Paraphrase Generation from Latent-Variable PCFGs for Semantic Parsing

Shashi Narayan, Siva Reddy, Shay B. Cohen

School of Informatics, University of Edinburgh

INLG, September 2016

THE UNIVERSITY of EDINBURGH

Semantic Parsing for Question Answering

Semantically parsing questions into Freebase logical forms for the goal of question answering

- task-specific grammars (Berant et al., 2013)
- strongly-typed CCG grammars (Kwiatkowski et al., 2013; Reddy et al., 2014, 2016)
- neural networks without requiring any grammar (Yih et al., 2015)

Semantic Parsing for Question Answering

Semantically parsing questions into Freebase logical forms for the goal of question answering

- task-specific grammars (Berant et al., 2013)
- strongly-typed CCG grammars (Kwiatkowski et al., 2013; Reddy et al., 2014, 2016)
- neural networks without requiring any grammar (Yih et al., 2015)

Sensitive to words used in a question and their word order

Vulnerable to unseen words and phrases

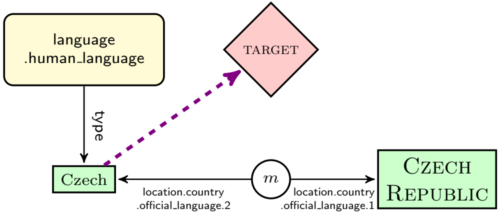
Semantic Parsing for Question Answering: An Example

What language do people in Czech Republic speak?

Semantic Parsing for Question Answering: An Example

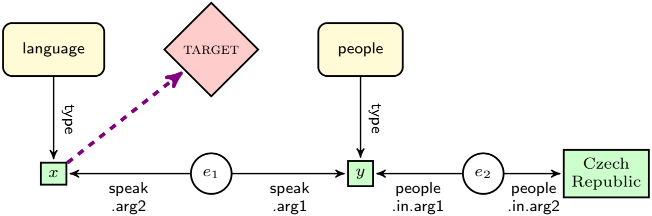
What language do people in Czech Republic speak?

Freebase knowledge graph

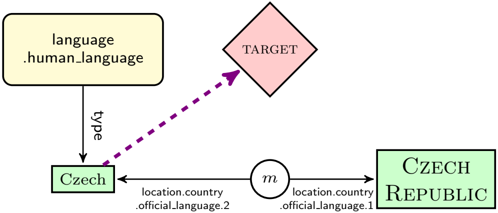


Graph Matching Problem

What language do people in Czech Republic speak?

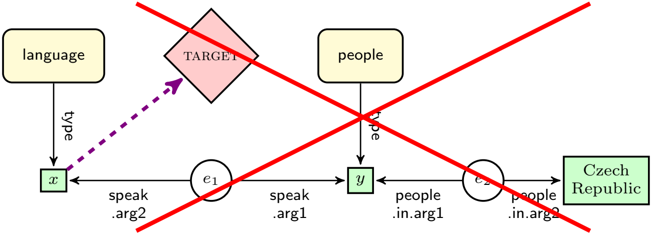


Freebase knowledge graph

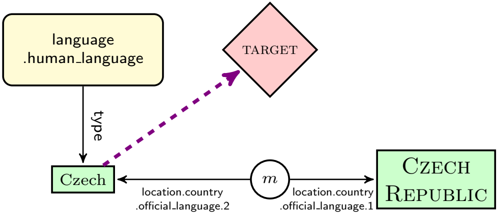


Graph Matching Problem

What language do people in Czech Republic speak?

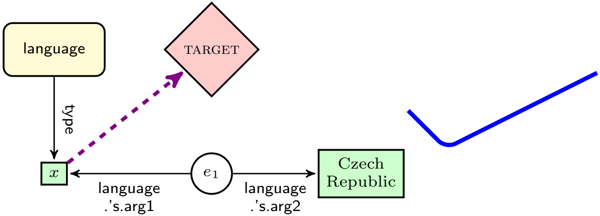


Freebase knowledge graph

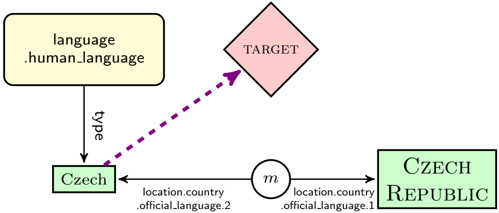


Graph Matching Problem with Paraphrases

What is Czech Republic's language?

