Textual Entailment
Part 1: Introduction

Sebastian Pado
Institut für Computerlinguistik
Universität Heidelberg, Germany

Rui Wang
Language Technology
DFKI, Saarbrücken, Germany

AAAI 2013, Bellevue, WA

Thanks to Ido Dagan and Dan Roth for permission to use slides

About Us

- Sebastian Pado

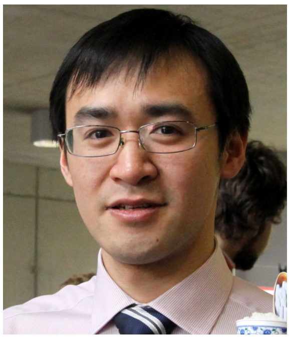
  Professor of Computational Linguistics

  Heidelberg University, Heidelberg, Germany

- Rui Wang

  Researcher in Language Technology

  German Research Center for Artificial Intelligence, Saarbrücken, Germany



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

Part 1: Overview

- Language Processing
  - Variability in Language
- Textual Entailment
  - What is it and what is it good for?
- The Textual Entailment ecosystem
  - The “Recognizing Textual Entailment” Challenges

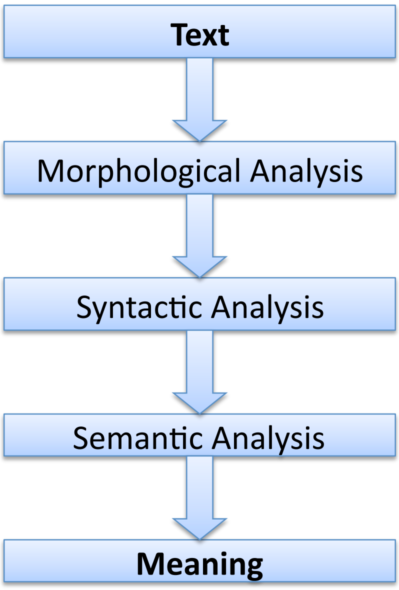
Natural Language Processing

- Text is the dominant modality to represent **knowledge** in many fields (science, industry, …)
- Text is the dominant modality in which users **interact** with computers

- We (and our computers) need to be able to
  - **extract** knowledge from texts and
  - **draw inferences**

Language Processing as Analysis

- Input: Text
- Output: Formal meaning representation
  - E.g. predicate logics, description logics, modal logics, …
- Inference: Logical calculus defined by meaning representation



Text

Morphological Analysis

Syntactic Analysis

Semantic Analysis

Meaning

Logical Entailment

- “A hypothesis H is entailed by a premise P (P ⊨ H) iff in every model where P holds, H holds as well”
  - Relevant devices: Theorem provers, model checkers, deduction systems, …

Problems of Representation

- The analysis approach formalizes language meaning **as precisely as possible: complete disambiguation**
- Language is **imprecise** and **incomplete**
  - Ambiguity:
*Yesterday, Peter passed by the **bank***
*I saw the man **with the telescope***
  - Deictic expressions:
you, he, yesterday
- Full analysis difficult and often highly ambiguous

Problems of Inference

- People are willing to accept “loose” inferences [Norvig 1987]:
    1. The cobbler sold a pair of study boots to the alpinist.
    2. The cobbler made the sturdy boots
- People use “loose speak” [Fan & Porter 2004] to formulate search queries

Is All Disambiguation Necessary?

- Consider concrete instances of inference

1. Obama addressed the general assembly yesterday
2. The president gave a speech at the UN

- To decide whether (1) implies (2), we do NOT care whether…
    - … “address” also has other senses
    - … there are other referents for “the president”
    - … what the exact date of “yesterday” is

Application-specific Processing

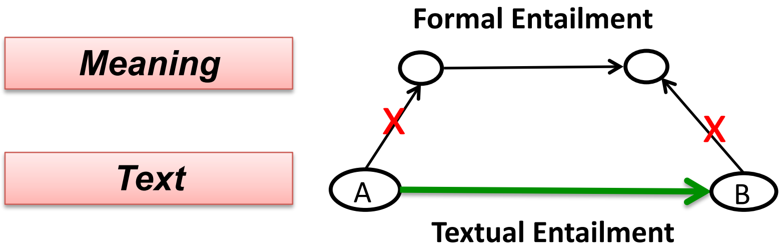
- Current dominant paradigm in language processing
  - Build task-specific models for semantic processing: Only treat **relevant** phenomena for given task
    - Semantic similarity → Distributional Methods
    - Semantic types → Named Entity Recognition
    - …
- Robust, often accurate, models for individual tasks
- BUT huge no generalization / consolidation

Fragmentation of processing, no "theory"

Reimagining Semantic Processing

- The goal of processing is **not** to analyze individual texts
- Instead: determine the **relationships** that hold among texts

- Most important relationship: **Entailment**
  - Does Text A imply Text B? (including common sense cases)



What Is Textual Entailment?

- TE is a **framework** for semantic language processing
  - **Not a concrete model!**

- Components:
  1. Concept of entailment (and its properties)
  2. Perspective on language processing centered around **variability**
  3. Body of research, community

Entailment

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

t ***entails*** *h (t⇒h)* if humans reading *t* will infer that *h* is most likely true [Dagan & Glickman 2004]

Textual vs. Logical Entailment

- Logical Entailment:
  - Define formal representation language
  - Define translation into formal language
  - **Entailment is what the representations say it is**
- Textual Entailment:
  - Collect entailment judgments for text pairs
  - Develop processing methods that can reproduce these judgments
  - **Entailment is what the speakers say it is**

Textual vs. Logical Entailment

"**Loose**" entailment: Textual but not logical

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

"**Uninformative**" entailment :Logical but not textual

T: The technological triumph known as GPS was incubated in the mind of Ivan Getting.
H: Two plus two equals four.

Entailment and Variability

- Variability is a central fact of language
  - TE can be seen as the task of distinguishing **meaning-preserving** from **meaning-changing** variability

The Global Positioning System was incubated in the mind of an American physicist, Ivan Getting. ⇒ **Ivan Getting invented GPS.**

Abbreviations, Paraphrases, Change of Voice, Apposition, …

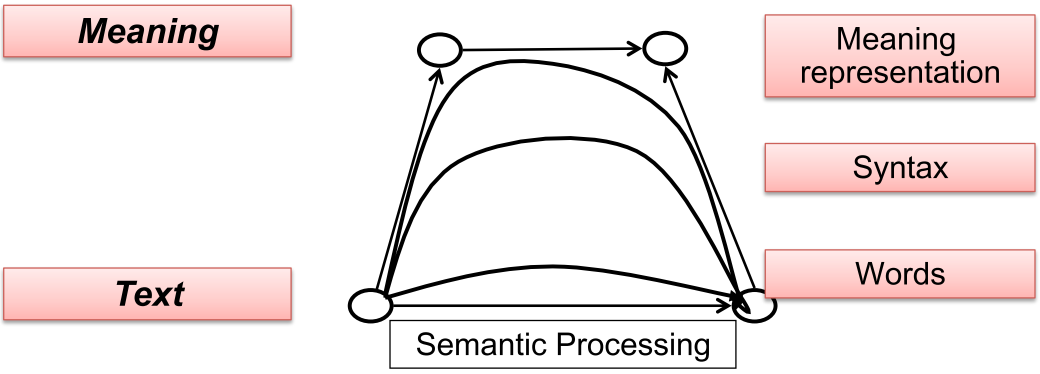
Variability and Inference

- Variability is important in, but not all of, inference:
  - Inferences about language variability
    - I **bought** a watch => I **purchased** a watch
  - Inferences about the extra-linguistic world
    - it **rained** yesterday => it **was wet** yesterday
- Most (Text, Hypothesis) pairs involve both
  - No definite boundary between the two
- Crucial role of both kinds of knowledge (**cf. Part 3**)

Recognizing Textual Entailment

- “Common ground” for processing approaches
  - Contrast to analysis-centered approach
    - No abstract gold standard
- Allows direct comparison of different processing approaches (**cf. Part 2)**
  - “Depth of analysis” up to each approach
- Mid-term goal: Identification and combination of best strategies from various approaches (**cf. Part 5**)

“Easy-first processing”



Meaning

Text

Meaning representation

Syntax

Words

Semantic Processing

- Perform as many inferences over natural language representations as possible
- Resort to formal meaning representation when necessary

Why Work With Textual Entailment?

- Conceptual benefits:
  - A concept of “common sense” inference
  - Alternatively, framework to address language variability
  - Novel perspective on the needs of language processing
- Practical benefits:
  - An attractive “meta framework” for language processing
  - A unified perspective on many research questions at the boundary of language processing, machine learning, and knowledge representation

Textual Inference in Applications

QA:

Question: What affects blood pressure?

“Salt causes an increase in blood pressure”

IR:

Query: symptoms of IBS

“IBS is characterized by vomiting”

Story Comprehension

(ENGLAND, June, 1989) - Christopher Robin is alive and well. He lives in England. He is the same person that you read about in the book Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. […]

1. Christopher Robin was born in England.
2. Winnie the Pooh is a title of a book.
3. Christopher Robin's dad was a magician

cf. also Part 4

Practical Role of Textual Entailment

- Young task: Introduced about 10 years ago
- A prominent concept in semantic processing
  - 20000 Google Scholar hits for "Textual Entailment"
- Important role: The "Recognizing Textual Entailment" Challenges (PASCAL/NIST)
  - Yearly preparation of new datasets
    - Created utilizing (or simulating) reductions from real systems' output
  - Shared task: Practical and conceptual advances

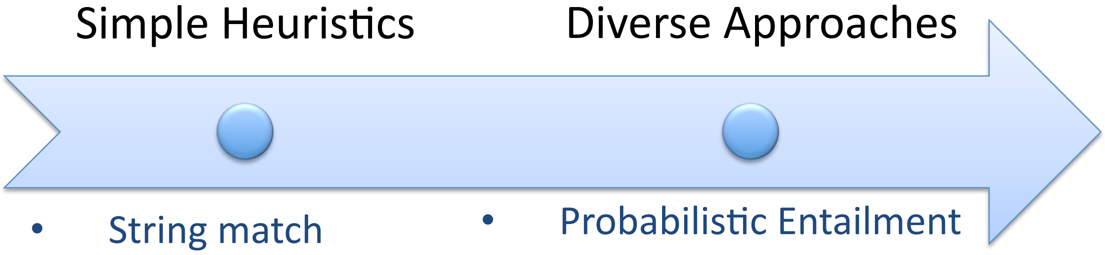
RTE Data

	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

Developments of the Task

- RTE 1, 2: Single-sentence T-H pairs
- RTE 3+: Longer texts
- RTE 4: Contradiction
  – Generalization to more relations
- RTE 5: Search Task (single H, multiple Ts)
- RTE 6+: Application-specific datasets
  – RTE 8 (2013): Student Response Analysis

Development of Methods



Early years:
Simple Heuristics

- String match
- Lexical coverage
- etc.

Now:
More Principled,
Diverse Approaches

- Probabilistic Entailment [Shnarch et al, 2011]
- Tree Edit Models [Heilman & Smith, 2010]
- Entailment as Search [Stern & Dagan 2011, 2012]

Remainder of this Tutorial

- Part 2 [RW]: Classes of Strategies and Learning
  - Which methods can be used to decide entailment?
- Part 3 [SP]: Knowledge and Knowledge Acquisition
  - What kinds of knowledge are necessary? Where can we find them or how can we learn them?
- Part 4 [SP]: Applications
  - How can language processing applications use entailment?
- Part 5 [RW]: Multilingual, Component-based System Building
  - How can we develop sustainable entailment systems?

Reference List

- I. Dagan and O. Glickman (2004). Probabilistic textual entailment: Generic applied modeling of language variability. Proceedings of the PASCAL workshop on Learning Methods for Text Understanding and Mining.
- J. Fan and B. Porter (2004). Interpreting Loosely Encoded Questions. Proceedings of AAAI, 399-405.
- Heilman, M. and N. Smith (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. Proceedings of NAACL, 1011–1019.
- P. Norvig (1987). Inference in text understanding. Proceedings of AAAI, 561–565.

Reference List

- Shnarch, E., J. Goldberger, and I. Dagan (2011). A probabilistic modeling framework for lexical entailment. Proceedings of ACL, 558–563.
- Stern, A. and I. Dagan (2011). A confidence model for syntactically-motivated entailment proofs. Proceedings of RANLP, 455–462.
- Stern, A. , R. Stern, I. Dagan, and A. Felner (2012). Efficient search for transformation-based inference. In Proceedings of ACL, 283-291.

