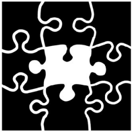
Learning Compositional Semantics for Open Domain Semantic Parsing

Phong Le, Willem Zuidema

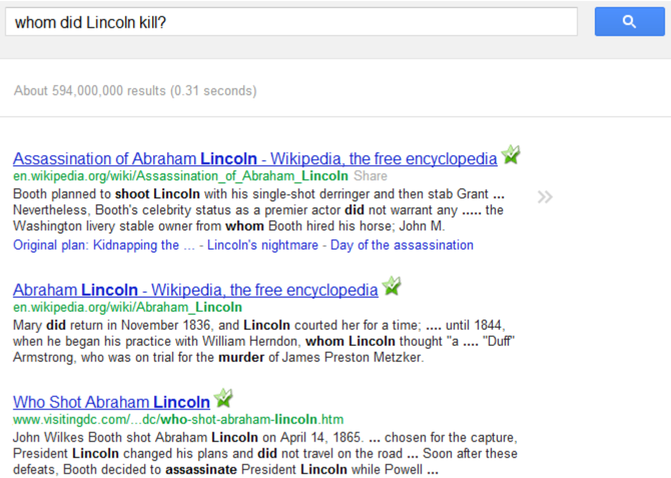
Institute for Logic, Language and Computation
University of Amsterdam



October 31, 2012

Outline

Does Google understand what I mean?



Even people misunderstand...

What should we do?

Semantic Parsing (or Semantic Analysis)

Translate natural language sentences into their **computer executable** *meaning representations*.

Example

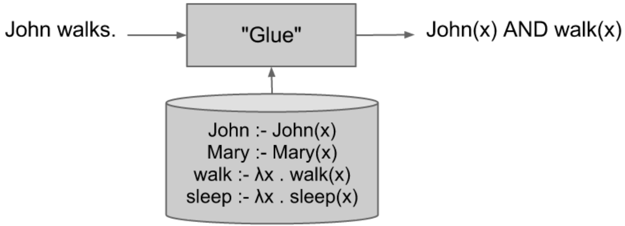
Which states border Arizona ?

`answer(A,(state(A),const(B,stateid(arizona)),next_to(A,B)))`

Common Strategy

Principle of Compositionality

"The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined."



Traditional approach: with lambda calculus

Lambda calculus

is an elegant tool for semantic composition in a bottom up manner

$$\text{John} :\!-\ \lambda x.john(x)$$
$$\text{walks} :\!-\ \lambda P.\lambda y.walks(y) \wedge P\ y$$

$$\begin{aligned}\text{John walks} &:\!-\ (\lambda P.\lambda y.walks(y) \wedge P\ y)\ \ (\lambda x.john(x))\\ &:\!-\ \lambda y.walks(y) \wedge (\lambda x.john(x))\ y\\ &:\!-\ \lambda y.walks(y) \wedge john(y)\end{aligned}$$

Why **learning** semantic parsing?

Speech recognition and syntactic analysis have had significant development under the umbrella of machine learning, thanks to

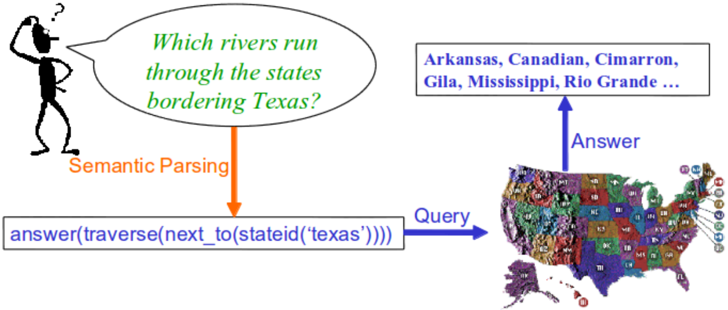
- the power of machine learning tools (e.g. Hidden Markov Model, Expectation Maximization)
- large corpora (e.g. WSJ)

How about semantic parsing?

a complicated story...

Domain-dependent semantic parsing

Geoquery



Features

closed world, simple present tense, wh-question

No need to handle

anaphora, possibility/necessity, tense, event,...

