

Plnění ontologie pomocí GATE 5

Tomáš Kliegr

21.4.2009

Plnění ontologií (Ontology population)

- Do existující ontologie přidáváme instance konceptů
- Tato úloha může probíhat ručně, např. v Protege, nebo je možné instance extrahovat z textů automaticky
- Přístupy k automatické extrakci jsou založeny na statistické analýze textu nebo na lexiko-syntaktických vzorech
- GATE je vhodný především pro extrakci pomocí lexiko-syntaktických vzorů

Přehled procesu

- Aby mohla být ontologie naplněna, je třeba identifikovat instance cílových konceptů v textu.
- Instance je typicky reprezentována jedním nebo vícero po sobě následujícími *tokens*.
- To zda je skupina tokenů identifikována jako instance určitého konceptu se určuje na základě podmínek.
- Tyto podmínky nepracují typicky s konkrétními řetězci, ale s vlastnostmi tokenů zachycenými ve formě anotací přiřazených různými lingvistickými nástroji (processing resources) obsaženými v GATE.
- Anotace lze zpracovat pomocí jazyka JAPE

Příklady vlastností

- Příkladem vlastností, které lze využít je:
 - Slovní druh tokenu (Part-Of-Speech)
 - Kmen (stem) nebo Kořen (lemma) tokenu
 - Seskupení tokenů do jmenné fráze (noun phrase chunking)- výraz jehož hlavou je podstatné jméno
 - Přítomnost tokenu nebo skupiny tokenů na seznamu (gazetteer)
 - Řešení anafor - typicky na co odkazuje zájmeno (v Gate Pronominal coreference)
 - Token je součástí nepřímé řeči (Reported Speech)
 - Token je součástí klíčové fráze (Keyphrase)
 - Název elementu v původním dokumentu, v jehož obsahu byl token umístěn (Original Markup)
- Tyto vlastnosti jsou zjišťovány moduly (PRs) a ukládány do anotací
- Operace nad anotacemi vč. tvorby nových anotací: JAPE: Java Annotations Pattern Engine

Příklad

- Následující ontologii chceme naplnit jmény politických stran

The screenshot shows an ontology editor interface. On the left, a tree view displays the class hierarchy under 'Classes and Instances'. The 'MoneyAmount' class is highlighted. On the right, the 'Properties' tab is active, showing details for the selected class.

Classes & Instances | **Properties**

Classes and Instances

- Date
- Location
 - City
 - Country
 - Province
 - Region
- MoneyAmount**
- Organization
 - Charity
 - Company
 - Government
 - Department
 - Ministry
 - Party
- Person
 - Businessman
 - MediaPerson
 - Politician
 - Sportsman

▼ Resource Information

- MoneyAmount
 - URI: <http://www.owl-ontologies.com/unname>
 - TYPE: Ontology Class

▼ Direct Super Classes

▼ All Super Classes

▼ Direct Sub Classes

▼ All Sub Classes

▼ Equivalent Classes

▼ Property Types

- comment [ALL RESOURCES]
- isDefinedBy [ALL RESOURCES]
- label [ALL RESOURCES]
- seeAlso [ALL RESOURCES]
- versionInfo [ALL RESOURCES]

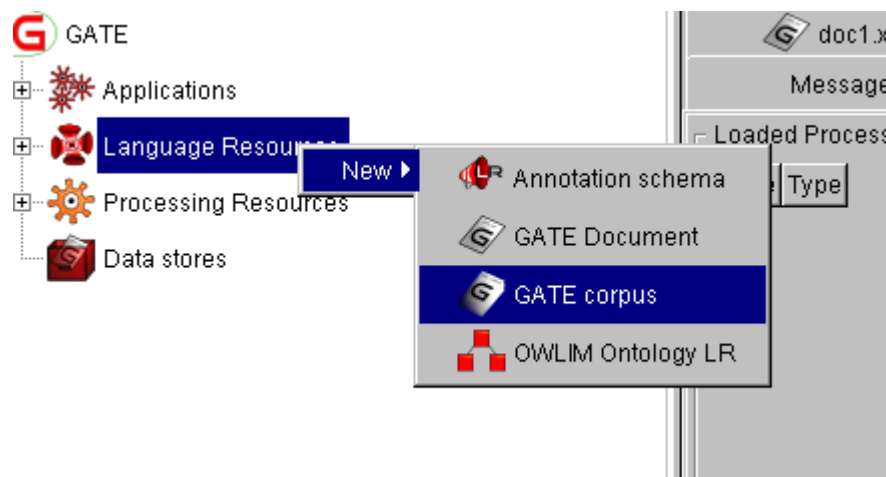
▼ Property Values

▼ Instances

Načtení dat

- Konkrétní politické strany lze najít v textu, který máme k dispozici
- **<P>Tributes poured in from around the world Thursday to the late Labour Party leader John Smith, who died earlier from a massive heart attack aged 55.</P>**
- **<P>In Washington, the US State Department issued a statement regretting "the untimely death" of the rapier-tongued Scottish barrister and parliamentarian.</P>**
- ...

