

---

# Unsupervised Ontology-based Semantic Tagging for Knowledge Markup

---

***Paul Buitelaar, Srikanth Ramaka***

*DFKI GmbH – Language Technology Lab &*

*Competence Center Semantic Web*

*Saarbrücken, Germany*

---

# Overview

- n Introduction
  - q Knowledge Markup and Semantic Tagging
- n System Architecture
  - q Ontology
  - q Classifiers
- n Evaluation
  - q Strategy
  - q Results
- n Related Work and Conclusions

---

# Introduction

## n Knowledge Markup

- q Annotating Documents with Semantic Metadata
  - n Ontology Classes
- q Use in Semantic Web Applications
  - n Semantic / Concept-based Search and Similar

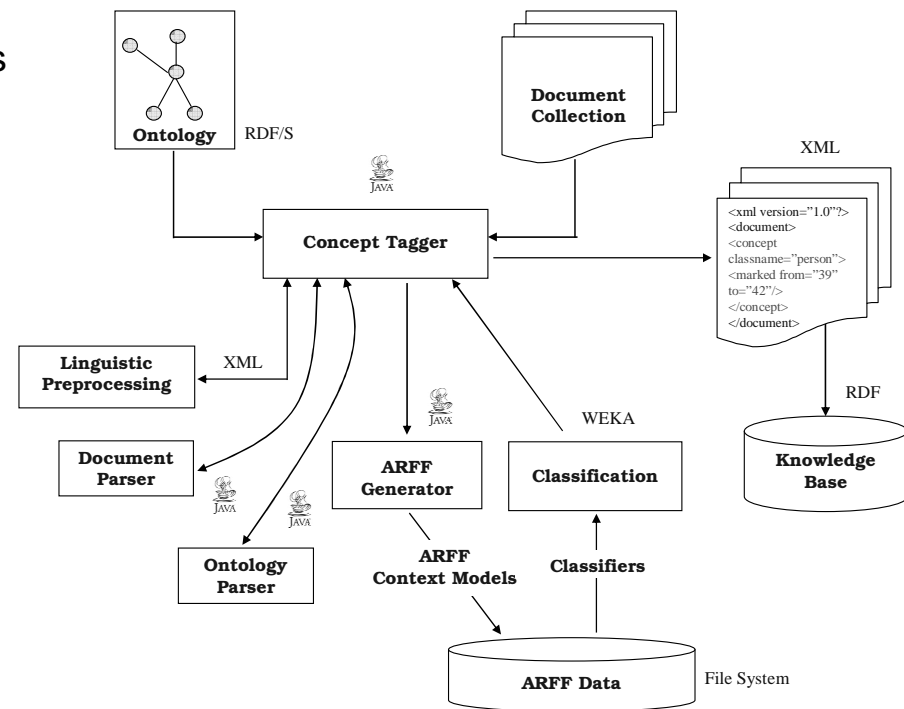
## n Semantic Tagging

- q Classifying Terms with Ontology Classes
- q Rule-based or with Trained Classifiers

# System Architecture

n System Architecture consists of:

- q a set of hierarchically organized classes from a domain ontology
- q a domain-relevant document collection for training and classification
- q a shallow linguistic module for preprocessing class labels and documents
- q a machine learning environment for generating context models and classifiers
- q a knowledge base to store marked up concept instantiations



---

# System Architecture – Ontology

## n Ontologies in RDFS, DAML, OWL

For experiments we took the SmartWeb corpus of soccer reports and a publicly available Ontology on Soccer

<http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml>

### action

#### player action

#### other player action

..., breakaway, center, charge, clear, cross, dribble, half volley shot, juggling, kickoff, lob, overlap, pass, run, shoot, tackle, throw in, trap, volley, ...

