# Lightly-Supervised Attribute Extraction

Yuri Osokin

Problem:

Get entity-attribue pairs from the text with lightly human interaction

Solution:

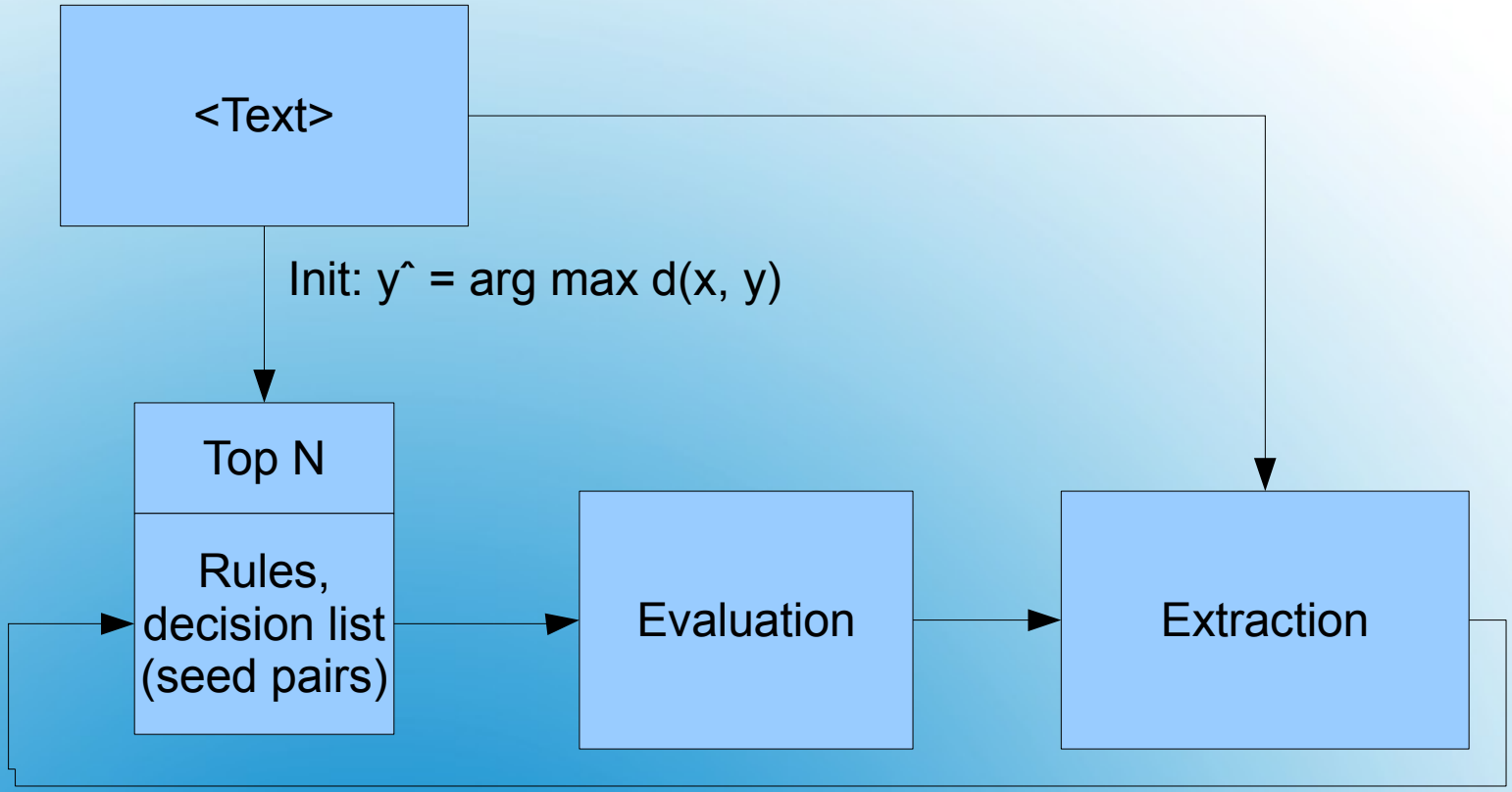
Set of alorithms based on rules extraction and entropy

What for:

Topic specific query search engine optimisation

- Decision List Co-training (next slides)
- Maximum entropy
  - Based on user-provided pairs iterative rule is applied to evaluate the probability of the pair to be good based on features.
- Generic pattern extractors:
  - Learning context specific features from the examples, <attr> of <entity> (height of John), <entity> CEO and <attr> (Bill Gates CEO and Chairman)

# Methods




# Decision List Co-training Classifier

**Evaluation**:

2 types of features: X1 – context, X2 – context
N iterations of self-consistent equations (each of C in [0,1].
In the end maximum of C is choosed as confidence value.

$$C(x_1) = \frac{\sum_{x_2 \in \mathcal{X}_2} \left( \frac{MI(x_1, x_2)}{MI_{max}} \times C(x_2) \right)}{|\mathcal{X}_2|}$$

$$C(x_2) = \frac{\sum_{x_1 \in \mathcal{X}_1} \left( \frac{MI(x_1, x_2)}{MI_{max}} \times C(x_1) \right)}{|\mathcal{X}_1|}$$

**Extraction**:

Top n pairs that match the rules are transformed into rules, and become part of the decision list.

# Decision List Co-training Classifier

- Complex rules extraction ( height precision, not general):

(*surrctxt=chairman and && <ATTR> of <COMPANY>*)

This is a feature. In experiment we limit the classifier to 3000 features.

# Decision List rules (extraction)

- The result is a sorted list of rules (pairs) which could be used to refine sort results
- Post processing stage which is not a part of algorithm is applied. It called re-ranking and will be described futher in slides.

# Decision List Results

Final optimisation.

The idea behind this re-ranking is that we should have confidence in an attribute value which is strongly associated with many reliable entities.

$$R(e,a) = c(e,a) \times \mathrm{C}(\mathrm{ent} = e) \times \mathrm{C}(\mathrm{attr} = a)$$

So the key idea is to make the entity and attribute universal inside the topic.

## Re-ranking

Two texts were used: countries and companies.

| *Relation* | *Key Seeds* | *Value Seeds* |
|---|---|---|
| *(Company, Attribute)* | Top 100 Fortune-500 companies | *type, headquarters, chairman, ceo, products, revenue, operating income, net income, employees, subsidiaries, website, headquarter* |
| *(Country, Attribute)* | 191 UN member countries. | *capital, largest city, official language, president, area, population, gdp, currency* |

Table 2: *(Company, Attribute)* and *(Country, Attribute)* seeds used in the experiments.

(a) (*Company, Attr*)

| *System* | *Precision* | | | |
|---|---|---|---|---|
| | @10 | @20 | @50 | @100 |
| $PT_+ + R$ | **100%** | **85%** | 70% | **70%** |
| $ME + R$ | 60% | 65% | **72%** | 47% |
| $SE + R$ | 90% | 75% | 56% | 40% |

(b) (*Country, Attr*)

| *System* | *Precision* | | | |
|---|---|---|---|---|
| | @10 | @20 | @50 | @100 |
| $PT_+ + R$ | 40% | 65% | 64% | 58% |
| $ME + R$ | **80%** | 75% | 80% | 77% |
| $SE + R$ | **80%** | **90%** | **88%** | **82%** |

Table 3: Non-seed attribute precision at various ranks.

# Experimental evaluation