An Incrementally Trainable Statistical Approach to Information Extraction Based on Token Classification and Rich Context Models

Christian Siefkes

Freie Universität Berlin
Berlin-Brandenburg Graduate School in Distributed Information Systems
(DFG grant no. GRK 316)

16th February 2007

Overview

Introduction

Information Extraction

- Information Retrieval makes huge amounts of *textual* data accessible, but supports only simple keyword-based queries
- Structured Query Languages allow a far richer set of queries, but only on data stored in *databases*
- Information Extraction (IE) is the task of making relevant parts of *texts* available for structured querying,
  by finding desired pieces of information in texts and storing them in a *database,*
  in an automatic or semi-automatic fashion

Information Extraction (2)

- Information Extraction (IE) is the task of making relevant parts of *texts* available for structured querying,
  by finding desired pieces of information in texts and storing them in a *database,*
  in an automatic or semi-automatic fashion
- Requires metadata:
  Target schema defining what kinds of information to extract
- and sample data:
  annotated training data and/or human supervision
  (correction of proposed extractions)

Example

Subject: **CEDA Spring Lecture Series**
Date: 9 Feb 2004 10:18
From: Edmund J. Delaney <ed@andrew.cmu.edu>

The Center for Electronic Design Automation, CEDA, in the department of Electrical and Computer Engineering will offer its first lecture in its Spring lecture series on February 13, in the *Adamson Wing, Baker Hall*.

The lecture begins at *3:30 p.m* followed by a reception in Hamerschlag Hall, Room 1112. *Professors Rob A. Rutenbar* and *Wojciech Maly* will speak on "The State of the Center for Electronic Design Automation".

Extracted information:

- speaker:
  Professors Rob A. Rutenbar
  Wojciech Maly
- location:
  Adamson Wing, Baker Hall
- start time:
  3:30 p.m
- end time:
  –

Approach

Approach

Information extraction can be modeled as a token classification task:

- Build context representation of each word or token
- Classify context representation of each token using a trainable classifier (e.g. *Winnow+OSB*)
- Use tagging strategy to combine classification results

Contribution: Designed and implemented a generic framework for classification-based information extraction that allows modifying and exchanging all core components independently of each other.

Which Items to Classify?

- The information we want to extract is contained in text fragments
- Each text fragment spans one or several tokens (words)
- So we classify each token in a text whether it is part of relevant information

...	**lecture**	**begins**	**at**	**3:30**	**p.m**	**followed**	**by**	...
	speaker	*speaker*	*speaker*	*speaker*	*speaker*	*speaker*	*speaker*	
	location	*location*	*location*	*location*	*location*	*location*	*location*	
	s(tart) time	*stime*	*stime*	*stime*	*stime*	*stime*	*stime*	
	e(nd) time	*etime*	*etime*	*etime*	*etime*	*etime*	*etime*	
	O(ther)	*O*	*O*	*O*	*O*	*O*	*O*	

Which Classes to Use?

- Just classify among each possible attribute (SPEAKER, LOCATION, ...) and other (O)? ($n + 1$ classes for n attributes)
- But consider:

HCI & G SEMINAR
Wednesday, October 21, 1992
3:30 - 5:00pm
Wean Hall 5409
Mode preference in a data-retrieval task
by
Emilie Roth
Alexander I. Rudnicky
(Carnegie Mellon University, School of Computer Science)

→ There are two separate SPEAKERS here, not just one! How to avoid them collapsing into one?

Which Classes to Use? (2)

- We can use a distinct prefix for the first token (begin) of each attribute value ($2n + 1$ classes for n attributes)

...	by	Emilie	Roth	Alexander	I.	Rudnicky	(	...
	O	*B-speaker*	*I-speaker*	*B-speaker*	*I-speaker*	*I-speaker*	*O*	

- There are various other ways of tagging word sequences in an unambiguous way (tagging strategies)

Text	Our	meeting	with	Mr.	Irfan	Ali
Triv	O	O	O	speaker	speaker	speaker
IOB2	O	O	O	B-speaker	I-speaker	I-speaker
IOB1	O	O	O	I-speaker	I-speaker	I-speaker
BIE	O	O	O	B-speaker	I-speaker	E-speaker
BIA	O	O	O	B-speaker	I-speaker	I-speaker
BE	O/O	O/O	O/O	B-speaker/O	O/O	O/E-speaker

