Semantic Annotation Semantically: Using a Shareable Extraction Ontology and a Reasoner

Jan Dědek Peter Vojtáš

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

SEMAPRO 2011
The Fifth International Conference on Advances
in Semantic Processing
November 20-25, 2011, Lisbon, Portugal

Outline

Semantic Annotation of Text (Problem)

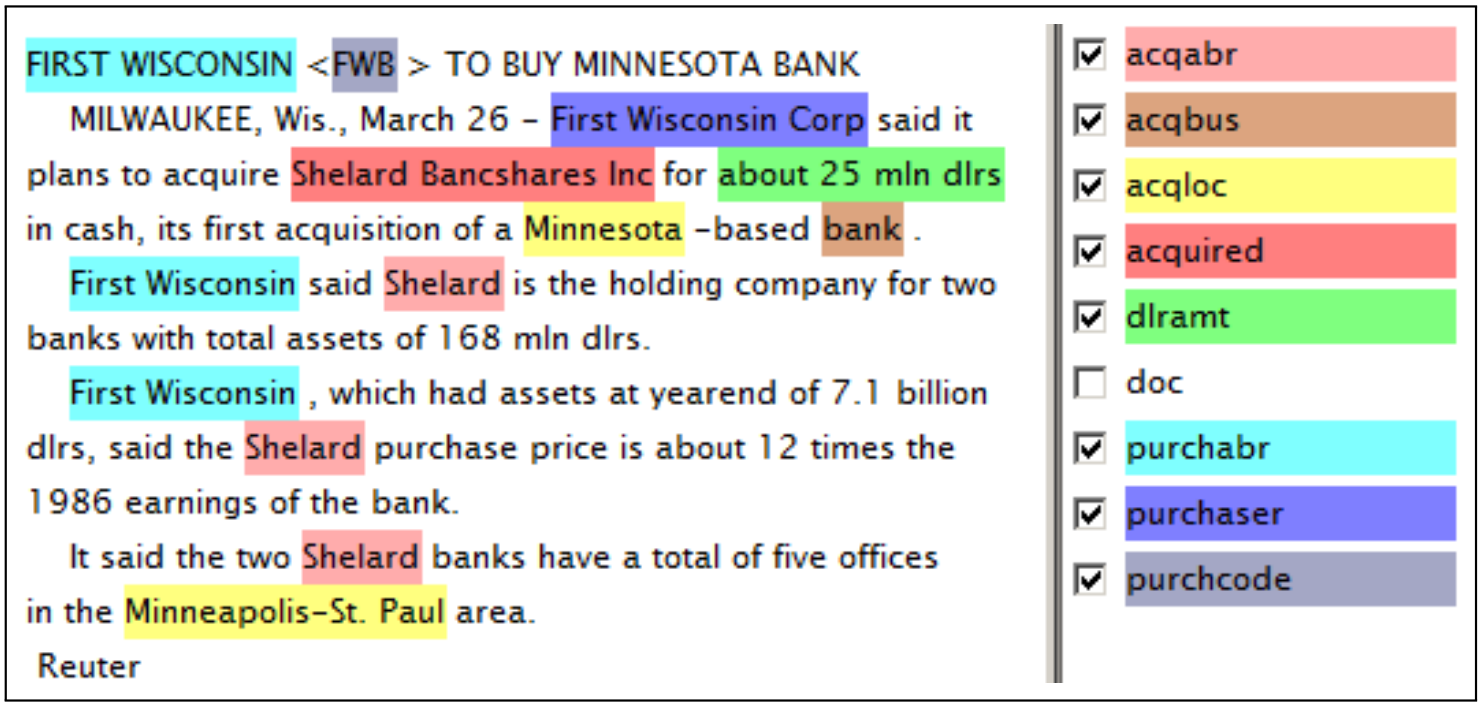
- Let's have a text describing an acquisition event.

> 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?

Semantic Annotation of Text (Solution)

- Well, there are Information Extraction tools that can identify and extract such information.
  - Of course not 100% accurete...



- The tools can also interpret such information in terms of a Semantic Web Ontology.

Extraction Ontology

- And even more!
- The knowledge (extraction model) used in the extraction process can itself be saved in an ontology.
  - So called Extraction Ontology

- D. W. Embley, “Toward semantic understanding: an approach based on information extraction ontologies,” in *ADC ’04*. Darlinghurst: ACS, 2004, pp. 3–12.
- M. Labský et al., “The Ex Project: Web Information Extraction Using Extraction Ontologies,” in *Knowledge Discovery Enhanced with Semantic and Social Information*, ser. Studies in Comput. Intellig. Springer, 2009, vol. 220, pp. 71–88.

