Automatic Extraction of Inference Rules form Text

M.Sc. Seminar: Recent Developments in Computational Semantics

Nikolina Koleva

Saarland University
Department of Computational Linguistics

November 4, 2013

Overview

Inference Rules

- Variants
- Paraphrases
- Entailment Relations

Inference Rules

- Variants
- Paraphrases
- Entailment Relations

Definition

A paraphrase is a restatement of the meaning of a text or passage using other words.

Inference Rules

- Variants
- Paraphrases
- Entailment Relations

Definition

A paraphrase is a restatement of the meaning of a text or passage using other words.

Example

Francis Scott Key **wrote** the “Star Spangled Banner”.
Francis Scott Key **is the author of** the “Star Spangled Banner”.

Inference Rules

- Variants
- Paraphrases
- Entailment Relations

Definition

A paraphrase is a restatement of the meaning of a text or passage using other words.

Example

Francis Scott Key **wrote** the “Star Spangled Banner”.
Francis Scott Key **is the author of** the “Star Spangled Banner”.

X writes Y ≈ X is the author of Y
X writes Y ⇒ X is the author of Y
X writes Y ⇐ X is the author of Y

NLP Applications

- Information Retrieval
- Information Extraction
- Question Answering
- Natural Language Generation
- Machine Translation
- Text Summarization

Collecting Inference Rules

How to collect linguistic templates that convey the same meaning?

1. Manual creation of knowledge base
   - costly (time consuming, experts availability)
   - low coverage

Collecting Inference Rules

How to collect linguistic templates that convey the same meaning?

1. Manual creation of knowledge base
   - costly (time consuming, experts availability)
   - low coverage

2. Automatic acquisition from a text corpus
   - cheeper (text corpus needed)
   - higher coverage

Discovery of Inference Rules from Text (DIRT)

Idea

Detect similar paths in dependency trees.

Discovery of Inference Rules from Text (DIRT)

Idea

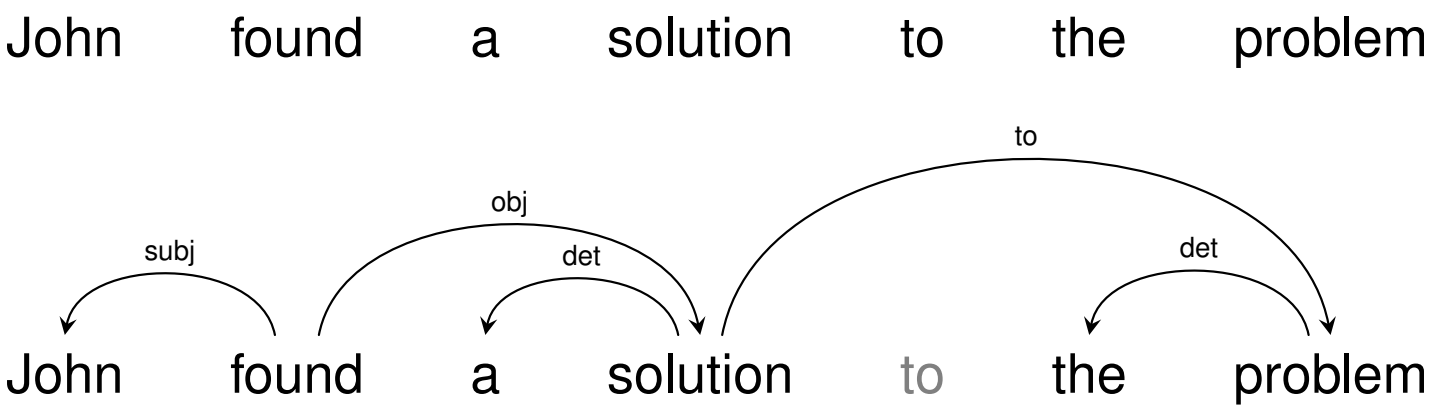
Detect similar paths in dependency trees.

John found a solution to the problem

Discovery of Inference Rules from Text (DIRT)

