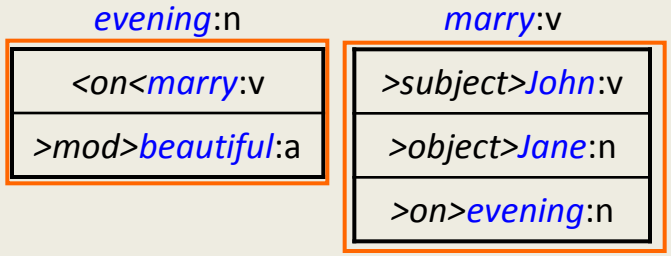
- General approach:
  1. Construct a feature vector for each term from its occurrences
     - Co-occurring words in the same sentence
  2. Score each feature
  3. Measure similarity between term vectors
     - Keep the top-N similar terms for each term

LIN SIMILARITY (LIN, 1998)

- Features – words in dependency relations with the target term

John married Jane on a beautiful evening





- Feature score – pointwise-mutual-information (PMI):

$$s(t, f) = pmi(t, r_f, w_f) = \log\left(\frac{\|t, r_f, w_f\| \cdot \|*, r_f, *\|}{\|t, r_f, *\| \cdot \|*, r_f, w_f\|}\right)$$

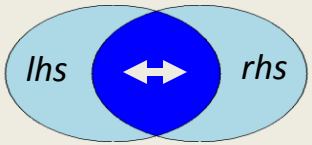
- Lin's similarity measure :

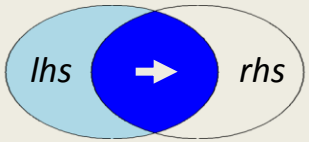
$$\mathrm{Lin}(l, r) = \frac{\sum_{f \in V_l \cap V_r} [s_l(f) + s_r(f)]}{\sum_{f \in V_l} s_l(f) + \sum_{f \in Vr} s_r(f)}$$

night:n	*evening*:n
>mod>*late*:a	>mod>*late*:a
>mod>*beautiful*:a	<on<*marry*:v
<obj<*party*:v	>mod>*beautiful*:a

Directional Similarity Measures

- Lin's measure is symmetric, but entailment is not
- How to find the direction of asymmetric relations?
  - Feature distribution (Lee, 1999)
  - Feature inclusion (Weeds and Weir, 2003; Geffet and Dagan, 2005)

symmetric: 

inclusion: 

Top-10 entailing words for *food*

symmetric (Lin 1998)	*meat, beverage, goods, medicine, drink, clothing, food stuff, textile, fruit, feed*
directional (Kotlerman et al. 2009)	*food stuff, food product, food company, noodle, canned food, feed, salad dressing, bread, food aid, drink*

Inclusion-based directional measures

- Basic inclusion formula
  (Weeds and Weir, 2003; Clarke 2009)

$$\text{Precision}(l \rightarrow r) = \frac{\sum_{f \in V_l \cap V_r} s_l(f)}{\sum_{f \in V_l} s_l(f)}$$

- Balance between symmetric and directional measures
  (Szpektor and Dagan, 2008)

$$\text{BalInc}(l \rightarrow r) = \text{Lin}(l, r) \cdot \text{Precision}(l \rightarrow r)$$

- Relative feature position instead of absolute score
  (Kotlerman et al., 2009)

$$\text{balAPinc}(l \rightarrow r) = \text{Lin}(l, r) \cdot \frac{\sum_{i=1}^{|FV_l|} \text{Prec}(i) \cdot rel(i \mid f_i \in FV_r)}{|FV_l|}$$

VERBOCEAN (CHKLOVSKI AND PANTEL, 2004)

- Pattern-based approach for broad-coverage semantic network of verbs

similarity	(*produce* :: *create*)
strength	(*permit* :: *authorize*)
antonymy	(*open* :: *close*)
happens-before	(*buy* :: *own*)
enablement	(*fight* :: *win*)

1. Start with highly associated candidate verb-pair (*fight* :: *win*)
2. Query the Web with manually constructed patterns for each relation
   - *enablement*: *X*ed * by *Y*ing the (*won by fighting the*)
   - *happens-before*: *X*ed and then *Y*ed (*fought and then won*)
3. Score each verb-pair/pattern co-occurrence (PMI)
   - A relation is considered correct if its pattern score exceeds a threshold
4. Prune based on consistency of selected relations with each other
   - "If *happens-before* is not detected, ignore detection of *enablement*"

Template-based rule acquisition from manually constructed resources

FRAMENET (BAKER ET AL., 1998)

- Conceptual structures called frames, describing prototypical situations
  - Predicates that evoke each frame
  - Semantic roles for each frame (*frame elements*)
  - Annotated sentences for many predicates
  - Semantic relations between frames – some useful for entailment

