Annotation guidelines for Serendipity assessment of WikiWooW-generated entity-pairs

Creation date: 18.02.2025
Final version: 31.03.2025
Authors: Cosimo Palma, Bence Molnár, Maria Pia Di Buono

This document contains the guidelines for annotating the entity-pairs serendipity extracted through the **WIKIWOOW** project.

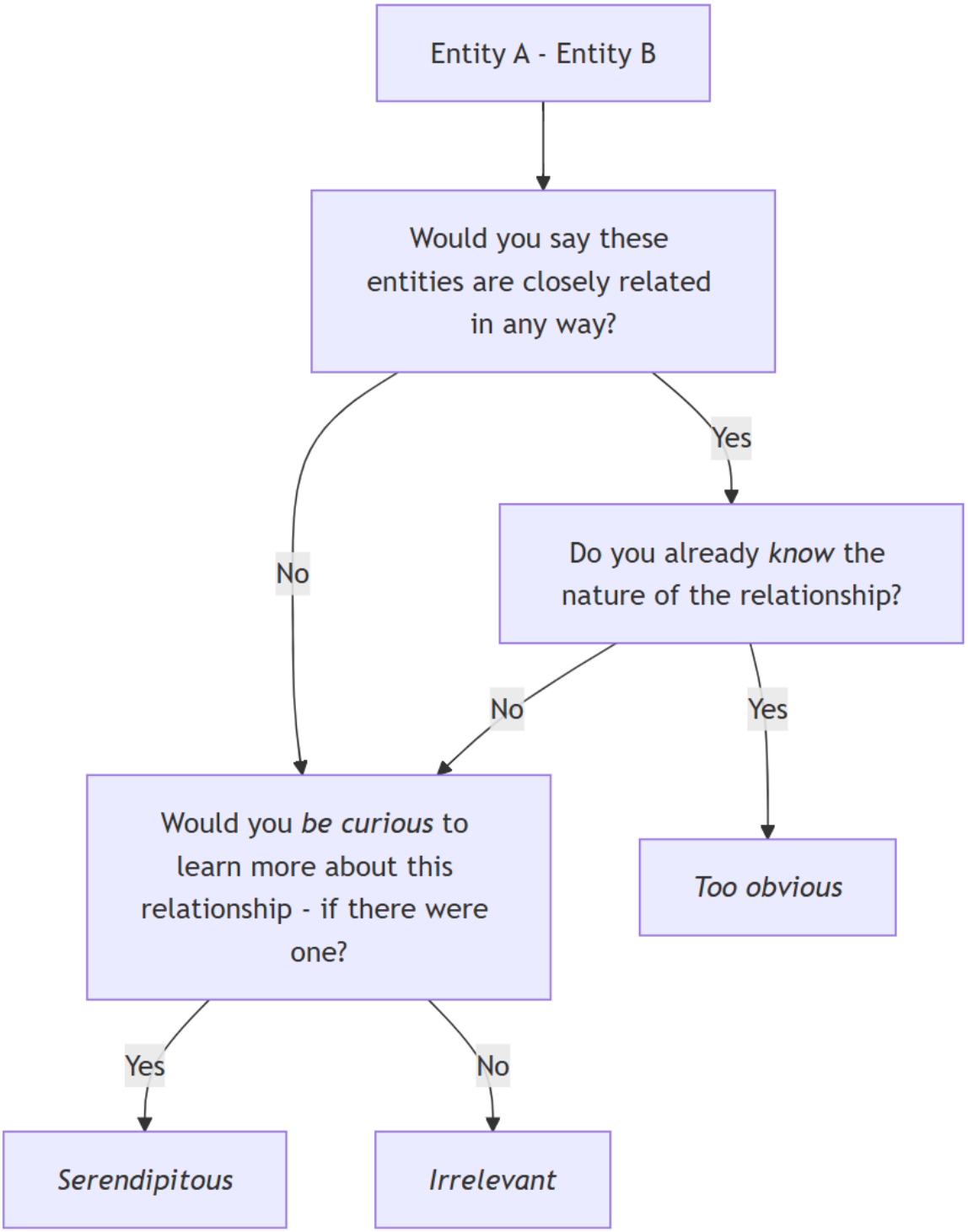
Definitions

These guidelines provide instructions for annotating pairs of entities extracted from the *English Wikipedia*.
The task aims at assessing the degree of *serendipity* between entity pairs.
The interpretation of *Serendipity* assumed for the design of this task is:
"the property of an entity pair whose relationship is at the same time regarded as ***unexpected*** *and* ***relevant*** or ***interesting***".

We also define as *close* a non-trivial relationship not involving other entities. Examples of such relationships are: "child of", "successor of", "grown in", "coeval with", and so on. *Trivial* relationships include "related to", "same as", "are things", "are persons", etc.

Guidelines

The flowchart below outlines the decision-making process you should follow:



By each annotation, a pair of entities is presented, as in the following example:

'Mark Anthony'; 'Alexander the Great'

For each pair the annotators are required to answer the questions displayed in the flowchart above with "yes" or "no".

Depending on the answer, the entity pair will automatically be annotated with a 1, a 0, or a 0.5, where 1 indicates a *serendipitous* relationship and 0 a not serendipitous one (in the flowchart, *too obvious*). 0.5 captures relationships that are regarded as not relevant, although still unexpected (*irrelevant*).

In the example:

'Classical antiquity'; 'Alexander the Great'

the pair will be marked as **0**, as the annotator would probably know the nature of the link: it belongs to the general knowledge.
Following the questions flow presented in the flowchart:
Q1: Do you suppose there could be a close relationship between these two entities?
Y: move to Q2
Q2: Do you already know the nature of the relationship?
Y: annotated as *too obvious*

In the example:

'Chanakya (TV series)'; 'Alexander the Great'

the pair shall be marked at least as **0.5**, as you probably do not imagine how these entities are related. If you also would be interested in further exploring their eventual relationship, then you should flag it as a **1**.