Learning approaches

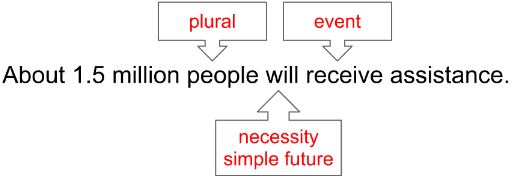
- Supervised
  - fully supervised (MRs are available)
    - Structured learning with CCG
    - Syntax-based Machine translation
    - Kernel-based approach
    - Integrating syntax and semantics
  - weakly supervised (response-driven)
    - Clarke et al. (2010)
    - Liang et al. (2011)
- Semi-supervised
  - Kernel-based approach
- Unsupervised
  - Confidence driven semantic parsing

Open-domain semantic parsing

Learning open-domain semantic parsing

is still largely unexplored, because of many difficulties

- need to handle various linguistics phenomena and syntactic structures



  In addition: presupposition, anaphora, etc.

- lack large standard corpora

In this paper

We want to bridge this gap!

by introducing a new learning open-domain semantic parsing approach: Dependency-based Semantic Composition using Graphs (DeSCoG)

Outline

- Meaning representation with **graph-based variant** of Discourse Representation Structures
  - remove the need of the lambda calculus
- Semantic composition
  - use existing state-of-the-art syntactic dependency parsers
  - with a probability model
- Experimental results on
  - Groningen Meaning Bank
  - Geoquery

Why abandon the lambda calculus?

How to learn lexicon?

Given

$$\text{John walks :- } \lambda y.walks(y) \wedge john(y)$$

how to find lambda forms for *John* and *walks*? Notorious problem!!!

$\Rightarrow$ Easy for composition, but difficult for learning lexicon!

Our idea

Not so difficult for composition, but easy for learning lexicon!

Why use existing syntactic dependency parsers?

- dependency structures encode predicate-argument relations which are strongly related to semantics
- the total complexity is reduced significantly compared with parsing syntax and semantics simultaneously
- prior knowledge of syntax is particularly helpful when sentences are long and complex

Outline

Discourse Representation Structure (DRS)

is used to represent a mental representation of the hearer as the discourse unfolds.

Example

Mary loves a man.

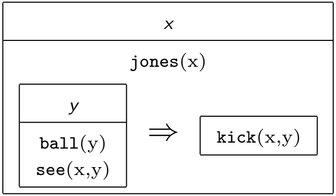
x, *y*
mary(x)
man(y)
love(x,y)

Our goal is

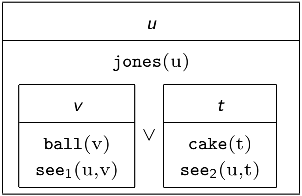
to assign as-good-as-possible DRS to unseen sentences.

How to evaluate success?

1. *If Jones sees a ball, he will kick it.*



2. *Jones will see a ball or a cake.*



The best alignment A is

$$A(x) = u, A(y) = v,$$
$$A(\texttt{jones}) = \texttt{jones}, A(\texttt{ball}) = \texttt{ball}, A(\texttt{see}) = \texttt{see}_2$$
$$A(\textit{outerbox}) = \textit{outerbox}, A(\textit{leftbox}_{\Rightarrow}) = \textit{leftbox}_{\vee}$$
$$A(\textit{rightbox}_{\Rightarrow}) = \textit{rightbox}_{\vee}$$

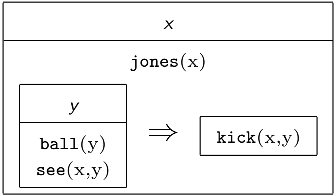
1. *If Jones sees a ball, he will kick it.*

$$\begin{array}{|c|}\hline x \\ \hline \texttt{jones}(x) \\ \begin{array}{|c|}\hline y \\ \hline \texttt{ball}(y) \\ \texttt{see}(x,y) \\ \hline \end{array} \Rightarrow \begin{array}{|c|}\hline \texttt{kick}(x,y) \\ \hline \end{array} \\ \hline \end{array}$$