Idea

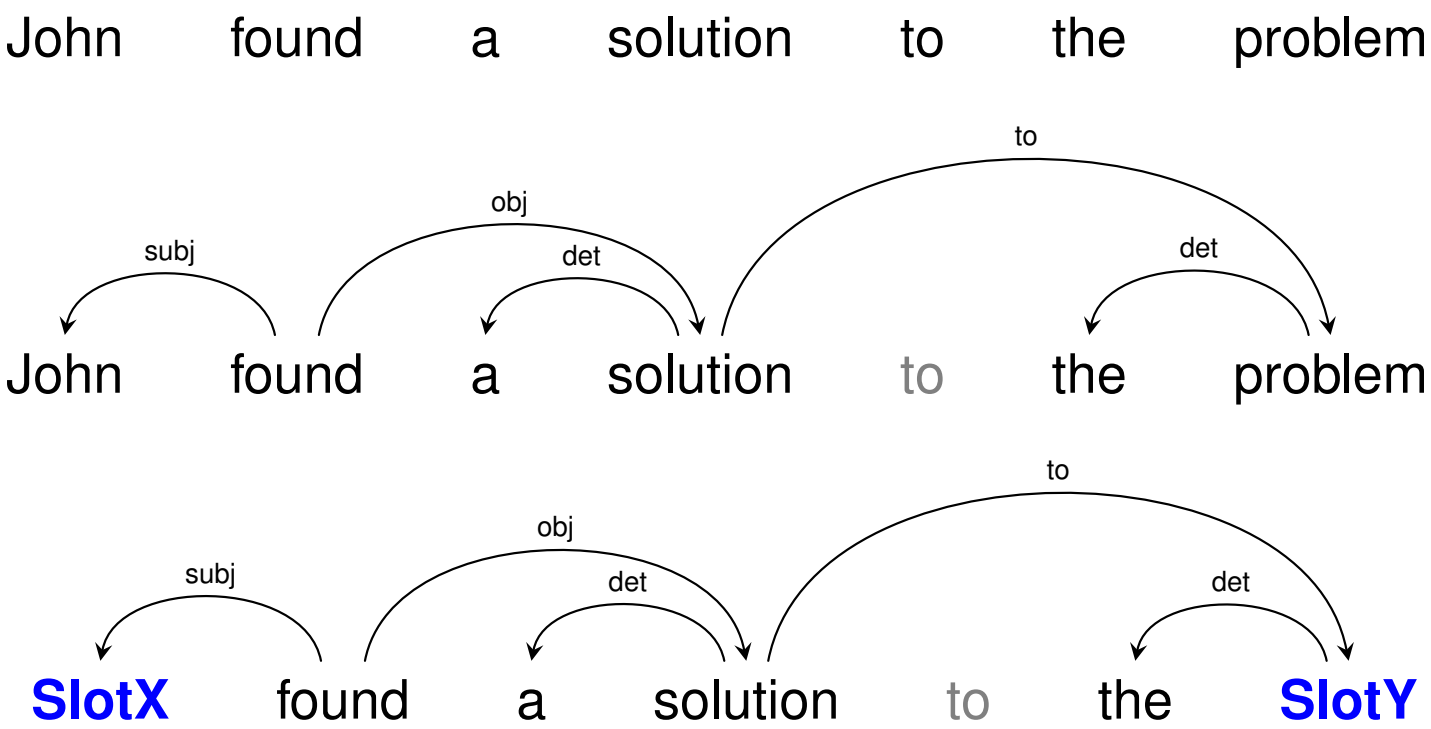
Detect similar paths in dependency trees.



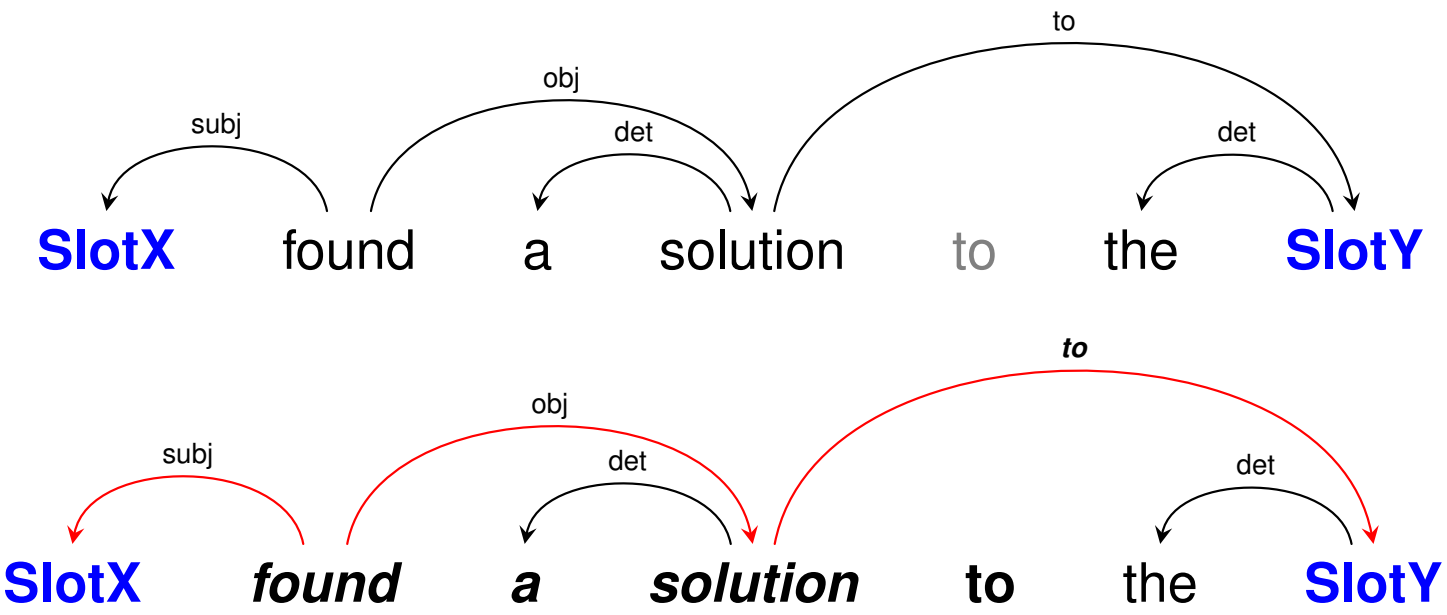
Discovery of Inference Rules from Text (DIRT)