Are Such Extraction Ontologies Shareable?

Question

- But are these Extraction Ontologies shareable?
- Is it possible to use them outside of the original tool?

Answer

? ...

Example [Embley]

CarAds Extraction Ontology

```
<ObjectSet x="329" y="51" lexical="true" name="Mileage" id="osmx50">
    <DataFrame>
      <InternalRepresentation>
        <DataType typeName="String"/>
      </InternalRepresentation>
      <ValuePhraseList>
        <ValuePhrase hint="Mileage Pattern 1">
          <ValueExpression color="ffffff">
            <ExpressionText>[1-9]\d{0,2}[kK]</ExpressionText>
          </ValueExpression>
          <LeftContextExpression color="ffffff">
          ...
      <KeywordPhraseList>
        <KeywordPhrase hint="New phrase 1">
          <KeywordExpression color="ffffff">
            <ExpressionText>\bmiles\b</ExpressionText>
            ...
```

http://www.deg.byu.edu/presentations/ColloqSemanticUnderstanding.Jun2006.ppt

Example [Labský]

```
<attribute id="city" type="name" card="0-1" eng="0.50">
  <pattern id="eu_big_cities" src="eu_cities.txt" encoding="utf-8"/>
  <pattern id="us_big_cities" src="us_cities.txt" encoding="iso-8859-1" />
  <pattern id="all_cities">
    <pattern ref="europe_big_cities"/> | <pattern ref="us_big_cities"/>
  </pattern>
  <pattern id="city_suffix"> City | Village | Town </pattern>
  <value>
    <pattern p="0.55" cover="0.05">
      <tok case="CA|UC"/>{1,2} <pattern ref="city_suffix"/>
    </pattern>
    <pattern p="0.80" cover="0.50" ignore="case" case="^CA|UC">
      <tok case="all_cities"/> <pattern ref="city_suffix"/>?
    </pattern>
  </value>
```

... pattern for matching city names which will utilize large lists of known cities.

`http://eso.vse.cz/~labsky/ex/ex_tutorial.pdf`

Are Such Extraction Ontologies Shareable?

Question

- But are these Extraction Ontologies shareable?
- Is it possible to use them outside of the original tool?

Answer

Not yet.
Although they are conceptually modeled, the native tool is the only tool capable of interpretation of these models (ontologies).

Our Idea: Shareable Extraction Ontologies

- **Publish** your extraction ontology on the web and **anybody can use it with a standard reasoner**!
- How this can be done?
- Can a reasoner process textual reports?

Text Processing by a Reasoner

- Can a reasoner process textual reports?
- Well, not from the beginning :-(
- Semantic Web reasoners work only with Semantic Web ontologies.
  (Not with textual documents)
- They can read RDF & OWL files.
- And also XHTML and XML files, but ...
  - But only the semantic part provided by RDFa annotations or by a GRDDL transformation.
- So we need some preprocessing...

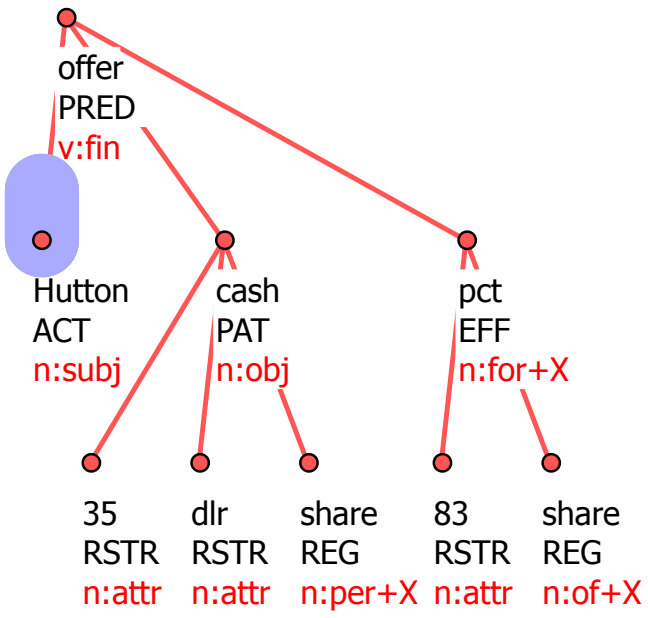
Document Ontologies

- ...we need some preprocessing.
- That will convert a (textual) document to a ontology.
- TXT (PDF, HTML) $\rightarrow$ RDF (OWL)
- Document $\rightarrow$ Document Ontology (reasoner readable)
- We call the ontological representation of a document a Document Ontology.
- Document Ontology contains:
  - All words of a document
  - Other units necessary for:
    - Information Extraction
    - Document reconstruction (Document Ontology $\rightarrow$ Document)

From now the idea is strait forward!