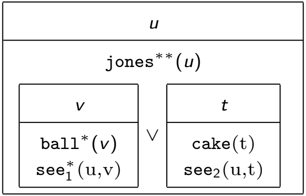
2. *Jones will see a ball or a cake.*

$$\begin{array}{|c|}\hline u \\ \hline \texttt{jones}^{**}(u) \\ \begin{array}{|c|}\hline v \\ \hline \texttt{ball}^{*}(v) \\ \texttt{see}_1^{*}(u,v) \\ \hline \end{array} \vee \begin{array}{|c|}\hline t \\ \hline \texttt{cake}(t) \\ \texttt{see}_2(u,t) \\ \hline \end{array} \\ \hline \end{array}$$

1. *If Jones sees a ball, he will kick it.*



2. *Jones will see a ball or a cake.*

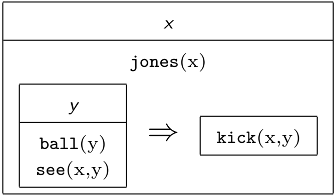


$\Omega(DRS1, DRS2) = 4$

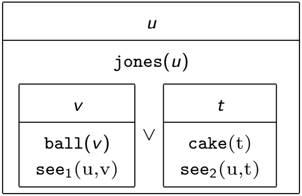
$\text{recall} = \frac{\Omega(DRS1,DRS2)}{\Omega(DRS1,DRS1)} = \frac{4}{10}$, $\text{prec} = \frac{\Omega(DRS1,DRS2)}{\Omega(DRS2,DRS2)} = \frac{4}{12}$, $\text{fscore} = 0.36$

Does it fit our intuition?

1. *If Jones sees a ball, he will kick it.*

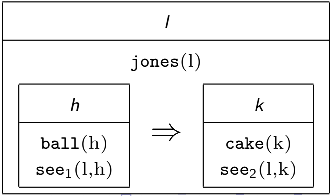


2. *Jones will see a ball or a cake.*



which one is more similar to

3 *If Jones sees a ball, he will see a cake.*



Does it fit our intuition?

Human intuition

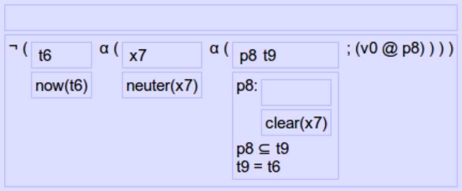
DRS1 is more similar to DRS3 than DRS2 to DRS3

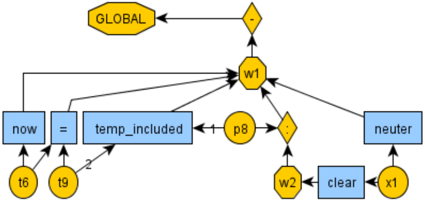
The measure

f-score$(DRS_1, DRS_3) = \frac{16}{22} = 0.73$ and
f-score$(DRS_2, DRS_3) = \frac{12}{24} = 0.5$; hence
f-score$(DRS_1, DRS_3) >$ f-score(DRS_2, DRS_3)

Semantic Graph

Representing a DRS by a graph.





Easy for composing and breaking components: simply by removing/adding links/nodes.

Outline

Introduction

Meaning Representation

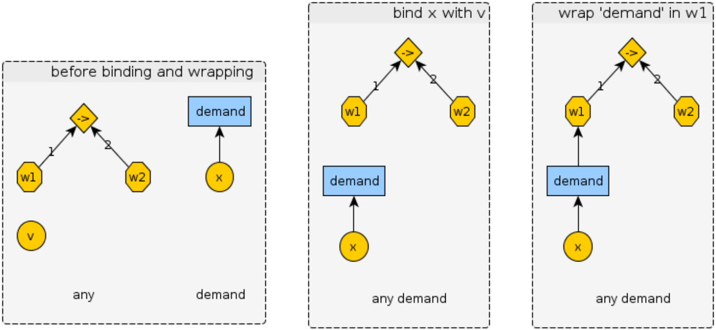
Semantic Composition

Experimental results
- Groningen Meaning Bank
- Geoquery

Conclusion

Combinatory Operators

- **Binding** is to bind a referent node x with another referent node v, denoted by $x \bowtie v$,
- **Wrapping** is to link a predicate/operator node p to a wrapper node w, denoted by $p \odot w$.



