HyLaP-AM – Semantic Search in Scientific Documents

Ulrich Schäfer, Hans Uszkoreit,

Christian Federmann, Yajing Zhang, Torsten Marek

DFKI Language Technology Lab

Talk Outline

- ☆ Extracting facts form scientific papers
  - Text extraction from PDF
  - Hybrid parsing with Heart of Gold on a Linux cluster
  - Quriple extraction: simplified semantic (predicate argument) structure per sentence
- ☆ Storage in Apache Solr server (plus bib metadata and NEs), with query expansion based on WordNet
- ☆ Quriples also for structured Question Answering (QA)
- ☆ Search and QA integration in HyLaP scientist's workbench application (demo)

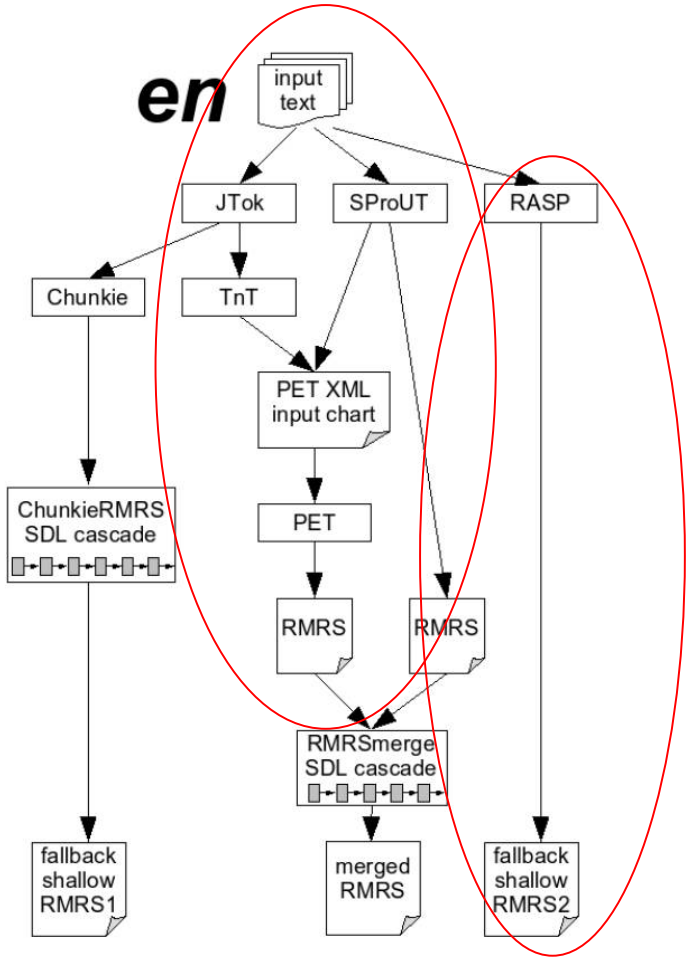
Paper Parsing Infrastructure

- ACL Anthology corpus subset of recent 6100 papers (2002-08)
- PDF extraction from scratch based on PDFbox (ACL Anthology text versions were not good enough)
- Additional postprocessing for better sentence extraction:
  - unboxing
  - de-hyphenation
  - text cleaning
  - removal of figures, tables, references, affiliation lines
  - addition of proper bibliographic metadata from bibtex
  - XML encoding/structuring

Hybrid Parsing

Parsing with Heart of Gold

- ☆ most of integration as in predecessor project QUETAL:
  - Tokenization (JTok)
  - PoS Tagging (TnT)
  - Named Entitiy Recognition (SProUT with LTWorld 2007 extensions)
- ☆ PET with ERG as of May 2008
- ☆ new RASP module (wrapper) for shallow fallback
- ☆ deep parsing coverage: 63% avg. sentence length of parsed sentences: 18.9 (abstracts)
- ☆ with RASP fallback almost 100%



HoG configuration for MRX output

Hybrid NLP Middleware

☆ Heart of Gold (http://heartofgold.dfki.de)

Component	NLP Type	Languages	Implemented in
JTok	tokenizer	de, en, it,...	Java
ChaSen	Japanese morph.	ja	C
TnT	statistical tagger	de, en,...	C
Treetagger	statistical tagger	en, de, es, it,...	C
Chunkie	stat. chunker	de, en,...	C
ChunkieRMRS	chunk RMRSes	de, en	XSLT, SDL/Java
LingPipe	statistical NER	en, es,...	Java
Sleepy	shallow parser	de	OCaml
SProUT	shallow NLP/NER	de, el, en, ja,...	Java
LoPar/wbtopo	PCFG parser	de	C, XSLT
Corcy	coref resolver	en	Python
RASP	shallow NLP	en	C, Lisp
PET	HPSG parser	de, el, en, ja,...	C, C++, Lisp
RMRSmerge	RMRS merger	de, en,...	XSLT, SDL/Java
SDL	sub-architectures		SDL/Java

Quriple Extraction

- ☆ Quriple = query-oriented triple/quintuple of subject, predicate, direct object, indirect object, pp/modifiers
- ☆ computed either from PET MRX or RASP RMRS
- ☆ algorithm: intermediate transformation into isomorphic Java objects (serialized; persistent) for efficient graph manipulation and quriple extraction (Java)
- ☆ special handling of negation, passive, coordination
- ☆ voting mechanism for 3 best readings (according to Redwoods) at quriple level (may collapse to 1)
- ☆ WordNet synsets of predicates are computed offline

Quriple Example 1

"We evaluate the efficiency and performance against the corpus."

