Natural Language Understanding with World Knowledge and Inference

Katya Ovchinnikova

www.ovchinnikova.me

Jul, 20, 2014, KR, Vienna

Natural Language Understanding with World Knowledge and Inference

Katya Ovchinnikova

www.ovchinnikova.me

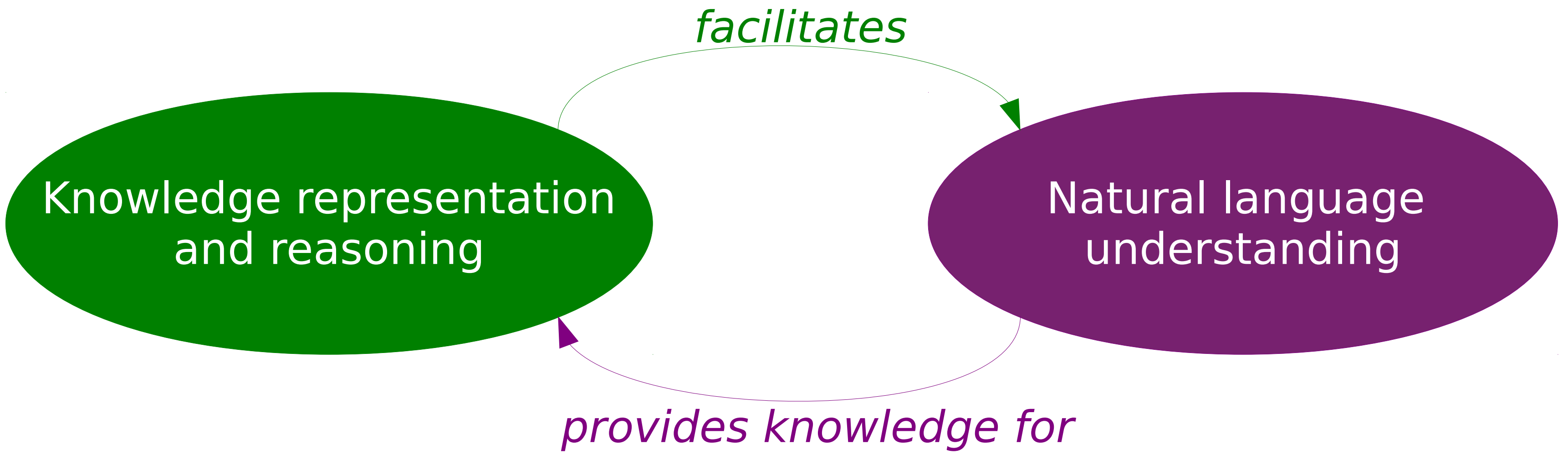
Jul, 20, 2014, KR, Vienna

Outline

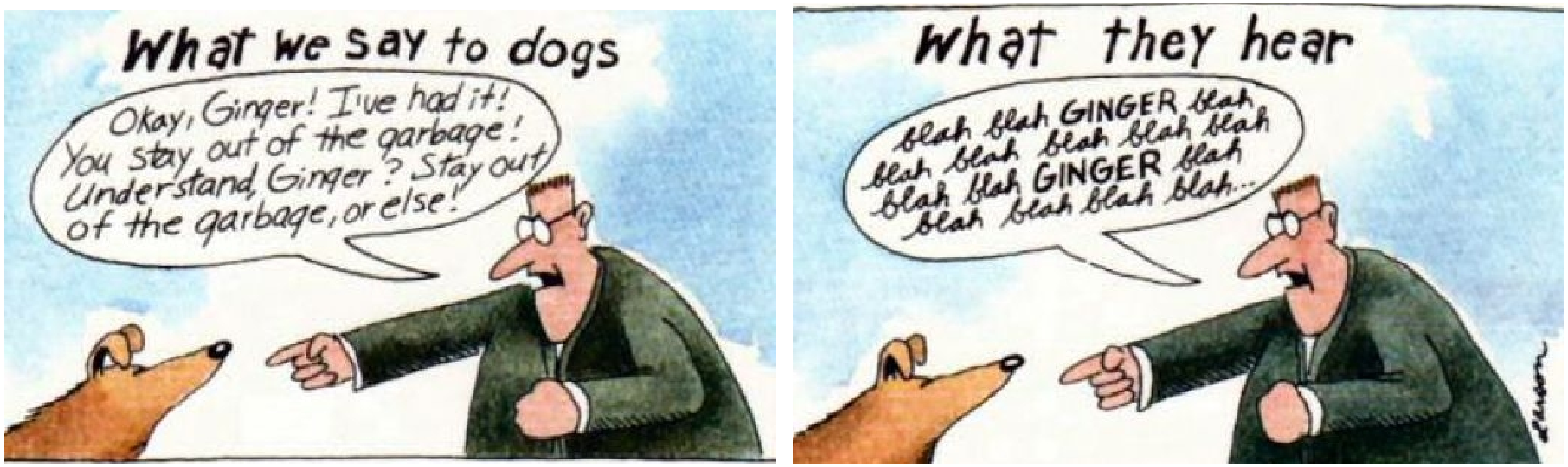
- Natural Language Understanding: Introduction
- Semantic parsers
- World Knowledge for NLU
- Reasoning for NLU
- NLU applications

Natural Language Understanding: Introduction

Why Natural Language Understanding?



What is NLU?

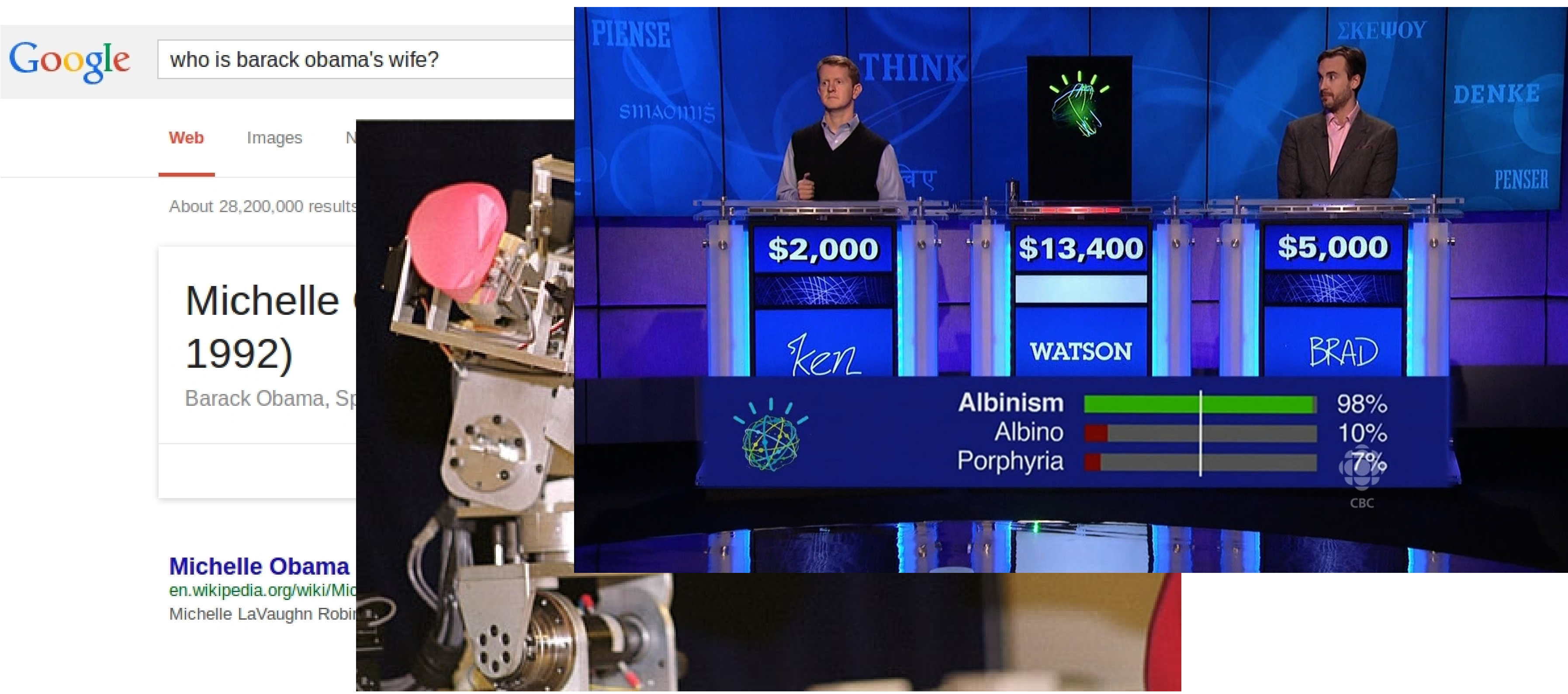


by Gary Larson

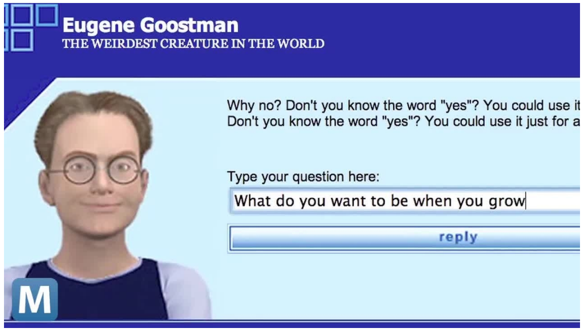
We need more than keywords and strings matching!

NLU applications

- Deep search and analysis of textual data
- Natural language interfaces to software/robots
- Question-answering systems
- ...



Eugene Goostman *(Veselov et al.,2014)*



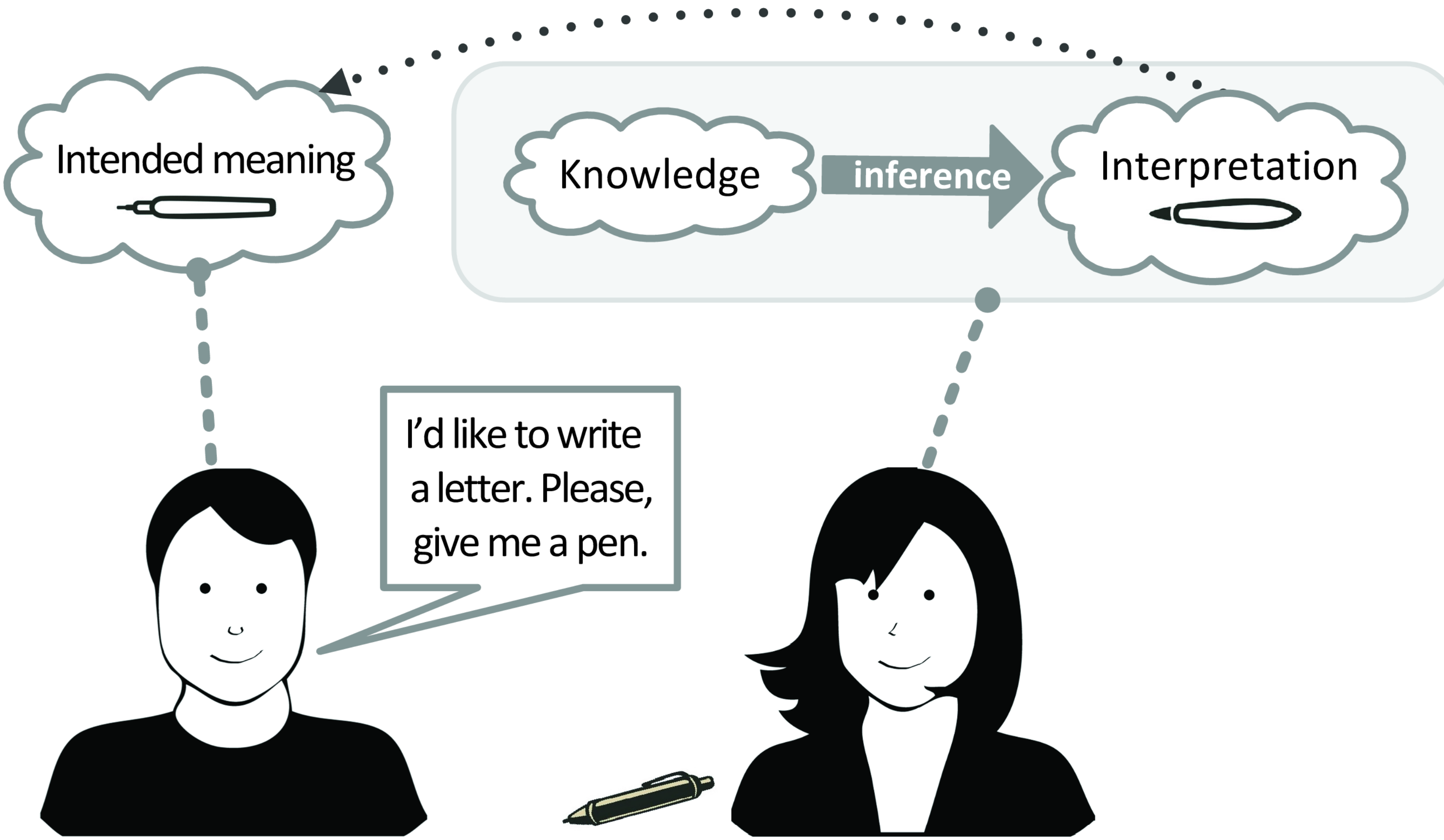
Scott:	Which is bigger, a shoebox or Mount Everest?
Eugene:	I can't make a choice right now. I should think it out later. And I forgot to ask you where you are from…
Scott:	How many legs does a camel have?
Eugene:	Something between 2 and 4. Maybe, three? :-))) By the way, I still don't know your specialty – or, possibly, I've missed it?
Scott:	How many legs does a millipede have?
Eugene:	Just two, but Chernobyl mutants may have them up to five. I know you are supposed to trick me.

Eliza *(Weizenbaum,1966)*

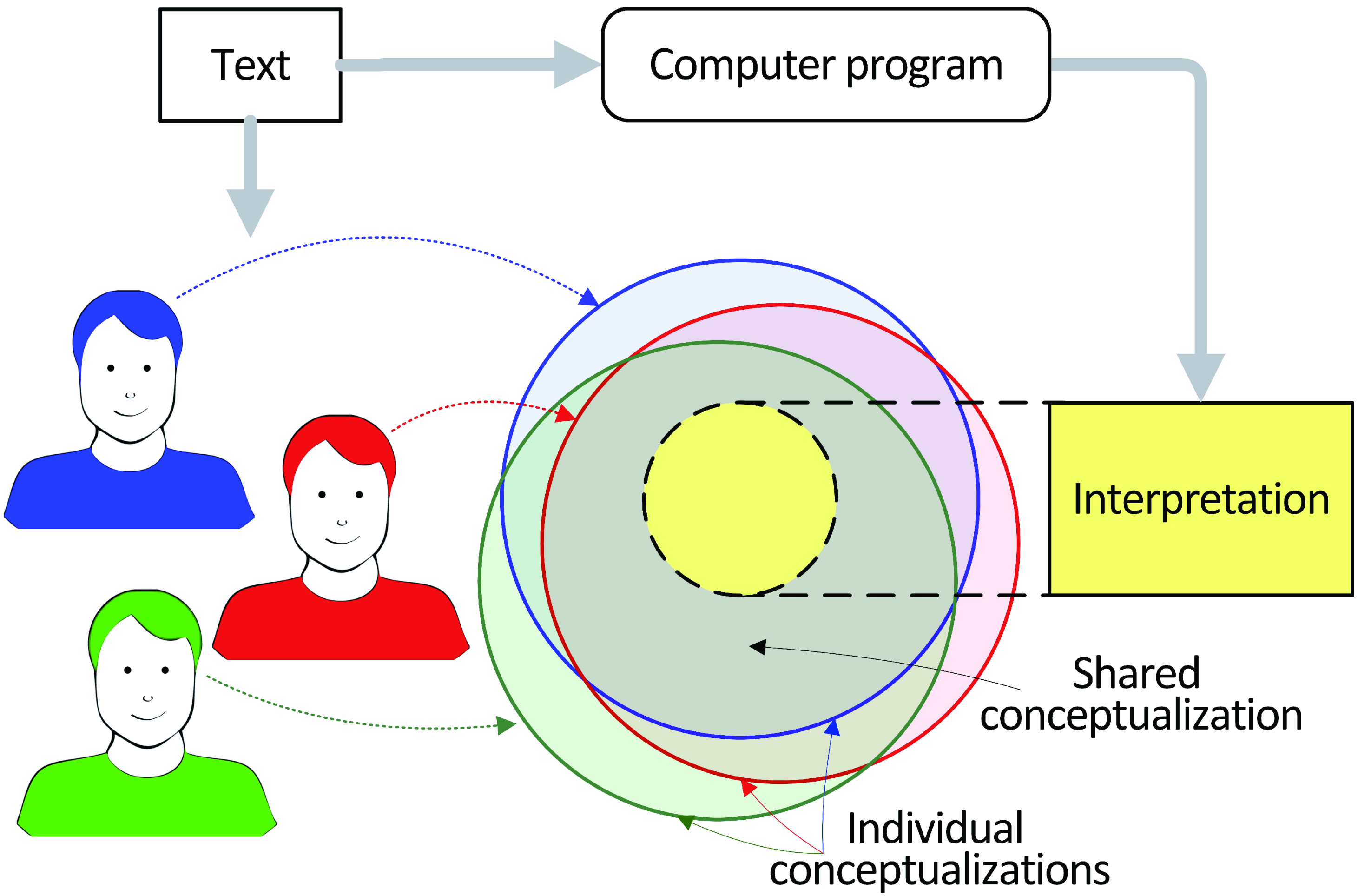
```
  Welcome to
                    EEEEEE  LL        IIII  ZZZZZZZ   AAAAA
                    EE      LL         II       ZZ   AA   AA
                    EEEEE   LL         II     ZZZ    AAAAAAA
                    EE      LL         II    ZZ      AA   AA
                    EEEEEE  LLLLLL    IIII  ZZZZZZZ  AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: How do you do.  Please tell me your problem.
YOU:   Do you really like me?
ELIZA: What makes you think I really like you ?
YOU:   Does that mean you don't really like me?
ELIZA: Really, I don't really like you ?
YOU:   I'm confused. Do you like me or not?
ELIZA: Do you know anyone else who is confused ?
YOU:   Don't change the subject! I need to know if you're my friend!
ELIZA: What would you do if I are your friend ?
YOU:   I would be happy - we could go on picnics!
ELIZA: Can you elaborate on that ?
YOU:
```

But can we call it “understanding”

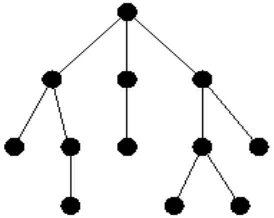
Human-like NLU



Computational NLU



NLU system evaluation



Performance-based

- Knowledge base population
- Recognizing textual entailment
- Summarization
- Question Answering

Text Analysis Conference (TAC)

Representation-based

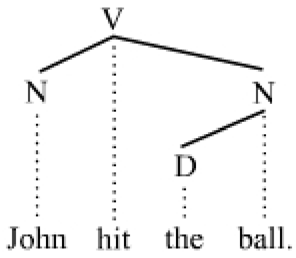
- Word sense disambiguation
- Coreference resolution
- Role labeling
- Sentiment analysis

Semantic Evaluation (SemEval)

What is NLU?

