# Annotation guidelines

The Task: Combination Mining

Recombination is the process of combining existing ideas/concepts/components/technologies to create something new - a whole greater than the sum of its parts. It is a great catalyst for innovation: discovering that two known concepts can work together in a synergistic way inspires creative new solutions to problems.

Wheels + Suitcase = Suitcase on Wheels!

Analogies (or inspirations) are a special type of recombination: associating similar ideas from different domains.
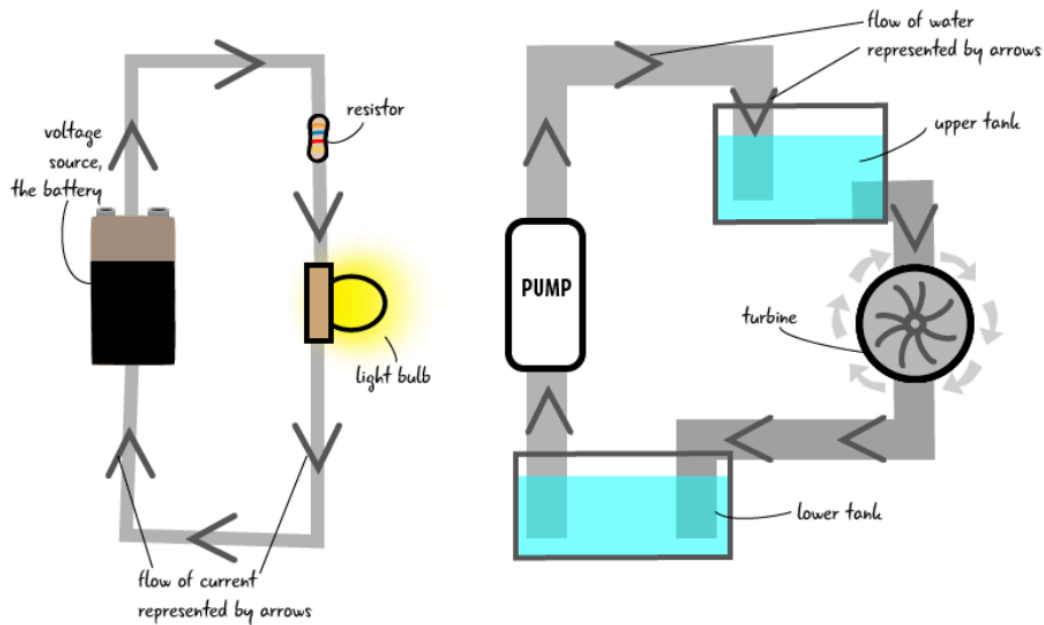
Analogies help transfer novel ideas between different fields.

This annotation task aims to collect training data for a model that would **learn to recombine ideas successfully**.

Use Cases

Researchers can use this model as a helpful tool to find new and interesting research directions. Imagine a system that allows you to explore recombinations across science, and suggest combinations of ideas for specific problems.

Guidelines

The aim of this task is to gather data that helps identify combinations automatically. Your job is to spot the different parts of combinations (including analogies) in the text and connect them based on how the authors have combined them.

General Guidelines for Entity Annotation

- When In doubt → avoid annotating.

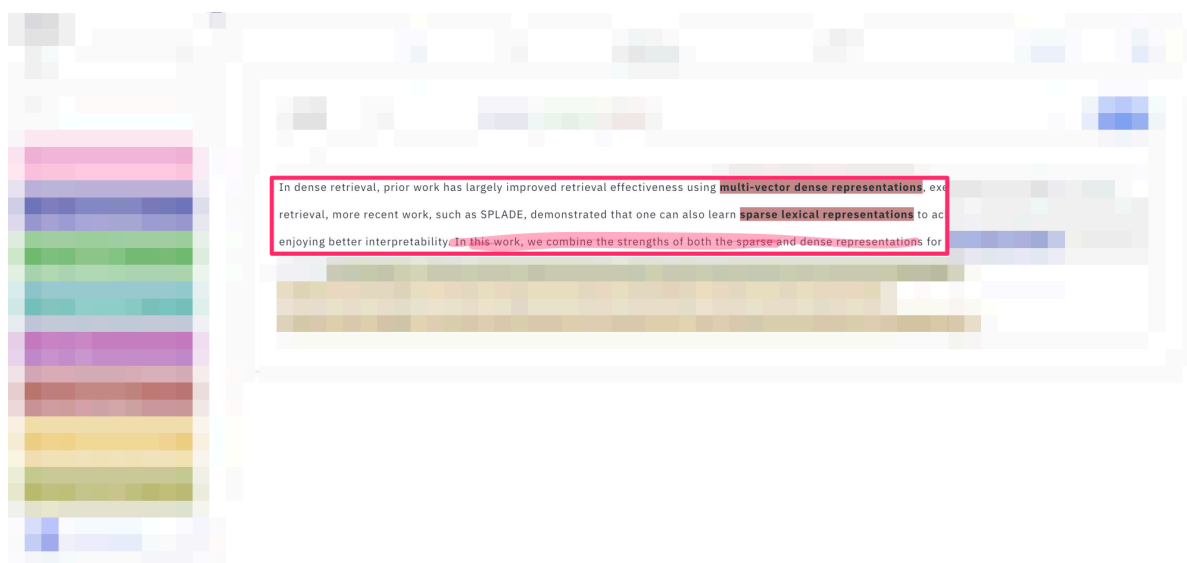- When contemplating the length of a span → take the largest one possible.

