Úvod do GATE

Marek Schmidt
mailto:xschmi01@stud.fit.vutbr.cz

22.4.2008

Co je GATE?

- General Architecture for Text Engineering

  http://gate.ac.uk/

- Framework (Java)
- Komponenty
- GUI

Framework 1/2

- Language Resources (LR)
  - (dokumenty, korpusy, ontologie)
- Processing Resources (PR)
  - (tokeniser, tagger, parser)
- Visual Resources
  - (grafická udělátka)

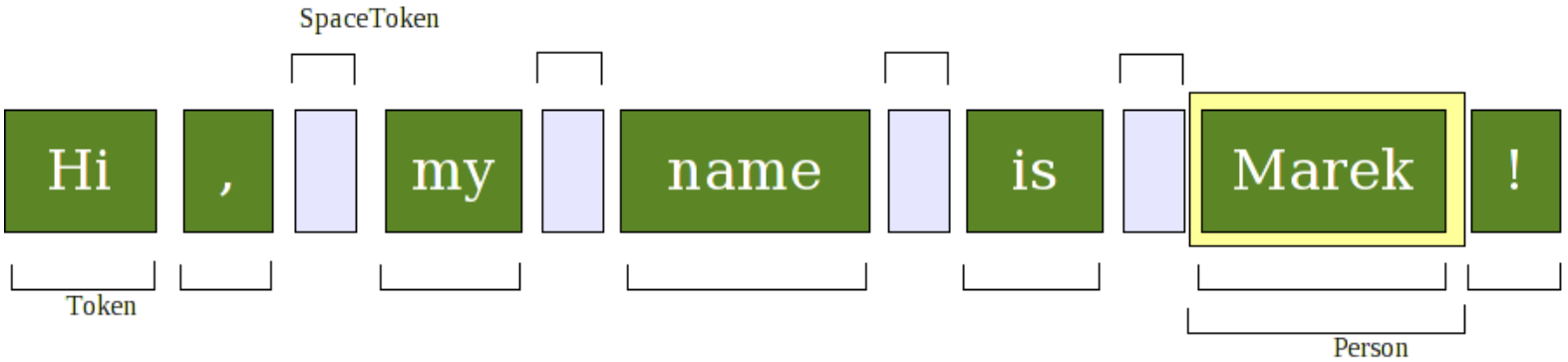
Framework 2/2

- Application
  - Seskupení LR, Pipeline
- Datastore
  - perzistentní uložení LR
    - Serial Datastore (Filesystem)
    - Database Datastore (SQL)

Annotation

- Orientovaný acyklický graf (DAG)
- Uzly
  - Pozice (znak) v textu
- Hrany
  - ID
  - Typ
  - FeatureMap

Annotation



Type	Set	Start	End	Features
Token		0	2	{category=NNP, kind=word, length=2, orth=upperInitial, string=Hi}
Token		2	3	{category=,, kind=punctuation, length=1, string=,}
SpaceToken		3	4	{kind=space, length=1, string= }
Token		4	6	{category=PRP$, kind=word, length=2, orth=lowercase, string=my}
SpaceToken		6	7	{kind=space, length=1, string= }
Token		7	11	{category=NN, kind=word, length=4, orth=lowercase, string=name}
SpaceToken		11	12	{kind=space, length=1, string= }
Token		12	14	{category=VBZ, kind=word, length=2, orth=lowercase, string=is}
SpaceToken		14	15	{kind=space, length=1, string= }
Token		15	20	{category=NNP, kind=word, length=5, orth=upperInitial, string=Marek}
Person		15	20	{gender=male, rule=PersonFinal, rule1=GazPersonFirst}
Token		20	21	{category=., kind=punctuation, length=1, string=!}

JAPE

- Java Annotation Patterns Engine
- Transdukce nad anotacemi
- Regulární výrazy

JAPE

```
Phase: jobs
Input: Person JobTitle Token
Rule: jobs1
(
 ({Person}):person
 {Token.string == "is"}
 {Token.string == "a"}
 ({JobTitle}):jobtitle
)
-->
 :person.PersonWithJob = {
    rule = "jobs1",
    string=:person.Person.string,
    title=:jobtitle.JobTitle.string}
```

ANNIE

A Nearly-New Information Extraction system

- Tokeniser
- Gazetteer
- Sentence Splitter
- Part of Speech Tagger
- ... a další

Tokeniser

Rozlišuje

- Slova
- Čísla
- Bílé znaky
- Interpunkce

He is 22 years old.

Type	Set	Start	End	Features
Token		0	2	{kind=word, length=2, orth=upperInitial, string=He}
SpaceToken		2	3	{kind=space, length=1, string= }
Token		3	5	{kind=word, length=2, orth=lowercase, string=is}
SpaceToken		5	6	{kind=space, length=1, string= }
Token		6	8	{kind=number, length=2, string=22}
SpaceToken		8	9	{kind=space, length=1, string= }
Token		9	14	{kind=word, length=5, orth=lowercase, string=years}
SpaceToken		14	15	{kind=space, length=1, string= }
Token		15	18	{kind=word, length=3, orth=lowercase, string=old}
Token		18	19	{kind=punctuation, length=1, string=.}

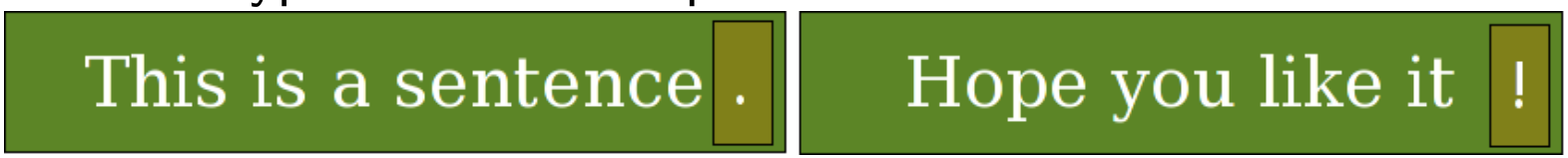
Gazetteer

- Seznamy slov
- Vytváří anotaci typu Lookup
- Tříděny do kategorií a podkategorií (features majorType, minorType)

```
John ... Lookup{majorType=person_first, minorType=male}
```

Sentence Splitter

Anotace typu Sentence a Split



Part of Speech Tagger

- Přiřazuje tokenům jejich POS kategorie.
- Brill Tagger

```
JJ      Adjective
NN      Noun (singular)
NNS     Noun (plural)
NNP     Proper Noun (singular)
NNPS    Proper Noun (plural)
...
```

Named Entity Transducer

Na základě gazetteeru + JAPE pravidel

- Vlastní jména (Person)
- Datum (Date)
- Názvy organizací (Organization)
- Povolání (JobTitle)
- ...

Praktická ukázka