Idea

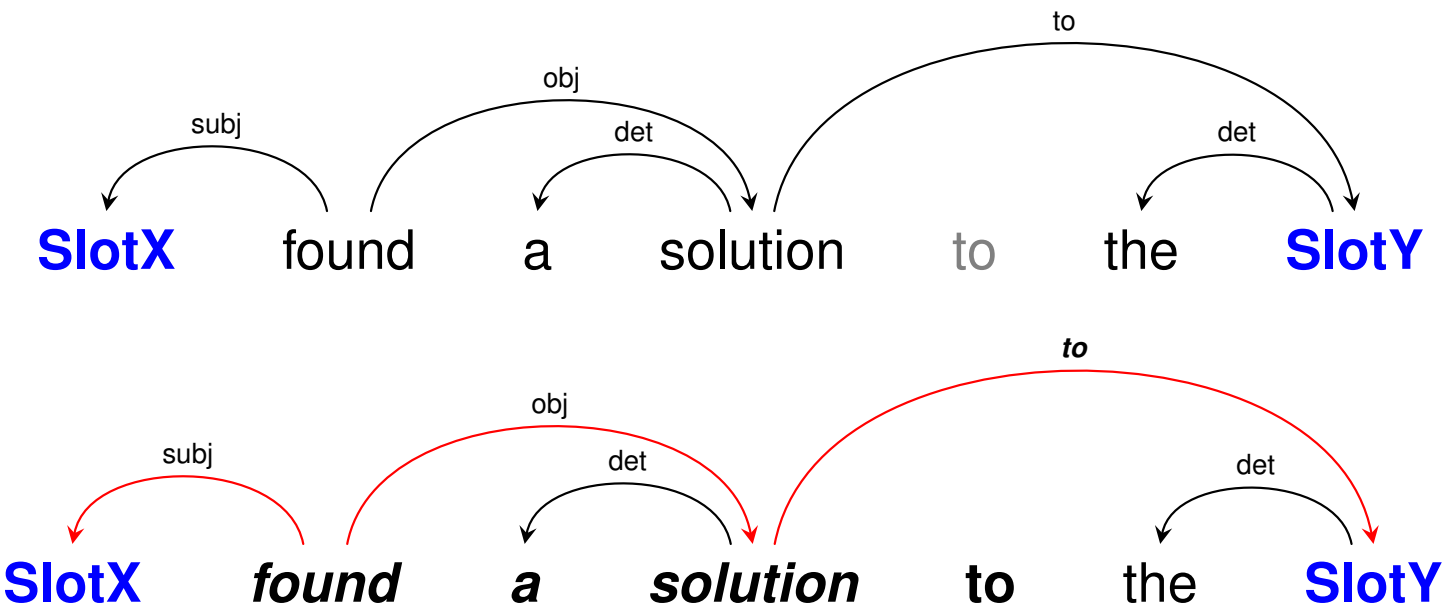
Detect similar paths in dependency trees.



Discovery of Inference Rules from Text (DIRT)



Discovery of Inference Rules from Text (DIRT)



- a path represents (indirect) ***semantic relation*** between two words
- slots fillers must be nouns
- relations containing stop words are excluded from a path
- ***internal relations*** are not connected to the slots
  → must exceed a frequency threshold

Hypothesis

Distributional Hypothesis

Words that occur in the same contexts tend to have similar meanings.