Text	will	be	at	1:30	pm	in ...
Triv	O	O	O	stime	stime	O
IOB2	O	O	O	B-stime	I-stime	O
IOB1	O	O	O	I-stime	I-stime	O
BIE	O	O	O	B-stime	E-stime	O
BIA	A-speaker	O	O	B-stime	I-stime	A-stime
BE	O/O	O/O	O	B-stime/O	O/E-stime	O/O

- **Contribution:** Analyzed and evaluated the various tagging strategies used for information extraction
  (also introduced a new one that turned out to be competitive with the best other strategies) [Sie06].

Which Features to Use for Classification?

3:30 ???

The Center for Electronic Design Automation, CEDA, in the department of Electrical and Computer Engineering will offer its first lecture in its Spring lecture series on February 13, in the *Adamson Wing, Baker Hall*.

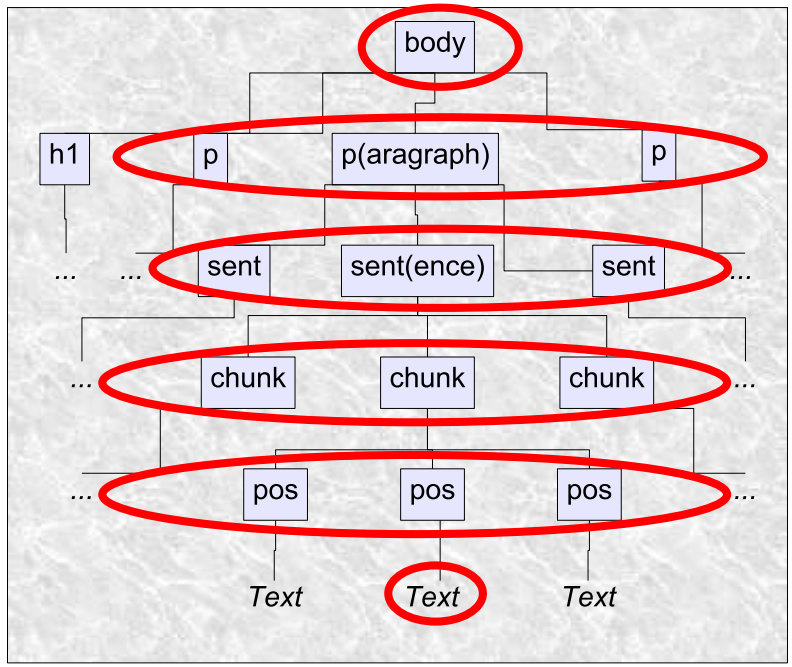
The lecture begins at *3:30 p.m* followed by a reception in Hamerschlag Hall, Room 1112. *Professors Rob A. Rutenbar* and *Wojciech Maly* will speak on "The State of the Center for Electronic Design Automation".

- Obviously, a token *alone* wouldn't provide much usable information to the classifier
- We need a *feature vector* representing various aspects of the tokens and its context (context representation)
- Which features to include in the context representation? Note that even surrounding *sentences* can provide valuable clues!

Which Features to Use for Classification? (2)

How to generate rich context representations that even take long-distance information (head words of surrounding sentences, description list labels etc.) into account?

- Capture document structure (including linguistic structure) in a tree representation
- Include key features from surrounding nodes in the XML tree (Inverted subtree)



- Also add features regarding the token itself (incl. linguistic, morphological, and semantic information)

Which Features to Use for Classification? (3)

- **Contribution:** Introduced rich tree-based context representations that combine document structure with linguistic and semantic sources of information.
- Problem: What if there are conflicts between document structure and linguistic structure?
  → **Contribution:** Developed an XML merging and repair algorithm that can resolve nesting errors (overlaps) and related problems
- Do the various sources of information actually help? How much?
  → **Contribution:** Performed a detailed ablation study measuring the influence of the various factors on the overall results.

How to Provide Training Data?

Portability problem: how to adapt an IE system to a new application domain?

