University of Glasgow

Kumar Kondreddi, S., Triantafillou, P., and Weikum, G. (2013) HIGGINS: where knowledge acquisition meets the crowds. In: International Conference on the World Wide Web, 13-17 May 2013, Rio de Janeiro, Brazil.

Copyright © 2013 International World Wide Web Conference Committee

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

The content must not be changed in any way or reproduced in any format or medium without the formal permission of the copyright holder(s)

When referring to this work, full bibliographic details must be given

http://eprints.gla.ac.uk/80439/

Deposited on: 13 September 2013

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

HIGGINS: where Knowledge Acquisition meets the Crowds

Sarath Kumar Kondreddi
Max Planck Institute for Informatics
Saarbrücken, Germany
skondred@mpi-inf.mpg.de

Peter Triantafillou
University of Patras
Rio-Patra, Greece
peter@ceid.upatras.gr

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

We present HIGGINS, an engine for high quality *Knowledge Acquisition (KA)*, placing special emphasis on its architecture. The distinguishing characteristic and novelty of HIGGINS lies in its special blending of two engines: An automated *Information Extraction (IE)* engine, aided by *semantic resources*, and a game-based, *Human Computing engine (HC)*. We focus on KA from web data and text sources and, in particular, on deriving relationships between entities. As a running application we utilise movie narratives, using which we wish to derive relationships among movie characters.

1. THE RATIONALE

KA critically relies on IE technology, combining methods from pattern matching, computational linguistics, and statistical learning. *Open IE* methods can derive a wide diversity of relational facts (instances of binary relationships) based on detecting and analyzing noun phrases (for entities) and verb-centric phrases (for relations), such as: "Vesper" "finally falls for" "Bond". Hence, fully-automated IE plays a key role within HIGGINS.

However, there exist fundamental limitations of IE technology. IE methods can yield noisy or non-sensical relationships (knowledge triples) such as: "Vesper" "certainly has" "Bond" (from the sentence "... certainly has ... sized up"). This occurs because fully automated IE generally faces fundamental obstacles as input sentences often have extremely complex structures, use of pronouns and other anaphoras, and ambiguous wording. The following snippet (from `imdb.com`) is a daunting example: "He quickly grabs Vesper and they kiss in the stairway entrance to cover themselves."

Crowdsourcing (a.k.a *Human Computing (HC)*) recently has been employed successfully as a means to help with tasks where computer-performed, automated solutions are deemed inadequate. Our thesis is that HC is a natural alternative to overcome fundamental limitations of automated IE. It can tap human intelligence and knowledge to assess candidate facts and to correct errors. Further, humans can readily recognize "surreal" relationships, (e.g., occurring in dreams, being imaginary, contained in intentions, etc.), such as "Vesper Lynd" *"pretends to love"* "James Bond", or "Voldermort" *"plans to kill"* "Harry Potter" and note that they do not correspond to true facts. Human *intelligence* can help to resolve pronouns in complex sentences or to identify erroneous paths in the dependency-parsing of natural language. Human *knowledge* on special topics such as movies, books, or medicine can add new facts, that may be entirely missing from the text, or even help compile/derive facts, e.g., relationships between characters, that are virtually impossible to extract automatically, as they may not be explicitly mentioned. Despite this great potential, to our knowledge there have been no previous attempts to employ HC and crowdsourcing platforms for the difficult KA task of extracting relationships between individual entities.

For applications with human experts, (e.g., movies with movie aficionados, or book lovers, or experts in diseases and medicine, etc.), one would expect that HC can be nicely cast into game form, thus enticing more users to contribute on the KA task. However, despite the inherent promises of HC for KA, humans alone cannot carry this burden. First, the number of real experts is typically limited. Second, these experts are not so likely to participate in online games. Hence, inevitably, HC output will contain a wide range from high-quality to highly noisy and incorrect facts. One may think that these HC errors could be compensated by large-scale crowdsourcing, with redundant Human Intelligence Tasks (HITS) and statistic reasoning over many contributors. However, there is still the issue of the total cost: each HIT may cost a few cents only, but paying for hundreds of thousands or millions of HITs quickly becomes prohibitive.

2. STRATEGY, DESIGN PRINCIPLES, AND ARCHITECTURE

HIGINS is a novel architecture that *blends* HC inputs with machine-centric IE and thus overcomes the limitations of either a purely HC-based or purely IE-based approach to advanced KA. The key idea is to use automatic IE to generate questions and candidate answers (for multiple-choice questions), for a KA game. Our expectation is that this can improve the quality of user contributions and reduce the overall cost of crowdsourcing.

HIGGINS' strategy for KA of entity-relationship facts proceeds in two phases:

1. IE phase: First we employ automated IE on appropriate Web corpora, in order to derive candidates for relation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WWW '13, Rio de Janeiro, Brazil
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$10.00.