Hypothesis

Distributional Hypothesis

Words that occur in the same contexts tend to have similar meanings.

Extended Distributional Hypothesis

If two paths tend to occur in similar **contexts**, the meanings of the paths tend to be similar.

☛ Paths similarity depends on the similarity of the **sets of slot fillers**

Paths as Triples

- Count frequencies of all dependency paths and slot fillers in the corpus
- If path p connects the words w_1 and w_2 → increment counts of the triples $(p, SlotX, w_1)$ and $(p, SlotY, w_2)$
- $(SlotX, w_1)$ and $(SlotY, w_2)$ are features of p

Paths as Triples

- Count frequencies of all dependency paths and slot fillers in the corpus
- If path p connects the words w_1 and w_2 → increment counts of the triples $(p, SlotX, w_1)$ and $(p, SlotY, w_2)$
- $(SlotX, w_1)$ and $(SlotY, w_2)$ are features of p

☛ The more common features two paths have, the more similar they are

Paths as Triples

- Count frequencies of all dependency paths and slot fillers in the corpus
- If path p connects the words w_1 and w_2 → increment counts of the triples $(p, SlotX, w_1)$ and $(p, SlotY, w_2)$
- $(SlotX, w_1)$ and $(SlotY, w_2)$ are features of p

☛ The more common features two paths have, the more similar they are

path p : X finds solution to Y		
Slot	*Slot Filler*	*Count*
SlotX	government	4
	he	7
	...	...
SlotY	problem	4
	crisis	2
	...	...

Mutual Information (MI) between Path, Slot and Slot Filler

- compute MI between all pairs of paths and features
- MI measures the association strength between a slot and a filler

$$mi(p,s,w) = log\left(\frac{P(p,s,w)}{P(s)P(p|s)P(w|s)}\right)$$

where p: path; s: slot; w: word /slot filler

Mutual Information (MI) between Path, Slot and Slot Filler

$|p, s, w|$ = frequency count of the triple (p, s, w)

$|p, s, *| = \sum_{w} |p, s, w|$

$|*, *, *| = \sum_{p,s,w} |p, s, w|$ (total number of triples)

$$mi(p, s, w) = log\left(\frac{P(p, s, w)}{P(s)P(p|s)P(w|s)}\right)$$

Mutual Information (MI) between Path, Slot and Slot Filler

$|p, s, w|$ = frequency count of the triple (p, s, w)

$|p, s, *| = \sum_{w} |p, s, w|$

$|*, *, *| = \sum_{p,s,w} |p, s, w|$ (total number of triples)

$$mi(p, s, w) = log\left(\frac{P(p, s, w)}{P(s)P(p|s)P(w|s)}\right)$$

$$= log\left(\frac{\frac{|p,s,w|}{|*,*,*|}}{\frac{|*,s,*|}{|*,*,*|}\frac{|p,s,*|}{|*,s,*|}\frac{|*,s,w|}{|*,s,*|}}\right)$$

$$= log\left(\frac{|p, s, w| \times |*, s, *|}{|p, s, *| \times |*, s, w|}\right)$$

Mutual Information (MI) between Path, Slot and Slot Filler

$|p, s, w|$ = frequency count of the triple (p, s, w)

$|p, s, *| = \sum_{w} |p, s, w|$

$|*, *, *| = \sum_{p,s,w} |p, s, w|$ (total number of triples)

$$mi(p, s, w) = log\left(\frac{|p, s, w| \times |*, s, *|}{|p, s, *| \times |*, s, w|}\right)$$

path p : X finds solution to Y			
Slot	*Slot Filler*	*Count*	*MI*
SlotX	government	4	3.45
	he	7	1.32
	...	...	...
SlotY	problem	4	3.09
	crisis	2	1.75
	...	...	...

Similarity between a Pair of Slots

$$sim(s_1, s_2) = \frac{\sum_{w \in T(p_1,s) \cap w \in T(p_2,s)} mi(p_1, s, w) + mi(p_2, s, w)}{\sum_{w \in T(p_1,s)} mi(p_1, s, w) + \sum_{w \in T(p_2,s)} mi(p_2, s, w)}$$

$s_1 = (p_1, s)$; $s_2 = (p_2, s)$;

