Textual Entailment
Part 2: Classes of Strategies and Learning

Sebastian Pado
Institut für Computerlinguistik
Universität Heidelberg, Germany

Rui Wang
Language Technology
DFKI, Saarbrücken, Germany

Tutorial at AAAI 2013, Bellevue, WA

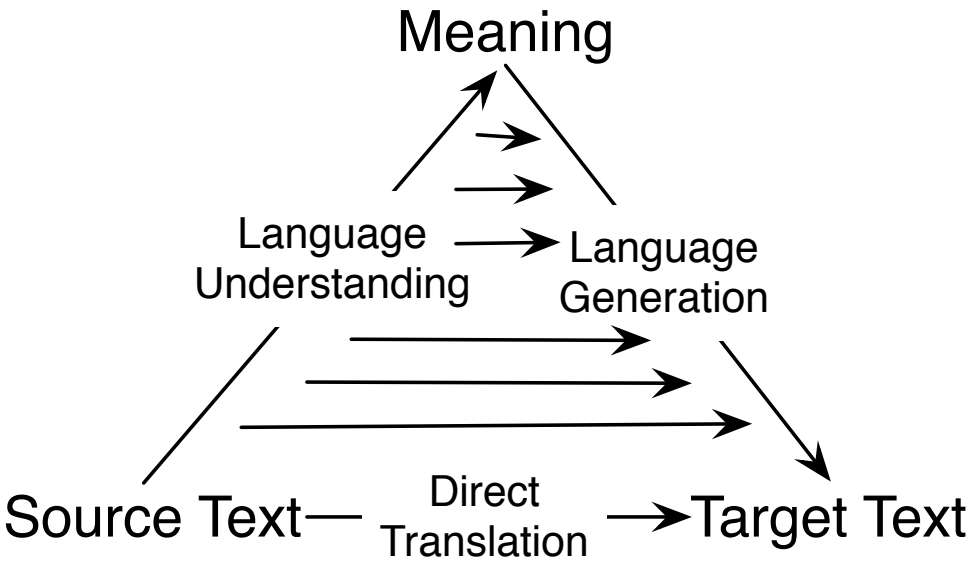
Thanks to Ido Dagan for permission to use slide material

Structure of the Tutorial

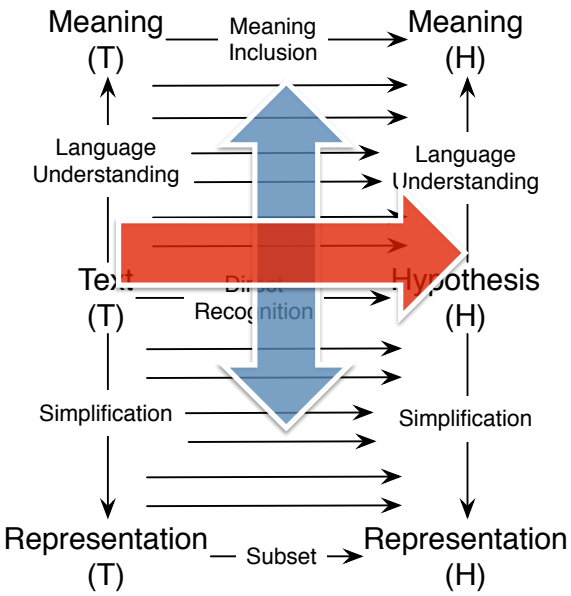
- Part 1 [SP]: Introduction and Basics
- Part 2 [RW]: Classes of Strategies and Learning

  * BREAK*
- Part 3 [SP]: Knowledge and Knowledge Acquisition
- Part 4 [SP]: Applications
- Part 5 [RW]: Multilingual, Component-based System Building

MT Triangle



RTE Rectangle



Architecture

- Linguistic analysis pipeline (LAP)
- Entailment decision algorithm (EDA)
  - Classification-based
  - Transformation-based
- Knowledge base (KB) (next section)

Architecture

- Linguistic analysis pipeline (LAP)
- Entailment decision algorithm (EDA)
  - Classification-based
  - Transformation-based
- Knowledge base (KB) (next section)

Overview of LAPs

- Tokenization (Word Segmentation)
- Part-of-Speech (POS) Tagging
- Lemmatization
- Named-Entity Recognition
- Syntactic Parsing
  - Constituent Parsing
  - Dependency Parsing
- Semantic Role Labeling
- Coreference Resolution
- …

Token-Level Processing

- Tokenization
  - Word segmentation
- Lemmatization
  - Morphological analysis
- POS Tagging
- Lexical Semantics
  - WordNet, distributional similarity, etc.