- Example: *Commercial_sell*
  - Predicates: *retail:v*, *retailer:n*, *sale:n*, *sell:v*, *vend:v*, *vender:n*
  - Frame elements: *Seller*, *Goods*, *Buyer*
  - *"We can not sell the property in Kent to Mr. Cooper"*
  - Inherits from: *Giving* , perspective on: *Commerce_goods-transfer*

Rule-sets from FrameNet

- Generate rules between FrameNet predicates with their argument mappings (Coyne and Rambo, 2009; Ben Aharon et al., 2010)

Algorithm:

1. Extract lexical entailment rules between predicates
   - Taken from FrameNet or WordNet

   cure:v → *recovery:n*

2. Add predicate argument mapping
   - Based on FrameNet elements shared between predicates

   cure $X_{Patient}$ → $X_{Patient}$*'s recovery*

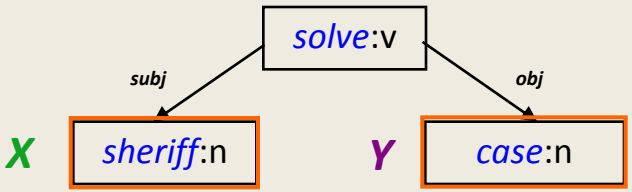
   cure of $Y_{Affliction}$ → *recovery from* $Y_{Affliction}$

Corpus-based statistical template-based rule acquisition

Distributional Similarity between Templates

- Similar to the lexical case
  - Templates – often paths in dependency parse-trees
  - Features – argument instantiations
- DIRT: (Lin and Pantel, 2001)
  1. Create a word co-occurrence vector for each variable in a binary template
  2. Templates with similar vectors are considered semantically related
     - Lin similarity measure

The sheriff solved the case



X find a solution to Y		*X solve Y*	
Slot X	**Slot Y**	**Slot X**	**Slot Y**
commission	*strike*	*committee*	*problem*
committee	*crisis*	*clout*	*crisis*
government	*problem*	*government*	*mystery*
legislator	*budget deficit*	*petition*	*woe*
sheriff	*murder*	*sheriff*	*case*

Directional Distributional Similarity for Templates

- (LEDIR: Bhagat et al., 2007)
  - Find the direction of rules learned by DIRT

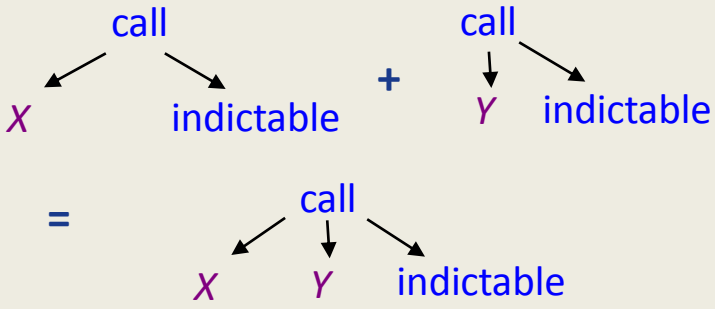
  1. Generate semantic classes for the arguments from their instantiations, based on a taxonomy (e.g. WordNet)
     - *X* own *Y*: I_Y:{*3Com*, *Sun, a car*} $\Rightarrow$ C_Y:{*software company*, *company, vehicle*}
  2. Significant context size difference indicates the rule's directionality
     - $|C(X \text{ own } Y)| > |C(X \text{ acquire } Y)| \Rightarrow X \text{ acquire } Y \rightarrow X \text{ own } Y$

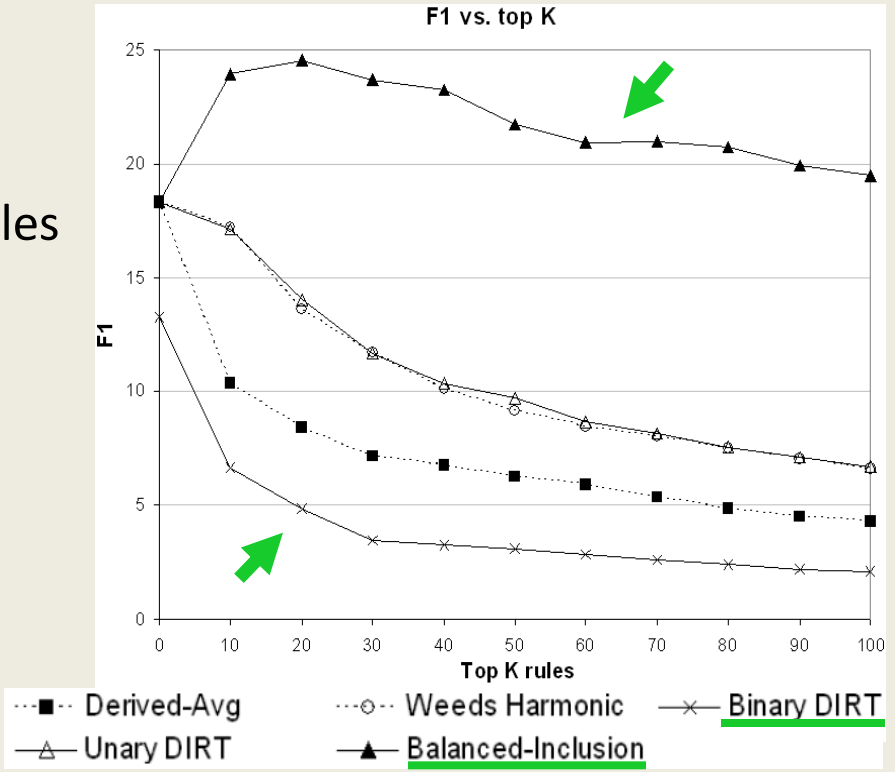
- (Szpektor and Dagan, 2008)
  - Unary templates instead of binary templates
    - *X* acquire ; acquire *Y* instead of *X* acquire *Y*
  - *Balanced-Inclusion* directional similarity
    - counter-suite against *X* → *X* sue

