Natural Language Inference in Natural Language Terms

Ido Dagan
Bar-Ilan University, Israel

BIU NLP lab - Acknowledgments

What is **inference**?

From *dictionary.com*:

- *inferring*:
  to derive by reasoning; conclude or judge from premises or evidence.

- *reasoning*:
  the process of forming conclusions, judgments,
  or inferences from facts or premises.

What is **inference**?

From *dictionary.com*:

- ***inferring***:
  to **derive** by reasoning; **conclude** or judge **from premises** or evidence.

- *reasoning*:
  the process of **forming conclusions**, judgments, or **inferences from** facts or **premises**.

Textual Inference

- Inferring new *textual expressions* from given ones
- Captures two types of inference:
  1. Inferences about the “extra-linguistic” world
     - *it rained yesterday => it was wet yesterday*
  2. Inferences about *language variability*
     - *I <u>bought</u> a watch => I <u>purchased</u> a watch*
- No definite boundary between the two

Textual Entailment – a definition capturing textual inference

- 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:
  - Human gold standard
  - Assuming common background knowledge
    - Language & world knowledge

Motivation: Inference in Applications

QA:

Question: What affects blood pressure?

"Salt causes an increase in blood pressure"

IE: X purchase Y

IBM	Coremetrics
Google	reMail
Yahoo	Overture

IR:

Query: symptoms of IBS

"IBS is characterized by vomiting"

Entailment in Multi-document Summarization

Hypothesis

Obama gave a speech last night in the Israeli lobby conference

In his speech at the American Israel Public Affairs Committee yesterday, the president challenged …

Barack Obama's AIPAC address yesterday ...

Texts

Appeal of *textual entailment* definition

- Became a prominent view on textual inference
  - RTE 1-7; 1950 hits in Google Scholar

- Much more concrete than:
  - "paraphrase"
    - *bi-directional entailment / equivalence*
    - "partial highly-covering entailment"
  - "similarity" – very **vague** (non-scientific?) notion

- Additional textual inference types may be defined
  - But they should be **defined**, reasonably precisely

Evaluation: PASCAL RTE Challenges

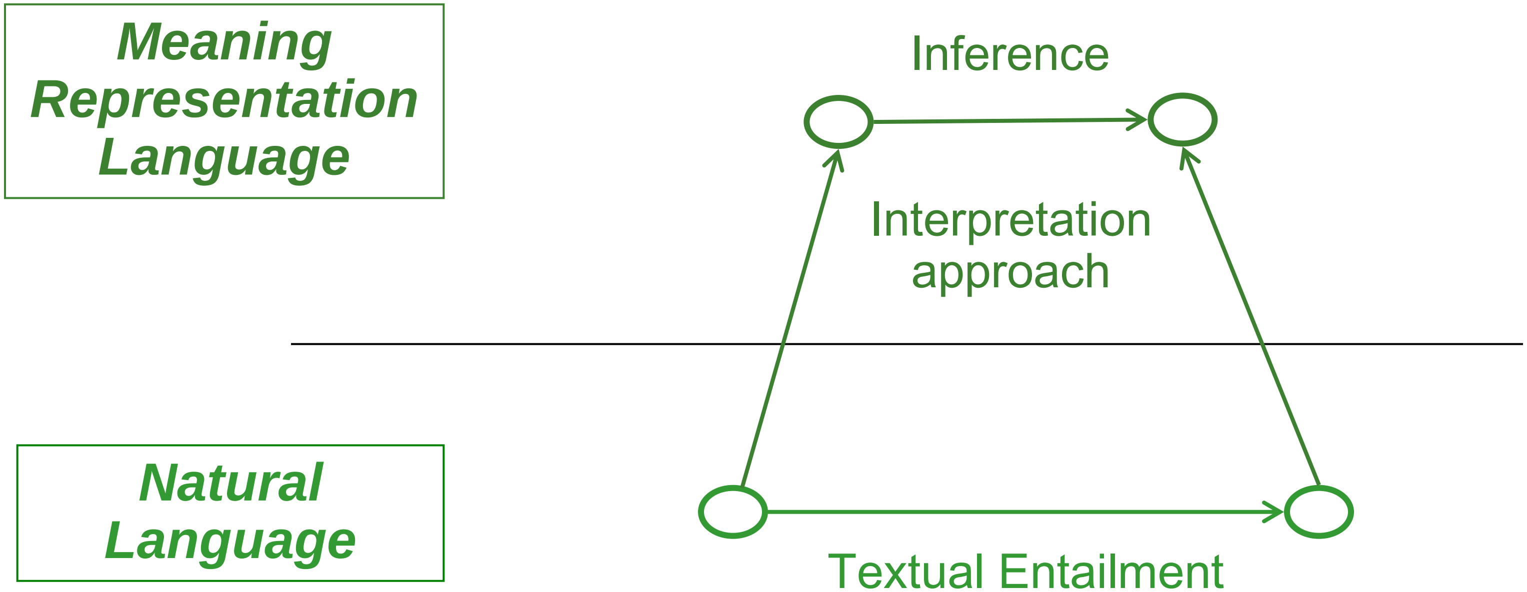
	TEXT	HYPOTHESIS	TASK	ENTAIL-MENT
1	*Regan attended a ceremony in Washington to commemorate the landings in Normandy.*	*Washington is located in Normandy.*	IE	False
2	*Google files for its long awaited IPO.*	*Google goes public.*	IR	True
3	*….: a shootout at the Guadalajara airport in May, 1993, that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

- Created utilizing (or simulating) reductions from real systems' output

Initial use of RTE systems in applications

- QA
  - Harabagiu & Hickl, ACL-06
  - Answer Validation Exercise (AVE) at CLEF
  - QallMe (FBK-irst, Magnini et al.)
- Relation extraction
  - Romano et al., EACL-06
- Educational applications
  - Nielsen et al., ACL-08 education workshop, SemEval/RTE-8
- Summarization
  - Harabagiu et al. 2007, Information Processing and Management
- MT evaluation and paraphrasing for MT (two ACL-2009 papers)