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

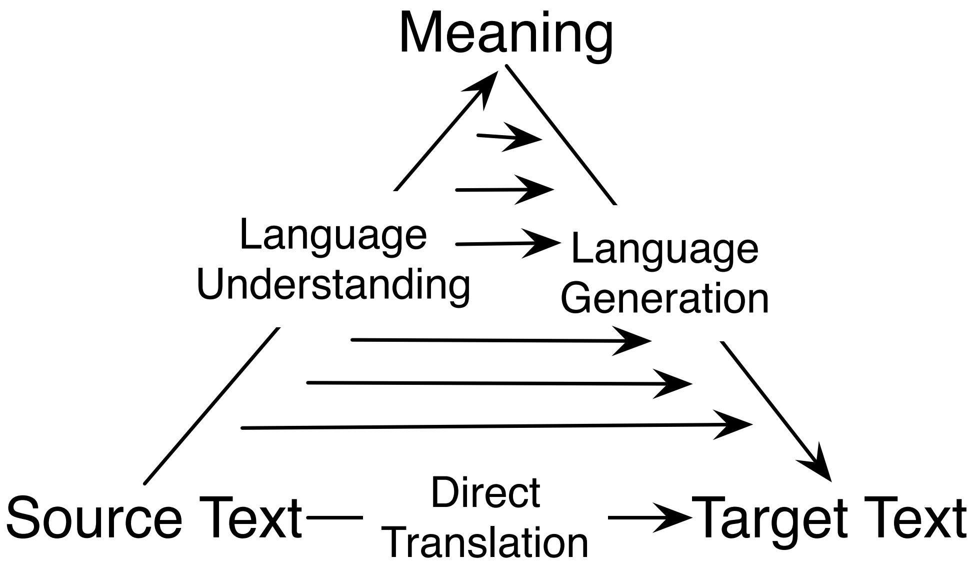
Thanks to Ido Dagan for permission to use slide material

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



Meaning

Language Understanding

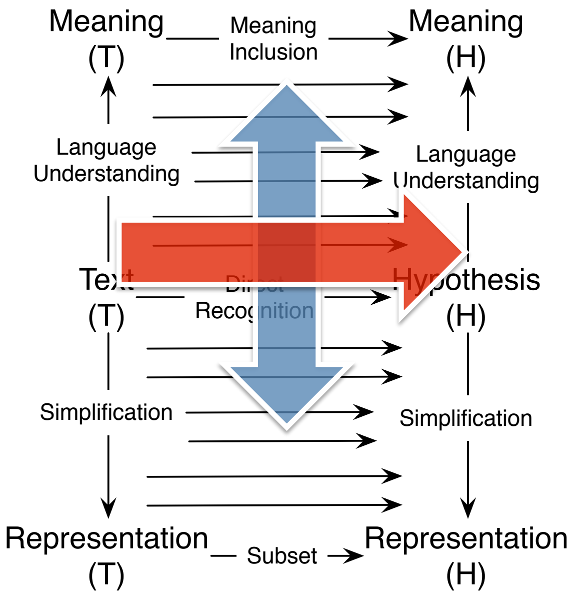
Language Generation

Source Text

Direct Translation

Target Text

RTE Rectangle



Meaning (T)

Meaning Inclusion

Meaning (H)

Language Understanding

Language Understanding

Text (T)

Direct Recognition

Hypothesis (H)

Simplification

Simplification

Representation (T)

Subset

Representation (H)

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	.

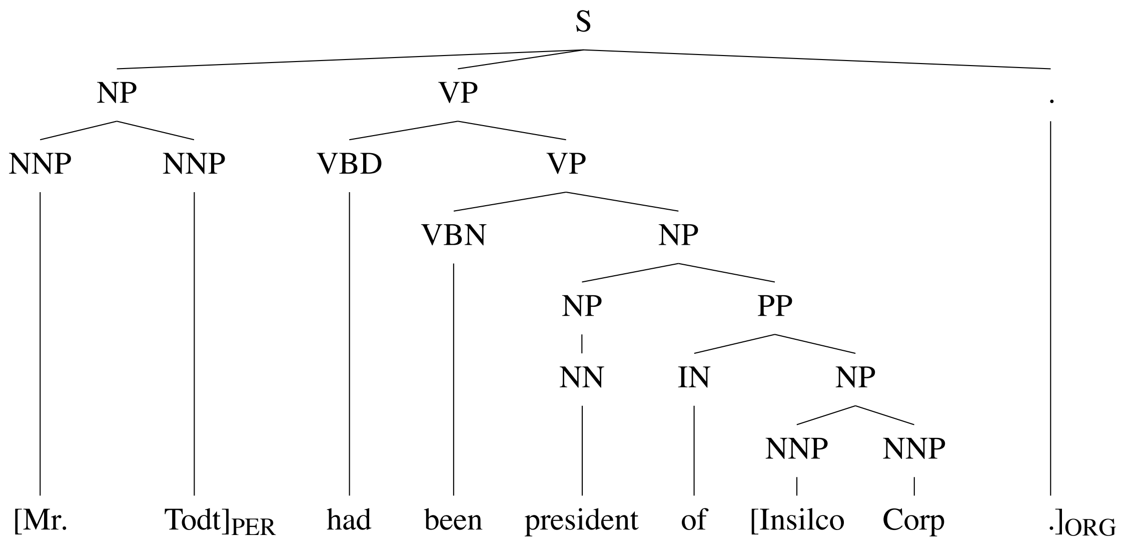
From TreeTagger website

Constituents

- Chunking
- Named-Entity Recognition
- Constituent Parsing

NER:
70~90%

An Example



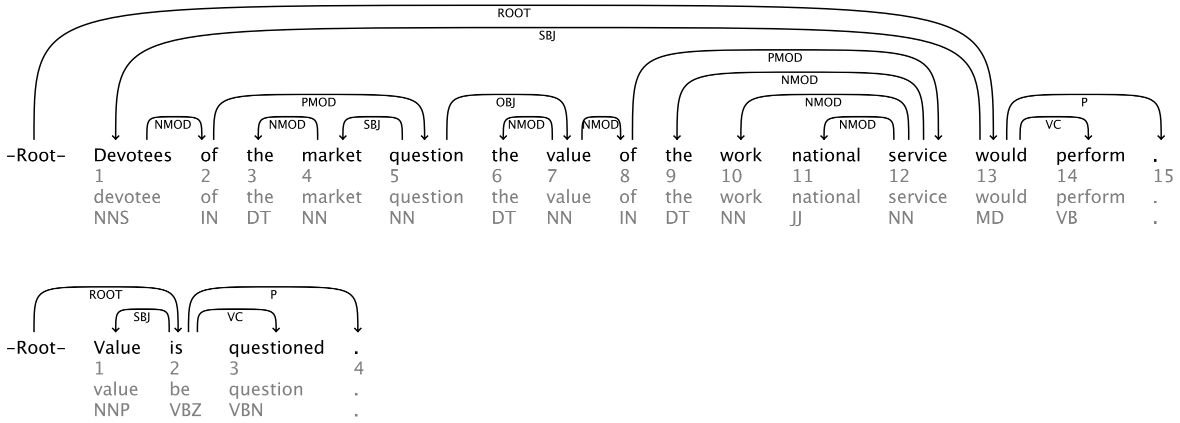
From Stanford NER (Finkel and Manning, 2009)

Dependency

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

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

An Example

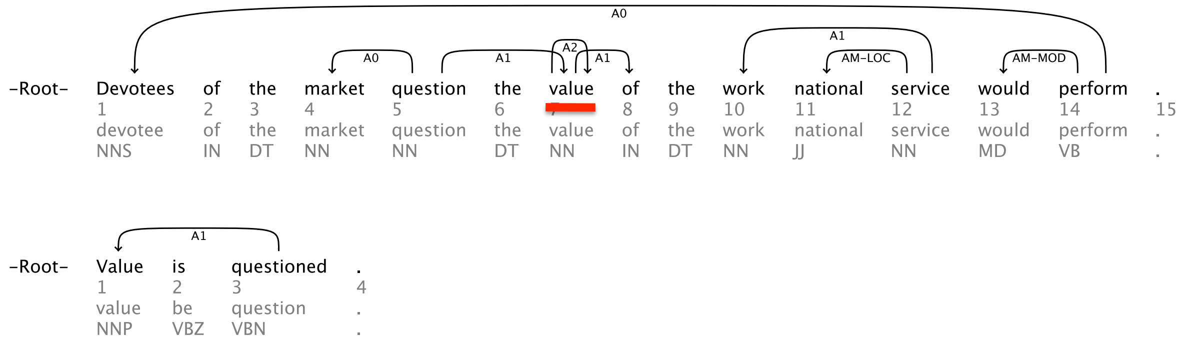


-Root-	Devotees	of	the	market	question	the	value	of	the	work	national	service	would	perform	.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	devotee	of	the	market	question	the	value	of	the	work	national	service	would	perform	.
	NNS	IN	DT	NN	NN	DT	NN	IN	DT	NN	JJ	NN	MD	VB	.

-Root-	Value	is	questioned	.
	1	2	3	4
	value	be	question	.
	NNP	VBZ	VBN	.

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

An Example (cont.)



-Root-	Devotees	of	the	market	question	the	value	of	the	work	national	service	would	perform	.
	1	2	3	4	5	6	[redacted]	8	9	10	11	12	13	14	15
	devotee	of	the	market	question	the	value	of	the	work	national	service	would	perform	.
	NNS	IN	DT	NN	NN	DT	NN	IN	DT	NN	JJ	NN	MD	VB	.

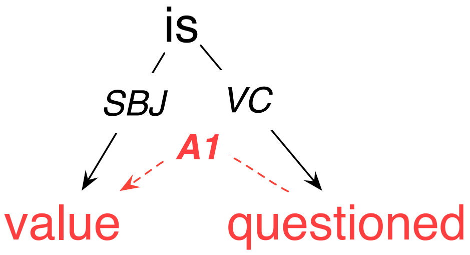
-Root-	Value	is	questioned	.
	1	2	3	4
	value	be	question	.
	NNP	VBZ	VBN	.

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)

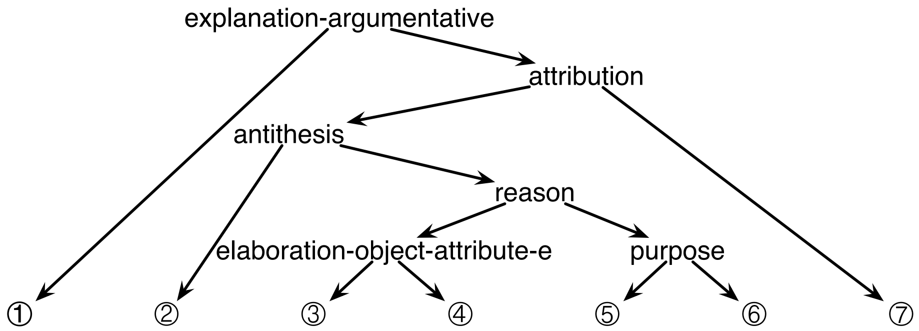
- Core arguments: A0-A5
  - different semantics for each verb
  - specified in the PropBank Frame files

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

Discourse

- Coreference Resolution
- Event Structure
- Discourse Parsing

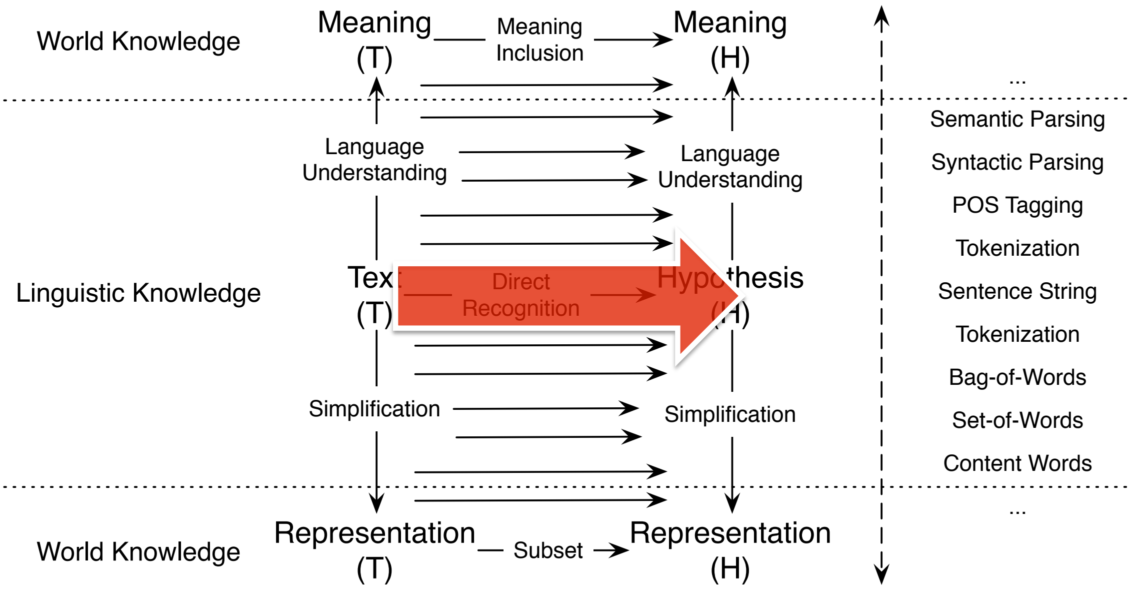
An Example



[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)



World Knowledge

Meaning (T) — Meaning Inclusion → Meaning (H)

Language Understanding → Language Understanding

Linguistic Knowledge

Text (T) — Direct Recognition → Hypothesis (H)

Simplification → Simplification

World Knowledge

Representation (T) — Subset → Representation (H)

...