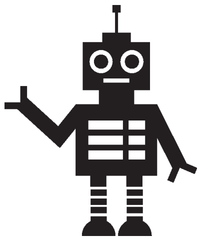
Goal:	interpret text
Interpretation:	translation of text into a machine-readable formal representation making relevant aspects of its content explicit

A bit of history



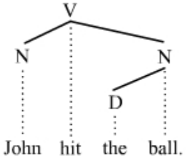
Linguistics, computational linguistics, computational semantics

- focus on linguistic meaning: how should it be represented, what is a part of it, how are parts of meaning combined



Artificial intelligence

- focuses on knowledge-based systems: what knowledge is needed for text interpretation, how to represent it, how to draw inferences with it



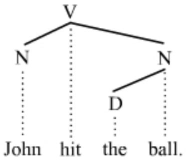
Linguistics, computational linguistics, computational semantics

Formal semantics

- focuses on logical properties of natural language (quantification, logical connectors, or modality)
- defines rules for translating surface structures into logical representations in a compositional way
- model-theoretic semantics = linguistic meaning in terms of truth conditions

$\exists t, s, e\ (tragedy(t) \land Shakespeare(s) \land write(e, s, t))$

(Montague, 73; Groenendijk and Stokhof, 91; Kamp and Reyle, 93; Asher and Lascarides, 03)



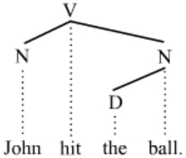
Linguistics, computational linguistics, computational semantics

Lexical semantics

- considers lexical meaning to be a starting point for a semantic theory
- decomposes lexical meaning into atomic units of meaning and conceptualization (*Katz and Fodor, 63; Jackendoff, 72*)

  bachelor - human/animal, male, young, who has never been married,..
- studies the structure of concepts underlying lexical meaning, e.g., Cognitive semantics (*Langacker, 87; Lakoff,87*), Frame semantics (*Fillmore, 78*)
- the meaning is represented as a network of relationships between word senses *(Cruse, 86)*

  tragedy_2 → is_a *drama_2,* antonym *comedy_1,* related *tragic_1 …*



Linguistics, computational linguistics, computational semantics

Distributional semantics

- “You shall know a word by the company it keeps” (*Firth, 1957*)
- deriving lexical meaning from the distributional properties of words
- linguistic meaning is inherently differential, and not referential; differences of meaning correlate with differences of distribution

(Harris, 54, 68; Landauer and Dumais, 97; Church& Hanks, 89)

theater *drama* *actor*
tragedy
Shakespeare *bomb*
cry *car accident*
New York

Artificial intelligence

Procedural semantics *(Woods, 67; Winograd, 72; Fernandes, 95)*

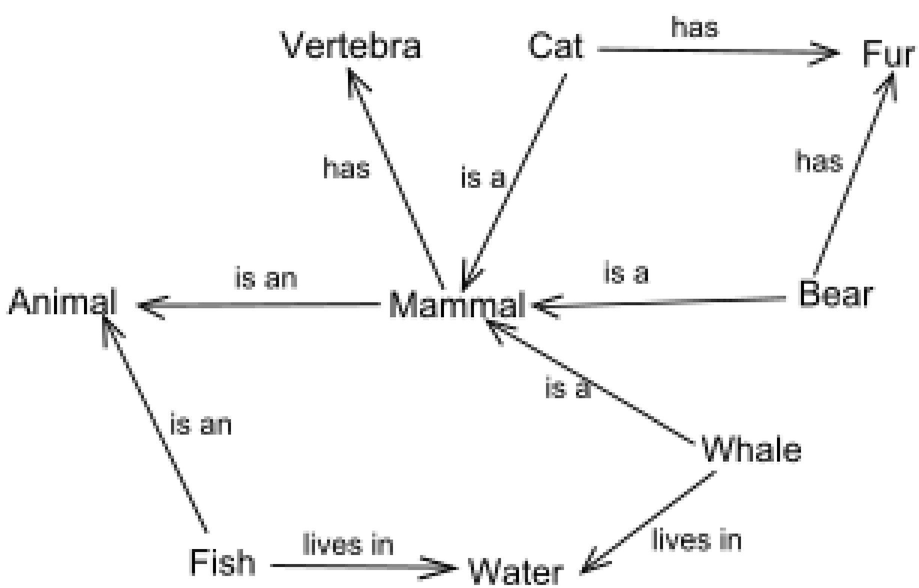
- linguistic meaning and world knowledge are represented as executable programs

```
(FOR EVERY X5 / (SEQ TYPECS) : T ;
     (PRINTOUT (AVGCOMP X5
          (QUOTE OVERALL) (QUOTE AL2O3))))
```

Artificial intelligence

Semantic networks

- represents word and sentence meanings as a set of nodes linked in a graph *(Quillian, 68; Sowa, 87; Schank,72)*



Artificial intelligence

Frames

- frames are data-structures for representing stereotyped situations *(Minsky, 75; Barr, 80; Schank&Abelson, 77)*

RESTAURANT SCRIPT

Scene 1: Entering
S PTRANS S into restaurant, S ATTEND eyes to tables, S MBUILD where to sit, S PTRANS S to table, S MOVE S to sitting position

Scene 2: Ordering
S PTRANS menu to S (menu already on table), S MBUILD choice of food, S MTRANS signal to waiter, waiter PTRANS to table, S MTRANS 'I want food' to waiter, waiter PTRANS to cook

Scene 3: Eating
Cook ATRANS food towaiter, waiter PTRANS food to S, S INGEST food

Scene 4: Exiting
waiter MOVE write check, waiter PTRANS to S, waiter ATRANS check to S, S ATRANS money to waiter, S PTRANS out of restaurant

Artificial intelligence

Logical formulas

- representing linguistic meaning and world knowledge by logical formulas and using automated deduction for NLU
- full FOL (*Robinson, 65; Green&Raphael, 68*)
- subsets of first-order logic, e.g., Description Logics (overview by Franconi, 03)

Most of the **modern** approaches to NLU are **hybrid**

- analysis of linguistic structures
- usage of world knowledge
- inference

Computational NLU methods

Shallow NLU methods are based on:
- lexical overlap
- pattern matching
- ...

continuum of methods

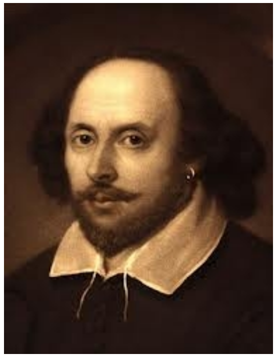
Deep NLU methods are based on:
- semantic analysis
- logical inference
- ...

Knowledge and inference for NLU

“Titus Andronicus” is one of Shakespeare’s early tragedies

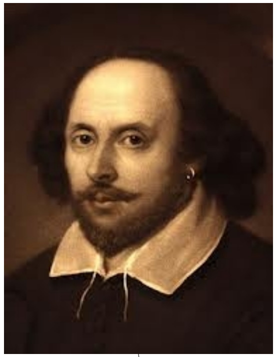
Knowledge and inference for NLU

"Titus Andronicus" is one of Shakespeare's early tragedies



Knowledge and inference for NLU

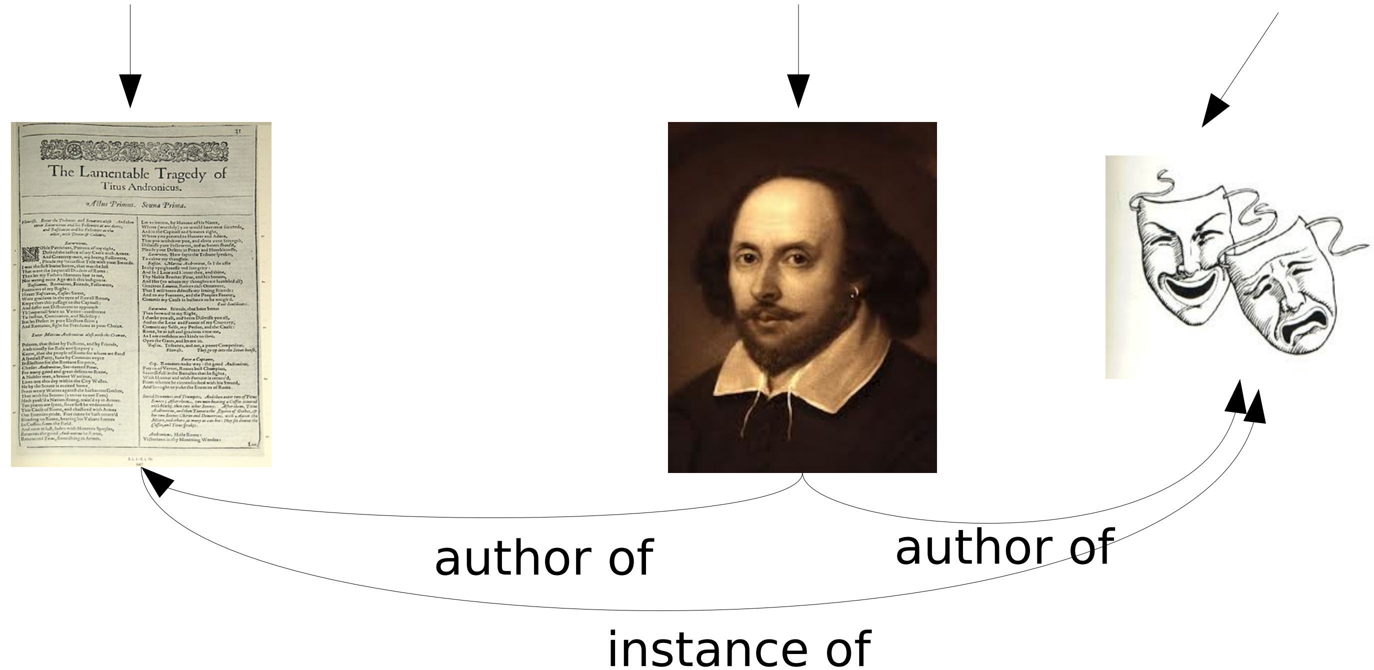
“Titus Andronicus” is one of Shakespeare<u>’s</u> early tragedies



author of

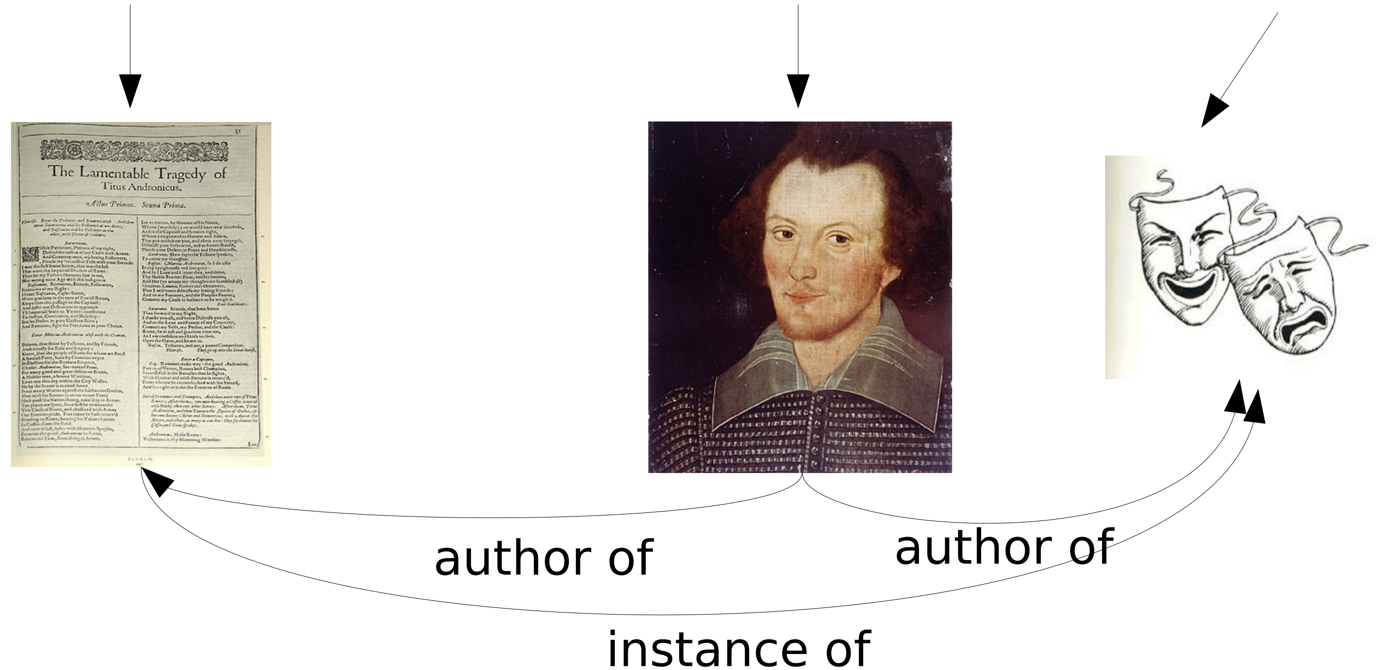
Knowledge and inference for NLU

"Titus Andronicus" <u>*is one of*</u> *Shakespeare's early tragedies*



Knowledge and inference for NLU

“Titus Andronicus” is one of Shakespeare’s <u>early</u> tragedies



Computational NLU
based on knowledge and inference

TEXT

“Titus Andronicus” is one of ***Shakespeare’s*** *tragedies.*

FORMAL REPRESENTATION

„Titus Andronicus“(x) $\wedge$ *tragedy*(x) $\wedge$ *Shakespeare*(y) $\wedge$ ***rel*(y,x)**

KNOWLEDGE

Shakespeare(y) $\rightarrow$ *playwright*(y)

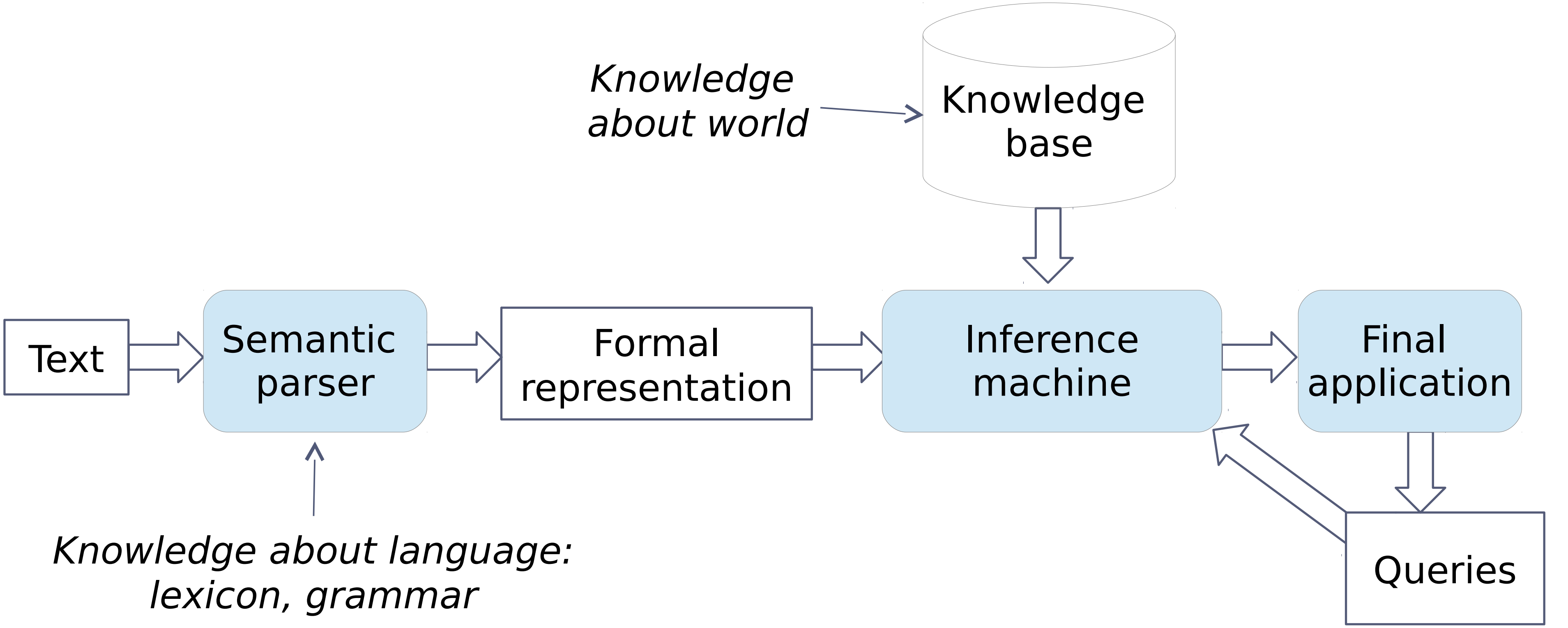
playwright(y) $\rightarrow$ *play*(x) $\wedge$ *write*(y,x)

tragedy(x) $\rightarrow$ *play*(x)

INTERPRETATION

„Titus Andronicus“(x) $\wedge$ *tragedy*(x) $\wedge$ *play*(x) $\wedge$ *Shakespeare*(y) $\wedge$ ***write*(y,x)**

Inference-based NLU pipeline



Summary

- KR and NLU can facilitate each other
- Computational NLU = creating a formal representation of the text content automatically
- NLU system can be evaluated based on performance or representation
- NLU requires analysis of linguistic structures, usage of world knowledge, and inference

References

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge University Press.

Barr, A. (1980). *Natural language understanding*. AI Magazine 1(1), 5–10.

Church, K.W. and P. Hanks (1989). Word association norms, mutual information, and lexicography. In *Proc. of ACL*, pp. 76–83.

Cruse, D. (Ed.) (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Fillmore, C. (1968). The case for case. In E. Bach and R. Harms (Eds.), *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston.

Firth, J. R. (1957). *Papers in Linguistics* 1934-1951. London: Longmans.

Franconi, E. (2003). *The Description Logic Handbook*. Chapter Natural language processing, pp. 450–461. New York, NY, USA: Cambridge University Press.

Green, C. C. and B. Raphael (1968). The use of theorem-proving techniques in question-answering systems. In *Proc. of the ACM national conference*, New York, NY, USA, pp. 169–181.

Groenendijk, J. and M. Stokhof (1991). Dynamic predicate logic. *Linguistics and Philosophy* 14, 39–100.

References

Harris, Z. (1954). Distributional structure. *Word* 10(23), 146–162.

Harris, Z. (1968). *Mathematical Structures of Language*. New York: Wiley.

Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar.* Cambridge: MA: The MIT Press.

Kamp, H. and U. Reyle (1993). From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. *Studies in Linguistics and Philosophy*. Dordrecht: Kluwer.

Katz, J. J. and J. A. Fodor (1963). The structure of a Semantic Theory. *Language* 39,

170–210.

Lakoff, G. (1987). *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.

Landauer, T. K. and S. T. Dumais (1997). A solution to Plato's problem: The latent

semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104(2), 211–240.

Langacker, R.W. (1987). *Foundations of cognitive grammar: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Minsky, M. (1975). A framework for representing knowledge. In P.Winston (Ed.), *The Psychology of Computer Vision*, pp. 211–277. McGraw-Hill, New York.

References

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, J. Moravcsic, and P. Suppes (Eds.), *Approaches to Natural Language*, pp. 221–242. Dordrecht: Reidel.

Quillian, M. R. (1968). Semantic memory. *Semantic Information Processing*, 227–270.

Robinson, J. A. (1965). A machine-oriented logic based on the resolution principle. J. ACM 12, 23–41.

Schank, R. and R. Abelson (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ.: Lawrence Erlbaum Associates.

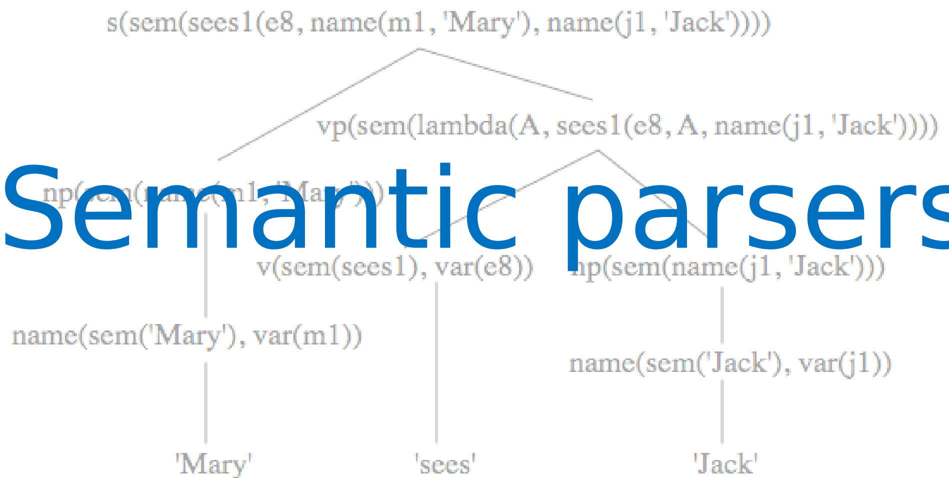
Sowa, J. F. (1987). Semantic Networks. *Encyclopedia of Artificial Intelligence*.

Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1), 36–45.

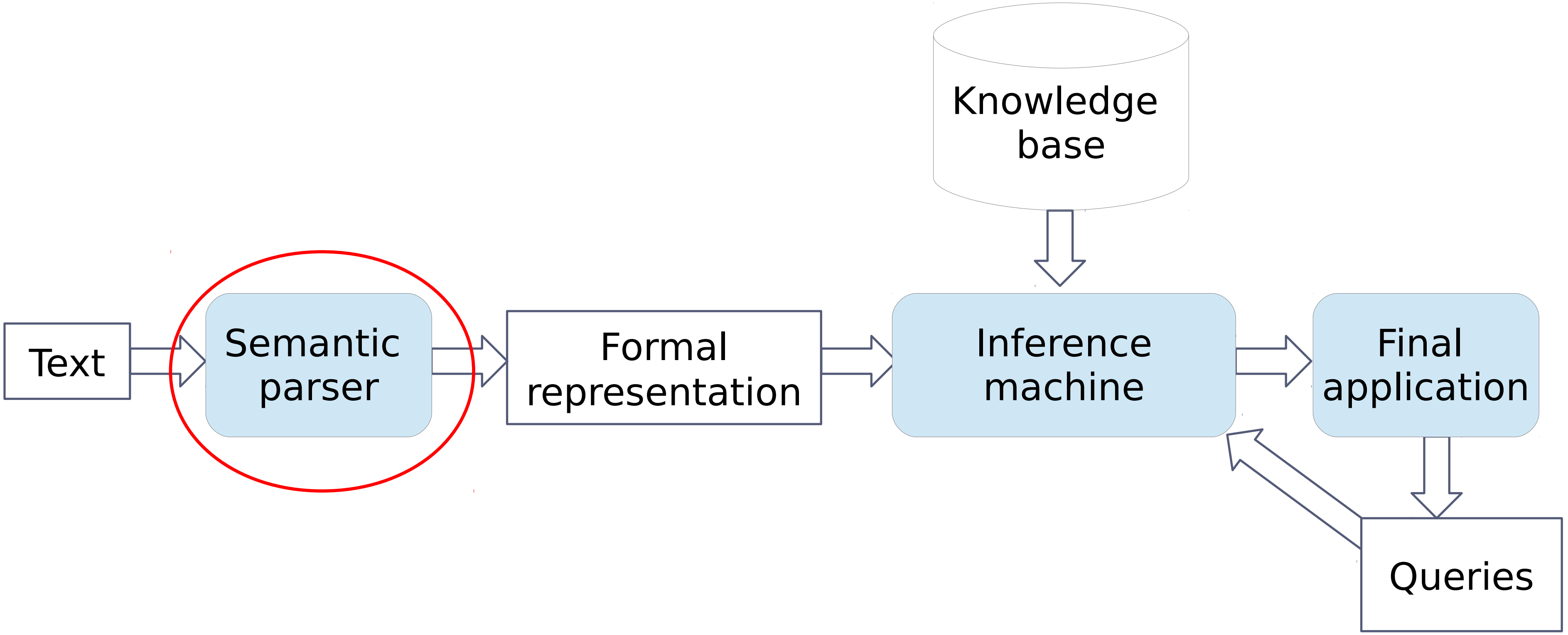
Winograd, T. (1972). *Understanding Natural Language*. Orlando, FL, USA: Academic Press, Inc.

Woods, W. A., R. Kaplan, and N. B. Webber (1972). The LUNAR sciences natural language information system: Final report. Technical Report BBN Report No. 2378, Bolt Beranek and Newman, Cambridge, Massachusetts.