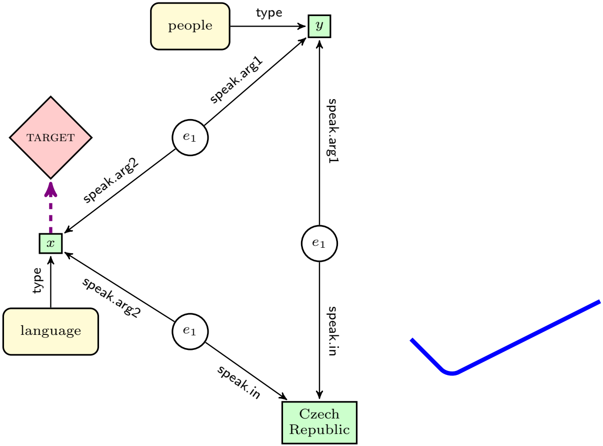


Freebase knowledge graph



Graph Matching Problem with Paraphrases

What language do people speak in Czech Republic?



Question Answering with Paraphrases

Paraphrasing with **phrase-based machine translation** for text-based QA (Duboue and Chu-Carroll, 2006; Riezler et al., 2007)

Paraphrasing with **hand annotated grammars** for KB-based QA (Berant and Liang, 2014)

This talk ...

Paraphrase Generation with **Latent-Variable PCFGs (L-PCFGs)**

This talk ...

Paraphrase Generation with **Latent-Variable PCFGs (L-PCFGs)**

- Uses **spectral method** of Narayan and Cohen (EMNLP 2015) to learn **sparse and robust** grammar to sample paraphrases, and

This talk ...

Paraphrase Generation with **Latent-Variable PCFGs (L-PCFGs)**

- Uses **spectral method** of Narayan and Cohen (EMNLP 2015) to learn **sparse and robust** grammar to sample paraphrases, and
- generates lexically and syntactically diverse paraphrases

This talk ...

Paraphrase Generation with **Latent-Variable PCFGs (L-PCFGs)**

- Uses **spectral method** of Narayan and Cohen (EMNLP 2015) to learn **sparse and robust** grammar to sample paraphrases, and
- generates lexically and syntactically diverse paraphrases

Improving semantic parsing of questions into Freebase logical forms using paraphrases

Outline of this talk

Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

Semantic Parsing using Paraphrases

Results and Discussion

Outline of this talk

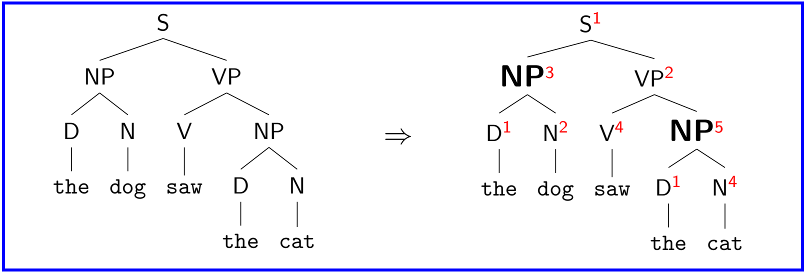
Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

Semantic Parsing using Paraphrases

Results and Discussion

Probabilistic CFGs with Latent States (Matsuzaki et al., 2005; Prescher 2005)



Latent states play the role of nonterminal subcategorization, e.g., NP $\rightarrow \{\text{NP}^1, \text{NP}^2, \ldots, \text{NP}^{24}\}$

- analogous to syntactic heads as in lexicalization (Charniak 1997) ?

They are not part of the observed data in the treebank

Estimating PCFGs with Latent States (L-PCFGs)

EM Algorithm (Matsuzaki et al., 2005; Petrov et al., 2006)

$\Downarrow$ **Problems with local maxima**; it fails to provide certain type of theoretical guarantees as it doesn't find global maximum of the log-likelihood

Estimating PCFGs with Latent States (L-PCFGs)

EM Algorithm (Matsuzaki et al., 2005; Petrov et al., 2006)

$\Downarrow$ **Problems with local maxima**; it fails to provide certain type of theoretical guarantees as it doesn't find global maximum of the log-likelihood

Spectral Algorithm (Cohen et al., 2012, 2014, Narayan and Cohen, 2015, 2016)

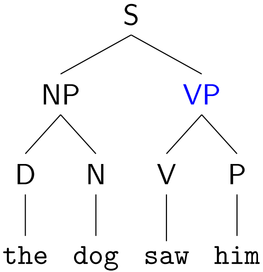
$\Uparrow$ Statistically consistent algorithms that make use of spectral decomposition

$\Uparrow$ Much faster training than the EM algorithm

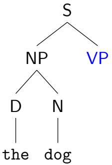
Intuition behind the Spectral Algorithm

Inside and outside trees

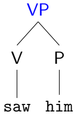
At node VP:



Outside tree $o =$



Inside tree $t =$

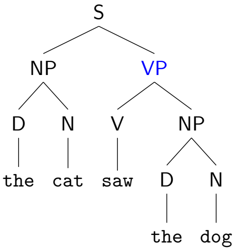


Conditionally independent given the label and the hidden state

$$p(o, t|\text{VP}, h) = p(o|\text{VP}, h) \times p(t|\text{VP}, h)$$

Inside Features used

Consider the VP node in the following tree:

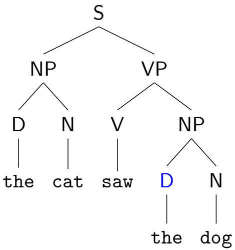


The inside features consist of:

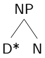
- The pairs `(VP, V)` and `(VP, NP)`
- The rule `VP` $\rightarrow$ `V NP`
- The tree fragment `(VP (V saw) NP)`
- The tree fragment `(VP V (NP D N))`
- The pair of head part-of-speech tag with `VP`: `(VP, V)`

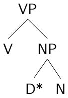
Outside Features used

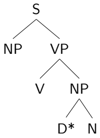
Consider the D node in the following tree:



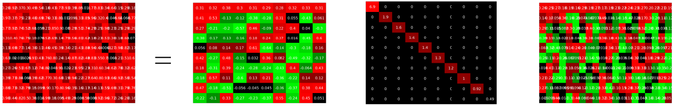
The outside features consist of:

- The pairs `(D, NP)` and `(D, NP, VP)`
- The pair of head part-of-speech tag with `D: (D, N)`
- The tree fragments 

, 

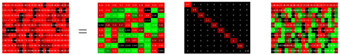
and 

Recent Advances in Spectral Estimation



Singular value decomposition (SVD) of cross-covariance matrix for each nonterminal

Recent Advances in Spectral Estimation

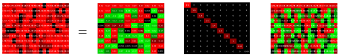


SVD Step

Method of moments (Cohen et al., 2012, 2014)

- Averaging with SVD parameters ⇒ Dense estimates

Recent Advances in Spectral Estimation

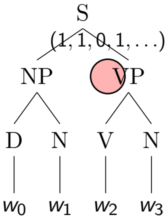


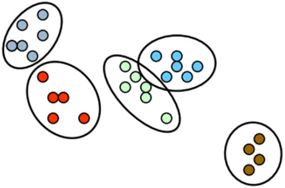
SVD Step

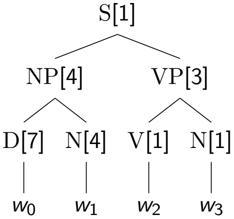
Method of moments (Cohen et al., 2012, 2014)

- Averaging with SVD parameters ⇒ Dense estimates

Clustering variants (Narayan and Cohen 2015)







Sparse estimates

Outline of this talk

Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

Semantic Parsing using Paraphrases

Results and Discussion

Outline of this talk

Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

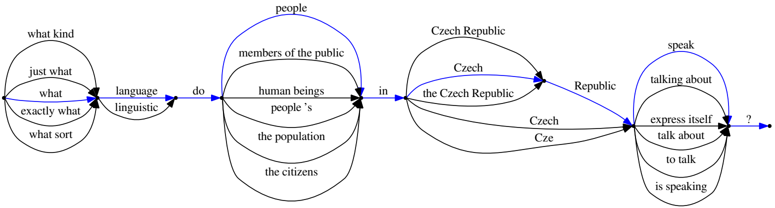
Semantic Parsing using Paraphrases

Results and Discussion

Paraphrase Generation Algorithm

Given an input sentence

- **Word lattice construction** to constrain our paraphrases to a specific choice of words and phrases



What language do people in Czech Republic speak?

Paraphrase Generation Algorithm

Given an input sentence

- **Word lattice construction** to constrain our paraphrases to a specific choice of words and phrases
- **Sampling paraphrases** using L-PCFGs, constrained by the word lattice

Paraphrase Generation Algorithm

Given an input sentence

- **Word lattice construction** to constrain our paraphrases to a specific choice of words and phrases
- **Sampling paraphrases** using L-PCFGs, constrained by the word lattice
- **Paraphrase classification** to improve precision

L-PCFG Estimation for Sampling Paraphrases

The PARALEX Corpus, 18m paraphrase pairs with 2.4M distinct questions (Fader et. al. 2013)

Who wrote the Winnie the Pooh books? Who is the author of winnie the pooh? What was the name of the authur of winnie the pooh? Who wrote the series of books for Winnie the poo? Who wrote the children's storybook 'Winnie the Pooh'? Who is poohs creator?
What relieves a hangover? What is the best cure for a hangover? The best way to recover from a hangover? Best remedy for a hangover? What takes away a hangover? How do you lose a hangover? What helps hangover symptoms?
What are social networking sites used for? Why do people use social networking sites worldwide? Advantages of using social network sites? Why do people use social networks a lot? Why do people communicate on social networking sites? What are the pros and cons of social networking sites?
How do you say Santa Claus in Sweden? Say santa clause in sweden? How do you say santa clause in swedish? How do they say santa in Sweden? In Sweden what is santa called? Who is sweden santa?

