HIGGINS: Knowledge Acquisition meets the Crowds

Sarath Kumar Kondreddi*, Peter Triantafillou & Gerhard Weikum

**Rationale**

Knowledge Acquisition efforts critically rely on IE techniques to build useful knowledge bases from unstructured text. However, automated IE cannot cope with sentences of complex structure, use of co-references and ambiguous wording. The following snippet from [www.imdb.com](http://www.imdb.com) is a daunting example: “He quickly grabs Vesper and they kiss in the stairway entrance to cover themselves.”

**Key Idea**

Employ human computing (HC) for discovering facts! Human knowledge and intelligence can detect relations between entities from special topics such as movies, books or medicine and can extract relations between entities unambiguously even if they are not explicitly mentioned. The key is to tap human potential to assess automatically generated facts and to correct errors. But employing HC to correct every extracted fact is prohibitively expensive.

**HIGGINS Architecture**

WE Engine

- Web Document Collections
- Statistical Language Models
- Entities & Relational Phrases

HC Engine

- Game questions / Crowdsourcing NITS
- MoveIEWizard
- Aggregate logs and statistics

**Our contribution**

We built HIGGINS, a principled framework to combine IE and HC for advances KA tasks such as relation extraction. It contains:

- An IE component that utilizes entity lexicons, statistical LMs, combined with semantic resources for relational phrases.
- A complementing HC component that employs ranking and diversification based on IE-generated statistics to produce fact candidates.

**Design Principles**

1. Build an aggressive IE engine that captures as many entity-centric and relational phrases as possible aiming for high recall. Expand this set by specifically designed statistical language models.
2. Use statistics and heuristics to generate important questions featuring important entities and salient but not obvious relationships.
3. Overcome sparseness by complementing with relational phrases from semantic resources such as WordNet and ConceptNet, and ReVerb and Patty collections. Use mixture models to combine generate and rank them.
4. Generate candidate answers as multiple-choice by ranking judiciously using corpus-collected statistics. Additionally diversify the choices to avoid boring the player.

**Player Perspective**

- a) Answer the factoid questions on existing movie stories by choosing an option or by typing in.
- b) Awarded points based on player agreement.

**System Perspective**

- a) Identify, resolve and pair up entities in the document D, based on their context S (e.g., entities occurring in same sentence).
- b) Come up with plausible relations for an entity pair mentioned in the context by ranking candidates using a Statistical Language Model,

\[
P_r(R) = (\alpha \cdot P_r(R | S)) + (1 - \alpha \cdot P_r(R | D))
\]

where \(P_r(R | S) = \sum_{C \in C(r)} P_r(C | S) \cdot P(C | r) \) or textual patterns in relation \(R\).

Evaluated using Wikipedia movie plots with “in-house” and “in-the-wild” crowds. Results show high precision/recall and quick convergence.

<table>
<thead>
<tr>
<th></th>
<th>HIGGINS Games</th>
<th>CrowdFlower</th>
<th>No. of Players</th>
<th>No. of questions</th>
<th>Fleiss Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.8160</td>
<td>0.7372</td>
<td>2</td>
<td>209</td>
<td>0.71</td>
</tr>
<tr>
<td>Recall</td>
<td>0.8930</td>
<td>0.7522</td>
<td>4</td>
<td>172</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>118</td>
<td></td>
<td></td>
<td>0.69</td>
</tr>
</tbody>
</table>

* Primary contact: sarath.kondreddi@mpi-inf.mpg.de
Contact attending WWW 2013: Marc Spaniol (mspaniol@mpi-inf.mpg.de)