Directional Distributional Similarity (cont.)

IE experiment

- Directional measure outperformed symmetric measures
- Unary rules outperformed binary rules
  - Unary templates occur more
  - Unary paths are more expressive





Other Learning Methods for Template-Based Rules

- Extract rules from news articles on the same topic (Shinyama et al., 2002)
- Paraphrase using pivot languages in aligned multilingual corpora (Callison-Burch, 2008; Zhao et al., 2009)
  - Extension of the lexical case (Bannard and Callison-Burch, 2005)
- Learn rules from the Web with complex features (Szpektor et al., 2004)
- Discourse analysis for argument mapping (Pekar, 2008)
  - Related work: narrative schemas (Chambers and Jurafsky, 2009)
- Combine WordNet and distributional similarity (Szpektor and Dagan, 2009; Dinu and Wang, 2009)

Reported Resource Contributions (RTE 4,5)

Resource	Relative Resource Contribution (%)
WordNet	-2, -0.5, 0.8, 1.0, 2.5, 3.2, 4.0, 5.6
Wikipedia	-1.0, 1.0, 1.17, 1.3, 1.5, 3.3
Moby thesaurus	2.8
Acronyms	0.2, 0.3
Gazetteer	-0.8
VerbOcean	-0.2, 0.2, 0.3, 0.5
FrameNet	2.0
DIRT	-1.2 , 0.2, 0.5, 0. 7, 0.9, 1.3

http://www.aclweb.org/aclwiki/index.php?title=RTE5_-_Ablation_Tests

Entailment Rule Application

Ambiguity in Rule Application

- A rule is considered correct if it yields correct inferences when applied in valid contexts

  X charge *Y* → *X* bill *Y*

  valid context: *"Telemarketers charged the account"*
  → *Telemarketers billed the account*

  invalid context: *"Prosecutors charged Nichols with bombing"*
  ~~→~~ *Prosecutors billed Nichols*

- Problem: term disambiguation in context
  - Known problem in many NLP apps, e.g. QA, IE, RTE search task
  - Less dominant in classic RTE datasets
    - The T-H pairs were usually chosen within the same context

Unsupervised Context Models

- Task: decide whether a context is valid for rule application
  - ~~*t*~~: *Children acquire new languages*
  - *r*: *acquire* → *own*

- Typical Word Sense Disambiguation (WSD) is not enough
  - No sense-annotated training data for large-scale resources
  - Inference applicability goes beyond senses
    produce milk vs. *produce eggs* for *produce* → *lay*

- Use unsupervised context models
  - Strategy: detect contexts that are common to *lhs* and *rhs*
  - Unlike WordNet, "senses" are modelled by surface words
    - Not explicit sense-ids

Inferential Selectional Preferences (Pantel et al., 2007)

- Model valid argument instantiations by semantic classes (Resnik, 1996)
  1. Find shared instances of *lhs* and *rhs* in a corpus (*X* acquire *Y* → *X* own *Y*)
     I_X:{*HP*, *Oracle*, *Teva*} , I_Y:{*3Com*, *Sun*, *Barr*}
  2. Extract valid semantic classes of instances from a taxonomy (e.g. WordNet)
     C_X:{*company*, *software company*, *pharmaceutical company*}

- A lhs occurrence is valid if its arguments belong to valid classes

  Microsoft acquired Farecast for $115M

  {*X:company*} {*Y:company*}

  Children acquire new languages quickly

  {*X:juvenile, person*} {*Y:communication*}

- Additional work:
  - (Agirre and Martinez, 2002): selectional preferences for WordNet verb classes
  - (Erk and Pado, 2008): word meaning as a structured combination of vectors

Unsupervised Classifier Training

- Instead of heuristic context matching criteria, train a classifier
  - No labeled data → unsupervised training-set generation