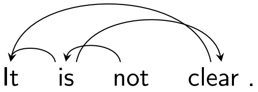
Composition Procedure

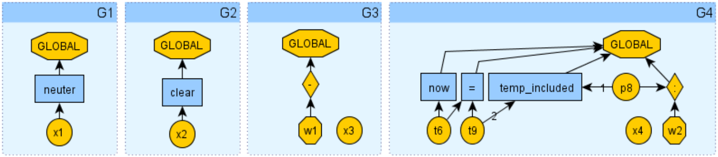
3 steps

1. select lexical elements
2. apply binding operations
3. apply wrapping operations

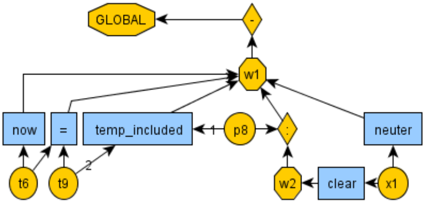
following a dependency structure

Given a dependency structure and a bag of partial graphs

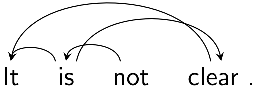


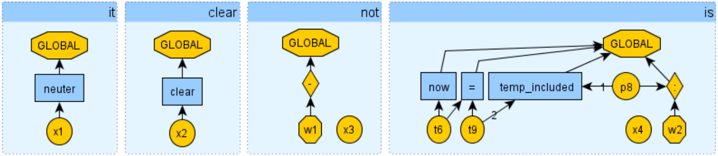


Target

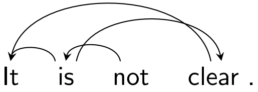


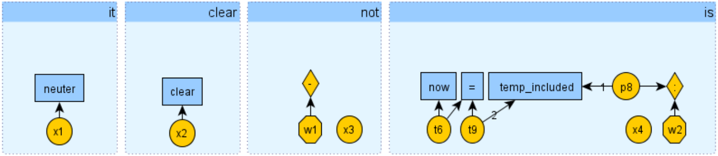
Step 1: Selecting lexical elements



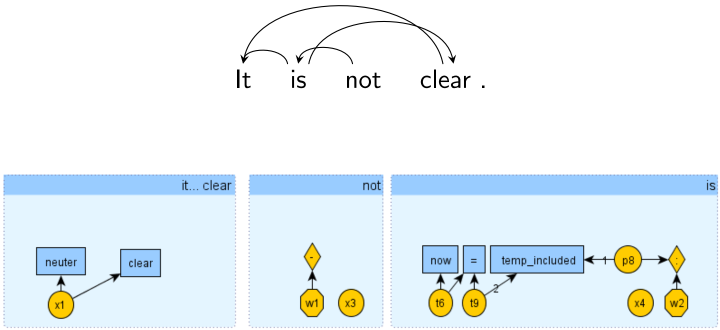


Step 2: Binding

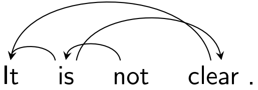


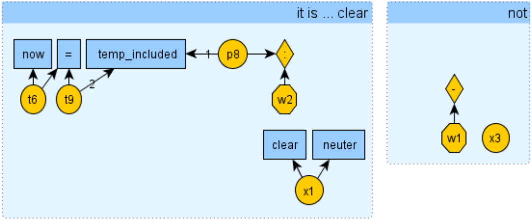


Step 2: Binding

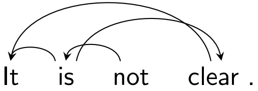


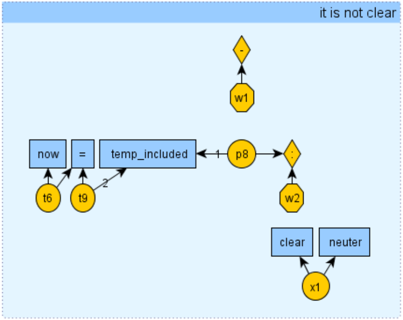
Step 2: Binding



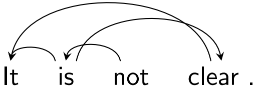


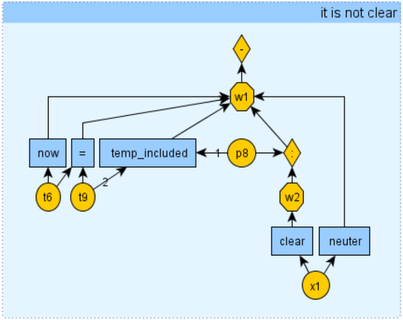
Step 2: Binding



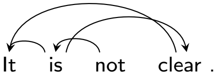


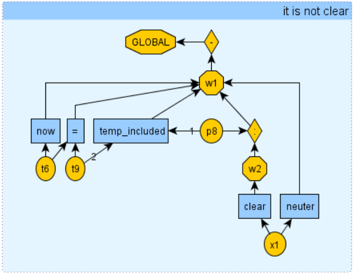
Step 3: Wrapping



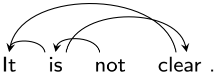


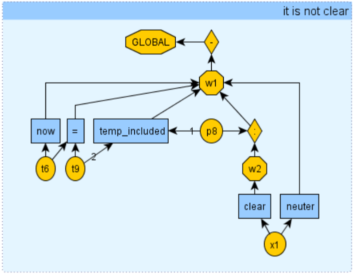
Step 3: Wrapping





Step 3: Wrapping





How to prohibit *clear* from linking to *GLOBAL*?

Wrapping constraint for all dependencies $s_i \curvearrowright s_j \in D$, if a referent node v in G^j binds with a referent node u in G^i then all the predicate/operator nodes in G^i linked from u must link to wrapper nodes which have access to v.

