Transformation- and Logic-Based Approaches in RTE

LSA Institute Workshop on Semantics for Textual Inference

Mark Sammons

University of Illinois, Urbana-Champaign

mssammon@illinois.edu

http://cogcomp.cs.illinois.edu

What do we hope to get from RTE?

What kind of solution would be intellectually appealing?

Text: The Cassini spacecraft has taken images that show rivers on Saturn's moon Titan.

Hyp: The Cassini spacecraft has reached Titan.

What do we hope to get from RTE?

What kind of solution would be intellectually appealing?

Text: The Cassini spacecraft has taken images that show rivers on Saturn's moon Titan.

Hyp: The Cassini spacecraft has reached Titan.

The Cassini spacecraft has taken images that show rivers on Saturn's moon Titan.

1. |= The Cassini spacecraft take images of rivers on Saturn's Moon Titan
2. |= The Cassini spacecraft take images of Saturn's moon Titan
3. |= The Cassini spacecraft take images of Titan
4. |= The Cassini spacecraft is at Titan
5. |= The Cassini spacecraft reach Titan

In this presentation…

- Overview of transformation-based/proof-theoretic approaches to RTE
- Analysis of problems with these systems and with the RTE task
- Proposals for moving forward in a direction that supports/encourages development of Natural Language Understanding capabilities
  - Generating “simple” or “specialized” RTE corpora

Outline

- Defining Semantic Entailment
- Transformation-based approaches to RTE
- Logic-based approaches to RTE
- The State of the RTE challenge
- Re(de)fining the RTE task
- Ongoing work: generating “simple” RTE Corpora

Outline

- Defining Semantic Entailment
- Transformation-based approaches to RTE
- Logic-based approaches to RTE
- The State of the RTE challenge
- Re(de)fining the RTE task
- Ongoing work: generating “simple” RTE Corpora

Defining Semantic Entailment

- R - a knowledge representation language, with a well defined syntax and semantics for a domain D.
- For text snippets t, h:
  - r_t, r_h - their representations in R.
  - $M(r_t)$, $M(r_h)$ their model theoretic representations
- There is a well defined notion of subsumption in R, defined model theoretically
- $u, v \in R$: u is subsumed by v when $M(u) \subseteq M(v)$
- Not an algorithm; need a proof theory.

Defining Semantic Entailment (2)

- $r \in R$ is faithful to s if $M(r_t) = M(r)$

Definition: Let t, h, be text snippets with representations $r_t, r_h \in R$.

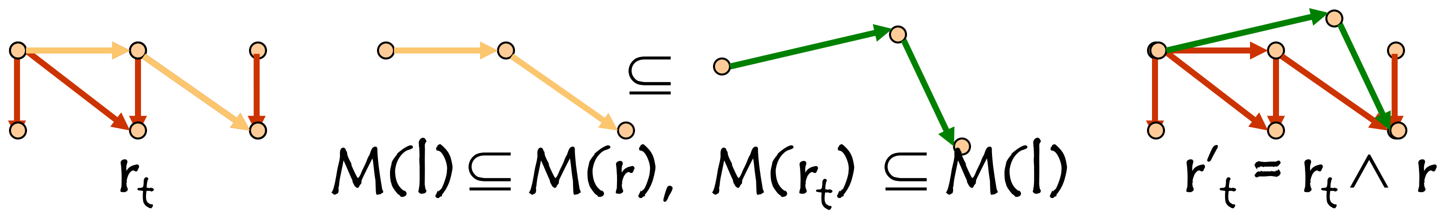
We say that t textually entails h if there is a representation $r \in R$ that is faithful to , for which we can prove that $M(r) \subseteq M(r_t)$

- Given r_t one needs to generate many equivalent representations r'_t and test $M(r'_t) \subseteq M(r_h)$

Cannot be done exhaustively
How to generate alternative representations?

The Role of Knowledge: Refining Representations

- A rewrite rule (l,r) is a pair of expressions in R such that $M(l) \subseteq M(r)$
- Given a representation r_t of t and a rule (l,r) for which $M(r_t) \subseteq M(l)$ the augmentation of r_t via (l,r) is $r'_t = r_t \wedge r$.



Claim: r'_t is faithful to t.

Proof: In general, since $r'_t = r_t \wedge r$ then $M(r'_t) = M(r_t) \cap M(r)$

However, since $M(r_t) \subseteq M(l) \subseteq M(r)$ then $M(r_t) \subseteq M(r)$.

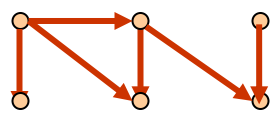
Consequently: $M(r'_t) = M(r_t)$

And the augmented representation is faithful to t.

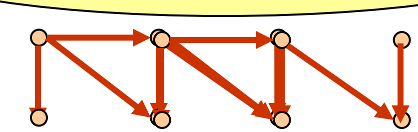
General Strategy

Given a sentence T (answer)

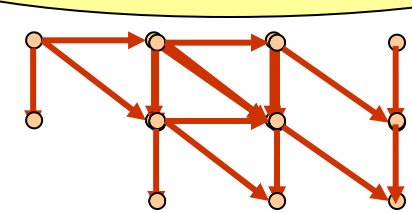
Induce an abstract representation of T (a concept graph)



Re-represent T



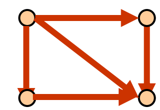
Re-represent T



$\subseteq_e$

Given a sentence H (question)

Induce an abstract representation of H (a concept graph)



Given a KB of semantic; structural and pragmatic **transformations (rules).**

Find the **optimal set of transformations** that maps one sentence to the target sentence.