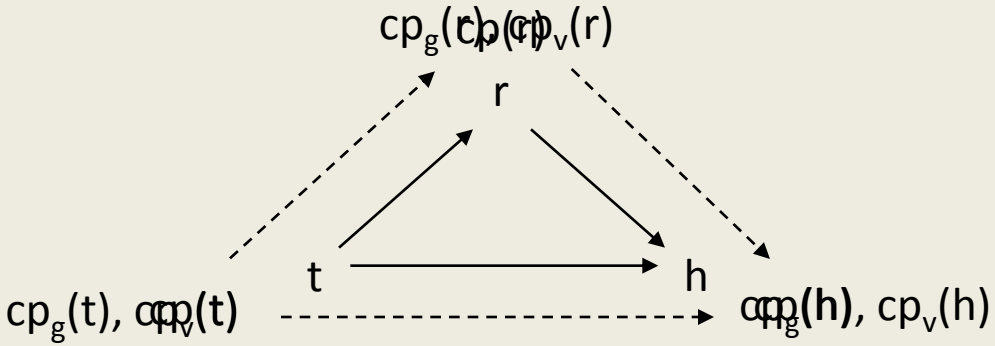
A local classifier for each rule (Kauchak and Barzilay, 2006; Bergsma et al., 2008)

- Context features: terms in a window around a word
- Positive examples: sentences containing *rhs* (*difficult* → *hard*)
  - *It is hard to believe that the …*
- Negative examples: sentences not containing *rhs*
  - *… apply this method to …*
- Task: classify the context features around an *lhs* occurrence
  - *It is difficult to believe this kind …*

- A global classifier for all rules (Connor and Roth, 2007)
- Features: similarity measures
  - Similarity between *lhs* and *rhs* training contexts and a tested sentence
  - Classifier learns to combine similarity measures
- Positive and negative examples by bootstrapping from local classifiers

CONTEXTUAL PREFERENCES

(SZPEKTOR ET AL., 2008)



- Enrich object representation with contextual information, denoted *cp*( )
- CP are intended to constrain or disambiguate object meaning

- During inference, CP should match as well (on top of structural matching)

- Two components within *cp*( ):
  - cp_v( ): preferences or constraints on object's variable instantiations
    cp_g( ): global ("topical") context in which an object typically occurs

CONTEXTUAL PREFERENCES

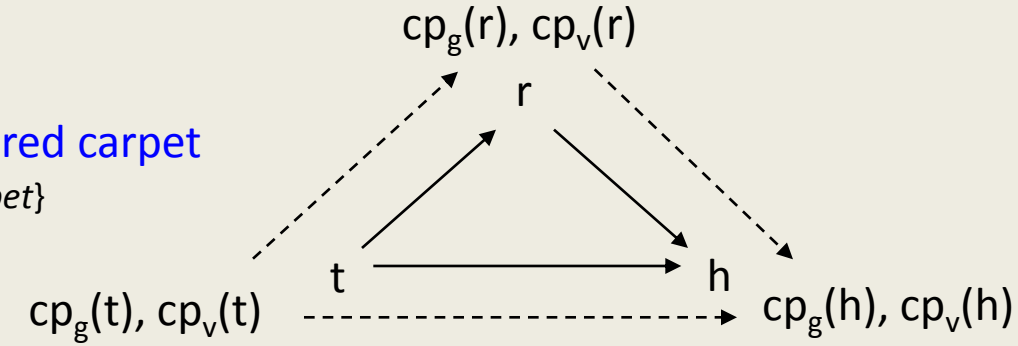
(SZPEKTOR ET AL., 2008)

X lay *Y* ⇒ *X* produce *Y*

cp_v(r:*Y*) = {*eggs*}

X accuse *Y* ⇒ *X* attack *Y*

cp_g(r) = {*criticism,decision*}



Bengal to lay red carpet

cp_v(t:*Y*) = {*carpet*}

X attack *Y*

cp_g(h) = {*war,injury*}

Children acquire new languages

cp_v(t:*X*) = {*child*}
cp_v(t:*Y*) = {*language*}

X acquire *Y*

cp_v(h:*X*) = {*company*}
cp_v(h:*Y*) = {*company*}

- Two components within *cp*( ):
  - cp_g( ): global ("topical") context in which an object typically occurs
  - cp_v( ): preferences on possible variable instantiations

ENTAILMENT RULE EVALUATION

Manual Evaluation Approaches

Rule-based approach

- Human annotators evaluate the correctness of each rule
  - The judge should think of reasonable contexts under which the rule holds
- Correctness criterion is not well-defined and hard to apply
  ⇒ low annotator agreement (*X* set *Y* →? *X* allow *Y*)

Instance-based approach (Pantel et al., 2007; Szpektor et al., 2007)

- Judges evaluate the correctness of rule applications
- Rule application correctness criterion follows textual entailment
  - *The committee set the following refunds on Monday* →?
    The committee allow the following refunds
- Easier and well-defined setup for annotators ⇒ higher annotator agreement

Application-based evaluation

- Manual evaluation limitations:
  - Measures only rule accuracy, not rule coverage
  - All rules are equal, ignoring rule frequency
  - Hard labor

