Combining Information Extraction and Human Computing for Crowdsourced Knowledge Acquisition

Sarah Kumar Kondreddi*, Peter Triantafillou & Gerhard Weikum

**Automated Information Extraction (IE)**

IE helps to build large knowledge bases from Web and other sources (e.g., DBpedia, YAGO, NELL, KnowItAll, etc.). However, IE from natural language text (e.g., news, books, blogs, etc.) is challenging:
- complex structure of sentences
- ambiguous wording
- diversity of textual expressions

**Pros and cons:**
- Operates at scale, produces millions of facts
- Low output quality for complex inputs

**Motivation**

Humans outperform machines in understanding the use of language, resolving ambiguity and performing reasoning. Human Intelligence and Knowledge can help fact acquisition from text:
- by generating new facts
- correcting extraction errors
- validating automatically generated facts

**Pros and cons:**
- Higher output quality
- Costs increase with scale (in terms of money & time)

**Human Computing & Crowdsourcing (HC)**

**Experiments**

- Delivers higher precision and recall
- At reduced crowdsourcing costs!

**Higgins System**

Main Idea: Use automated IE to generate questions and meaningful answers for HC

**Information Extraction Engine**

Web Document Collections

Statistical Language Models

Entities & Relational Phrases

**Extract candidates and rank**

Given any textual context \( \mathbf{S} \),

1. detect entity mentions
2. rank relations using statistical translation model over phrases

\[
Pr(\mathbf{R}|\mathbf{S}) = \sum_{\mathbf{P} \in \mathbf{R}} Pr(\mathbf{P}|\mathbf{C}) Pr(C|\mathbf{S})
\]

**Extract candidates and rank**

Example context:

"Bond develops feelings for Vesper"

\[ \text{[Bond: } \text{[develops feelings for: } \text{[Vesper: ]} \text{]}
\]

**Game questions / Crowdsourcing HITS**

**Human Computing Engine**

**Evaluate using crowds**

- HC engine creates crowdsourcing HITS or game questions from extracted fact candidates
- Provides top-k ranked candidates as answers
- Diversify answer choices
- Provide the extraction context
- Interesting HC games possible:
  - Anonymized slots for entities
  - Interaction games with multiple players
  - Support player with clues and callbacks

**Coverage higher than state-of-the-art IE**

**Comparisons with automated IE**

61% 84% 34%

**Human Intelligence**

Max Planck Institute for Informatics