L-PCFG Estimation for Sampling Paraphrases

The PARALEX Corpus, 18m paraphrase pairs with 2.4M distinct questions (Fader et. al. 2013)

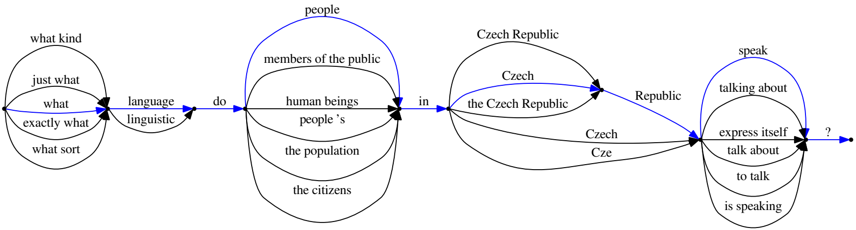
Parse all the questions using the BLLIP Parser (Charniak and Johnson, 2005)

Estimate a **robust** and **sparse** L-PCFG G_{syn} with $m = 24$ (Narayan and Cohen 2015)

Sampling Sentential Paraphrases using L-PCFG G_{syn}

Given an input word lattice and a grammar G_{syn}:

Lexical pruning: Extract a grammar G'_{syn} from G_{syn} which is constrained to the lattice

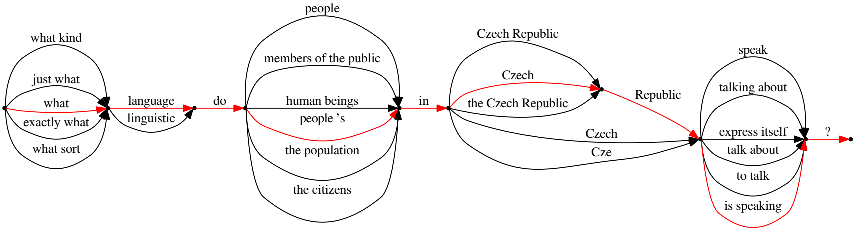


What language do people in Czech Republic speak?

Sampling Sentential Paraphrases using L-PCFG G_{syn}

Given an input word lattice and a grammar G_{syn}:

Controlled sampling: Sample a question from G'_{syn} by recursively sampling nodes in the derivation tree, together with their latent states, over the lattice



(what, language, do, people 's, in, Czech, Republic, is speaking, ?)

$\Downarrow$

what is Czech Republic 's language?

Paraphrase Classification

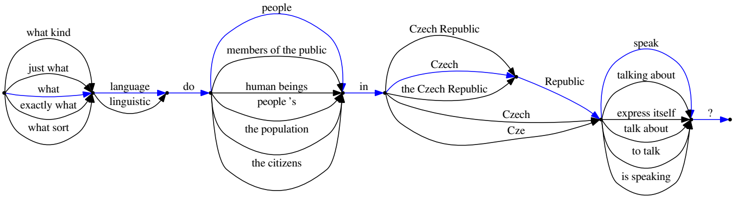
Filter with a classifier to improve the precision of the generated paraphrases

MT metrics for paraphrase identification (Madnani et al. 2012)

Word Lattice Construction

Two approaches:

1. Lexical and phrasal paraphrase rules from the **Paraphrase Database** (Ganitkevitch et al., 2013)



What language do people in Czech Republic speak?

Word Lattice Construction

Two approaches:

1. Lexical and phrasal paraphrase rules from the **Paraphrase Database** (Ganitkevitch et al., 2013)

2. Lexical paraphrases from **Bi-layered L-PCFG**

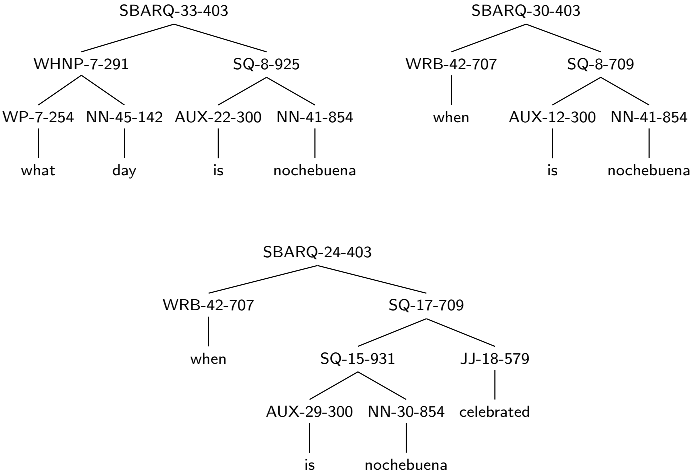
Inducing Paraphrases with a Bi-layered L-PCFG

