Leveraging Knowledge Graphs for Web Search

Part 4 - Crowdsourcing for Knowledge Graphs
Gianluca Demartini
University of Sheffield
gianluca.demartini.net
Crowdsourcing

• Leverage human intelligence at scale to solve
  – Tasks simple for humans, complex for machines
  – With a large number of humans (the Crowd)
  – Small problems: micro-tasks (Amazon MTurk)

• Examples
  – Wikipedia, Image tagging, reCaptcha

• Incentives
  – Financial, fun, visibility

• See also my tutorial at ESWC 2013 and ISWC 2013

Slides: gianlucademartini.net/kg
# Types of Crowdsourcing Tasks

<table>
<thead>
<tr>
<th>Task Granularity</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex Tasks</td>
<td>• Build a website</td>
</tr>
<tr>
<td></td>
<td>• Develop a software system</td>
</tr>
<tr>
<td></td>
<td>• Overthrow a government?</td>
</tr>
<tr>
<td>Simple Projects</td>
<td>• Design a logo and visual identity</td>
</tr>
<tr>
<td></td>
<td>• Write a term paper</td>
</tr>
<tr>
<td>Macro Tasks</td>
<td>• Write a restaurant review</td>
</tr>
<tr>
<td></td>
<td>• Test a new website feature</td>
</tr>
<tr>
<td></td>
<td>• Identify a galaxy</td>
</tr>
<tr>
<td>Micro Tasks</td>
<td>• Label an image</td>
</tr>
<tr>
<td></td>
<td>• Verify an address</td>
</tr>
<tr>
<td></td>
<td>• Simple entity resolution</td>
</tr>
</tbody>
</table>

Background

A Crowdsourcing Platform allows *requesters* to publish a crowdsourcing request (*batch*) composed of multiple tasks (*HITs*)

Programmatically Invoke the crowd with APIs or using a website

*Workers* in the crowd complete tasks and obtain a monetary reward
Case-Study: Amazon MTurk

• Micro-task crowdsourcing marketplace
• On-demand, scalable, real-time workforce
• Online since 2005 (still in “beta”)
• Currently the most popular platform
• Developer’s API as well as GUI
Amazon MTurk

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results
MTurk is a Marketplace for HITs

<table>
<thead>
<tr>
<th>HITs Available (most first)</th>
<th>Show all details</th>
<th>Hide all details</th>
<th>1 2 3 4 5</th>
<th>Next</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private Information about a Product</td>
<td>HIT Expiration Date: May 23, 2015 (4 weeks 1 day)</td>
<td>Reward: $0.06</td>
<td>Hits Available: 11526</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: requester</td>
<td>Time Allotted: 25 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product Attribute Tagging - April 17th</td>
<td>HIT Expiration Date: May 23, 2015 (4 weeks 2 days)</td>
<td>Reward: $0.03</td>
<td>Hits Available: 23857</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: rsg</td>
<td>Time Allotted: 60 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any B.2</td>
<td>HIT Expiration Date: May 22, 2015 (4 weeks 1 day)</td>
<td>Reward: $0.00</td>
<td>Hits Available: 19740</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: rhabdf</td>
<td>Time Allotted: 48 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geo Result Relevance: Tue Apr 21 10:48:14 PDT 2015</td>
<td>HIT Expiration Date: May 22, 2015 (4 weeks 1 day)</td>
<td>Reward: $0.00</td>
<td>Hits Available: 10774</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: Amazon Requester Inc.</td>
<td>Time Allotted: 60 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type the text from this image. Carefully. Productivity and bonuses guaranteed</td>
<td>HIT Expiration Date: Apr 30, 2015 (6 days 23 hours)</td>
<td>Reward: $0.01</td>
<td>Hits Available: 10510</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: CopyText Inc.</td>
<td>Time Allotted: 10 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcribe up to 25 Seconds of Media to Text - Earn up to $0.12 per HIT</td>
<td>HIT Expiration Date: Apr 21, 2016 (51 weeks 6 days)</td>
<td>Reward: $0.08</td>
<td>Hits Available: 6792</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: Crowdsurf Support</td>
<td>Time Allotted: 15 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fun and Fast Fashion Tagging</td>
<td>HIT Expiration Date: Apr 28, 2015 (5 days 11 hours)</td>
<td>Reward: $0.02</td>
<td>Hits Available: 6460</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: gfrt</td>
<td>Time Allotted: 60 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geo Result Relevance: Wed Apr 08 14:30:08 PDT 2015</td>
<td>HIT Expiration Date: May 10, 2015 (2 weeks 2 days)</td>
<td>Reward: $0.00</td>
<td>Hits Available: 6182</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: Amazon Requester Inc.</td>
<td>Time Allotted: 60 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcribe up to 25 Seconds of General Content to Text - Earn up to $0.14 per HIT</td>
<td>HIT Expiration Date: Apr 21, 2016 (51 weeks 6 days)</td>
<td>Reward: $0.09</td>
<td>Hits Available: 6043</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: Crowdsurf Support</td>
<td>Time Allotted: 15 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What a meal by Gaze [hard mode]</td>
<td>HIT Expiration Date: Apr 23, 2015 (8 hours 40 minutes)</td>
<td>Reward: $0.10</td>
<td>Hits Available: 4582</td>
<td>View a HIT in this group</td>
<td></td>
</tr>
<tr>
<td>Requester: at</td>
<td>Time Allotted: 60 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Amazon MTurk