Semantic Parsing

Syntactic Parsing

POS Tagging

Tokenization

Sentence String

Tokenization

Bag-of-Words

Set-of-Words

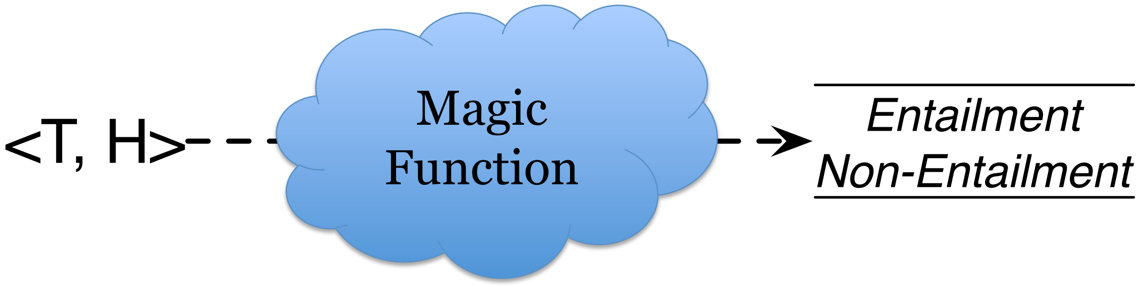
Content Words

...

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

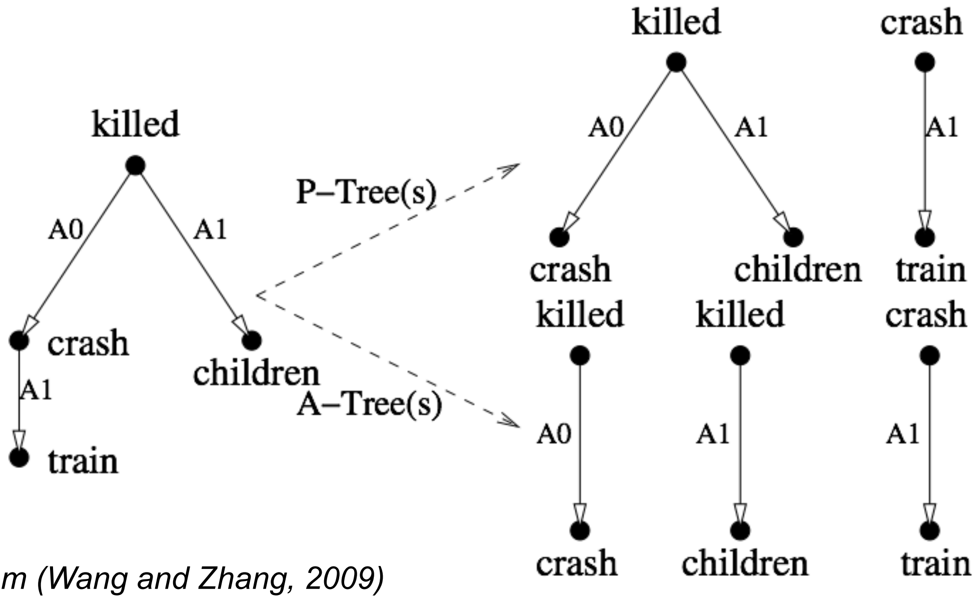
From (Mehdad and Magnini, 2009)

Dependencies

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

- E_1’ = |Triple(T)⋀Triple (H)| / |Triple(H)|

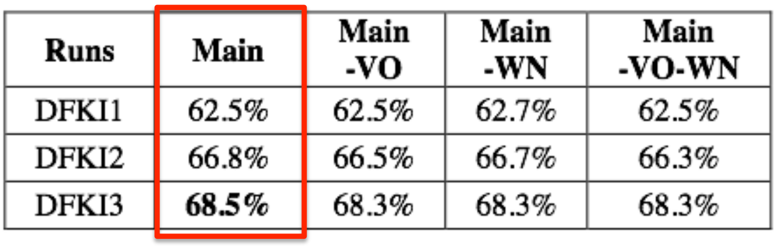
Dependencies (cont.)



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'' = $|\mathrm{Subtree(T)} \wedge \mathrm{Subtree(H)}| / |\mathrm{Subtree(H)}|$

Subtrees

$T_1 \Rightarrow H_1$
T_1 *"Farmers feed cows animal extracts"*
H_1 *"Cows eat animal extracts"*

$T_2 \Rightarrow H_2$
T_2 *"They feed dolphins fish"*
H_2 *"Fish eat dolphins"*

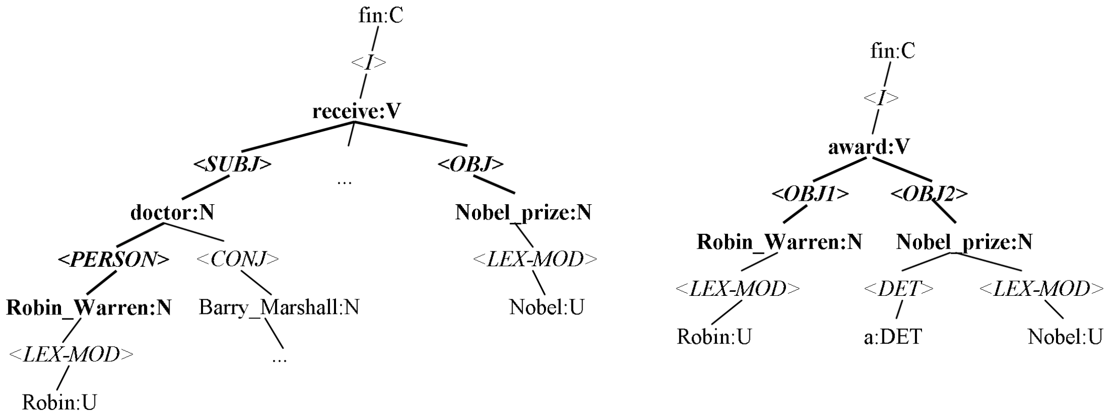
feed X Y → X eat Y

$T_3 \Rightarrow H_3$
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)

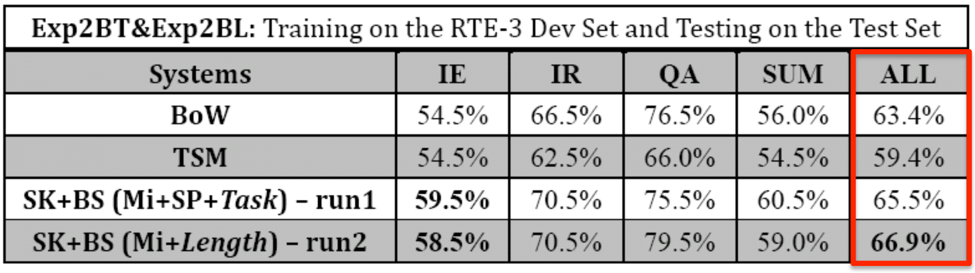
Results

- RTE-2

Bag-of-Words Similarity

Tree Skeleton

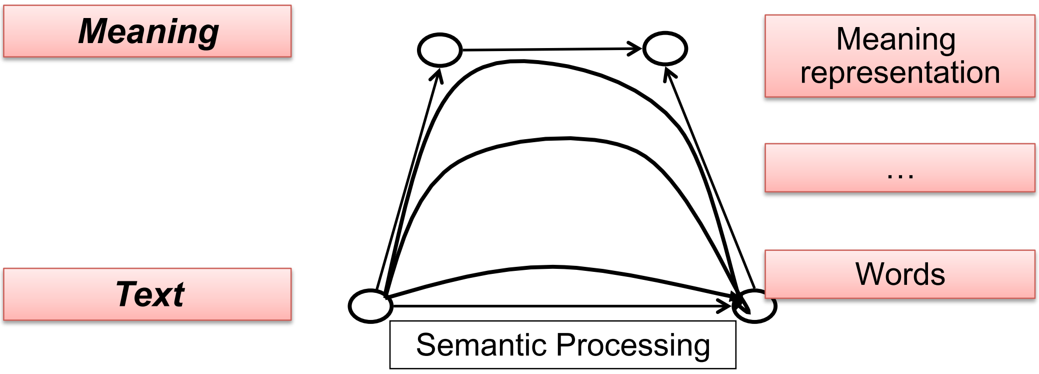
- RTE-3



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

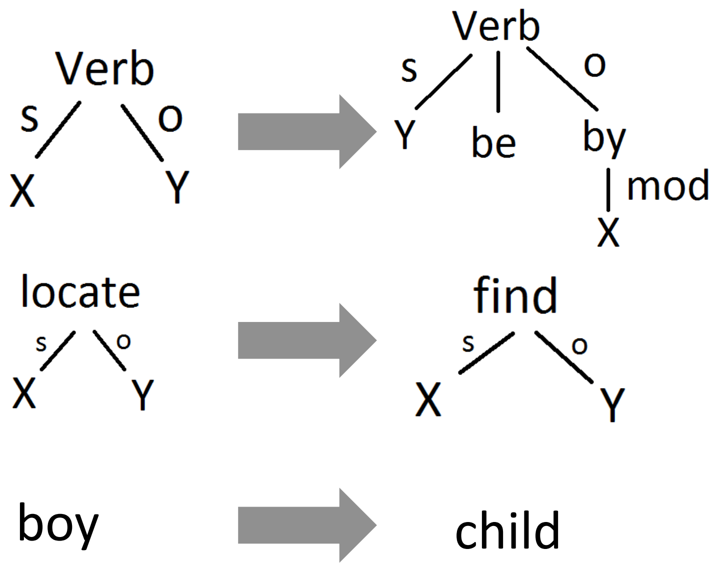
- BIUTEE (Stern and Dagan, 2011)
  - On-the-fly operations
  - Cost model
  - Search for the best inference

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

Verb s o X Y → Verb s o Y be by mod X

Lexical Syntactic

locate s o X Y → find s o X Y

Lexical

boy → child

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

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

Learn

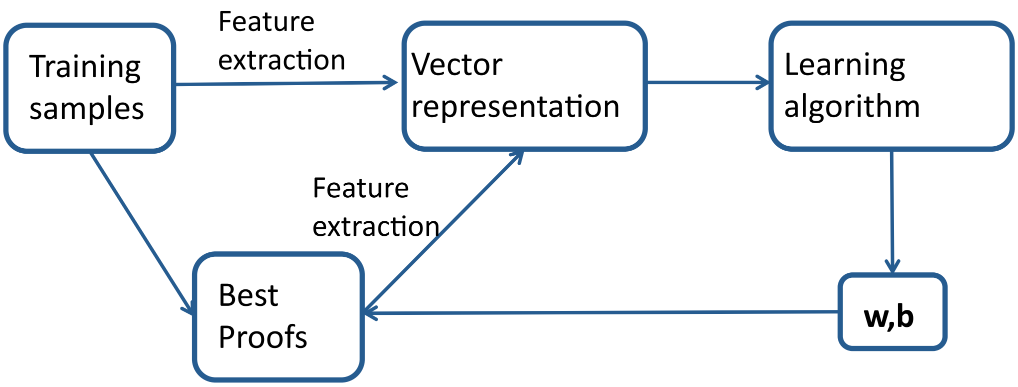
Variant of (Raina et al., 2005)

Search the Best Proof

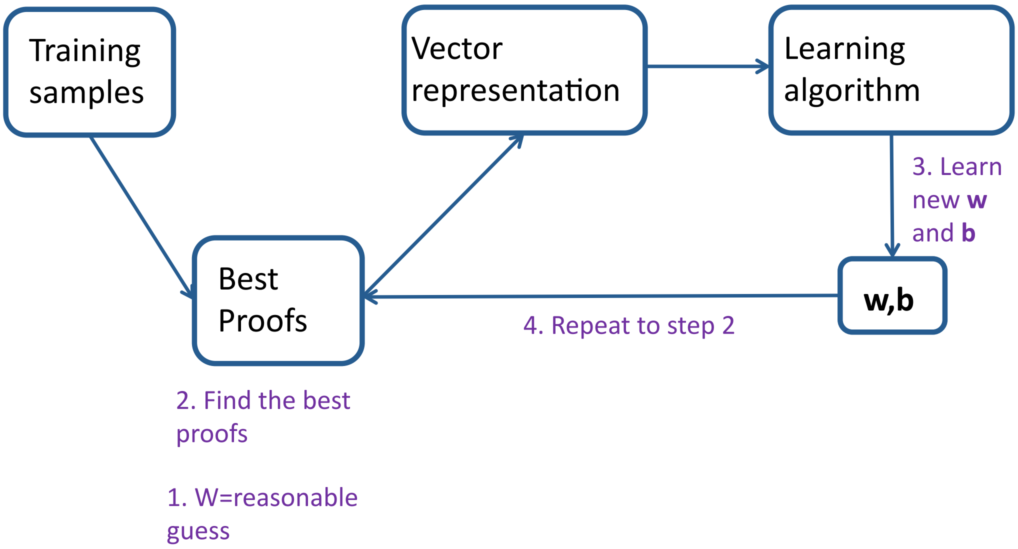
$T \rightarrow H$			$T \nrightarrow H$		
Proof #1	$T \rightsquigarrow H$	✗	Proof #1	$T \rightsquigarrow H$	✗
Proof #2	$T \rightsquigarrow H$	✓	Proof #2	$T \rightsquigarrow H$	✗
Proof #3	$T \rightsquigarrow H$	✗	Proof #3	$T \rightsquigarrow H$	✗
Proof #4	$T \rightsquigarrow H$	✗	Proof #4	$T \rightsquigarrow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