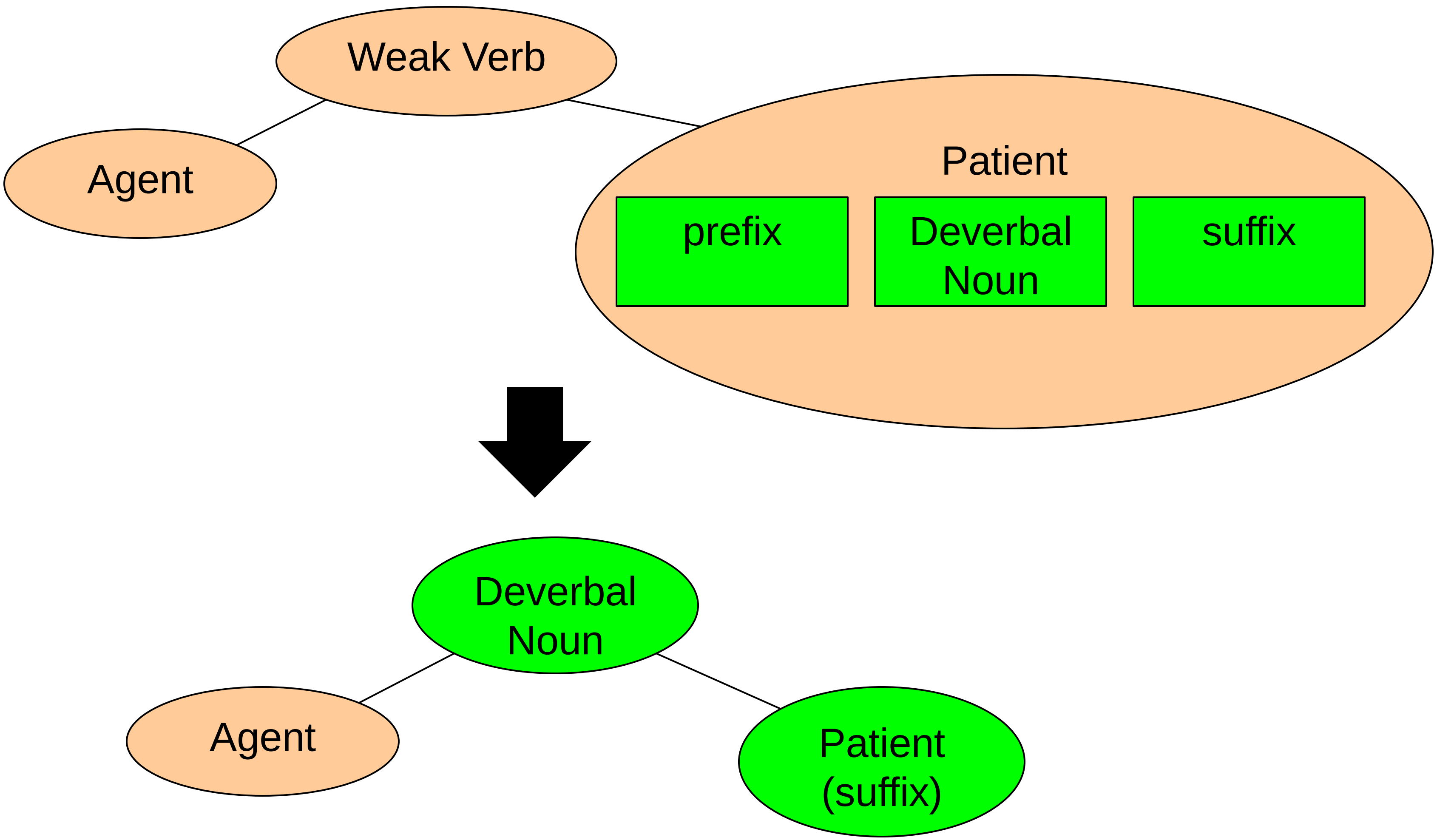
Outline

- Defining Semantic Entailment
- Transformation-based approaches to RTE
- Logic-based approaches to RTE
- The State of the RTE challenge
- Re(de)fining the RTE task
- Ongoing work: generating “simple” RTE Corpora

Transformation-based Approaches: Braz et al. 2005

- SRL, dependency parse and phrases marked in hierarchical representation
- Hand-coded rules based on various levels of representation, incl. lexical
  - Weak Verb rewrite (make, do, begin, ... + nominalized verb)
  - Embedding verb rewrite (fail, manage, want, …)
  - Quantifiers
  - Negation
  - Apposition
  - Conjunction
  - Lexical mappings handled separately ("functional subsumption"), using WordNet

Sample Rule: Weak verb rewrite



Sample Rule: Weak verb rewrite

police | began | an investigation into the robbery

↓

police | investigate | into the robbery

Braz et al. (cont'd)