- The **semantics of the Extraction Model** has to be converted to the **semantics of the Extraction Ontology**.
- A reasoner can then interpret the Extraction Ontology in the same way as the original tool would interpret the Extraction Model.
- Application of the Extraction Ontology on the Document Ontology by a reasoner will result in "Annotated Document Ontology"
  - Annotated Document can be reconstructed from it.

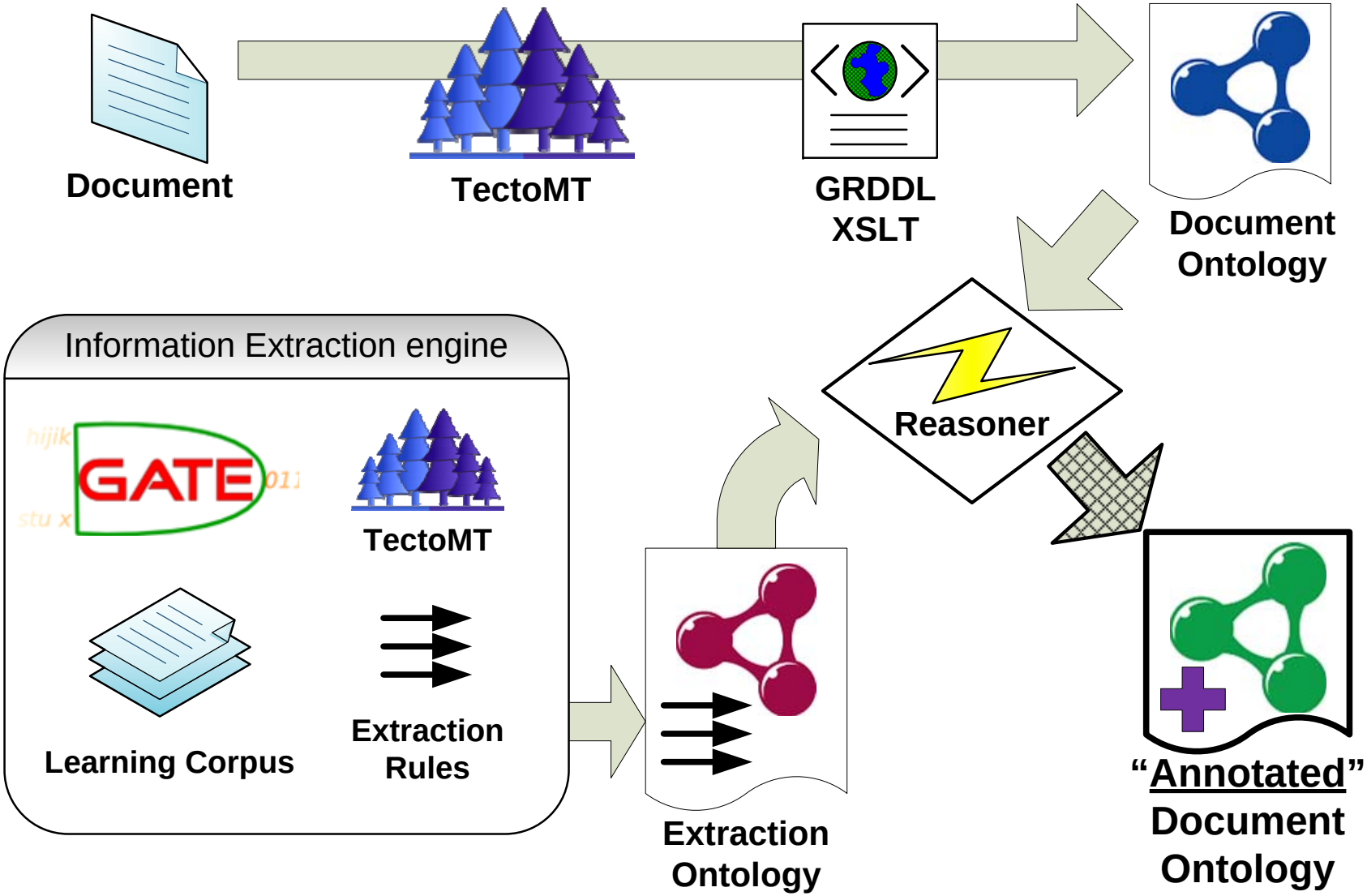
Information Extraction Engine



"Hutton is offering 35 dlrs cash per share for 83 pct of the shares."