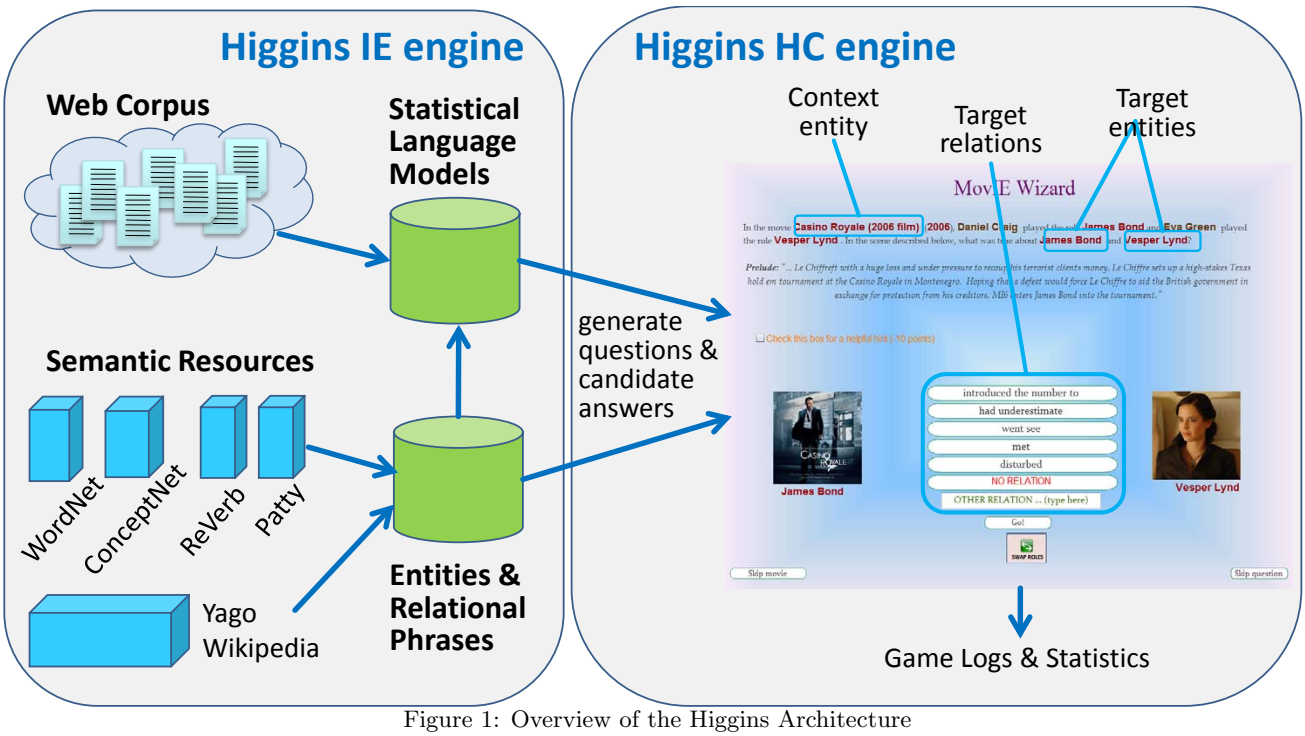

Figure 1: Overview of the Higgins Architecture

instances, with an open set of potential relations. Here, we use a suite of techniques from computational linguistics, including dependency parsing (with the Stanford Parser) and pronoun resolution (with our own customized method). The resulting triples are usually of very mixed quality, necessitating the second stage.

2. HC phase: The sets of candidates from the IE stage and their underlying patterns are then used to generate HITs in game form. Abstractly, each HIT presents the user with a *knowledge quad* of the form $(c, e1, r, e2)$ where $e1$ and $e2$ are entities, r is a relation, and c is a cue or textual context. One or more of the components $e1$, r, and $e2$ can be empty slots (variables) to be filled by the user; we may present a multiple-choice list to the user to pick the missing value. Concretely, the quads are presented in the form of questions, with candidate answers and additional free-text fields for entering further values. We focus on the case where the relationship r is left to be filled, and both entities and the context are given.

The design principles for HIGGINS are as follows:

1. The IE engine is tuned to work very aggressively (aiming for high recall) capturing as many relational patterns as possible, and we expand this set by specifically designed *statistical (translation) language models (LM's).*
2. We use statistics and heuristics to generate interesting questions about *important* entities and *salient but not obvious* relationships. Candidate answers for multiple-choice input are judiciously *ranked*, using corpus-collected statistics, and an additional *diversification* step serves to avoid boring the user with near-duplicate choices.
3. The statistically derived relational phrases for candidate answers are complemented by phrases from *semantic resources*, specifically, the WordNet and ConceptNet thesauri and the ReVerb and PATTY collections of relational patterns All this information is combined by a mixture model for the LM that generates, expands, and ranks relationships for a given context.

Architecture. Figure 1 depicts the architecture and main components of HIGGINS.

3. RESULTS AND CONCLUSIONS

HIGGINS contributes a systematic and principled way for generating HITs, based on novel statistical language models relying on extensive large corpora, semantic resources, and on an extensive dictionary. The latter serves as a conduit between free-form, human expressions for entities and relationships and the corresponding system "proper" names for these. The result is a meaningful engagement of human game players, facilitating the validation and assessment of facts obtained from the IE engine, and the derivation of new knowledge.

We have conducted extensive experiments using Wikipedia movie plots and crowds "in the wild" (with CrowdFlower) and crowds of students in our lab. Our results show that (i) the full HIGGINS architecture can more than double recall and precision of derived facts (achieving recall and precision numbers over 70%) compared against a purely statistical IE engine or a purely semantics-based IE engine. Further, only a small number of game players are required in order to achieve this performance - a fact that translates into reduced dollar-costs for human engagement.