The Textual Entailment Task vs. Classical Approach to Inference

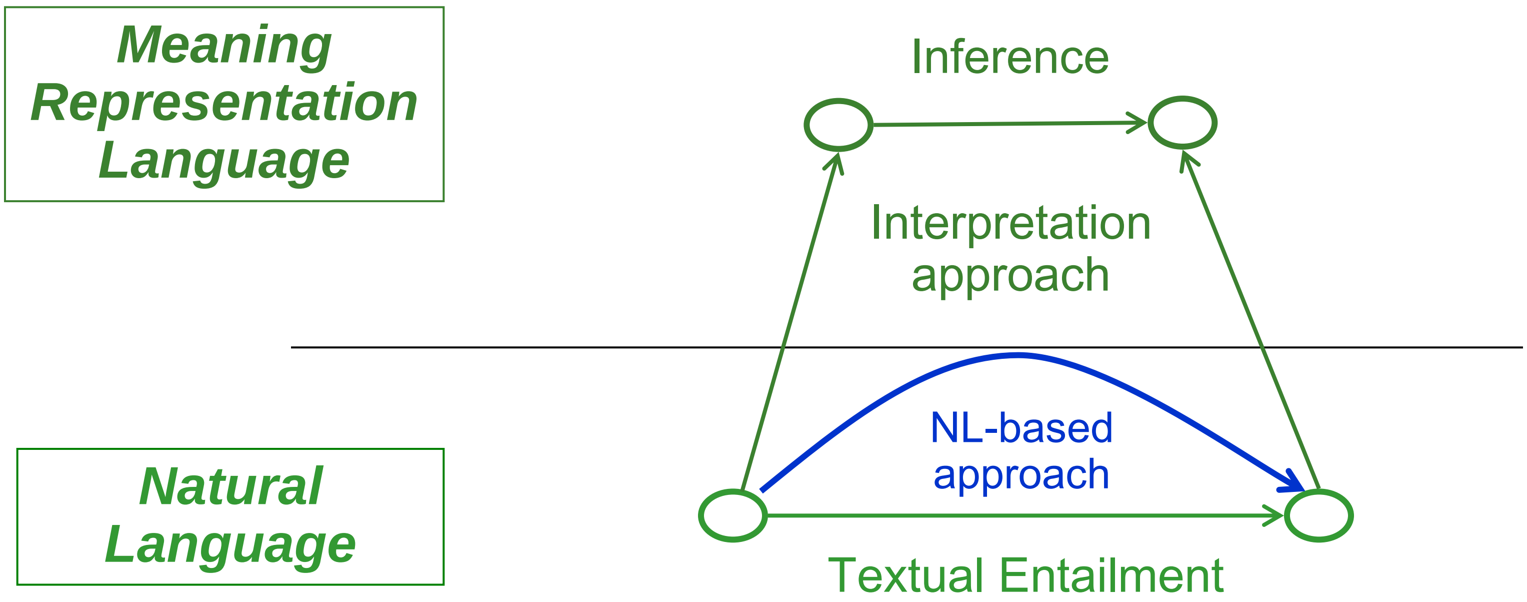


- Classical logic-based approach:
  - *Natural language isn't suitable for conducting inference*
    - *Too vague, ambiguous, …*
  - *We need to invent artificial languages (logics) that support inference*

Textual inferences in practice – are based on NL representations

- Recognizing target expressions (QA, IE, …)
  - Lexical substitutions and similarity
  - Matching syntactic and predicate patterns, semantic roles
  - Machine learning based on linguistic features
- Co-reference chains for discourse objects
- “Natural” name classes (vs. abstract classes)
  - “football player”, “coffee producer”, …

How should computers infer?



- Alternative language-based approach:
  - *- Perform many inferences over natural language representations*
  - *- May resort to extra-linguistic representations/inference when needed*

Appeal of NL representations

If people think in NL, why shouldn't computers?...

- Saves the need of logic interpretation
  - And the need to invent (and agree on!) logics
  - NL representations are consensual and obtainable

- Easier to acquire inference knowledge
  - Particularly with unsupervised learning methods
    - A great challenge – more later…

Desiderata

1. Develop **principled & practical** inference over NL representations
   - Analogous to principled "logics" (learning based)
   - Most current applied inferences are ad-hoc
     (in RTE or application-specific)

2. Develop methods for acquiring vast inference knowledge
   - Represented in language structures

3. Develop generic platforms/engines that implement both of the above

** Other fields as role models: MT, parsing – similar investment needed!*

Principled Learning-based Inference Mechanisms

- *over language structures*

BIUTEE

Knowledge and Tree-Edits in Learnable Entailment Proofs

Asher Stern and Ido Dagan

(earlier partial version by Roy Bar-Haim)

Download at: http://www.cs.biu.ac.il/~nlp/downloads/biutee

Transformation-based Inference

- Sequence of transformations (A proof)

  $$T = T_0 \rightarrow T_1 \rightarrow T_2 \rightarrow \ldots \rightarrow T_n = H$$

  - Tree-Edits
    - Complete proofs – by limited pre-defined set of operations
    - Estimate confidence in each operation
  - Knowledge based Entailment Rules
    - Arbitrary knowledge-based transformations
    - Formalize many types of knowledge

Transformation based RTE - Example

$$T = T_0 \rightarrow T_1 \rightarrow T_2 \rightarrow \ldots \rightarrow T_n = H$$

Text: The boy was located by the police.

Hypothesis: Eventually, the police found the child.

Transformation based RTE - Example

$$T = T_0 \rightarrow T_1 \rightarrow T_2 \rightarrow \ldots \rightarrow T_n = H$$

Text: The boy was located by the police.

⇩

<u>The police located the boy.</u>

⇩

The police <u>found</u> the boy.

⇩

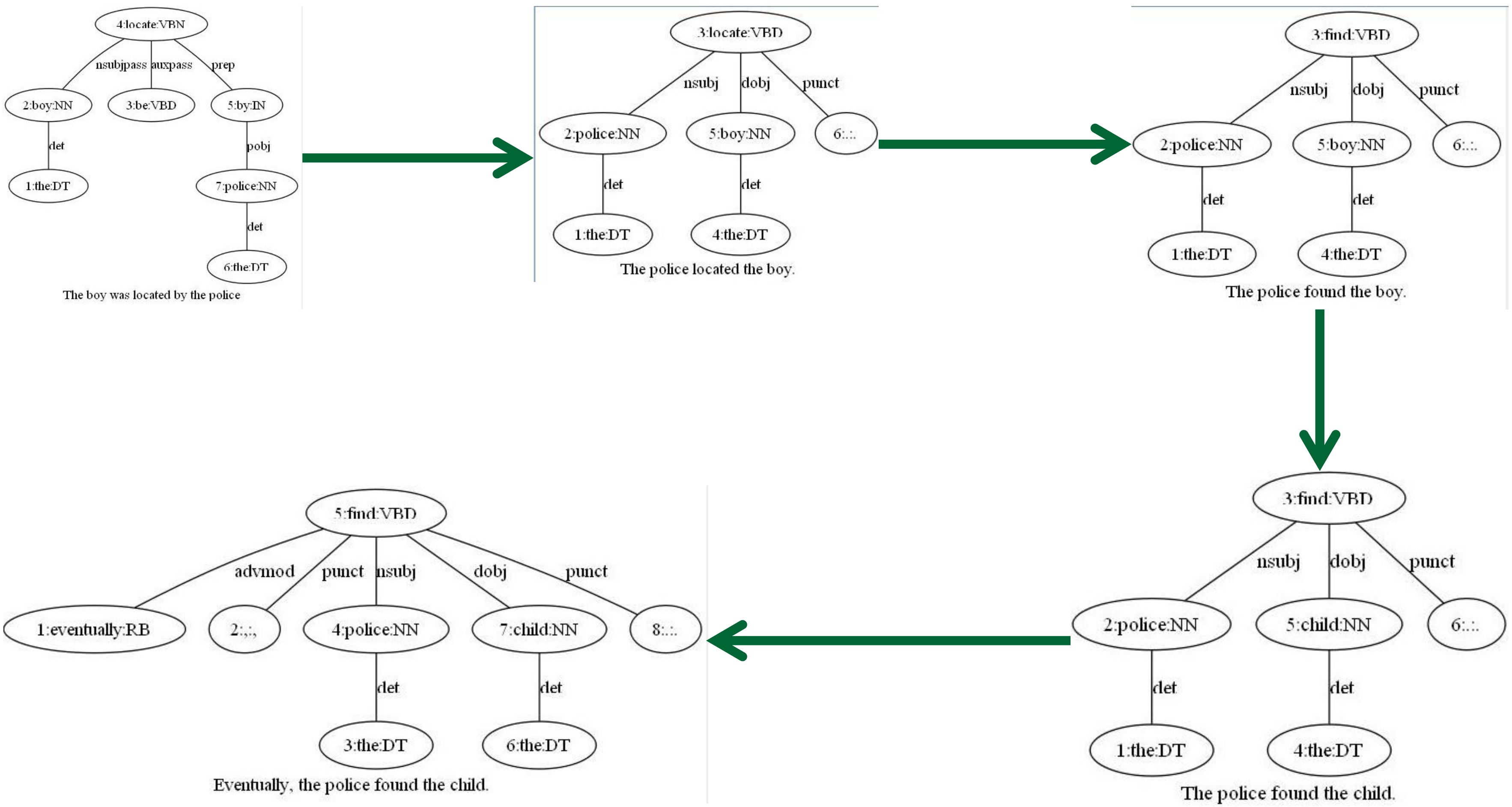
The police found the <u>child</u>.

⇩

Hypothesis: <u>Eventually</u>, the police found the child.