- “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



Training samples

Vector representation

Learning algorithm

3. Learn new **w** and **b**

w,b

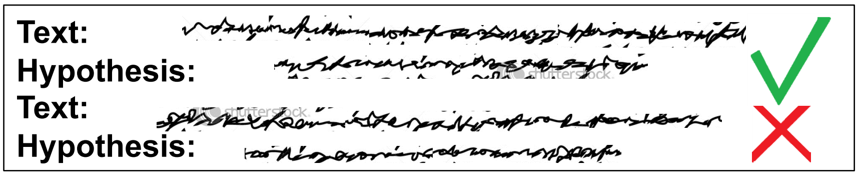
4. Repeat to step 2

Best Proofs

2. Find the best proofs

1. W=reasonable guess

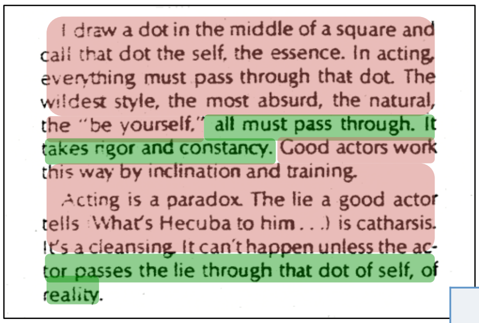
Performance (Classification)



Text:
Hypothesis:
Text:
Hypothesis:

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.

Unbalanced!

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.*

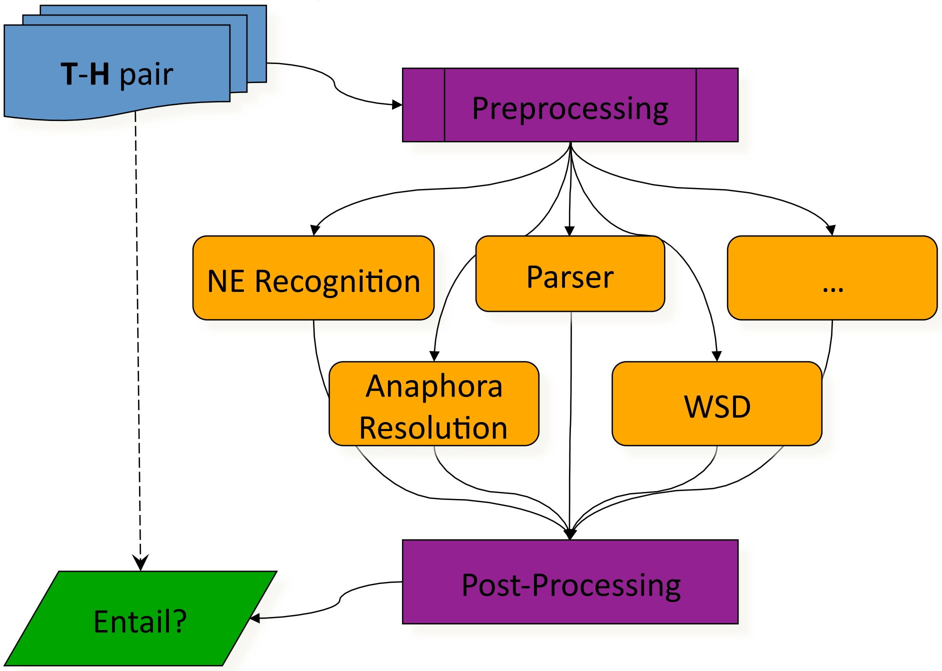
- **H**: *Mr. Bush is proposing that workers be allowed to divert their payroll taxes into private accounts.*

An Example

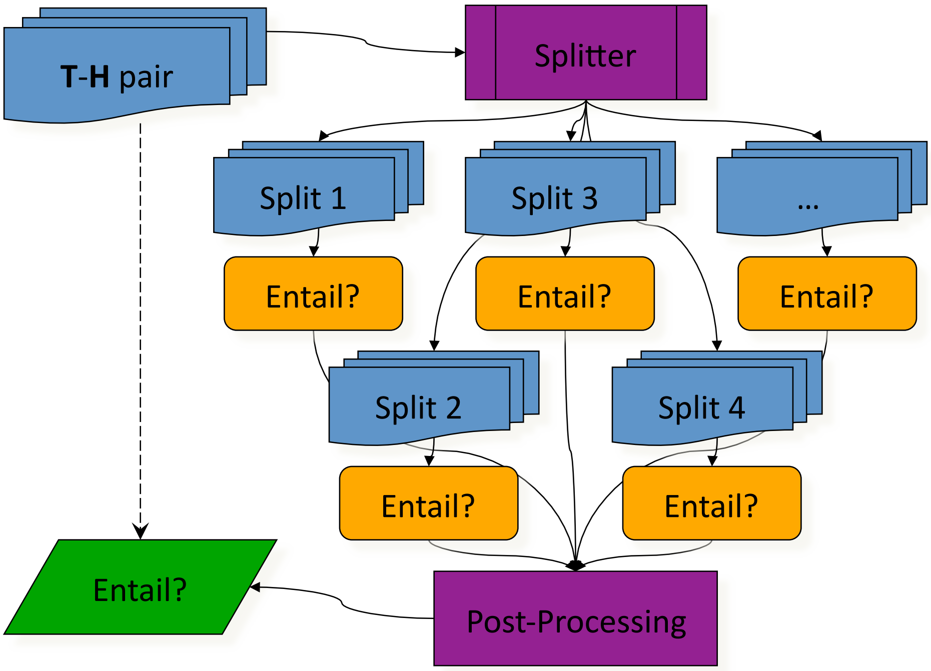
- **T**: *Bush used his weekly radio address to try to build support for his plan to allow workers to divert part of their Social Security payroll taxes into private investment accounts.*

- **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

Temporal Anchoring

Tree Skeleton Matching

Named-Entity Matching

- 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.

Textual Entailment
Part 3: Knowledge Resources and Knowledge Acquisition

Sebastian Pado	Rui Wang
Institut für Computerlinguistik	Language Technology
Universität Heidelberg, Germany	DFKI, Saarbrücken, Germany

AAAI 2013, Bellevue, WA

Thanks to Ido Dagan for permission to use slide material

This part of the tutorial

1. Overview: Types of Inference Knowledge
2. Use Case 1: Acquiring Asymmetrical Similarity
3. Use Case 2: Truth Status in Context

Part 1: Types of Inference Knowledge

Inference Rules

- TE assesses if H can be inferred from T
  - Requires linguistic knowledge, world knowledge
- Sentence-level entailment is always *decomposed* into atomic (subsentential) inference steps
  - Corresponding to *compositional* meaning construction
- Valid atomic inference steps can be represented as **inference rules** *a* → *b*
  - a, b almost arbitrary linguistic representations
  - Various linguistic levels (lexical, syntactic, phrasal, …)

Application of Inference Rules

- Resources with inference rules are used in virtually every single Textual Entailment system:
  - Transformation-based approaches:
    Inference rules motivate proof steps
  - Classification-based approaches:
    Inference rules inform similarity features
- What types of inference knowledge is helpful?
  - Clark et al. (2006): analysis of knowledge types
  - Mirkin et al. (2009): ablation tests for various knowledge resources on entailment

The Challenge for Knowledge

- Textual Entailment requires its inference rules to have both *high precision* and *high recall*
  – Low precision: rules do more harm than good
  – Low recall: rules are irrelevant

- Complementary behavior of resources:
  – Manually constructed resources often lack recall
  – Automatically constructed resources often lack precision

Normalization Knowledge

- Named Entities, Abbreviations, Acronyms, etc.
  - Sources: Machine-readable dictionaries
  - Status: Relatively unproblematic

Mr. Clinton ⇔
Bill Clinton ⇔
President Clinton

US ⇔
U.S. ⇔
United States

Lexical Knowledge

- **Nominal** relations: Synonymy, Hyponymy
  - Sources: WordNet, Distributional Thesauri
  - Status: most widely used type of knowledge, still recall problems
    Use case 1
- **Verbal** relations: Causation, Presupposition
  - Sources: WordNet, VerbOcean
  - Status: also widely used, but both recall and precision problems

Peter owns a kitchen **table** ⇒
Peter owns an **object**

Peter **buys** a kitchen table ⇒
Peter **owns** a kitchen table

Syntactic Knowledge

- Structural variation (relative clauses, genitives, active/passive, etc.)
  - Sources: syntactic rule bases
  - Status: often used, but limited recall

Peter**, who** sleeps soundly, … ⇒
Peter sleeps soundly

Peter **broke** the vase. ⇒
The vase **was broken** by Peter.

Paraphrase Knowledge

- Inferences that cannot be captured at word level
  - Variety of phenomena
  - Range from simple to very difficult
- Sources: Corpora (both monolingual and parallel)
- Status: Very difficult to balance precision and recall

X **buys** Y from Z ⇒
Z **sells** Y to X

X **gave** me **a hand** ⇒
X **helped** me

X was **a Yorkshireman by birth** ⇒
Y was **born in Yorkshire**

World knowledge

- **Factual** Knowledge
  - Sources: Gazetteers, Wikipedia
- “**Core theories**”
  [Clark et al. 2006]
  - Sources: mostly hand-coded

T: Paris is **in France** ⇒
H: Paris is **in Europe**

T: Easter 2011 was **on April 24** ⇒
H: Easter 2011 was **between April 20 and 30**

- Status: Superficial treatment in most TE systems
  - Interesting direction: Unstructured vs. structured data – compare IBM Watson (Kalyanpur et al. 2012)

Sentential Context

- Sentential context influences inference
- Variety of factors
  - Monotonicity
  - Clause Truth Status
    - **Use case 2**
  - Presupposition
- Status: Current research

T: Peter sees **a** poodle ⇒
H: Peter sees a dog

T: Peter sees **no** poodles ⇏
H: Peter sees no dogs

T: Peter **managed** to come ⇒
H: Peter came

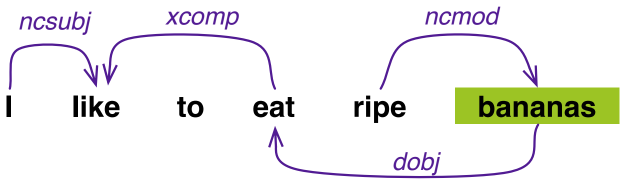
T: Peter **promised** to come ⇒**?**
H: Peter came

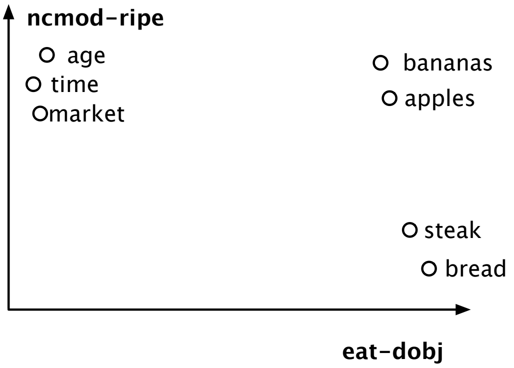
Use Case 1: Asymmetrical Similarity (Kotlerman et al. 2010)

Distributional Semantics

- Goal: Learn lexical inference rules *a* ⇒ *b* from corpora
- Distributional Semantics: “You shall know a word by the company it keeps” [Firth, 1957]
- An unsupervised way to model word meaning:
  - Observe in which contexts a word occurs
  - Represent words as vectors in high-dimensional space
  - Vector similarity correlates with semantic similarity
- Applied to many tasks in language processing [Turney & Pantel 2010]

Distributional Similarity





Standard Similarity Measures

- Cosine: Angle between vectors

$$cos(\vec{u},\vec{v}) = \frac{\sum_i u_i \cdot v_i}{\sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2}}$$

- Lin’s similarity: Pointwise mutual information of shared features

$$PMI(u,f) = \log \frac{P(u,f)}{P(u)P(f)}$$

$$lin(\vec{u},\vec{v}) \frac{\sum_{i:u_i>0,v_i>0}[PMI(u,f_i)+PMI(v,f_i)]}{\sum_{i:u_i>0} PMI(u,f_i) + \sum_{i:v_i>0} PMI(v,f_i)}$$

Acquiring Entailment Rules

- Standard approach: For each target word, find the highest-similarity neighbors
  - Synonyms (and other close semantic relations): Lexical entailment rules [Lin 1998]
  - Generalization from words to dependency paths: Paraphrase rules [Pantel and Lin 2001]

Asymmetry of Inference Rules

- Standard similarity measures are symmetrical…
- …Inference rules are asymmetrical!