- Old static rule-based systems: manually rewrite rules—huge amount of work for hard-to-get specialists
- Current trainable systems: provide manually annotated training data—easier task, but still time-consuming

Solution to reduce training burden:

- Incremental learning: documents are annotated sequentially by a user and immediately incorporated into the extraction model—system can support user by proposing annotations

→ **Contribution:** introduced incremental learning (“train as you go”, as in spam filtering) into the field of IE and evaluated its usefulness

Evaluation

Test Corpus I: Seminar Announcements

- 485 seminar announcements from university newsgroups
- *Semi-structured texts:* informal, quickly written e-mail-style messages
- Extract (if present):
  - SPEAKER (409)
  - LOCATION (464)
  - START TIME (485)
  - END TIME (228)

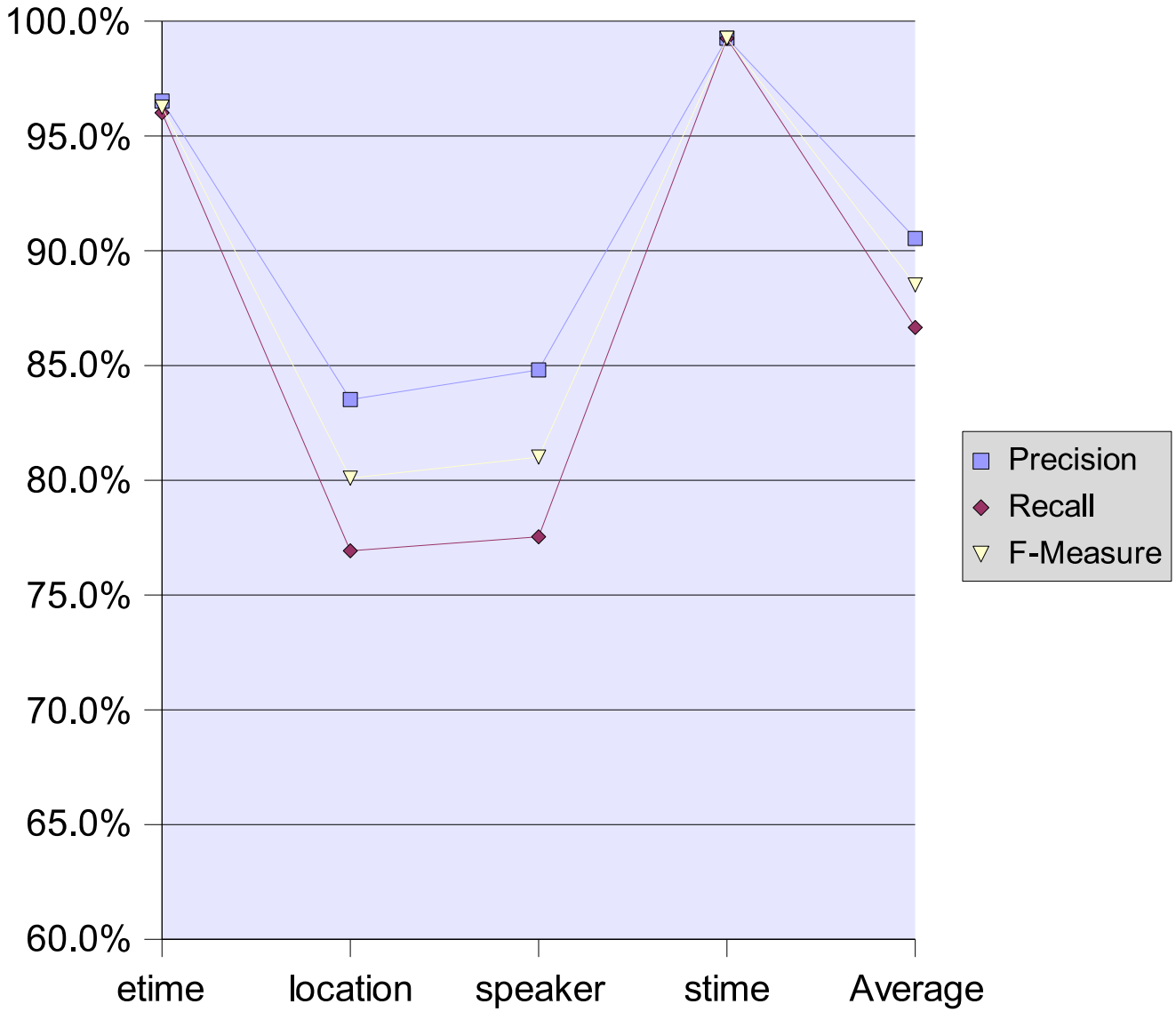
Test Corpus 2: Corporate Acquisitions

- 600 newspaper articles about mergers and acquisitions (Reuters corpus)
- *Free texts*: formally written, strictly grammatical, almost no structured information
- Nine attributes to extracted:
  - Official names of the parties to an acquisition: ACQUIRED (593), PURCHASER (545), SELLER (235)
  - Corresponding abbreviated names: ACQABR (437), PURCHABR (445), SELLERABR (182)
  - Location of the acquired company: ACQLOC (178)
  - Price paid: DLRAMT (259)
  - Information about the status of negotiations: STATUS (453)