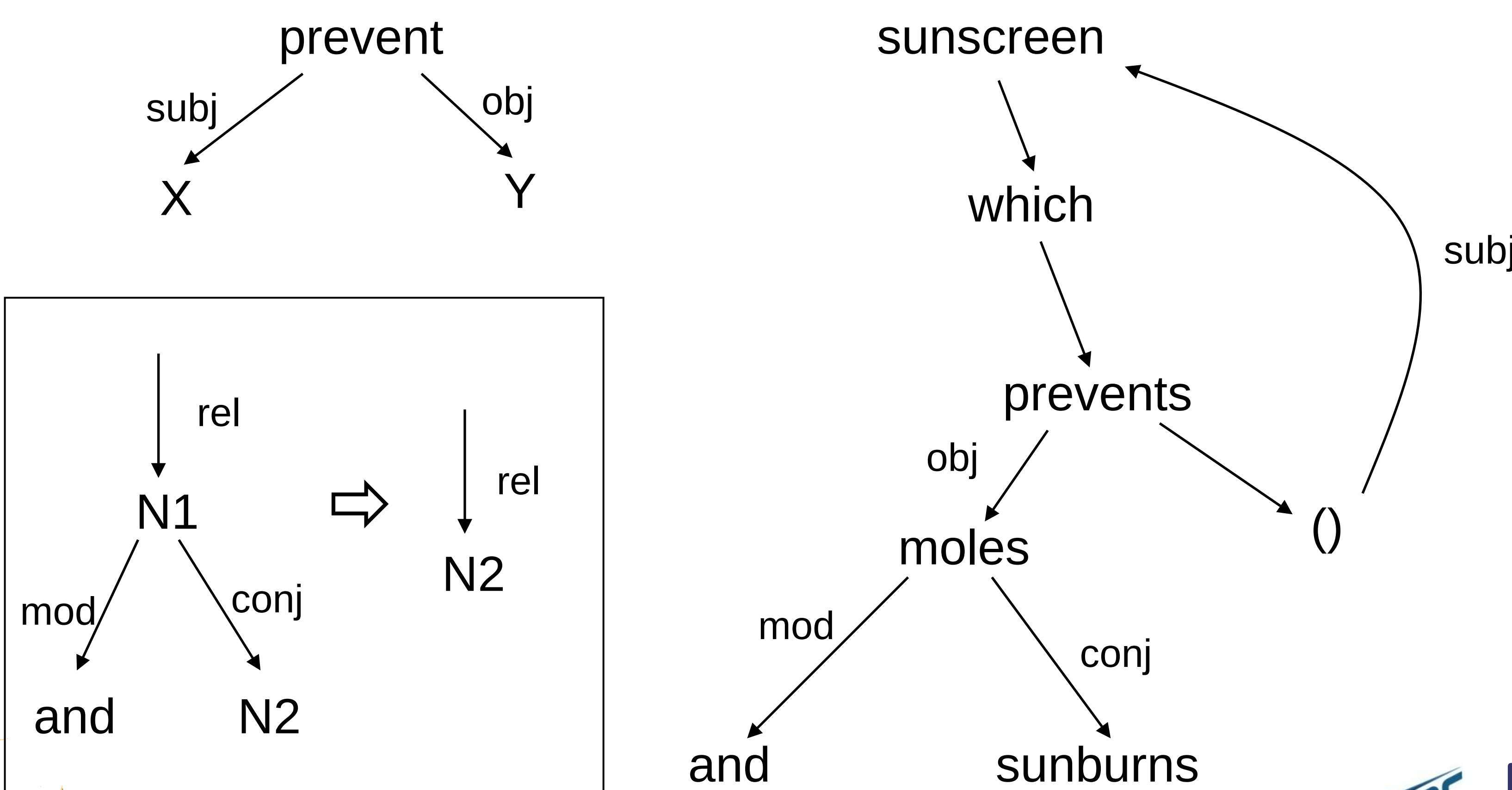
- Abduction-like operator for dropping unmatched terms with some cost
- ILP formulation: find optimal sequence of transformations
- Problems:
  - Knowledge coverage
  - Interpretation errors
  - Some noisy rules (e.g. apposition)
  - Slow inference step

Transformation-based Approaches: Bar-Haim et al. 2007, 2008, …

- Syntactic parse-based representation
- Syntax-based transformations:
  - Passive-Active
  - Conjunctions (simplify to single conjuncts)
  - Determiners ("They sold their house" → "They sold a house")
  - Clausal modifiers ("They watched as the men burned the books." → "the men burned the books.")
  - Relative clauses ("They shot at the car which carried Mr. Smith" → "the car carried Mr. Smith")
  - Genetives ("Mr. Smith's lantern" → "The lantern of Mr. Smith")
- Abstractions:
  - Polarity/negation/modality (mark nodes in tree)

Syntactic Transformation Rules Example: conjunctions

- Sunscreen, which prevents moles and sunburns, ….



Syntax-based Transformation System

- Some success on Information Extraction task
  - Large corpus (multiple representations of same information)
  - Precision-oriented evaluation
- Problems when processing Textual Entailment corpus:
  - Incomplete knowledge → seldom finds a proof
  - Presumably, noise in interpretation results in further errors (missed opportunities, incorrectly applied rules)

Outline

- Defining Semantic Entailment
- Transformation-based approaches to RTE
- Logic-based approaches to RTE
- The State of the RTE challenge
- Re(de)fining the RTE task
- Ongoing work: generating “simple” RTE Corpora

Straightforward Approach: Bos and Markert '06

- Text: *Vincent loves Mia*.
- DRT:

x y
vincent(x)
mia(y)
love(x,y)

- FOL: $\exists x \exists y(\text{vincent}(x)\ \&\ \text{mia}(y)\ \&\ \text{love}(x,y))$
- BK: $\forall x\ (\text{vincent}(x) \rightarrow \text{man}(x))$
  $\forall x\ (\text{mia}(x) \rightarrow \text{woman}(x))$
  $\forall x\ (\text{man}(x) \rightarrow \neg\ \text{woman}(x))$
- Model: D = {d1,d2} F(vincent)={d1}
  F(mia)={d2}
  F(love)={(d1,d2)}

Bos and Markert '06 (cont'd)

- Set of 115 hand-engineered rules representing linguistic and world knowledge PLUS automatically-derived lexical rules from WordNet
- Some (ad-hoc?) modeling of conventional implicature
- Very low coverage of "strict" system: based on '05 report, 0.767 precision and 0.058 recall (f1=0.10)
  - Errors in induced representation affected accuracy even when system had relevant knowledge
  - Knowledge base is inadequate

Theorem Proving with Abduction: Raina et al. 05

- Concept: learn weights for abduction operations
- Induce graphs encoding syntactic dependencies over Text, Hypothesis, map to logical form; enrich with (ad-hoc) semantic annotations for e.g. negation
- Represent as Horn clauses, use Unit Resolution
  - Negate hypothesis and try to derive empty clause
- Add set of abductive operations based on type of constituent being dropped
- Machine-Learning method to optimize weights of abduction operations using RTE development data set, and set of features defined for operations

Raina et al. cont'd:

TEXT: Bob purchased an old convertible.