"Peter has a pet dog"

⇒ ⇍

"Peter has a pet poodle"

bank* ⇒ *company

company* ⇏ *bank

Symmetric Similarity - Results

- Most similar words for *food*:

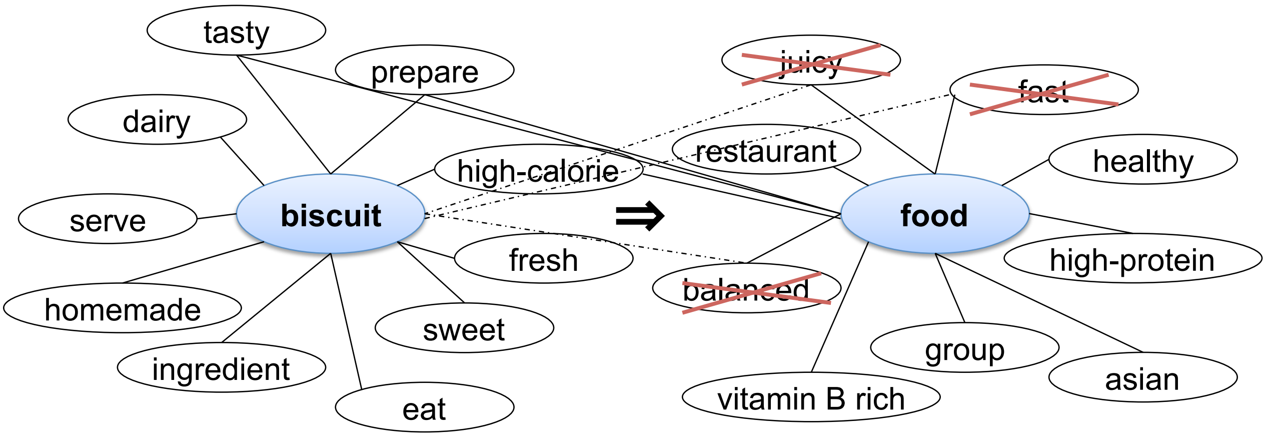
meat clothing water sugar
beverage foodstuff coffee material
goods textile meal chemical
medicine fruit tobacco equipment

- Evaluation of resources for entailment (Mirkin et al. 2009)

Resource	*Precision*	*Recall*
WordNet	55%	20%
Wikipedia	45%	7%
Dist.sim.(Lin)	28%	43%

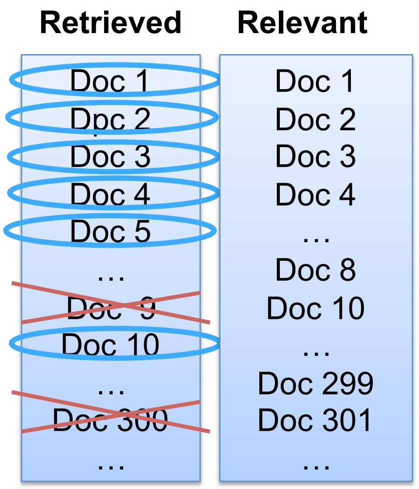
Distributional Inclusion

- If u ⇒ v, then the characteristic contexts of u are expected to be characteristic for v, but not vice versa [Weeds et al., 2004]



Average Precision

- Average Precision: Measure from Information Retrieval to assess search engine output (ranked list)
- Goals:
  - retrieve many relevant documents
  - retrieve few irrelevant documents
  - retrieve relevant docs early in list



Retrieved

Doc 1
Doc 2
Doc 3
Doc 4
Doc 5
…
~~Doc 9~~
Doc 10
…
~~Doc 300~~
…

Relevant

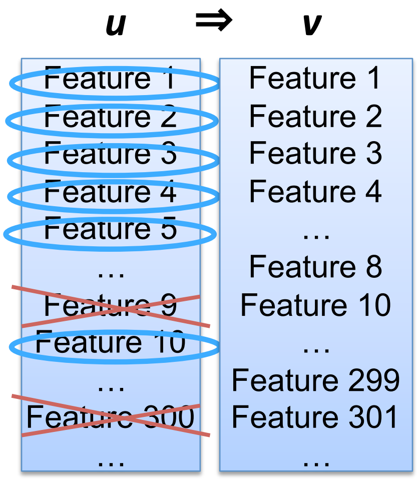
Doc 1
Doc 2
Doc 3
Doc 4
…
Doc 8
Doc 10
…
Doc 299
Doc 301
…

$$AP = \frac{\sum_i Prec(d_1, \ldots, d_i) \cdot rel(d_i)}{\sum_i rel(d_i)}$$

(where *rel* is 1 if doc is relevant)

Balanced Average Precision

- Average Precision can applied to vectors to measure **feature inclusion**:
  - Retrieved, Relevant ⇒ u,v
  - Documents ⇒ Features
- u ⇒v if top features of v are shared by u and u has few other top features
- Modifications: *rel'* is graded relevance based on rank; balance with Lin similarity to alleviate sparse *v* vectors



u ⇒ *v*

Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
…
~~Feature 9~~
Feature 10
…
~~Feature 300~~
…

Feature 1
Feature 2
Feature 3
Feature 4
…
Feature 8
Feature 10
…
Feature 299
Feature 301
…

$$balAPinc(\vec{u}, \vec{v}) = \sqrt{lin(\vec{u}, \vec{v}) \cdot \frac{\sum_i Prec(f_1^u, \ldots, f_i^u) \cdot rel'(f_i^u)}{|\vec{u}|}}$$

Directional similarity - results

- The most similar words to *food*:

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

- For more evaluation, see Kotlerman et al. (2010)

Use Case 2: Truth Status in Context (Lotan et al. 2013)

Motivation

- Reminder: Context influences entailment patterns
  - Complex phenomenon
- Subproblem: *Truth status* of *clauses* in context
  - Case 1: Clause is true (positively entailed) **[+]**
  - Case 2: Clause is false (negatively entailed) **[-]**
  - Case 3: Clause truth is unknown **[?]**

T: Peter managed to sleep. ⇒
H: Peter slept.

T: Peter failed to sleep. ⇒
H: Peter did not sleep.

T: Peter promised to sleep.⇒**?**
H: Peter slept.

Determining Clause Truth

- Clause Truth is determined primarily by three factors:
  - Embedding words/phrases
  - Modifiers
  - Specific structures (presuppositions)
- Each factors can be associated with a *signature*
  - Description of its influence on clause truth

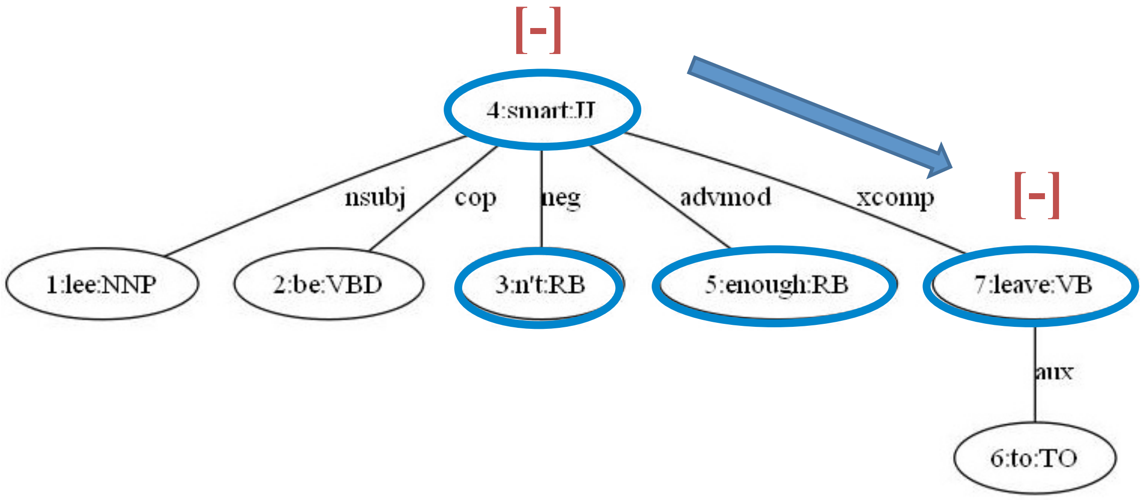
Signatures

1. Negation: **[+] → [-], [-] →[+]**
2. Modality markers (many adverbs, modal verbs):
   [+] → [?], [-] →[?]
3. Factive embeddings (knowledge/perception/emotion):
   [+] → [+], [-] →[+]
4. Presuppositions (relative clauses, definite NPs): **[+]**
   [Kiparsky and Kiparsky 1970]
5. Implicative embeddings: various patterns
   [Karttunen 1971, 2012]
   - have the time: **[+] → [+], [-] →[-]**
   - make sure: **[+] → [+], [-] →[?]**

The "TruthTeller" system

- Lexicon of modifiers and embedding words/phrases
  - About 2000 entries with signatures
  - Constructed semi-automatically
- Recursive algorithm inspired by Natural Logic [Lakoff 1970, MacCartney & Manning 2009]
  - Determine truth status of all clauses in a sentence
- http://u.cs.biu.ac.il/~nlp/downloads/TruthTeller/

TruthTeller Example



Lee wasn't smart enough to leave

Evaluation

- Evaluation against manual truth status labels
- Most frequent class baseline: 77% accuracy (**[+]**)
- Total accuracy: 89%

Truth Status	Recall	Precision	Occurrences
[+]	87.3%	98%	120
[-]	74%	83%	50
[?]	91.4%	70%	48

- No evaluation integrated in RTE system yet

Take-Home Messages

- Knowledge plays a central role in deciding TE
  - Can be represented uniformly with entailment rules
  - Multiple layers of linguistic and world knowledge
- Manual resources (coverage issue) vs. automatically acquired resources (accuracy issue)
- Use Cases:
  - Better automatic acquisition with asymmetrical similarity
  - More precise context modeling with truth status

Reference List

- Clark, P., Murray, W. R., Thompson, J., Harrison, P., Hobbs, J. R., Fellbaum, C. (2007). On the Role of Lexical and World Knowledge in RTE-3. Proceedings of the ACL Workshop on Textual Entailment and Paraphrasing, 54–59.
- J.R. Firth (1957). Papers in Linguistics 1934–1951. London: Oxford University Press.
- Geffet, M. and Dagan, I. (2004). Feature Vector Quality and Distributional Similarity. Proceedings of COLING, 247-254.

Reference List

- Kalyanpur, A., Boguraev, BK., Patwardhan, S., Murdock, JW. , Lally, A., Welty, C., Prager, JM. , Coppola, B., Fokoue-Nkoutche, A., Zhang, L. (2012). Structured data and inference in DeepQA. IBM Journal of Research and Development, vol. 56 3.4: 10–1.
- Karttunen, L. (1971). Implicative Verbs. Language (47), 340-58.
- Karttunen. L. (2012). Simple and Phrasal Implicatives. Proceedings of *SEM, 124-131.
- Kiparsky, P. and Kiparsky, C. (1970). Fact. In M. Bierwisch and K.E. Heidolph (eds), Progress in Linguistics, 143-73.

Reference List

- Kotlerman, L., and Dagan, I., and Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. Natural Language Engineering 16(4), 359-389.
- Lakoff, G. (1970). Linguistics and natural logic. Synthese 22, 151-271.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. Proceedings of ACL/COLING, 768–774.
- Lin, D. and Pantel, P. (2002). Discovery of Inference Rules for Question Answering. Natural Language Engineering 7(4), 343–360.

- A. Lotan, A. Stern, and I. Dagan (2013). TruthTeller: Annotating Predicate Truth. Proceedings of NAACL, 752-757.
- MacCartney, W., and Manning, C. (2009). An extended model of natural logic. Proceedings of IWCS, 140-156.
- Mirkin, S., Dagan, I., and Shnarch, E. (2009). Evaluating the inferential utility of lexical-semantic resources. In Proceedings of EACL, 558–566.
- Turney, P. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. JAIR 37(1):141–188
- Weeds, J., Weir, D., McCarthy, D. (2004). Characterizing Measures of Distributional Similarity. Proceedings of COLING, 1015-1021.

Textual Entailment
Part 4: Applications

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

Content of Part 4

- Overview: Four paradigms for using Textual Entailment in Natural Language Processing Applications
- Use Cases for two of the paradigms:
  - Use Case 1: Machine Translation Evaluation
  - Use Case 2: Entailment Graphs for Text Exploration

Overview

Applications of Textual Entailment

- Assumption (cf. Part 1): TE can cover a substantial part of the semantic processing in NLP applications
  - Mapping of semantic (sub)tasks onto textual entailment queries
- If large datasets are involved, **division of labor**:
  1. Shallow (e.g. word based) methods generate candidates
  2. Textual Entailment methods act as filter/(re)scorer
     - Integrates “deeper” algorithms / knowledge
     - Allow shallow methods to be more liberal

Applications of Textual Entailment

- Mapping of semantic (sub)tasks onto textual entailment queries
  - Part 1: What are the Text and the Hypothesis?
  - Part 2: How is the output of the TE system used?