- Based on dependency parsing
- Linguistic trees
  - Transfered to Document Ontology
- Extraction rules
  - Tree patterns
- Machine learning of rules using Inductive Logic Programming (ILP)
  - It works – really :-)
  - See e.g. [8] (self citation)
- Rules can be also handcrafted.
- These extraction rules can be directly exported to ontology using SWRL!

System Architecture



Examples of native extraction rules

[Rule 1] [Pos cover = 23 Neg cover = 6]

```
mention_root(acquired,A) :-
    'lex.rf'(B,A), t_lemma(B,'Inc'),
    tDependency(C,B), tDependency(C,D),
    formeme(D,'n:in+X'), tDependency(E,C).
```

[Rule 11] [Pos cover = 25 Neg cover = 6]

```
mention_root(acquired,A) :-
    'lex.rf'(B,A), t_lemma(B,'Inc'),
    tDependency(C,B), formeme(C,'n:obj'),
    tDependency(C,D), functor(D,'APP').
```

[Rule 75] [Pos cover = 14 Neg cover = 1]

```
mention_root(acquired,A) :-
    'lex.rf'(B,A), t_lemma(B,'Inc'),
    functor(B,'APP'), tDependency(C,B),
    number(C,pl).
```

SWRL (OWL/XML) Representation

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Ontology [
  <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
  <!ENTITY pml "http://ufal.mff.cuni.cz/pdt/pml/" >
]>
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
  ontologyIRI="http://czsem.berlios.de/onto ... rules.owl">
  <DLSafeRule>
    <Body>
      <ObjectPropertyAtom>
        <ObjectProperty IRI="&pml;lex.rf" />
        <Variable IRI="urn:swrl#b" />
        <Variable IRI="urn:swrl#a" />
      </ObjectPropertyAtom>
...
      <DataPropertyAtom>
        <DataProperty IRI="&pml;number" />
        <Variable IRI="urn:swrl#c" />
        <Literal>pl</Literal>
      </DataPropertyAtom>
    </Body>
    <Head>
      <DataPropertyAtom>
        <DataProperty IRI="&pml;mention_root" />
        <Literal>acquired</Literal>
        <Variable IRI="urn:swrl#a" />
      </DataPropertyAtom>
    </Head>
  </DLSafeRule>
</Ontology>
```

The same in Jena rules format

```
@prefix pml: <http://ufal.mff.cuni.cz/pdt/pml/>.
[rule-75:
        ( ?b pml:lex.rf ?a )
        ( ?c pml:tDependency ?b )
        ( ?b pml:functor 'APP' )
        ( ?c pml:number 'pl' )
        ( ?b pml:t_lemma 'Inc' )
    ->
        ( ?a pml:mention_root 'acquired' )
]
```

Datasets & Reasoners

dataset	domain	language	num of files	data size (MB)	num of rules
czech_fireman	accidents	Czech	50	16	2
acquisitions	finance	English	600	126	113

reasoner	**czech_fireman**	stdev	**acquisitions-v1.1**	stdev
Jena	161 s	0.226	1259 s	3.579
HermiT	219 s	1.636	$\gg$ 13 hours	
Pellet	11 s	0.062	503 s	4.145
FaCT++	Does not support rules.			

- How poor is the poor performance? :-)

Conclusion

- Idea of Shareable Extraction Ontologies presented
  - With the drawback of the necessity of document preprocessing (TXT $\rightarrow$ RDF)
- Realization of the idea demonstrated by adaptation of our IE system
- Performance evaluation experiment done
  - Poor performance (as expected)
  - But bearable
- New public OWL (+SWRL) reasoning benchmark created as a side effect

Future Work

- Compare the performance with rules translated to SPARQL
  - Increased performance could be expected
- Annotated Document Ontologies $\rightarrow$ Fact Ontologies
  - Data integration and duplicity of information issues
  - Technological problems: creating new individuals during (safe) reasoning
- General Shareable Extraction Ontology creation guidelines
  - E.g. how to encode a gazetteer list this way

Thank you for your attention!

Questions?