HYP: Bob bought an old car.

- Dependency parse: e.g.

Bob purchased an old convertible.

- Induce Logical form:

T: $(\exists\ A,B,C)\ Bob(A) \wedge convertible(B) \wedge old(B) \wedge purchased(C,A,B)$

H: $(\forall\ X, Y, Z)\ \neg Bob(X) \vee \neg car(Y) \vee \neg old(Y) \vee \neg bought(Z,X,Y)$

Raina et al. (cont'd)

- Refinements:
  - Abduction operators: match non-identical terms in T, H or drop some term from hypothesis
  - Associate a set of contextual features with operators
- Learn weights for operators: at each step
  - Using current operator weights, derive "best" (min-cost) proof for each example
  - Using set of best proofs, compute weights maximizing likelihood of training data such that positive examples have lower proof costs than negative examples

Comment on Logic-based Approaches

- For the most part, these logical representations are very close to syntactic and shallow semantic parses
- Arguably, simply the same process as transformation-based approaches, but in a different representation
  - Meaning Representation is fundamentally lexical
  - Systems rely on WordNet or similar resources to provide mappings between lexical terms

Outline

- Defining Semantic Entailment
- Transformation-based approaches to RTE
- Logic-based approaches to RTE
- The State of the RTE challenge
- Re(de)fining the RTE task
- Ongoing work: generating “simple” RTE Corpora

Performance of Different Approaches

	accuracy on 2-way RTE task				
	average	**lexical***	**proof/ transform***	**pr/tr plus backoff**	**best**
RTE1	55.1	55.4	< 52.9	57.0	58.6
RTE2	58.5	54.4	51.38	73.5	75.4
RTE3	61.6	62.4	51.5	72.5	80
RTE4	58.0	56.6	51.5	60.5	74.6
RTE5	60.3	57.5	56.7	62.7	73.5

Lexical is a simple lexical baseline based on lexical overlap, allowing stemming

Proof/Transform includes only systems using abstraction of structure

Characteristics of successful systems

- Combined many heterogeneous resources
  - NLP analytics; Relation extraction; similarity measures
- Focused the entailment decision via an alignment step
- Applied machine learning to a small feature set derived from comparison of Text with Hypothesis
- Leveraged augmented training data
- My interpretation: most gains come from being more robust to Interpretation errors, by using global similarity and/or Machine Learning

Observations about Progress in RTE

- Proof-theoretic approaches are outperformed by systems using machine-learning approaches
  - Interpretation noise and knowledge coverage problems are too hard to overcome
  - E.g. Out-of-domain parse tree accuracy: Likely to be ceiling of 80% ParsEval score, which means for any long sentence, there is a very high probability that multiple errors exist
  - Even when machine learning introduced into proof-theoretic approaches, they underperform compared to the best systems
- No standard, Open-Source system
  - Engineering effort is a significant barrier to entry
  - No real re-use of RTE systems/components
  - No agreement on underlying model on which to base such a system

Assessing Component Contributions

- Last two RTE challenges required ablation studies: “leave-one-out” approach to knowledge resources
- For the most part, systems showed limited benefits of most knowledge resources (e.g. VerbOcean, DIRT) from the perspective of system performance on RTE task

Ablated Resource	# of ablation tests	Impact on systems		
		positive	null	negative
Wordnet	19	9 (+1.48%)	3	7 (-0.71%)
VerbOcean	6	2 (+0.25%)	3	1 (-0.16%)
Wikipedia	4 (+1.17%)	3	0	1 (-1.0%)
FrameNet	3 (+1.16%)	1 (+0.16%)	1	1 (-0.17%)
DIRT	3 (+0.75%)	2	0	1 (-1.17%)

Assessing Component Contributions

- But this is not the whole story…
  - Were the systems "closer" to getting some answers right, even when the final answer was wrong?
  - How can we diagnose this behavior?
  - How do we know which components of a system are making a positive contribution?
- If we had a reliable way to assess component contributions, this might encourage specialized module development and use
- If we had enough good components, we ***might*** start to see significant, consistent improvement in RTE results…

Why Assessing Components is Hard

- Noise in interpretation presents significant obstacle
- Intuition: long tail of entailment phenomena
  - Each phenomenon is active in relatively few examples: hard limit on demonstrable improvement based on end-to-end RTE task
  - Most examples require multiple phenomena to be correctly handled: improving performance on one phenomenon will have an even lower global impact
- Hard to show improvement using model for local inference phenomena on RTE corpus
  - No(?) large 'focused' corpora available
  - Some interest in RTE-based evaluation for focused task: e.g. SemEval parsing task 2010

Outline

- Defining Semantic Entailment
- Transformation-based approaches to RTE
- Logic-based approaches to RTE
- The State of the RTE challenge
- Re(de)fining the RTE task
- Ongoing work: generating "simple" RTE Corpora

Possible Focus 1: Robustness against Interpretation Errors

- We have some principled Proof-Theoretic approaches; why not just improve them?
  - Most are strongly dependent on clean Interpretation
  - We could focus on making these work better with state-of-the-art Interpretation, i.e. make them more robust
- Problem: there is a second large deficiency limiting RTE performance: coverage of Knowledge resources
  - We can work on both problems at once, or try to isolate them
  - If the former, progress on one problem alone is likely to have limited impact on system performance