### entity

#### stoppage

corner, fault, free kick, goal, goal kick, offside, out of bounds, penalty

---

# System Architecture – Classifiers

- n Instance-based Learning

- q Generate Context Models

- n **Consider Subclasses as Terms for a Class**

- n Match Terms in Text as Instances for this Class

- q Unsupervised Training – No Manual Annotation

- n Collect N-gram Context for each Class-Instance

- q N-gram Context over Part-of-Speech, Morphological Stems

- n Generate a Context Model for each Class

- q Classification

- n Use the Context Model to Classify Terms in Text

- q Note: unseen classified terms may be new subclasses – ontology learning

# Classifiers – An Example

n Instances for the ontology class “Football Player” by match on subclass:

- q *Even during those early minutes Palace's former Carlisle **attacker** Matt Jansen looked up for a big game, and no wonder as he was facing his boyhood idols!*
- q *Arsenal's new French **midfielder** Patrick Vieira started the rot for Leeds this time after only 44 seconds.*
- q *That they went home empty-handed was largely down to another of Gullit's instant imported hits, former Strasbourg **sweeper** Frank Leboeuf.*

n Instance to be classified (unknown term ‘*striker*’):

- q *The big French **striker** stepped up to drill home the penalty himself.*

Context Model (N-gram with N=5) for “Football Player”			
-2	-1	+1	+2
<i>former</i>	<i>Carlisle</i>	<i>Matt</i>	<i>Jansen</i>
<i>new</i>	<i>French</i>	<i>Patrick</i>	<i>Vieira</i>
<i>former</i>	<i>Strassbourg</i>	<i>Frank</i>	<i>Leboeuf</i>

# Evaluation – Strategy

- n Evaluation set by pooling
  - q Runs with different parameters
  - q Set of 869 classified instances
  - q 3 judges evaluated classifications – for 863 with agreement
    - n ‘very good’, ‘good’, ‘incorrect’

	<i>very good</i>	<i>good</i>	<i>incorrect</i>
<code>other_player_action</code>	47	32	104
<code>person</code>	50	4	57
<code>place</code>	24	14	118
<code>stoppage</code>	4	2	407
<b>Total</b>	125	52	686

- n ‘very goods’ selected for “strict” evaluation set
- n ‘very goods’ + ‘goods’ selected for “relaxed” evaluation set

# Evaluation – Results

<i>N</i>	<i>k</i>	<i>Strict Set</i>				<i>Relaxed Set</i>			
		<i>Recall</i>		<i>Precision</i>		<i>Recall</i>		<i>Precision</i>	
<b>1</b>	<b>1</b>	<b>111</b>	<b>89%</b>	<b>99</b>	<b>89%</b>	<b>162</b>	<b>92%</b>	<b>144</b>	<b>89%</b>
	2	111	89%	97	87%	162	92%	141	87%
	10	109	87%	90	83%	158	89%	132	84%
2	1	86	69%	71	83%	117	66%	96	82%
	2	82	66%	70	85%	114	64%	94	82%
	10	83	66%	70	84%	115	65%	94	82%
5	1	21	17%	17	81%	31	18%	25	81%
	2	19	15%	16	84%	29	16%	22	76%
	10	18	14%	15	83%	27	15%	22	81%

**Results with N-gram Size 3 to 11 and  $k = 1, 2, 10$**

<i>N</i>	<i>k</i>	<i>Strict Set</i>		<i>Relaxed Set</i>	
		<i>Rec.</i>	<i>Prec.</i>	<i>Rec.</i>	<i>Prec.</i>
<b>1</b>	<b>1</b>	74%	85%	76%	84%
		(92)	(78)	(135)	(113)
	2	74%	83%	76%	81%
		(92)	(76)	(135)	(110)
	<b>10</b>	74%	79%	75%	79%
		(92)	(73)	(132)	(104)

**Results without Linguistic Preprocessing**

---

# Related Work and Conclusions

- n Word Sense Disambiguation
  - n Same methods - classification of words according to context
  - n Nature of classes is different: WordNet vs. Ontology
  
- n Named-Entity Recognition
  - n Same methods apply - ...
  - n Number of classes is different: 3-4 vs. 100-1000+
  
- n Large-Scale, Ontology-based Semantic Tagging
  - n Unsupervised Training – bootstrapping method based on a domain-specific Ontology hierarchy