Semantic Annotations

Jan Dědek

Department of Software Engineering
Faculty of Mathematics and Physics
Charles University in Prague

Defence of Doctoral Thesis
21st September
2012

Outline

Information Extraction (Problem)

- Let's have a text describing an acquisition event.

FIRST WISCONSIN <**FWB** > TO BUY MINNESOTA BANK

MILWAUKEE, Wis., March 26 – **First Wisconsin Corp** said it plans to acquire **Shelard Bancshares Inc** for **about** 25 **mln dlrs** in cash, its first acquisition of a Minnesota -based **bank** .

First Wisconsin said **Shelard** is the holding company for two banks with total assets of 168 mln dlrs.

First Wisconsin , which had assets at yearend of 7.1 billion dlrs, said the **Shelard** purchase price is about 12 times the 1986 earnings of the bank.

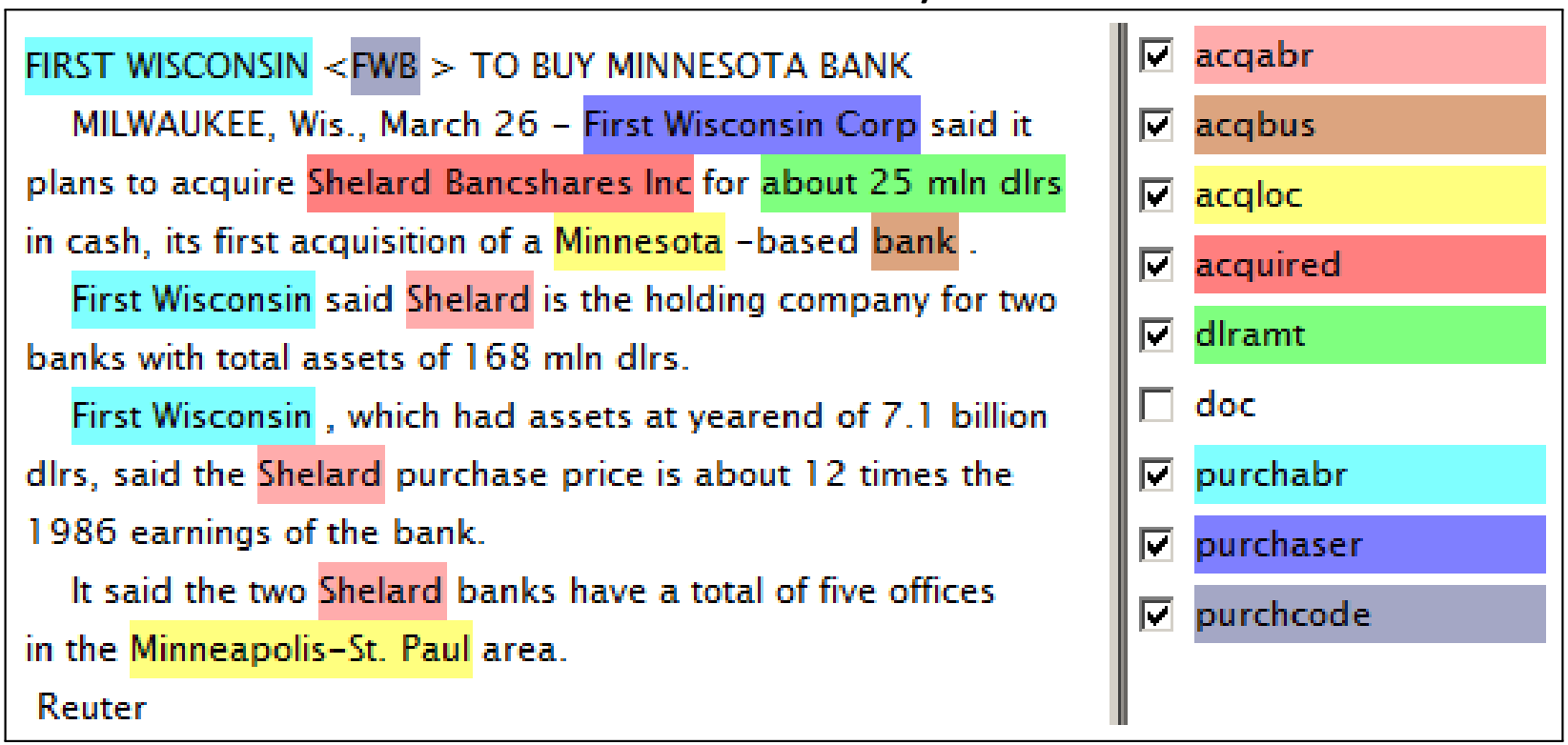
It said the two **Shelard** banks have a total of five offices in the Minneapolis-St. Paul area.