Nejdříve pod Language Resources založit nový Corpus, následně načíst dokument (New Document) a přiřadit ho do nově vytvořeného korpusu.



Načtení PRs

- Před definicí součástí Pipeline je třeba načíst použité lingvistické moduly

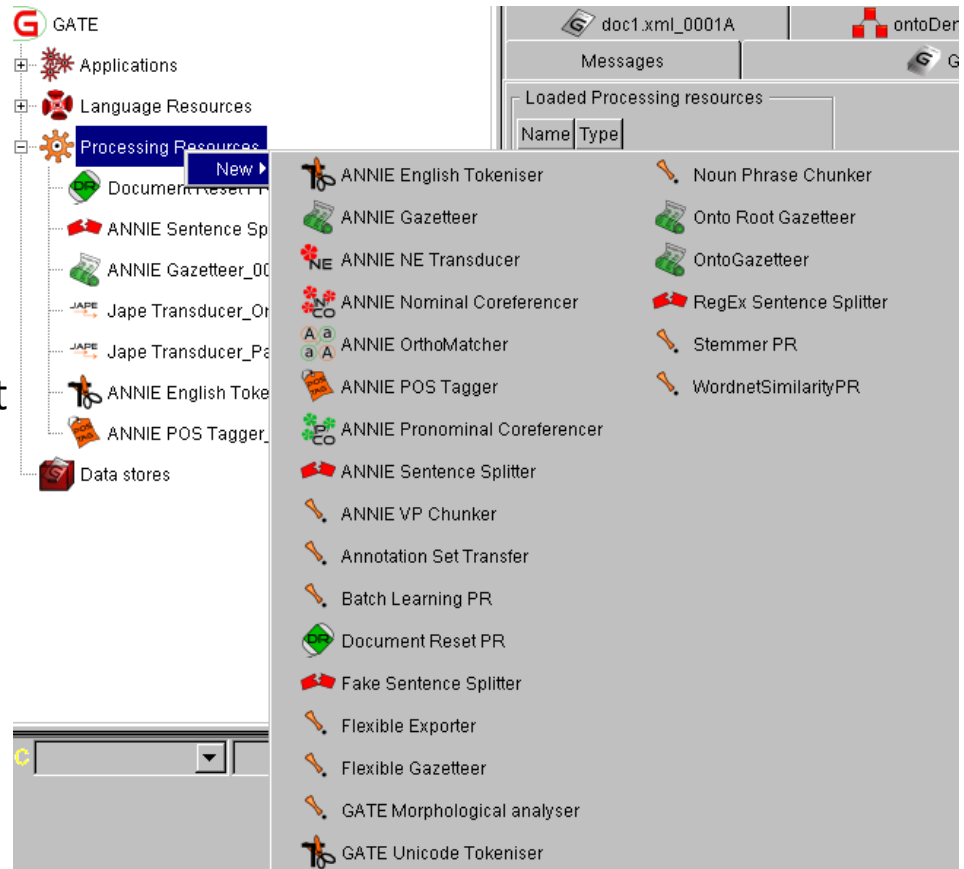
Nejsou-li v nabídce, je třeba je nahrát přes File-manage CREOLE plugins

JAPE Transducery vyžadují nastavení cest K JAPE Gramatikám

Použijte:

Onto5.Jape













Party3.Jape



Nastavení Pipeline

- Provádí se pomocí Processing Resources (PRs), které jsou sekvenčně spouštěny pomocí Pipeline.
- Pod uzlem Application vytvořit New Corpus Pipeline. Parametrem je korpus vytvořený v předcházejícím kroku.