L-PCFG $G_{layered}$ with two layers of latent states:

one layer is intended to capture the usual syntactic information (traditional G_{syn} with $m = 24$), and

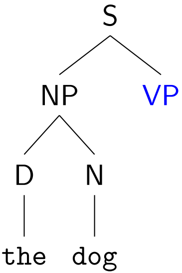
the other aims to capture semantic and topical information by using a large set of states (G_{par} with $m = 1000$)

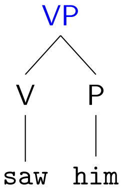
Training trees for Bi-layered L-PCFG Training



Features for Second Layer

Design feature functions ψ and ϕ:





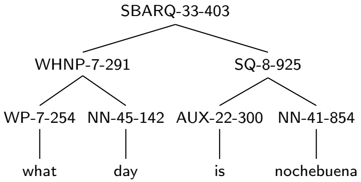
Outside tree $o \Rightarrow$	Inside tree $t \Rightarrow$
$\psi(o) = [0, 1, 0, 0, \ldots, 0, 1] \in \mathbb{R}^{d'}$	$\phi(t) = [1, 0, 0, 0, \ldots, 1, 0] \in \mathbb{R}^{d}$
Bag of aligned words (the, dog, pet, ...)	Bag of aligned words (saw, him, notice, ...)

Training a Bi-layered L-PCFG

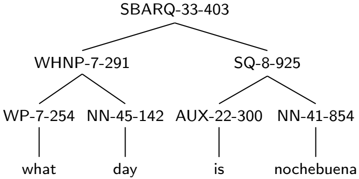
The PARALEX Corpus, 18m paraphrase pairs (Fader et. al. 2013)

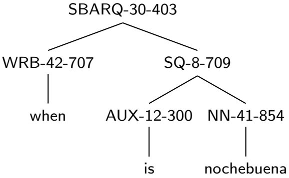
Who wrote the Winnie the Pooh books? Who is the author of winnie the pooh? What was the name of the authur of winnie the pooh? Who wrote the series of books for Winnie the poo? Who wrote the children's storybook 'Winnie the Pooh'? Who is poohs creator?
What relieves a hangover? What is the best cure for a hangover? The best way to recover from a hangover? Best remedy for a hangover? What takes away a hangover? How do you lose a hangover? What helps hangover symptoms?
What are social networking sites used for? Why do people use social networking sites worldwide? Advantages of using social network sites? Why do people use social networks a lot? Why do people communicate on social networking sites? What are the pros and cons of social networking sites?
How do you say Santa Claus in Sweden? Say santa clause in sweden? How do you say santa clause in swedish? How do they say santa in Sweden? In Sweden what is santa called? Who is sweden santa?

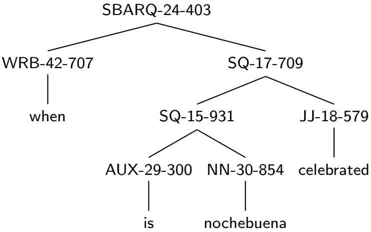
Inducing Paraphrases with a Bi-layered L-PCFG



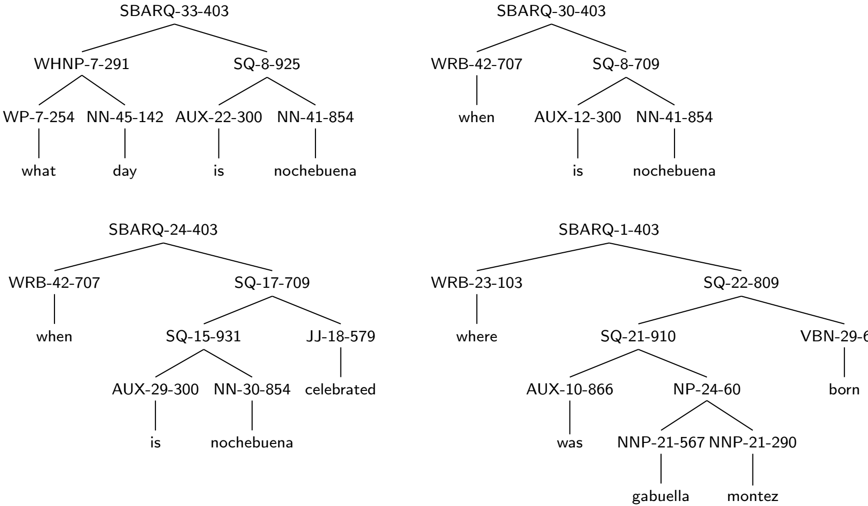
Inducing Paraphrases with a Bi-layered L-PCFG



