Social Computing Project

Link to video presentation -
https://youtu.be/xdebhmMqkXU

Name - Himanshu Maheshwari
Roll No - 20171033

Problem Statement

Study the role of Wikidata in enhancing the productivity of the Wikipedia writers and enhance Wikidata for Hindi and Telugu using gamification.

Mentors:

Tushar Abhishek
Nikhil Pattisapu

Role of Wikidata in enhancing the productivity of the Wikipedia writers

There are basically two Wikipedia writers:

1. **Humans**
   Human writers could use Wikidata as a knowledge pool to take reliable data from.

2. **Bots**
   Main role that wikidata plays in comes from its machine friendly nature. This machine friendly nature could be leveraged by bots to create wikipedia pages.

On top of this Wikidata is free and its data could be use without copyright restrictions.

Enhance Telugu and Hindi Wikidata using gamification.

Motivation:

- 56% of the webpages are written in English while only 0.1% of the webpages are written in Hindi or Telugu.

- Thus there is a lack of Hindi or Telugu representation in various web entries and thus we would like to enhance Hindi or Telugu Wikidata to do away with English dominance.

Gamification:

Gamification is the application of typical elements of game playing (e.g. point scoring, competition with others, rules of play) to other areas of activity (here enhancing Wikidata for Hindi and Telugu), to encourage engagement with a product or service.

Existing Solutions

There are a lot of existing solutions, however all of them are of same kind. These games are basic questions and answer games.

- **First user is asked for his/her choice of field.**
- **After that there are some fixed questions with some probable answers. Users are then asked to choose from those options.**
- **Once some trust is reached relevant wikidata edit is made.**

All of these games are basic MCQ/Fill in the blanks type question where user is asked to choose from a pool of options.

Solutions Proposed

- I take the basic idea from these games and create more sophisticated game with more reliability on users.
- Since the existing game has already created options, it brings in some bias of the creator/programmer. I want to take down that bias.
- Thus the game seeks potential answers from user and once threshold trust is reached in an answer, that answer is used to make relevant edit in Wikidata.
- The aim is not to make many edits but to make quality edits.

Target Domain: Science & Technology

Target Audience: High School Kids

Target Domain

As already said that the target domain is Science and Technology, yet in this domain we are specifically focusing on following 10 sub-categories:

a. Chemical Compound
b. Chemical Elements
c. Stars
d. Theorems
e. Diseases
f. Softwares
g. Biological Process
h. Rivers
i. Mountains
j. Countries

Solutions Proposed

The Entire Solution has two parts:

1. **Question Creator - It's aim is to create questions about something that should be there in Wikidata and yet is missing.**
2. **The Game - It's aim is to crowdsource answer from the user and make relevant edits.**

Question Creator

- It’s aim is to create questions about something that should be there in Wikidata and yet is missing.
- There are 10 subcategories, all the items that belong of one sub-categories are very closely related. So all the items that are instance of these domains them have some common properties.
- We leverage the fact that all these items that are instance of these domains are closely related and use it to create questions.

Question Creator - An Example

- Let us focus on sub-category elements. There are certain properties that are common to all the elements, like atomic weight, number of protons, location of discovery/creation etc.
- Thus we find all those properties that occur in more than 75%(different for other domains) of the elements. The argument is these properties are general to all the elements and should be there.
- Thus for elements we find following properties that occur in more than 75% of the elements - ['फ्रीबेस पहचानकर्ता', 'कॉमन्स श्रेणी', 'GND अभिज्ञापक', 'LCCN अभिज्ञापक', 'स्रोत द्वारा वर्णित', 'उदहारण है', 'का उपवर्ग', 'विषय की मुख्य श्रेणी', 'कॉमन्स गैलरी', 'चित्र', 'खोजकर्ता या आविष्कारक']

Question Creator - An Example

- Now suppose an element A does not have any of this property say '**खोजकर्ता या आविष्कारक**'(Discoverer or Inventor) then we create question something like this using template based mechanism:

  तत्त्व A का खोजकर्ता या आविष्कारक क्या है? (Google Translation - What is the 'discoverer or inventor' of element A?)

- Thus the template for elements is:

  तत्त्व + Element Name + **का** + Missing Property + **क्या है?** This though might not generate grammatically correct sentences yet they are understable.

Question Creator

- **This way I was able to generate about 1636 questions.**
- **However there is a chance that generated questions may not have an answer or inherently wrong. Such questions are handled in the game (will be described later).**

The Game

The game is divided into two parts:

1. **Multiplayer Scenario - The aim is to present the questions created by the question creator to users and get a list of probable answers with relevant sources from them.**
2. **Singleplayer Scenario- The aim is to present the questions that have been to multiplayer scenario to users and ask them to choose correct answer. Once a threshold trust score is reached on an option it is used to make relevant edits.**

Multiplayer Scenario

- **The aim is to present the questions created by the question creator to users and get a list of probable answers with relevant sources from them.**
- **It is asynchronous multiplayer game meaning all the players do not play simultaneously.**
- **Users are presented with a question and they have to provide answer and source. Users are incentivised to not take a guess and rather leave the question if they do not know the answer. Similarly the are incentivised to provide a source.**