Application-based evaluation

- Measure rule-set contribution to entailment system performance
- Need setups that isolate rule-set performance from other system components:
  - Simple test hypotheses
  - Gold standard annotation of texts inferring the hypotheses in a corpus
    $\Rightarrow$ coverage (*recall*) and accuracy (*precision*)
  - Entailment systems whose components are decoupled

➔ Current RTE datasets should not be the primary test-bed for application-based evaluation

Successful Example: ACE Evaluation

- ACE event extraction dataset:
  - 33 target events, e.g. *Attack*, *Marry*, *Sue*, and their arguments
    - *Attack*: *Attacker*, *Target*, *Time*
  - All event mentions are annotated in a corpus: *recall* + *precision*

- Utilize ACE for rule-set evaluation (Szpektor and Dagan, 2008)
  - Map events to seed templates
    - $X_{Attacker}$ *attack* ; *attack* Y_{Target}
  - Match seeds or entailing templates from a tested rule-set in the corpus
    assult Y → *attack* Y_{Target} :: "*Police stations were assaulted by insurgents*"

- Useful for error analysis
  - Example: reasons for incorrect extractions by binary DIRT

	Invalid Context	Partial Template	Incorrect	Total
# rules	16	27	157	200
# incorrect applications	70	**2665**	2584	5319

(partial lhs template: *take* ~~→~~ *arrest* vs. *take into custody* → *arrest*)

Part V:

Research Directions in RTE

Mark Sammons

Idan Szpektor

V.G. Vinod Vydiswaran

Some Conclusions

- Structure is important in understanding text!
  - Systems using structure do much better than lexical overlap measures, even with n-grams/semantic similarity measures
  - Need (reliable!) richer structure for enrichment/application of background knowledge
- Contradiction is not a subset of Unknown
- Rigid two-stage system problematic

More Conclusions

- Need good quality, broad coverage entailment knowledge resources
  - Currently, a lot of targeted knowledge engineering
- Current systems good for backing off when knowledge missing – next step is to provide/incorporate that knowledge
- Much engineering effort:
  - Big barrier to entry
  - Open-source RTE engine framework very desirable, but non-trivial effort needed

Current systems: "2nd wave"

- 1st wave:
  - lexical overlap approach
  - pushed using lexical similarity resources
  - Intuition: irrelevant word in H means T != H
  - Intuition 2: structure not very important
- 2nd wave:
  - shallow structure constrains lexical comparisons
  - locally engineered, noisy knowledge
  - Shallow, broad coverage lexical resources
  - Intuition: structure is important, but deep structure is not sufficiently reliable

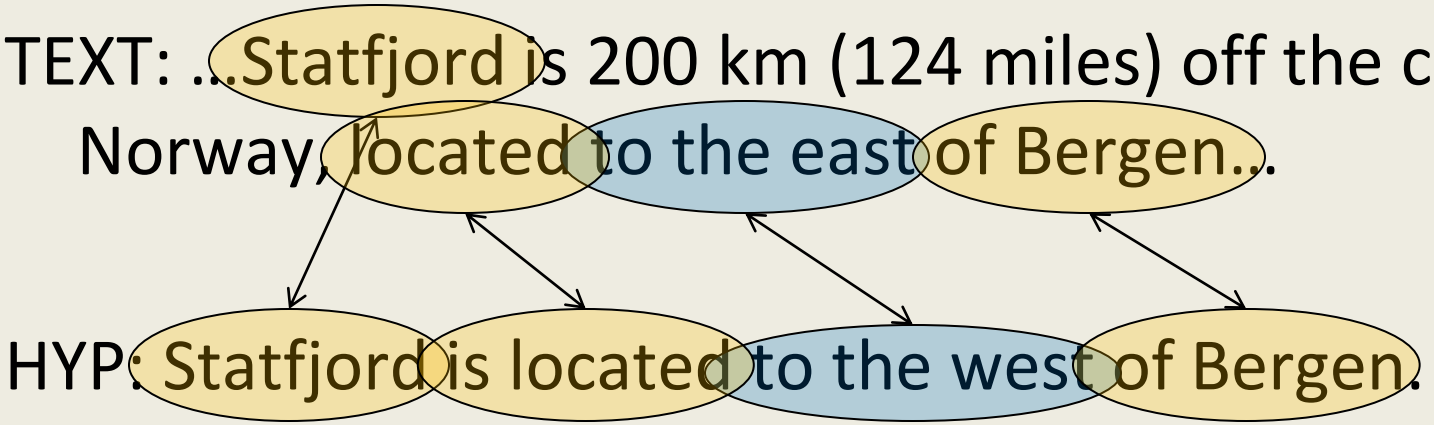
The "3rd Wave"

- Deeper structure
  - Move beyond sentence boundaries
  - Improve precision (can we maintain recall? cf. Fail examples from error analysis)
  - Employ more structured (and therefore precise) knowledge
  - More informed "alignment" model
- Need (better) tools for extracting deep structure!
  - More applicable discourse models/resources
- Need (better) knowledge resources & acquisition techniques!