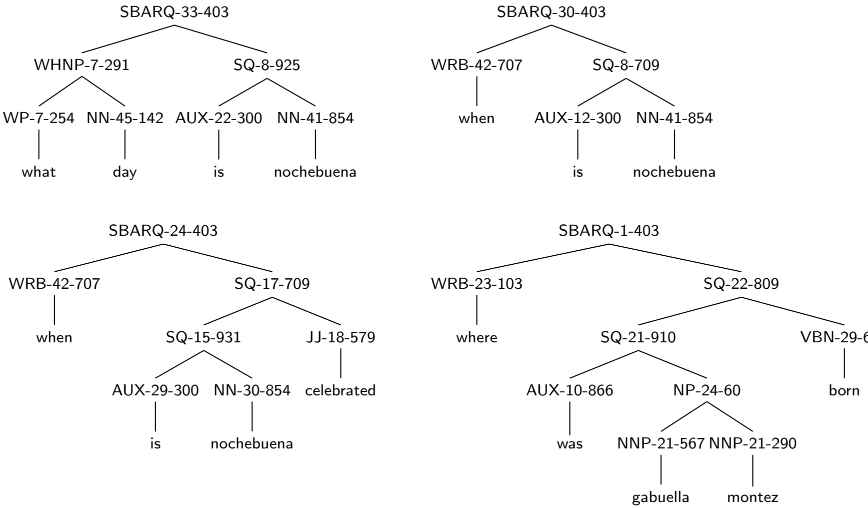




Inducing Paraphrases with a Bi-layered L-PCFG



Inducing Paraphrases with a Bi-layered L-PCFG



Inducing lexical paraphrases only

Outline of this talk

Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

Semantic Parsing using Paraphrases

Results and Discussion

Outline of this talk

Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

Semantic Parsing using Paraphrases

Results and Discussion

Semantic Parsing using Paraphrases (Reddy et. al., 2014)

What language do people in Czech Republic speak?

Semantic Parsing using Paraphrases (Reddy et. al., 2014)

What language do people in Czech Republic speak?

What language do people in Czech Republic speak? What is Czech Republic's language? What language do people speak in Czech Republic? ...

Semantic Parsing using Paraphrases (Reddy et. al., 2014)

What language do people in Czech Republic speak?

What language do people in Czech Republic speak? What is Czech Republic's language? What language do people speak in Czech Republic? …

$\xrightarrow{\text{CCG}}$

$\lambda e.\text{speak.arg1}(e, \text{people})$
$\wedge \text{speak.arg2}(e, \text{language?})$
$\wedge \text{speak.in}(e, \text{CzechRepublic})$

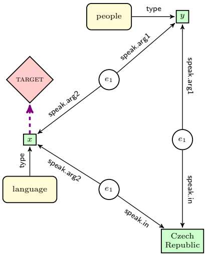
Semantic Parsing using Paraphrases (Reddy et. al., 2014)

What language do people in Czech Republic speak?

What language do people in Czech Republic speak? What is Czech Republic's language? What language do people speak in Czech Republic? ...

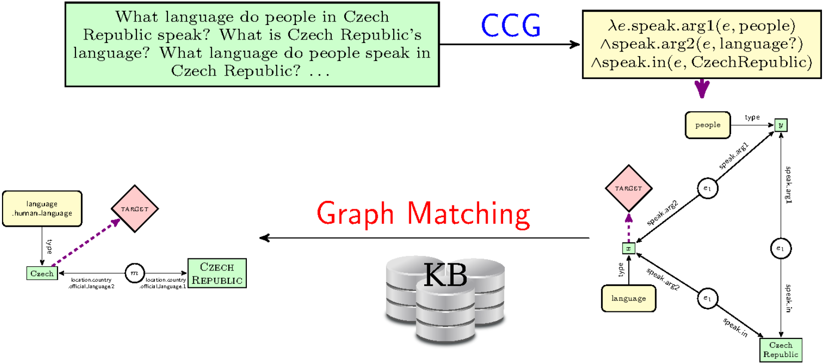
CCG

$\lambda e.\mathrm{speak.arg1}(e, \mathrm{people})$
$\wedge \mathrm{speak.arg2}(e, \mathrm{language?})$
$\wedge \mathrm{speak.in}(e, \mathrm{CzechRepublic})$