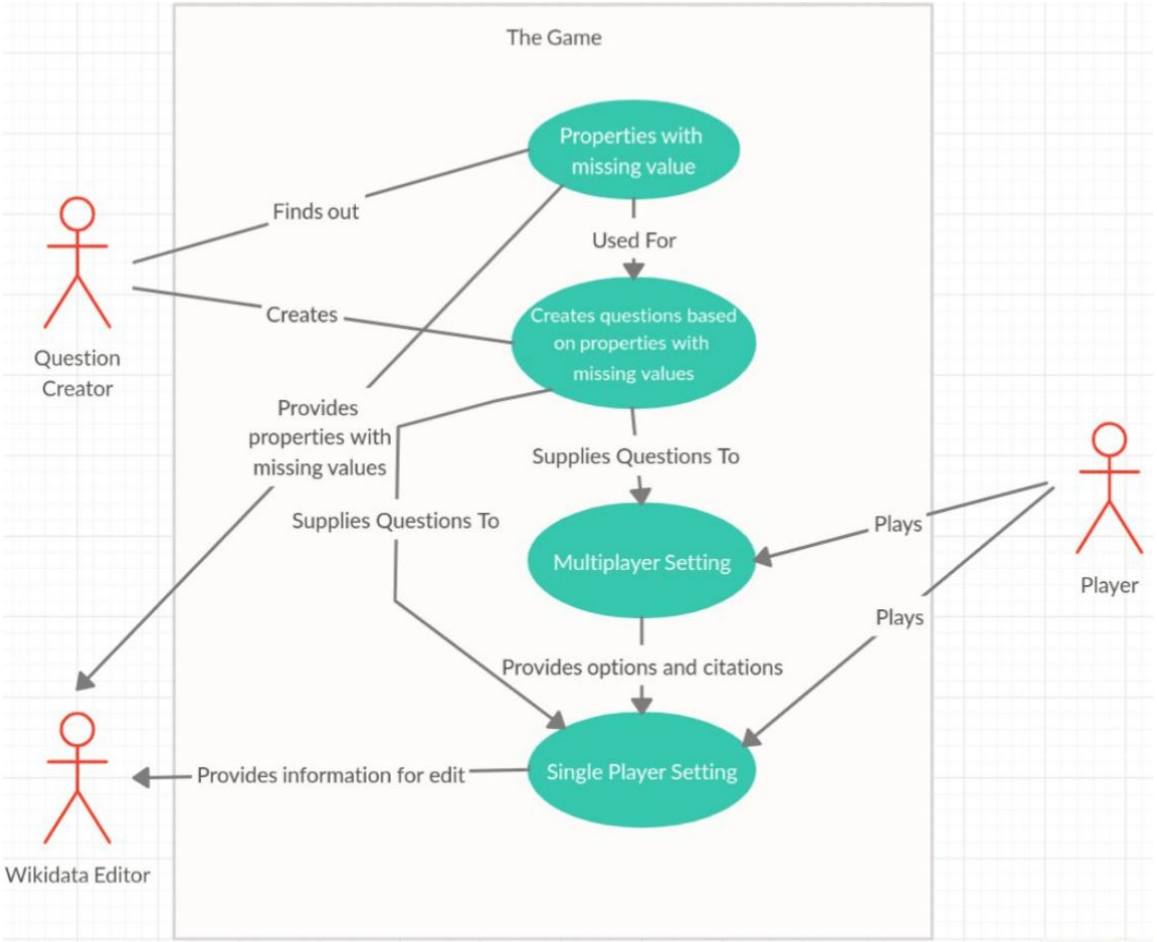
Multiplayer Scenario

- Once all the players have played we check the sources. At the moment I am using simple regex to check the source, later we can include more sophisticated techniques to check the source. Once that is done, we give trust score to all the options and points to users. Users that have cited the correct source gets highest points. Similarly users that has left the answer are given more points than users that took a guess.
- As we talked earlier some questions might not have any possible answers. So if for a question only a few users have provided answers and that too different answer than that question is discarded.

Singleplayer Scenario

- Now once a question has been to multiplayer scenario we have potential answers to that question.
- Now the user is provided with these options and asked to choose. Once again they are incentivised to not guess.
- Finally an option that reaches a threshold of trust score is considered to be correct and is then used to make wikidata entry along with the source(if any).
- As mentioned earlier the aim is not to make many edits but to make quality edits.

Use Case Diagram



Future Direction

1. Better and quality question creator.
2. Fully automated and sophisticated source checking.
3. Introducing aging in questions i.e. if a question is there for a long time it is either presented more to users or deleted.
4. Taking into account users expertize and past contributions in giving trust scores.
5. Currently only Test Wikidata is being edited because my bot account hasn't been approved yet. In future it should be able to edit main Wikidata.
6. Currently only string data types statements are being added, it needs to extended to other Wikidata data types.
7. Extensive testing, I could not do it because I did not have enough player to play.

Resources

The entire source code, literature survey, Problem outline and defining expected outcome (both version 1 & 2) etc. could be found at - https://github.com/him-mah10/social-computing-project

I faced some difficulties setting up pywikibot, using it to add statements and references. So I wrote the following article, please check: https://medium.com/@himanshumaheshwari_41605/bots-to-update-wikidata-181ef932e2dc

Thank - you