Transformation based RTE - Example

$$T = T_0 \rightarrow T_1 \rightarrow T_2 \rightarrow \ldots \rightarrow T_n = H$$



BIUTEE's Inference Formalism

- Analogy to logic proof systems:

Propositions	*Parse Trees*
Inference Steps	*Tree transformation/generation*
Proof	*Sequence of generated trees:* $T \ldots T_i \ldots H$

BIUTEE Goals

- Rely on Entailment Rules
  - Supported by many types of knowledge
- Tree Edits
  - Allow complete proofs
- BIUTEE
  - Integrates the benefits of both
  - Estimate confidence of both

Challenges / System Components

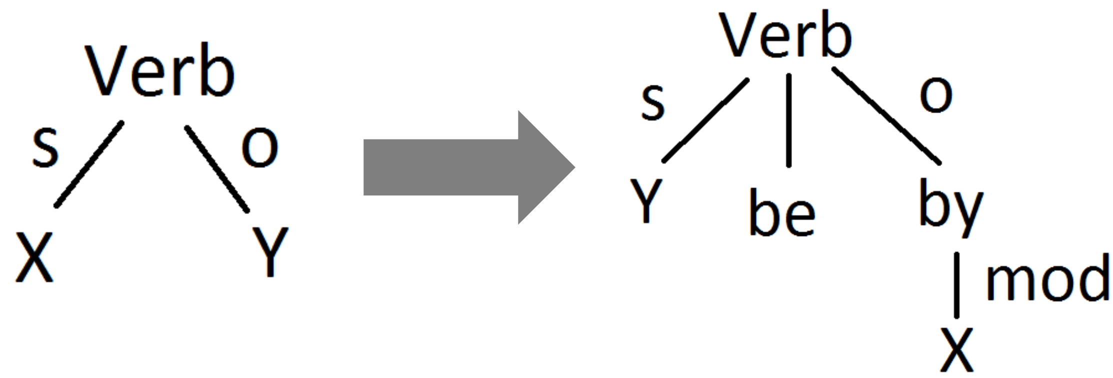
How to …

1. generate linguistically motivated complete proofs?
2. estimate proof confidence?
3. find the best proof?
4. learn the model parameters?

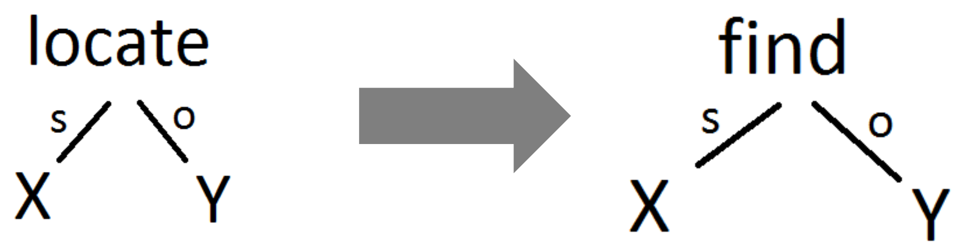
1. Generate linguistically motivated complete proofs

Knowledge-based Entailment Rules

Generic Syntactic



Lexical Syntactic



Lexical

boy → child

Bar-Haim et al. 2007. ***Semantic inference at the lexical-syntactic level***.

Extended Tree Edits (On The Fly Operations)

- Predefined custom tree edits
  - Insert node on the fly
  - Move node / move sub-tree on the fly
  - Flip part of speech
  - …
- Heuristically capture linguistic phenomena
  - Operation definition
  - Features – to estimate confidence

Proof over Parse Trees - Example

$$T = T_0 \rightarrow T_1 \rightarrow T_2 \rightarrow \ldots \rightarrow T_n = H$$

Text: The boy was located by the police.

Passive to active

The police located the boy.

X locate Y → X find Y

The police found the boy.

Boy → child

The police found the child.

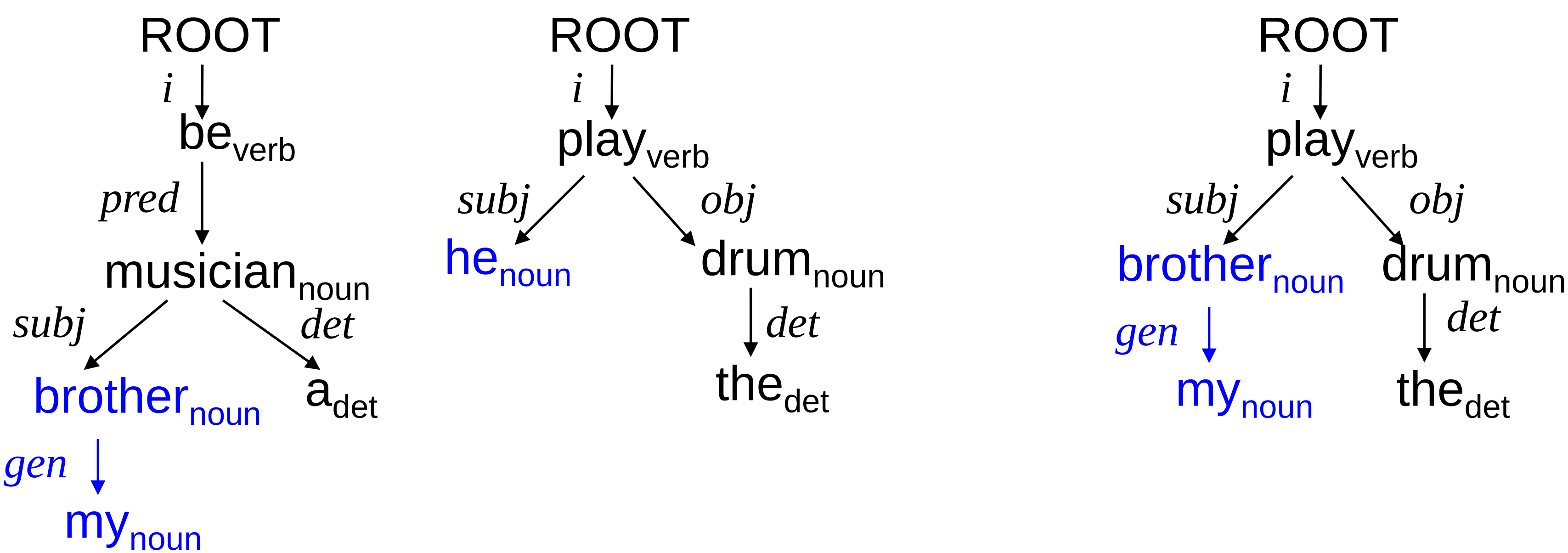
Tree-edit insertion

Hypothesis: Eventually, the police found the child.

Co-reference Substitution

- For co-referring subtrees S_1 , S_2:
  - Copy source tree containing S_1 while replacing it with S_2

__My brother__ is a musician.
__He__ plays the drums. ⇒ *__My brother__ plays the drums.*



2. Estimate proof confidence

Cost based Model (Variant of Raina et al., 2005)

- Define **operation cost**
  - Represent each operation as a feature vector
  - Cost is linear combination of feature values
- Define **proof cost** as the sum of the operations' costs
- Classify: *entailment* if and only if proof cost is lower than a threshold

Feature vector representation

- Define **operation feature value**
  - Represent each operation as a feature vector

Features (Insert-Named-Entity, Insert-Verb, … , WordNet, Lin, DIRT, …)

An operation

The police located the boy.

DIRT: *X locate Y → X find Y (score = 0.9)*

The police found the boy.

A downward function of score

(0,0,…,0.257,…,0)

Feature vector that represents the operation

Cost based Model

- Define **operation cost**
  - Cost is standard linear combination of feature values

Cost = weight-vector * feature-vector

- Weight-vector is learned automatically

$$C_w(f(o)) = w^T \cdot f(o)$$

Confidence Model

- Define **operation cost**
  - Represent each operation as a feature vector
- Define **proof cost** as the sum of the operations' costs

$$C_w(P) \equiv \sum_{i=1}^{n} C_w(o_i) = \sum_{i=1}^{n} w^T \cdot f(o_i) = w^T \cdot f(P)$$

Cost of proof

Weight vector

Vector represents the proof. Define $\sum_{i=1}^{n} f(o_i) \equiv f(P)$

Feature vector representation - example

$$T = T_0 \rightarrow T_1 \rightarrow T_2 \rightarrow \ldots \rightarrow T_n = H$$

(0,0,……………………….,1,0)

\+

(0,0,………………0.457,..,0,0)

\+

(0,0,..…0.5,.……………….,0,0)

\+

(0,0,1,……………………..,0,0)

=

(0,0,1..0.5.…0.457,....…1,0)

Text: The boy was located by the police.