-> quriple:

SUBJ	We
PRED	evaluate
DOBJ	the efficiency and performance
OCMP	-
ADJU	against the corpus

Quriple Example 2: Coordination

"The system automatically extracts pairs of syntactic units from a text and assigns a semantic relation to each pair."

-> introduce 2 quriples:

SUBJ	The system
PRED	extract
DOBJ	pairs of syntactic units
OCMP	from a text
ADJU	against the corpus

SUBJ	The system
PRED	assign
DOBJ	a semantic relation
OCMP	to each pair
ADJU	automatically

Quriple Example 3: Passive

"Unseen input was classified by trained neural networks with varying error rates depending on corpus type."

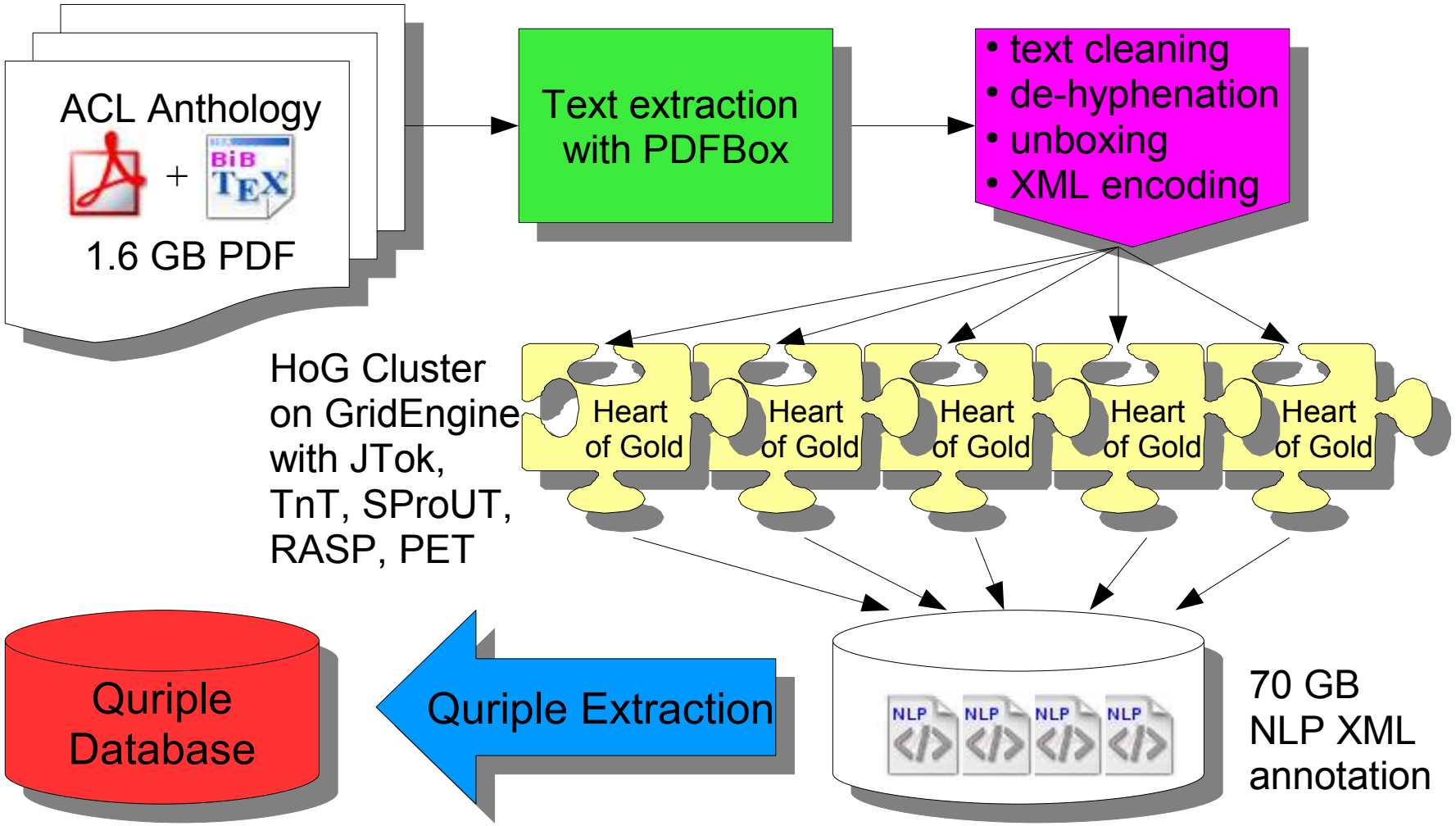
SUBJ	trained neural networks with varying error rates depending on corpus type
PRED	classify
DOBJ	unseen input
OCMP	-
ADJU	-