Semantic parsers



Inference-based NLU pipeline

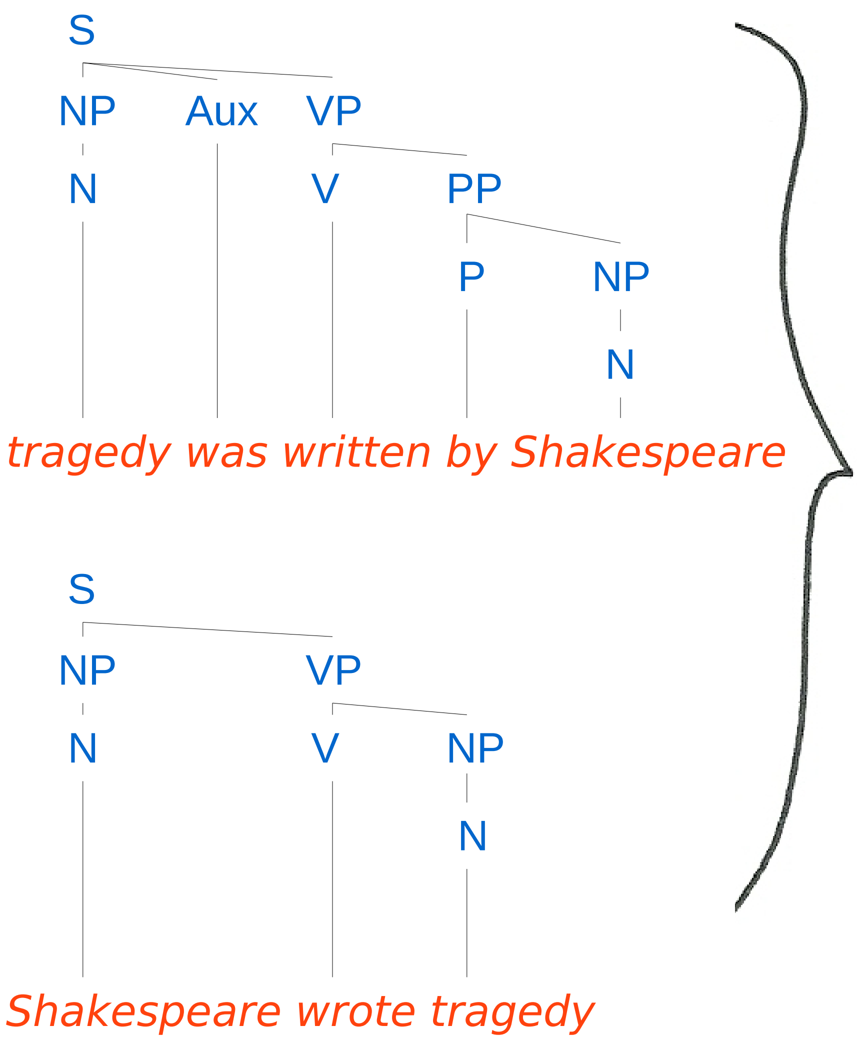


Semantic parsing

“Semantic Parsing” is an ambiguous term:

- mapping a natural language sentence to a formal representation abstracting from superficial linguistic structures (syntax)
- ...
- ...
- transforming a natural language sentence into its meaning representation

Example

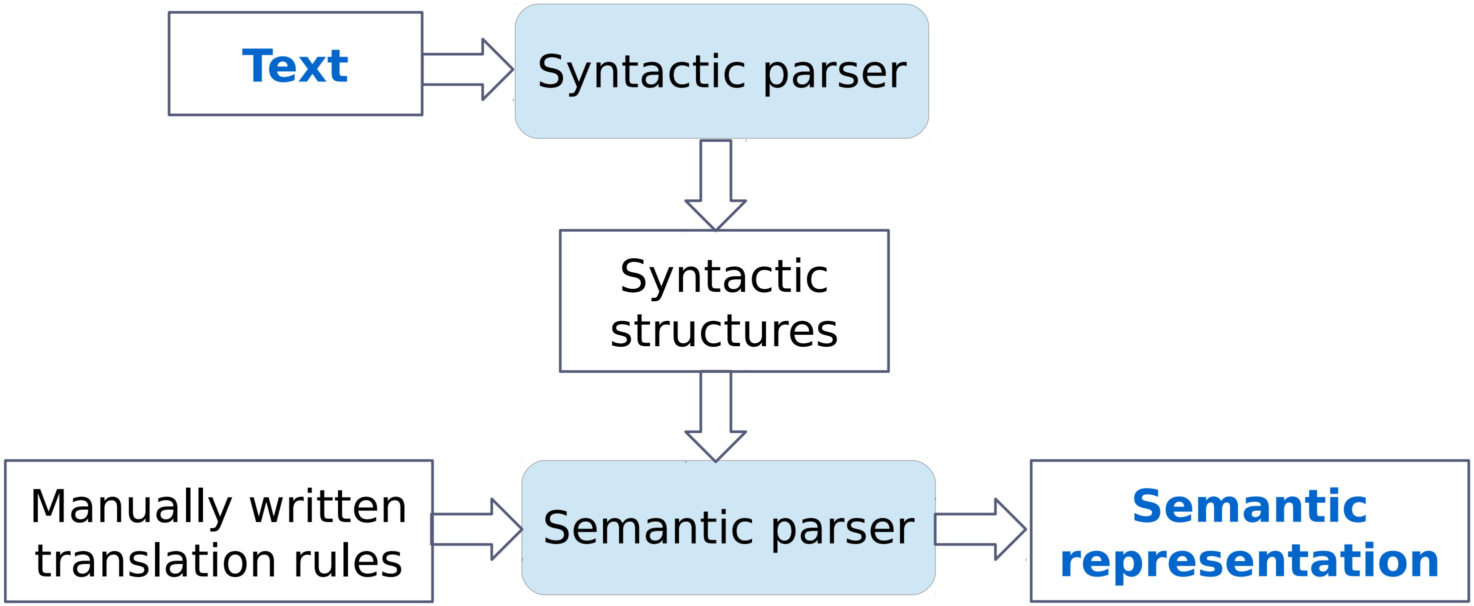


$\exists\ s, t\ (Shakespeare(s) \wedge tragedy(t) \wedge write(s,t))$

ART_CREATION [
Type: *write*
Creator: *Shakespeare,*
Work_of_art: *tragedy,*
]

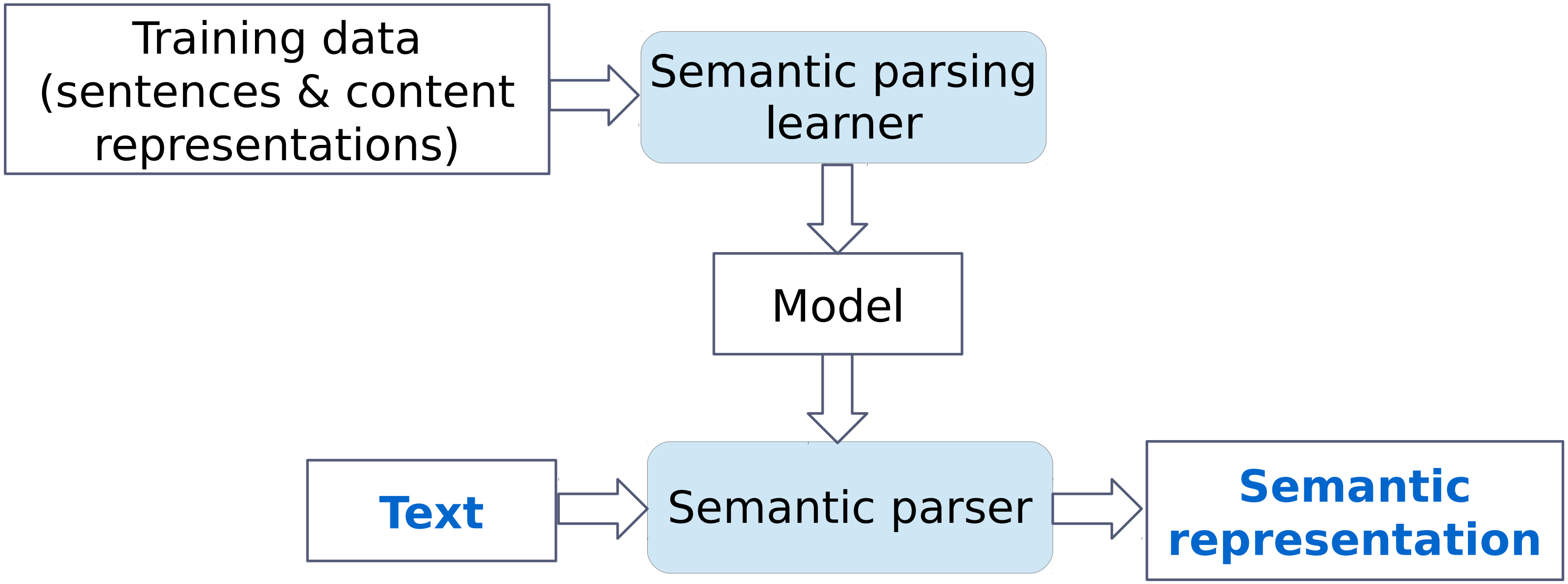
<rdf:Description
rdf:about="http://www.../Romeo&Juliet">
<cd:author>*Shakespeare*</cd:author>
<cd:type>*tragedy*</cd:play>
</rdf:Description>

Rule-based semantic parsing



Manual writing of rules
Generality

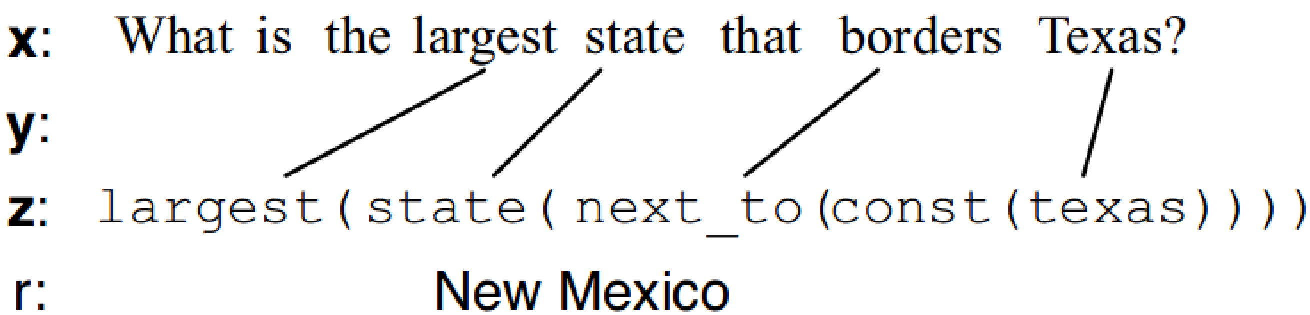
Learning semantic parsing



Lack of large training data
Domain-specific knowledge

Learning from question-answering pairs

Training on gold-standard answers (*Clarke et al., 10; Liang et al., 11; Cai&Yates, 13; Kwiatkowski et al., 13; Berant et al., 13*)



Learning from clarification dialogs

Parse harder sentences by using user interaction to break them down into simpler components through “clarification dialogs” (*Artzi&Zettlemoyer, 11*)

SYSTEM: how can I help you?

USER: I would like to fly from atlanta georgia to london england on september twenty fourth in the early evening I would like to return on october first departing from london in the late morning

SYSTEM: leaving what city?

USER: atlanta georgia

SYSTEM: leaving atlanta. going to which city?

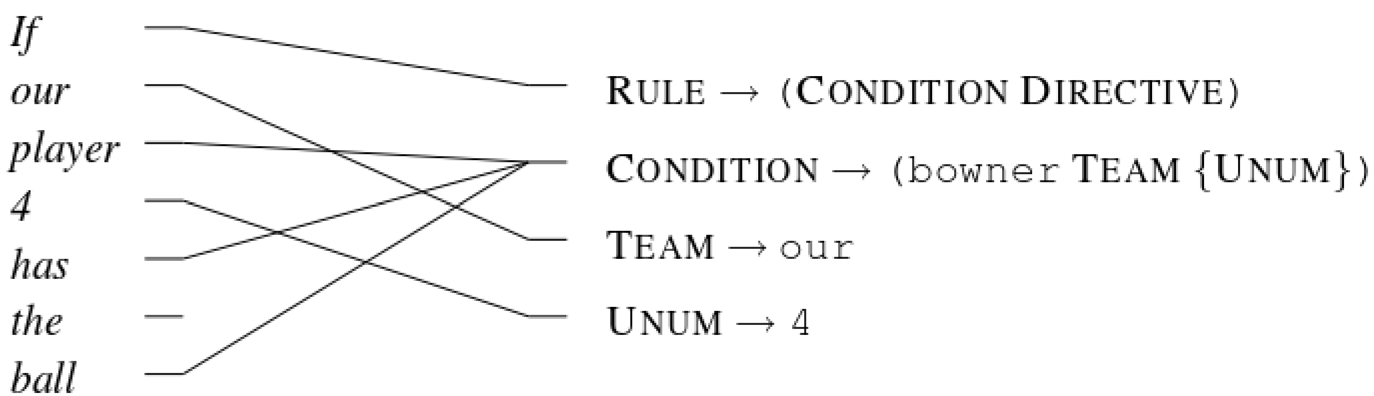
USER: london

SYSTEM: arriving in london england. what date would you like to depart atlanta?

Semantic parsing as machine translation

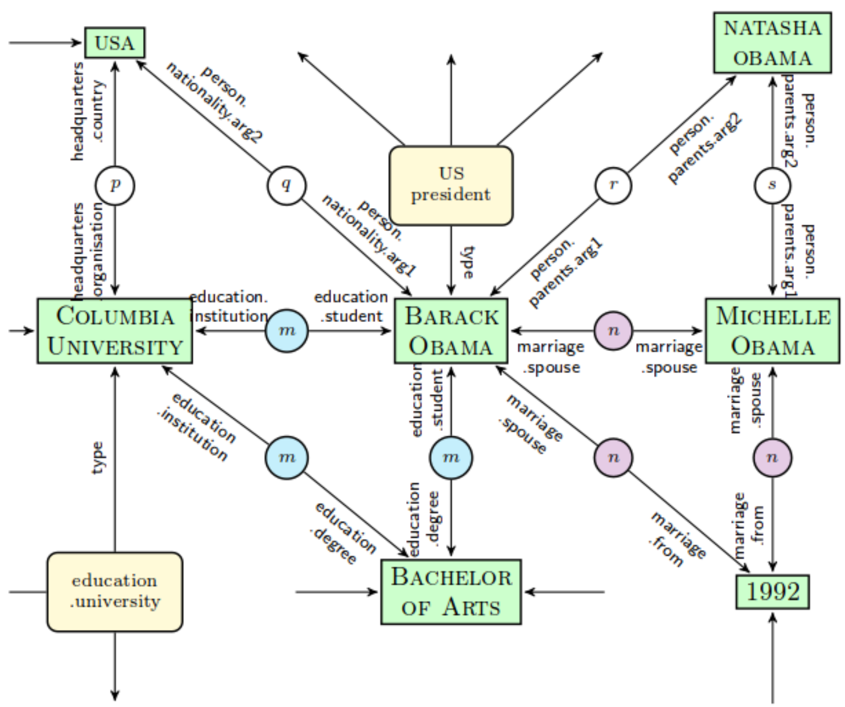
Uses machine translation techniques, e.g. word alignment

(*Wong & Mooney, 07*)



Learning using knowledge graphs

Take a parser that builds semantic representations and learn the relation between those representations and the knowledge graph *(Reddy, 14)*

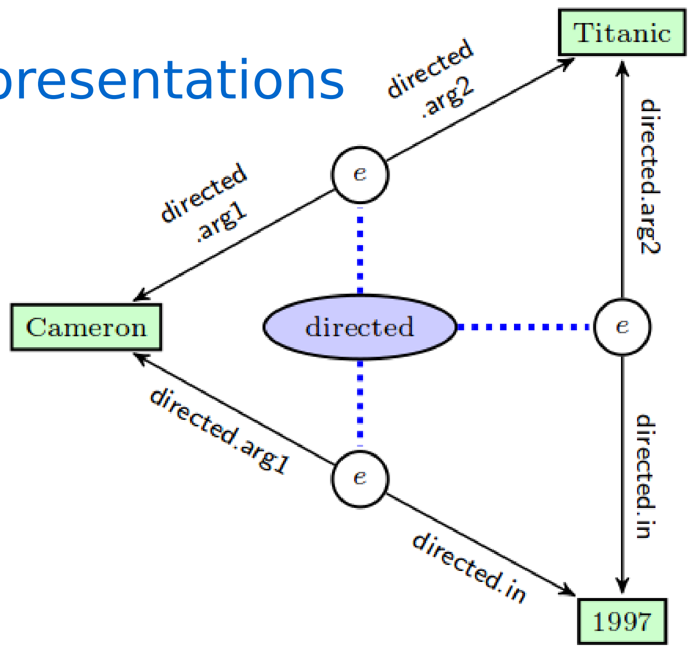


pictures are taken from Steedman's presentation at SP14

Learning using knowledge graphs

Take a parser that builds semantic representations and learn the relation between those representations and the knowledge graph *(Reddy, 14)*

Map logical representations to LF graphs



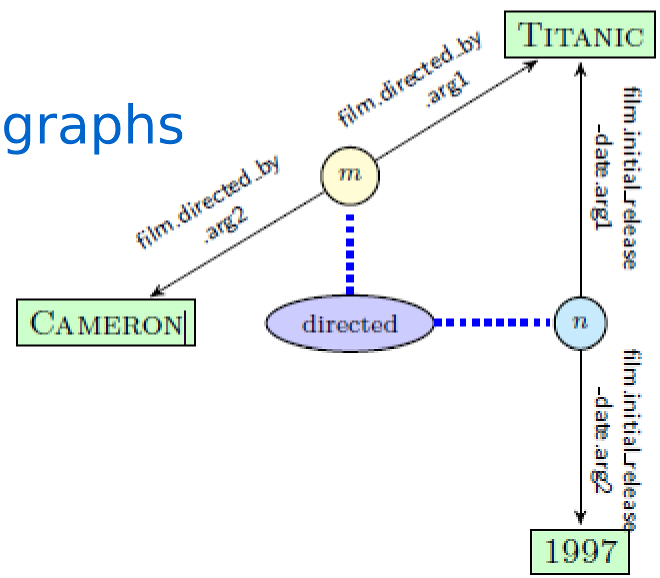
directed.arg1(e, Cameron) $\wedge$
directed.arg2(e, Titanic) $\wedge$ directed.in(e, 1997)

pictures are taken from Steedman's presentation at SP14

Learning using knowledge graphs

Take a parser that builds semantic representations and learn the relation between those representations and the knowledge graph (*Reddy, 14*)

Map LF
to knowledge graphs



film.directed_by.arg2(m, CAMERON) $\wedge$
film.directed_by.arg1(m, TITANIC) $\wedge$
film.initial_release_date.arg1(n, TITANIC) $\wedge$
film.initial_release_date.arg2(n, 1997)

pictures are taken from Steedman's presentation at SP14

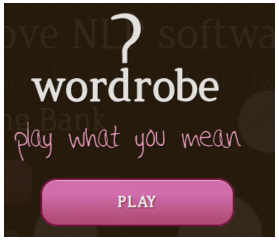
Learning from human annotations

Learn semantic parser from NL sentences paired with their respective semantic representations (*Kate & Mooney, 06)*

- Groningen Meaning Bank (*Basile et al., 12*)

  - freely available semantically annotated English corpus of currently around 1 million tokens in 7,600 documents, made up mainly of political news, country descriptions, fables and legal text.

  - populated through games for purpose



Ready-to-use parsers

- **Boxer** (http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer) - Discourse Representation Structures in FOL
- **English Slot Grammar Parser** (*http://preview.tinyurl.com/kcq68f9*) - Horn clauses
- **Epilog** (http://cs.rochester.edu/research/epilog/) - Episodic Logic
- **NL2KR** (http://nl2kr.engineering.asu.edu/) - FOL Lambda Calculus

Summary

- If you need a general semantic parser, use one of the existing rule-based tools or wait for a large annotated corpus to be released
- If you need to work in a specific domain, you can train your own parser
- To learn more about semantic parsers, see Workshop on Semantic Parsing website: *http://sp14.ws/*

References

Basile, V., J. Bos, K. Evang, N. Venhuizen (2012). A platform for collaborative semantic annotation. In *Proc. of EACL*, pp 92–96, Avignon, France.

Berant, J., A. Chou, R. Frostig and P. Liang (2013). Semantic Parsing on Freebase from Question-Answer Pairs. In *Proc. of EMNLP*. Seattle:ACL, 1533–1544.

Cai, Q. and A. Yates (2013). Semantic Parsing Freebase: Towards Open-domain Semantic Parsing. In *Second Joint Conference on Lexical and Computational Semantics*, Volume 1: Proc. of the Main Conference and the Shared Task: Semantic Textual Similarity. Atlanta: ACL, 328–338.

Clarke, J., D. Goldwasser, M.-W. Chang, and D. Roth (2010). Driving Semantic Parsing from the World's Response. In *Proc. of the 14th Conf. on Computational Natural Language Learning*. Uppsala:ACL, 18-27.

Ge, R. and R. J. Mooney (2009). Learning a compositional semantic parser using an existing syntactic parser. In *Proc. of ACL*, pp. 611-619, Suntec, Singapore.

Hirschman, L. (1992). Multi-site data collection for a spoken language corpus. In *Proc. of HLT Workshop on Speech and Natural Language*, pp. 7-14. Harriman, NY.

Kate, R. J. and R. J. Mooney (2006). Using string-kernels for learning semantic parsers. In *Proc. of COLING/ACL*, pp. 913-920, Sydney, Australia.

References

Kuhn, R., R. De Mori (1995). The application of semantic classification trees to natural language understanding. *IEEE Trans. on PAMI*, 17(5):449-460.

Kwiatkowski, T., E. Choi, Y. Artzi, and L. Zettlemoyer (2013). Scaling Semantic Parsers with On-the-Fly Ontology Matching. In *Proc. of EMNLP*. Seattle: ACL, 1545–1556.

Percy, L., M. Jordan, and D. Klein (2011). Learning Dependency-Based Compositional Semantics. In *Proc. of ACL: Human Language Technologies*. Portland, OR: ACL, 590–599.