Reuter

- What was the object of the acquisition?
- Who was the buyer?
- What was the deal amount?

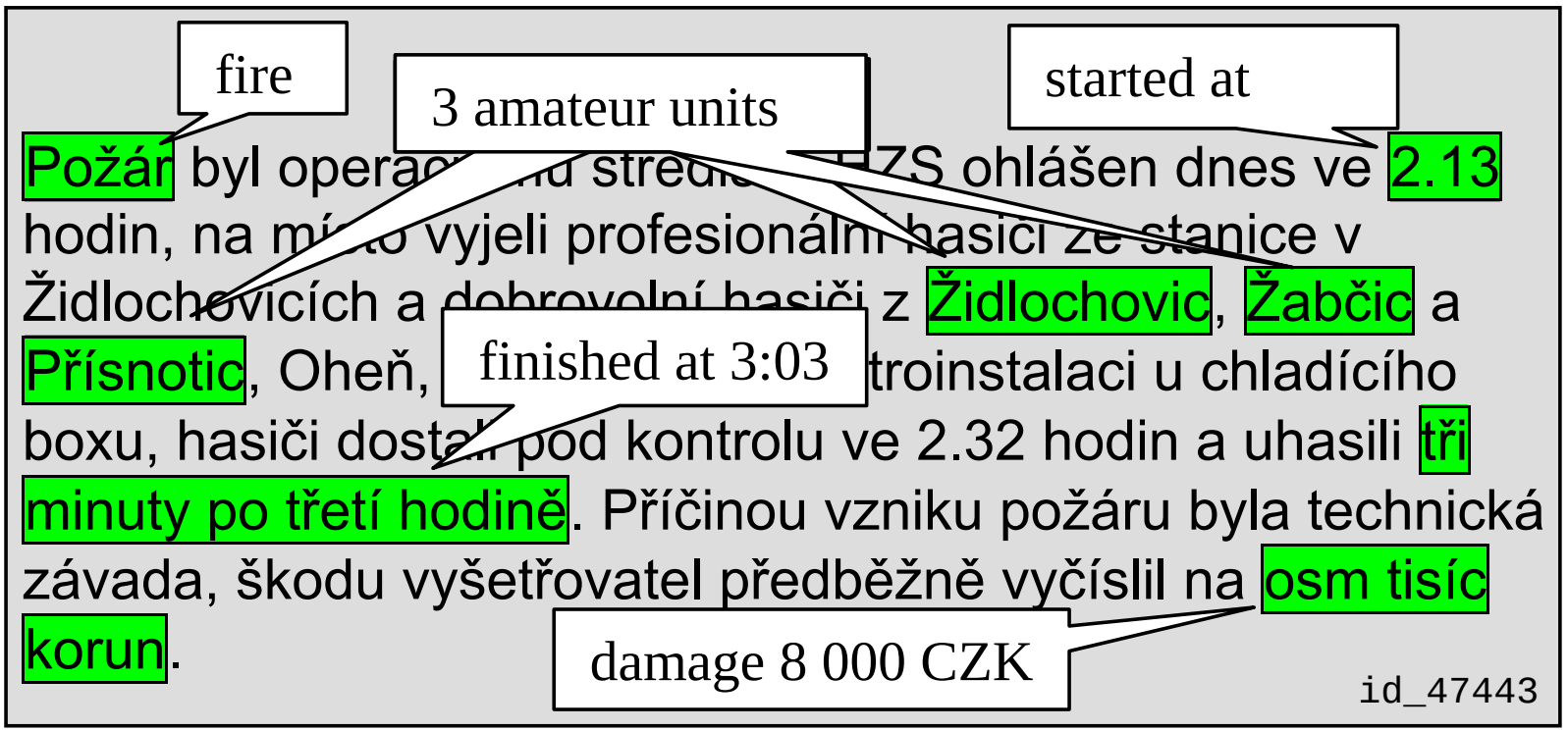
Information Extraction (Solution)

- Information Extraction tools can identify and extract such information.



