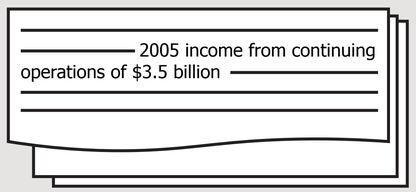
Seer: Automatically Learning Rules for Information Extraction

Maeda Hanafi (maeda.hanafi@nyu.edu), Azza Abouzied (azza@nyu.edu), Laura Chiticariu (chiti@us.ibm.com), Yunyao Li (yunyaoli@us.ibm.com)

INFORMATION EXTRACTION

- How does extracting information from a collection of documents work?
  - Extract revenue per year from financial reports.



- Rule developers create extraction rules that can extract what they need from the documents.

Revenue Rule 1:

Document	Revenue Rule
FinancialReport2005.txt	2005 income from continuing operations of $3.5 billion
FinancialReport2007.txt	2007: Total revenues of $26 billion
FinancialReport2008.txt	2008 net income of $7.9 billion
FinancialReport2009.txt	2009 reduced the assets by $92 billion

PROBLEM

- Creating rules is overwhelming to novice developers.
- It is time consuming for rule developers to create a rule that captures all instances of the kinds of text they want.
- Rule developers know what kinds of text a rule should capture but don't know the right rule refinement.

THE SOLUTION: Seer

- Users provide examples by highlighting areas of the text they wish to extract.
- Seer suggests and refine extraction rules based on example text from user.
- Seer learns rules and refinements to suggest to the user.

1 HIGHLIGHT

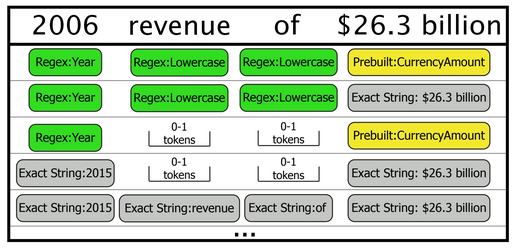
The user highlights examples of what to extract.

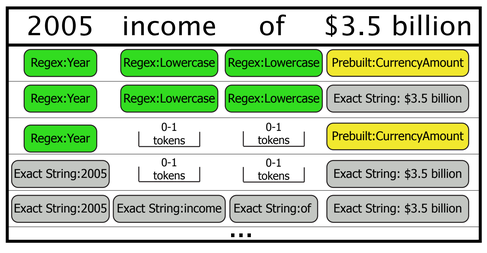
continuing operations for the fourth quarter grew 2 percent compared with the fourth-quarter 2005 income of $3.4 billion, excluding the one-time charge. Total revenues for the fourth quarter of 2006 revenue of $26.3 billion increased 7 percent (4 percent, adjusting for currency) from the fourth quarter of 2005.

2 LEARN EXTRACTION RULES

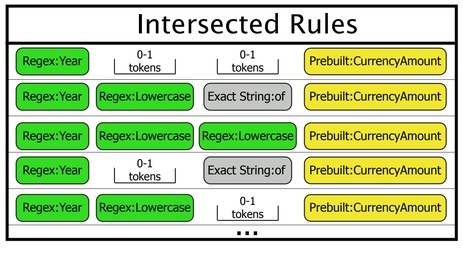
The goal of learning is to show the user a list of distinct rules that covers all examples.

- a) For each example, generate all possible rules.





- b) Intersect all learned rules.

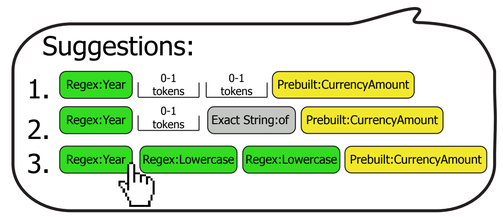


- c) Identify distinct rules based on the component types e.g. Prebuillt, Regex, etc.

Rules to Suggest	Rule Type
Regex:Year 0-1 tokens 0-1 tokens Prebuilt:CurrencyAmount	Regex, Token, Prebuilt
Regex:Year 0-1 tokens Exact String:of Prebuilt:CurrencyAmount	Regex, Exact String, Token, Prebuilt
Regex:Year Regex:Lowercase Regex:Lowercase Prebuilt:CurrencyAmount	Regex, Prebuilt

3 SUGGEST RULES TO USER

User selects and accepts a rule by clicking on it.



4 RULE EXTRACTS DATA

The rule is run on all documents and the extraction results are highlighted.

continuing operations for the fourth quarter grew 2 percent compared with the fourth-quarter 2005 income of $3.4 billion, excluding the one-time charge. Total revenues for the fourth quarter of 2006 revenue of $26.3 billion increased 7 percent (4 percent, adjusting for currency) from the fourth quarter of 2005.
The 2005 revenue of $5.6 billion was from the Software segment, higher than 2004; systems revenues $642 million, a 2% decrease.

5 REFINE A RULE

A rule can be refined by highlighting more examples or rejecting existing highlights.

- Example: Reject a data extraction result.

The 2005 revenue of $5.6 billion was from the Software segment, higher than ~~2004; systems revenues $642 million~~, a 2% decrease.

- Seer refines the rule to capture or not capture the new examples through the learning process in step 2.