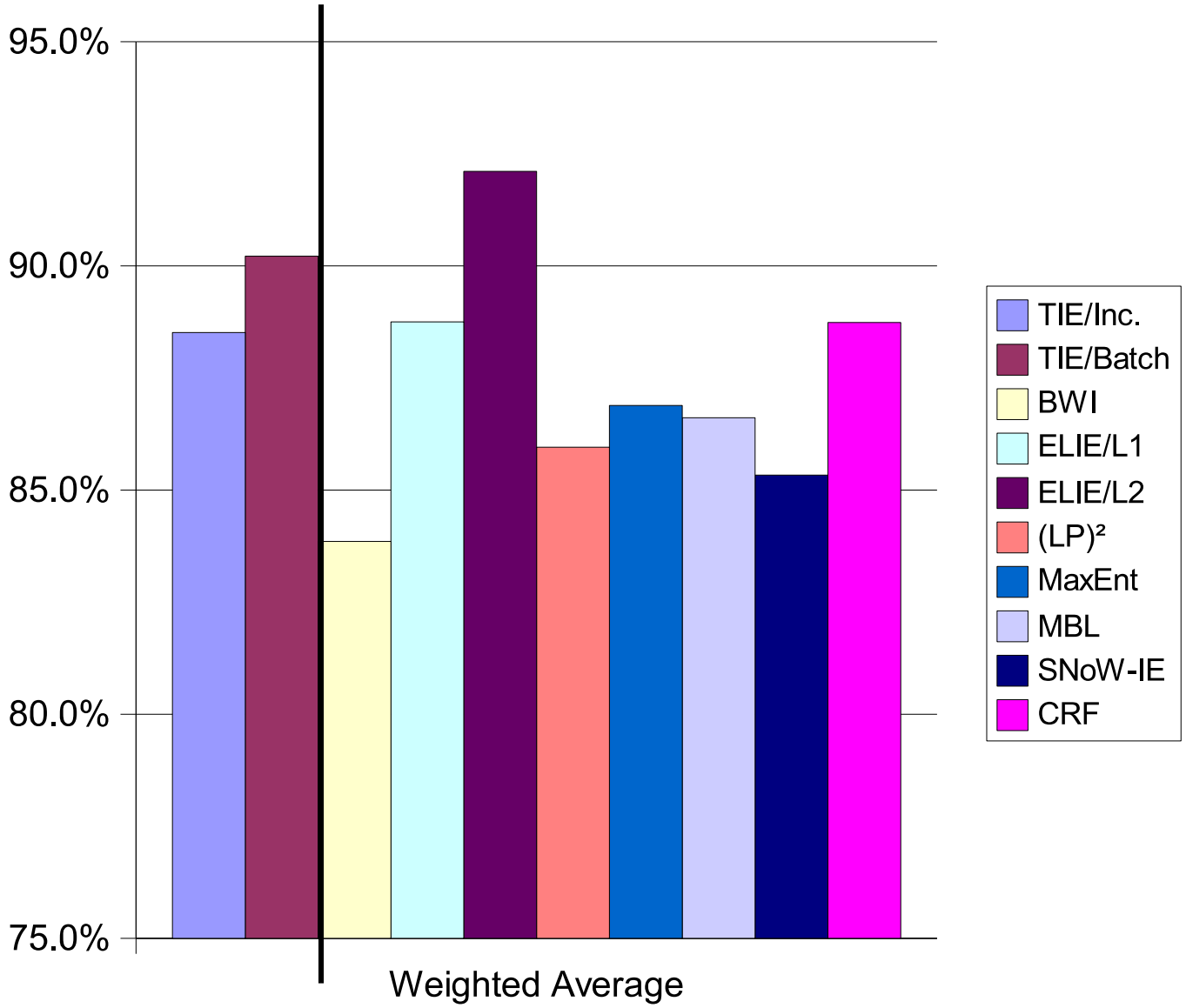
Experimental Setup

- Using the standard setup for both corpora:
  - 50% of texts used for training, 50% for evaluation
  - Results averaged over 5 resp. 10 random splits
  - Evaluation mode: *One answer per attribute* (“match-best”)
  - Matches must be exact (partial matches are errors)
- Evaluation metrics:
  - *Precision* $P = \frac{true\ positives}{true\ positives + false\ positives}$