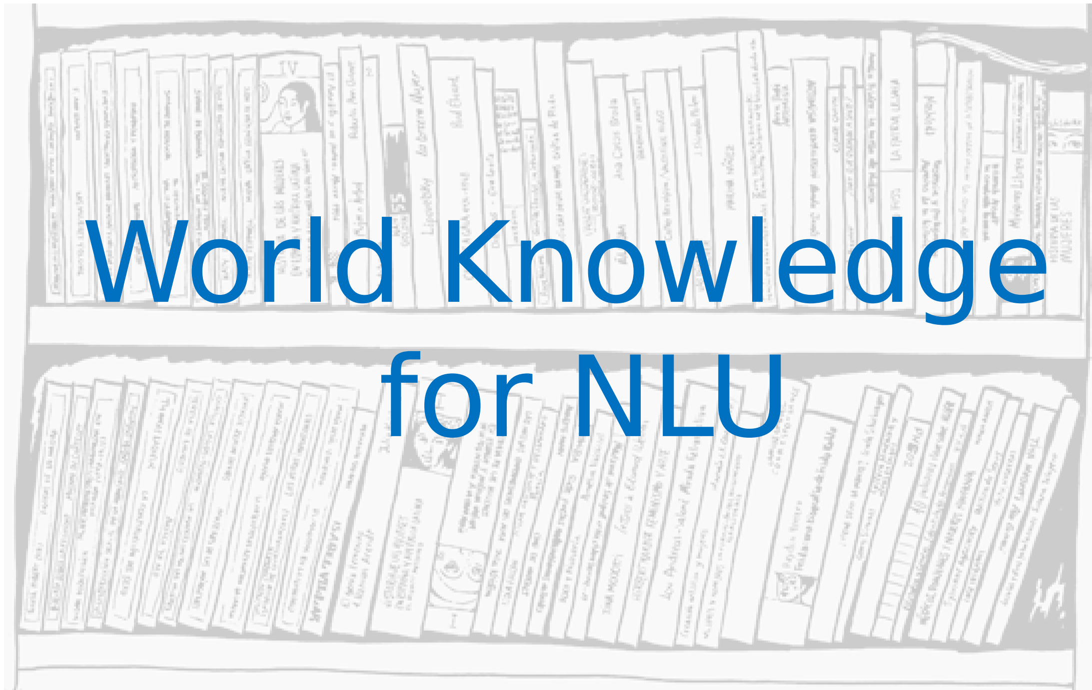
Lu, W., H. T. Ng, W. S. Lee and L. S. Zettlemoyer (2008). A generative model for parsing natural language to meaning representations. In *Proc. of EMNLP*, Waikiki, Honolulu, Hawaii.

Reddy, S. (2014). Large-scale Semantic Parsing without Question-Answer Pairs. *TACL*, subject to revisions.

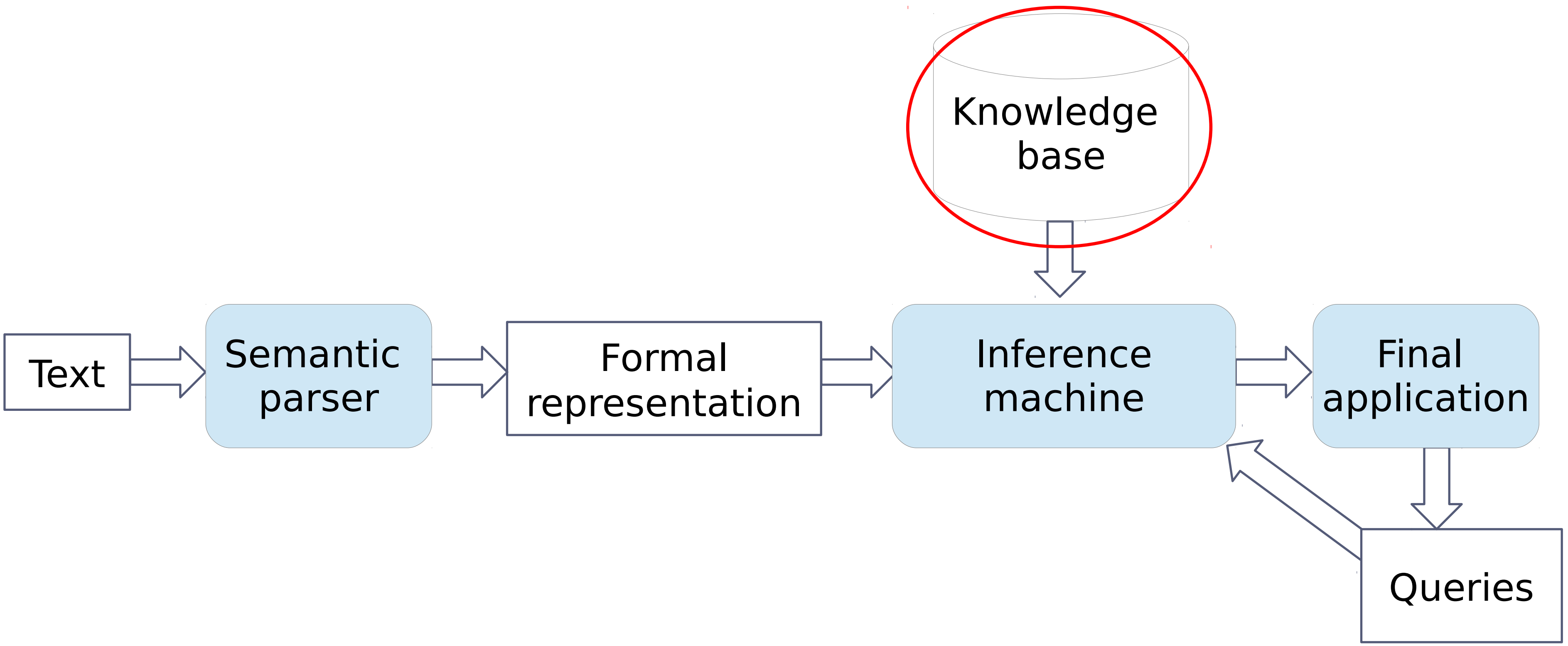
L. Zettlemoyer, M. Collins (2007). Online learning of relaxed CCG grammars for parsing to logical form. In *Proc. of EMNLP-CoNLL*, pp. 678-687. Prague, Czech Republic.

Wong, Y. W. and R. Mooney (2007). Generation by inverting a semantic parser that uses statistical machine translation. In *Proc. of NAACL-HLT*, pp. 172-179. Rochester, NY.

World Knowledge for NLU



Inference-based NLU pipeline



A bit of history

- Interest to model world knowledge arose in AI in the late 1960s (*Quillian, 68; Minsky, 75; Bobrow et al., 77; Woods et al., 80*)
- Later, two lines of research developed:

  - "clean" theory based KBs, efficient reasoning, sufficient conceptual coverage (*ontologies*)

  - KBs based on words instead of artificial concepts, result from corpus studies and psycholinguistic experiments (*lexical-semantic dictionaries*)
- Starting from the 1990s, progress of the statistical approaches allowed to learn knowledge from corpora automatically
- In the 2000s, global spread of the Internet facilitated community-based development of knowledge resources

Lexical-semantic dictionaries

- Words are linked to a set of **word senses**, which are united into groups of semantically similar senses.
- Different types of **semantic relations** are then defined on such groups, e.g., taxonomic, part-whole, causal, etc.
- Resources are created **manually** based on corpus annotation, psycholinguistic experiments, and dictionary comparison.

WordNet family

(*http://www.globalwordnet.org/, http://wordnet.princeton.edu/*)

- Network-like structure
- Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms called *synsets*
- Semantic relations defined between synsets
- English WN:

POS	Unique words/phrases	Synsets	Word-synset pairs
Nouns	117798	82115	146312
Verbs	11529	13767	25047
Adjectives	21479	18156	30002
Adverbs	4481	3621	5580
Total	155287	117659	206941


WordNet Search - 3.1
- WordNet home page - Glossary - Help
Word to search for: tragedy
Search WordNet
Display Options: (Select option to change)
Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"
Noun
S: (n) calamity, catastrophe, disaster, tragedy, cataclysm (an event resulting in great loss and misfortune) "the whole city was affected by the irremediable calamity"; "the earthquake was a disaster"
direct hyponym / full hyponym
direct hypernym / inherited hypernym / sister term
derivationally related form
S: (n) tragedy (drama in which the protagonist is overcome by some superior force or circumstance; excites terror or pity)
direct hyponym / full hyponym
direct hypernym / inherited hypernym / sister term
antonym
derivationally related form

Usage of WordNet

Usage (*Morato et al., 04; http://wordnet.princeton.edu/wordnet/related-projects/*):

- word sense disambiguation (training using WN-annotated corpora)
- computing semantic similarity
- simple inference with semantic relations
- deriving concept axiomatization from synset definitions (e.g. Extended WordNet, *http://www.hlt.utdallas.edu/~xwn/about.html*)
- ...

Criticism:

- word sense distinctions are too fine-grained (*Agirre&Lacalle, 03*)
- no conceptual consistency (*Oltramari et al., 02*)
- semantic relations between synsets with the same POS

Nevertheless:

- Huge lexical and conceptual coverage
- Simple structure, easy to use (Prolog format)
- The most popular resource so far!

FrameNet family

(*https://framenet.icsi.berkeley.edu*)

- based on Fillmore's frame semantics (*Fillmore, 68*)
- meaning of predicates is expressed in terms of **frames**, which describe prototypical situations spoken about in natural language
- frame contains a set of **roles** corresponding to the participants of the described situation
- **frame relations** defined on frames
- based on **annotating examples** of how words are used in actual texts
- English FN:

POS	Lexical units	Frames	Frame relations
Nouns	5206		
Verbs	4998		
Adjectives	2271		
Other POS	390		
Total	12865	1182	1755

Giving

Lexical Unit Index

Definition:

A Donor transfers a Theme from a Donorto a Recipient. This frame includes only actions that are initiated by the Donor (the one that starts out owning the Theme). Sentences (even metaphorical ones) must meet the following entailments: the Donor first has possession of the Theme. Following the transfer the Donor no longer has the Theme and the Recipient does.

> Barney GAVE the beer to Moe.

> $300 was ENDOWED to the universityto build a new performing arts building.

Usage of FrameNet

Usage (*https://framenet.icsi.berkeley.edu/fndrupal/framenet_users*):

- semantic role labeling (*https://framenet.icsi.berkeley.edu/fndrupal/ASRL*)
- word sense disambiguation
- question answering
- recognizing textual entailment
- ...

Criticism:

- low coverage (*Shen and Lapata, 07; Cao et al., 08*)
- no axiomatization of frame relations (*Ovchinnikova et al., 10*)
- complicated format

Solutions:

- Automatic extension of lexical coverage (*Burchardt et al., 05; Cao et al., 08*)
- ontology-based axiomatization (*Ovchinnikova et al., 10*)

Ontologies

The term “ontology” (originating in philosophy) is ambiguous:

- theory about how to model the world
  “An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world” (*Guarino, 98*)

- specific world models
  “an ontology is an explicit specification of a conceptualization” (*Gruber, 93*)

Ontology Modeling

Ontologies are intended to represent **one particular view** of the modeled domain in an **unambiguous** and **well-defined** way.

- usually do not tolerate inconsistencies and ambiguities
- provide valid inferences
- are much closer to “scientific” theories than to fuzzy common sense knowledge

Ontology Representation

- Complex knowledge representation

$\forall i\ (Pacific\_Island(i) \rightarrow Island(i) \wedge \exists o(Ocean(o) \wedge locatedIn(i, o)))$

- Most of the ontology representation languages are based on logical formalisms *(Bruijn, 03)*
- Trade-off between expressivity and complexity

Interface between Ontologies and Lexicons

In order to be used in an NLU application, ontologies need to have an interface to a natural language lexicon.

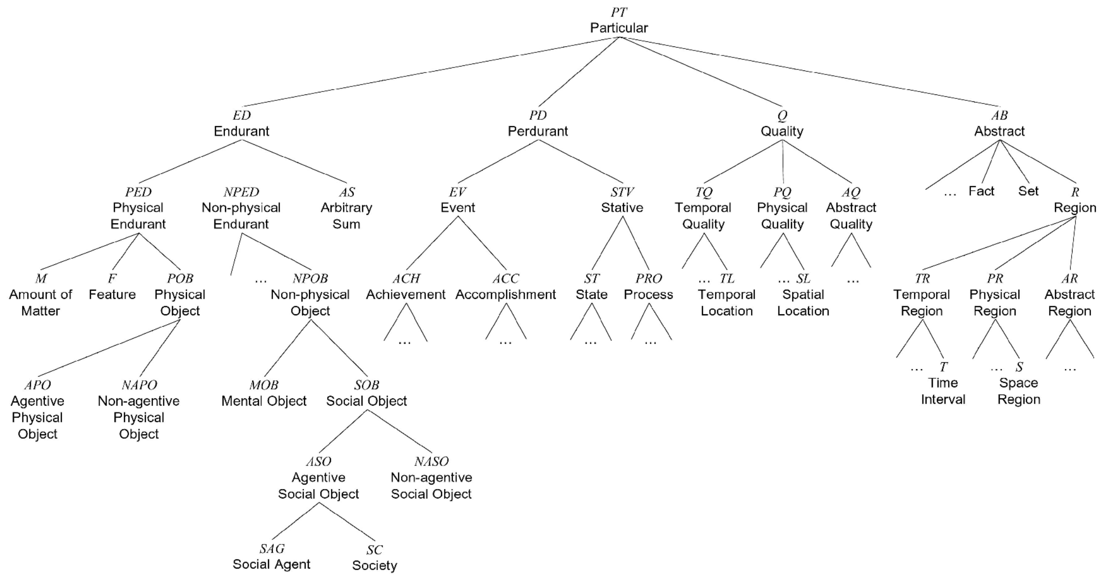
Methods of interfacing (*Prevot et al., 05*):

- Restructuring a computational lexicon on the basis of ontological-driven principles
- Populating an ontology with lexical information
- Aligning an ontology and a lexical resource

Expert-developed ontologies

DOLCE (*http://www.loa.istc.cnr.it/old/DOLCE.html*) - aims at capturing the upper ontological categories underlying natural language and human common sense.

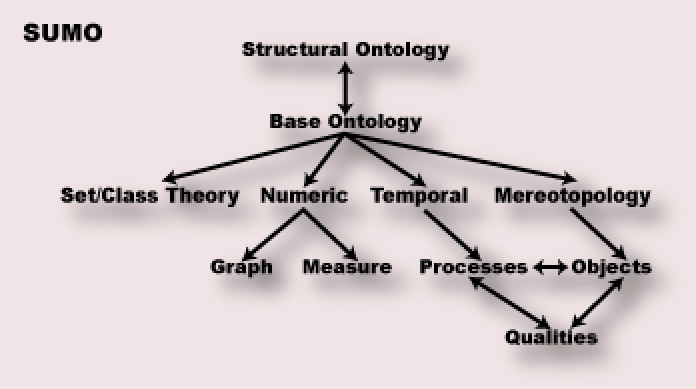
- conceptually sound and explicit about its ontological choices
- no interface to lexicon
- used for interfacing domain-specific ontologies



Expert-developed ontologies

SUMO (*http://www.ontologyportal.org/*) - is an integrative database created “by merging publicly available ontological content into a single structure”

- has been criticized for messy conceptualization (*Oberle et al., 2007*)
- linked to the WordNet lexicon (*Niles et al., 2003*)
- used by a couple of QA systems (*Harabagiu et al., 2005; Suchanek, 2008*)



Expert-developed ontologies

Extensive development of domain-specific ontologies was stimulated by the progress of **Semantic Web**

- knowledge representation standards (e.g., OWL)
- reasoning tools mostly based on Description Logics (*Baader et al., 03*)

NLU applications that employ reasoning with domain ontologies:

- information retrieval (*Andreasen&Nilsson, 04; Buitelaar&Siegel, 06*)
- question answering (*Mollá &Vicedo, 07*)
- dialog systems (*Estival et al., 04*)
- automatic summarization (*Morales et al., 08*)

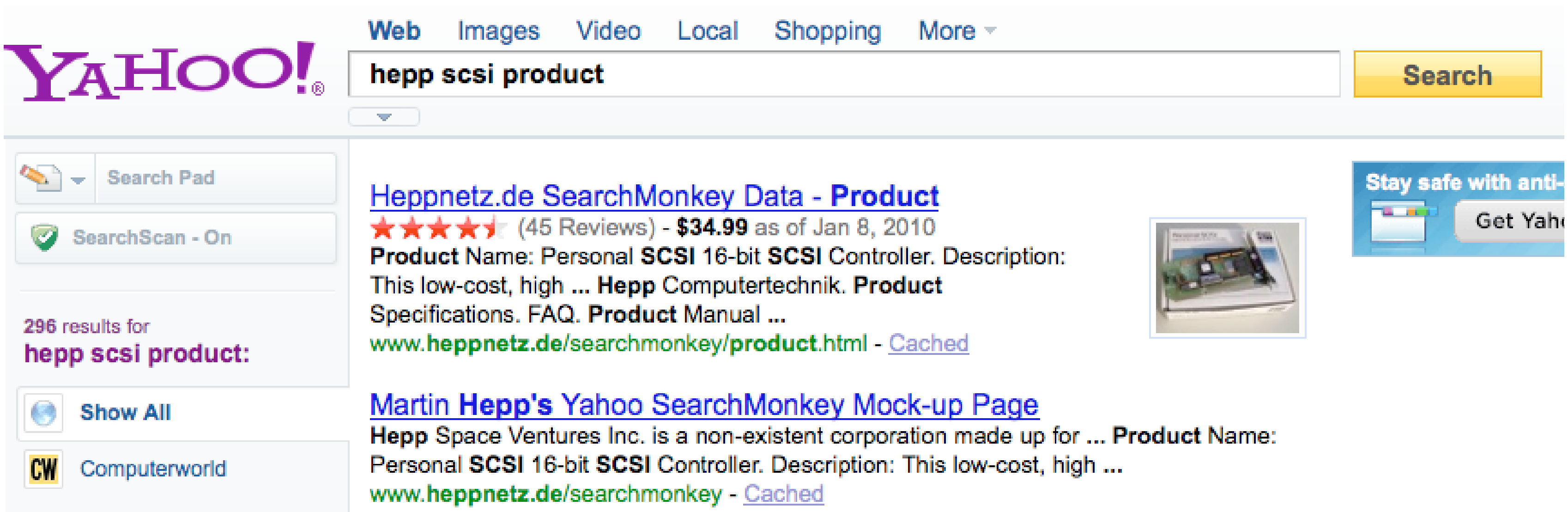
However, the full power of OWL ontologies is hardly used in NLU (*Lehmann&Völker, 14*)

- low coverage
- lack of links to lexicon
- no need for expressive knowledge (yet!)

Expert-developed ontologies

GoodRelations (*http://www.heppnetz.de/projects/goodrelations/*) - is a lightweight ontology for annotating offerings and other aspects of e-commerce on the Web.

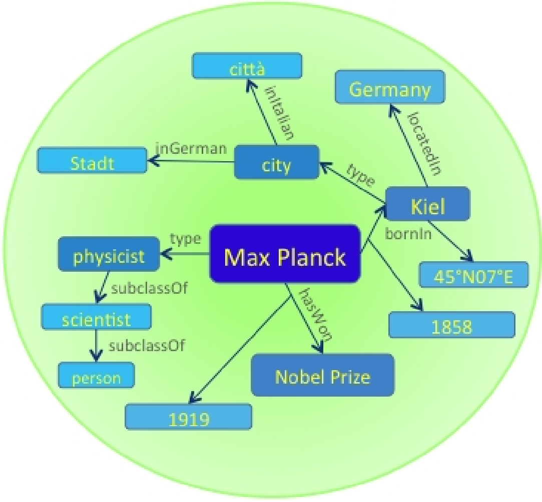
- used by Google, Yahoo!, BestBuy, sears.com, kmart.com, … to provide rich snippets



Community-developed ontologies

YAGO (*www.mpi-inf.mpg.de/yago/*) - is a KB derived from Wikipedia, WordNet, and Geonames

- 10 million entities (persons, organizations, cities, etc.), 120 million facts about these entities, 350 000 classes
- attaches a temporal and spacial dimensions to facts
- contains a taxonomy as well as domains (e.g. "music" or "science")



Used by Watson
and many other NLU systems,
facilitates Freebase and DBPedia

Community-developed ontologies

Freebase (*http://www.freebase.com/*) - is a community-curated database of well-known people, places, and things.

- 1B+ facts, 40M+ topics, 2k+ types
- data derived from Wikipedia and added by users
- A source of Google's Knowledge Graph
- provides search API
- geosearch

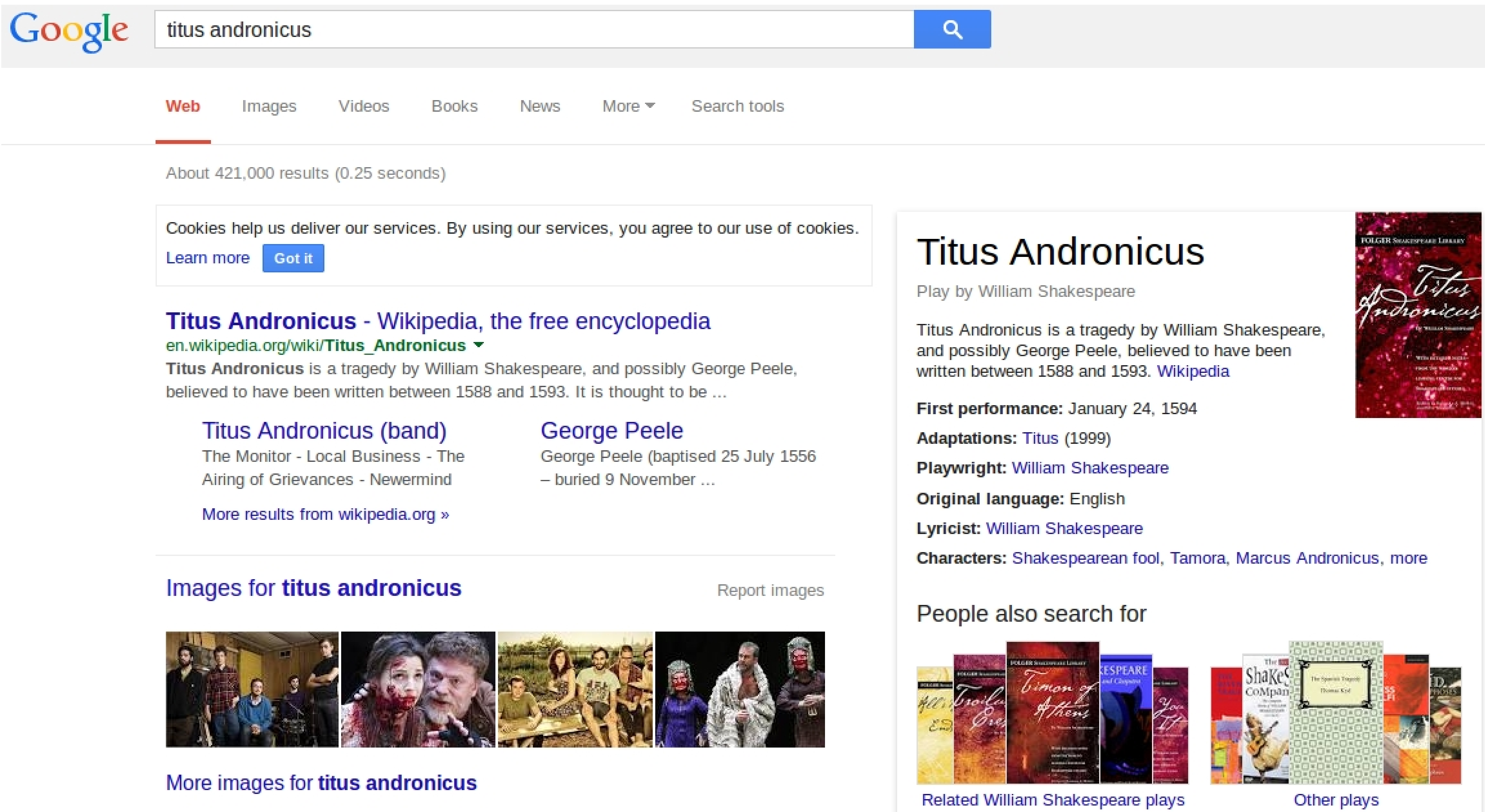
Community-developed ontologies

Google Knowledge Graph a knowledge base used by Google to enhance its search engine.