WHERE DO WE WANT TO BE?

- TE as plug-and-play inference engine
  - Textual inference service for other NLP tasks
- Well-defined RTE subtasks
  - Incl. focused development of RTE knowledge resources
  - Needs definition, data, and evaluation methodology
  - Lower barrier to entry to RTE research (don't build a new system from scratch)
- Improved evaluation setting for Knowledge resources and RTE subtasks
  - Confidence that if you solve a problem, can demonstrate meaningful improvement (i.e., can publish!)

ID: 5T-582: CONTRADICTION

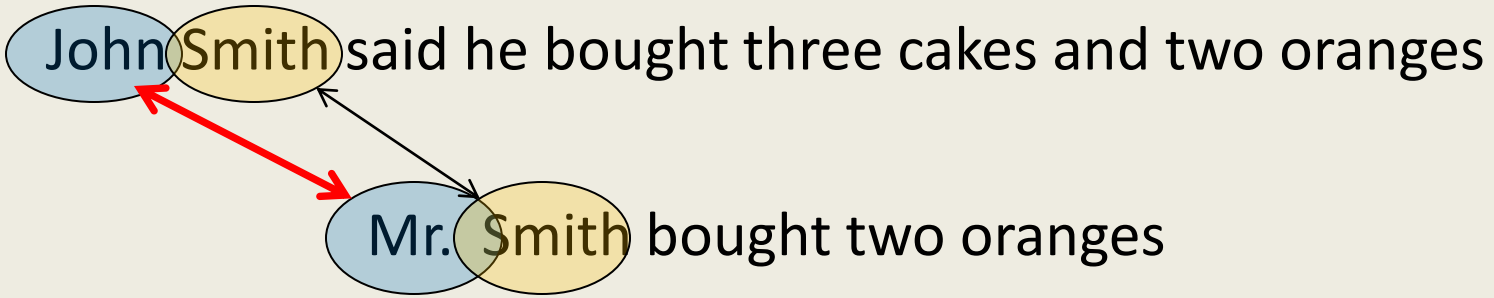


From alignment perspective, this is simple,
given correct predicate-argument structure

(and given structure-first approach)

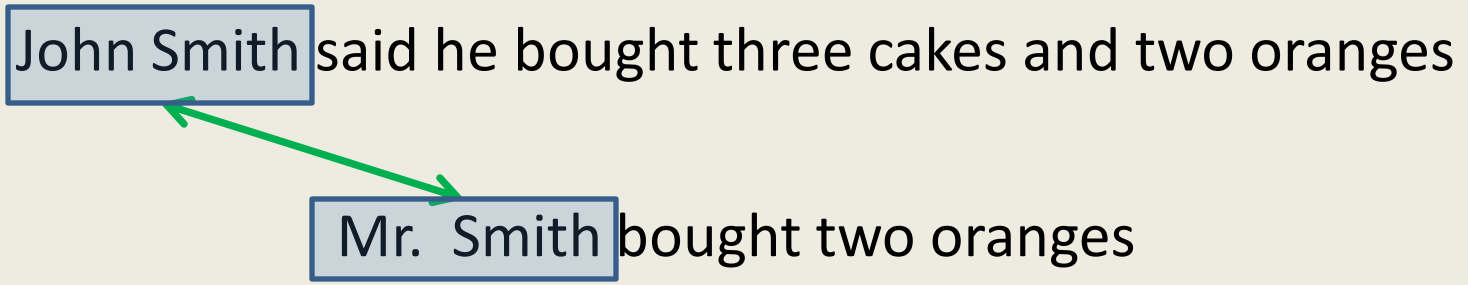
Alignment and Metrics

- The metric abstraction fits nicely into alignment model
  - Pair specialized annotation (constituent type) with specialized comparison resource



Alignment and Metrics

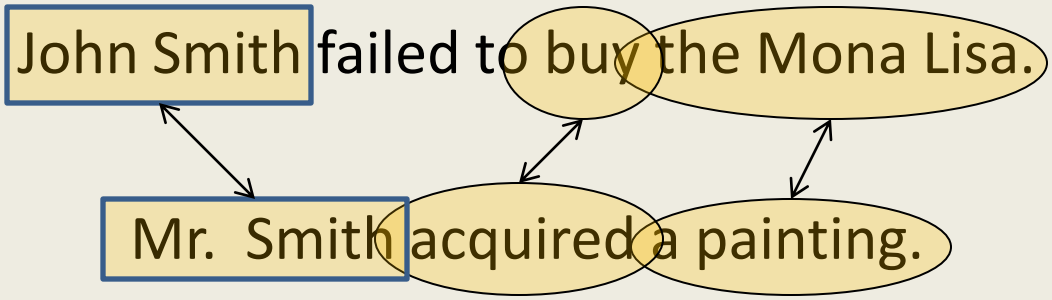
- The metric abstraction fits nicely into alignment model
  - Pair specialized annotation (constituent type) with specialized comparison resource



