Faceted Relevances - Annotator Guidelines

Sheshera Mysore and Timothy O'Gorman

January 26, 2020.
Updated: September 25, 2021

1 Background

1.1 Broader goal of the project

Given the ever increasing scientific literature, tools that help search and navigate this increasing amount of literature are of increasing importance. Expert users of these tools demand tools which are able to search beyond simple ad-hoc retrieval paradigms where experts may wish to explore documents through an enriched semantic representation of the documents or may not know exactly what they might be looking for. In this work we concern ourselves with models which are able to retrieve scientific papers analogous to a query scientific paper along specifically chosen rhetorical structure elements. These rhetorical structure elements, which we refer to as *facets*, indicate aspects of a scientific paper such as the "background", "method", "results" or other facets. Our model for retrieval builds vector representations for the paper disentangled along these facets and uses these representations for retrieval.

1.2 Goals of the annotated dataset and terms

The dataset being annotated represents a *test-collection* which will be used for evaluation of the faceted retrieval models described in Section 1.1. Our evaluation time task will consist of a *query abstract* and a *facet* which will be used to indicate the sentences in the abstract being used as the query. Models will rank a set of *candidate abstracts* with respect to the sentences in the query abstract. The annotation task will involve annotators indicating the similarity of a query and candidate abstract along the specified facet on a graded similarity scale (elaborated on in Section 2).

For this project a *facet* for a research paper is the aspect of a paper corresponding to the typical steps involved in carrying out scientific research. Such as, the identification of a research problem/question, formation of hypothesis, testing of the hypothesis, and formation of conclusions. A specific instance of these facets applied in our case is described in Section 1.3. We will assume that the abstract of papers will contain a sentences corresponding to most of these facets. Next, we describe the specific instances of facets we work with.

1.3 Facet Definitions

Our facets definitions have a bent toward methodological NLP papers and are modified from the labelled dataset of Cohan et al. (2019). These facets are:

`Background/Objective`: Most often sets up the motivation for the work, states how it relates to prior work and states what the problem or research question being asked is.

`Method`: Describes the method being proposed or adopted in the paper. The method could be described at a very high level or it might be specified at a very fine grained level depending on the type of paper. Note that our definitions of methods are broad and will include methods of analysis of a phenomena, a model, data, or procedural descriptions of the experiments carried out. The specific interpretation of method will also depend on the type of paper (paper types are expanded upon in Section 1.4).

`Result`: This may be a detailed statement of the findings of analysis, a statement of results or a concluding set of statements based on the kind of paper.

The facets were predicted using the model of Cohan et al. (2019) into the set of labels: {`background`, `objective`, `method`, `result`, `other`}. Prior to relevance annotation, facets `objective` and `background` are merged into one facet called `background`. The `other` facet isn't considered for annotation. Further, incorrect facet labels for the query abstracts will be manually corrected.

1.4 Selection of data being annotated

Following established practice in the information retrieval community our dataset consists of query abstracts for which a *pool* of candidate abstracts have been generated using a range of methods (Soboroff, 2017).

Candidate pools: Our set of candidates per query are drawn from the following pool of methods: TF-IDF, averaged `word2vec` embeddings, and TF-IDF weighted `word2vec` based similarities run on titles, and abstracts giving us a set total of 6 methods. Further, a state of the art BERT based model, SPECTER, trained for scientific paper representation was also part the set of methods used to generate our pool (Cohan et al., 2020). Finally, papers cited in the query paper are also added to the pool. The set of methods is meant to represent a diverse range of similarities, ranging from term-overlap to dense embedding based similarities at the word to whole paragraph level and our analysis reveals that each of the methods retrieve largely different kinds of candidate papers.

Query abstracts: The set of query abstracts have been manually selected from the set of papers appearing in the ACL Anthalogy[1]. The query papers are picked based on a mix of random and curatorial criteria based on the following process:

1. Randomly select 20 papers from the approximately 5 year periods of 1994-1999, 2000-2004, 2005-2009, 2010-2014, and 2015-2019. Yielding a total of 100 papers.
2. Classify the 100 papers into the following broad class of paper types based on the call for papers in the ACL 2015 conference[2]: "Application/Tool", "Empirical/data-driven approaches", "Resources and evaluation", "Theoretical", and "Survey Papers". Given our facet definitions we exclude "Survey Papers" from the query papers.
3. Rate the papers on a scale of 1-5 based on the following criteria: contain substantial content in atleast one facet, meaning of abstract does not rely entirely on a paper being cited in the abstract, fits the description of atleast one of the paper types listed above.
4. Following this, query papers rated 3 or higher are selected to be approximately uniformly distributed across years, paper types, and facets.