Performance >97%

An Example

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

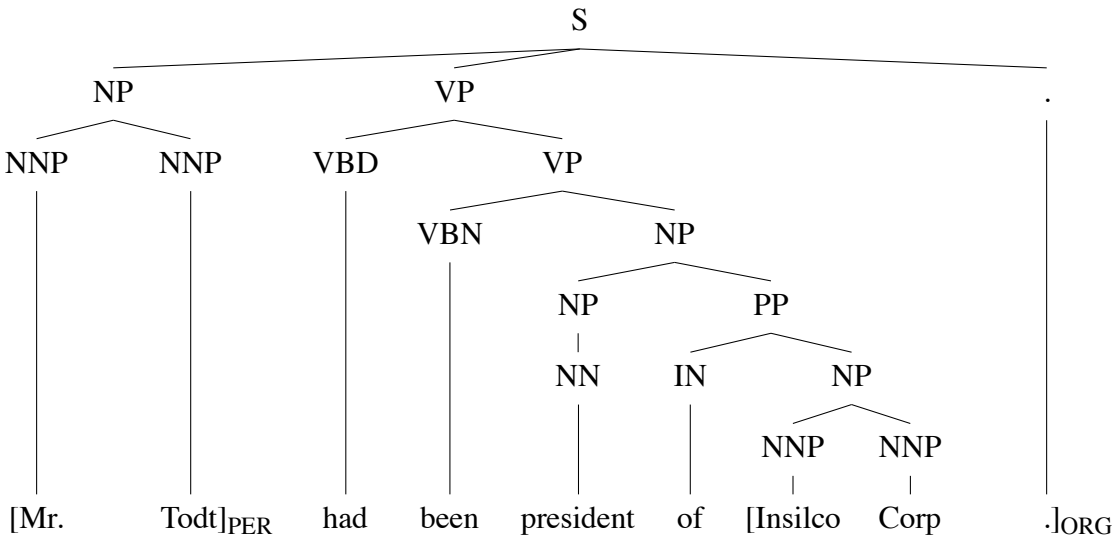
From TreeTagger website

Constituents

- Chunking
- Named-Entity Recognition
- Constituent Parsing

NER:
70~90%

An Example



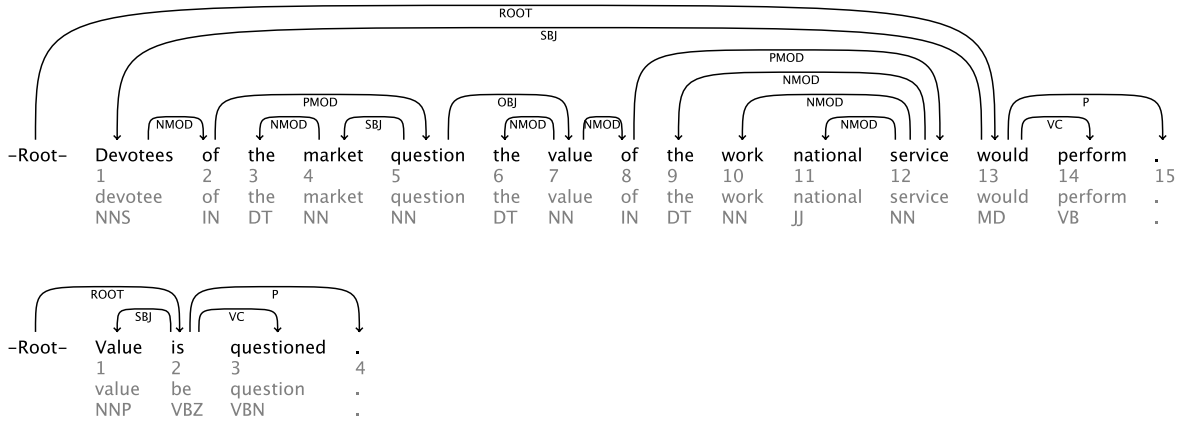
From Stanford NER (Finkel and Manning, 2009)

Dependency

- Syntactic Dependency Parsing
- Semantic Dependency Parsing
  - Semantic Role Labeling
  - Predicate-Argument Structure
- Logic Form Composition

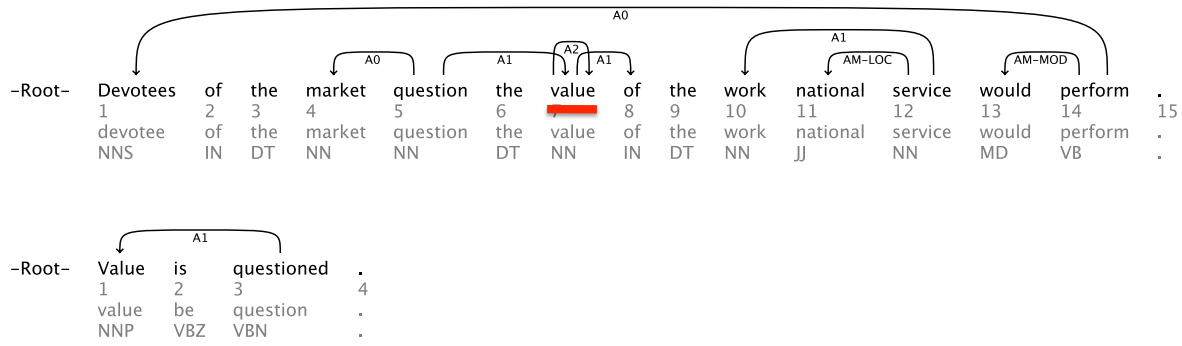
Syn: 80~90%
Sem: 75~85%

An Example



From MSTParser (McDonald et al., 2005); Visualized by https://code.google.com/p/whatswrong/

An Example (cont.)

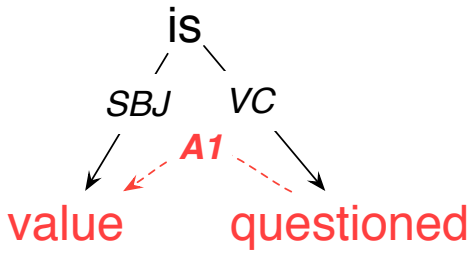


From Laputa SRL (Zhang et al., 2008); Visualized by https://code.google.com/p/whatswrong/

An Example (cont.)