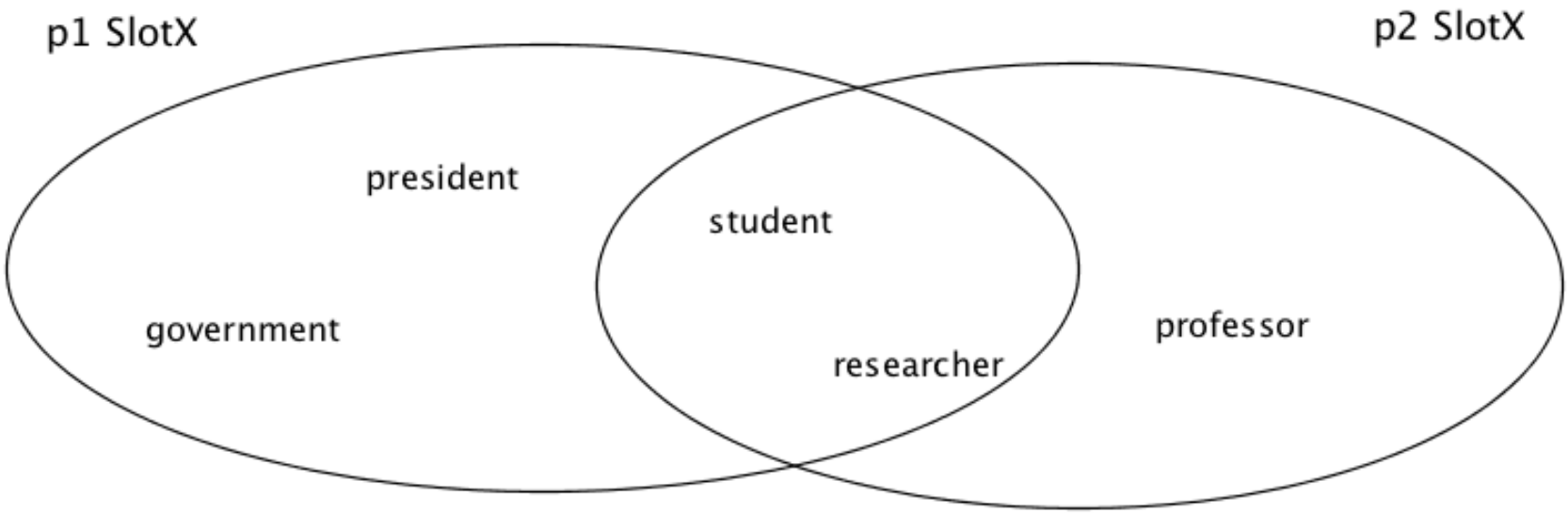
$T(p_i, s)$: set of words that fill in slot s in path p

- sum MI values of common slot fillers
- normalize by the sum of all MI values for both slots

