Automatic Ontology Extraction

Miloš Husák
xhusak@mail.muni.cz

RASLAN 2010

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

EVROPSKÁ UNIE esf MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost 2007-13 UNIVERSITAS MASARYKIANA BRUNENSIS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

EVROPSKÁ UNIE

esf

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

UNIVERSITAS MASARYKIANA BRUNENSIS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

My apologies for “I believe” and “I think”

Motivation

“I believe that ontologies can improve performance of word-sense desambiguation, and therefore improve parsing and many other tasks.”

“Existing ontologies do not seem to be suitable for word-sense desambiguation”

“Manual building of language resources is slow, expensive, tiresome and inaccurate.”

What kind of information do YOU think we need for word-sense desambiguation?

Comparison of existing technologies

- Semantic hierarchy: WordNet
- Semantic roles: FrameNet
- Grammars: Synt
- Syntactic properties: Verbalex

Problems: Incomplete/overly complete, inconsistent, contradicting, expensive, no signs of use for analysis

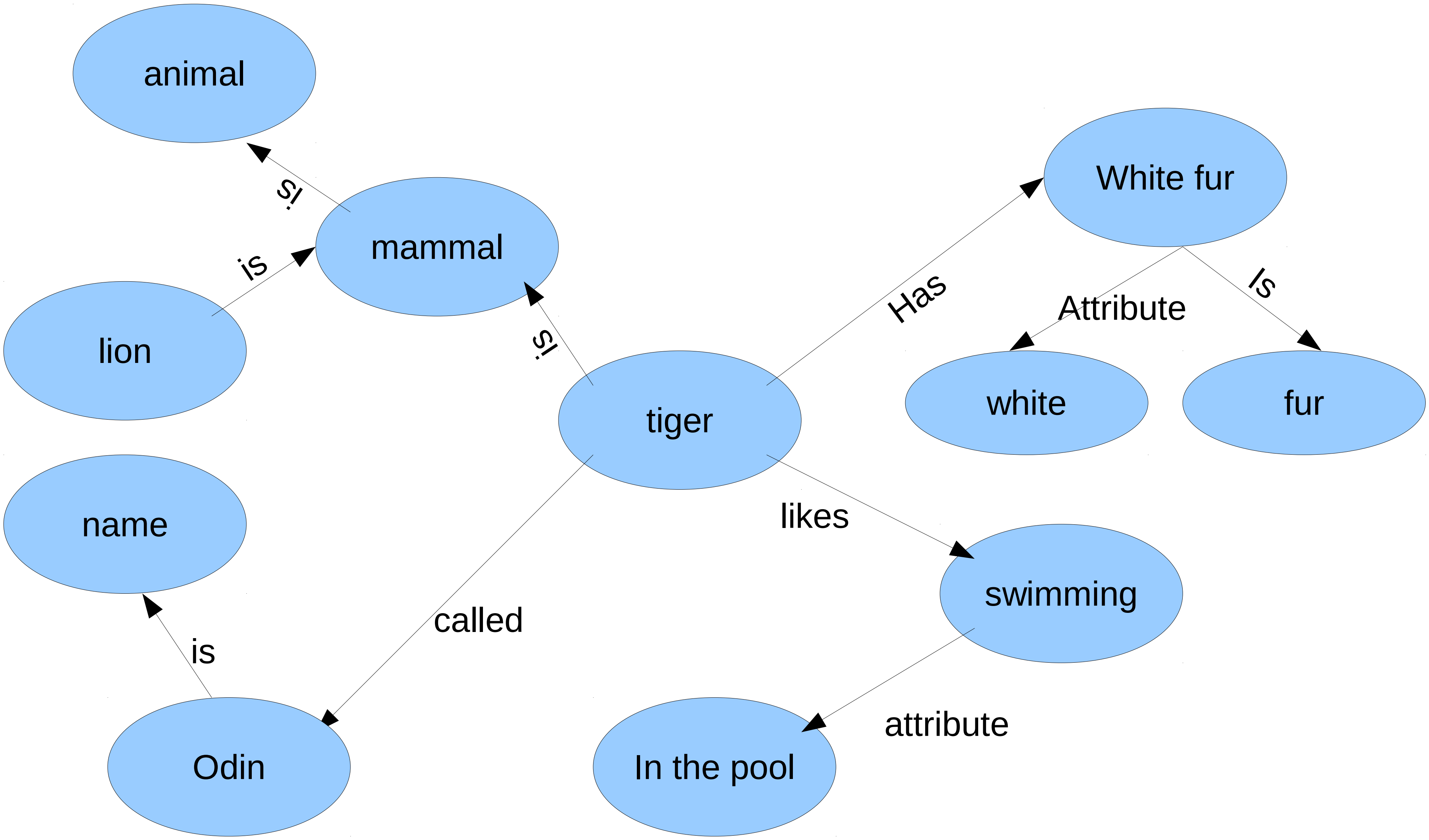
What is the ontology?

”Ontology deals with questions concerning whether entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.”

“Some claim that ontology should contain only nouns, some claim it should contain all words, some people prefer concepts or synstes.”

“I see ontology as a sort of knowledge representation – dense network of multitude of relations between any relevant objects, that can help desambiguate between senses (or at least parts of speech) .

Ontology


animal
is
is
mammal
lion
is
tiger
Has
White fur
Attribute
Is
white
fur
name
likes
called
swimming
is
Odin
In the pool
attribute

Three approaches

From the least ambitious to the least probable:

- Use of semantically annotated patterns
- Use of syntactic parser with semantically annotated rules
- Generation of the on-the-fly grammar

Which of them get the best results?

Semantically annotated patterns

- Patterns simillar to the word-sketch definitions with relations
- The alghoritm would probably extract spare ontology for small corpora... (the more the better)

Semantically annotated grammar

- Can use parser – e.g. Synt to analyze sentences.
- It would remember frequencies of all possible parses.
- It should be possible to decide, which of the possible pareses is correct based on the statistics for each word (in contrast with statistics for grammar rules).

On-the-fly grammar generation

- Extract new words from sentences that contain only one unknown word.
- Put the new words into ontology based on the location of known fitting words.
- Define the relations by used grammar rules as well as the expressions itself.
- Remember all derivation trees in the ontology along with the frequencies of every word in every leaf.
- Learn the relations on unambiguous simple sentences and rememeber statistics for each node.
- Use the learned values to decide ambiguous sentences.

Additional notes

- As a postprocessing, it should be possible to express the relations between different forms of otherwise identical relations (e.g. passive and active verb forms, cases etc...) and cross-add the relations that were learned for one form, but not for the other one.

Example Application

Anti-nuclear protestors released live cockroaches inside the
White House, and these were arrested when they left and blocked a
security gate.

Questions?
Suggestions?

Thank you for your attention!

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

EVROPSKÁ UNIE esf

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání
pro konkurenceschopnost

UNIVERSITAS MASARYKIANA BRUNENSIS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Miloš Husák
xhusak@mail.muni.cz

References

- CONCEPT ACQUISITION IN EXAMPLE-BASED GRAMMAR AUTHORING; Ye-Yi Wang and Alex Acero; Microsoft Research (2003)