PRs nabízené v okně Loaded LRs uspořádejte násl. způsobem:

!	Name	Type
	 Document Reset PR_00049	Document Reset PR
	 ANNIE English Tokeniser_0002E	ANNIE English Tokeniser
	 ANNIE Sentence Splitter_00044	ANNIE Sentence Splitter
	 ANNIE POS Tagger_0002B	ANNIE POS Tagger
	 ANNIE Gazetteer_0003C	ANNIE Gazetteer
	 Jape Transducer_Party3	Jape Transducer

- Pipeline spustíte

Výsledek 1

The screenshot displays the GATE (General Architecture for Text Engineering) software interface. On the left, a sidebar shows a project tree with 'Language Resources' containing 'ontoDemo.rdf_00040', 'doc1.xml_0001A', and 'GATE corpus_00016', and 'Processing Resources' containing 'Data stores'. At the bottom left, a table shows document properties:

C	MimeType	▼	text/xml
C	gate.SourceURL	▼	file:/D:/KI
C		▼	

The central text area contains the following content with NLP annotations:

Tributes pour in for late **British Labour Party** leader
UNDATED, May 12 (AFP) |

Tributes poured in from around the world Thursday to the late **Labour Party** leader John Smith, who died earlier from a massive heart attack aged 55.

In Washington, the US State Department issued a statement regretting "the untimely death" of the rapier-tongued Scottish barrister and parliamentarian.

"Mr. Smith, throughout his distinguished career in government and in opposition, left a profound impression on the history of his party and his country," State Department spokesman Michael McCurry said.

"Secretary (of State Warren) Christopher extends his deepest condolences to Mrs. Smith and to the Smith children."

In Bonn, the head of the **German Social Democratic Party**, Rudolf Scharping, said in a statement he was "very affected by the sudden death of John Smith.

On the right, a filter menu is visible with the following options:

- Lookup
- Party
- Sentence
- SpaceToken
- Split
- Token
- ▶ Original markups

JAPE pro extrakci politických stran

```
Phase: party
Input: Token
```

```
Rule: party_name2
```

```
(
  ({Token.category=="NNP"} | {Token.category=="NNPS"})+
  ({Token.string == "Party"})
):party
-->
  :party.Party = { rule = "party_name2", class="Party" }
```

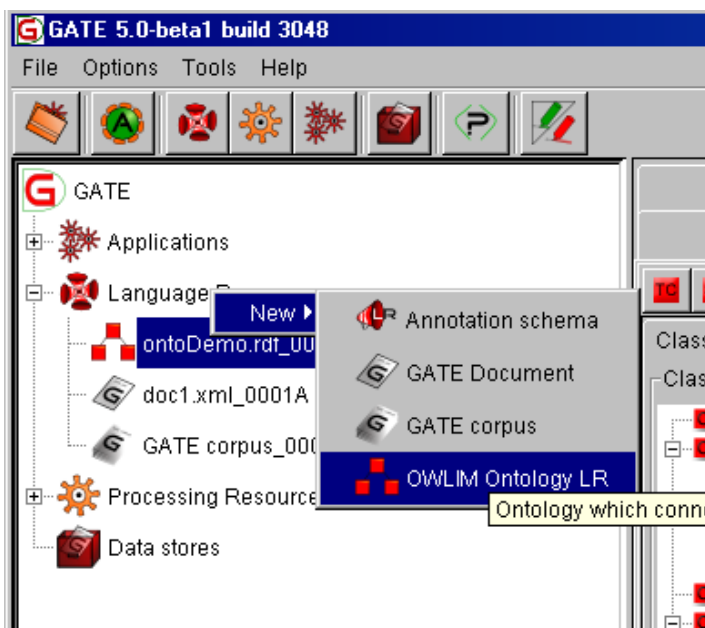
```
Rule: party_name3
```

```
(
  {Token.string == "Party"}
  {Token.string == "of"}
  ({Token.category == "NNP"} | {Token.category == "NNPS"})*
):party
-->
  :party.Party = { rule = "party_name3", class="Party" }
```

Vložení výsledku do ontologie

- Načíst ontologii pomocí OWLIM pluginu
- Vytvořit manipulační JAPE gramatiku, která vezme anotace určitého typu a vloží je jako instance předdefinovaného konceptu

Použijte soubor
ontoDemo.rdf



JAPE pro vkládání do ontologie

- Část 1: vytažení jména strany

```
Phase: onto
Input: Party

Rule: FindEntities
({Party}):party
-->
{
  //find the annotation matched by LHS
  //we know the annotation set returned
  //will always contain a single annotation
  Annotation partyAnn = (Annotation)
    ((AnnotationSet)bindings.get("party")).
    iterator().next();

  //find the class of the party
  String className = (String)partyAnn.getFeatures().
    get(gate.creole.ANNIEConstants.LOOKUP_CLASS_FEATURE_NAME);

  //find the text covered by the annotation
  String partyName;
  try {
    partyName = doc.getContent().
      getContent(
        partyAnn.getStartNode().getOffset(),
        partyAnn.getEndNode().getOffset()).
      toString();
  } catch (InvalidOffsetException e) {
    throw new GateRuntimeException(e); //this should never happen
  }
}
```