- **H**: *Value is questioned.*

- Syntactic dependency
  - <is, SBJ, value>
  - <is, VC, questioned>



- Semantic dependency
  - <questioned, A1, value>

Semantic Roles

- PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004)

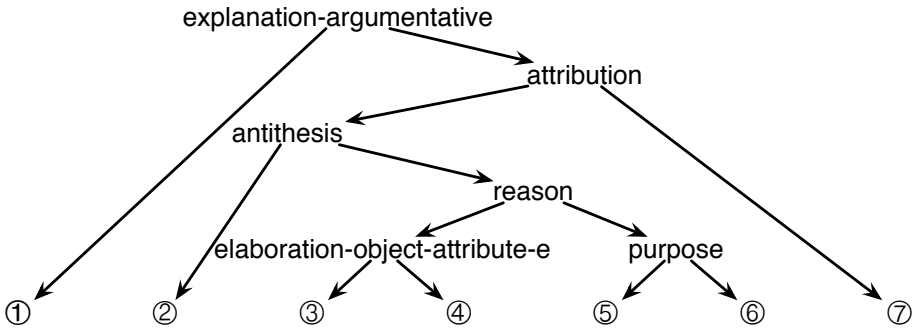
- Core arguments: A0-A5
  - different semantics for each verb
  - specified in the PropBank Frame files

- 13 types of adjuncts labeled as AM-*arg*
  - where *arg* specifies the adjunct type

Discourse

- Coreference Resolution
- Event Structure
- Discourse Parsing

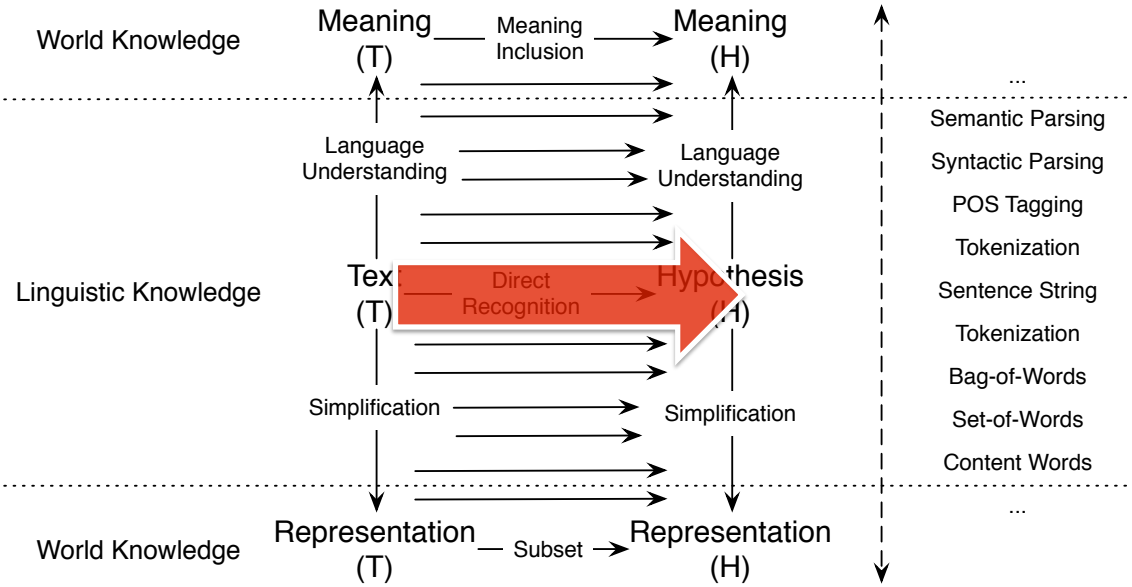
An Example



[① Ford Motor Co. and Chrysler Corp. representatives criticized Mr. Tonkin's plan as unworkable.] [② It "is going to sound neat to the dealer] [③ except when his 15-day car supply doesn't include the bright red one] [④ that the lady wants to buy] [⑤ and she goes up the street] [⑥ to buy one,"] [⑦ a Chrysler spokesman said.]

From RST Discourse Treebank (Carlson et al., 2002)

RTE Rectangle (more details)



Overview of EDAs

- Classification-based
  - Score / Threshold
  - Structure / Alignment
- Transformation-based
  - Edit distance
  - (Knowledge) rule application
- Meta-EDA

Classification (RTE Style)

Popular Classifiers

Model	Perceptron/SVM	Naïve Bayes	Logistic Regression
Type	Discriminative	Generative	Discriminative
Distribution	N/A	P(X, Y)	P(Y|X)
Independence	None	Strong	None
Features	Ex/Impilicit	Explicit	Explicit
Speed	Fast/Slow	Fast	Intermediate

Kernel-Based Methods

- Kernel Function
  - Mapping between spaces
  - Cross-combination of features (implicitly!)
  - Intro-pair features → cross-pair features

- Subsequence Kernel (Lodhi et al., 2002; Wang and Neumann, 2007a)
- Tree Kernel (Collins and Duffy, 2001; Zanzotto et al., 2007)

Linguistic Features

- Measure ~~something~~similarity between *t* and *h*:
  - Lexical overlap (unigram, N-gram, subsequence)
    - Assisted by lexical resources like WordNet
  - Syntactic matching
  - Lexical-syntactic variations ("paraphrases")
  - Semantic role matching
  - Global similarity parameters (e.g. negation, modality)

- Detect mismatch (for non-entailment)