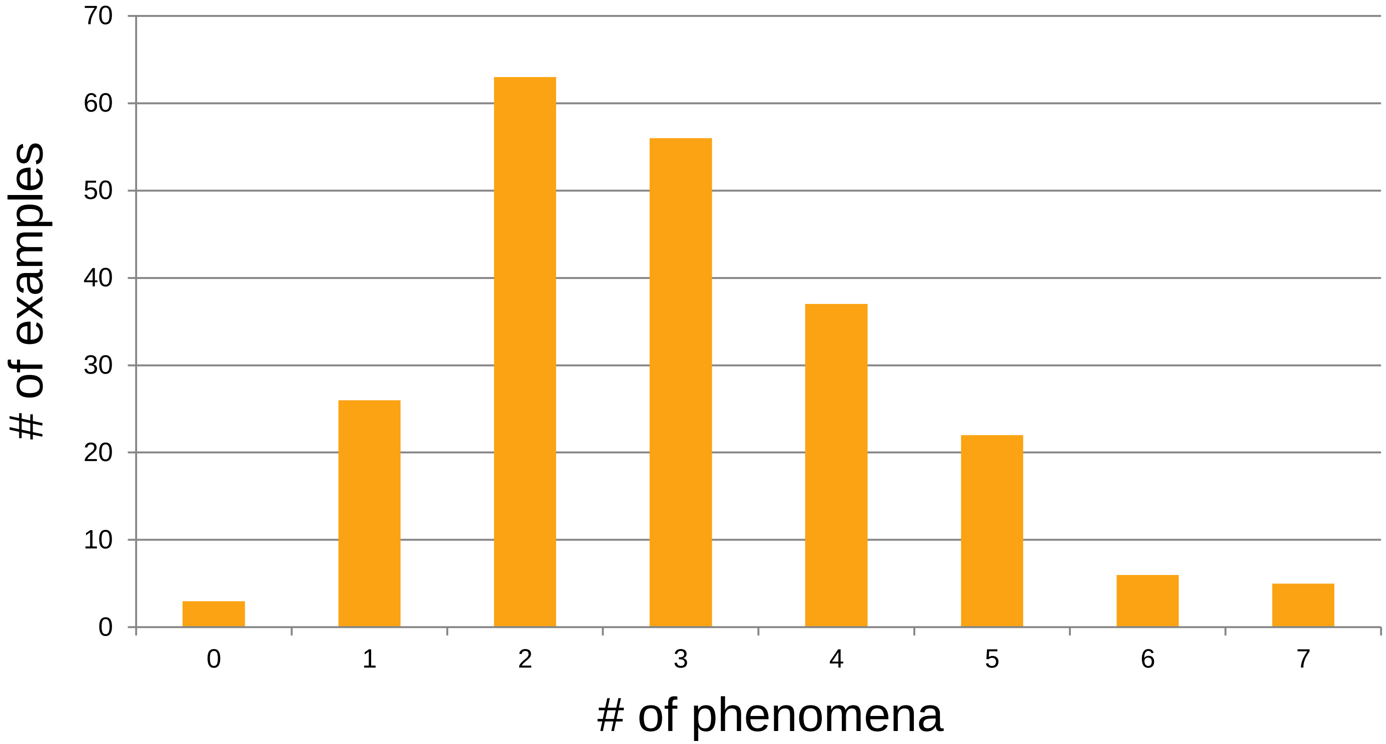
Possible Focus 2: Knowledge (coverage)

- Try to minimize effect of incorrect Interpretation: use simple sentences
- Encourage development of specialized resources
  - Try to isolate domains/entailment phenomena: Generate a sufficient number of examples to…
  - …allow for variability of language: more robust test of RTE systems (more likely to translate into overall performance gain), and of the proposed solution
  - ...allow for statistically significant evaluation of solution (in isolation, and as part of overall system)

Knowledge: What do we need to know?

- Pilot annotation effort (Sammons et al. 2010)
- While there was much anecdotal support for the need for certain types of linguistic & domain knowledge, there were few systematic assessments
- Identify and list phenomena required to prove entailment result for ~200 entailment examples (roughly balanced between positive and negative, and btw. RTE ‘tasks’)
  - Not an easy annotation task – but encouraging initial agreement levels on many phenomena
- Outline a human inference process we hope that annotators can agree on
  - Did not try to order inference steps
  - Allowed for multiple proofs for same example

Number of Phenomena Histogram



- Variance = 2.09; mean = 2.98 (210 RTE examples)
- Undercount – ignores "perfect" interpretation

Entailment Phenomena

Phenomenon	Occurrence	Agreement
coreference	35.00%	0.698
simple rewrite rule	32.62%	0.580
lexical relation	25.00%	0.738
implicit relation	23.33%	0.633
factoid	15.00%	0.412
genetive relation	9.29%	0.608
nominalization	8.33%	0.514
numeric reasoning	4.05%	0.847
spatial reasoning	3.57%	0.720

Negative and Contradiction Phenomena

Phenomenon	Occurrence	Agreement
missing argument	16.19%	0.763
missing relation	14.76%	0.708
excluding argument	10.48%	0.952
Named Entity mismatch	9.29%	0.921
excluding relation	5.00%	0.870
disconnected relation	4.52%	0.580
mismatched modifier	3.81%	0.465
disconnected argument	3.33%	0.764
Numeric Quant. mismatch	3.33%	0.882

T: UberSoft CEO Bill Jobs

H: Frank N. Furter is CEO of Ubersoft

Knowledge Domains (210 examples)

Domain	Occurrence	Agreement
work	16.90%	0.918
name	12.38%	0.833
die kill injure	12.14%	0.979
group	9.52%	0.794
be in	8.57%	0.888
kinship	7.14%	1.000
create	6.19%	1.000
cause	6.19%	0.854
come from	5.48%	0.879
win compete	3.10%	0.813
Others	29.52%	0.864