- data derived from CIA World Factbook, Freebase, and Wikipedia

Community-developed ontologies

Google Knowledge Graph a knowledge base used by Google to enhance its search engine.



Community-developed ontologies

Google Knowledge Graph a knowledge base used by Google to enhance its search engine.

Extracting knowledge from corpora

- The Distributional Hypothesis: "You shall know a word by the company it keeps" (*Firth, 57*)
- Two forms are similar if these are found in similar contexts
- Types of contexts:

  - context window
  - document
  - syntactic structure

Two useful ideas:

- patterns (*Hearst, 92*)

  dogs, cats <u>and other</u> animals
  malaria infection <u>results in</u> the death ...

- pointwise mutual information (*Church&Hanks, 90*)

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

What we can learn

- Semantic/ontological relations between nouns (*Hearst, 92; Girju et al., 07; Navigli et al., 11*)
  dog <u>is_a</u> animal, Shakespeare <u>instance_of</u> playwright, branch <u>part_of</u> tree

- Verb relations, e.g., causal and temporal (*Kozareva, 12*)
  chemotherapy <u>causes</u> tumors to shrink

- Selectional preferences (*Resnik, 96; Schulte im Walde, 10*)
  people <u>fly to</u> cities

- Paraphrases (*Lin&Pantel, 01*)
  X writes Y - X is the author of Y

- Entailment rules (*Berant et al., 11*)
  X killed Y → Y died

- Narrative event chains (*Chambers&Jurafsky, 09*)
  X arrest, X charge, X raid, X seize, X confiscate, X detain, X deport

What we cannot learn yet

- Relations between abstract concepts/words
  idea, shape, relation
- Negation, quantification, modality
  X is independent → there is nothing X depends on
- Complex concept definitions
  Space - a continuous area or expanse which is free, available, or unoccupied

  but see (*Völker et al., 07*)
- Abstract knowledge
  X blocks Y → X causes some action by Y not being performed

Available large corpora

- English **Gigaword** (*https://catalog.ldc.upenn.edu/LDC2011T07*)
  10-million English documents from seven news outlets
- **ClueWeb** '09, '12 (*http://lemurproject.org/clueweb09/, http://www.lemurproject.org/clueweb12.php/*)
  - '09: 1 billion web pages, in 10 languages
  - '12: 733 million documents
- **Google ngram corpus** (*http://storage.googleapis.com/books/ngrams/books/datasetsv2.html*)
  3.5 million English books containing about 345 billion words, parsed, tagged and frequency counted
- **Wikipedia** dumps (*http://dumps.wikimedia.org/*)
  4.5 million articles in 287 languages
- **Spinn3r** Dataset (*http://www.icwsm.org/data/*)
  386 million blog posts, news articles, classifieds, forum posts and social media content

Some useful resources learned automatically

- **VerbOcean**: verb-based paraphrases (*http://demo.patrickpantel.com/demos/verbocean/*)
  X outrage Y happens-after/is stronger than X shock Y
- **wikiRules**: lexical reference rules (*http://u.cs.biu.ac.il/~nlp/resources/downloads/lexical-reference-rules-from-wikipedia*)
  Bentley –> luxury car, physician –> medicine, Abbey Road –> The Beatles
- **Reverb** (*http://reverb.cs.washington.edu/*): binary relationships
  Cabbage also contains significant amounts of Vitamin A
- **Proposition stores** (*http://colo-vm19.isi.edu/#/*)
  subj_verb_dirobj people prevent-VB tragedy-NN
- Database of **factoids mined by KNEXT** (*http://cs.rochester.edu/research/knext/*)
  A tragedy can be horrible [⟨det tragedy.n⟩ horrible.a]

World knowledge resources

	Lexical-semantic dictionaries	Expert-developed ontologies	Community-developed ontologies	Corpora
knowledge obtained	manually	manually	manually	automatically
relations defined on	word senses	concepts	concepts	words
language-dependence	yes	no	no	yes
domain-dependence	no	yes/no	yes/no	yes/no
structure	simple	complex	simple	simple
coverage	small	small	large	large
consistency	no (defeasible)	yes	yes	no (defeasible)
examples	WordNet, FrameNet, VerbNet	SUMO, Cyc, DOLCE, GoodRelations	YAGO, Freebase, GoogleGraph	Gigaword, Clueweb, Google ngram corpus

Knowledge resources at work

Recognizing Textual Entailment resources:

http://www.aclweb.org/aclwiki/index.php?title=RTE_Knowledge_Resources

Resource	Type	Author	Brief description	PAST Users <ref name:"rtethree">RTE-3 data have been provided by participants by means of a questionnaire.</ref>	RTE4 Users<ref name:"rtefour">RTE-4 data have been provided by participants and have been integrated with information extracted from the related proceedings.</ref>	RTE5 Users<ref name:"rtefive">RTE-5 data have been provided by participants and have been integrated with information extracted from the related proceedings.</ref>	RTE6 Users<ref name:"rtesix">RTE-6 data have been provided by participants and have been integrated with information extracted from the related proceedings.</ref>
WordNet	Lexical DB	Princeton University	Lexical database of English nouns, verbs, adjectives and adverbs	3	21	18	22
VerbOcean	Lexical DB	Information Sciences Institute, University of Southern California	Broad-coverage semantic network of verbs	2	3	6	7
Wikipedia	Encyclopedia		Free encyclopedia. Used for extraction of lexical-semantic rules (from its more structured parts), named entity recognition, geographical information etc.	0	3	6	6
DIRT Paraphrase Collection	Collection of paraphrases	University of Alberta	DIRT (Discovery of Inference Rules from Text) is both an algorithm and a resulting knowledge collection. The DIRT knowledge collection is the output of the DIRT algorithm over a 1GB set of newspaper text.	2	4	3	4
FrameNet	Lexical DB	ICSI (International Computer Science Institute) - Berkley University	Lexical resource for English words, based on frame semantics (valences) and supported by corpus evidence	1	1	2	3

Summary

- What NLU needs and can provide right now:
  - defeasible knowledge bases
  - with simple structure
  - and high coverage
- Most useful resources so far:
  - large lexical-semantic dictionaries (WordNet)
  - community-curated knowledge graphs
- Large-scale NLU currently neither uses nor provides expressive ontologies
- Note: resources of different types can be successfully used in combination (*Ovchinnikova, 12*)

References

Agirre, E. and O. L. D. Lacalle (2003). Clustering WordNet word senses. In *Proc. of the Conference on Recent Advances on Natural Language*, 121–130.

Andreasen, T. and J. F. Nilsson (2004). Grammatical specification of domain ontologies. *Data and Knowledge Engineering* 48, 221–230.

Baader, F., D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider (Eds.) (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. NY: Cambridge University Press.

Berant, J., Dagan, I., and Goldberger, J. (2011). Global learning of typed entailment rules. In Proc. of ACL, 610-619.

Bobrow, D., R. Kaplan, M. Kay, D. Norman, H. Thompson, and T. Winograd (1977). GUS, A Frame-Driven Dialogue System. *Artificial Intelligence* 8, 155–173.

Buitelaar, P. andM. Siegel (2006). Ontology-based Information Extraction with SOBA. In *Proc. of LREC*, pp. 2321–2324.

Burchardt, A., K. Erk, and A. Frank (2005). A WordNet Detour to FrameNet. In B. Fisseni, H.-C. Schmitz, B. Schrder, and P. Wagner (Eds.), *Sprachtechnologie*, mobile. Kommunikation und linguistische Resourcen, Frankfurt am Main, pp. 16. Lang, Peter.

Cao, D. D., D. Croce, M. Pennacchiotti, and R. Basili (2008). Combining word sense and usage for modeling frame semantics. In *Proc. of the Semantics in Text Processing Conference*, 85–101.

References

Chambers, N. and D. Jurafsky (2009). Unsupervised learning of narrative schemas and their participants. In *Proc. of ACL*.

Church, K. W. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist*. 16 (1): 22–29.

Estival, D., C. Nowak, and A. Zschorn (2004). Towards Ontology-based Natural Language Processing. In *Proc. of the 4th Workshop on NLP and XML: RDF/RDFS and OWL in Language Technology*, 59–66. Markowski.

Fillmore, C. (1968). The case for case. In E. Bach and R. Harms (Eds.), *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston.

Girju, R., Nakov, P., Nastaste, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). SemEval-2007 task 04: Classification of semantic relations between nominals. In Proc. Of SemEval 2007.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2), 199–220.

Guarino, N. (1998). Formal ontology and information systems. In *Proc. of the International Conference on Formal Ontologies in Information Systems*, 3–15. Amsterdam, IOS Press.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th conference on Computational linguistics*, 539–545.

Kozareva, Z. (2012) Learning Verbs on the Fly. In Proc. Of COLING, 599-609.

References

Lehmann, J., and J. Völker(Eds.). (2014). Perspectives on Ontology Learning. Akademische Verlagsgesellschaft AKA.

Lin, D. and P. Pantel (2001). Discovery of inference rules for question-answering. *Natural Language Engineering* 7 (4), 343–360.

Mollá, D. and J. L. Vicedo (2007). Question answering in restricted domains: An overview. *Computational Linguistics* 33, 41–61.

Morales, L. P., A. D. Esteban, and P. Gerv´as (2008). Concept-graph based biomedical automatic summarization using ontologies. In Proc. ofTextGraphs, Morristown, NJ, USA, 53–56. ACL.

Morato, J., M. N. Marzal, J. Llorns, and J. Moreiro (2004). Wordnet applications. In *Proc. of the Global Wordnet Conference*, Brno, Czech Republic.

Navigli, R., P. Velardi, and S. Faralli (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proc. of IJCAI*, 1872–1877.

Oltramari, A., A. Gangemi, N. Guarino, and C. Masolo (2002). Restructuring WordNet's top-level: The OntoClean approach. In *Proc. of OntoLex*, 17–26.

Ovchinnikova, E. (2012) *Integration of World Knowledge for Natural Language Understanding*, Atlantis Press, Springer.

Ovchinnikova, E., L. Vieu, A. Oltramari, S. Borgo, and T. Alexandrov (2010). Data-Driven and Ontological Analysis of FrameNet for Natural Language Reasoning. In *Proc. of LREC*.

References

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1-2):127 – 159

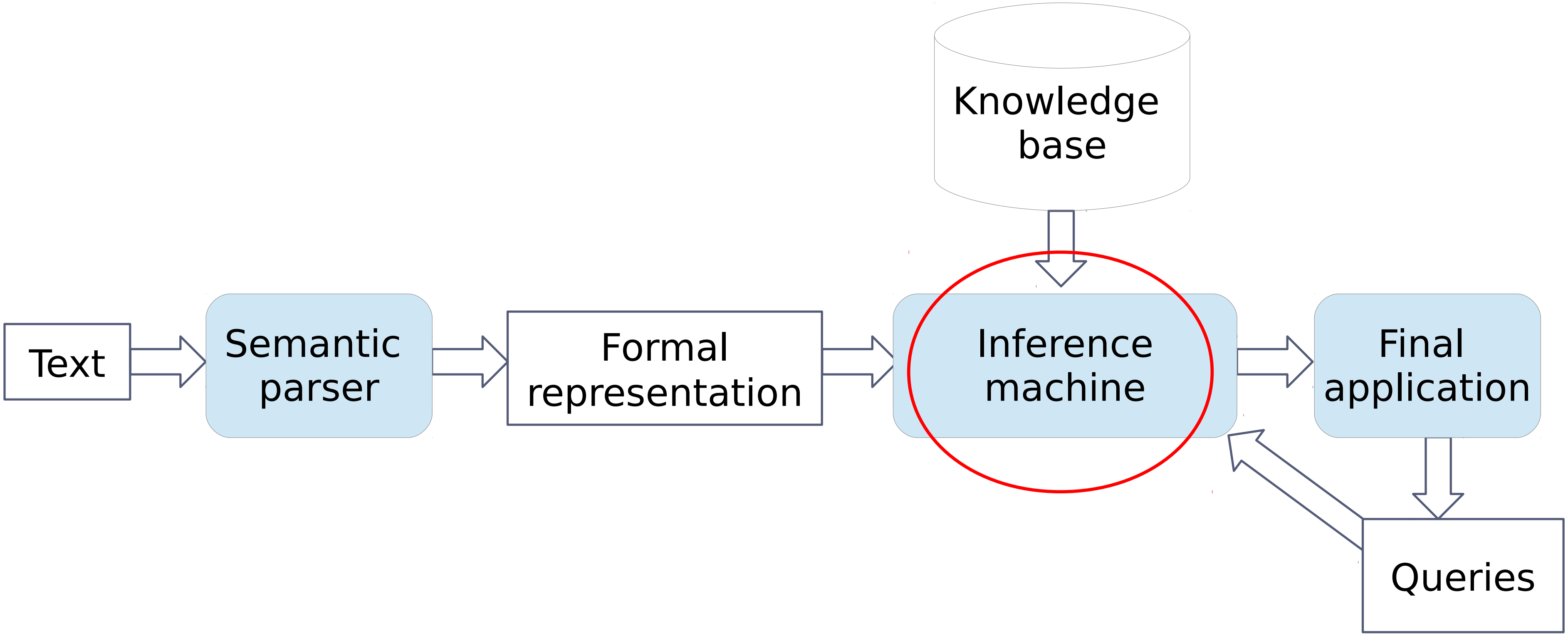
Schulte im Walde, S. (2010). Comparing Computational Approaches to Selectional Preferences – Second-Order Co-Occurrence vs. Latent Semantic Clusters. In *Proc. of LREC*.

Shen, D. andM. Lapata (2007). Using Semantic Roles to Improve Question Answering. In Proc. of EMNLP, 12–21.

Völker, J., P. Hitzler, and P. Cimiano. Acquisition of owl dl axioms from lexical resources. In *Proc. of ESWC*, 670–685.

Reasoning for NLU

Inference-based NLU pipeline



Inference

- the process of deriving conclusions from premises known or assumed to be true.

Symbolic – knowledge is encoded in the form of verbal rules

Theorem provers

Expert systems

Constraint solvers

Support Vector Machines

Neural networks

Sub-symbolic – knowledge is encoded as a set of numerical patterns

Logical inference for NLU

Deduction is valid logical inference.

If X is true, what else is true?

$\forall x(p(x) \rightarrow q(x))$	*Dogs are animals.*
$p(A)$	*Pluto is a dog.*
$q(A)$	*Pluto is an animal.*

Abduction is inference to the best explanation.

If X is true, why is it true?

$\forall x(p(x) \rightarrow q(x))$	*If it rains then the grass is wet.*
$q(A)$	*The grass is wet.*
$p(A)$	*It rains.*

Deduction for NLU

- The idea of applying deduction to NLU originated in the context of question answering (*Black, 64; Green&Raphael, 68*) and story understanding (*Winograd, 72; Charniak, 72*).

- Two main directions (*Gardent&Webber, 01*):

  - check satisfiability (*Bos, 09*)

  - build models (*Bos, 03; Cimiano, 03*)

Satisfiability check

Filter out unwanted interpretations (*Bos, 09*)

The dog ate the bone. It was hungry.

Two interpretations:

$\exists d, b, e\ (dog(d) \land eat(e,d,b) \land hungry(d))$ The dog was hungry.

~~$\exists d, b, e\ (dog(d) \land eat(e,d,b) \land hungry(b))$ The bone was hungry.~~

Knowledge:

$\forall x(hungry(x) \rightarrow living\_being(x))$ Only living beings can be hungry.

$\forall d(dog(d) \rightarrow living\_being(d))$ Dogs are living beings.

$\forall b(bone(d) \rightarrow \neg\ living\_being(b))$ Bones are not living beings.

Model building

- More specific representation is constructed in the course of proving the underspecified one (*Bos, 03; Cimiano, 03*)
- Model builder - a program that takes a set of logical formulas Φ and tries to build a model that satisfies Φ.
- Consistency check “for free”
- Minimal models are favored

Model building

John saw the house. The door was open.

Logical representation:

$\exists\, j, s, h, e, d\ (John(j) \wedge see(e,j,h) \wedge house(h) \wedge door(d) \wedge open(d))$

Knowledge:

$\forall x(house(x) \rightarrow \exists\, d(\ door(d) \wedge part\_of(y,x))$ Houses have doors.

Two models:

M1 = $\{John(J), see(E,J,H) \wedge house(H) \wedge has\_part(H,D1) \wedge door(D1) \wedge door(D2) \wedge open(D2)\}$

M2 = $\{John(J), see(E,J,H) \wedge house(H) \wedge has\_part(H,D) \wedge door(D) \wedge open(D)\}$

Theorem provers

Nice comparison of existing theorem provers available at

http://en.wikipedia.org/wiki/Automated_theorem_prover

Applications of theorem proving to NLU

- Dialog systems *(Bos, 09)*
- Recognizing textual entailment *(Bos&Markert, 06; Tatu&Moldovan, 07)*

Problems

- Unable to choose between alternative interpretations if both are consistent
- Model minimality criteria is problematic
- Unable to reason with inconsistent knowledge
- If a piece of knowledge is missing, fails to find a proof
- Unlimited inference chains
- Reasoning is computationally complex

Markov Logic Networks

- First-order inference in a probabilistic way
- FOL formulas are assigned weights
- An instantiation of Markov Network, where logical formulas determine the network structure
- MLN – template for constructing Markov Network

(*Richardson and Domingos, 2006*)

Markov Logic Networks

A Markov Logic Network L is a set of pairs (F_i,w_i), where F_i is a formula in FOL and wi is a real number. Together with a finite set of constraints $C=\{c_1,..,c_n\}$ it defines a Markov Network $M_{L,C}$ as follows:

- $M_{L,C}$ contains one binary node for each possible grounding of each predicate occurring in L. The value of the node is 1 if grounding is true, and 0 otherwise.
- $M_{L,C}$ contains one feature for each possible grounding of each formula F_i in L. The value of this feature is 1 if th ground formula is true, and 0 otherwise. The weight of the feature is w_i.

Probability distribution

$$P(x) = \frac{1}{Z} \exp\left(\sum_i \boxed{w_i} \boxed{n_i(x)} \right)$$

Weight of formula *i*

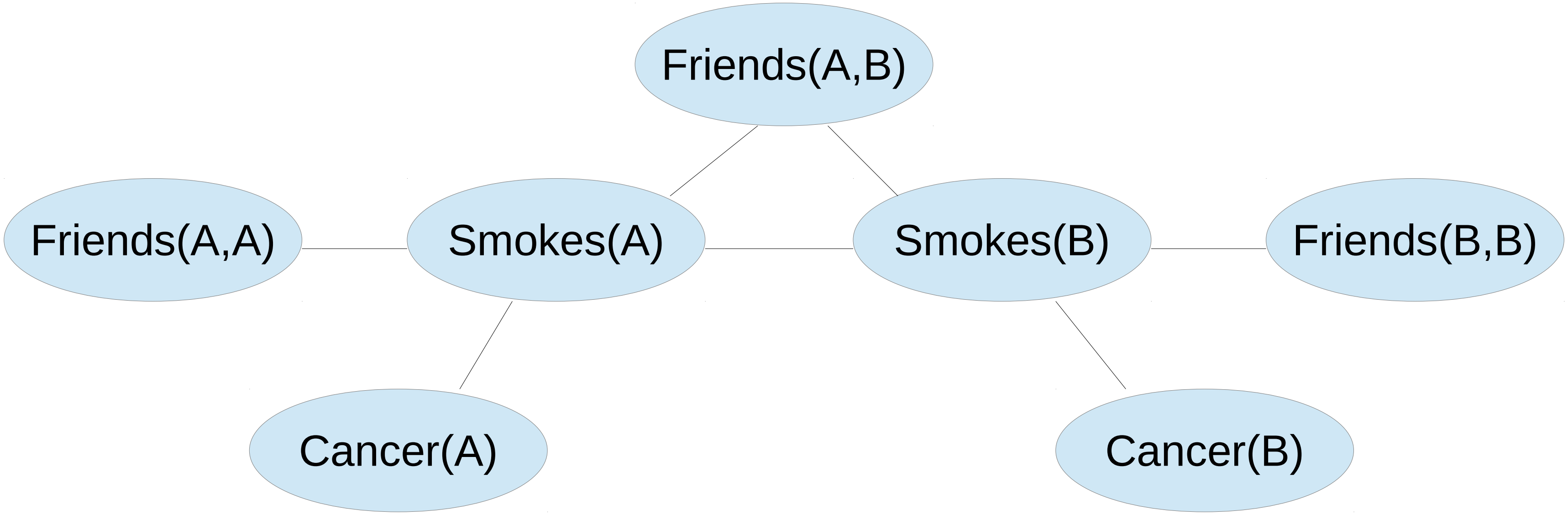
No. of true groundings of formula *i* in *x*

Example

0.7 Smokes(x) → Cancer(x)

0.6 Friends(x,y) → (Smokes(x) ∧ Smokes(y))

Two constants: A and B



MLN software

- Alchemy (*http://alchemy.cs.washington.edu/*)
- Probcog (*http://ias.cs.tum.edu/software/probcog*)

MLN applications to NLU

- General discourse interpretation (*Garrette et al., 11; Beltagy et al., 14*)
- Recognizing textual entailment (*Qiu et al., 12*)

Still problems

- Unable to choose between alternative interpretations if both are consistent
- Model minimality criteria is problematic
- Unable to reason with inconsistent knowledge
- If a piece of knowledge is missing, fails to find a proof
- Unlimited inference chains
- Reasoning is computationally complex (too many groundings)

Probabilistic Soft Logic

(Kimmig et al., 2012)

Similar to Markov Logic Networks. Differences:

- ground atoms have soft, continuous truth values in the interval [0, 1] rather than binary truth values
- Inference algorithm (Most Probable Explanation) can be implemented efficiently in polynomial time.