Data Structures

- String-to-String rewriting
  - String edit distance (MacCartney and Manning, 2007)
  - Tree skeleton difference (Wang and Neumann, 2007a)
- Tree-to-Tree editing
  - Tree edit distance (Kouylekov and Magnini, 2005)
- Graph-to-Graph mapping
  - Graph matching (Haghighi et al., 2005)

Word Overlap

- |T|: number of words in T
- |H|: number of words in H
- $E_1 = |T \wedge H| / |H|$
- $E_2 = |T \wedge H| / |T|$
- $E_3 = (2 * E_1 * E_2) / E_1 + E_2$

57.2 on average

- Content words only
- Lemmatization

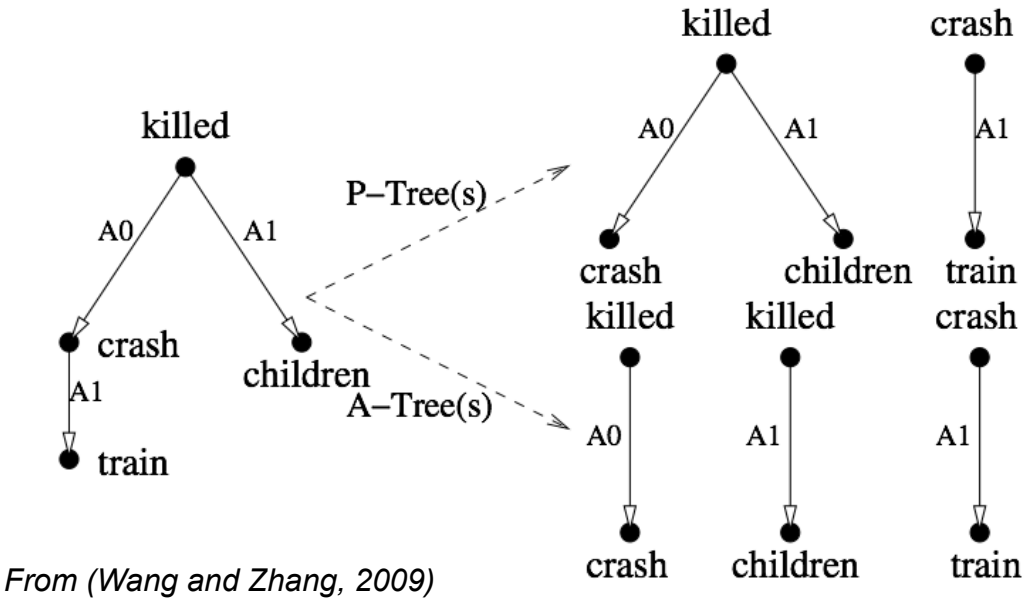
From (Mehdad and Magnini, 2009)

Dependencies

- Syntactic dependency trees
  - Dependency triples *<Node, Relation, Head>*
  - Bag of such triples

- E_1' = |Triple(T) ∧ Triple (H)| / |Triple(H)|

Dependencies (cont.)



From (Wang and Zhang, 2009)

Results (RTE-5)

- DFKI1: BoW and syntactic dependency
- DFKI2: BoW, syntactic, and semantic dependency
- DFKI3: BoW and joint syntactic and semantic representation

Runs	Main	Main -VO	Main -WN	Main -VO-WN
DFKI1	62.5%	62.5%	62.7%	62.5%
DFKI2	66.8%	66.5%	66.7%	66.3%
DFKI3	**68.5%**	68.3%	68.3%	68.3%

From (Wang et al., 2009)

Larger Sub-Structures

- Dependency paths
  - Common sub-paths
- Subtrees
- E_1'' = |Subtree(T) ∧ Subtree(H)| / |Subtree(H)|

Subtrees

$T_1 \Rightarrow H_1$	$T_2 \Rightarrow H_2$
T_1 *"Farmers feed cows animal extracts"*	T_2 *"They feed dolphins fish"*
H_1 *"Cows eat animal extracts"*	H_2 *"Fish eat dolphins"*

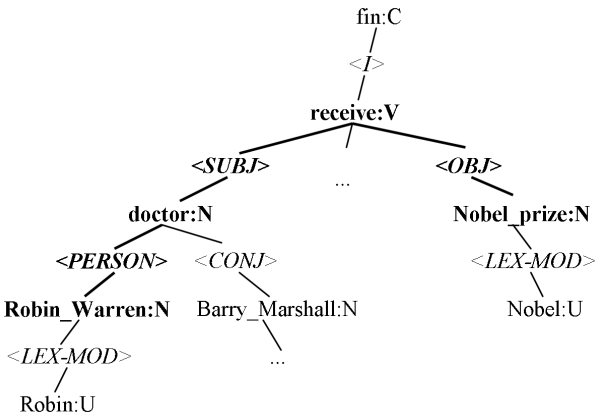
feed X Y → X eat Y

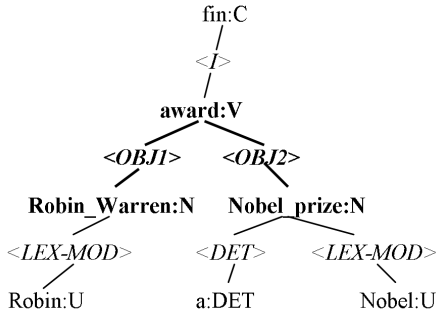
$T_3 \Rightarrow H_3$
T_3 *"Mothers feed babies milk"*
H_3 *"Babies eat milk"*

From (Zanzotto and Dell'Arciprete, 2009)

