Flowers are soft. But, How Would Computers Know?

Niket Tandon

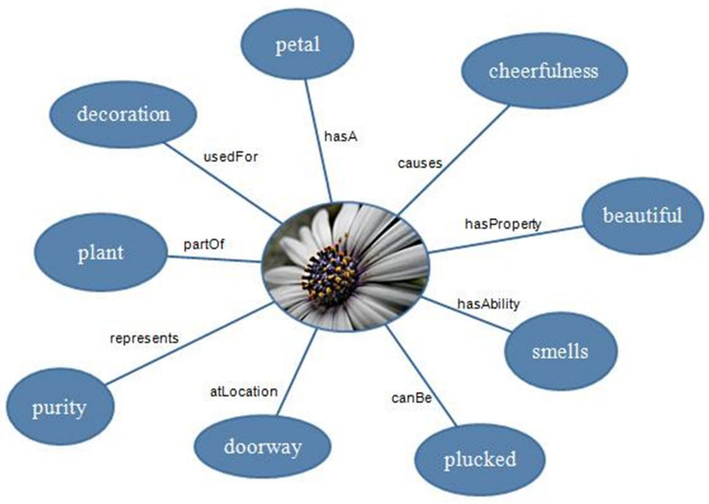
OVERVIEW

Making Computers Commonsense enabled.

Motivation & Objective

Motivation: Machines lack commonsense knowledge(CSK)!

Objective: Harvest CSK from text and *make it highly structured*.

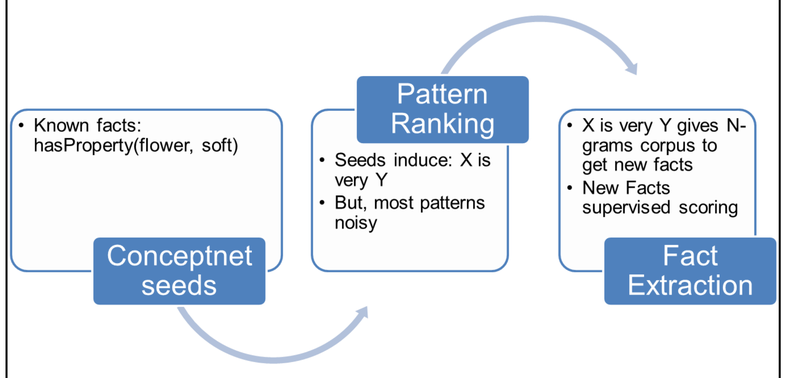


Challenges

- ❑ CSK is rarely mentioned in text
- ❑ Difficult to procure large corpus
- ❑ Natural language text is noisy

Overview: Our Approach

Pattern-based IE over N-grams



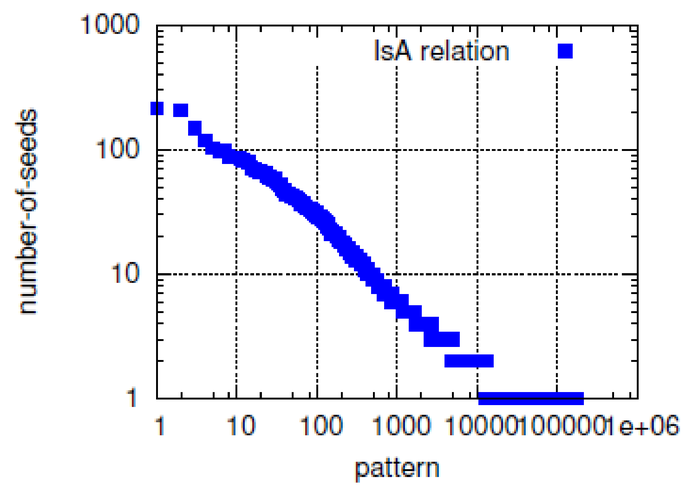
Contributions

1. We harvest over easily available N-grams corpus instead of hard to procure text
2. Existing approaches give low accuracy on our corpus, we design novel pattern scoring

CONTRIBUTION-1

Novel Pattern Scoring adapted for N-gram corpus

Observations



1. Power-law distribution for #seeds a pattern matches. Bad patterns in the tail.
2. Some patterns match too many relation's seeds. Penalize such patterns.

Our Pattern Scoring

Score based on Observation 1,2:

$$\theta(\mathbf{R_i}, \mathbf{p}) = \frac{\mathbf{e}^{\phi(\mathbf{R_i},\mathbf{p})}}{\mathbf{1} + \mathbf{e}^{\phi(\mathbf{R_i},\mathbf{p})}} \cdot \frac{|\frac{\mathbf{d}}{\mathbf{dx}}\mathbf{s}(\mathbf{x})|}{\mathbf{1} + |\frac{\mathbf{d}}{\mathbf{dx}}\mathbf{s}(\mathbf{x})|}$$

$\phi(R_i,p)$: pattern specificity.
s(x): slope of the power law graph

Suitability to N-grams

Beyond raw frequency, we consider distinct seeds, and down-weigh generic patterns.

CONTRIBUTION-2

Harvest commonsense facts from N-gram corpus

Our Fact Scoring

Naïve score: # patterns matched.

Facts supported by many patterns are accurate	Facts supported by few patterns may or may not be accurate

Supervised learning approach: Instead, we learn a decision tree over patterns as features.

Extraction Results

Over 200 million facts extracted.

CSK Relation	Precision (%)	#Facts Extracted
CapableOf	77	**907,173**
Causes	88	**3,218,388**
HasProperty	62	**2,976,028**
…	…	…

Future Work

1. Highly structure the extracted CSK (e.g. from hasProperty to hasColor, hasSize…)
2. Make the Knowledge base accurate.

References

1. Deriving a Web-Scale Common sense Fact Database: Niket Tandon, et . al, Proc. *AAAI 2011*
2. Information Extraction from Web-Scale N-Gram Data: Niket Tandon et . al. Web N-gram *Workshop at SIGIR 2010*

max planck institut informatik
UNIVERSITÄT DES SAARLANDES
International Max Planck Research School for Computer Science