Similarity between a Pair of Slots

p_1: X finds solution to Y

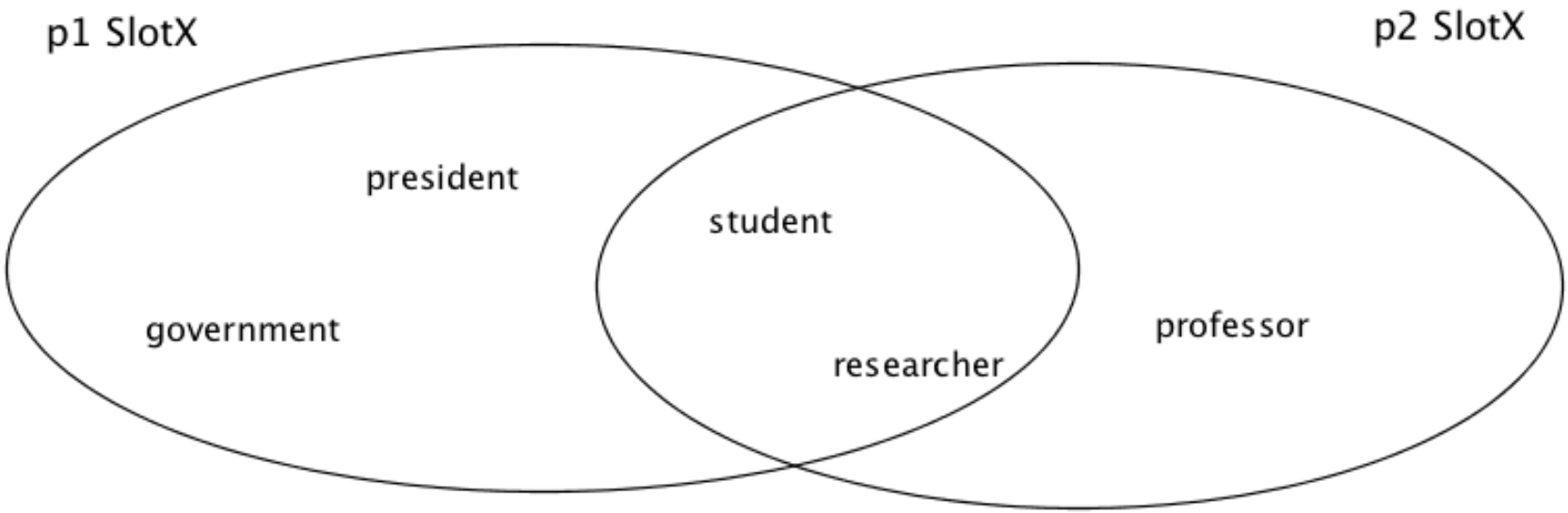
p_2: X solves Y



Similarity between a Pair of Slots

p_1: X finds solution to Y

p_2: X solves Y



$$sim(SlotX_{p_1}, SlotX_{p_2}) =$$

$$\frac{mi(p_1, SlotX, researcher) + mi(p_1, SlotX, student) + mi(p_2, SlotX, researcher) + mi(p_2, SlotX, student)}{mi(p_1, SlotX, researcher) + mi(p_1, SlotX, student) + mi(p_1, SlotX, president) + mi(p_1, SlotX, government) + mi(p_2, SlotX, researcher) + mi(p_2, SlotX, student) + mi(p_2, SlotX, professor)}$$

Similarity between a Pair of Paths

Defined as:

The geometric average of SlotX and SlotY similarities of the two paths.

$$sim(p_1, p_2) = \sqrt{sim(SlotX_{p_1}, SlotX_{p_2}) \times sim(SlotY_{p_1}, SlotY_{p_2})}$$

Efficient Detection of Similar Patterns

- Large number of paths in the database
- For a path p
  1. fetch all paths that share at least one slot filler value with p, candidate paths C
  2. count shared features of $c \in C$ and p and drop c if only they are too few
  3. compute the similarity between p and c, rank candidates according to their similarity

Efficient Detection of Similar Patterns

- Large number of paths in the database
- For a path *p*
  1. fetch all paths that share at least one slot filler value with *p*, candidate paths *C*
  2. count shared features of $c \in C$ and *p* and drop *c* if only they are too few
  3. compute the similarity between *p* and *c*, rank candidates according to their similarity

X solves Y
Y is solved by X
X resolves Y
X finds solution Y
...

Evaluation

- Apply DIRT on 1GB newspaper text (7 million paths)
  `input`: patterns of the first six questions from the TREC-8[1]
- Comparison with paraphrases created by humans
- Manual inspection of the top 40 DIRT outputs for each question
  correct or incorrect

[1] Text REtrieval Conference

Questions

Table : **First six questions from TREC-8**

id	Questions
Q1	Who is the author of the book, “The Iron Lady: A Biography of Margaret Thatcher"?
Q2	What is the monetary value of the Nobel Peace Prize in 1989?
Q3	What does the Peugeot company manufacture?
Q4	How much did Mercury spend on advertising in 1993?
Q5	What is the name of the managing director of Apricot Computer?
Q6	Why did David Koresh ask the FBI for a word processor?