Application of PSL to NLU:

- Semantic Textual Similarity (*Beltagy et al., 14*)

Abduction for NLU

- Abduction – inference to the best (most economical) explanation.
- Idea:
  - New text = observation
  - Context = background knowledge
  - Interpreting text = providing the best explanation of why it would be true
- Early abduction-based approaches to discourse interpretation (*Norvig, 83; Wilensky, 83; Charniak&Goldman, 89; Stickel, 90; Hobbs et al., 93*)
  - Disambiguation
  - Metonymy/metaphor resolution
  - Coreference resolution
  - ...

Abduction: definition

Given: Background knowledge B, observations O, where both B and O are sets of first-order logical formulas,

Find: A hypothesis H such that $H \cup B \models O$; $H \cup B \not\models \perp$, where H is a set of first-order logical formulas.

Typically,

- observations are conjunctions of propositions and variable inequalities existentially quantied with the widest possible scope
- background knowledge is a set of Horn clauses

Inference operations

Backchaining (introduction of new assumptions)

$$\frac{\bigwedge_{i=1}^{n} P_n \rightarrow \bigwedge_{j=1}^{m} Q_j \in B \ and \ \bigwedge_{j=1}^{m} Q_j \ occurs \ in \ O \wedge H, where \ H \in \mathcal{H}}{\mathcal{H} := \mathcal{H} \cup \{H \wedge \bigwedge_{i=1}^{n} P_n\}}$$

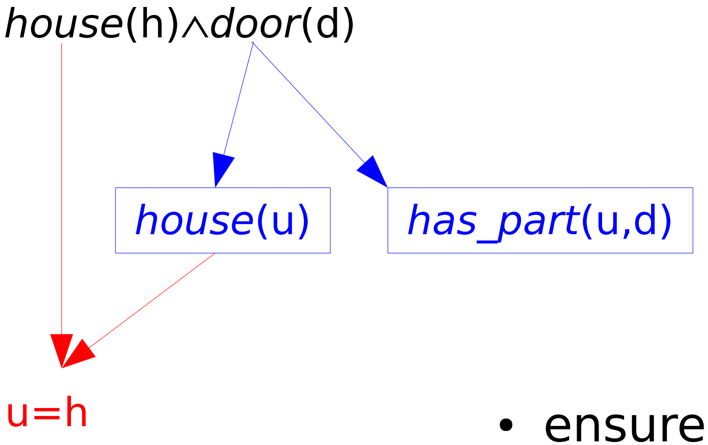
Unification (merging of prepositions)

set of hypotheses

$$\frac{p(X) \wedge p(Y) \ occurs \ in \ O \wedge H, where \ H \in \mathcal{H}}{\mathcal{H} := \mathcal{H} \cup \{H \wedge X = Y\}}$$

Example

John saw a house. The door was open.



- ensures discourse coherence
- allows incomplete knowledge
- supports defeasible knowledge

Estimating Hypothesis Likelihood

Many explanations can be found for the same observation.

Shakespeare's tragedy :

Did Shakespeare write a play or experience a drama?

How to chose the best one?

- Cost-based abduction (*Charniak&Shimony, 90*)
- Bayesian Networks-based abduction (*Pearl, 88; Charniak&Goldman, 89; Raghavan&Mooney, 10*)
- Markov Logic Networks-based abduction (*Kate and Mooney, 09*)
- Weighted abduction (*Hobbs, 93*)

Discussion of these approaches: (*Ovchinnikova et al., 13*)

Cost propagation scheme in weighted abduction

Each observable is assigned a cost (how probable it is to be explained vs. assumed)

$O = \{q(\mathrm{A})^{\$10}\}$

- Each assumption in KB is assigned a weight (how probable it is that it explains given literal)
  $B = \{p(\mathrm{x})^{1.2} \wedge \mathrm{s}(\mathrm{y})^{0.2} \rightarrow q(\mathrm{x})\}$
- Cost of the new assumption is a function of its weight and the cost of the explained literal. Usually $f(w,c) = w*c$ is used. Given O, assuming $p(\mathrm{A})$ costs $12.
- If a literal is explained, its cost = 0
  $O = \{q(\mathrm{A})^{\$10}\} \rightarrow H_0 = q(\mathrm{A})^{\$0} \wedge p(\mathrm{A})^{\$12}$
- If two literals are unified, then the cost of unification is the minimal cost out.
  $O = \{q(\mathrm{x})^{\$10} \wedge q(\mathrm{y})^{\$20}\} \rightarrow H_1 = q(\mathrm{x})^{\$0} \wedge q(\mathrm{y})^{\$0} \wedge \mathrm{x}{=}\mathrm{y}^{\$10}$
- Interpretation cost = sum of costs of all assumptions
  $\mathrm{cost}(H_0) = \$12$

Example

$\mathit{Shakespeare}(x2)^{\$10} \wedge \mathit{of}(x3,x2)^{\$10} \wedge \mathit{tragedy}(x3)^{\$0}$

$\text{Shakespeare}(x) \rightarrow \text{playwright}(x)^{1.2}$
$\text{Shakespeare}(x) \rightarrow \text{person}(x)^{1.1}$
$\text{playwright}(x) \rightarrow \text{author}(x,y)^{0.5} \wedge \text{play}(y)^{0.5}$
$\text{person}(x) \wedge \text{of}(x,y) \wedge \text{play}(y) \rightarrow \text{author}(x,y)^{2.0}$
$\text{tragedy}(x) \rightarrow \text{play}(x)^{1.2}$
$\text{tragedy}(x) \rightarrow \text{dramatic\_event}(x)^{1.2}$
$\text{person}(x) \wedge \text{of}(x,y) \wedge \text{dramatic\_event}(y) \rightarrow \text{experiencer}(x,y)^{2.0}$

OBSERVATION COST = $30

Example

Shakespeare(x2)$^{\$10}$ ∧ *of*(x3,x2)$^{\$10}$ ∧ *tragedy*(x3)$^{\$0}$

dramatic event(x3)$^{\$12}$

Shakespeare(x) → playwright(x)$^{1.2}$
Shakespeare(x) → person(x)$^{1.1}$
playwright(x) → author(x,y)$^{0.5}$ ∧ play(y)$^{0.5}$
person(x) ∧ of(x,y) ∧ play(y) → author(x,y)$^{2.0}$
tragedy(x) → play(x)$^{1.2}$
tragedy(x) → dramatic_event(x)$^{1.2}$
person(x) ∧ of(x,y) ∧ dramatic_event(y) → experiencer(x,y)$^{2.0}$

OBSERVATION COST = $32

Example

$\textit{Shakespeare}(x2)^{\$0} \wedge \textit{of}(x3,x2)^{\$10} \wedge \textit{tragedy}(x3)^{\$0}$

$\textit{person}(x2)^{\$11}$

$\textit{dramatic event}(x3)^{\$12}$

$\text{Shakespeare}(x) \rightarrow \text{playwright}(x)^{1.2}$

$\mathbf{Shakespeare(x) \rightarrow person(x)^{1.1}}$

$\text{playwright}(x) \rightarrow \text{author}(x,y)^{0.5} \wedge \text{play}(y)^{0.5}$

$\text{person}(x) \wedge \text{of}(x,y) \wedge \text{play}(y) \rightarrow \text{author}(x,y)^{2.0}$

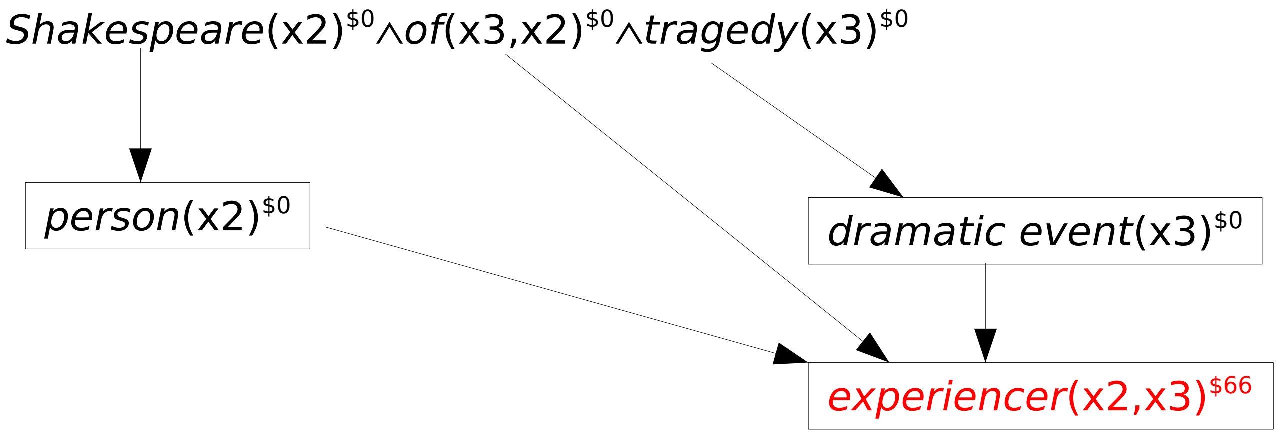
$\text{tragedy}(x) \rightarrow \text{play}(x)^{1.2}$

$\mathbf{tragedy(x) \rightarrow dramatic\_event(x)^{1.2}}$

$\text{person}(x) \wedge \text{of}(x,y) \wedge \text{dramatic\_event}(y) \rightarrow \text{experiencer}(x,y)^{2.0}$

OBSERVATION COST = $33

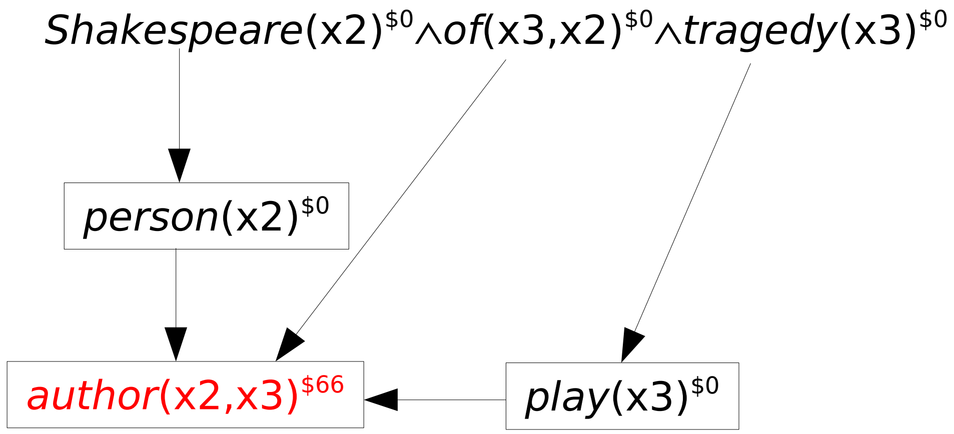
Example



Shakespeare(x) → playwright(x)$^{1.2}$
Shakespeare(x) → person(x)$^{1.1}$
playwright(x) → author(x,y)$^{0.5}$ ∧ play(y)$^{0.5}$
person(x) ∧ of(x,y) ∧ play(y) → author(x,y)$^{2.0}$
tragedy(x) → play(x)$^{1.2}$
tragedy(x) → dramatic_event(x)$^{1.2}$
person(x) ∧ of(x,y) ∧ dramatic_event(y) → experiencer(x,y)$^{2.0}$

OBSERVATION COST = $66

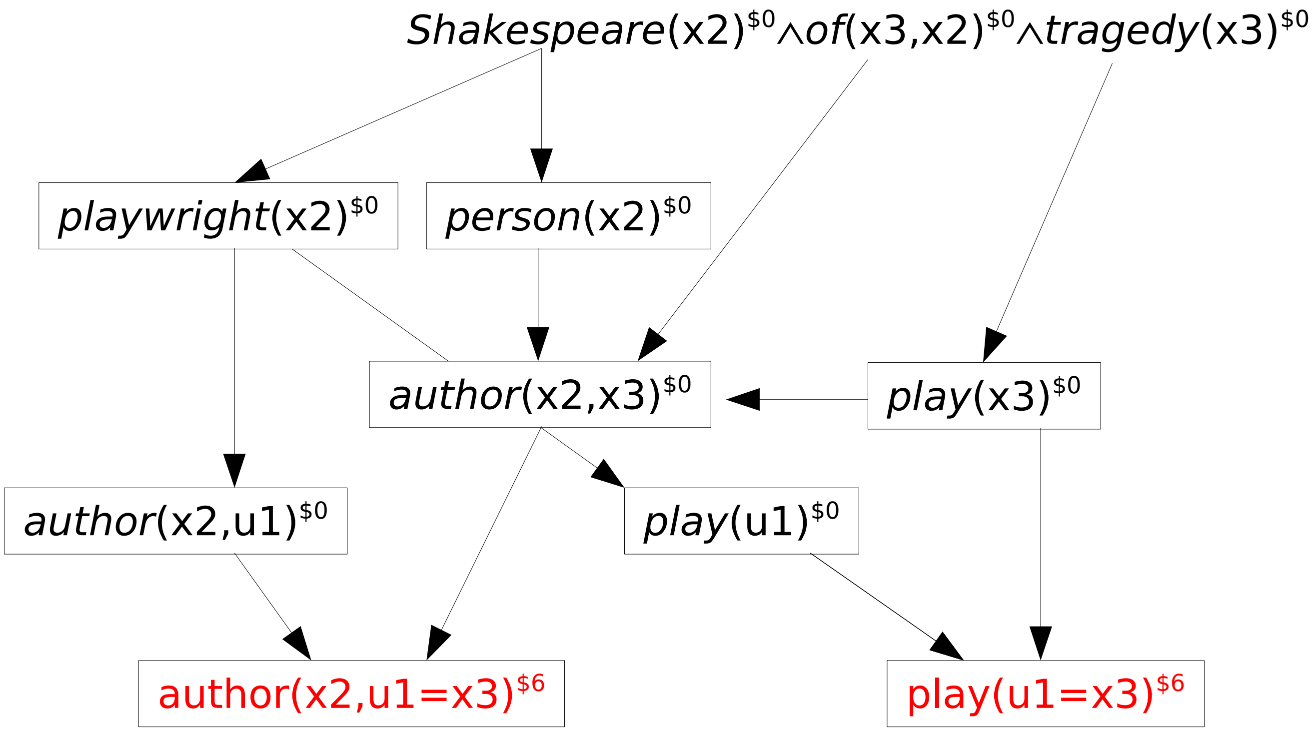
Example



Shakespeare(x) → playwright(x)$^{1.2}$
Shakespeare(x) → person(x)$^{1.1}$
playwright(x) → author(x,y)$^{0.5}$ ∧ play(y)$^{0.5}$
person(x) ∧ of(x,y) ∧ play(y) → author(x,y)$^{2.0}$
tragedy(x) → play(x)$^{1.2}$
tragedy(x) → dramatic_event(x)$^{1.2}$
person(x) ∧ of(x,y) ∧ dramatic_event(y) → experiencer(x,y)$^{2.0}$

OBSERVATION COST = $66

Example



Shakespeare(x) → playwright(x)$^{1.2}$
Shakespeare(x) → person(x)$^{1.1}$
playwright(x) → author(x,y)$^{0.5}$ ∧ play(y)$^{0.5}$
person(x) ∧ of(x,y) ∧ play(y) → author(x,y)$^{2.0}$
tragedy(x) → play(x)$^{1.2}$
tragedy(x) → dramatic_event(x)$^{1.2}$
person(x) ∧ of(x,y) ∧ dramatic_event(y) → experiencer(x,y)$^{2.0}$

OBSERVATION COST = $12

Why is it a nice framework for NLU?

- Allows assumptions
- Can deal with incomplete, inconsistent, and defeasible knowledge
- Supports discourse coherence (favors explanations with more unifications)
- Restricts inference chains

Problems

- Lack of expressivity (Horn clauses)
- No consistency check

Complexity of reasoning

- Generating search space has exponential complexity
- Previous implementation: Mini-TACITUS (*Mulkar et. al, 07*)
  around 30 min per sentence/1000 axioms

Solution:
implementation based on Integer Linear Programming
(*Inoue and Inui, 11*)

Integer Linear Programming (ILP)

- a technique for the optimization of a linear objective function, subject to linear equality and linear inequality constraints.

maximize	$\mathbf{c^T x}$
subject to	$\mathbf{Ax \le b}$
and	$\mathbf{x \ge 0}$

Example:

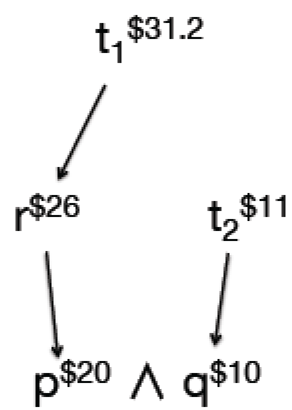
maximize	$\mathbf{S_1 x_1 + S_2 x_2}$
subject to	$\mathbf{0 \le x_1 + x_2 \le L}$

Weighed abduction as ILP problem

1. Apply all possible axioms generating new assumptions.
2. Candidate interpretations can be represented as an arbitrary combination of assumptions.

Potential elemental hypotheses:

$P = \{\mathrm{p}, \mathrm{q}, \mathrm{r}, \mathrm{t}_1, \mathrm{t}_2\}$



Observations:

$O = \mathrm{p}^{\$20} \wedge \mathrm{q}^{\$10}$

Background knowledge:

$B = \{\mathrm{r}^{1.3} \Rightarrow \mathrm{p}, \; \mathrm{t}^{1.1} \Rightarrow \mathrm{q}, \; \mathrm{t}^{1.2} \Rightarrow \mathrm{r}\}$

Candidate hypotheses:

$\mathcal{H}$		P: p	q	r	t_1	t_2	$c(H)$	
H_1	$p \wedge q$	●	●				\$30	
H_2	$p \wedge t_2$	●	○			●	\$31	
H_3	$r \wedge q$	○	●	●			\$36	
H_4	$r \wedge t_2$	○	○	●		●	\$37	
H_5	$t_1 \wedge q$	○	●	○	●		\$41.2	
H_6	$t_1 \wedge t_2$	○	○	○	●	●	\$42.2	
H_7	t_2	○	○	○	○	●	\$11	← H^*

* t_1 and t_2 are unified in H_7.

Weighted abduction as ILP problem

3. Introduce variables for each predication p which define whether p is included into the best interpretation, unified with other predications, etc.

   $h_p = 1$, if p is included into the interpretation, otherwise $h_p=0$

   $r_p = 1$, if p does not pay its cost, otherwise $r_p=0$

   $u_{p,q} = 1$, if p is merged with u, otherwise $u_{p,q}=0$

4. Define constraints on these variables

   $h_p = 1$ for each input p

   $u_{p,q} \leq \frac{1}{2}(h_p + h_q)$ for each p, q

5. Represent cost of hypothesis as linear function of 0-1 variables

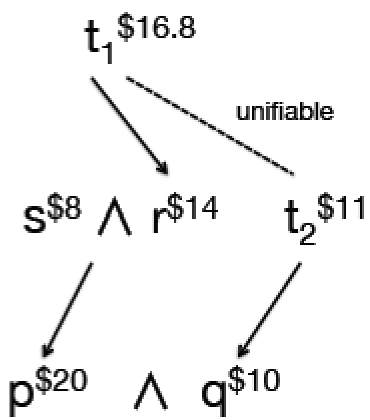
   $$cost(H) = c_1 \cdot h_{p1} + \ldots + c_n \cdot h_{pn}$$

6. Use state-of-the-art ILP solver for finding assignments of the variables, which minimize the objective function

Weighed abduction as ILP problem

Potential elemental hypotheses:

$P = \{\mathrm{p, q, r, s, t_1, t_2}\}$



Observations:

$O = \mathrm{p}^{\$20} \wedge \mathrm{q}^{\$10}$

Background knowledge:

$B = \{\mathrm{s}^{0.4} \wedge \mathrm{r}^{0.7} \Rightarrow \mathrm{p},$
$\mathrm{t}^{1.1} \Rightarrow \mathrm{q},$
$\mathrm{t}^{1.2} \Rightarrow \mathrm{r}\}$

Example of constraints:

C1: $h_p=1,\ h_q=1$

C2: $r_p \leq h_s + h_r,\ h_s = h_r \cdot 1$

$r_{t1} \leq u_{t1,t2}$

C3: $u_{t1,t2} \leq ½\,(h_{t1}+h_{t2})$

Candidate hypotheses:

$\mathcal{H}$		h_p	r_p	h_q	r_q	h_r	r_r	h_s	r_s	h_{t1}	r_{t1}	h_{t2}	r_{t2}	$u_{t1,t2}$	$c(H)$
H_1	$p \wedge q$	1	0	1	0	0	0	0	0	0	0	0	0	0	\$30
H_2	$p \wedge t_2$	1	0	1	1	0	0	0	0	0	0	1	0	0	\$31
H_3	$s \wedge r \wedge q$	1	1	1	0	1	0	1	0	0	0	0	0	0	\$32
H_4	$s \wedge r \wedge t_2$	1	1	1	1	1	0	1	0	0	0	1	0	0	\$33
H_5	$s \wedge t_1 \wedge q$	1	1	1	0	1	1	1	0	1	0	0	0	0	\$34.8
H_6	$s \wedge t_1 \wedge t_2$	1	1	1	1	1	1	1	0	1	0	1	0	0	\$35.8
H_7	$s \wedge t_2$	1	1	1	1	1	1	1	0	1	1	1	0	1	\$19

Comparison with *Mini-TACITUS*

(Inoue&Inui, 11)

- Dataset: 50 plan recognition problems, 107 axioms (evaluation dataset for ACCEL)

System	Depth	% solved	Time [sec]	Precision	Recall	F-measure
Mini-Tacitus	1	28	8.3	43	61	50
	2	20	10.2	38	64	47
	3	20	10.2	38	64	47
ILP-system	1	100	0.03	57	69	62
	2	100	0.36	53	76	62
	3	100	0.96	53	77	62

Applications of weighted abduction to NLU

- Recognizing textual entailment (*Ovchinnikova et al., 11; Inoue et al., 14*)
- Coreference resolution (*Inoue et al., 12*)
- Recognizing Implicit Discourse Relations (*Sugiura et al., 13*)
- Metaphor interpretation (*Ovchinnikova et al., 14*)

Summary

- Automatic theorem proving has significant limitations as applied to NLU
- Probabilistic deduction is promising
- Abduction has some nice features relevant for NLU, e.g., it supports discourse coherence
- Integer linear programming can help with the complexity issue

References

Beltagy, I., K. Erk, and R. Mooney (2014). Semantic Parsing Using Distributional Semantics and Probabilistic Logic. In *Proc. of SP14*. 7-11.