– Main paradigms:
  - Entailment for Validation
  - Entailment for Scoring
  - Entailment for Generation
  - Entailment for Structuring

Entailment for Validation

- Example: Question Answering [Hickl et al. 2007]
  - Step 1: Identify promising answer candidates
    - Shallow methods
  - Step 2: Turn question into statement
    - Replace question word
      (who → someone, which book → a book)
  - Step 3: **Use Textual Entailment to verify that the answer candidate entails the question-as-statement**
    - Binary decision

Example: Question Answering

Question: Who discovered Australia?
Text snippet (T): The first European to reach Australia was Willem Jansszon.
Question-as-statement (H): Someone discovered Australia.

Entailment query: The first European to reach Australia was Willem Jansszon. **⇒?** Someone discovered Australia

- Other application: Relation Extraction [Roth et al. 2009]

Entailment for Scoring

- Example: Machine Translation Evaluation [Pado et al. 2009]
  - Step 1: Create System translation with MT system
  - Hypothesis: Good system translation is *semantically equivalent* to reference translation
  - Step 2: **Use TE to verify that the reference translation entails the system translation – and vice versa!**
    - Graded decision: Degree of semantic equivalence
      - Typically easy to obtain from Textual Entailment systems
    - Details: see **Use Case 1**

Example: MT Evaluation

MT System Translation (ST): Today I will consider this reality.
MT Reference Translation (RT) : I shall face that fact today.

Entailment query 1: ST ⇒? RT

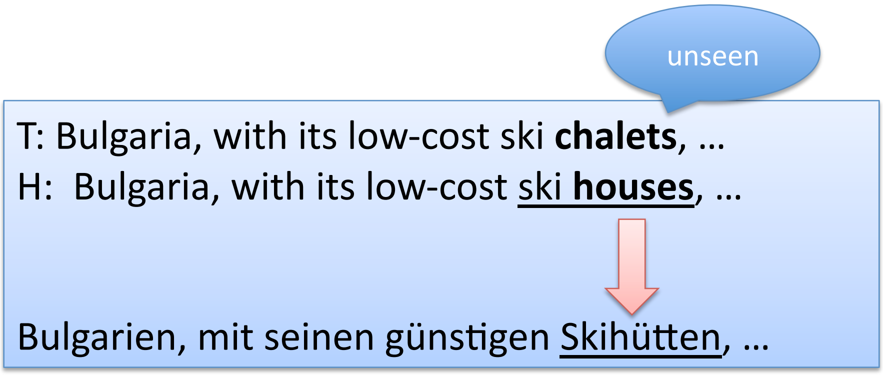
Entailment query 2: RT ⇒? ST

- Other application: Student Answer Assessment [Nielsen et al. 2009]

Entailment for Generation

- Example: Machine Translation “Smoothing” [Mirkin et al. 2009]
  - Source language terms missing from the phrase table cannot be translated
  - Parallel corpora much smaller than monolingual corpora
- **Use entailment to generate entailed “replacements” for unknown source language terms**
  - Sentence may lose some information but is translatable
    - Prefer terms that retain maximal information
  - Requires entailment system that can generate H for given T

Example: Term Replacement in MT



unseen

T: Bulgaria, with its low-cost ski **chalets**, …

H: Bulgaria, with its low-cost ski **houses**, …

Bulgarien, mit seinen günstigen Skihütten, …

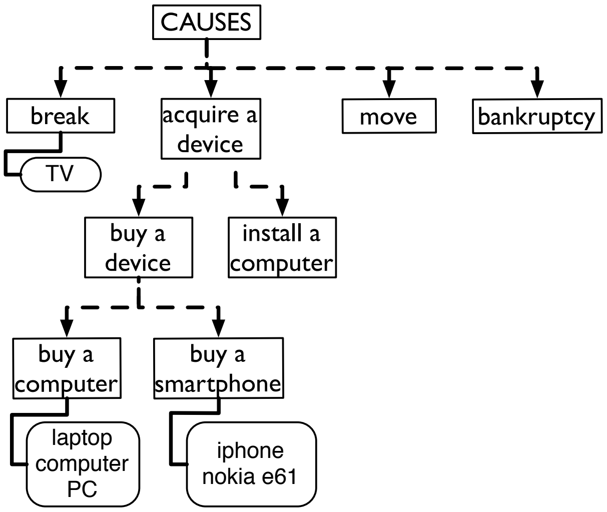
Entailment for Structuring

- Example: Information Presentation [Berant et al. 2012, **Use case 2**]
  - Starting point: Large amount of unstructured data about some concept
  - Goal: Make information easily human-accessible: Build hierarchical structure
- Step 1: Acquire atomic propositions
- Step 2: **Apply entailment queries to each pair of propositions**

- Other applications: Multi-document summarization [Harabagiu et al. 2007]

Example: Information Presentation

own a computer
we got another PC
got a laptop
bankruptcy
i couldn't download
just installed the computer
haven't used the service
TV broke
i have a nokia e 61 now
i don't use it
i'm gonna be moving
decided to buy an iphone



Use Case 1: Machine Translation Evaluation (Padó et al. 2009)

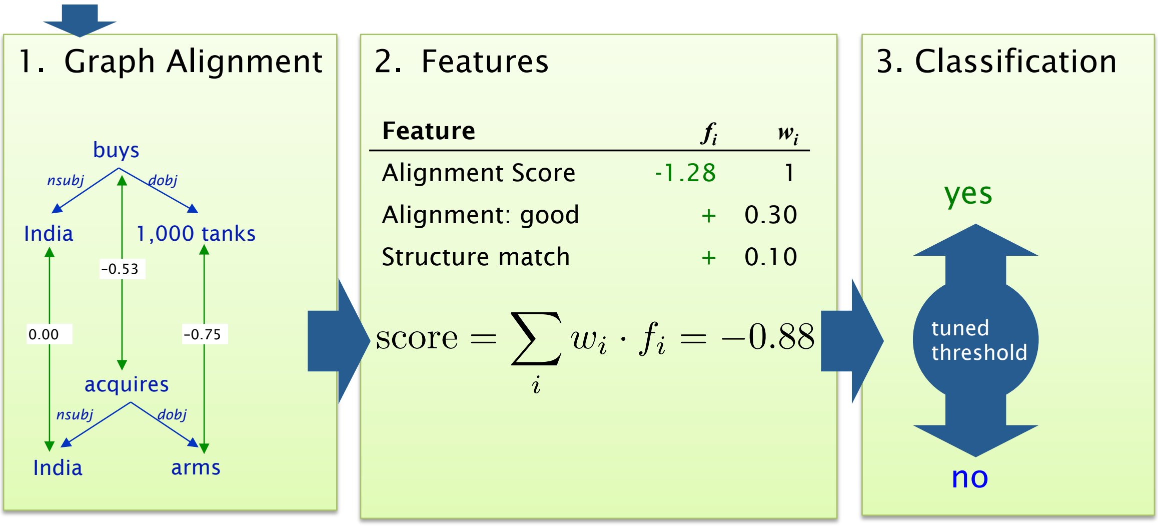
(Entailment for Scoring)

Automatic Evaluation

- Important role in Machine Translation
  - Objective *large-scale* assessment of system quality
  - Minimum Error Rate Training [Och 2002]
- Most widely used metric: BLEU
  - Pure n-gram matching
  - Problems recognizing very different translations [Callison-Burch et al. 2006, etc.]
- METEOR, TER, etc. attempt to make matching more intelligent
  - Still surface-oriented
  - Metrics should evaluate for **semantic equivalence**: TE

The Stanford Textual Entailment System

T: India buys 1,000 tanks.
H: India acquires arms.



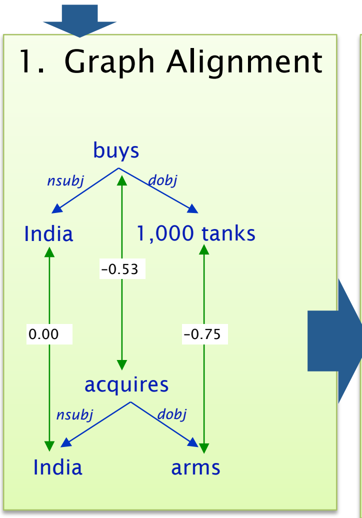
Feature	f_i	w_i
Alignment Score	-1.28	1
Alignment: good	+	0.30
Structure match	+	0.10

$$\text{score} = \sum_i w_i \cdot f_i = -0.88$$

Use for MT Evaluation

T: India buys 1,000 tanks.
H: India acquires arms.

1. Graph Alignment



2. Features

Feature	f_i	w_i
Alignment Score	-1.28	1
Alignment: good	+	0.30
Structure match	+	0.10

$$\text{score} = \sum_i w_i \cdot f_i = -0.88$$

Linear regression score =
“Degree of entailment”

Technical points

- 1. How to combine two entailment directions?
  - Option 1: Compute directions separately: Not good
  - Option 2: Combine features of both directions into one “bidirectional” regression model: Better
    - Deletion vs. addition features
- 2. How to learn feature weights?
  - Supervised learning from translation quality annotations
    - NIST OpenMT corpora: Newswire (Arabic, Chinese)
    - SMT workshop corpora: EUROPARL transcriptions (F, ES, D)

Evaluation

- Correlation with human sentence-level judgments
  - 10-fold cross validation
- Baselines:
  - BLEU
  - “TradMetrics” regression model: BLEU, TER, METEOR, NIST

Corpora	BLEU	TradMetrics (regression)	RTE (regression)	TradMetrics + RTE (regression)
NIST	60.0	65.6	63.1	**68.3**
SMT	35.9	39.6	42.3	**45.7**

RTE features and “traditional” metrics are complementary!

We’re getting something right

Ref:	U.S. Treasury Offers $14 billion of 30-Year Treasury Bonds	
Sys:	American treasury posing 14 billion from bonds with maturity 30 years	
Human: 6	RTE: 5.77	BLEU: 3.4

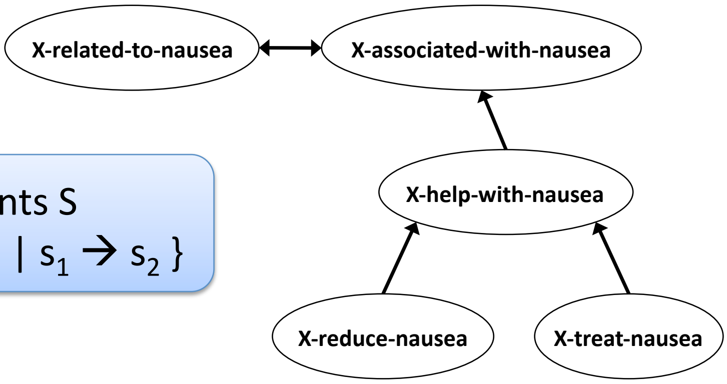
Ref:	What does BBC’s Haroon Rasheed say after a visit to Lal Masjid Jamia Hafsa complex? There are no un- derground tunnels in Lal Masjid or Jamia Hafsa.	
Sys:	BBC Haroon Rasheed Lal Masjid, Jamia Hafsa after his visit to Auob Medical Complex says Lal Masjid and seminary in under a land mine	
Human: 1	RTE: 1.2	METEOR: 4.5

Use Case 2: Entailment Graphs [Berant et al. 2012]

(Entailment for Structuring)

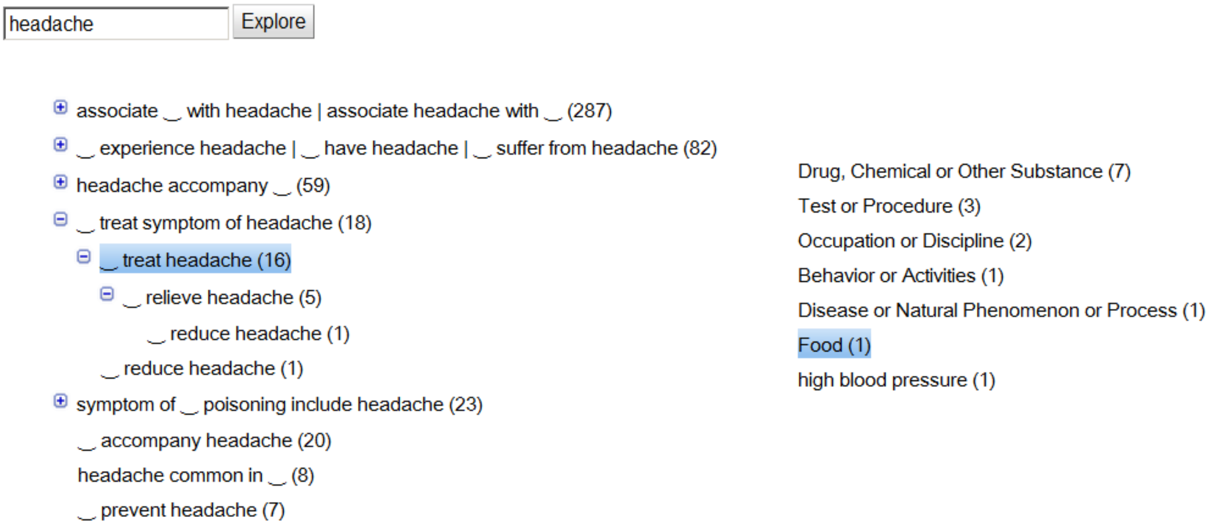
Evaluation: Information Presentation

- Guide users through facts about unfamiliar concept
  - Statements about the target concept collected "Open IE style" [Etzioni et al. 2011]
- Traditional answer: keyword-based presentation
- Proposal: Organize knowledge as **entailment graph**