- Don't include punctuation marks on the spans start/end if you can (for example, avoid including a full stop at the end of a sentence). The same applies to spaces, tabs and so on.

- To consider whether a certain annotation is good, it's often helpful to think of an end user of the model (like the researcher in): Would a user of this model be happy to get this output? Given the first combination element, is it helpful to get the second?

Entity Types

comb-element

- A comb-element should be an idea (e.g. a method/model/technique/approach) that the authors combine with other comb-elements (one or more) to address a goal.

- The combination itself should be part of the core idea of the paper (suggested by the authors) or a core idea in a cited work.

- The authors should say **explicitly** that they combine the elements.

▼ **Example**

The authors suggest combining "multi-vector dense representation" (one **combination element**) with "sparse lexical representations" (another **combination element**)

▼ Example (more than 2 combination elements)

proposed approach combines saturation-based lexicon construction, statistical corpus analysis methods and graph collocations

▼ Example (vague combination elements)

Graph data is ubiquitous in academia and industry, from social networks to bioinformatics. The pervasiveness of graphs today has raised the demand for algorithms that can answer various questions: Which products would a user lik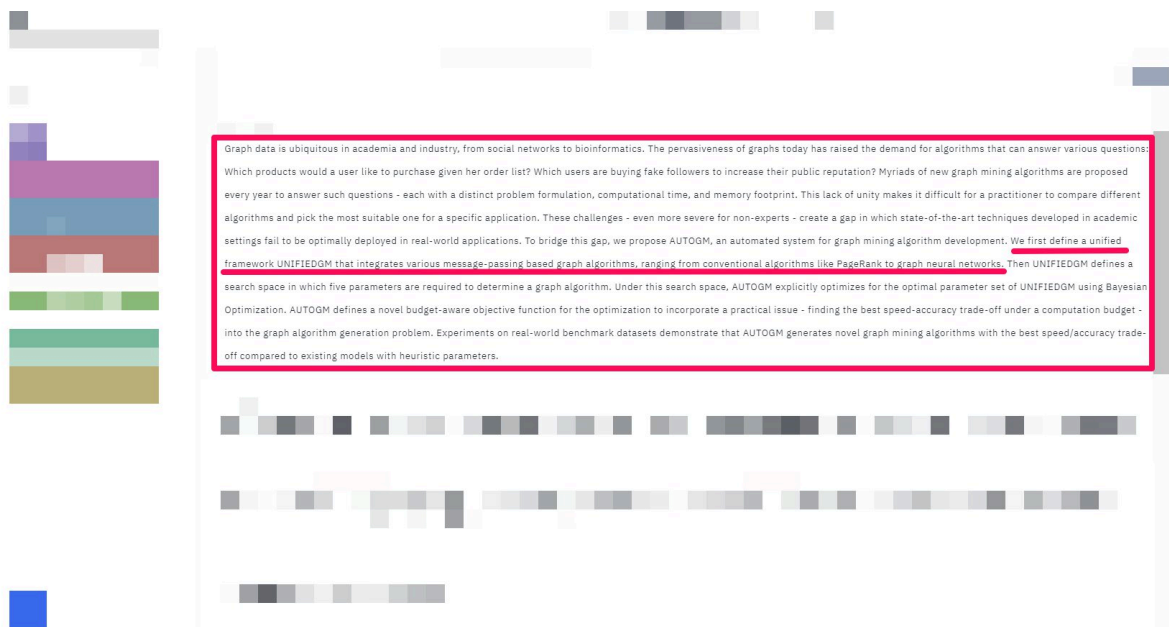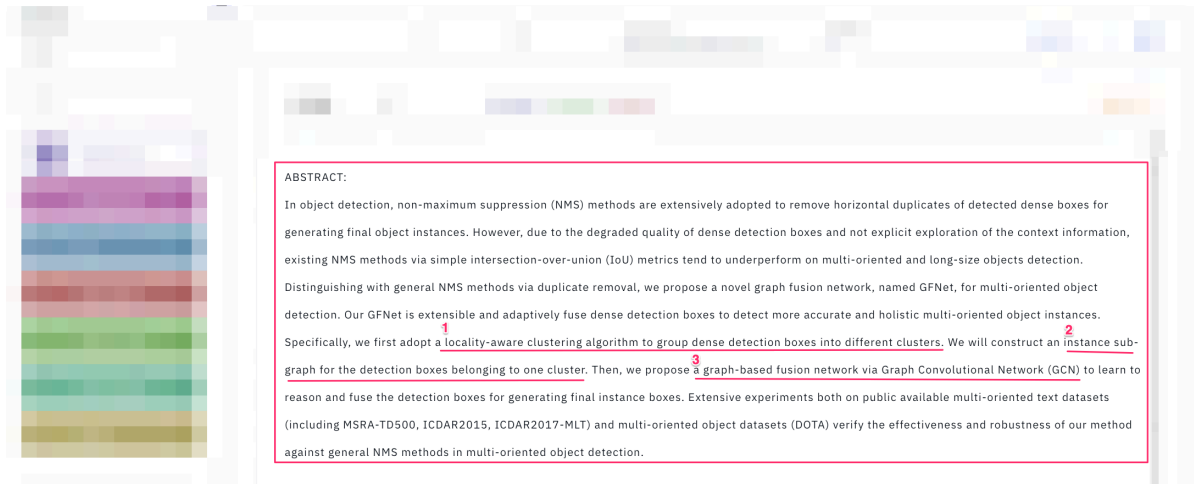e to purchase given her order list? Which users are buying fake followers to increase their public reputation? Myriads of new graph mining algorithms are proposed every year to answer such questions - each with a distinct problem formulation, computational time, and memory footprint. This lack of unity makes it difficult for a practitioner to compare different algorithms and pick the most suitable one for a specific application. These challenges - even more severe for non-experts - create a gap in which state-of-the-art techniques developed in academic settings fail to be optimally deployed in real-world applications. To bridge this gap, we propose AUTOGM, an automated system for graph mining algorithm development. We first define a unified framework UNIFIEDGM that integrates various message-passing based graph algorithms, ranging