2 Relevance Annotation Guidelines

While the crux of the similarity guidelines are meant to be facet dependent a more overarching guidelines will also apply, these are described next.

2.1 General Guidelines

1. **Focus on the query facet:** Focus on marking the similarity between the querying facet alone rather than the similarity of the entire abstract. In making annotations, a good rule of thumb is to look at the highlighted query facet and attempt to mark similarity and look at the rest of the abstract if you think it might contain important contextual information based on a glance at the rest of abstract or the title for example.

[1] https://www.aclweb.org/anthology/

[2] https://mirror.aclweb.org/acl2015/call_for_papers.html

2. **Importance of broader context depends on the facet:** The strictness of the above guideline will depend on the facet being annotated. In order, it will be important to consider the broader context for the `result`, then the `method` and then the `background` facet.
3. **Multiple salient aspects in a facet**: Sentences of an abstract marked with the same facet will often speak of slightly different aspects, or abstracts will speak of multiple distinct "methods" both of which would fairly be called a "method" (or any other facet). For example, slightly different aspects might look like one where a paper which speaks of two aspects of the result, time complexity of a method and accuracy performance of the method. Or, Ex 5 speaks of the method of dataset construction and of the reinforcement learning agent trained on the collected dataset. In marking similarities attempt to spot all such aspects and mark similarities if any one of the aspects overlaps between a candidate and query. Largely, however, we leave the determination of these to annotators. But note that in initial annotations these are the sources for difference in annotators ratings.
4. **Look terms up if needed:** In judging your similarity only rely on the text in the abstracts but if you think there are important terms in the abstract that you are not aware of but are crucial to understanding the abstract then feel free to look the terms up.
5. **Mentally correct incorrect facet labels in candidate abstracts**: The facet labels on sentences in the candidates and query abstracts are model predicted, this will mean they have errors. While we will strive to correct egregious mistakes in the query documents in your relevance rating mentally correct wrong labels and judge candidate facet sentence similarity.
6. **Do not collaborate with other annotators**: The annotation process will include meetings between batches of annotations for all the annotators to align on definitions of similarity and to talk about outlier cases. Do not collaborate on the annotations outside of these periods.
7. **Don't overthink the similarity**: While we would like to encourage you to have well thought judgements of similarity if you find yourself spending many minutes judging similarities for nearly all the papers you might be over thinking the judgements. While some papers will invariably take longer clearly irrelevant ones will take a few seconds. However if you find any set of papers too difficult to understand bring it up with us.

2.2 Background/Objective Similarity

Note that the following facet dependent criteria will only list the three grades of relevant papers. Papers which don't fit these criteria should in all likelihood be marked "Irrelevant (0)".

2.2.1 Near-identical background/objective (3)

"Near-identical" implies that both papers are trying to accomplish the same specific goal, are solving the same specific modeling problem or the same machine learning task, or are motivated in a specific similar manner.

E.g: Both papers below focus on the goal of creating a natural language interface for querying of tabular data.

(1) a. **Query**: *"Naturalizing a Programming Language via Interactive Learning", Our goal is to create a convenient natural language interface for performing well-specified but complex actions such as analyzing data, manipulating text, and querying databases. However, existing natural language interfaces for such tasks are quite primitive compared to the power one wields with a programming language.*

b. **Candidate**: *"NLyze: interactive programming by natural language for spreadsheet* *data analysis and manipulation” Millions of computer end users need to perform tasks over tabular spreadsheet data, yet lack the programming knowledge to do such tasks automatically. This paper describes the design and implementation of a robust natural language based interface to spreadsheet programming. Our methodology involves designing a typed domain-specific language (DSL) that supports an expressive algebra of map, filter, reduce, join, and formatting capabilities at a level of abstraction appropriate for non-expert users.*

2.2.2 Similar background/objective (2)

“Similar” implies that papers have a similar goal or are motivated by similar underlying theoretical/prior background on a higher level without too much overlap of the specifics of the goal or motivation.

E.g. Both papers have a similar goal, enabling the use of natural-language for programming/commands. But they differ in the specifics, while one modifies a programming language the other restricts a natural language. Further, while the query intends to do this for tabular data the candidate does not mention that specific goal.