Black, F. (1964). A Deductive Question Answering System. Ph. D. thesis, Harvard University.

Bos, J. (2003). Exploring Model Building for Natural Language Understanding. In *Proc. of IWCS*, 41–55.

Bos, J., K. Markert (2006): When logical inference helps determining textual entailment (and when it doesn't). In: B. Magnini, I. Dagan (eds): The Second PASCAL Recognising Textual Entailment Challenge. Proc. of the Challenges Workshop, pp 98–103, Venice, Italy.

Bos, J. (2009). Applying automated deduction to natural language understanding. *Journal of Applied Logic* 7(1), 100–112.

Charniak, E. (1972). Toward a Model of Children's Story Comprehension. Ph. D. thesis, MIT.

Charniak, E. and R. P. Goldman (1989). A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding. In *Proc. of the IJCAI*, 1074–1079.

References

Charniak E. and S. E. Shimony. 1990. Probabilistic semantics for cost-based abduction. In *Proc. of the 8th National Conference on AI*, 106–111.

Cimiano, P. (2003). Building models for bridges. In *Proc. of the Workshop on Inference in Computational Semantics*, 57–71.

Gardent, C. and B. L. Webber (2001). Towards the Use of Automated Reasoning in Discourse Disambiguation. *Journal of Logic, Language and Information* 10(4), 487–509.

Garrette, D., K. Erk.,and R. Mooney. (2011). Integrating logical representations with probabilistic information using Markov logic. In *Proc. of IWCS*. 105–114.

Green, C. C. and B. Raphael (1968). The use of theorem-proving techniques in question-answering systems. In *Proc. of ACM*, New York, NY, USA, 169–181.

Hobbs, J. R., M. Stickel, D. Appelt, and P. Martin (1993). Interpretation as abduction. *Artificial Intelligence* 63, 69–142.

Inoue, N. and K. Inui (2011). ILP-Based Reasoning for Weighted Abduction. In *Proc. of AAAI Workshop on Plan, Activity and Intent Recognition.*

References

Inoue. N., E. Ovchinnikova, K.Inui, and J. R. Hobbs (2012). Coreference Resolution with ILP-based Weighted Abduction. In Proc. of COLING, 1291-1308.

Inoue. N., E. Ovchinnikova, K.Inui, and J. R. Hobbs (2014). Weighted Abduction for Discourse Processing Based on Integer Linear Programming. In Gita Sukthankar, Christopher Geib, Robert P. Goldman, Hung Bui, and David V. Pynadath (Eds), *Plan, Activity, and Intent Recognition*, Elsevier, pp.33-55.

Kate. R.J. and R. J. Mooney (2009). Probabilistic abduction using markov logic networks. In *Proc. Of PAIR'09*, Pasadena, CA.

Kimmig. A., S. H. Bach, M. Broecheler, B. Huang, and L. Getoor (2012). A short introduction to Probabilistic Soft Logic. In *Proc. of NIPS Workshop on Probabilistic Programming: Foundations and Applications*.

Mulkar, R., J. R. Hobbs, and E. Hovy (2007). Learning from Reading Syntactically Complex Biology Texts. In *Proc. of the International Symposium on Logical Formalizations of Commonsense Reasoning,* Palo Alto.

Norvig, P. (1987). Inference in text understanding. In *Proc. of National Conference on Artificial Intelligence*, 561–565.

References

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Raghavan, S. and R. Mooney (2010). Bayesian abductive logic programs. In *Proc. of Star-AI'10*, 82–87, Atlanta, GA.

Richardson, M. and P. Domingos (2006). Markov logic networks. *Machine Learning* 62(1-2), 107–136.

Stickel, M. E. (1990). Rationale and methods for abductive reasoning in naturallanguage interpretation. In Studer (Ed.), *Natural Language and Logic*, Volume 459 of LNCS, pp. 233–252. Springer.

Sugiura, J., N. Inoue. and K. Inui. Recognizing Implicit Discourse Relations through Abductive Reasoning with Large-scale Lexical Knowledge (2013). In Proc. of the 1st Workshop on Natural Language Processing and Automated Reasoning.

Tatu,M. and D. Moldovan (2007). COGEX at RTE3. In *Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. 22–27.

Ovchinnikova, E., N. Montazeri, T. Alexandrov, J. R Hobbs, M. McCord, and R. Mulkar-Mehta (2011). Abductive Reasoning with a Large Knowledge Base for Discourse Processing. In *Proc. of IWCS*, 225–234, Oxford, UK.

References

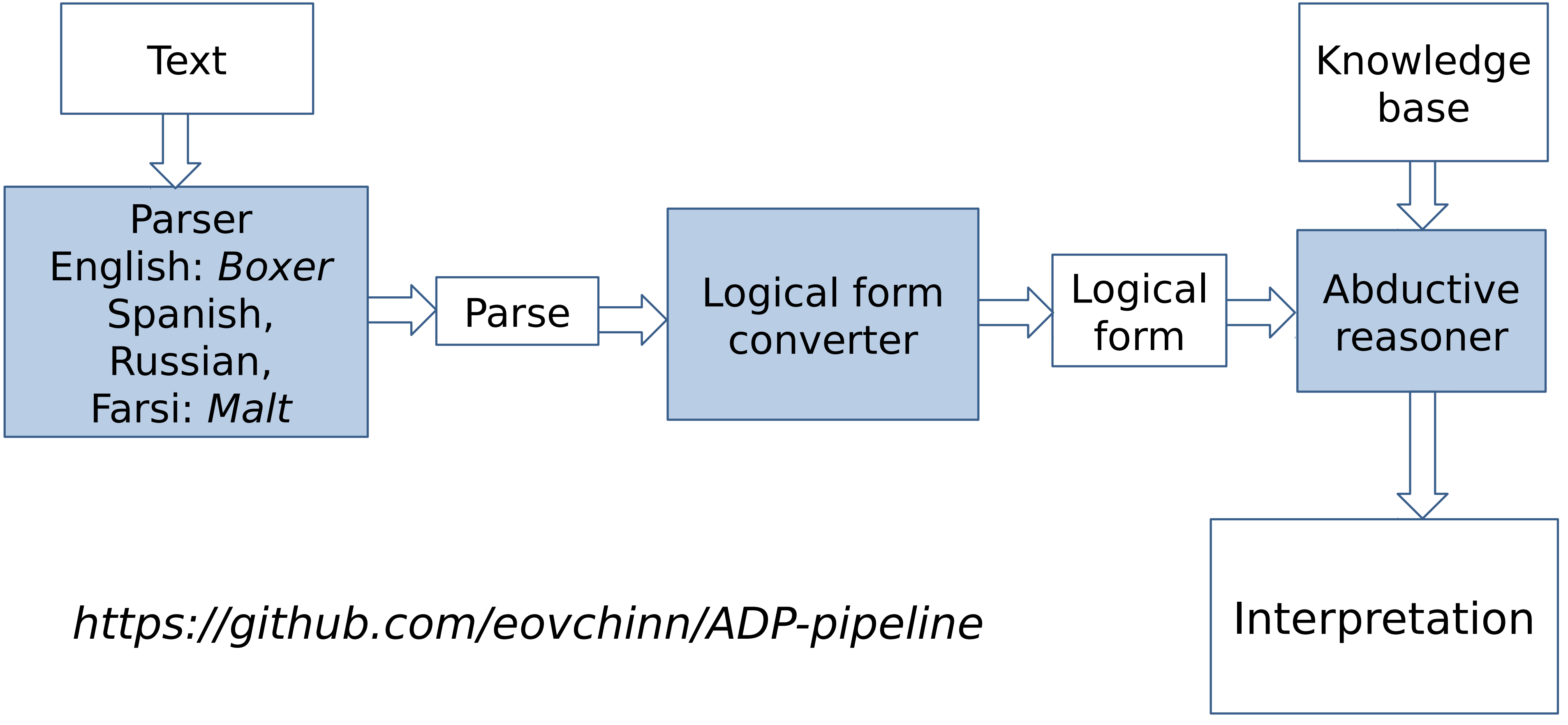
Ovchinnikova, E., Israel, R., Wertheim, S., Zaytsev, V., Montazeri, N., and Hobbs, J. (2014). Abductive Inference for Interpretation of Metaphors. In *Proc. of ACL 2014 Workshop on Metaphor in NLP*. Baltimore, MD.,, to appear.

Qiu, X., L. Cao, Z. Liu, and X. Huang (2012). Recognizing Inference in Texts with Markov Logic Networks. ACM Trans. Asian Lang. Inf. Process. 11(4): 15.

Wilensky, R. (1983). *Planning and Understanding: A Computational Approach to Human Reasoning*. Reading, MA: Addison-Wesley.

End-to-end NLU system

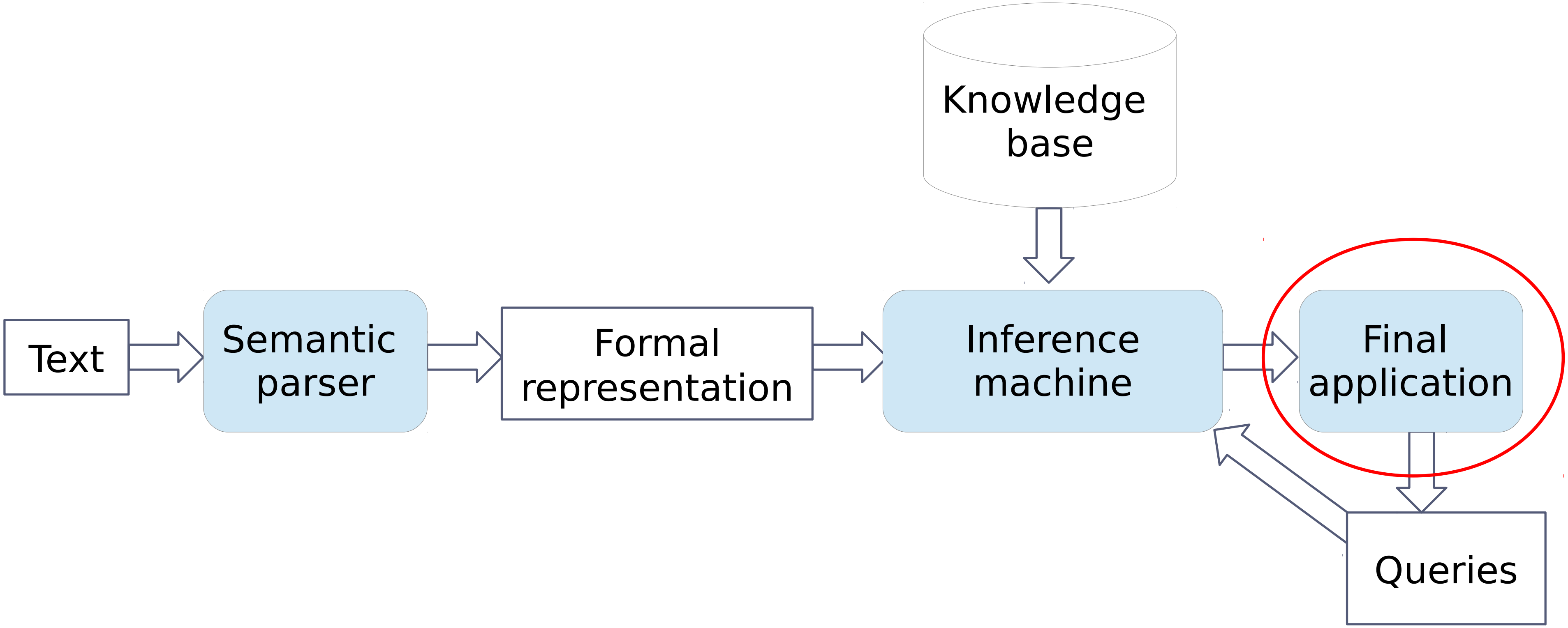
Implemented for: English, Spanish, Russian, Farsi



https://github.com/eovchinn/ADP-pipeline

NLU applications

Inference-based NLU pipeline



Recognizing textual entailment (RTE)

(*Dagan&Glickman, 05; Dagan et al., 13; Bos, 13; Bos, 14*)

Text :	*John gave a book to Mary.*
Hypothesis :	*Mary got a book.*
Entailment:	YES

Text :	*John gave a book to Mary.*
Hypothesis:	*Mary read a book.*
Entailment:	NO

Task: given a Text-Hypothesis pair predict entailment

Recognizing textual entailment (RTE)

- captures major semantic inference needs in natural language understanding
- generic for several NLU applications:
  - information extraction: extracted information should be entailed by the corresponding text.
  - question answering: the answer is entailed by the supporting text fragment.
  - summarization: the text should entail its summary.

Deduction for RTE

Nutcracker system (http://svn.ask.it.usyd.edu.au/trac/candc/wiki/nutcracker)

Theorem prover:

- $T \rightarrow H$ Entailment
- $\neg(T \wedge H)$ Inconsistency, no entailment

  T: *His family has steadfastly denied the charges.*
  H: *The charges were denied by his family.*
- $KB \wedge T \rightarrow H$ Entailment
- $\neg(KB \wedge T \wedge H)$ Inconsistency, no entailment

  T: *Crude oil prices soared to record levels.*
  H: *Crude oil prices rise.*

Model builder:

- $\neg(T \wedge H)$ No entailment possible
- $T \wedge H$ Entailment possible

Deduction for RTE

Nutcracker was evaluated on RTE-2 Challenge dataset (*Bos and Markert, 06*).

In this evaluation:

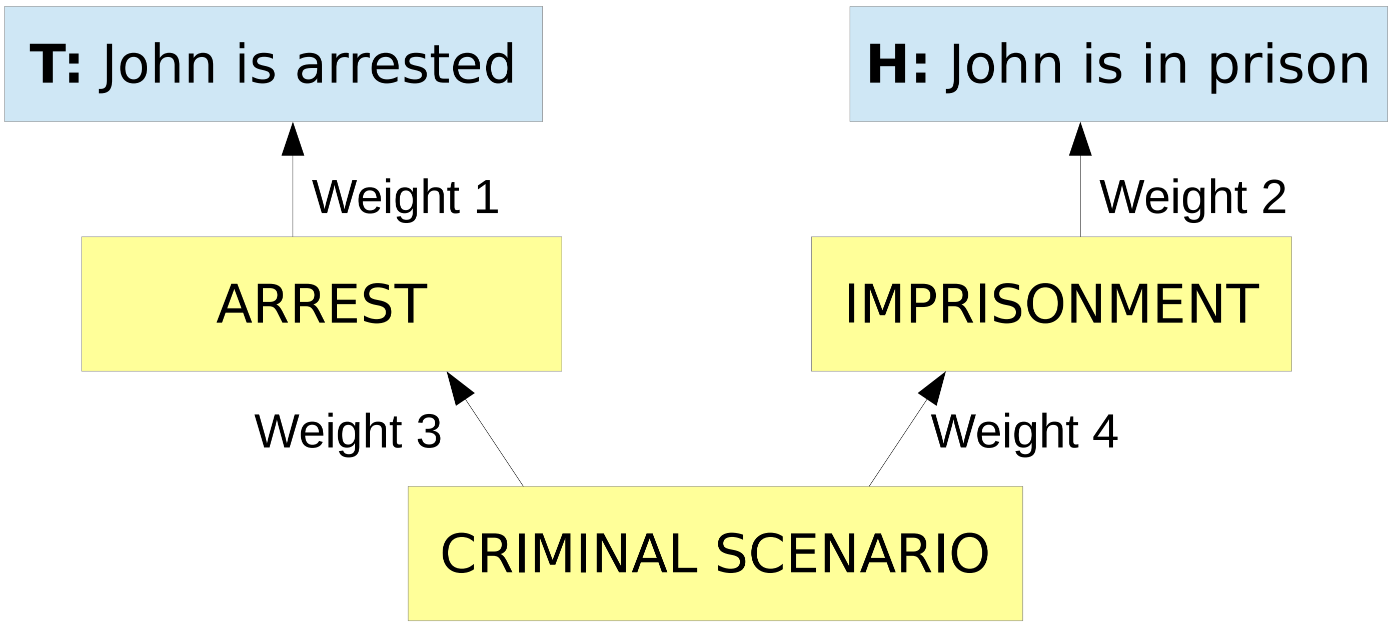
- The dev and test datasets contain 800 T-H pairs each.
- Shallow features (lexical overlap) were together with deep features (logical proofs)
- Small KB of world knowledge created manually
- Difference of T and H model sizes used as another feature

Results:

- Overall performance without deep features was better!
- 29 proofs found (22 correct proofs)
- 19 proofs without KB
- 10 proofs with a small manually created KB

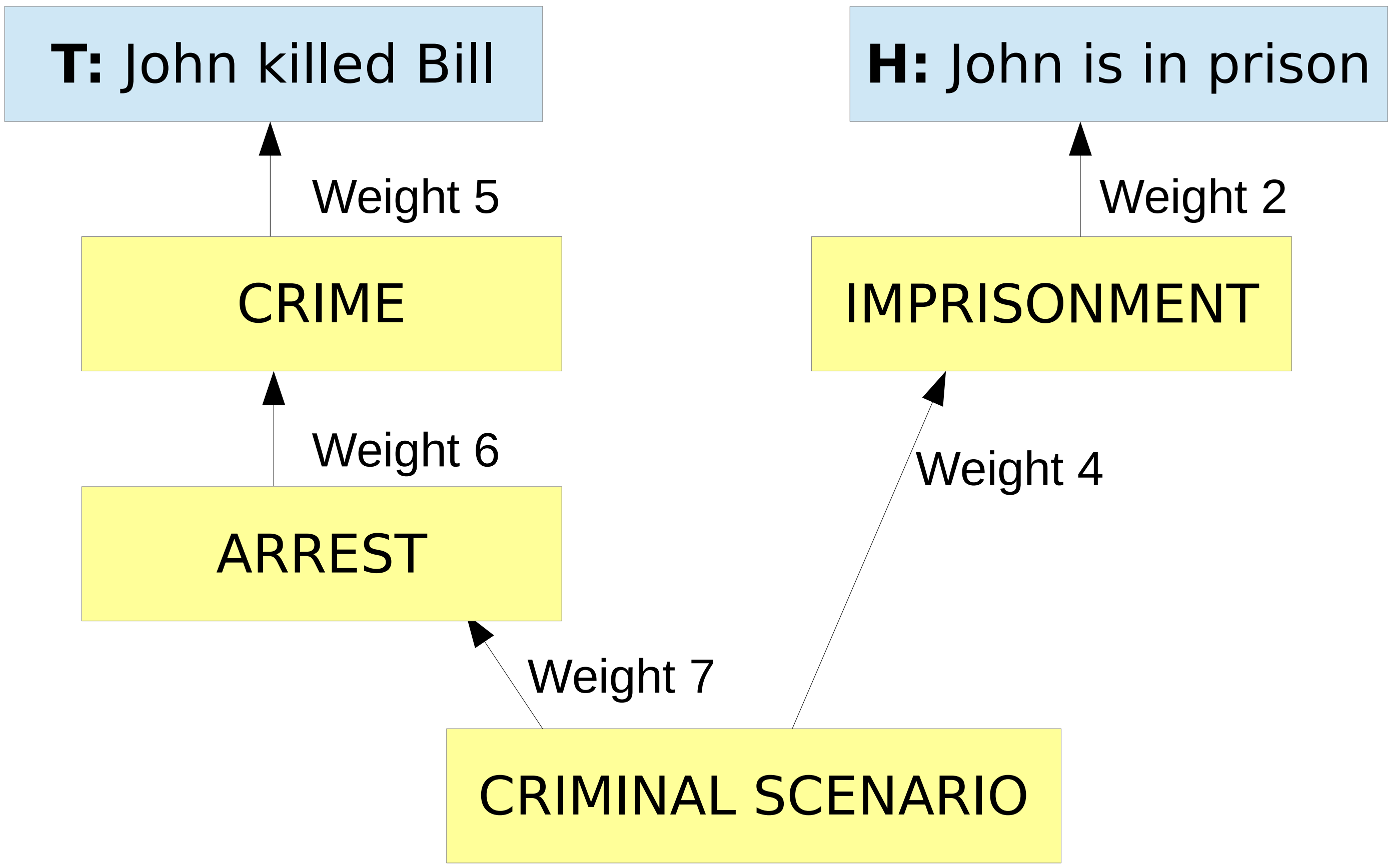
Reason: missing knowledge, hard YES/NO inference

RTE as Discourse Interpretation



Does knowing T helps to understand H?
→ How much does T reduce the cost of interpreting H?

RTE as Discourse Interpretation



Abduction for RTE

Procedure:

1. compute best interpretation of T towards KB:

   $KB \Rightarrow Int(T)$

2. compute best interpretation of H towards KB:

   $KB \Rightarrow Int(H)$

3. add best interpretation of T to KB:

   $KB + Int(T)$

4. compute best interpretation of H towards KB + I(T):

   $KB + Int(T) \Rightarrow Int_{KB+Int(T)}(H)$

5. is $cost\ (Int_{KB}\ (H) - I_{KB+Int(T)}\ (H)\ ) > threshold$?