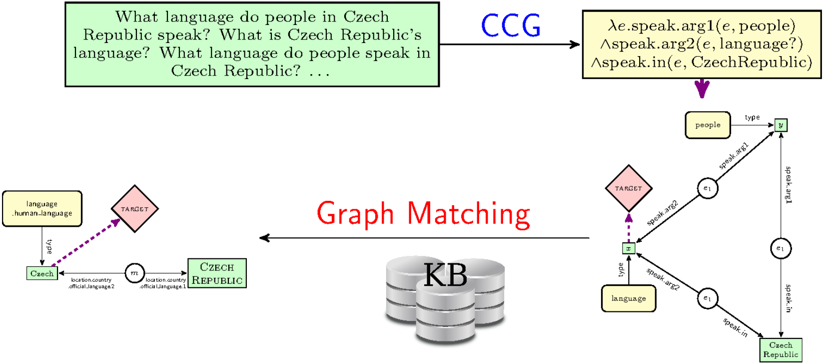
Semantic Parsing using Paraphrases (Reddy et. al., 2014)

What language do people in Czech Republic speak?



Semantic Parsing using Paraphrases (Reddy et. al., 2014)

What language do people in Czech Republic speak?



$$(\hat{p}, \hat{u}, \hat{g}) = \underset{(p,u,g)}{\arg\max}\, \theta \cdot \Phi(p, u, g, q, \mathcal{K})$$

where, $\Phi(p, u, g, q, \mathcal{K}) \in \mathbb{R}^n$ denotes the features for the tuple of paraphrase p, ungrounded u and grounded g graphs

Model

Structured Perceptron: Ranks a tuple of paraphrase, grounded and ungrounded graph.

$$(\hat{p}, \hat{u}, \hat{g}) = \arg\max_{(p,u,g)} \theta \cdot \Phi(p, u, g, q, \mathcal{K})$$

Features: Φ is defined over sentence, grounded and ungrounded graph.

Training: Use surrogate gold graph to update weights

$$\theta^{t+1} \leftarrow \theta^{t} + \Phi(p^{+}, u^{+}, g^{+}, q, \mathcal{K}) - \Phi(\hat{p}, \hat{u}, \hat{g}, q, \mathcal{K}),$$

More details: We use Margin-Sensitive Averaged Peceptron.

Outline of this talk

Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

Semantic Parsing using Paraphrases

Results and Discussion

Outline of this talk

Spectral Learning of Latent-variable PCFGs

Paraphrase Generation using L-PCFGs

Semantic Parsing using Paraphrases

Results and Discussion

Experimental Setup

WebQuestions (Berant et al., 2013)

- Google search queries starting with wh question words
- 5,810 question-answer pairs (3,778 training and 2,032 test)
- *Development experiments*: held-out data consisting of 30% training questions

Experimental Setup

WebQuestions (Berant et al., 2013)

- Google search queries starting with wh question words
- 5,810 question-answer pairs (3,778 training and 2,032 test)
- *Development experiments*: held-out data consisting of 30% training questions

Evaluation metric

- Average precision, average recall and average F_1 (Berant et al., 2013)

Experimental Setup

Our systems

- NAIVE: Word lattice representing the input sentence itself
- PPDB: Word lattice constructed using the PPDB rules
- BILAYERED: Word lattice constructed using the bi-layered L-PCFG

Experimental Setup

Our systems

- NAIVE: Word lattice representing the input sentence itself
- PPDB: Word lattice constructed using the PPDB rules
- BILAYERED: Word lattice constructed using the bi-layered L-PCFG

Baselines

- ORIGINAL: Semantic parser (Reddy et. al., 2014) without paraphrases
- MT: Monolingual machine translation based model for paraphrase generation (Quirk et al., 2004; Wubben et al., 2010)

Results on the Development Set

Oracle statistics and Average F_1 Scores

Method	avg oracle F_1	# oracle graphs	avg F_1
ORIGINAL	65.1	11.0	44.7
MT	71.5	77.2	47.0
NAIVE	71.2	53.6	47.5
PPDB	71.8	59.8	47.9
BILAYERED	71.6	55.0	47.1

Results on the Development Set

Oracle statistics and Average F_1 Scores

Method	avg oracle F_1	# oracle graphs	avg F_1
ORIGINAL	65.1	11.0	44.7
MT	71.5	77.2	47.0
NAIVE	71.2	53.6	47.5
PPDB	71.8	59.8	47.9
BILAYERED	71.6	55.0	47.1

Results on the Development Set

Oracle statistics and Average F_1 Scores

Method	avg oracle F_1	# oracle graphs	avg F_1
ORIGINAL	65.1	11.0	44.7
MT	71.5	77.2	47.0
NAIVE	71.2	53.6	47.5
PPDB	71.8	59.8	47.9
BILAYERED	71.6	55.0	47.1