Probability Model

Let $G = (G_c, B, W)_{S,D}$

- $G_c = \{G_c^1, ..., G_c^n\}$ be a set of assigned partial graphs,
- $B = \{u \bowtie v\}$ be a set of binding operations, and
- $W = \{f \hat{\odot}_k o\}$ be a set of in-wrapper relations

Probability Model

Given a sentence S and a dependency structure D, find the most probable semantic graph G^*

$$
\begin{aligned}
G^* &= \arg\max_{G} Pr(G|S,D) \\
&= \arg\max_{G=(G_c,B,W)_{S,D}} Pr(G_c|S,D)Pr(B|G_c,S,D)Pr(W|G_c,B,S,D)
\end{aligned}
$$

Probability Model

$G = (G_c, B, W)_{S,D}$

- $G_c = \{G_c^1, ..., G_c^n\}$ be a set of assigned partial graphs,
- $B = \{u \bowtie v\}$ be a set of binding operations, and
- $W = \{f \hat{\odot}_k o\}$ be a set of in-wrapper relations

Under some independence assumption.

$$Pr(G_c|S,D) = \prod_{i=1}^{n} Pr_l(G_c^i|s_i, POS(s_i), POS(Dep(s_i)))$$

$$Pr(B|G_c,S,D) = \prod_{u \bowtie v \in B} Pr_b(u \bowtie v|G_c(u), G_c(v), POS(s(u)), POS(s(v)))$$

$$Pr(W|G_c,B,S,D) = Z \times \psi(W) \times \prod_{f \hat{\odot}_k o \in W} Pr_w(f \hat{\odot}_k o|G_c(f), G_c(o), POS(path(s(f), s(o))))$$

$\psi(W) = 1$ if the wrapper constraint is satisfied, $= 0$ otherwise

Searching

$G = (G_c, B, W)_{S,D}$

- $G_c = \{G_c^1, ..., G_c^n\}$ be a set of assigned partial graphs,
- $B = \{u \bowtie v\}$ be a set of binding operations, and
- $W = \{f \hat{\odot}_k o\}$ be a set of in-wrapper relations