  - *Recall* $R = \frac{true\ positives}{true\ positives + false\ negatives}$
  - *F-measure* $F = \frac{2 \times P \times R}{P + R}$
- Evaluated with *incremental training* and with *batch training* (default)

Results on Seminar Corpus

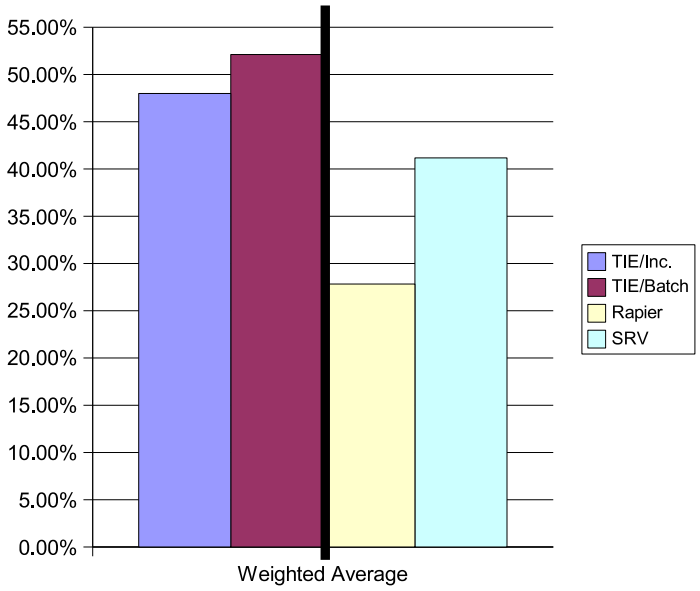


System Comparison (F-measure)

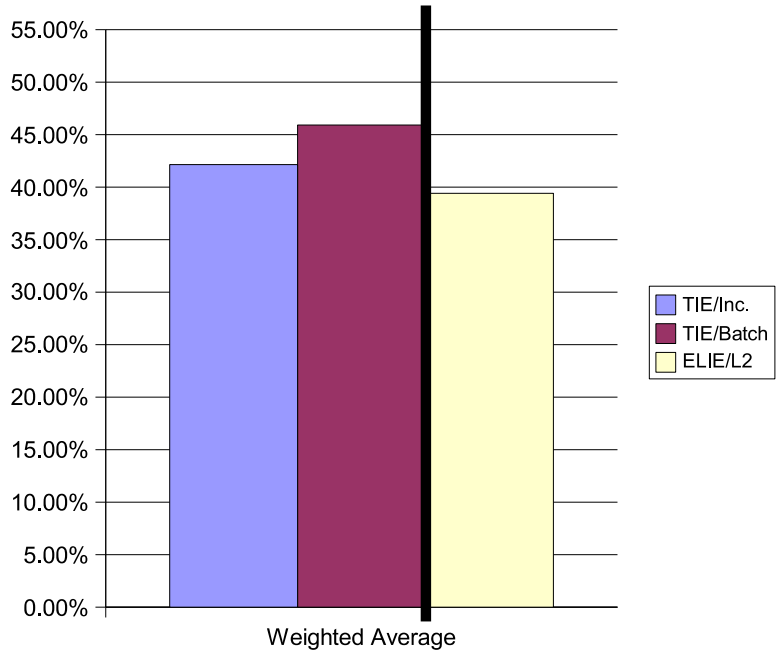


Results on Acquisitions Corpus

System Comparison (F-measure)