Pilot annotation effort: conclusions

- Confirms many different entailment phenomena need to be solved in RTE
- Confirms that typically, multiple inference steps are required to determine the entailment label
- Generally, each phenomenon is active in relatively few examples: hard limit on demonstrable improvement based on end-to-end RTE task
- Most examples require multiple phenomena to be correctly handled: improving performance on one phenomenon will have an even lower global impact

Proposals for Change 1: Explanation-based RTE

- Based on Pilot Annotation effort, suggested an RTE pilot task with closed set of inference steps
  - Annotate all operations – possibly with partial ordering – required to solve inference for entailment pair
- Motivation is from an engineering perspective: what problems can we isolate that are solvable, and will have an impact?
- Allows partial credit – for getting ***closer*** to correct answer, both positive and negative
  - Encourage component development and reuse
- Encourage systematic development based on (hopefully) agreeable, human-interpretable inference model

Explanation-based RTE task?

- Benefits:
  - Gain information about types, distribution of phenomena
  - If successful, evaluate the impact of components that successfully target focused inference problems
  - …and encourage reuse of successful components, reduce duplication of effort
- Drawbacks:
  - Significant burden for RTE systems to provide explanations in common format
  - Interpretation errors will interfere with successful application of specialized resources
  - Distribution of phenomena (very long tail) will make it hard to meaningfully evaluate solutions for many sub-problems

Outline

- Defining Semantic Entailment
- Transformation-based approaches to RTE
- Logic-based approaches to RTE
- The State of the RTE challenge
- Re(de)fining the RTE task
- Ongoing work: generating “simple” RTE Corpora

Proposals for Change 2: Simple Entailment Corpora

- Intuition:
  remove Interpretation errors from consideration, to focus on understanding capabilities

- Principal goals:
  - Define "simple" in a useful, practical, and defensible way
  - Develop defensible protocols for generating positive and negative examples exhibiting phenomena of interest
  - Develop defensible methodology for meaningful evaluation of component/system performance on simple corpora

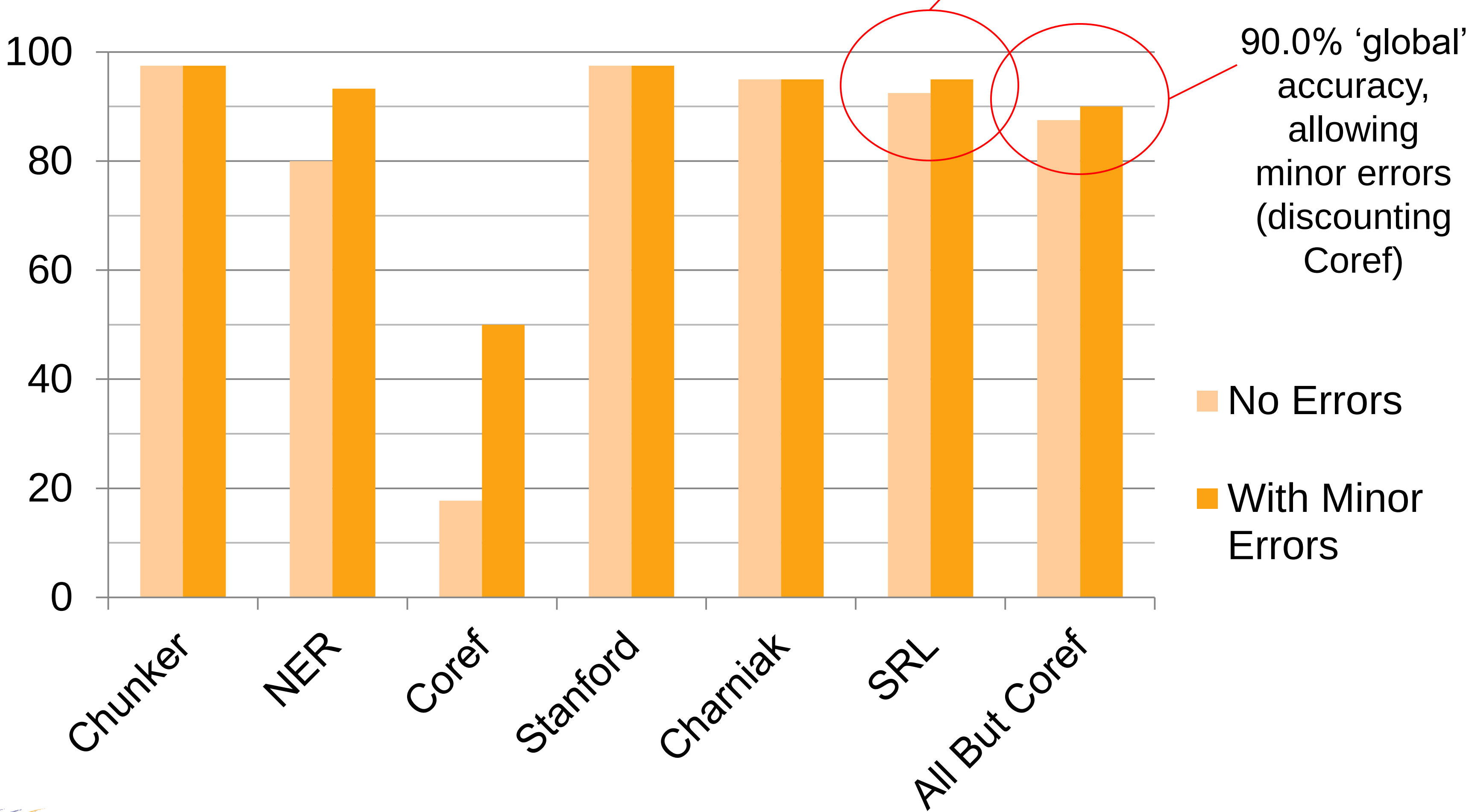
Sanity Check: can we control Interpretation noise?