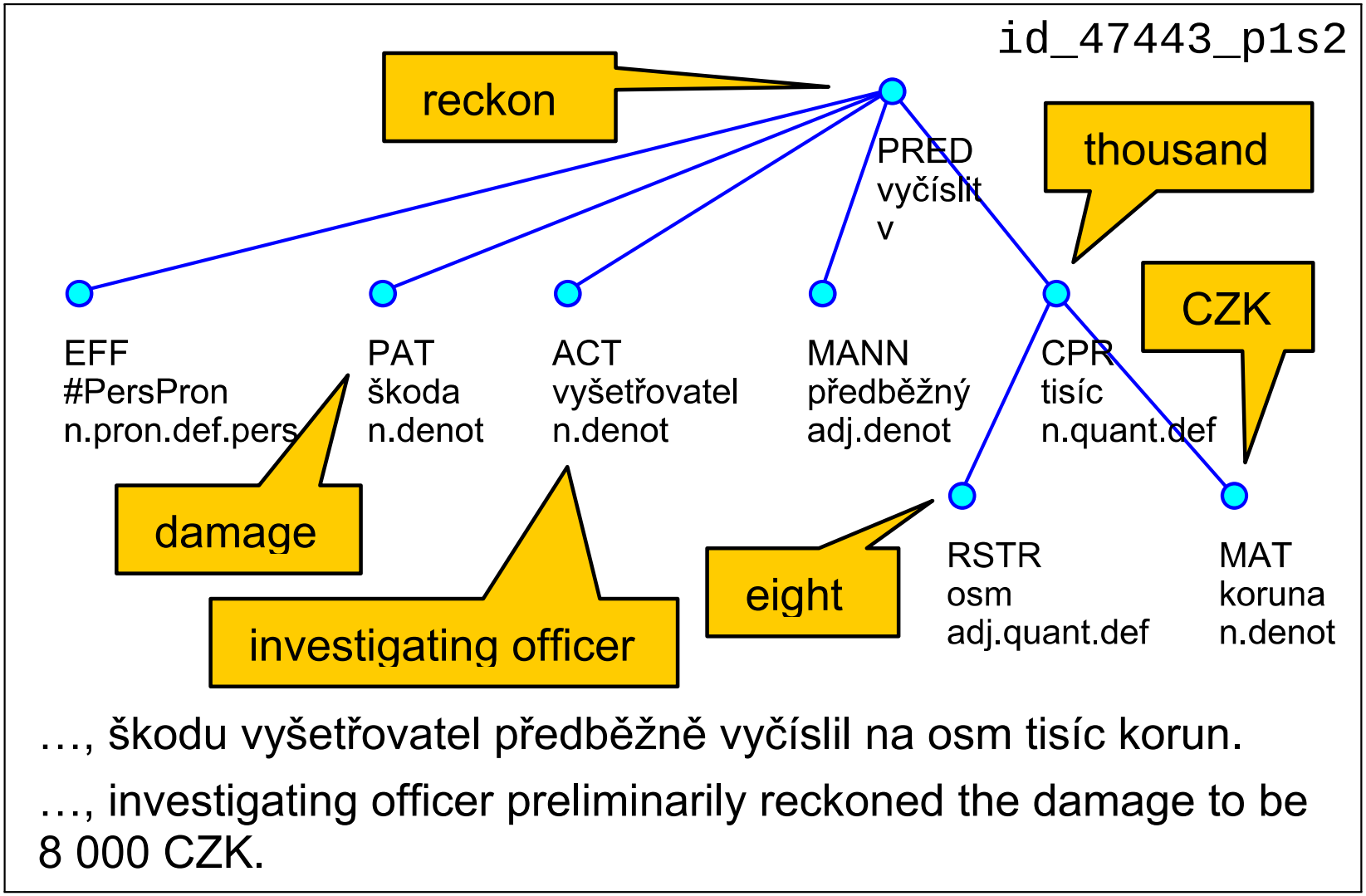
Information Extraction (Czech Example)

- Information Extraction tools can identify and extract such information.



Deep Language Parsing (Czech Example)

- Linguistic tools perform automated linguistic analysis.
- Producing so called *dependency trees*.



…, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

…, investigating officer preliminarily reckoned the damage to be 8 000 CZK.

Inductive Logic Programming

- Learning examples $E = P \cup N$ (Positive and Negative)
  - E.g. relevant and irrelevant pieces of text w.r.t. particular extraction tasks
- Background knowledge B
  - E.g. linguistic structure connecting individual words
- ILP task: To find logical program or hypothesis H such that all positive examples are covered and none negative

$$(\forall e \in P)(B \cup H \models e) \ \& \ (\forall n \in N)(B \cup H \not\models n).$$

  - E.g. to find common pattern (in the linguistic structure) present around every relevant piece of text and none irrelevant.

Four Main Topics

- **Manual Design of Extraction Rules**
- **Induction of Extraction Rules**
- **Shareable Extraction Ontologies**
- **Fuzzy ILP Document Classification**

Manual Design of Extraction Rules

Slides about the topic *Manual Design of Extraction Rules* will have **brown** headline background.

Induction of Extraction Rules

Slides about the topic *Induction of Extraction Rules* will have **green** headline background.

Shareable Extraction Ontologies

Slides about the topic *Shareable Extraction Ontologies* will have **cyan** headline background.

Fuzzy ILP Document Classification

Slides about the topic *Fuzzy ILP Document Classification* will have **magenta** headline background.

Ordinary