JAPE pro vkládání do ontologie

- Část 2: nalezení konceptu „Party“ v ontologii a vložení instance

```
//add the instance to the ontology
//get the first class with that name
gate.creole.ontology.OCClass aClass = null;
for (gate.creole.ontology.OResource aResource :
     ontology.getOResourcesByName(className)) {
    if (aResource instanceof gate.creole.ontology.OCClass) {
        aClass = (gate.creole.ontology.OCClass) aResource;
        break;
    }
}
if (aClass == null) {
    System.err.println("Error class " + className + " does not exist!");
} else {
    //check if the instance already exists
    //assume that the partyName instances are unique instances
    gate.creole.ontology.URI uri = gate.creole.ontology.OntologyUtilities
        .createURI(ontology, partyName, false);
    if (!ontology.containsOInstance(uri)) {
        // create the instance in the ontology
        ontology.addOInstance(uri, aClass);
    }
}
}
```

Aktualizace pipeline

The screenshot displays the GATE 5.0-beta1 build 3048 interface. The left sidebar shows a tree view of resources, with 'Corpus Pipeline_00027' selected under 'Applications'. The main workspace shows the 'Messages' tab with 'GATE corpus_00016' selected. Below this, the 'Loaded Processing resources' and 'Selected Processing resources' sections are visible. The 'Selected Processing resources' table lists the following components:

!	Name	Type
	Document Reset PR_00049	Document Reset PR
	ANNIE English Tokeniser_0002E	ANNIE English Toker
	ANNIE Sentence Splitter_00044	ANNIE Sentence Spli
	ANNIE POS Tagger_0002B	ANNIE POS Tagger
	ANNIE Gazetteer_0003C	ANNIE Gazetteer
	Jape Transducer_Party3	Jape Transducer
	Jape Transducer_Onto5	Jape Transducer

Below the resource lists, the 'Corpus' dropdown is set to 'GATE corpus_00016'. A message states: 'The **corpus** and **document** parameters are not available as they are automatically set by the controller!'. The parameters for the 'Jape Transducer_Onto5' Jape Transducer are shown in the following table:

Name	Type	Required	Value
inputASName	java.lang.String		
ontology	gate.creole.ontology.Ontology		ontoDemo.rdf_00040
outputASName	java.lang.String		

A 'Run' button is located at the bottom right of the configuration area. The bottom status bar indicates 'Serial Application Editor Initialisation Parameters'.

Výsledek 2

The screenshot displays the Jape (Jape Transducer) interface for editing an ontology. The main window is titled 'ontoDemo.rdf_00040' and shows a tree view of classes and instances. The 'Party' class is selected and highlighted in blue.

Classes & Instances

- Classes and Instances
 - Date
 - Location
 - City
 - Country
 - Province
 - Region
 - MoneyAmount
 - Organization
 - Charity
 - Company
 - Government
 - Department
 - Ministry
 - Party (selected)
 - British Labour Party
 - LabourParty
 - German Social Democratic Party
 - Labour Party
 - Conservative Party
 - French Socialist Party
 - Party of
 - Party of European Socialists
 - Party of European
 - Portuguese Socialist Party
 - Person
 - Businessman
 - MediaPerson
 - Politician
 - Sportsman

Resource Information

- Party
 - URI: <http://www.owl-ontologies.com>
 - TYPE: Ontology Class

Direct Super Classes

- Government

All Super Classes

- Government
- Organization

Direct Sub Classes

All Sub Classes

Equivalent Classes

Property Types

- comment [ALL RESOURCES]
- isDefinedBy [ALL RESOURCES]
- label [ALL RESOURCES]
- seeAlso [ALL RESOURCES]
- versionInfo [ALL RESOURCES]

Property Values

Instances

- British Labour Party
- Conservative Party
- French Socialist Party
- German Social Democratic Party
- LabourParty
- Labour Party
- Party of
- Party of European

Úkol

- Jakým způsobem je možné výsledky zlepšit pomocí gazetteru?
- Seznam typů a subtypů uložených v předdefinovaném gazetteeru naleznete po kliknutí na gazetteer pod uzlem Processing Resources.