Passive to active

The police located the boy.

X locate Y → X find Y

The police found the boy.

Boy → child

The police found the child.

Insertion on the fly

Hypothesis: Eventually, the police found the child.

Cost based Model

- Define **operation cost**
  - Represent each operation as a feature vector
- Define **proof cost** as the sum of the operations' costs
- Classify: "entailing" if and only if proof cost is smaller than a threshold

$$w^T \cdot f(P) < b$$

Learn

3. Find the best proof

Search the best proof

T - H

Proof #1	$T \rightsquigarrow H$
Proof #2	$T \rightsquigarrow H$
Proof #3	$T \rightsquigarrow H$
Proof #4	$T \rightsquigarrow H$

Search the best proof

$T \rightarrow H$		$T \nrightarrow H$	
Proof #1	$T \rightsquigarrow H$ ✗	Proof #1	$T \rightsquigarrow H$ ✗
Proof #2	$T \rightsquigarrow H$ ✓	Proof #2	$T \rightsquigarrow H$ ✗
Proof #3	$T \rightsquigarrow H$ ✗	Proof #3	$T \rightsquigarrow H$ ✗
Proof #4	$T \rightsquigarrow H$ ✗	Proof #4	$T \rightsquigarrow H$ ✗

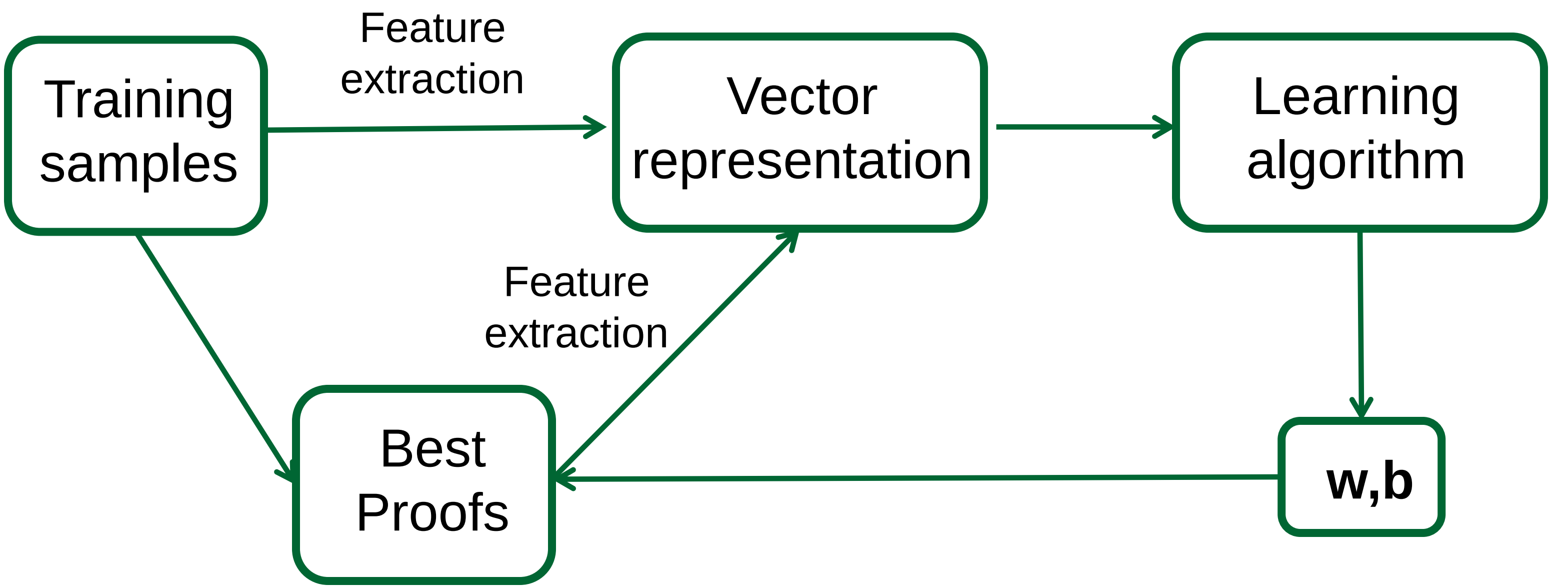
- Need to consider the “best” proof for the positive pairs
- “Best Proof” = proof with lowest cost
  - Assuming a weight vector is given
- Search space exponential – AI-style search (ACL-12)
  - Gradient-based evaluation function
  - Local lookahead for “complex” operations

4. Learn model parameters

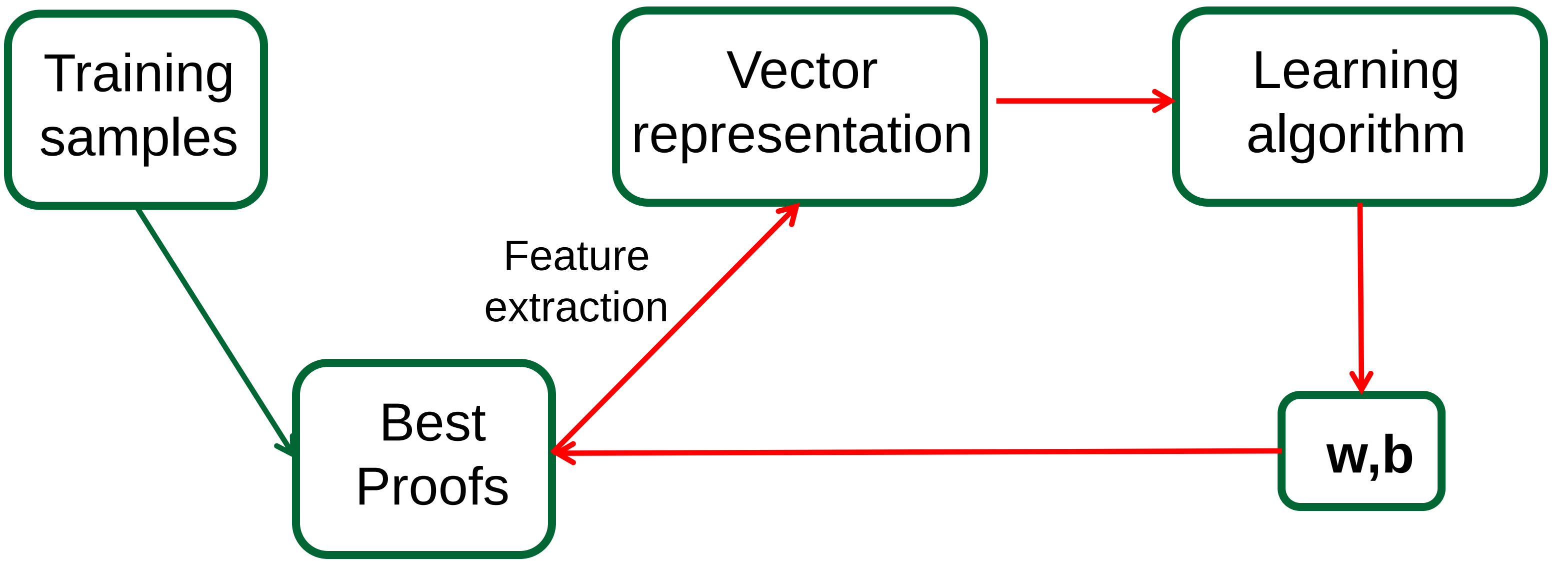
Learning

- Goal: Learn parameters **(w, b)**
- Use a linear learning algorithm
  - logistic regression

