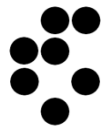
High-Coverage Extraction of Semantic Assertions from Text

Mitja Trampuš, **Dunja Mladenić**

AI Lab, Jožef Stefan Institute



The Task

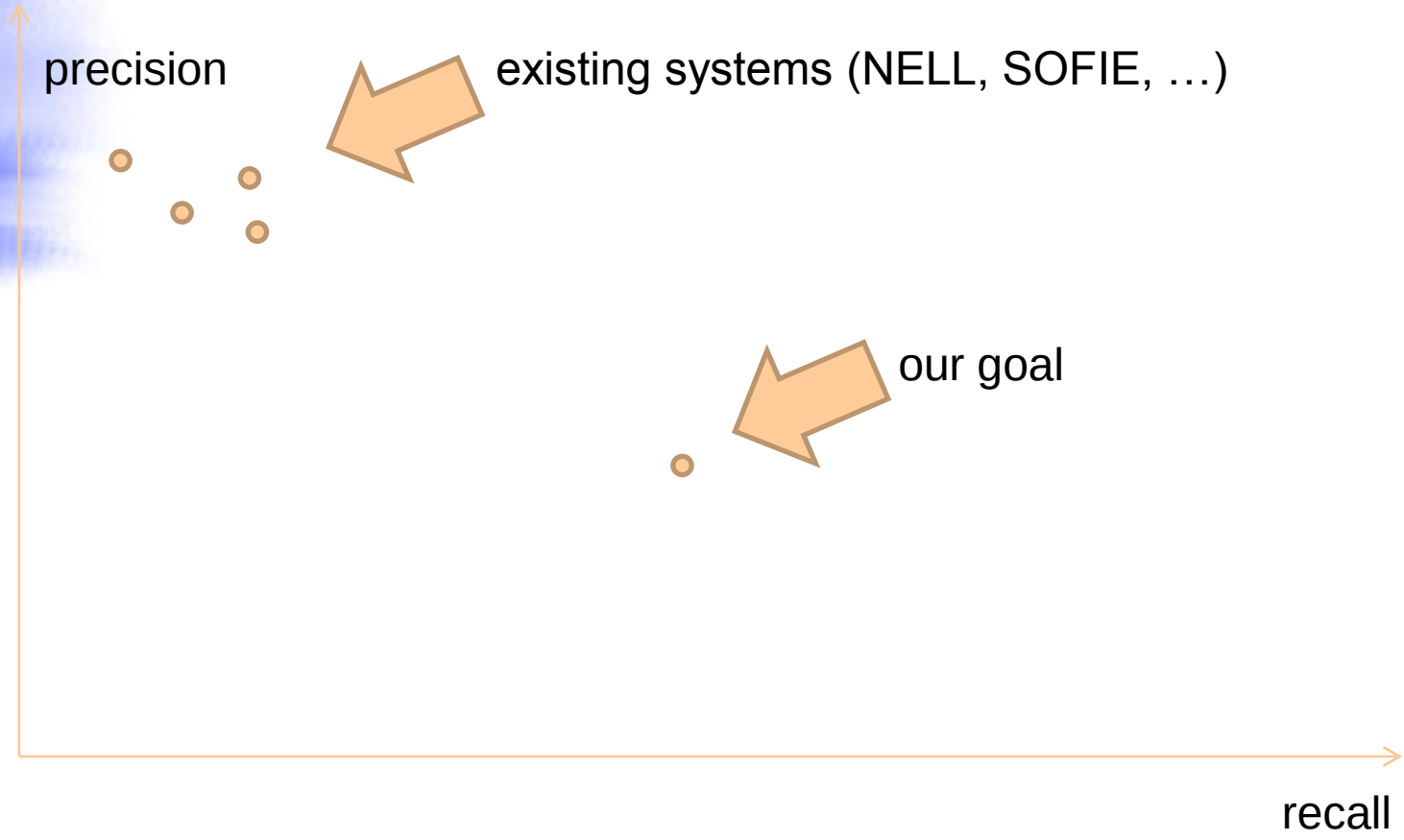
- Input: unstructured text
- Output: semantic assertions made in the text

- Simplification: extract only the simplest / most salient assertions

- **Example:**
  "John rode a bike to work yesterday." → ride(John, B), isa(B, bike)

This is a sad panda. The panda is sad because nobody knows how to extract most of the assertions while maintaining high accuracy.