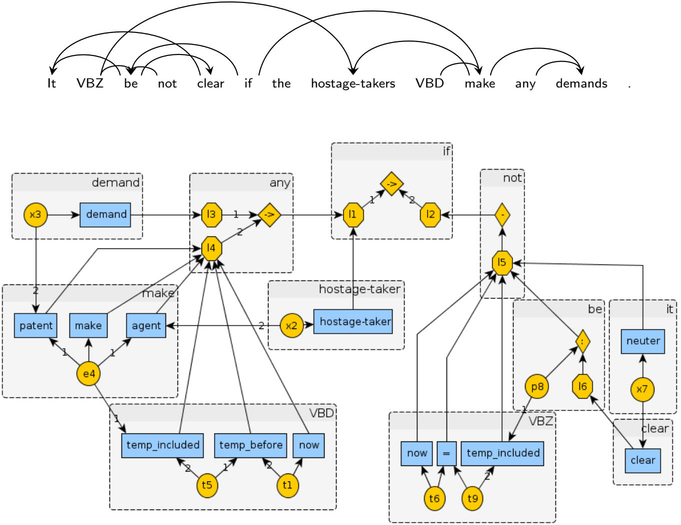
$$G^* = \underset{G=(G_c,B,W)_{S,D}}{\arg\max} Pr(G_c|S,D)Pr(B|G_c,S,D)Pr(W|G_c,B,S,D)$$

2-stage beam search

- **stage 1** maximize $Pr(G_c|S,D)Pr(B|G_c,S,D)$, output a list of N-best (G_c, B)'s
- **stage 2** maximize $Pr(W|G_c,B,S,D)$, look for the best W for each of those N-best (G_c, B)'s.
  - using Linear Integer Programming

Learning lexicon

Word-to-graph alignment. Using A-star algorithm, based on $Pr(node|word)$ (Giza++).



Parameter estimation

Using relative frequencies

$$Pr_l(G|s, POS(s), POS(Dep(s))) \approx \frac{\#(G, s, POS(s), POS(Dep(s)))}{\#(s, POS(s), POS(Dep(s)))}$$

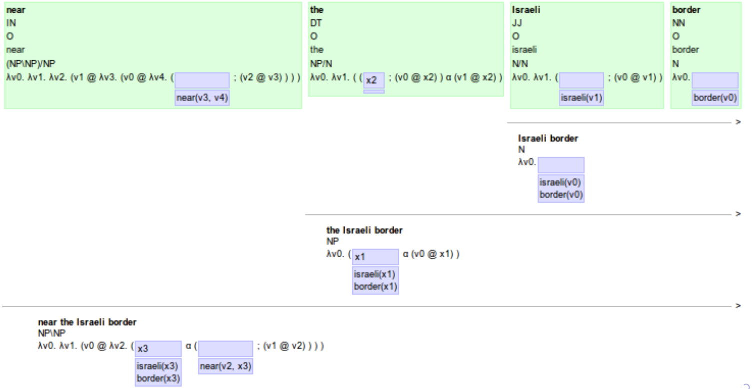
with smoothing

- Good-Turing
- multilevel back-off

Outline

GMB 1.1 corpus

- contains 2000 documents with 9418 sentences
- from many public sources: Voice of America, fables, CIA World Factbook, and MASC Full
- MR language: Partial DRS
- automatically parsed with Boxer and partly hand-corrected



Settings

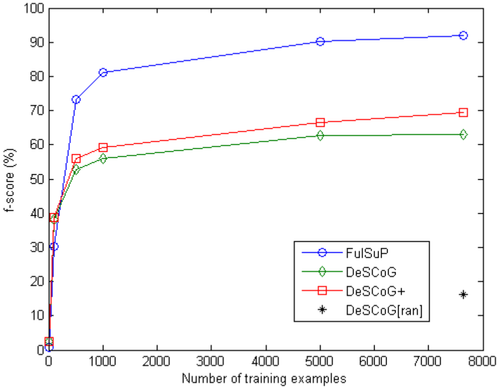
Dataset

- **Training** (GMB.0-79) 7642 examples in the sections from 0 to 79 for training
- **Testing** (GMB.80-99) 1776 examples

Alternatives

- **FulSuP** (Fully Supervised Parser) is a parser that was trained with the semantic lexicon given by GMB.
- **DeSCoG+** is DeSCoG with the help from an "oracle" for the alignment process beforehand thanks to the semantic lexicon given by GMB.
- **DeSCoG[ran]** (baseline) is DeSCoG with random parameters