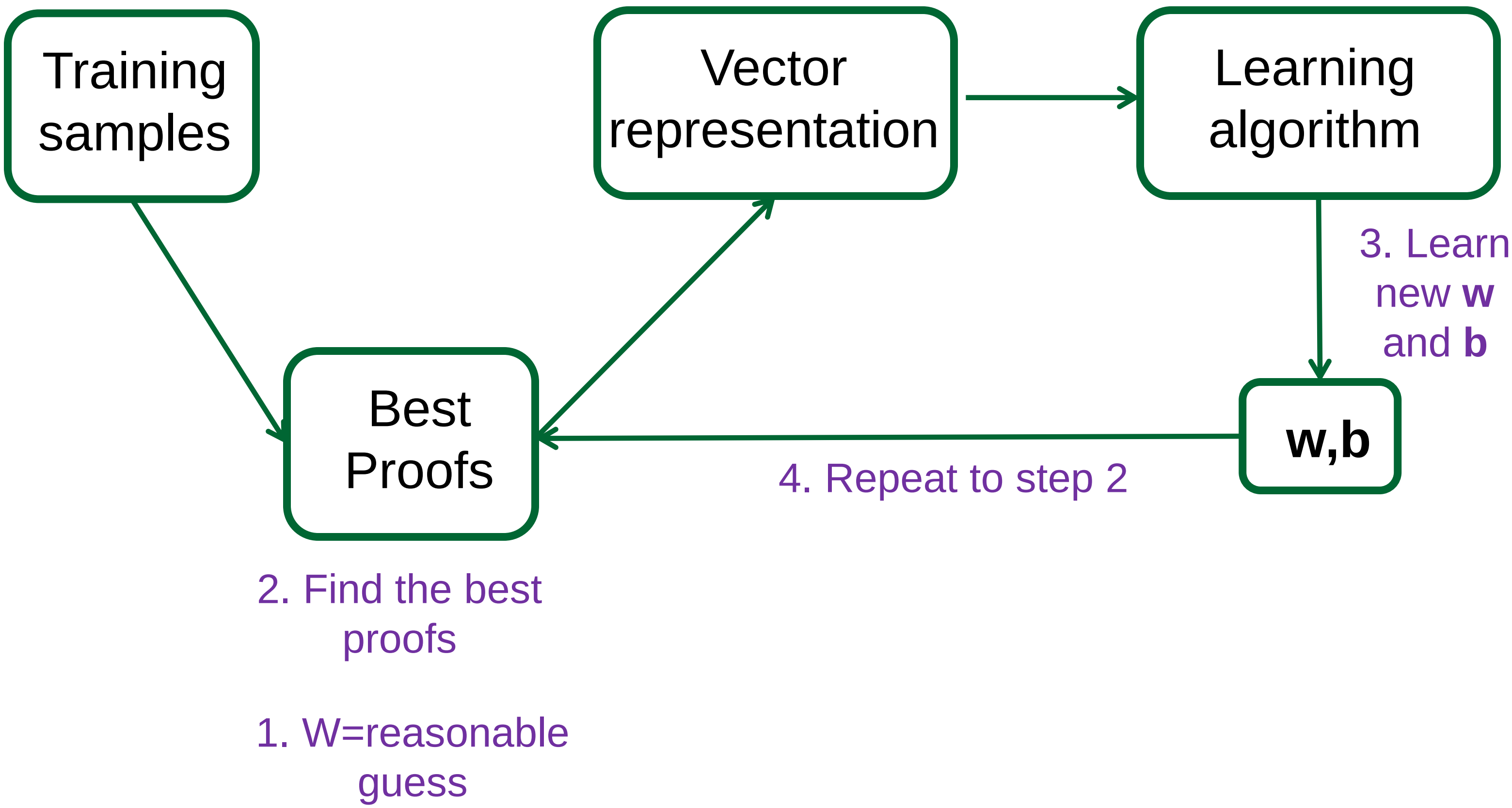
Inference vs. Learning



Inference vs. Learning



Iterative Learning Scheme



Summary- System Components

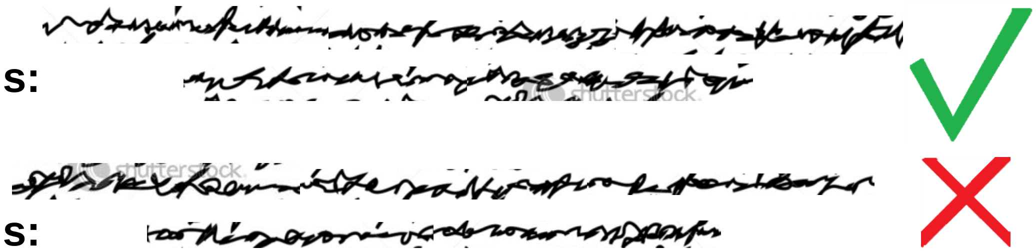
How to

1. Generate syntactically motivated complete proofs?
   - *Entailment rules*
   - *On the fly operations (Extended Tree Edit Operations)*
2. Estimate proof validity?
   - *Confidence Model*
3. Find the best proof?
   - *Novel search Algorithm*
4. Learn the model parameters?
   - *Iterative Learning Scheme*

Results RTE 1-5

Text:
Hypothesis:

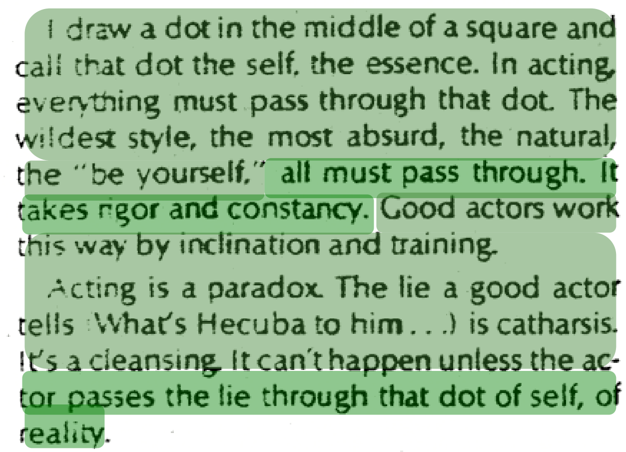
Text:
Hypothesis:



Evaluation by accuracy – comparison with transformation-based systems

System	RTE-1	RTE-2	RTE-3	RTE-5
Raina et al. 2005	57.0			
Harmeling, 2009		56.39	57.88	
Wang and Manning, 2010		**63.0**	61.10	
Bar-Haim et al., 2007			61.12	**63.80**
Mehdad and Magnini, 2009	**58.62**	59.87	62.4	60.2
Our System	57.13	61.63	**67.13**	63.50

Results RTE 6



I draw a dot in the middle of a square and call that dot the self, the essence. In acting, everything must pass through that dot. The wildest style, the most absurd, the natural, the "be yourself," all must pass through. It takes rigor and constancy. Good actors work this way by inclination and training.

Acting is a paradox. The lie a good actor tells (What's Hecuba to him . . .) is catharsis. It's a cleansing. It can't happen unless the actor passes the lie through that dot of self, of reality.