Input: Set of statements S
Goal: Find E = { (s_1,s_2) | $s_1 \rightarrow s_2$ }

BIU Healthcare Explorer [Adler et al. 2012]



headache Explore

- associate _ with headache | associate headache with _ (287)
- _ experience headache | _ have headache | _ suffer from headache (82)
- headache accompany _ (59)
- _ treat symptom of headache (18)
  - _ treat headache (16)
    - _ relieve headache (5)
      - _ reduce headache (1)
    - _ reduce headache (1)
- symptom of _ poisoning include headache (23)
- _ accompany headache (20)
- headache common in _ (8)
- _ prevent headache (7)

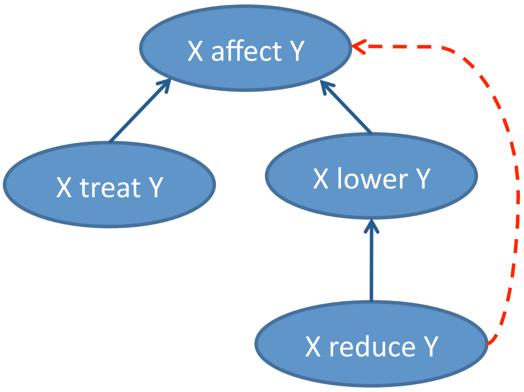
Drug, Chemical or Other Substance (7)
Test or Procedure (3)
Occupation or Discipline (2)
Behavior or Activities (1)
Disease or Natural Phenomenon or Process (1)
Food (1)
high blood pressure (1)

http://irsrv2.cs.biu.ac.il:8080/exploration/

Building Graphs

- Naïve graph construction: Decide entailment for each pair of statements
- Problem: “Local” decisions are not guaranteed to conform to properties of the entailment relation: **transitivity**

X affect Y ⇒ X treat Y	✓
X treat Y ⇒ X affect Y	✗
. . .	
X lower Y ⇒ X affect Y	✓
X reduce Y ⇒ X lower Y	✓
X reduce Y ⇒ X affect Y	✗



Learning Entailment Graphs

- Input: Corpus C
- Output: Entailment graph G = (P,E)
  1. Extract statements S from C
  2. Use a local entailment classifier to estimate $P_{ij} = P(s_i \rightarrow s_j)$ for each pair (s_i, s_j)
     - Techniques from Part 2
  3. **Find the most probable transitive graph**
     - **Part 1: Define objective function for graph**
     - **Part 2: Identify best graph**

Graph Objective Function

$$\hat{G} = \arg\max \sum_{i \neq j} w_{ij} \cdot x_{ij}$$

$$\begin{cases} 1 & i \rightarrow j \\ 0 & \text{else} \end{cases}$$

$$w_{ij} = \log \frac{p_{ij} \cdot \theta}{(1 - p_{ij}) \cdot (1 - \theta)}$$

“density” prior

- Still assumes independence between edges

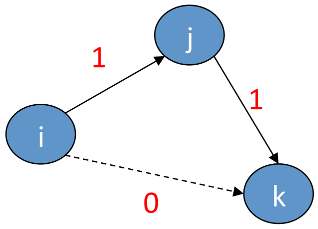
Integer Linear Program

$$\hat{G} = \arg\max \sum_{i \neq j} w_{ij} \cdot \boxed{x_{ij}}$$

$$\forall i, j, k : x_{ij} + x_{jk} - x_{ik} \leq 1$$

$$x_{ij} \in \{0, 1\}$$

1+1-0 = 2 > 1



- NP hard
  - Optimization: Decompose sparse graph
    - Details: [Berant et al. 2012]

Experimental Evaluation

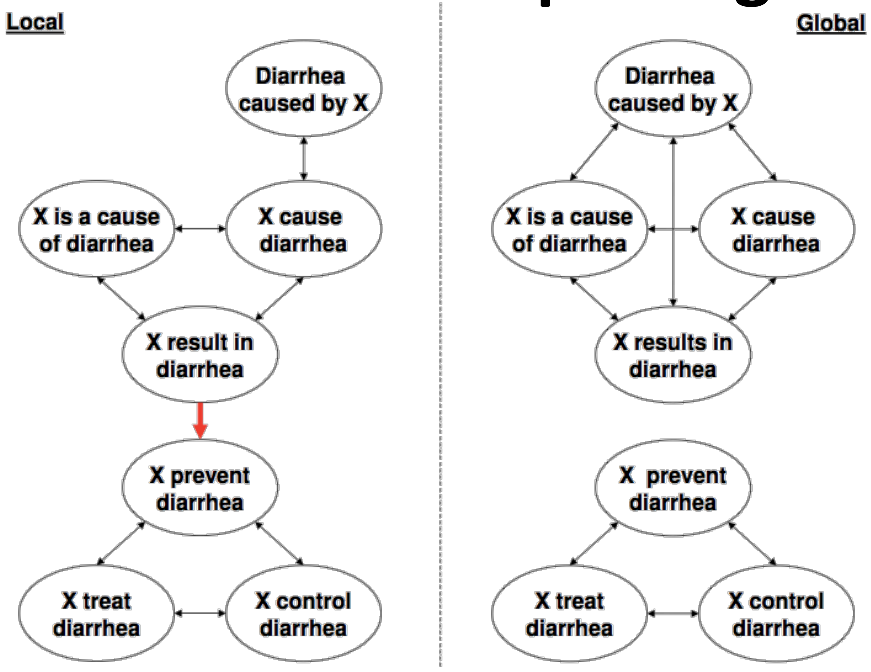
- 50 million word tokens **healthcare** corpus
- Gold standard entailment graphs for 23 medical concepts
  - Smoking, seizure, headache, lungs, diarrhea, chemotherapy, HPV, Salmonella, Asthma, etc.
- Evaluation: F_1 on learned edges vs. gold standard
- Baselines:
  - WordNet as source of entailments between predicates
  - “Local” model without enforcing transitivity

Results

	Recall	Precision	F_1
WordNet	10.8	44.1	13.2
Local	**53.5**	38.0	39.8
Global (ILP)	46.0	**50.1**	**43.8**

- Global algorithm avoids false positives
  - High precision

Illustration – Graph Fragment



Local

Diarrhea caused by X

X is a cause of diarrhea

X cause diarrhea

X result in diarrhea

X prevent diarrhea

X treat diarrhea

X control diarrhea

Global

Diarrhea caused by X

X is a cause of diarrhea

X cause diarrhea

X results in diarrhea

X prevent diarrhea

X treat diarrhea

X control diarrhea

Take-home Message

- Many applications can be mapped (partially) onto Textual Entailment
  – Four paradigms: verify, score, generate, structure
  – Large datasets: Division of labor between shallow methods (generators) and Textual Entailment (filter)
- Two Use Cases:
  – MT Evaluation: TE to measure semantic equivalence
  – Entailment Graphs: Global learning for information presentation

Reference List

- Berant, J., Dagan, I., and Goldberger, J. (2012). Learning entailment relations by global graph structure optimization. Computational Linguistics, 38(1):73–111.
- Callison-Burch, C., Osborne, M., and Koehn, P. P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In Proceedings of EACL, pages 249–256.
- Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam, M. (2011). Open information extraction: the second generation. Proceedings of IJCAI, pages 3–10.
- Harabagiu, S., Hickl, A., and Lacatusu, F. (2007). Satisfying information needs with multi-document summaries. Information Processing and Management, 43(6):1619–1642.

Reference List

- Hickl, A. and Bensley, J. (2007). A Discourse Commitment-Based Framework for Recognizing Textual Entailment. Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 171–176.
- Mirkin, S., Specia, L., Cancedda, N., Dagan, I., Dymetman, M., and Szpektor, I. (2009). Source-Language Entailment Modeling for Translating Unknown Terms. Proceedings of ACL, pages 791–799.
- Nielsen, R. D., Ward, W., and Martin, J. H. (2009). Recognizing Entailment in Intelligent Tutoring Systems. Natural Language Engineering, 15(4):479–501.
- Och, F. J. (2003). Minimum error rate training in statistical machine translation. In Proceedings of ACL, pages 160–167.

Reference List

- Padó, S., Cer, D., Galley, M., Manning, C. D., and Jurafsky, D. (2009). Measuring Machine Translation Quality as Semantic Equivalence: A Metric based on Entailment Features. Machine Translation, 23(2–3):181–193.
- Roth, D., Sammons, M., and Vydiswaran, V. V. (2009). A Framework for Entailed Relation Recognition. Proceedings of ACL, pages 57-60.

Textual Entailment
Part 5: Multilingual, Component-based System Building

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

State of the Art

- What is the state of the TE community in 2013?
  - Almost ten years of research
  - Where do we go from here?

- **Evaluation**: gain insights on what works
- **Sustainable development**: build systems that reflect these insights
- **Application**: make a difference for NLP with TE

State of the Art (cont.)

- In MT, there is a “universal platform”
  - MOSES (Koehn et al., 2007)

- There are two open source systems for TE:
  - EDITS, an alignment-based system
  - BIUTEE, a translation-based system

- So people can download these systems, experiment with them, and use them in applications?
  - In principle yes…
  - …but there are a couple of problems

Problems

- Systems are prototypes of specific algorithms
  - Hard-wired preprocessing tools
  - Hard-wired assumptions about language
  - No modularization of algorithmic parts
  - No interchange format for inference rules

In sum:

Evaluation, development, application are difficult

Are we back at square one?

Summary

- Theoretically
  - Reusability of Algorithms and Resources
  - Framework Generality
- Practically
  - Systematic Evaluation
  - Multilinguality, and Integration in Applications

The EXCITEMENT Project

EXCITEMENT EXploring Customer Interactions through Textual EntailMENT

- EXCITEMENT Open Platform (EOP)
  - – Multilingual
  - – Component-based
  - – Open source
- http://www.excitement-project.eu

The EXCITEMENT Project

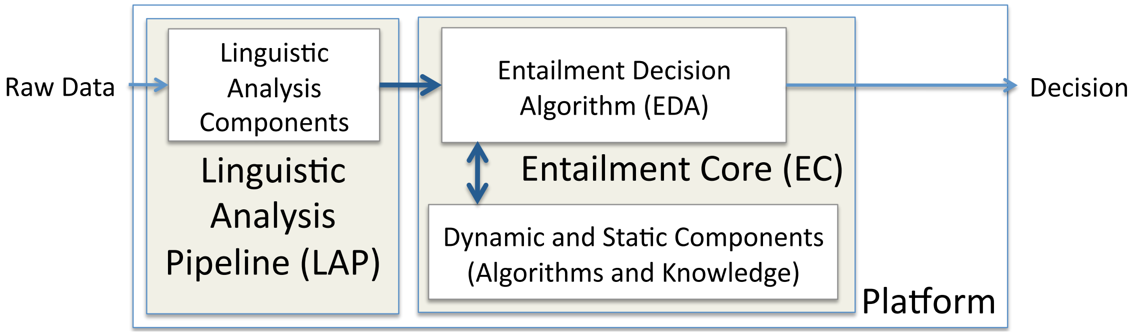
- EU FP 7 Project
  - HEI, DFKI, Bar-Ilan, FBK + industrial partners
- Goal: Provide the necessary infrastructure for sustainable research in Textual Entailment
  - **Specification**: Modular architecture for TE systems
    - Reusability of algorithms, resources through interfaces
    - Towards "plug and play" construction of systems

Complete

  - **Platform**: Implementation of modular specification
    - Working for English, German, Italian

Complete

The EOP Architecture


Raw Data
Linguistic Analysis Components
Linguistic Analysis Pipeline (LAP)
Entailment Decision Algorithm (EDA)
Entailment Core (EC)
Dynamic and Static Components (Algorithms and Knowledge)
Platform
Decision

Specification

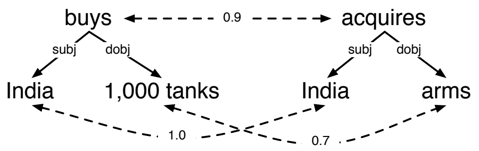
- Linguistic Analysis Pipeline
  - Apache UIMA: linguistic analysis = enrichment of document with strongly typed annotation
  - DKPro type system: language-independent representation of (almost) all linguistic layers
- Entailment Core (Java-based)
  - Interfaces for relevant modules
- Also: “soft” constraints (“best practice” policies)
  - Initialization behavior, error handling, …