  - Cf. Named Entity similarity metrics, lexical similarity metrics…

Alignment and Rules

- The enrichment approach is compatible: makes implicit structure available for alignment



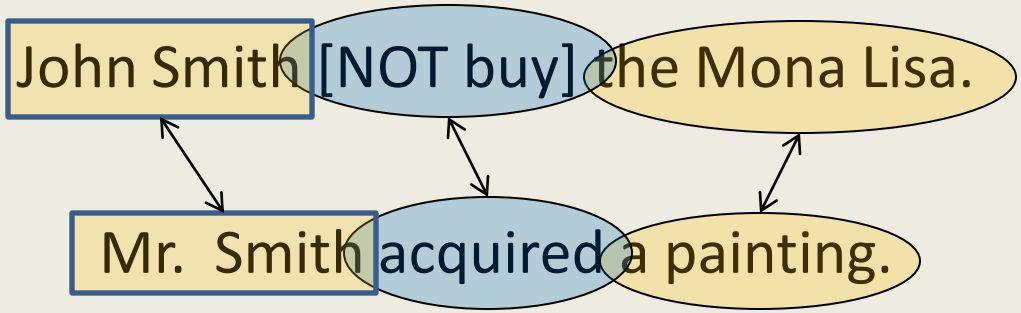
[X]/NP fail to [Y]/VP

=>

X [NOT Y]

Alignment and Rules

- The enrichment approach is compatible: makes implicit structure available for alignment



Alignment and Entailment

- Alignment is a nice abstraction for handling:
  - logical modifiers
  - polarity and monotonicity
  - contradiction

  -- assuming you have the necessary resources
- 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

What's in a metric?

- Ideally, use Natural Logic abstractions (MacCartney et al., 2008, 2009) to investigate/characterize semantic behavior
  - Semantic containment, monotonicity, exclusion, implicativity
  - Analogous to set relations
- Organizes knowledge resources/ontologies by defining a set of coarse, but universal, relations

Venn	symbol	name	example
	P = Q	equivalence	*couch* = *sofa*
	P ⊏ Q	forward entailment (strict)	*crow* ⊏ *bird*
	P ⊐ Q	reverse entailment (strict)	*European* ⊐ *French*
	P ^ Q	negation (exhaustive exclusion)	*human* ^ *nonhuman*
	P | Q	alternation (non-exhaustive exclusion)	*cat* | *dog*
	P _ Q	cover (exhaustive non-exclusion)	*animal* _ *nonhuman*
	P # Q	independence	*hungry* # *hippo*

NatLog and Metrics

- So far, systems focus on:
  - Equivalence (synonymy, paraphrase)
  - Containment (hypernymy, entailment)
  - (Simple) exclusion (negation, antonymy)
  - (Simple) cover (modal constructions)
- The NatLog operators provide focused direction for research; scale to structures as well as words
- Interactions/scope behaviors described by NatLog
  - Some phenomena (e.g. monotonicity/ quantifiers) not well represented in existing RTE corpora
  - Still important for general textual inference capabilities

Alignment as Framework

- Alignment is a natural framework for thinking about RTE
  - Intuitive model for contradiction vs. entailed and vs. unknown
  - Supports localizing knowledge as metrics, enrichment resources, structural normalization
  - Separate logical semantics from structural/lexical similarity
- Use NatLog operators to define metric behaviors
- Define RTE framework in these terms, promote development of framework and components

Proposal: Alignment-based RTE framework

- General framework is widely acknowledged as desirable in the RTE community
- Needs to accommodate wide range of approaches
  - **Most approaches have similar macro structure**
  - **Alignment is a dominant paradigm**
- Needs to support evaluation of impact of resources
  - **Metrics, enrichment are natural abstractions for focusing research**
  - **Many inference processes can sit on top of same alignment structure**
- Ideally, needs theoretical basis (cf. syntactic parsing)
  - **Natural Logic offers a good basis for practical, systematic textual inference**

Beyond Metrics and Natural Logic

- Implicit constituents ("traces", etc.) and relations are not recognized by standard NLP tools
  - **Most existing NLP resources work at level of verbs and nouns**
  - **Ellipsis cuts across multiple levels of linguistic analysis**
  - **Does not make sense to express all possible elisions in e.g. syntactic rules**
- Reliable deep structure is crucial
  - **Integral resource for alignment-based approaches**
  - **Facilitates normalization for global similarity approaches**
- Good domain adaptation framework would be a valuable resource
  - **Many NLP tools anecdotally underperform on RTE data**
  - **presumably due to domain shift**

Inference-based analysis

- (Sammons et al., 2010), analyzed 210 examples from RTE 5 using standardized inference process
- Identified relevant linguistic and semantic phenomena in Entailed, Contradicted, and Unknown entailment examples