Natural distribution of entailments

Evaluation by Recall / Precision / F1

RTE 6 (F1%)	
Base line (Use IR top-5 relevance)	34.63
Median (2010)	36.14
Best (2010)	48.01
Our system	**49.54**

Conclusions –
The BIUTEE Inference Engine

- Inference as proof over parse trees
  - Natural to incorporate many inference types
- Results - close to best or best on RTEs

- **Open Source**
  - Configurable
  - Extensible
  - Visual tracing
  - Support

Adding extra-linguistic inferences

- Some tasks may benefit from extra-linguistic "expert" inferences
  - Temporal / arithmetic / spatial reasoning / …
    - *2 soldiers and a civilian => 3 people*

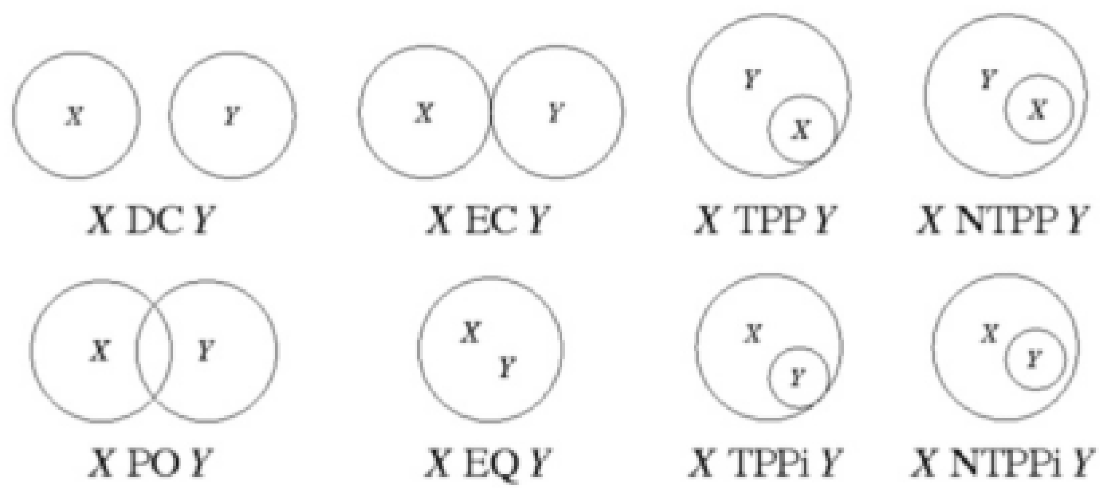
- Need to integrate with primary inference over language structures
  - "Expert" may detect on the fly inferences that would bridge text and hypothesis,
  - Interleaved within tree-generation process

Slide from Inderjeet Mani

RCC-8 Mereotopology

1. $\underline{\mathrm{DC}}(x, y) \stackrel{\mathrm{def}}{=} \backsim \mathrm{Connect}(x, y)$.
2. $\mathrm{Part}(x, y) \stackrel{\mathrm{def}}{=} \forall z\ \mathrm{Connect}(z, x) \rightarrow \mathrm{Connect}(z, y)$.
3. $\underline{\mathrm{EQ}}(x, y) \stackrel{\mathrm{def}}{=} \mathrm{Part}(x, y) \wedge \mathrm{Part}(y, x)$.
4. $\mathrm{Overlap}(x, y) \stackrel{\mathrm{def}}{=} \exists z\ \mathrm{Part}(z, x) \wedge \mathrm{Part}(z, y)$.
5. $\underline{\mathrm{EC}}(x, y) \stackrel{\mathrm{def}}{=} \mathrm{Connect}(x, y) \wedge \backsim \mathrm{Overlap}(x, y)$.
6. $\mathrm{PO}(x, y) \stackrel{\mathrm{def}}{=} \mathrm{Overlap}(x, y) \wedge \backsim \mathrm{Part}(x, y) \wedge \backsim \mathrm{Part}(y, x)$.
7. $\mathrm{PP}(x, y) \stackrel{\mathrm{def}}{=} \mathrm{Part}(x, y) \wedge \text{not } \mathrm{Part}(y, x)$.
8. $\underline{\mathrm{TPP}}(x, y) \stackrel{\mathrm{def}}{=} \mathrm{PP}(x, y) \wedge \exists z[\mathrm{EC}(z, x) \wedge \mathrm{EC}(z, y)]$
9. $\underline{\mathrm{NTPP}}(x, y) \stackrel{\mathrm{def}}{=} \mathrm{PP}(x, y) \wedge \backsim \exists z[\mathrm{EC}(z, x) \wedge \mathrm{EC}(z, y)]$.

Disconnected (DC): A and B do not touch each other.
Externally Connected (EC): A and B touch each other at their boundaries.
Partial Overlap (PO): A and B overlap each other in Euclidean space.
Equal (EQ): A and B occupy the exact same Euclidean space.
Tangential Proper Part (TPP): A is inside B and touches the boundary of B.
Non-tangential Proper Part (NTPP): A is inside B and does not touch the boundary of B.

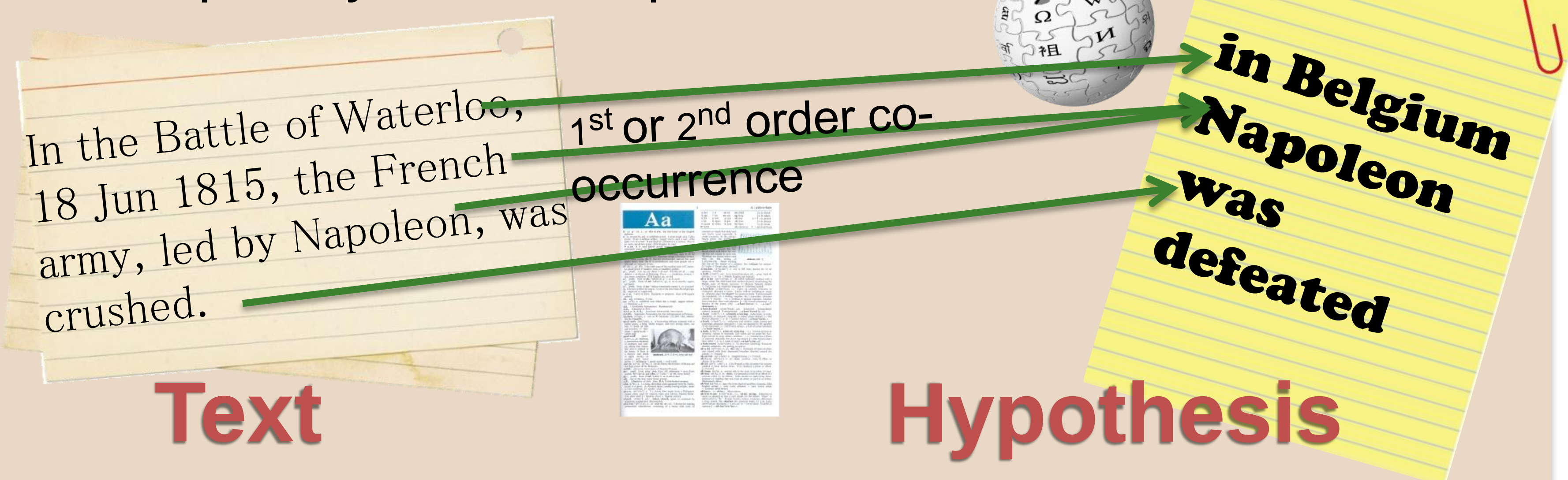


Related Research Lines

- RTE via tree edits
  - Learning edit costs, compute Tree-Edit-Distance
    - Mehdad & Magnini 2009; Heilman & Smith 2010; Wang & Manning 2010
- Text-to-text generation
  - Cf. ACL-2011 workshop, Smith's invited talk
- Paraphrasing – recognition, generation
- Richer discourse-level inferences
  - Mirkin at al. 2010 (merging, bridging)
  - Implicit argument detection
    - Gerber 2010, SemEval task 2010, Ruppenhofer et al. 2011
- Recovering implicit relations
  - Nakov & Kozareva 2011
- Natural logic
  - MacCartney & Manning 2009

Lexical Textual Inference [Eyal Shnarch]