• Requesters create tasks (HITs)
• The platform takes a fee (30% of the reward)
• Workers preview, accept, submit HITs
• Requesters approve, download results

• If the results are approved, workers are paid
mturk-tracker.com

- Collects metadata about each visible batch (Title, description, rewards, required qualifications, HITs available etc)

- Records batch progress (every ~20 minutes)

We note that the tracker reports data periodically only and does not reflect fine-grained information (e.g., real-time variations)
A 5-years analysis of the Amazon MTurk market evolution:
### Top requesters last week

**Top-1000 Requesters, report for July 25, 2015 to August 24, 2015**

<table>
<thead>
<tr>
<th>Requester name</th>
<th>hits</th>
<th>reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speechpad</td>
<td>39932</td>
<td>$270,902.29</td>
</tr>
<tr>
<td>WorkFusion</td>
<td>2215</td>
<td>$9,496.00</td>
</tr>
<tr>
<td>CastingWords</td>
<td>10229</td>
<td>$6,405.64</td>
</tr>
<tr>
<td>VidAngel</td>
<td>193</td>
<td>$2,757.55</td>
</tr>
<tr>
<td>Amazon Requester Inc - browse classification</td>
<td>39795</td>
<td>$2,387.70</td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>170</td>
<td>$2,210.00</td>
</tr>
<tr>
<td>p9r</td>
<td>20440</td>
<td>$1,853.52</td>
</tr>
<tr>
<td>Mark Yatskar</td>
<td>19206</td>
<td>$1,456.75</td>
</tr>
<tr>
<td>Mediaeval Search Hyper</td>
<td>13926</td>
<td>$1,392.60</td>
</tr>
<tr>
<td>World Vision International</td>
<td>19769</td>
<td>$1,262.00</td>
</tr>
</tbody>
</table>
SLAs are expensive
Why Crowdsourcing for IR?

• Easy, cheap and fast labeling
• Ready-to use infrastructure – MTurk payments, workforce, interface widgets – CrowdFlower quality control mechanisms, etc.
• Allows early, iterative, frequent experiments – Iteratively prototype and test new ideas – Try new tasks, test when you want & as you go
• Proven in major IR shared task evaluations – CLEF image, TREC, INEX, WWW/Yahoo SemSearch
Crowdsourcing Ontology Mapping

• Find a set of mappings between two ontologies

• Micro-tasks:
  – Verify/identify a mapping relationships:
    • Is concept A the same as concept B
    • A is a kind of B
    • B is a kind of A
    • No relation
Crowdsourcing Ontology Mapping

- Crowd-based outperforms purely automatic approaches
Crowdsourcing Ontology Engineering

• Ask the crowd to create/verify subClassOf relations
  – “Car” is a “vehicle”

• Does it work for domain specific ontologies?
  – A “protandrous hermaphroditic organism” is a “sequential hermaphroditic organism”

• Workers perform worse than experts
• Workers presented with concept definitions perform as good as experts
Application of Crowdsourcing to Knowledge Graphs

- Entity Linking (Demartini et al., WWW2012)
- Search Query Understanding (Demartini et al., CIDR2013)
- Search Result Extraction (Bernstein et al., CHI2012)
- KG enrichment (Ipeirotis and Gabrilovich, WWW2014)
Facebook is not waiting for its initial public offering to make its first big purchase.