from conventional algorithms like PageRank to graph neural networks. Then UNIFIEDGM defines a search space in which five parameters are required to determine a graph algorithm. Under this search space, AUTOGM explicitly optimizes for the optimal parameter set of UNIFIEDGM using Bayesian Optimization. AUTOGM defines a novel budget-aware objective function for the optimization to incorporate a practical issue - finding the best speed-accuracy trade-off under a computation budget - into the graph algorithm generation problem. Experiments on real-world benchmark datasets demonstrate that AUTOGM generates novel graph mining algorithms with the best speed/accuracy trade-off compared to existing models with heuristic parameters.

In the above examples, the authors integrate various graph-mining algorithms, but they are unknown, and it's impossible to annotate them as combination elements.

▼ Example (distinct elements aren't necessarily combination elements)

ABSTRACT:

In object detection, non-maximum suppression (NMS) methods are extensively adopted to remove horizontal duplicates of detected dense boxes for generating final object instances. However, due to the degraded quality of dense detection boxes and not explicit exploration of the context information, existing NMS methods via simple intersection-over-union (IoU) metrics tend to underperform on multi-oriented and long-size objects detection. Distinguishing with general NMS methods via duplicate removal, we propose a novel graph fusion network, named GFNet, for multi-oriented object detection. Our GFNet is extensible and adaptively fuse dense detection boxes to detect more accurate and holistic multi-oriented object instances. Specifically, we first adopt a locality-aware clustering algorithm to group dense detection boxes into different clusters. We will construct an instance sub-graph for the detection boxes belonging to one cluster. Then, we propose a graph-based fusion network via Graph Convolutional Network (GCN) to learn to reason and fuse the detection boxes for generating final instance boxes. Extensive experiments both on public available multi-oriented text datasets (including MSRA-TD500, ICDAR2015, ICDAR2017-MLT) and multi-oriented object datasets (DOTA) verify the effectiveness and robustness of our method against general NMS methods in multi-oriented object detection.