Quriple Search: Example

- Quriple search (also used by structured QA with Quantico):
- Quriple search expression "method improve baseline" is translated to an Apache Solr query subj:method +pred: (improve OR ameliorate OR better OR meliorate) + (dobj:baseline OR iobj:baseline OR rest:baseline)
- Answer example (1 out of 72) on the right

```
<doc>  <!-- each doc is a single quriple sentence here -->
  <float name="score">1.2502118</float>
  <date
     name="timestamp">2009-01-27T10:46:38.452Z
  </date>
  <str name="aclaid">W05-0814</str>
  <int name="offset">198</int>
  <int name="sentno">87</int>
  <int name="page">4</int>
  <str name="prefix">W05-0814-s87-p4</str>
  <str name="qgen">PET</str>
  <str name="sentence">Our model and training
     method improve upon a strong baseline for
     producing 1-to-many alignments.
  </str>
  <str name="subj">Our model training method</str>
  <int name="subj_start">0</int>
  <int name="subj_end">28</int>
  <str name="pred">improve</str>
  <int name="pred_start">30</int>
  <int name="pred_end">36</int>
  <str name="rest">upon a strong baseline for
     producing 1-to-many alignments
  </str>
  <int name="rest_start">38</int>
  <int name="rest_end">94</int>
</doc>
```

Grid-based Quriple Extraction



700 MB Solr Blob (953 000 sentences/quriples)

AIAMA Demo: In search of lost papers

Associative Information Access and Management Application

- Scientist's Workbench: document editing support for knowledge workers
- Integrated quriple search and structured QA on full content of 6100 scientific papers (ACL Anthology)
- Integration with QA servers and quriple data from papers as source for structured QA, additional interface to open domain QA (HyLaP-QA co-project)

AIAMA Architecture

Python Application Server

Online named entity analysis enhanced w/ ontology information

Ontology concept browser

LTWorld Ontology (Jena)

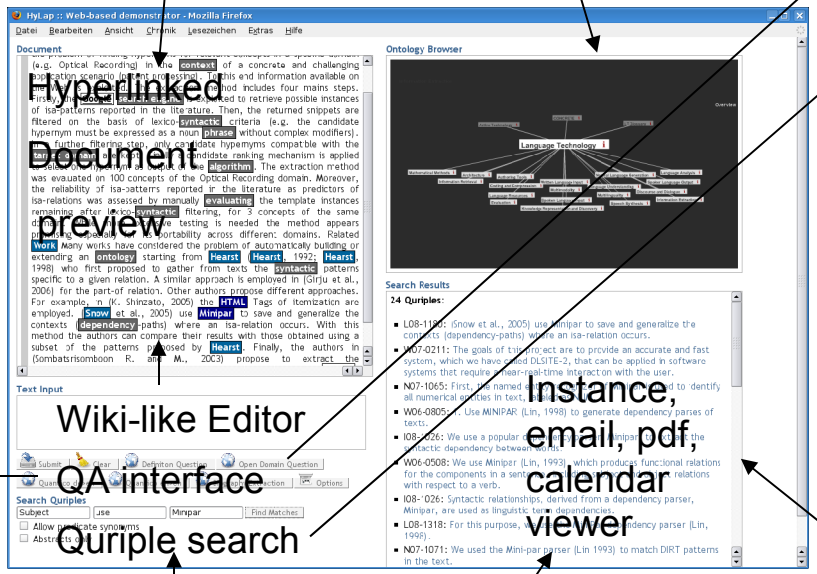
Quantico Structured QA, crosslingual

Quriple Store (Solr)

Quriple extraction

AIAMA GUI – A scientist's workbench

AJAX



Hyperlinked Document preview

Wiki-like Editor

QA interface

Quriple search

Instance, email, pdf, calendar viewer

Open-domain QA Definition and Biography Retrieval

Heart of Gold Cluster (offline) tokenization, SBR, PoS tagging, named entity recognition, shallow + deep parsing, quriple extraction

Web

WordNet synonyms

Facts, docs, links RDBMS

ACL Anthology AAN Michigan

From HyLaP to TAKE

- ☆ Last phase (year) of HyLaP: shift from personal memory to eScience application
- ☆ Continue/elaborate work in follow-up project TAKE
- ☆ Improve results various stages of the extraction process
  - PDF to text extraction (commercial tool or OCR integration)
  - shallow-deep integration (new chart mapping interface)
  - improve domain-specific handling: better integration of LT World ontology (via SProUT) and automatic extension
  - quriple extraction: cover more cases

More Info: http://hylap.dfki.de

¿Questions?