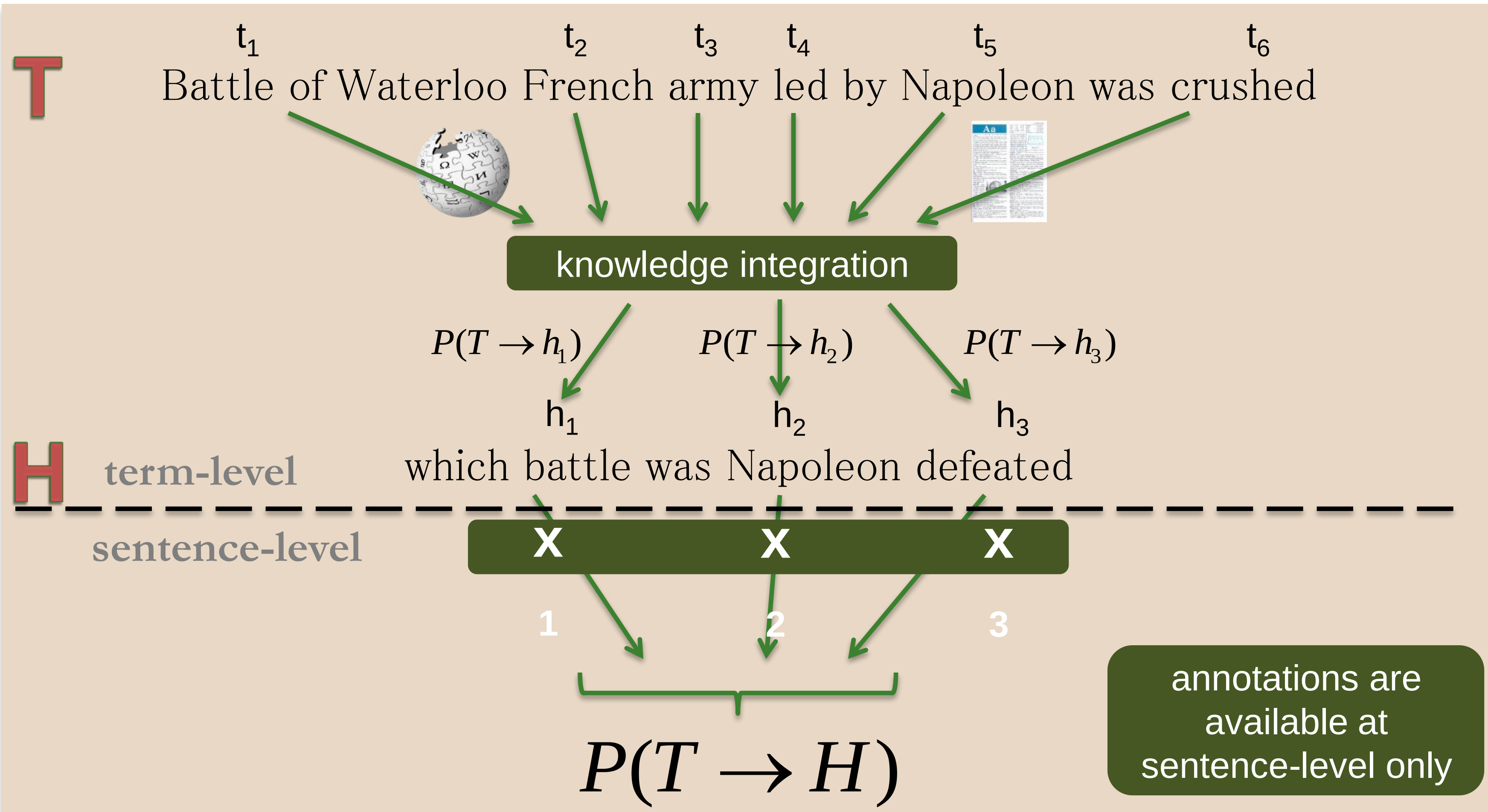
- Complex systems use parser



- ***Lexical inference rules*** link terms from *T* to *H*
- Lexical rules come from ***lexical resources***
- *H* is inferred from *T* iff all its terms are inferred

lexical textual inference ▷ principled probabilistic model ▷ Improves state-of-the-art

Probabilistic model – forward inference



lexical textual inference ▷ principled probabilistic model ▷ Improves state-of-the-art

Backward witnesses model

t_1 t_2 t_3 t_4 t_5 t_6

Battle of Waterloo French army led by Napoleon was crushed

$$\theta_w = P(w(x_i) = 1 | x_i = 1)$$

$$\tau_w = P(w(x_i) = 1 | x_i = 0)$$

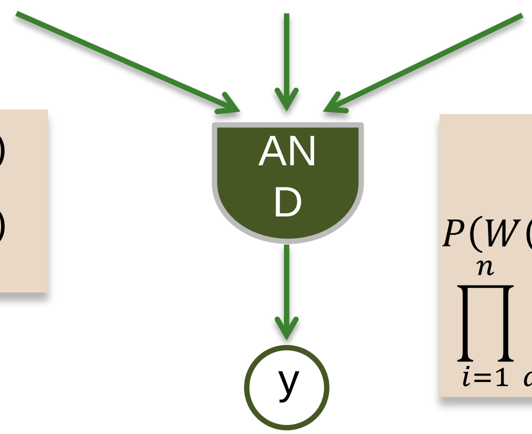
$$P(W(h_i)|x_i = 1) = \prod_{w \in W(h_i)} \theta_w \cdot \prod_{w \notin W(h_i)} (1 - \theta_w)$$

$$P(W(h_i)|x_i = 0) = \prod_{w \in W(h_i)} \tau_w \cdot \prod_{w \notin W(h_i)} (1 - \tau_w)$$

h_1 h_2 h_3

which battle was Napoleon defeated

x_1 x_2 x_3



$$\eta_0 \stackrel{\text{def}}{=} P(x_i = 1 | y = 0)$$

$$\eta_1 \stackrel{\text{def}}{=} P(x_i = 1 | y = 1)$$

Likelihood

$$P(W(H)|y = e) = \prod_{i=1}^{n} \sum_{a=0,1} P(W(h_i)|x_i = a) \cdot P(x_i = a | y = e)$$

lexical textual inference ▷ principled probabilistic model ▷ Improves state-of-the-art

Acquiring Inference Knowledge

- *over language structures*

Knowledge acquisition sources

- Learning from corpora
- Mining human-oriented knowledge resources
  - Wikipedia, dictionary definitions
- Computational NLP resources
  - WN, FrameNet, NOMLEX, …
- Manual knowledge engineering
  - Recent Mechanical Turk potential

Distributional similarity (Symmetric)

- Most similar words for ***food*** (Lin, 1998)
  - Symmetric measure often identifies “sister” terms

meat	clothing	water	sugar
beverage	foodstuff	coffee	material
goods	textile	meal	chemical
medicine	fruit	tobacco	equipment
drink	feed	fuel	rice

Directional similarity – Feature Inclusion

- Kotlerman et al. (2009)

Most directionally-similar words for ***food*** :

foodstuff	ration	blanket	margarine
food product	drinking water	soup	dessert
food company	wheat flour	biscuit	cookie
noodle	grocery	sweetener	sauce
canned food	beverage	meat	ingredient
feed	snack	agribusiness	meal
salad dressing	dairy product	diet	vegetable
bread	hamburger	medicine	vegetable oil
food aid	chocolate	food supply	herb
drink	seafood	fruit juice	milk

Extraction from Wikipedia

E.T. the Extra-Terrestrial

From Wikipedia, the free encyclopedia
(Redirected from E.T. (film))