(2) a. **Query**: *“Naturalizing a Programming Language via Interactive Learning”, Our goal is to create a convenient natural language interface for performing well-specified but complex actions such as analyzing data, manipulating text, and querying databases. However, existing natural language interfaces for such tasks are quite primitive compared to the power one wields with a programming language.*

b. **Candidate**: *“Controlled English to Facilitate Human/machine Analytical Processing”, Controlled English is a human-readable information representation format that is implemented using a restricted subset of the English language, but which is unambiguous and directly accessible by simple machine processes. We have been researching the capabilities of CE in a number of contexts, and exploring the degree to which a flexible and more human-friendly information representation format could aid the intelligence analyst in a multi-agent collaborative operational environment; especially in cases where the agents are a mixture of other human users and machine processes aimed at assisting the human users. CE itself is built upon a formal logic basis, but allows users to easily specify models for a domain of interest in a human-friendly language.*

2.2.3 Related background/objective (1)

“Related” implies that the goal or motivation are similar at a very high level with close to no overlaps of the specifics of the goal or motivation. Or the papers may be tackling different sub-problems of a larger very specific problem.

E.g. for the example below, the two problems or goals aren’t identical, but are related at a very high level – while one has the goal of modifying natural language for enacting data manipulation and querying, the other proposes a programming language intended for easing the expression of program synthesis problems. As with other “related” categories, this can be thorny and up for interpretation, and we want to just focus on getting calibrated.

(3) a. **Query**: *Naturalizing a Programming Language via Interactive Learning Our goal is to create a convenient natural language interface for performing well-specified but complex actions such as analyzing data, manipulating text, and querying databases. However, existing natural language interfaces for such tasks are quite primitive compared to the power one wields with a programming language.*

b. **Candidate**: *TerpreT: A Probabilistic Programming Language for Program Induction We study machine learning formulations of inductive program synthesis; given* *input-output examples, we try to synthesize source code that maps inputs to corresponding outputs. Our aims are to develop new machine learning approaches based on neural networks and graphical models, and to understand the capabilities of machine learning techniques relative to traditional alternatives, such as those based on constraint solving from the programming languages community. Our key contribution is the proposal of TerpreT, a domain-specific language for expressing program synthesis problems.*

2.3 Method Similarity

In the case of methods it will be important to focus on mechanistic similarities between query and candidate methods. Since methods often have multiple steps, actions or components within a procedure, similarity ratings should consider both any overarching patterns in how a sequence of steps manipulates a set of objects *and* similarity between objects involved in individual steps. Note that while the other facets (such as background) might be useful in determining similarity in the details of the method, the dominant driver of your similarity judgement must be the mechanistic aspect of the methods. For example, two papers proposing models for on sentiment classification will likely be similar in the details of their inputs and outputs but the mechanistic similarity between the methods might be significantly different.

2.3.1 Near-Identical method (3)

"Near-Identical" implies methods described share a similar over-arching mechanistic similarity, further the methods must also be similar in terms of the details of the objects being manipulated.

E.g. While both the below methods are motivated differently (efficiency vs preventing error propagation) and use different classes of dependency parsing methods (transition based vs graph based), both the below methods are similar in first generating an un-directed parse and then converting it to a directed parse.

(4) a. **Query**: *"Edge-Linear First-Order Dependency Parsing with Undirected Minimum Spanning Tree Inference", ... We propose such an inference algorithm for first-order models, which encodes the problem as a minimum spanning tree (MST) problem in an undirected graph. This allows us to utilize state-of-the-art undirected MST algorithms whose run time is O(m) at expectation and with a very high probability. A directed parse tree is then inferred from the undirected MST and is subsequently improved with respect to the directed parsing model through local greedy updates, both steps running in O(n) time...*

b. **Candidate**: *"Dependency Parsing with Undirected Graphs", We introduce a new approach to transition based dependency parsing in which the parser does not directly construct a dependency structure, but rather an undirected graph, which is then converted into a directed dependency tree in a post-processing step. This alleviates error propagation, since undirected parsers do not need to observe the single-head constraint. Undirected parsers can be obtained by simplifying existing transition-based parsers satisfying certain conditions. We apply this approach to obtain undirected variants of the planar and 2-planar parsers and of Covington's non-projective parser.*

Here the query described two methods, one for creating a dataset and the other describing an RL agent. The dataset collection process follows a very similar structure in being a chat between two individuals grounded in a task related to a visual object.