Tree Skeletons

- T: *Doctor Robin Warren and Barry Marshall received Nobel Prize …*
- H: *Robin Warren was awarded a Nobel Prize.*





From (Wang and Neumann, 2007)

Results

- RTE-2

Bag-of-Words

Tree Skeleton

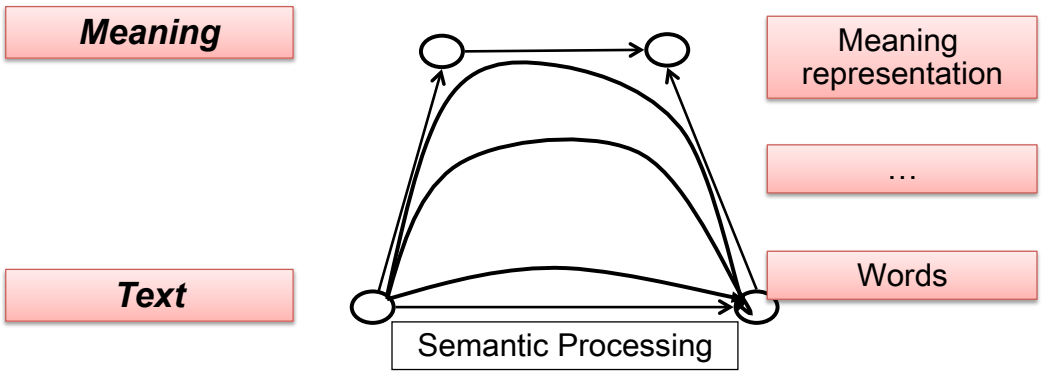
- RTE-3

Exp2BT&Exp2BL: Training on the RTE-3 Dev Set and Testing on the Test Set

Systems	IE	IR	QA	SUM	ALL
BoW	54.5%	66.5%	76.5%	56.0%	63.4%
TSM	54.5%	62.5%	66.0%	54.5%	59.4%
SK+BS (Mi+SP+*Task*) – run1	**59.5%**	70.5%	75.5%	60.5%	65.5%
SK+BS (Mi+*Length*) – run2	**58.5%**	70.5%	79.5%	59.0%	**66.9%**

From (Wang, 2007)

The RIGHT level



- Trade-offs between
  - *Competence* of the knowledge (deeper)
  - *Performance* of the processing (shallower)

Alignment-Based Approaches

- Word alignment (Glickman et al., 2006)
- Phrase alignment (chambers et al., 2007; MacCartney et al., 2008)
- Relation alignment (Sammons et al., 2009)

Overview of EDAs

- ~~Classification-based~~
  - ~~Score / Threshold~~
  - ~~Structure / Alignment~~
- Transformation-based
  - Edit distance
  - (Knowledge) rule application
- Meta-EDA

Matching vs. Transformations

- Direct matching (so far, no chaining)

- Sequence of transformations (A proof)

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

  – Tree-Edits

  – Knowledge based Entailment Rules

Edit Distance

- (Limited) pre-defined operators
  – Insertion
  – Deletion
  – Substition
- String-to-String
- Tree-to-Tree

Weakly linguistically motivated!

- The EDITS system (Kouylekov and Negri, 2010)
  – Estimate confidence in each operation
- Wang and Manning (2010), Heilman and Smith (2010), etc.

Knowledge-Based Rules

- Rule application
  - Arbitrary knowledge-based transformations
  - Formalize many types of knowledge

- BIUTEE (Stern and Dagan, 2011)
  - On-the-fly operations
  - Cost model
  - Search for the best inference

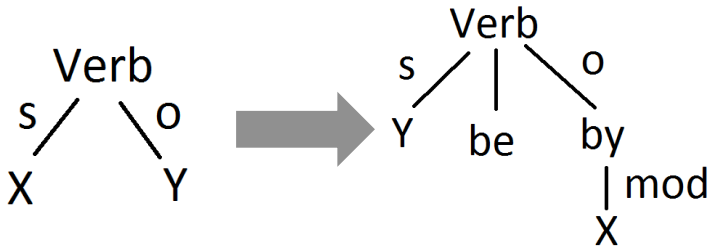
An Example

Id	Operation	Generated Text
0	-	He received the letter from the secretary.
1	Coreference substitution	The employee received the letter from the secretary.
2	X received Y from Z → Y was sent to X by Z	The letter was sent to the employee by the secretary.
3	Y [verb-passive] by X → X [verb-active] Y	The secretary sent the letter to the employee.
4	X send Y → X deliver Y	The secretary delivered the letter to the employee.
5	letter → message	The secretary delivered the message to the employee.

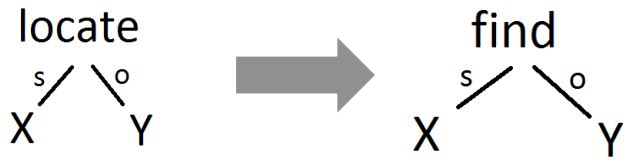
From (Stern et al., 2012)

Entailment Rules

Generic Syntactic



Lexical Syntactic



Lexical

From (Bar-Haim et al., 2007)

Cost Based Model

- Define **operation cost**
  - Represent each operation as a feature vector
  - Cost is linear combination of feature values