The new object detection approach is constructed from a few distinct components, but they are listed as sequential phases in a pipeline, and not working together in synergy. They don't create recombination.

▼ Example (no combination elements)

ABSTRACT:

Measuring a document's complexity level is an open challenge, particularly when one is working on a diverse corpus of documents rather than comparing several documents on a similar topic or working on a language other than English. In this paper, we define a methodology to measure the complexity of French documents, using a new general and diversified corpus of texts, the "French Canadian complexity level corpus", and a wide range of metrics. We compare different learning algorithms to this task and contrast their performances and their observations on which characteristics of the texts are more significant to their complexity. Our results show that our methodology gives a general-purpose measurement of text complexity in French.

PARAGRAPH-0:

This paper explores the design and training of a general-purpose text complexity measuring algorithm in French. To achieve this goal, we pick out a variety of metrics that the literature identified as significant to complexity and train multiple learning algorithms to combine them to represent the general, non-application-specific complexity of a wide range of documents. We focus specifically on French, using our own French document corpus and metrics specialized to the French language. Our methodology can be applied to other languages as well by using a different corpus and metrics.

Nothing in the text suggests recombination.

analogy-src & analogy-target

- The analogy source is the domain/idea/concept we derive inspiration from.

- The analogy target is the domain/idea/concept benefiting from this inspiration.

- The analogy should be part of the core idea of the paper.

- Avoid analogy annotation when the source and target are too ambiguous.

▼ Example (unclear source and target)

In Section 2, we start from recalling Dyson's original construction [CITE_0]. Then, we propose an alternative description, where the fundamental object is the characteristic polynomial. We show the advantages of such description, borrowing heavily from the analogies to the simplest model of turbulence, i.e., the so-called Burgers equation. We also briefly mention, how the seminal results for the Gaussian Unitary Ensemble (GUE) can be recovered from a Burgers-like description.

It's simple to notice that "Burgers equation" is a good candidate for an analogy-src entity, however, there's not enough information in the text to infer a meaningful description of the analogy target. In this case, we should avoid analogy annotation overall.

> ✍🏼 **Tip**: Try replacing **<SRC>** and **<TRGT>** in the following template with the analogy source and target you've annotated:
> "The authors took inspiration from
> **<SRC>** for **<TRGT>**"

Relation Types

combination

> ✍🏼 **(comb-element-1, comb-element-2, ...comb-element-n,)**

- Connects the elements combined together.

analogy

> ✍🏼 **(analogy-src, analogy-target)**

- Connects an analogy source to the corresponding analogy target.

Document class

Classes applied to the whole document.

relevant

A document is relevant if one of the following holds:

1. It discusses a combination. You should be able to annotate a combination relation.

2. It discusses an analogy. You should be able to annotate an analogy relation.

bad-data

The data is badly parsed. Common issues:

- Visible LaTeX - all math is replaced by `MATH-i` placeholders when preprocessing the data for annotation. When LaTeX shows in the text is a sign the entry wasn't processed correctly.

We prove that for a weakly mixing algebraic action $\sigma: G\curvearrowright(X,\nu)$ [MATH_0]n[MATH_1]H^n(G\curvearrowright X; \mathbb{T})[MATH_2]H^n(G,\mathbb{T})\$, contains [MATH_3] as a natural subgroup for [MATH_4]. If we further assume the diagonal actions [MATH_5] are [MATH_6]-cocycle superrigid and [MATH_7] is torsion free as an abelian group, then the above also holds true for [MATH_8]. Applying it for principal algebraic actions when [MATH_9], we show that [MATH_10] is torsion free as an abelian group when [MATH_11] has property (T) as a direct corollary of Sorin Popa's cocycle superrigidity theorem; we also use it (when [MATH_12]) to answer, negatively, a question of Sorin Popa on the 2nd cohomology group of Bernoulli shift actions of property (T) groups.

Example of visible LaTeX in a document.

- Chopped off sentences.
- Anything else out of the ordinary

> ✍🏻 **Please use this class in moderation.** Preferably only when the issue affects the annotated spans.

irrelevant

The document doesn't fall under the previous two categories.