Following the questions flow presented in the flowchart:
Q1: Do you suppose there could be a close relationship between these two entities?
Y: move to Q2
Q2: Do you already know the nature of the relationship?
N: move to Q3
Q3: Would you be curious to learn more about this relationship, if there were one?
Y: *Serendipitous*
N: *Irrelevant*

In the example:

'Mark Anthony'; 'Alexander the Great'

the pair shall probably be marked as serendipitous (**1**): it is well known they do not belong to the same historical period, yet they are very popular personalities that are most possibly linked. Very few know what they have in common. In case you do, but you think others would not, you shall mark this relationship as "serendipitous" as well.

Following the questions flow presented in the flowchart:

Q1: Do you suppose there could be a close relationship between these two entities?

Y: move to Q2

Q2: Do you already know the nature of the relationship?

N: move to Q3

Q3: Would you be curious to learn more about this relationship, if there were one?

Y: *Serendipitous*

Recommendations

Data annotation quality directly impacts machine learning model performance across all metrics. High-quality annotations create a foundation for reliable AI systems, while poor annotations propagate errors throughout the model lifecycle.

Models trained on imprecise or inconsistent data inevitably learn and amplify these inconsistencies, potentially requiring costly retraining cycles.

Furthermore, annotation quality becomes increasingly important as models scale.

For this reason, we hope you might take this task seriously.
In case of uncertainty in answering the last question, we advise you select “No”, unless the majority of your previous answers is not already “Irrelevant”. In that case, please choose “Yes”.

You may search Google for information about the entities, as long as it doesn't bias your response to any particular question. The questions are designed to accommodate any uncertainty or knowledge gaps you might have about a given pair.

Finally, we suggest you should take your time: at least 10 seconds per entity pair annotation. Take a brief pause every 30 pairs.

If you find there is an inconsistency, or a case that was not captured in these guidelines, we would appreciate if you could report to us, writing at:

cosimo.palma@phd.unipi.it