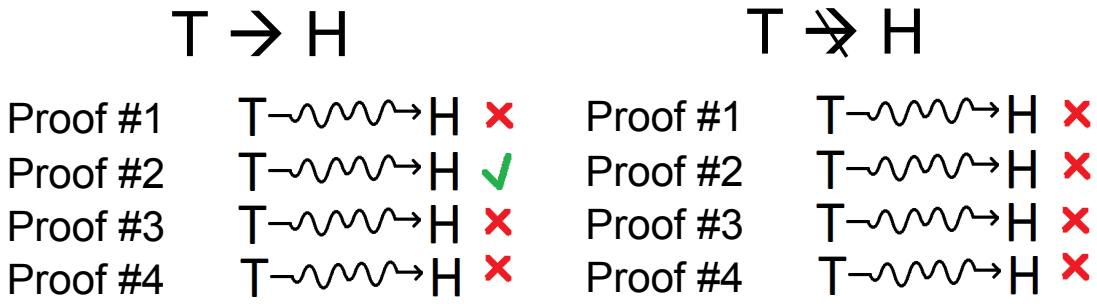
- Define **proof cost** as the sum of the operations' costs

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

Learn

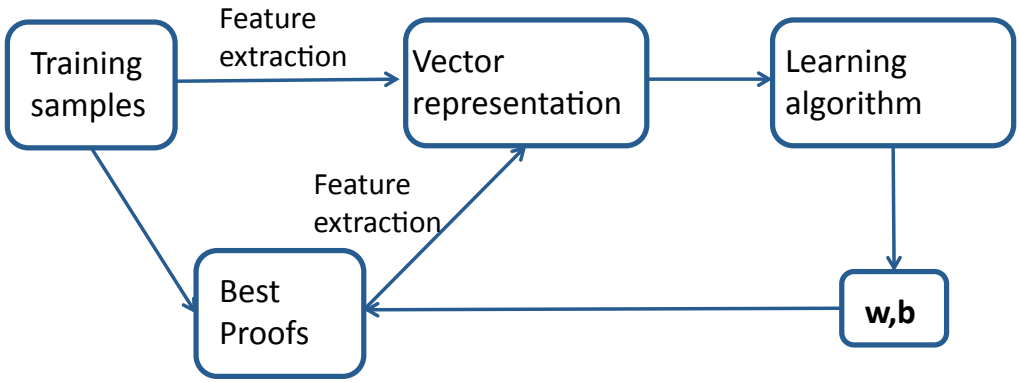
Variant of (Raina et al., 2005)

Search the Best Proof

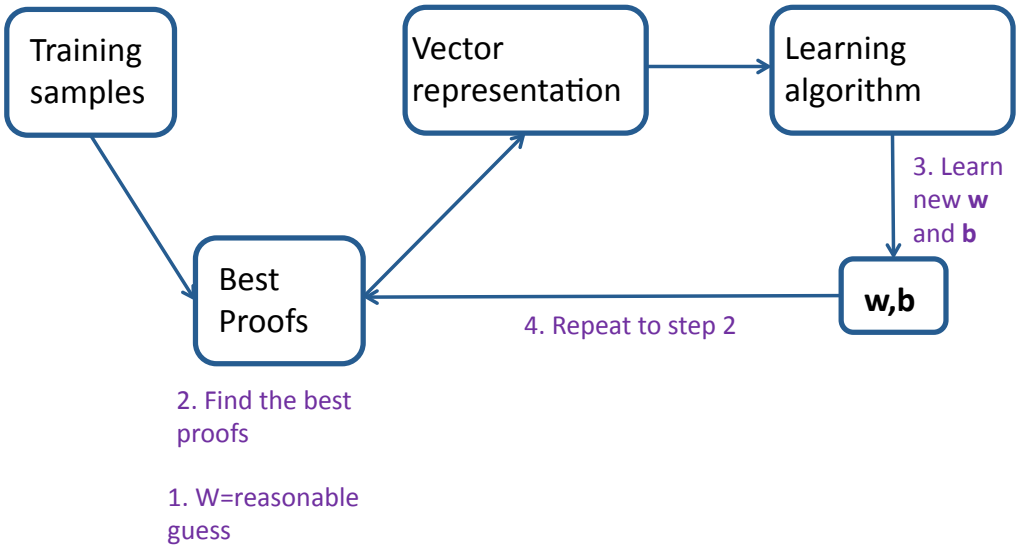


- "Best Proof" = proof with lowest cost
- Search space exponential – AI-style search (Stern et al., 2012)
  - Gradient-based evaluation function
  - Local look ahead for "complex" operations

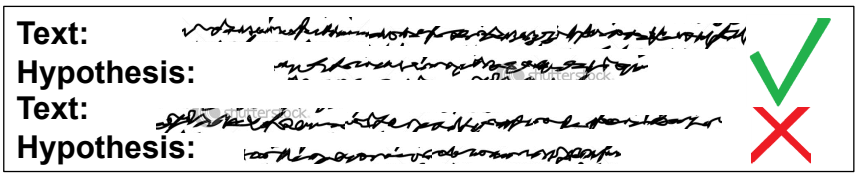
Inference vs. Learning



Iterative Learning Scheme

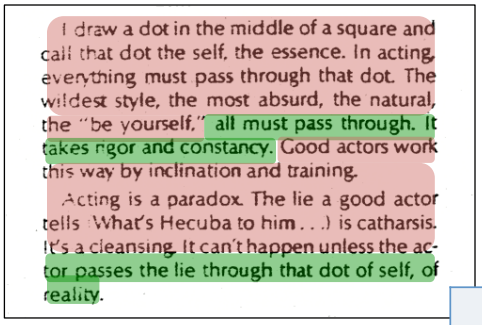


Performance (Classification)



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
BIUTEE (2011)	57.13	61.63	**67.13**	63.50

Performance (Search)


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.