In its largest acquisition to date, the social network has purchased Instagram, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.

Facebook (http://dbpedia.org/resource/Facebook) is not waiting for its initial public offering to make its first big purchase.

In its largest acquisition to date, the social network has purchased Instagram (http://dbpedia.org/resource/Instagram), the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.

Instagram for Android is now available

At long last, Instagram finally releases the Android version of its app. (http://dbpedia.org/resource/Instagram)

Instagram has been around since 2010, available only to iOS devices. Android users have been waiting patiently, with repeated promises of an Android version arriving soon.
ZenCrowd

• Combine both algorithmic and manual linking
• Automate manual linking via crowdsourcing
• Dynamically assess human workers with a probabilistic reasoning framework
Algorithmic Matching

• Inverted index over LOD entities
  – DBPedia, Freebase, Geonames, NYT

• TF-IDF (IR ranking function)

• Top ranked URIs linked to entities in docs

• Threshold on the ranking function or top N
Entity Factor Graphs

- **Graph components**
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints

- **Probabilistic Inference**
  - Select all links with posterior prob >τ

2 workers, 6 clicks, 3 candidate links
Entity Factor Graphs

• Training phase
  – Initialize worker priors
  – with $k$ matches on known answers

• Updating worker Priors
  – Use link decision as new observations
  – Compute new worker probabilities

• Identify (and discard) unreliable workers
Experimental Evaluation

• Datasets
  – 25 news articles from
    • CNN.com (Global news)
    • NYTimes.com (Global news)
    • Washington-post.com (US local news)
    • Timesofindia.indiatimes.com (India news)
    • Swissinfo.com (Switzerland local news)
  – 40M entities (Freebase, DBPedia, Geonames, NYT)

<table>
<thead>
<tr>
<th></th>
<th>US Workers</th>
<th>Indian Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>GL News</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>US News</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>IN News</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>SW News</td>
<td>0.72</td>
<td>0.80</td>
</tr>
<tr>
<td>All News</td>
<td>0.80</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Experimental Evaluation

- Entity Linking with Crowdsourcing and agreement vote (at least 2 out of 5 workers select the same URI)

<table>
<thead>
<tr>
<th></th>
<th>US Workers</th>
<th></th>
<th>Indian Workers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P  R  A</td>
<td>P  R  A</td>
<td></td>
<td>P  R  A</td>
</tr>
<tr>
<td>GL News</td>
<td>0.79 0.85 0.77</td>
<td>0.60 0.80 0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>US News</td>
<td>0.52 0.61 0.54</td>
<td>0.50 0.74 0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN News</td>
<td>0.62 0.76 0.65</td>
<td>0.64 0.86 0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SW News</td>
<td>0.69 0.82 0.69</td>
<td>0.50 0.69 0.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All News</td>
<td><strong>0.74</strong> 0.82 0.73</td>
<td><strong>0.57</strong> 0.78 0.59</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Top-1 precision: 0.70
Experimental Evaluation

• Entity Linking with ZenCrowd
  – Training with first 5 entities + 5% afterwards
  – 3 consecutive bad answers lead to blacklisting

<table>
<thead>
<tr>
<th></th>
<th>US Workers</th>
<th>Indian Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>GL News</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>US News</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>IN News</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>SW News</td>
<td>0.72</td>
<td>0.80</td>
</tr>
<tr>
<td>All News</td>
<td>0.80</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Experimental Evaluation

- Worker Selection

![Graph showing worker precision vs. number of tasks for US and IN workers. The graph includes a scatter plot with data points for each worker category and a legend indicating US Workers and IN Workers.](image)
Lessons Learnt

• Crowdsourcing + Prob reasoning works!

• But
  – Different worker communities perform differently
  – Many low quality workers
  – Completion time may vary (based on reward)

• Need to find the right workers for your task
  (see WWW13 paper)
ZenCrowd Summary

• ZenCrowd: Probabilistic reasoning over automatic and crowdsourcing methods for entity linking
• Standard crowdsourcing improves 6% over automatic
• 4% - 35% improvement over standard crowdsourcing
• 14% average improvement over automatic approaches
**Blocking** for Instance Matching

- Find the instances about the same real-world entity within two datasets
- Avoid Comparison of all possible pairs
  - Step 1: cluster similar items using a cheap similarity measure
  - Step 2: n*n comparison within the clusters with an expensive measure
Three-stage blocking with the Crowd for Data Integration