The Compromise



BACKGROUND KNOWLEDGE

(Yes / No ?)

Ontology: CYC

- A large general-purpose ontology
- Uses predicate logic in line with the *ride(John, B), isa(B, bike)* example

- **Problem: No training data** for expressing text in Cyc logic. Has a natural language component, but coverage is poor beyond single words

The Glue: FrameNet

- A database / shallow ontology of *semantic frames* and their *roles*. Example:
  John rode a bike to work yesterday

  Mover *Means* *Destination*

Frame: Movement
Mover
Means
Destination
Speed
Evoked by run, walk, drive, ride, stroll, …

- **Pro:** Has training data (sentences hand-tagged with frames and role fillers)
- **Con:** Ontology of frames and roles is too shallow and sparse for general-purpose stuff.

FrameNet and Cyc

- **Idea: the best of both worlds**
  First annotate text with FrameNet (training data!), then map annotations to Cyc (rich ontology!)

Annotating text with FrameNet frames and roles

SEMANTIC ROLE LABELING (SRL)

SRL: Standard Breakdown into Tasks

- Find which frames appear in a sentence
  - Right now: naïve, recall-oriented: if a sentence contains a word W that could evoke a frame F, assume F appears in the sentence
- Detect role boundaries (= identify phrases that fill a role)
- Classify roles (= decide *which* roles they fill)

The last two steps are often done together; we follow this approach as well

SRL: Role Detection and Classification

Text

↓ Charniak parser

Parse trees

↓ For every tree node: feature extractio

Features

↓ (classical supervised learning; SVM)

Model

Feature examples:

- lemma of frame-evoking word
- Penn Treebank tag of node
- parent node's tag
- passive/active voice of sentence.
- POS tag of node's head word.

MAPPING: FRAMENET → CYC

Mapping Frames

- cca 600 “interesting” frames
  - action-related ones that map to Cyc nicely
- the mapping was done in a semi-supervised manner
  - use Cyc natural language components to identify possible matches for a frame (based on its frame-evoking words)
  - choose the best candidate by hand

Mapping Roles

- Each frame has 5-10 roles, so several roles to map – too much work for doing it by hand
- The mapping is done automatically, by computing similarity between FrameNet roles and Cyc roles.
  - Similarity measure based on BOW of roles' descriptions and on their prevalent usage (subject or object or neither)

Mapping Role-Fillers

- Essentially the Word Sense Disambiguation task
- Currently: a quick solution; a separate WSD module being prepared at the department
- Two-step approach:
  - Identify the head word of a role filler (hand written rules)
  - Use Cyc's NLP predicates to map it onto

RESULTS

Results: the Numbers

- Biggest issue: long pipeline. Accuracies (approximate):
  - Tree parsing: 90%
  - Semantic Role Labeling: 65%
  - FrameNet-Cyc Alignment: 45%
    - Cannot do better than 75% due to discrepancies between the two ontologies
  - Word Sense Disambiguation: 60%
    - Not counting personal pronouns (he, she, him, …)

Results: an Example

To understand and appreciate the Bush administration's policy regarding Israeli Prime Minister Sharon's disengagement plan, we must briefly reexamine the record. For three and a half years now, the administration's attitude toward the Israeli-Palestinian conflict/peace process has been characterized by high rhetoric and little action.

Facts from the first sentence:

(#$objectImproved #$Comprehending* #$OrganizationPolicy*)
(#$performedBy #$Comprehending* (ObjectDenotedByFn "we"))
(#$evaluationInput #$Evaluating* #$OrganizationPolicy*)
(#$performedBy #$ExercisingAuthoritativeControlOverSomething*
(ObjectDenotedByFn "we"))
(#$performedBy #$PurposefulAction* (ObjectDenotedByFn "Sharon"))

Facts from the second sentence:

(#$eventOccursAt #$DescribingSomething* #$Attitude*)
(#$senderOfInfo #$DescribingSomething* #$Action*)
(#$performedBy #$ExercisingAuthoritativeControlOverSomething*
(ObjectDenotedByFn "constitutes"))

QUESTIONS?

More questions?
mitja.trampus@ijs.si