Results on the Development Set

Oracle statistics and Average F_1 Scores

Method	avg oracle F_1	# oracle graphs	avg F_1
ORIGINAL	65.1	11.0	44.7
MT	71.5	77.2	47.0
NAIVE	71.2	53.6	47.5
PPDB	71.8	59.8	47.9
BILAYERED	71.6	55.0	47.1

Results on the Development Set

Oracle statistics and Average F_1 Scores

Method	avg oracle F_1	# oracle graphs	avg F_1
ORIGINAL	65.1	11.0	44.7
MT	71.5	77.2	47.0
NAIVE	71.2	53.6	47.5
PPDB	71.8	59.8	47.9
BILAYERED	71.6	55.0	47.1

Results on the Development Set

Oracle statistics and Average F_1 Scores

Method	avg oracle F_1	# oracle graphs	avg F_1
ORIGINAL	65.1	11.0	44.7
MT	71.5	77.2	47.0
NAIVE	71.2	53.6	47.5
PPDB	71.8	59.8	47.9
BILAYERED	71.6	55.0	47.1

Results on the Development Set

Oracle statistics and Average F_1 Scores

Method	avg oracle F_1	# oracle graphs	avg F_1
ORIGINAL	65.1	11.0	44.7
MT	71.5	77.2	47.0
NAIVE	71.2	53.6	47.5
PPDB	71.8	59.8	47.9
BILAYERED	71.6	55.0	47.1

Results on the Test Set

Method	avg P.	avg R.	avg F_1
ORIGINAL	53.2	54.2	45.0
MT	48.0	56.9	47.1
NAIVE	48.1	57.7	47.2
PPDB	48.4	58.1	47.7
BILAYERED	47.0	57.6	47.2

Results on the Test Set

Method	avg P.	avg R.	avg F_1
ORIGINAL	53.2	54.2	45.0
MT	48.0	56.9	47.1
NAIVE	48.1	57.7	47.2
PPDB	48.4	58.1	47.7
BILAYERED	47.0	57.6	47.2
Others			
Berant and Liang '14	40.5	46.6	39.9
Bordes et al. '14	-	-	39.2
Dong et al. '15	-	-	40.8
Yao '15	52.6	54.5	44.3
Bao et al. '15	44.7	52.5	45.3
Bast and Haussmann '15	49.8	60.4	49.4
Berant and Liang '15	50.4	55.7	49.7
Yih et al. '15	52.8	60.7	52.5

Results on the Test Set

Method	avg P.	avg R.	avg F_1
ORIGINAL	53.2	54.2	45.0
MT	48.0	56.9	47.1
NAIVE	48.1	57.7	47.2
PPDB	48.4	58.1	47.7
BILAYERED	47.0	57.6	47.2
Others			
Berant and Liang '14	40.5	46.6	39.9
Bordes et al. '14	-	-	39.2
Dong et al. '15	-	-	40.8
Yao '15	52.6	54.5	44.3
Bao et al. '15	44.7	52.5	45.3
Bast and Haussmann '15	49.8	60.4	49.4
Berant and Liang '15	50.4	55.7	49.7
Yih et al. '15	52.8	60.7	52.5

Results on the Test Set

Method	avg P.	avg R.	avg F_1
ORIGINAL	53.2	54.2	45.0
MT	48.0	56.9	47.1
NAIVE	48.1	57.7	47.2
PPDB	48.4	58.1	47.7
BILAYERED	47.0	57.6	47.2
Others			
Berant and Liang '14	40.5	46.6	39.9
Bordes et al. '14	-	-	39.2
Dong et al. '15	-	-	40.8
Yao '15	52.6	54.5	44.3
Bao et al. '15	44.7	52.5	45.3
Bast and Haussmann '15	49.8	60.4	49.4
Berant and Liang '15	50.4	55.7	49.7
Yih et al. '15	52.8	60.7	52.5

Results on the Test Set

Method	avg P.	avg R.	avg F_1
ORIGINAL	53.2	54.2	45.0
MT	48.0	56.9	47.1
NAIVE	48.1	57.7	47.2
PPDB	48.4	58.1	47.7
BILAYERED	47.0	57.6	47.2
Others			
Berant and Liang '14	40.5	46.6	39.9
Bordes et al. '14	-	-	39.2
Dong et al. '15	-	-	40.8
Yao '15	52.6	54.5	44.3
Bao et al. '15	44.7	52.5	45.3
Bast and Haussmann '15	49.8	60.4	49.4
Berant and Liang '15	50.4	55.7	49.7
Yih et al. '15	52.8	60.7	52.5

Error Mining

78.4% of the errors are partially correct answers occurring due to incomplete gold answer annotations or partially correct groundings

13.5% are due to bad paraphrases, and

the rest 8.1% are due to wrong entity annotations

Conclusion

Our method is rather generic and can be applied to any question answering system

Bi-layered L-PCFG for semantic similarity tasks