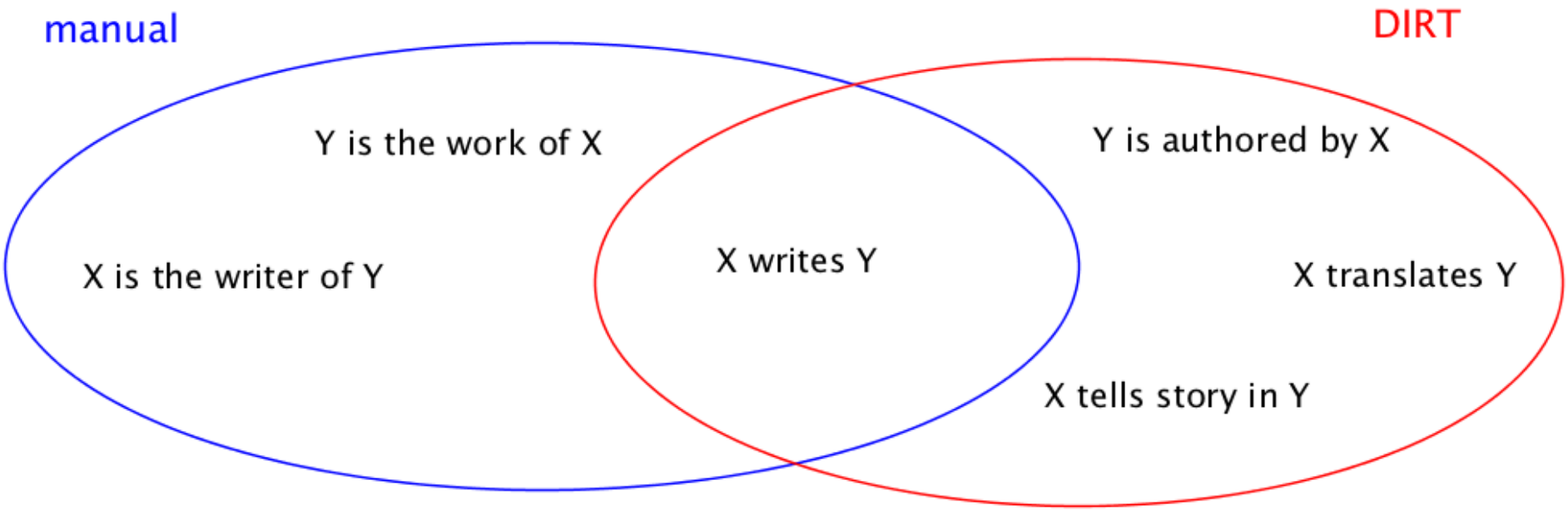
Experimental Results

Table : **Evaluation of the top 40 similar paths**

	input paths	**manual**	**DIRT**	∩	**Acc. %**
Q1	X is author of Y	7	21	2	52.5
Q2	X is monetary value of Y	6	0	0	N/A
Q3	X manufactures Y	13	37	4	92.5
Q4	X spend Y	7	16	2	40.0
	spend X on Y	8	15	3	37.5
Q5	X is managing director of Y	5	14	1	35.0
Q6	X asks Y	2	23	0	57.5
	asks X for Y	2	14	0	35.0
	X asks for Y	3	21	3	52.5

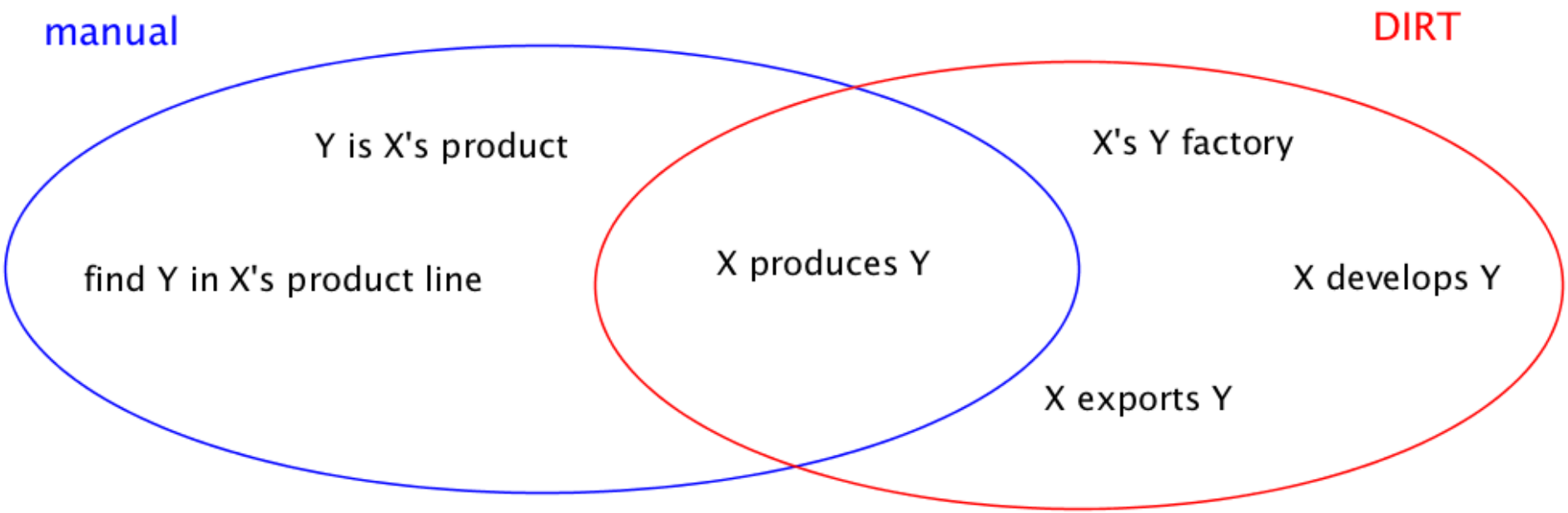
Manual vs. DIRT

Q1 path: X is the author of Y



Manual vs. DIRT

Q3 path: X manufactures Y



Observations

- little overlap between manually created paraphrases and automatically generated inference rules

☛ finding useful paraphrases/inference rules is a difficult task in general

- DIRT can be used as an assisting tool for building a knowledge base

LEearning Directionality of Inference Rules (LEDIR)

Inference rules are underspecified in directionality

DIRT: X eats Y ≈ X likes Y

but

X eats Y ⇒ X likes Y

X likes Y ⇏ X eats Y

LEearning Directionality of Inference Rules (LEDIR)

Inference rules are underspecified in directionality

DIRT: X eats Y ≈ X likes Y

but

X eats Y ⇒ X likes Y

X likes Y ⇏ X eats Y

Example

Bill eats apples. ⇒ Bill likes apples.