• 1. Cheap clustering/inverted index selection of candidates
• 2. Expensive similarity measure
• 3. Crowdsource low confidence matches

Crowd-powered Direct Answers and Query Understanding
Extract Direct Answers w/ Crowdsourcing

Bernstein et al., Direct Answers for Search Queries in the Long Tail, CHI 2012.
birthdate of the mayor of the capital city of Italy
capital city of Italy

Rome
Italy, Capital
mayor of rome

Ignazio Marino

The outgoing Mayor of Rome, Gianni Alemanno (PdL), stood for election for a second term as mayor. The center-left candidate, heart surgeon Ignazio Marino was be chosen by a multi-party primary election on 7 April 2013. Control of the 15 municipi of the Italian capital was decided in the election.

birthdate of Ignazio Marino

March 10, 1955 (age 60 years)
Ignazio Marino, Date of birth

Gianni Alemanno
March 3, 1958

Nicola Zingaretti
October 11, 1965

Salvatore Buzzi
November 15, 1955

Ignazio Roberto Maria Ma
transplant surgeon and th
of Rome. He is a member
Democrat Party and hel
Italian Senate from 2006 t
as mayor of Rome. Wiki
Born: March 10, 1955 (ag
Italy
Education: Catholic Uni
Sacred Heart (1979)
Party: Democratic Party
Motivation

• Web Search Engines can answer simple factual queries directly on the result page

• Users with complex information needs are often unsatisfied

• Purely automatic techniques are not enough
• We want to solve it with Crowdsourcing!
CrowdQ

- CrowdQ is the first system that uses crowdsourcing to
  - *Understand* the intended meaning
  - *Build* a structured query template
  - *Answer* the query over Linked Open Data

<table>
<thead>
<tr>
<th>City</th>
<th>Mayor</th>
<th>Birthdate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rome, Italy</td>
<td>Gianni Alemanno</td>
<td>March 3, 1958</td>
</tr>
<tr>
<td>Venice, Italy</td>
<td>Giorgio Orsoni</td>
<td>August 29, 1946</td>
</tr>
<tr>
<td>Milan, Italy</td>
<td>Giuliano Pisapia</td>
<td>May 20, 1949</td>
</tr>
</tbody>
</table>

Cities in Italy | Italy Travel Guide
www.italylogue.com/italian-cities

Learn about the best cities in Italy to visit, and some Italian cities you might never have heard of before. These cities in Italy are all great for visitors.

Top Ten Cities for Visitors to Italy - Top Italian Cities to See
goitaly.about.com/od/planningandinformation/tp/topcities.htm

Italy has many beautiful and historic cities that are well worth a visit. Here are our picks for the ten best cities for visitors to Italy.

Italian Cities and Towns - Italy
en.comuni-italiani.it/
CrowdQ Architecture

**Off-line**: query template generation with the help of the crowd

**On-line**: query template matching using NLP and search over open data
Hybrid Human-Machine Pipeline

Q= birthdate of actors of forrest gump

Query annotation

Noun

Noun

Named entity

Verification

Is forrest gump this entity in the query?

Entity Relations

Which is the relation between: actors and forrest gump

Starring <dbpedia-owl:starring>

Schema element

Verification

Is the relation between:
Indiana Jones – Harrison Ford
Back to the Future – Michael J. Fox
of the same type as
Forrest Gump - actors
Structured query generation

Q= birthdate of actors of Forrest Gump

SELECT ?y ?x
?z <rdfs:label> ‘Forrest Gump’ }

Results from BTC09:

<http://dbpedia.org/resource/Robin_Wright_Penn> 1966-04-08
<http://dbpedia.org/resource/Tom_Hanks> 1956-07-09
<http://dbpedia.org/resource/Sally_Field> 1946-11-06
<http://dbpedia.org/resource/Gary_Sinise> 1955-03-17
<http://dbpedia.org/resource/Mykelti_Williamson> 1960-03-04
Summary

• Crowdsourcing as a means to access a large number of on-line workers on-demand
• Hybrid human-machine systems to scale over large amount of data with high quality
  – Entity Linking
  – Data Integration
  – Answer extraction
  – Keyword query understanding

An overview of such systems:
I’m hiring

• A post-doctoral researcher to start Jan 2016 (or later)
• On Crowdsourcing and Human Computation systems
• Get in touch
  – g.demartini@sheffield.ac.uk