Results



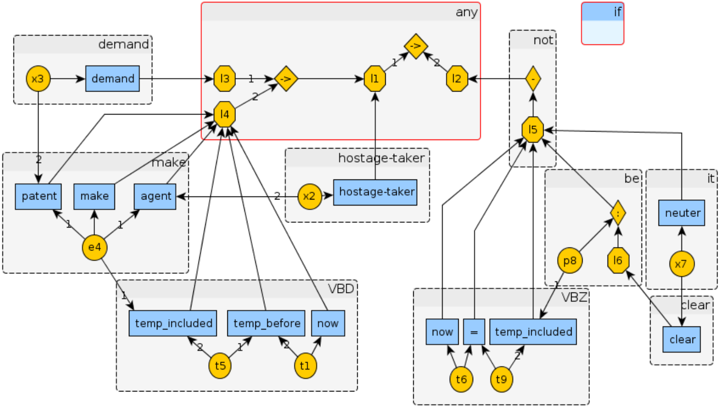
f-score (%)
Number of training examples
FulSuP
DeSCoG
DeSCoG+
DeSCoG[ran]

Analysing the alignment phase

The alignment phase succeeded 5725 times, which is 74.9%.

False alignment

$Pr(\rightarrow | any) = 0.69 > Pr(\rightarrow | if) = 0.48$

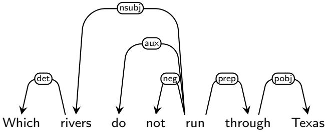


Geoquery

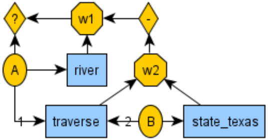
Geoquery corpus

contains 880 English queries and their manually annotated MRs in a Prolog-base first-order language and FUNQL

In our experiments




```
answer(A,(river(A),not((traverse(A,B),const(B,stateid(Texas))))))
```



Settings

- 10-fold cross validation
- A test MR is correct if it and the gold-standard MR receive the same answer
- Precision = # correct/total # parsed, Recall = # correct/total # examples

Alternatives

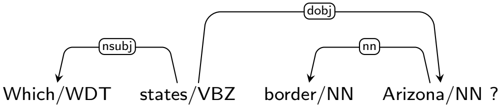
- SCISSOR (Ge and Mooney, 2005), an integrated syntactic-semantic parser,
- KRISP (Kate and Mooney, 2006), a SVM-based parser using string kernels,
- WASP (Wong and Mooney, 2006) and λ-WASP (Wong, 2007), two parsers based on synchronous grammars,
- Z&C05 (Zettlemoyer and Collins,2005), a parser using structural learning with CCG grammars, and
- SYN0 (Ge and Mooney, 2009), a parser using an existing syntactic parser.

Results

	Recall	Precision	Fscore
DeSCoG	74.89	87.40	80.66
SYN0	78.98	81.76	80.35
λWASP	86.59	91.95	89.19
Z&C05	79.29	96.25	86.95
SCISSOR	72.3	91.5	80.77
WASP	74.8	87.2	80.5
KRISP	71.7	93.3	81.1

Problem from wrong syntactic parses

Incorrect syntactic parse



leads to difficulty (or impossibility) creating

```
answer(A,(state(B),next_to(A,B),const(B,stateid(arizona))
```

- Parsing syntax and semantics simultaneously can overcome this problem by making use of the frequent appearance of the structure *A border B*.

Outline

Introduction

Meaning Representation

Semantic Composition

Experimental results

- Groningen Meaning Bank
- Geoquery

Conclusion

Conclusion

- Introduce new learning approach, DeSCoG, for open-domain semantic parsing
  - represent logical forms by graphs, which provide a flexible way to combine and break components
  - use dependency structures and a probabilistic model for semantic composition
- Introduce new method for measuring the similarity between two DRSs
- DeSCoG significantly outperformed the baseline on the Groningen Meaning Bank corpus, and performed equivalently with many parsers on Geoquery.

Future work

- Enhance the word-to-graph alignment
- Does the relative frequent estimate equal the maximum likelihood estimate?
- Embed unsupervised dependency parsing model in the current semantic parsing model
- Test DeSCoG on other corpora (e.g. CLang, ATIS)

Thank you!