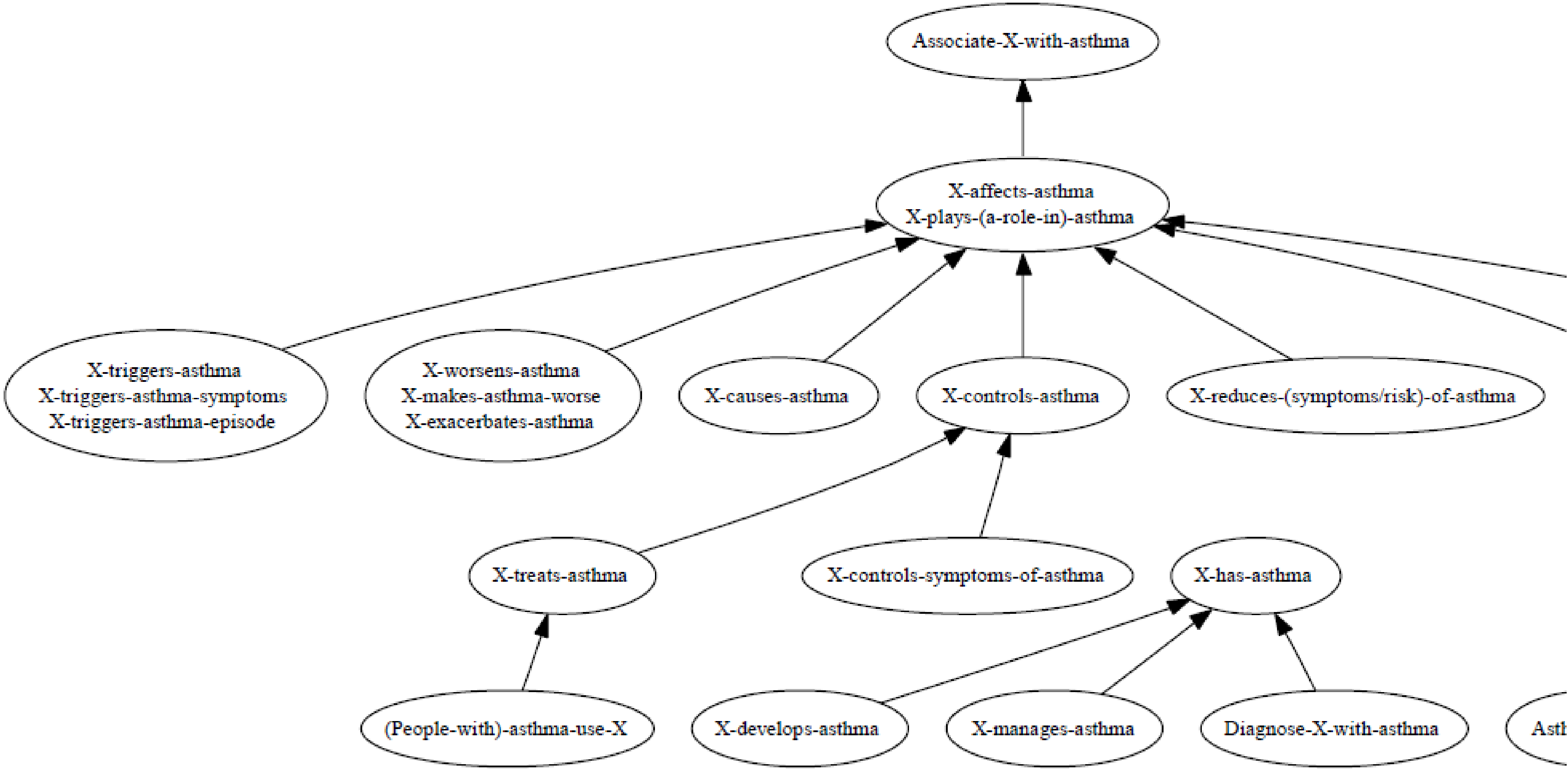
E.T. the Extra-Terrestrial is a 1982 science fiction film co-produced and directed by Steven Spielberg, written by Melissa Mathison and starring Henry Thomas, Robert MacNaughton, Drew Barrymore, Dee Wallace and Peter Coyote. It tells the story of Elliott (played by Thomas), a lonely boy who befriends a friendly alien, dubbed "E.T.", who is stranded on Earth. Elliott and hi

Extraterrestrial life

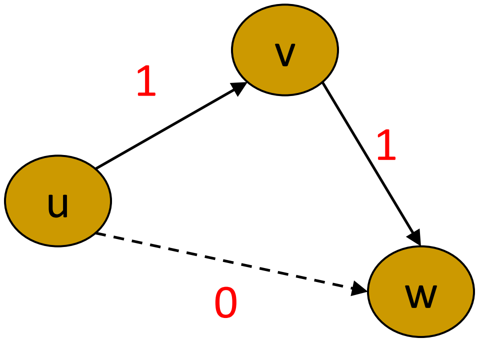
- *Be-complement*
- *Redirect*
- *Parenthesis*
- *Link*

Learning Entailment-rule Graphs

- Berant et al. series of works – increasing scalability: ACL-2010, ACL-2011 (best paper by student), ACL-2012
- Example target hierarchy (medical domain):



Global Optimization of Graph Edges



$$\hat{G} = \arg\max \sum_{u \neq v} f(u,v) \cdot I_{uv}$$

$$\forall u,v,w \in V.\, I_{uv} + I_{vw} - I_{uw} \leq 1$$

$$\forall (u,v) \in NEG.\, I_{uv} = 0$$

$$\forall (u,v) \in POS.\, I_{uv} = 1$$

$$I_{uv} \in \{0,1\}$$

- Integer Linear Program
  - Optimize global edge scores under transitivity and other constraints

Syntactic-driven Entailments

- Active-passive transformations
- Recover relative clause arguments
- Extract conjuncts
- Appositions
- …

- *TruthTeller*: annotate truth for predicates and clauses
  - Positive: John called(+) Mary.
  - Negative: John forgot to call(−) Mary.
  - Unknown: John wanted to call(?) Mary.

- Constructed via human linguistic engineering
  - May be combined with automatic learning

Mechanical Turk & Community Knowledge-engineering

- Validating automatically-learned rules
  - Generating paraphrases/entailments
  - Zeichner et al., ACL-2012
- Potential for community contribution
  - Stipulating domain knowledge in NL

Inference & ambiguity

H *The US accepts a large number of foreigners every year*

welcome ⇒ accept

alien ⇔ foreigner

T *The US welcomes hundreds of thousands of aliens yearly*

*If it's any consolation, dear, our **alien** abduction insurance is finally going to pay off*

Context matching

- ***Context Matching*** generalizes sense matching
  - Does *aliens* in T match the meaning of *outer-space* ?
  - Does 'children acquire English' match
    X acquire Y $\Rightarrow$ *X learn Y*?

- *Contextual Preferences*
  - A generic *context validation* framework for entailment rules
    - Szpektor & Dagan, ACL-08
    - Classification-based approach (Mirkin et al., TextInfer 2011)
  - Match hypothesis, rules and text only in suitable contexts
  - An alternative to explicit WSD

BIUTEE Demo

EXCITEMENT: towards *Textual-inference Platform*

- Open source & community

A Textual Inference Platform

- Starting with BIUTEE, moving to **EXCITEMENT**
  - Goal: build MOSES-like environment
    - Incorporate partners' inference systems
  - Addressing two types of research communities:
    - Applications which can benefit from textual inference
    - Technologies which can improve inference technology

- Partners:
  - Academic: FBK, Heidelberg, DFKI, Bar-Ilan
  - Industriacl: NICE (Israel), AlmaWave (Italy), OMQ (Germany)

Future: Extended Operation modes

- **Recognition:** recognize entailment given *T*/*H* pair
  - *Validation in applications*
- **Search:** given *H* and corpus/doc, find all entailing texts
  - *Multi-document summarization (RTE-5 pilot & RTE-6)*
  - *QA, IR, IE against corpus/doc*
  - *Use entailment knowledge to generate expanded queries*
- **Generation:** given text, generate all entailed statements
  - *Paraphrase generation for MT*
  - *Unsupervised IE – generate "canonical" propositions*
- **Functionality extensions**
  - Include **variables** in hypothesis (perform extraction - IE, QA, …)
  - **Partial entailment:** identify entailments of parts of *h*
  - *…*

Takeout

- Time to develop *textual inference*
  - Generic, applied, principled

- Proposal:
  - Base core inference on language-based representations
    - Parse trees, co-references, lexical contexts, …
  - Extra-linguistic/logical inference for specific suitable cases

- Breakthrough potential – current and future applications

- It’s a long-term endeavor, but it’s here!

http://www.cs.biu.ac.il/~nlp/downloads/biutee

Thank You!