- Hypothesis:
  if we keep sentences simple, we can get very high performance from NLP tools
- Sub-hypothesis:
  "short" sentences are also "simple" sentences
- Experiment: extracted 80 sentences and assessed performance of suite of NLP tools
  - 40 sentences between 6 and 9 tokens in length
  - 40 sentences between 10 and 15 tokens in length
  - NLP tools: POS, Chunker, Named Entity recognizer, Charniak parser, Stanford parser, Semantic Role Labeler

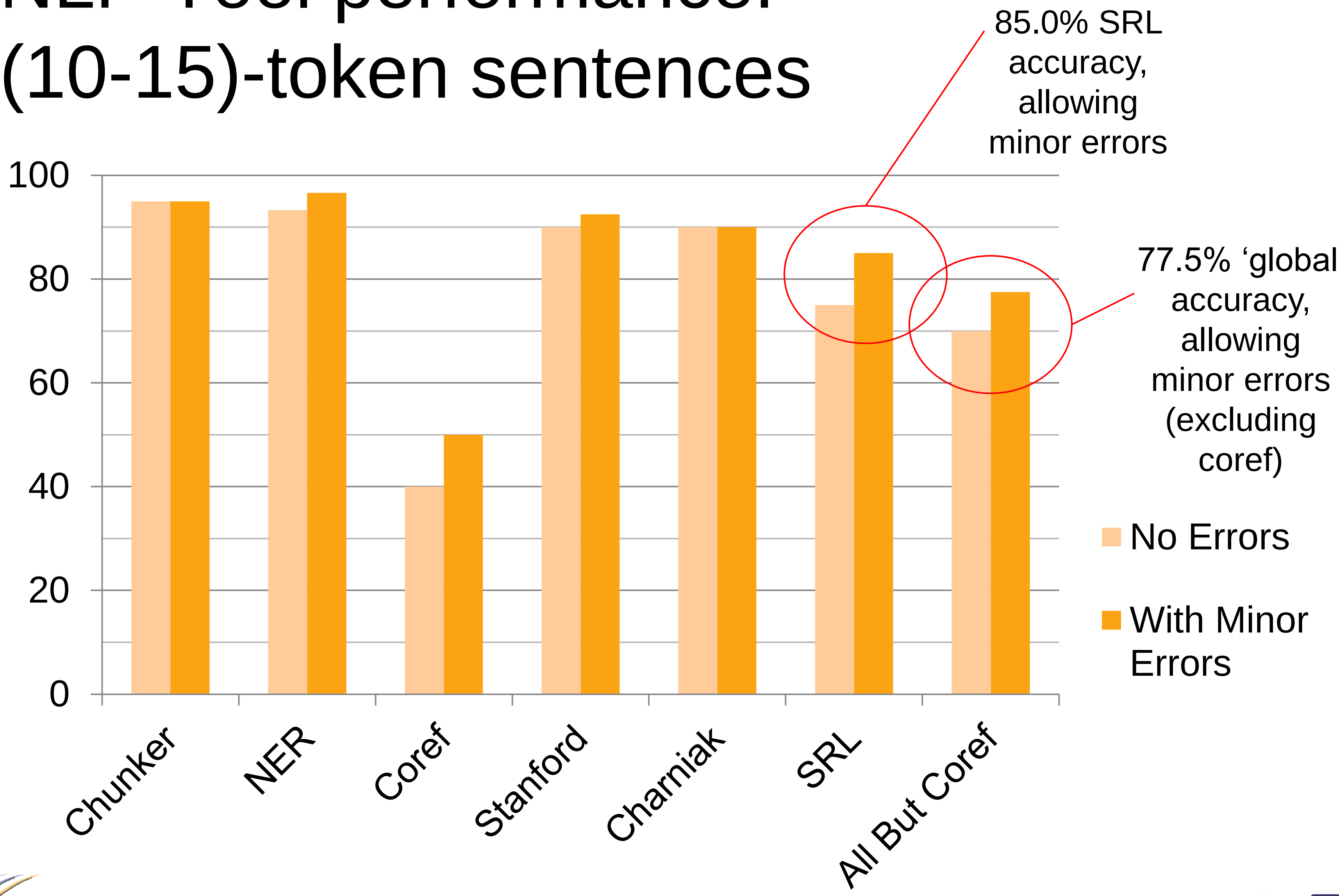
NLP Tool performance: (6-9)-token sentences

92.5% SRL accuracy

90.0% 'global' accuracy, allowing minor errors (discounting Coref)



NLP Tool performance: (10-15)-token sentences



Sample short sentences from NYT Annotated Corpus

Quantifiers:

Meanwhile, some club executives were discussing deals.

Monotonicity and Hypernymy:

I got a stomach virus.

Senator Leahy has snow tires.

Waiters never let a champagne glass get empty.

Name alternation:

So Paul A. Volcker caused all those deficits.

More short sentences…

Implicature:

Representative Wright's proposal recognizes this reality.

He is helping her find an apartment.

Metaphor:

The album is a quiet gem.

Negation:

The couple had no children.

All except Mr. Pendleton performed in the work.