Bill likes Mary. ⇏ Bill eats Mary.

Hypothesis

Directionality Hypothesis

If two binary semantic relations tend to occur in similar contexts and the first one occurs in significantly more contexts than the second, then the second most likely implies the first and not vice versa.

Example

One can like many more things than one can eat.
Thus: X eats Y ⇒ X likes Y

LEDIR steps

1. Modelling contexts of paths p_i and p_j as relational selectional preferences (RSPs)
2. Classify a rule as plausible or not
   by checking the number of shared RSPs
3. If plausible then determine directionality

Contexts of a relation (RSPs)

For a binary relation (x, p, y)
$C(x)$ and $C(y)$ are the sets of semantic classes of the words occurring in position x and y
☛ obtained by clustering or WordNet

Contexts of a relation (RSPs)

For a binary relation (x, p, y)
$C(x)$ and $C(y)$ are the sets of semantic classes of the words occurring in position x and y
☛ obtained by clustering or WordNet

Example

X likes Y
$C(x)$ = { individual, social group, ...}
$C(y)$ = { individual, food, sport, ...}

Joint vs. Independent **Relational Model** ((JRM) vs. (IRM))

1. Get all realizations of a relation p
2. For each (x, p, y)
   - get the sets $C(x)$ and $C(y)$of the semantic classes for x and y
   - **JRM**: considers each $(c(x), p, c(y))$ as candidate RSP, assuming that all $c(x)$ and $c(y)$ can co-occur
   - **IRM**: $(c(x), p, *)$ and $(*, p, c(y))$ are independent candidate RSPs
3. Rank candidates
   - **JRM**: Pointwise mutual information
   - **IRM**: Maximum likelihood estimates for $P(c(x)|p)$ and $P(c(y)|p)$

Inference plausibility and directionality model

$$sim(p_i, p_j) = \frac{|(C(x), p_i, C(y)) \cap (C(x), p_j, C(y))|}{min(|(C(x), p_i, C(y))|, |(C(x), p_j, C(y))|)}$$

Inference plausibility and directionality model

$$sim(p_i, p_j) = \frac{|(C(x), p_i, C(y)) \cap (C(x), p_j, C(y))|}{min(|(C(x), p_i, C(y))|, |(C(x), p_j, C(y))|)}$$

If $sim(p_i, p_j) < \alpha$

the inference rule is not plausible

else

the inference rule is plausible

If $\frac{|(C(x),p_i,C(y))|}{|(C(x),p_j,C(y))|} \geq \beta$ conclude $p_i \Leftarrow p_j$

else if $\frac{|(C(x),p_i,C(y))|}{|(C(x),p_j,C(y))|} \geq \frac{1}{\beta}$ conclude $p_i \Rightarrow p_j$

else conclude $p_i \Leftrightarrow p_j$

α and β (empirically gained threshold values)

LEDIR Evaluaton

- DIRT inference rules as resource
- Train models on newswire containing 31 million words
- Two sets of semantic classes
  - 1628 by clustering words
  - 1287 form WordNet
- Gold Standard creation
  - annotations for 57 DIRT rules
  - labels: $\Leftrightarrow$, $\Leftarrow$, $\Rightarrow$ or *NO*
- Best obtained accuracy ☛ 48 % IRM + sem classes by clustering

Summary

1. Extended Distributional Hypothesis
2. Dependency paths and their storage
3. Similarity of paths using MI
4. Evaluation of DIRT
5. Limitations of DIRT
6. Learning directionality (LEDIR)

Thank you for your attention!

Any questions?

Discussion

1. How to discard antonyms?
   X loves Y ⇔ X hates Y
2. Paraphrases can have different syntactic categories:
   her preference = what she prefers

References

Rahul Bhagat, Patrick Pantel, and Eduard Hovy.

Ledir: An unsupervised algorithm for learning directionality of inference rules.

In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 161–170, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Dekang Lin.

Extracting collocations from text corpora.

In *Proc. of 1st Workshop on Computational Terminology, Computerm '98*, pages 57–63, Montreal, Canada, 1998. COLING-ACL.

Dekang Lin and Patrick Pantel.

DIRT - Discovery of Inference Rules from Text.

In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 323–328, San Francisco, CA, 2001.