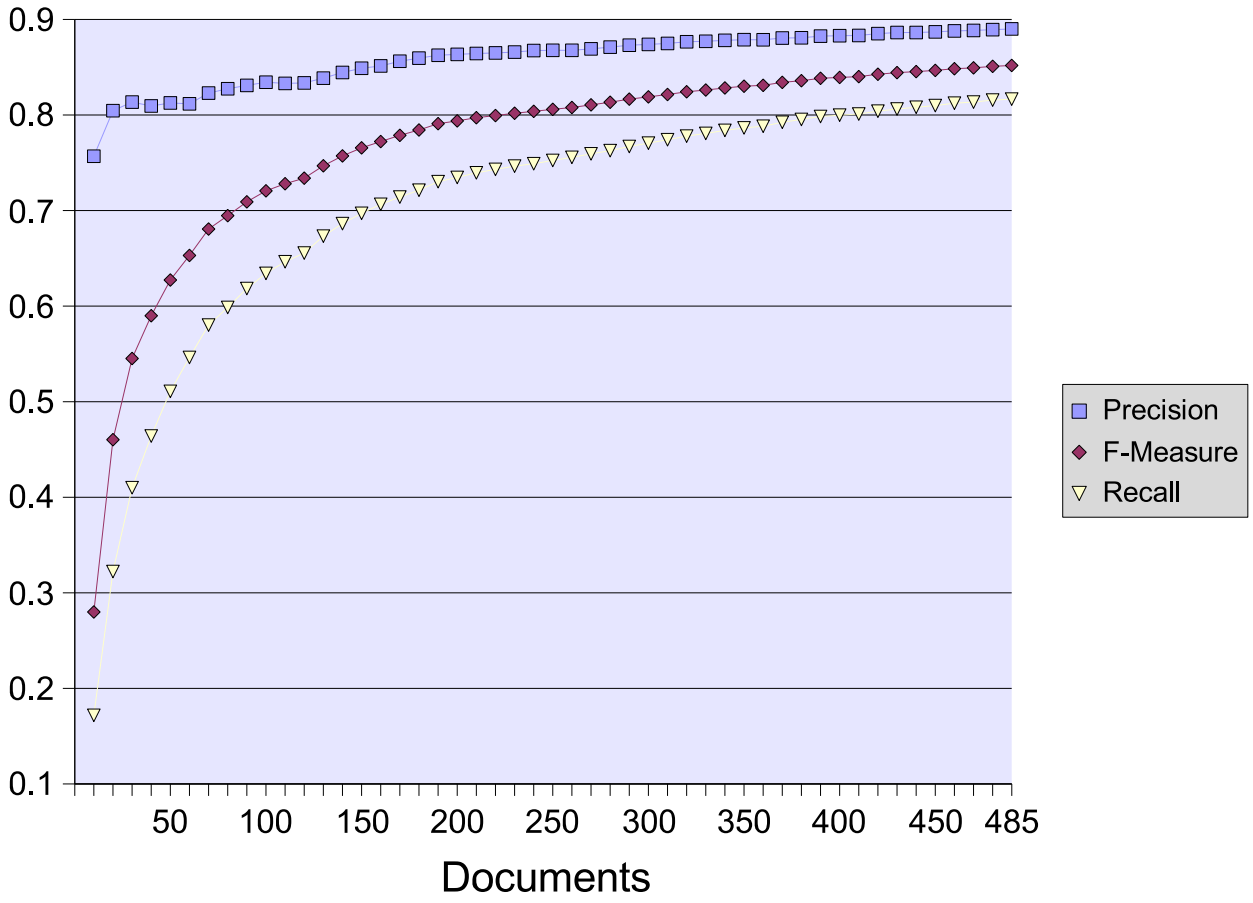
(a) Match-best Evaluation



(b) Match-all Evaluation

Incremental Learning Curve

Learning curve for (simulated) interactive incremental training:

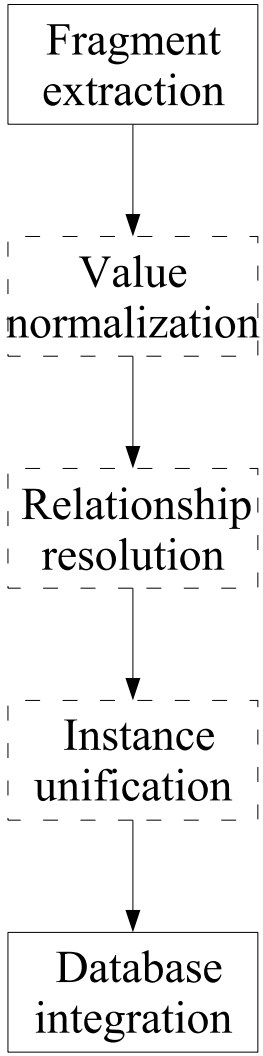


Future Work

Future Work

Most important challenges:

1. Learn more about which kinds of texts, attributes, and tasks are suitable for information extraction
2. *Fragment extraction* is only one step of a comprehensive approach for text-to-database integration
   $\longrightarrow$ work on the other ones!



Selected Publications

[Ass05] Fidelis Assis, William Yerazunis, Christian Siefkes, and Shalendra Chhabra.
CRM114 versus Mr. X: CRM114 notes for the TREC 2005 spam track.
In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*. 2005.
URL http://crm114.sourceforge.net/NIST_TREC_2005_paper.pdf.

[Ass06] Fidelis Assis, William Yerazunis, Christian Siefkes, and Shalendra Chhabra.
Exponential differential document count: A feature selection factor for improving bayesian filters accuracy.
In *2006 Spam Conference*. MIT, Cambridge, MA, 2006.
URL http://osbf-lua.luaforge.net/papers/osbf-eddc.pdf.

[Chh04] Shalendra Chhabra, William S. Yerazunis, and Christian Siefkes.
Spam filtering using a Markov random field model with variable weighting schemas.
In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM '04)*. 2004.
URL http://www.siefkes.net/papers/mrf-spamfiltering.pdf.

[Sie03] Christian Siefkes.
Learning to extract information for the Semantic Web.
In Robert Tolksdorf and Rainer Eckstein, eds., *Tagungsband Berliner XML Tage 2003*, pp. 452–459. 2003.
URL http://www.siefkes.net/papers/ie-semantic-web.pdf.

[Sie04a] Christian Siefkes.
A shallow algorithm for correcting nesting errors and other well-formedness violations in XML-like input.
In *Extreme Markup Languages (EML) 2004*. 2004.
URL http://www.siefkes.net/papers/eml/EML2004.pdf.

[Sie04b] Christian Siefkes, Fidelis Assis, Shalendra Chhabra, and William S. Yerazunis.
Combining Winnow and orthogonal sparse bigrams for incremental spam filtering.
In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, eds., *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004)*, vol. 3202 of *Lecture Notes in Artificial Intelligence*, pp. 410–421. Springer, 2004.
URL http://www.siefkes.net/papers/winnow-spam.pdf.

Selected Publications (2)

[Sie05a] Christian Siefkes.
Incremental information extraction using tree-based context representations.
In Alexander Gelbukh, ed., *Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2005)*, vol. 3406 of *Lecture Notes in Computer Science*, pp. 510–521. Springer, 2005.
URL http://www.siefkes.net/papers/incremental-ie.pdf.

[Sie05b] Christian Siefkes and Peter Siniakov.
An overview and classification of adaptive approaches to information extraction.
Journal on Data Semantics, IV:172–212, 2005.
URL http://www.siefkes.net/papers/overview-ie.pdf.
LNCS 3730.

[Sie06] Christian Siefkes.
A comparison of tagging strategies for statistical information extraction.
In *HLT-NAACL 2006*. 2006.
URL http://www.siefkes.net/papers/tagging-strategies-ie.pdf.

[Yer05] William S. Yerazunis, Shalendra Chhabra, Christian Siefkes, Fidelis Assis, and Dimitrios Gunopulos.
A unified model of spam filtration.
In *2005 Spam Conference*. MIT, Cambridge, MA, 2005.
URL http://crm114.sourceforge.net/UnifiedFilters.pdf.