Note: threshold is defined in training

Statistics of axioms used in the RTE&coref experiments

- lexeme-synset mappings (~ 422 000 axioms)
- WordNet synset relations (~ 141 000 axioms)
- WordNet derivational relations (~ 35 000 axioms)
- synset definitions (~ 120 500 axioms)
- mapping of lexemes to FrameNet frames (~ 35 000 axioms)
- frame relations (~ 5 900 axioms)

RTE experiment

(Inoue et al., 14)

- Datasets: RTE Challenge 1-5 datasets
- Axioms weights derived from annotated corpora

RTE	Dev	Test	Accuracy	Baseline	Average
1	567	800	54.2	53.6	54.6
2	800	800	61.4	59.2	60.3
3	800	800	62.7	62.8	54.4
4	-	1000	57.1	58.8	59.4
5	600	600	61.0	60.3	61.4

Main problem: coreference!

Coreference problem

Simple merging of predicates with the same name does not work

- *John eats an apple and Bill eats an apple.*
- *risk of conflict of interests*
- *John likes the red apple and the green apple.*

Solution: weighted unification

Weighted unification

Unification is modeled in a machine-learning framework

Negative features:

- Incompatible properties (black – white)
- Frequent predicates (of, go)
- Arguments of the same predicate (give(e1,x1,x2,x3))
- Explicit non-identity (similar to, different from)
- Functional relations (father of)
- Modality (not, believe)

Positive features:

- Common properties (*John was jogging, while Bill was sleeping. He jogs every day*)
- Derivational relations (buy(e1,x,y), buyer(x))

Coreference experiment

(Inoue et al., 12)

- Dataset: CoNLL-2011 dataset
- Axiom weights and features weights derived by learning

Results:

- Best performance (BLANC F-measure) with all features in combination
- **Outperforms the naive approach by more than 20% F-measure (60.4 vs. 39.9)**
- Some overmergings are were not captured
  - Different syntactic representation of the same property (*Japanese goods* vs. *goods from Germany*)
  - Discourse salience (*He sat near him*)

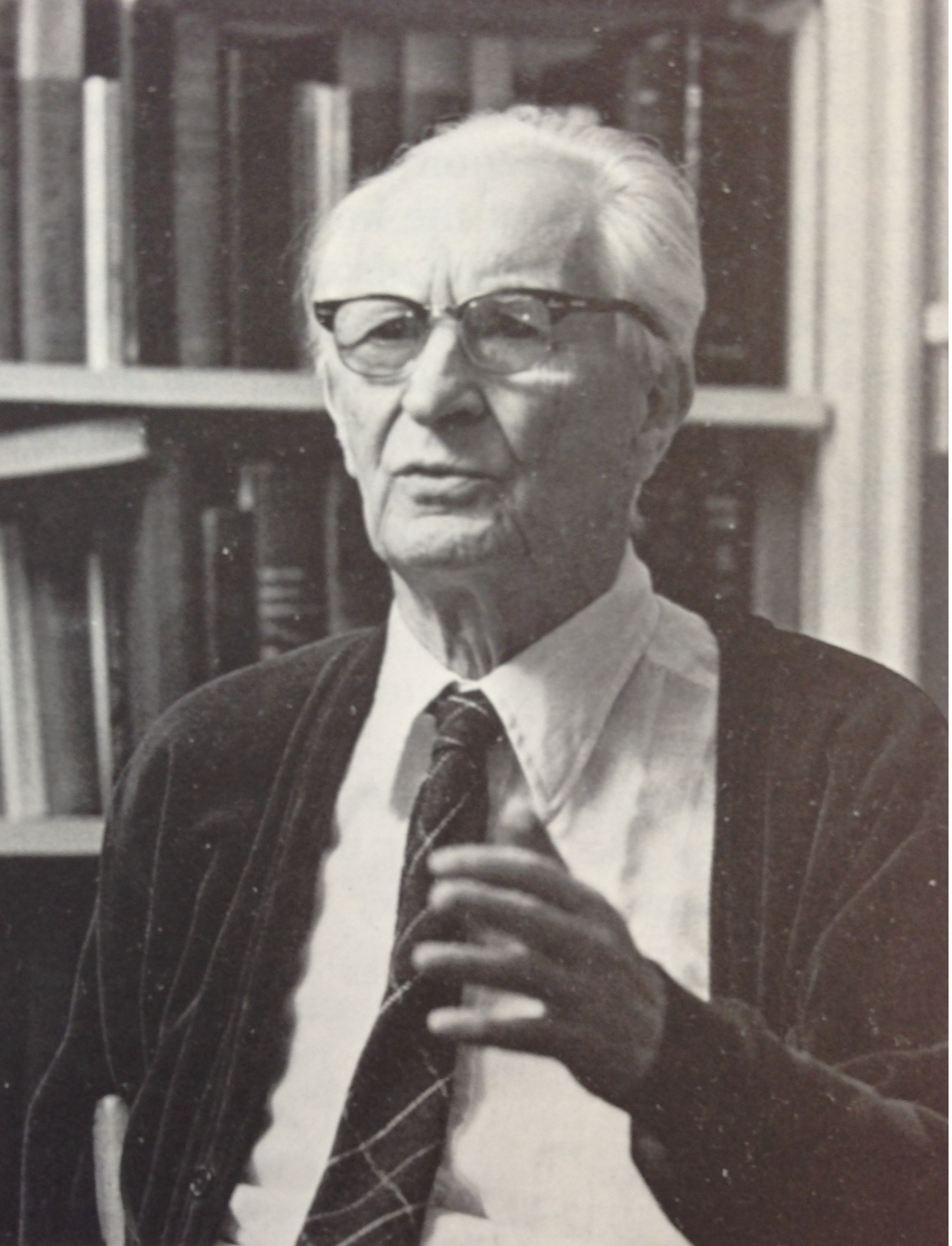
Lessons learned

- For the first time large-scale inference-based NLU is possible
- Just pumping in knowledge and running an inference machine is not enough: How to choose the best interpretation?
  - Which unifications/mergings do we allow?
  - Where to get knowledge about inconsistency?
  - How to estimate probabilities? *(Srikumar&Roth, 13)*

Narrativization of videos

(Heider-Simmel Interactive Theater at ICT/ISI)

Film by Fritz Heider and his student, Marianne Simmel, 1944



"…it has been impressive the way almost everybody who has watched it has perceived the picture in terms of human action and human feelings."

Fritz Heider

Heider-Simmel Interactive Theater: Project goal

Automatically interpret simple 2-dimensional videos (similar to the original Heider-Simmel video) in terms of mental states (goals, intentions, emotions) expressed by natural language narratives.

http://narrative.ict.usc.edu/heider-simmel-interactive-theater.html

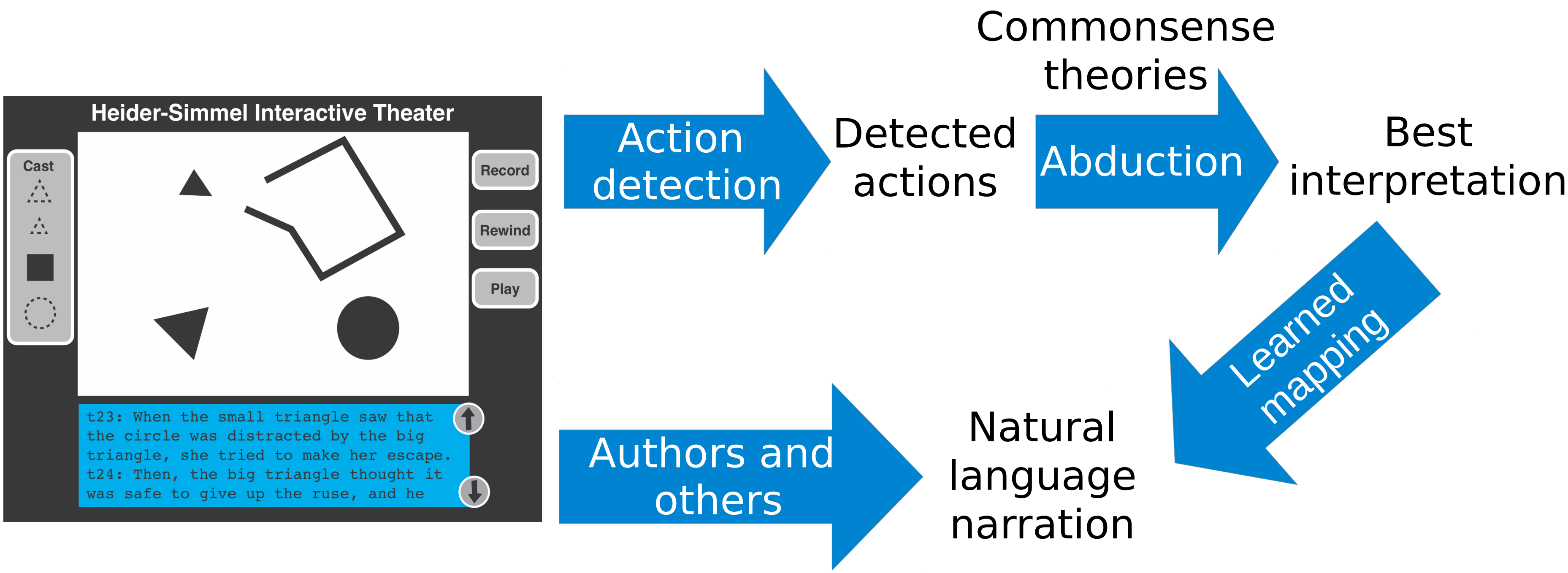
Solution

Action recognition
Actions are identified using contemporary Gesture Recognition methods

Interpretation as abduction
The internal causes are identified as the best proofs of the observed behaviors, using a formal theory of Commonsense Psychology in the reasoning framework of abductive inference

Data-driven narrative generation
Textual narratives are generated from the best proofs using contemporary grammar and data-driven language generation techniques, from thousands of example narratives

Solution


Heider-Simmel Interactive Theater
Cast
Record
Rewind
Play
t23: When the small triangle saw that the circle was distracted by the big triangle, she tried to make her escape. t24: Then, the big triangle thought it was safe to give up the ruse, and he
Action detection
Detected actions
Commonsense theories
Abduction
Best interpretation
Learned mapping
Authors and others
Natural language narration

Example

Example

<u>Observation</u>: chase(e1,BigT,Cir) & open(e2,LittleT,Door) & face(e3,LittleT,e1)

BigT is chasing Cir. LittleT opens Door and faces the chasing scene.

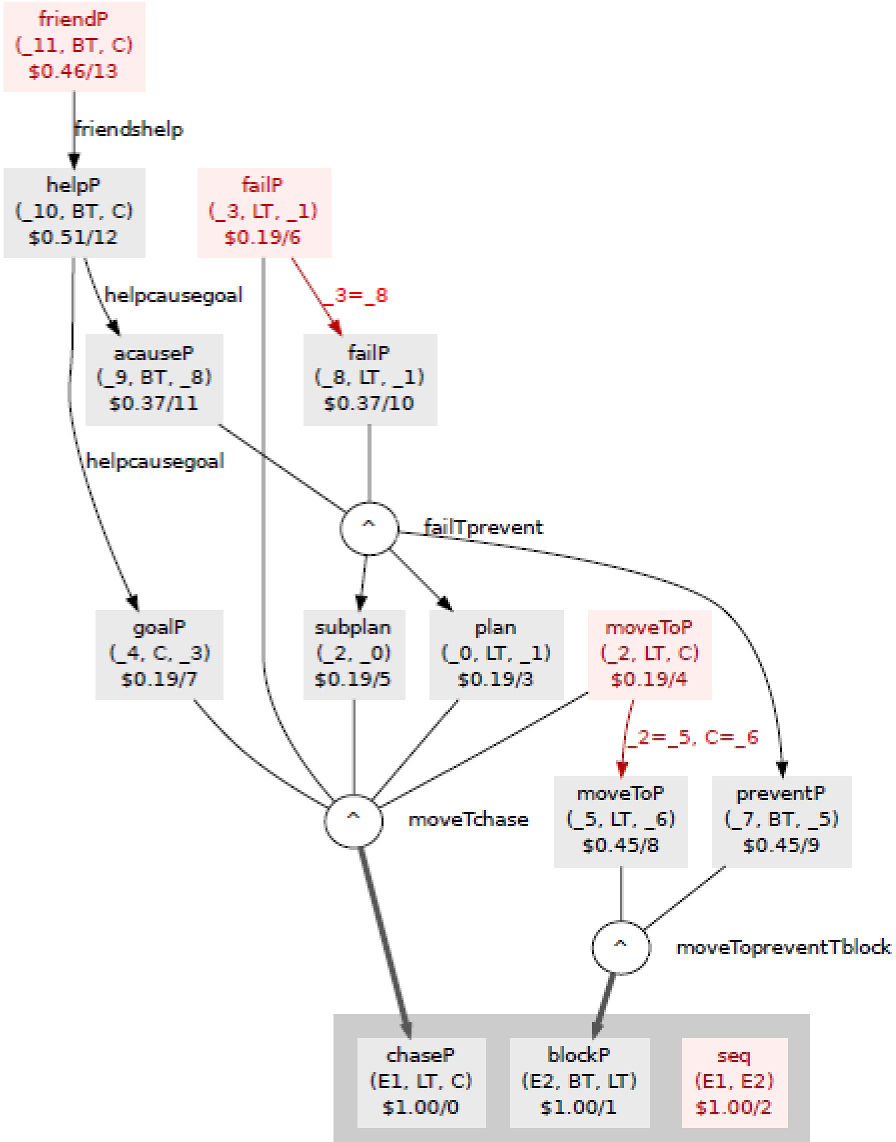
<u>Interpretation</u>: goal(e3,BigT,e4) & get(e4,BigT,Cir) & goal(e5,Cir,e4)& escape(e6,Cir,BigT) & frustrated(e7,BigT) & afraid(e8,Cir) & watch(e9,LittleT,e1) & pays_attention_to(e10,LittleT,e1)

The goal of BigT is to get Cir, the goal of Cir is to escape BigT, BigT is frustrated, Cir is afraid. LittleT is watching the chasing, it pays attention to it.

<u>Background knowledge (Commonsense theories)</u>:

1. People execute plans because they envision that doing so will cause their goals to be achieved
2. When people chase, they want to get
3. When people are chased, they want to escape
4. People feel fearful about an envisioned possible event that violates their goals
5. People feel frustrated about the failure of their plans to achieve their intended goals
6. If people face something, they watch it
7. If people watch something, they pay attention to it
8. …

Interpretation proofgraph



HEIDER-SIMMEL
INTERACTIVE THEATER

Make your own movies,
write your own stories
...with triangles

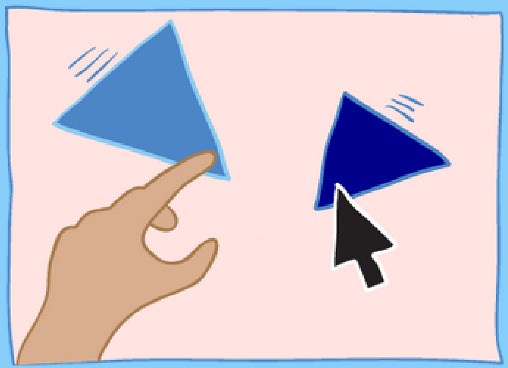
username

password

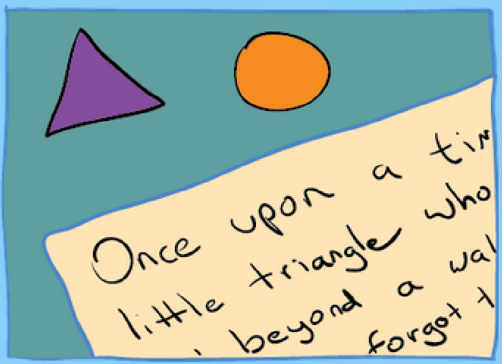
Log in

Want to try? Sign up for free!

Use the same username and password to play Triangle Charades



Make your own movies with a mouse or touch interface

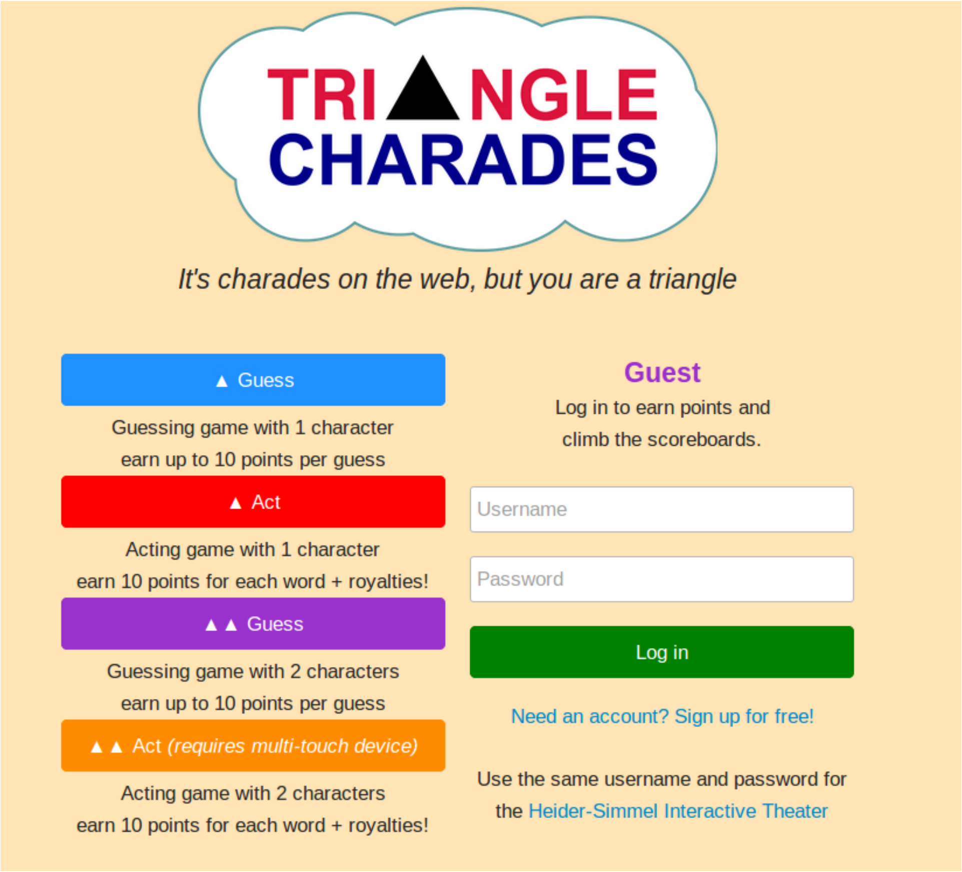


Write your own narrative interpretations of other people's movies

Triangle Charades Game

(Roemmele et al.. 14)

- for collecting training data
- use English verbs as action labels
- compute agreement and confusability



Textual+visual knowledge

- LEVAN project at Paul Allen's institute:

  http://levan.cs.washington.edu/

Summary

- Linguistic and visual input can be interpreted with the similar methods/in combination

References

Bos, J. (2014). Recognizing Textual Entailment and Computational Semantics. In *Computing meaning* (pp. 89-105). Springer Netherlands.

Bos, J. (2013): Is there a place for logic in recognizing textual entailment? *Linguistic Issues in Language Technology* 9(3): 1–18.

Bos, J., K. Markert (2006): When logical inference helps determining textual entailment (and when it doesn't). In: B. Magnini, I. Dagan (eds): The Second PASCAL Recognising Textual Entailment Challenge. Proc. of the Challenges Workshop, pp 98–103, Venice, Italy.

Dagan, I., O. Glickman, and B. Magnini (2005). The PASCAL Recognizing Textual Entailment Challenge. *In Machine Learning Challenges*, Volume 3944 of LNCS, pp. 177–190. Springer.

Dagan, I., Roth, D., and Sammons, M. (2013). *Recognizing textual entailment*.

Santosh Divvala, Ali Farhadi, Carlos Guestrin (2014). Learning Everything about Anything: Webly-Supervised Visual Concept Learning. In Proc. of CVPR.

References

Srikumar, V. and D. Roth, Modeling Semantic Relations Expressed by Prepositions. Transactions of the Association for Computational Linguistics (2013) pp. 231-242

Inoue, N., E. Ovchinnikova, K. Inui, and J. R. Hobbs (2012). Coreference Resolution with ILP-based Weighted Abduction. In Proc. of COLING, pp.1291-1308.

Inoue, N., E. Ovchinnikova, K. Inui, and J. R. Hobbs (2014). Weighted Abduction for Discourse Processing Based on Integer Linear Programming. In Gita Sukthankar, Christopher Geib, Robert P. Goldman, Hung Bui, and David V. Pynadath (Eds), *Plan, Activity, and Intent Recognition*, Elsevier, 33-55.

Roemmele, M., Archer-McClellan, H., and Gordon, A. (2014) Triangle Charades: A Data-Collection Game for Recognizing Actions in Motion Trajectories. In *Proc. Of International Conference on Intelligent User Interfaces*, Haifa, Israel.


ALMOST THERE!

Final summary

- NLU requires knowledge about linguistic structures and the world + ability to draw inferences

- Translating NL into logical representations is kind of solved
  - Use rule-based parsers for general domain applications
  - Train your own domain-specific parser

- World knowledge can be obtained from:
  - Lexical-semantic dictionaries
  - Expert- or community-developed ontologies
  - Corpora

- It's still not easy to obtain structurally complex knowledge

Final summary

- Logical inference is not yet really fit for NLU

  - it should be probabilistic, but what are other requirements?

- Deep approaches to NLU based on inference do not yet beat shallow approaches on a large scale.

- Observations/knowledge of different types (textual, visual) can be interpreted or used for interpretation in the same framework.


what's next?

- Now, when inference-based NLU work on a large scale, we should explore what logic can and cannot do in real applications.
- It is still unclear, what kind of structural complexity of knowledge we need for NLU applications (what cannot be learned, does not exist)
- Logical structure of NL can inform machine learning approaches.
- Multi-modal interpretation frameworks have a great potential.


Thank You