Entailment Core

- Top-level interface: Entailment Decision Algorithm
  - Text-Hypothesis pair (UIMA) in, Decision out
  - Existing systems can be wrapped trivially as EDAs
- Three major component types
  - Annotation components
  - Feature components
  - Knowledge components
    - (Don’t cover everything, but 95%)

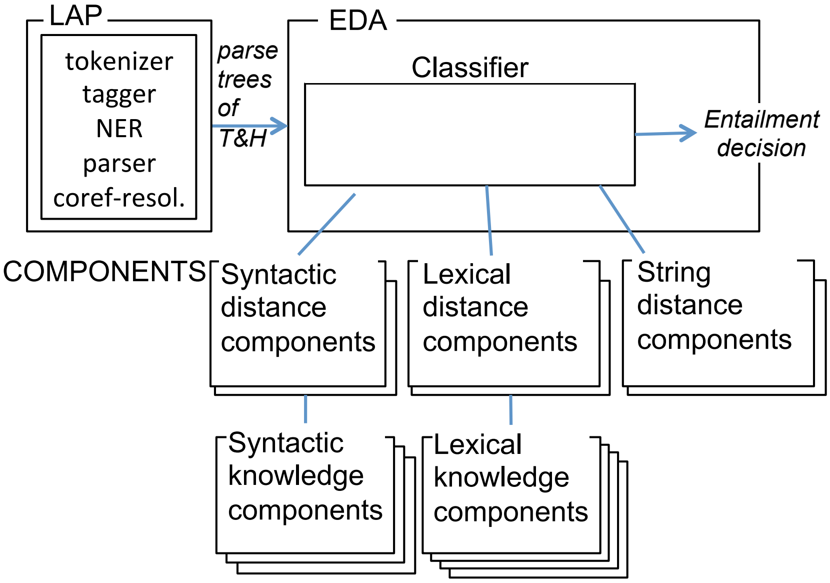
Components

- Annotation components
  - Add linguistic analysis to the P/H pair, e.g. alignment

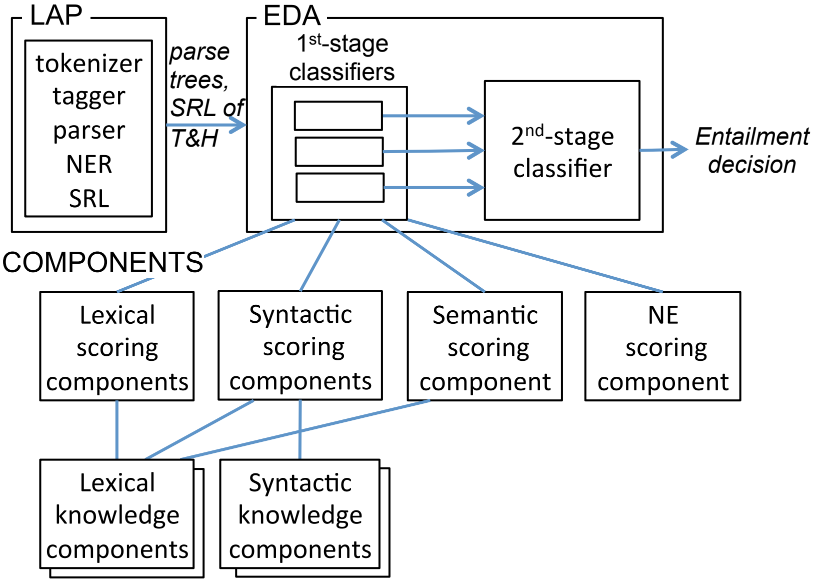


- Feature components
  - Compute match/mismatch features, distance/similarity features, scoring features, …
- Knowledge components
  - Provide access to inference rule bases

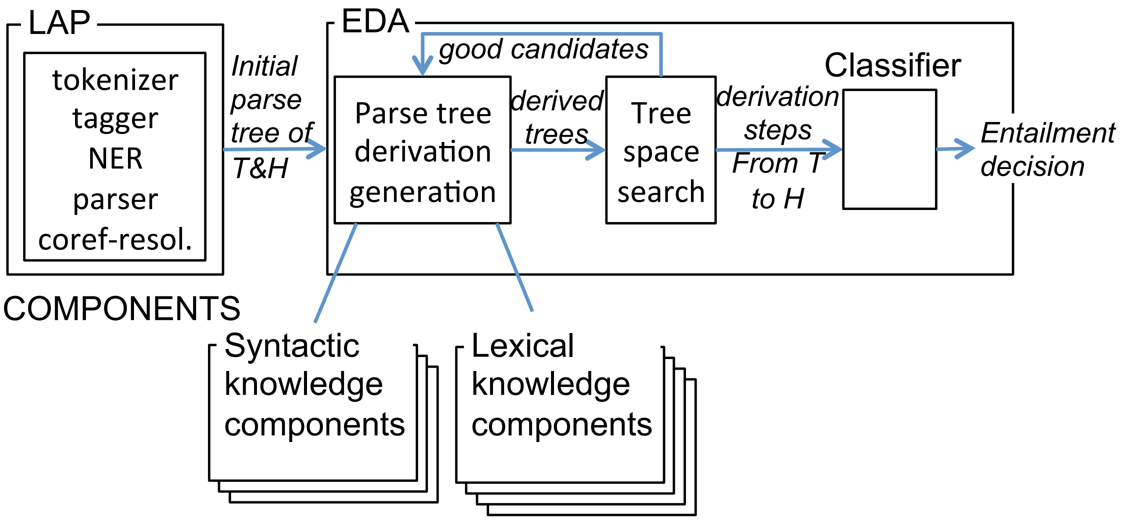
EDITS


LAP
tokenizer
tagger
NER
parser
coref-resol.
parse trees of T&H
EDA
Classifier
Entailment decision
COMPONENTS
Syntactic distance components
Lexical distance components
String distance components
Syntactic knowledge components
Lexical knowledge components

TIE


LAP
tokenizer
tagger
parser
NER
SRL
parse trees, SRL of T&H
EDA
1st-stage classifiers
2nd-stage classifier
Entailment decision
COMPONENTS
Lexical scoring components
Syntactic scoring components
Semantic scoring component
NE scoring component
Lexical knowledge components
Syntactic knowledge components

BIUTEE



LAP

tokenizer
tagger
NER
parser
coref-resol.

Initial parse tree of T&H

EDA

good candidates

Parse tree derivation generation

derived trees

Tree space search

derivation steps From T to H

Classifier

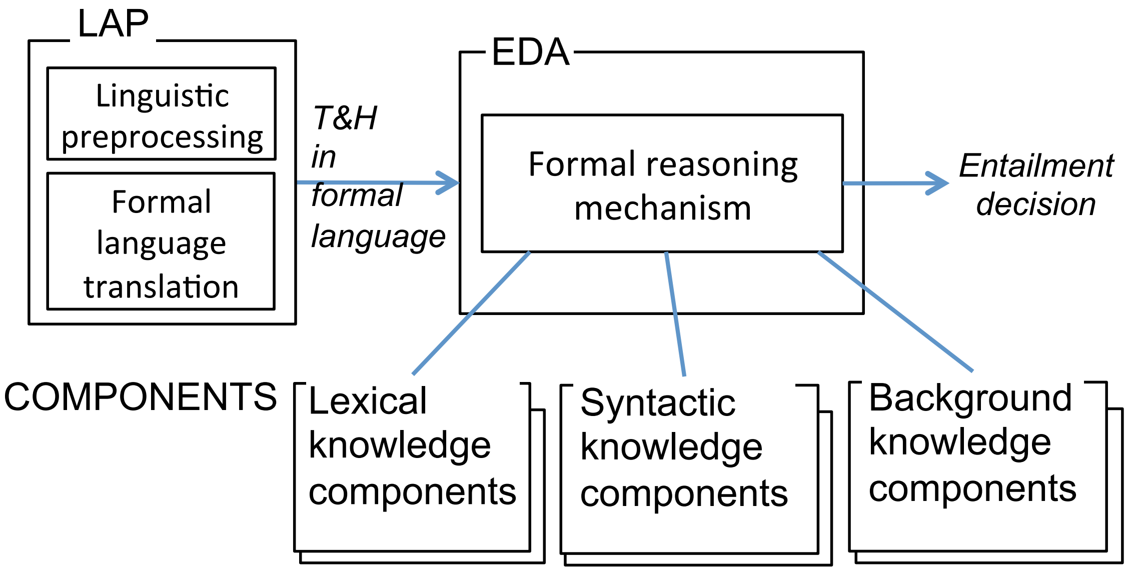
Entailment decision

COMPONENTS

Syntactic knowledge components

Lexical knowledge components

A Formal Reasoning System



LAP

Linguistic preprocessing

Formal language translation

T&H in formal language

EDA

Formal reasoning mechanism

Entailment decision

COMPONENTS

Lexical knowledge components

Syntactic knowledge components

Background knowledge components

Status

- Datasets (Based on RTE-3 data)
  - English, German, Italian, 1600 T-H pairs for each
- LAPs
  - For three languages
- EDAs
  - Three EDAs, EDITS, TIE, and BIUTEE
- Various components
- …and Many knowledge resources

Benefits and further plans

- Reusability
  - Import of BIUTEE's large lexical resources into EDITS for more informed syntactic distance measures
  - Use TIE's semantic role labeller to extend BIUTEE's knowledge resources
  - **"Toolbox" for future experiments**
- Comparable settings for experiments across EDAs
  - constant resources, constant preprocessing, …
- **Platform will be open-sourced**
  - Community of users

System Demo

Subscribe to:

http://hltfbk.github.io/Excitement-Open-Platform/mail-lists.html

Public release on **August 1st!**

Wrap-Up

Structure of the Tutorial

- Part 1 [SP]: Introduction and Basics
- Part 2 [RW]: Classes of Strategies and Learning
- Part 3 [SP]: Knowledge and Knowledge Acquisition
- Part 4 [SP]: Applications
- Part 5 [RW]: Multilingual, Component-based System

De
re
•
•

De
kn

Explore new application scenarios
- General semantic relation between texts

Not Covered in this Tutorial

- Formal reasoning methods
  – Tatu et al. (2006); Bos and Markert (2005); MacCartney and Manning (2007); Clark and Harrison (2009a,b)
- Corpus construction
  – Cooper et al. (1996); Burger and Ferro (2005); Wang and Sporleder (2010); Wang and Callison-Burch (2010)
- Related tasks: Paraphrase acquisition, Semantic textual similarity, etc.
- Crosslinguality: Mehdad et al. (2010)

Further Reference

- Tutorials
  - Dagan et al. ,ACL 2007
  - Sammons et al., NAACL 2010
  - Wang, HIT-MSRA Summer School 2012
    - http://mitlab.hit.edu.cn/2012summerschool/
  - Zanzotto, Web Intelligence 2012
    - http://art.uniroma2.it/zanzotto/teaching/tutorials/rte_at_web_intelligence/
- ACL RTE resource pool
  - http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

Further Reference

- Book
  - Dagan, I., Roth, D., and Zanzotto, F. M. (2012). Recognizing Textual Entailment: Models and Applications. Number 17 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.
- Book chapters & Journal Articles
  - Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. Natural Language Engineering, 15(4).

Further Reference

- Book chapters & Journal Articles
  - Androutsopoulos, I. and Malakasiotis, P. (2010). A Survey of Paraphrasing and Textual Entailment Methods. Artificial Intelligence Research, 38:135–187.
  - M. Sammons, V.G. Vydiswaran, and D. Roth (2012). Recognizing Textual Entailment. In: Multilingual Natural Language Applications: From Theory to Practice.
  - S. Pado & I. Dagan. (to appear). Textual Entailment. Oxford Handbook of Natural Language Processing.

Thank YOU!

Subscribe to:

http://hltfbk.github.io/Excitement-Open-Platform/mail-lists.html

Reference List

- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., and Bojar, O., Constantin, A., and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of ACL.
- Tatu, M., and Moldovan, D. 2007. Cogex at RTE3. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.
- Bos, J., and Markert, K. 2005. Recognising textual entailment with logical inference. In Proceedings of HLT-EMNLP.
- MacCartney, B., and Manning, C. D. 2007. Natural logic for textual inference. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.
- Clark, P., and Harrison, P. 2009. Large-scale extraction and use of knowledge from text. In Proceedings of the fifth international conference on Knowledge capture.

Reference List

- Clark, P., and Harrison, P. 2009. An inference-based approach to recognizing entailment. Proc. of TAC.
- Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. FraCaS Deliverable.
- Burger, J., and Ferro, L. 2005. Generating an entailment corpus from news headlines. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment.
- Wang, R., and Sporleder, C. 2010. Constructing a textual semantic relation corpus using a discourse treebank. In Proceedings of LREC.
- Wang, R., and Callison-Burch, C. 2010. Cheap facts and counter-facts. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.
- Mehdad, Y., Negri, M., and Federico, M. 2010. Towards cross-lingual textual entailment. In HLT-NAACL.