RTE 6 (F1%)	
Base line (Use IR top-5 relevance)	34.63
Median (2010)	36.14
Best (2010)	48.01
BIUTEE (2012)	49.54

Overview of EDAs

- ~~Classification-based~~
  - ~~Score / Threshold~~
  - ~~Structure / Alignment~~
- ~~Transformation-based~~
  - ~~Edit distance~~
  - ~~(Knowledge) rule application~~
- Meta-EDA

An Example

- **T**: *Bush used his weekly radio address to try to build support for his plan to allow workers to divert part of their Social Security payroll taxes into private investment accounts.*

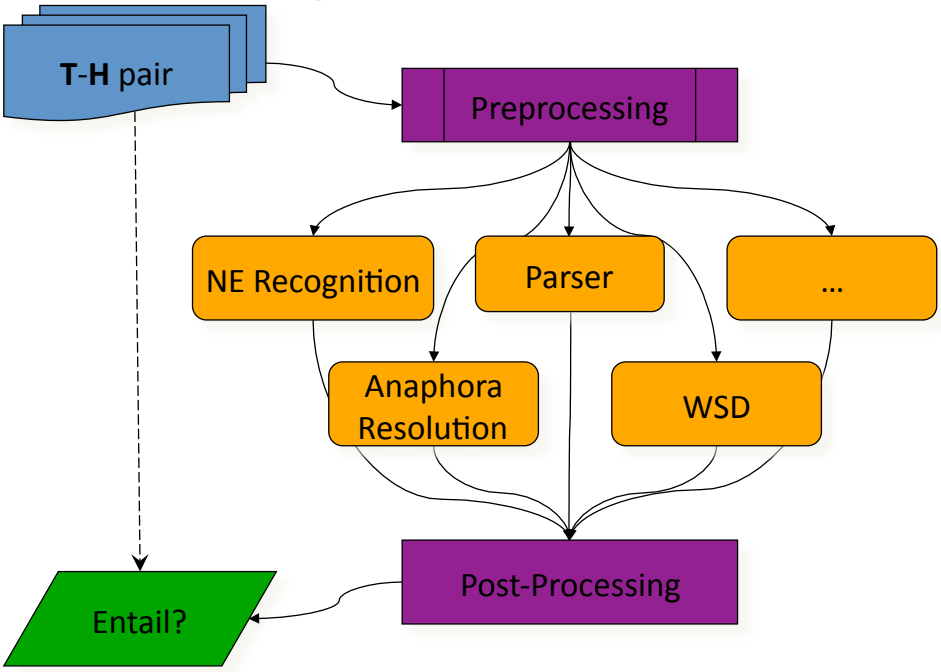
- **H**: *Mr. Bush is proposing that workers be allowed to divert their payroll taxes into private accounts.*

An Example

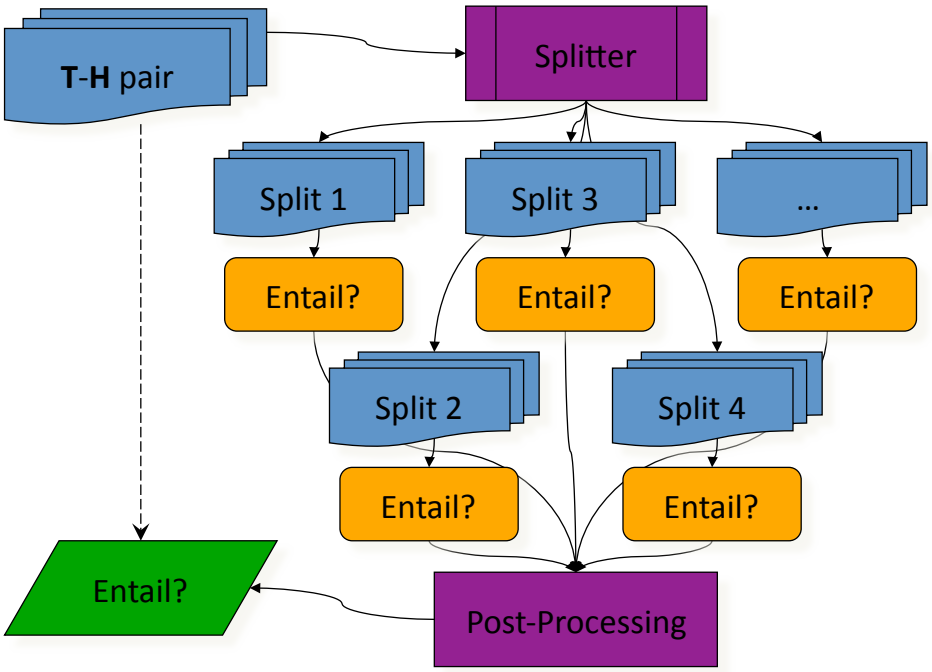
- **T**: *Bush used his weekly radio address to try to build support for his plan to allow workers to divert part of their Social Security payroll taxes into private investment accounts.*

- **H**: *Mr. Bush is proposing that workers be allowed to divert their payroll taxes into private accounts.*