Phenomenon	Coverage
Coreference, zero- and bridging anaphora	~35%
"Simple" entailment rules	~35%
Implicit relations	~25%

PROGRESS

- Some efforts to develop generic frameworks
  - lexical (LLM),
  - dependency structure (EDITS)
- Theoretical model for textual inference – Natural Logic – has been proposed
- Community buy-in
  - Machine Reading
  - SemEval: Parsing and Noun Compound resolution framed as Entailment Recognition
  - CLEF Answer Validation Exercise (RTE in QA)
  - EVALITA (Italian NLP evaluation)

LAST WORD

- Textual Entailment: a general paradigm for semantic understanding
- Alignment + syntactic structure : state-of-art
  - Applicable for many real world applications – QA, IE, paraphrase detection
- Modular components that handle smaller well-defined sub-problems well.
- Incorporating knowledge resources essential

FUTURE STEPS

- Discourse-level analysis of text
  - Includes annotating text across sentences
- Need to additional broad-scoped knowledge resources
  - Think how best to define, collect, and evaluate
- Alignment framework and plug-and-play style architecture

What we covered today

- Described the RTE task and its relevance as framework for textual inference
- Surveyed research in RTE in the context of the PASCAL/NIST TAC challenges
- Identified and explained the main approaches to inference in RTE:
  - Global similarity and Alignment
- Surveyed approaches to knowledge acquisition and application; proposed model for entailment rules
- Identified key areas for future RTE research

USEFUL RTE RESOURCES

- ACL Textual Entailment Portal
  - Systems, data, knowledge resources, publications
  - http://www.aclweb.org/aclwiki/index.php?title=Textual_Entailment_Portal
- CLEF Answer Validation Exercise
  - Spanish and English RTE corpora available
  - http://nlp.uned.es/clef-qa/ave/
- EVALITA
  - Italian NLP evaluation effort, including RTE
  - http://evalita.fbk.eu/
- NIST Text Analysis Conference
  - RTE Challenges, corpora, publications
  - http://www.nist.gov/tac/

THANK YOU

http://cogcomp.cs.illinois.edu/presentations/RTE_NAACL_2010.zip

MARK SAMMONS	mssammon@illinois.edu
IDAN SZPEKTOR	idan@yahoo-inc.com
V.G. VINOD VYDISWARAN	vgvinodv@illinois.edu

BIBLIOGRAPHY

- Agirre and Martinez, 2002
- Auer et al., 2007
- Baker et al., 1998
- Bannard and Callison-Burch 2005;
- Bar-Haim et al., 2006;
- Bar-Haim, et al, 2007
- Bayer et al., 2005
- Ben Aharon et al., 2010
- Bentivogli, et al., 2009
- Bergmair, 2009
- Bergsma et al., 2008
- Bhagat et al., 2007
- Bos and Makert, 2006
- Braz et al., 2005
- Callison-Burch, 2008
- Chambers and Jurafsky, 2009
- Chambers et al., 2007
- Chang, et al., 2010
- Chklovski and Pantel, 2004
- Clark and Harrison, 2009
- Clarke, 2009
- Connor and Roth, 2007
- Coyne and Rambow, 2009
- Dagan, Glickman, and Magnini, 2006
- deMarneffe et al., 2007
- de Marneffe, Rafferty, and Manning, 2008
- Dinu and Wang, 2009
- Erk and Pado, 2008
- Geffet and Dagan, 2005
- Giampiccolo et al., 2007
- Harabagiu and Hickl, 2006
- Hickl and Bensley, 2007
- Iftene and Moruz, 2009
- Harris, 1954
- Iftene and Balahur-Dobrescu, 2008
- Kauchak and Barzilay, 2006

BIBLIOGRAPHY

- Kotlerman et al., 2009
- Kozareva et al., 2008
- Lee, 1999
- Li, et al., 2009
- Lin, 1998
- Lin and Pantel, 2001
- MacCartney and Manning, 2007, 2008, 2009
- Mehdad , et al., 2009
- Mehdad, Zanzotto, and Moschitti, 2009
- Miller, 1995
- Mirkin, et al., 2009
- Moldovan and Rus, 2001
- Novischi and Moldovan, 2006
- Pado, et al., 2009
- Pantel et al., 2007
- Pekar, 2008
- Resnik, 1996
- Roth and Sammons, 2007
- Roth, Sammons, and Vydiswaran, 2009
- Sammons et al., 2009
- Sammons, Vydiswaran, and Roth, 2010
- Shinyama et al., 2002
- Shnarch et al., 2009
- Suchanek et al., 2007
- Szpektor et al., 2004
- Szpektor et al., 2007
- Szpektor et al., 2008
- Szpektor and Dagan, 2008, 2009
- Tatu et al., 2006
- Wang and Zhang, 2009
- Wang, Zhang, and Neumann, 2009
- Weeds and Weir, 2003
- Zhao et al., 2009
- Zanzotto and Dell'Arciprete, 2009
- Zanzotto and Moschitti, 2006