OPEN QUESTIONS: GENERATING FOCUSED RTE CORPORA

Criteria/wish list for Focused Corpora

- "Natural" Texts (i.e., instances from real corpora)
- A large number of examples for each individual phenomenon of interest
- A diverse population of examples that represent a plausible spectrum of natural occurrences of each phenomenon
  - Intuition: this leads to non-trivial solution that is broadly applicable to "natural" text
- A range of example complexities (in terms of number of active phenomena)
  - If we can generate examples that each require a single inference step, that seems like a good place to start
- A balanced corpus, with non-trivial negative examples

Generating Negative Examples

- Negative examples must be sufficiently adversarial to prevent overly general or intuitively irrelevant techniques from achieving good performance
- We are generating a ***focused*** corpus based around ***simple*** sentences…
- … so small differences can often be picked up with trivial features (e.g. lexical overlap).
- We want to probe a deeper level of linguistic performance.
- Trivial features are likely to give an incorrect signal for other types of entailment pairs.

Proposals for Generating Focused Corpora

- Single-inference-step Decomposition
- Custom Design
- Exhaustive Decomposition

Single-Phenomenon Corpora

- LREC Bentivogli et al.: for each entailment pair { T, H }, determine inference steps to determine label
- For each inference step, perturb the Text to generate a new Text_{mod} not requiring that inference step
- Now each such pair { Text, Text_{mod} } is an entailment pair requiring a single inference step, with the label ‘true’ or ‘contradiction’ (can’t easily generate proof for ‘unknown’)

Single-Phenomenon Corpora

T: British writer Doris Lessing, recipient of the 2007 Nobel Prize in Literature, has said in an interview [...]

H: Doris Lessing won the Nobel Prize in Literature in 2007

“Argument Realization”

T’: British writer Doris Lessing, recipient of the Nobel Prize in Literature in 2007, has said in an interview [...]

T entails T’ – a new, monothematic entailment pair.

Single-Phenomenon Corpora

- Benefits: Single-Phenomenon corpora would…
  - …provide a resource for developers of focused inference resources – identify range of contexts, evaluate performance of solution
  - …provide a resource that might help evaluate fine-grained capabilities of complete systems
- Problems:
  - Distribution of phenomena is tied to the original entailment corpus: rarer phenomena likely to be neglected
  - Length and complexity of many examples will result in Interpretation errors
  - Difficulty of designing negative examples that complement the positive examples in each specialized corpus

Custom Design

- Take short sentences extracted from corpus, and perturb them to generate hypotheses
  - Semi-systematic: analyze short sentences for active entailment-related phenomena
  - And/Or: given list of phenomena of interest, perturb sentences to exhibit them
  - Need to specify methods for generating “good” negatives
- Example:

T: Meanwhile, some club executives were discussing deals.

H: Some executives were discussing deals.

H: Some executives discussed deals.

H: Some club executives were ***not*** discussing deals.

Pros and Cons of Custom Design

- Pros:
  - "Natural" Texts, easy to collect
  - Using short sentences, largely eliminates Interpretation as a source of error
  - Some control over phenomena represented
- Cons:
  - Time consuming/expensive to generate entailment pairs
  - Short sentences may under-represent some phenomena more evident in longer sentences
  - Need (a) procedure(s) for generating negative examples

Exhaustive Decomposition

- Select (long) sentences from a corpus, and by hand extract every entailed “atomic” statement:

T: Mr. Smith, 63, who smoked for 22 years, became an advocate for cancer research.

H: Mr. Smith is 63 years old.

H: Mr. Smith advocated cancer research.

H: Mr. Smith used to smoke.

H: Mr. Smith smoked for 22 years.

Exhaustive Decomposition (cont'd)

- Generate plausible negative examples by reorganizing terms/introducing "reasonable" perturbations:

T: Mr. Smith, 63, who smoked for 22 years, became an advocate for cancer research.

H: Mr. Smith smoked for 63 years.

H: Mr. Smith advocated smoking.

H: Mrs. Smith used to smoke.

H: Mr. Smith is a cancer researcher.

Pros and Cons of Exhaustive Decomposition

- Pros:
  - Arguably, provides a good test of understanding
  - “Natural” Texts, easy to collect
- Cons:
  - Time consuming/expensive
  - No control over phenomena extracted:
    rarer phenomena likely to be under-represented
  - Intuition: inter-annotator agreement based purely on extraction will tend to be low – inter-annotator validation probably better
  - Intuition: biased generation of negative examples
    - Individuals tend toward focused set of perturbations
    - Expect reduced set of phenomena that substitute words/phrases or alternate syntactic structure

CONCLUSIONS AND QUESTIONS

Conclusions:

1. It seems practical to focus on short sentences as proxy for "simple" sentences.

2. Short sentences exhibit a variety of structural phenomena identified in our pilot RTE annotation.

3. "Short" means "less than 11 words", if we want (almost) perfect performance from Semantic Role Labeler, and don't want to analyze SRL output on every sentence

Other bases for "simple" corpora

Given likely limitations of "short" as proxy for simple, what are other directions we can pursue?

- Domain based definitions?
  - Find all the ways each one of a set of relations can be instantiated
  - Probably can be restricted to relatively short sentences
- Syntax-based definitions?
  - e.g. comma structures, noun compounds, …?
- Distributional similarity-based definition?

More questions

- Should we worry about rarer phenomena?
  - Can we characterize the “correct” distribution for entailment problems/natural language understanding? Rarity?
- Can we characterize phenomena of interest in a way that will allow us to capture a broad range of instantiations in a corpus?
  - Reduce cost of corpus building
- Is it defensible to design Texts as well as Hypotheses?
  - Some phenomena seem more likely to appear in long sentences – e.g. apposition
  - Perhaps we need to perturb longer sentences exhibiting phenomena of interest to make them simple enough to parse correctly

THANK YOU FOR YOUR ATTENTION

QUESTIONS?

References

- Mark Sammons, V.G.Vinod Vydiswaran and Dan Roth, "Ask not what Textual Entailment can do for you", *ACL* (2010)
- Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio and Bernardo Magnini, "Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference", LREC (2010)