Bag-of-Features

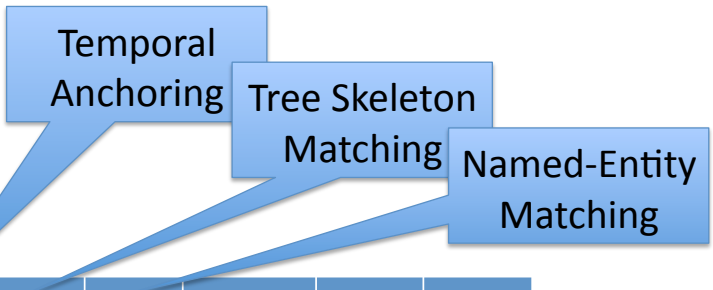


Specialized Modules



Divide-and-Conquer

- A specialized RTE module
  - A good target
  - A good tackle
- Results on RTE-4



Modules	TAC-M	TS-M	NE-M	BoW-BM	Tri-BM	Overall
Accuracy	**80.6%**	74.6%	**54.3%**	56.5%	52.8%	70.6%
Coverage	3.1%	34.6%	47.7%	100%	100%	100%

From (Wang and Neumann, 2009)

Summary

- Linguistic analysis pipeline
  - Various linguistic processing

imported

- Entailment decision algorithm
  - Classification & feature space
  - Transformation & knowledge bases (upcoming)

implemented

- Overall Strategy
  - Specialized modules

soon…

Reference List

- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.
- Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing.
- Jenny Rose Finkel and Christopher D. Manning. 2009. Joint Parsing and Named Entity Recognition. In Proceedings of NAACL.
- Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In Proceedings of HLT-EMNLP.
- Yi Zhang, Rui Wang, and Hans Uszkoreit. 2008. Hybrid Learning of Dependency Structures from Heterogeneous Linguistic Resources. In Proceedings of CoNLL.

Reference List

- Martha Palmer, Dan Gildea, Paul Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics Journal.
- A, Meyers, R. Reeves, C. Macleod, R, Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An Interim Report, Proc. of HLT-EACL Workshop: Frontiers in Corpus Annotation.
- Carlson, L., Okurowski, M. E., and Marcu, D. 2002. RST discourse treebank. Linguistic Data Consortium, University of Pennsylvania.
- Y. Mehdad and B. Magnini. 2009. A word overlap baseline for the recognizing textual entailment task. Online.
- Rui Wang and Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structures. In Proceedings of EMNLP.
- Rui Wang, Yi Zhang, and Günter Neumann. 2009. A joint syntactic-semantic representation for recognizing textual relatedness. In Text Analysis Conference TAC 2009 WORKSHOP Notebook Papers and Results.

Reference List

- Zanzotto, F. M. and Dell'Arciprete, L. 2009. Efficient kernels for sentence pair classification. In Proceedings of EMNLP.
- Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text Classification using String Kernels. Journal of Machine Learning Research.
- Rui Wang and Günter Neumann. 2007. Recognizing Textual Entailment Using a Subsequence Kernel Method. In Proceedings of AAAI.
- Michael Collins and Nigel Duffy. 2001. Convolution Kernels for Natural Language. Advances in Neural Information Processing Systems.
- Zanzotto, F. M., Pennacchiotti, M., and Moschitti, A. 2007. Shallow Semantic in Fast Textual Entailment Rule Learners. In Proceedings of the ACL-PASCAL Workshop on textual entailment and paraphrasing.
- Rui Wang. 2007. Textual entailment recognition: A data-driven approach. Master's thesis, Saarland University.

Reference List

- Oren Glickman and Ido Dagan. 2006. A Lexical Alignment Model for Probabilistic Textual Entailment. In Lecture Notes in Computer Science.
- Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning Alignments and Leveraging Natural Logic. In Proceedings of the ACL Workshop on Textual Entailment and Paraphrase.
- Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In Proceedings of EMNLP.
- Mark Sammons, V.G.Vinod Vydiswaran, Tim Vieira, Nikhil Johri, Ming-Wei Chang, Dan Goldwasser, Vivek Srikumar, Gourab Kundu, Yuancheng Tu, Kevin Small, Joshua Rule, Quang Do, and Dan Roth. 2009. Relation alignment for textual entailment recognition. In Proceedings of TAC.

Reference List

- Milen Kouylekov and Matteo Negri. 2010. An open-source package for recognizing textual entailment. In Proceedings of the ACL 2010 System Demonstrations.
- Mengqiu Wang and Christopher Manning. 2010. Probabilistic Tree-Edit Models with Structured Latent Variables for Textual Entailment and Question Answering. In Proceedings of COLING.
- Michael Heilman and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In Proceedings of NAACL-HLT.
- Asher Stern and Ido Dagan. 2011. A Confidence Model for Syntactically-Motivated Entailment Proofs. Proceedings of RANLP.
- Asher Stern, Roni Stern, Ido Dagan, and Ariel Felner. 2012. Efficient Search for Transformation-based Inference. In Proceedings of ACL.

Reference List

- Roy Bar-Haim, Ido Dagan, Iddo Greental, Idan Szpektor, and Moshe Friedman. Semantic inference at the lexical-syntactic level for textual entailment recognition. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.
- Rajat Raina, Aria Haghighi, Christopher Cox, Jenny Finkel, Jeff Michels, Kristina Toutanova, Bill MacCartney, Marie-Catherine de Marneffe, Christopher Manning, and Andrew Ng. 2005. Robust Textual Inference using Diverse Knowledge Sources. In Proceedings of the PASCAL RTE Challenge.
- Rui Wang and Günter Neumann. 2009. An accuracy-oriented divide-and-conquer strategy for recognizing textual entailment. In Proceedings of TAC RTE Track.