(5) a. **Query**: *"The BURCHAK corpus: a Challenge Data Set for Interactive Learning of Visually Grounded Word Meanings", We motivate and describe a new freely available human-human dialogue dataset for interactive learning of visually grounded word meanings through ostensive definition by a tutor to a learner.* ***The data has been collected using a novel, character-by-character variant of the DiET chat tool (Healey et al., 2003; Mills and Healey, submitted) with a novel task, where a Learner needs to learn invented visual attribute words (such as"burchak"for square) from a tutor.*** *As such, the text-based interactions closely resemble face-to-face conversation and thus contain many of the linguistic phenomena encountered in natural, spontaneous dialogue. These include self-and other-correction, mid-sentence continuations, interruptions, overlaps, fillers, and hedges. We also present a generic n-gram framework for building user (i.e. tutor) simulations from this type of incremental data, which is freely available to researchers. We show that the simulations produce outputs that are similar to the original data (e.g. 78% turn match similarity). Finally, we train and evaluate a Reinforcement Learning dialogue control agent for learning visually grounded word meanings, trained from the BURCHAK corpus.*

b. **Candidate**: *"The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue", This paper introduces the PhotoBook dataset, a large-scale collection of visually-grounded, task-oriented dialogues in English designed to investigate shared dialogue history accumulating during conversation.* ***Taking inspiration from seminal work on dialogue analysis, we propose a data-collection task formulated as a collaborative game prompting two online participants to refer to images utilising both their visual context as well as previously established referring expressions****. We provide a detailed description of the task setup and a thorough analysis of the 2,500 dialogues collected. To further illustrate the novel features of the dataset, we propose a baseline model for reference resolution which uses a simple method to take into account shared information accumulated in a reference chain.*

2.3.2 Similar method (2)

"Similar" implies that the methods are mechanistically similar and the details are only comparable between the query and candidate (Example 6). A stronger notion might be where methods are mechanistically similar but and the details would make the problem being solved similar but only in a very general sense (Example 7).

E.g. Both the methods below collect datasets using a chat tool and share the larger similarity of both being grounded in a visual task and but the details of the visual tasks are only mildly similar with one describing visual attribute words and the other involving navigating a visual environment with the goal of finding the other player.

(6) a. **Query**: *"The BURCHAK corpus: a Challenge Data Set for Interactive Learning of Visually Grounded Word Meanings", We motivate and describe a new freely available human-human dialogue dataset for interactive learning of visually grounded word meanings through ostensive definition by a tutor to a learner.* ***The data has been collected using a novel, character-by-character variant of the DiET chat tool (Healey et al., 2003; Mills and Healey, submitted) with a novel task, where a Learner needs to learn invented visual attribute words (such as"burchak"for square) from a tutor.*** *As such, the text-based interactions closely resemble face-to-face conversation and thus contain many of the linguistic phenomena encountered in natural, spontaneous dialogue. These include self-and other-correction, mid-sentence continuations, interruptions, overlaps, fillers, and hedges. We also present a generic n-gram framework for building*

user (i.e. tutor) simulations from this type of incremental data, which is freely available to researchers. We show that the simulations produce outputs that are similar to the original data (e.g. 78% turn match similarity). Finally, we train and evaluate a Reinforcement Learning dialogue control agent for learning visually grounded word meanings, trained from the BURCHAK corpus.

b. **Candidate**: *Building computer systems that can converse about their visual environment is one of the oldest concerns of research in Artificial Intelligence and Computational Linguistics (see, for example, Winograd's 1972 SHRDLU system). Only recently, however, have methods from computer vision and natural language processing become powerful enough to make this vision seem more attainable. Pushed especially by developments in computer vision, many data sets and collection environments have recently been published that bring together verbal interaction and visual processing. Here, we argue that these datasets tend to oversimplify the dialogue part, and we propose a task—MeetUp!—that requires both visual and conversational grounding, and that makes stronger demands on representations of the discourse.* ***MeetUp! is a two-player coordination game where players move in a visual environment, with the objective of finding each other. To do so, they must talk about what they see, and achieve mutual understanding....***

Both of the below methods might be seen as using a "bootstrap" method which involves starting with an initial seed set of training instances and then expanding these with a set of automatic methods. Note, however, the differences in the final intents of both the papers (analyzing the patterns vs improving final performance), the difference in the specifics of the task (fact/emotional vs sarcastic/nasty utterences), and the nature of initial supervision.

(7) a. **Query**: *"And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue", We investigate the characteristics of factual and emotional argumentation styles observed in online debates. Using an annotated set of"factual"and"feeling"debate forum posts, we extract patterns that are highly correlated with factual and emotional arguments, and then apply a bootstrapping methodology to find new patterns in a larger pool of unannotated forum posts. This process automatically produces a large set of patterns representing linguistic expressions that are highly correlated with factual and emotional language. Finally, we analyze the most discriminating patterns to better understand the defining characteristics of factual and emotional arguments.*

b. **Candidate**: *"Identifying Subjective and Figurative Language in Online Dialogue", More and more of the information on the web is dialogic, from Facebook newsfeeds, to forum conversations, to comment threads on news articles. In contrast to traditional, monologic resources such as news, highly social dialogue is very frequent in social media. We aim to automatically identify sarcastic and nasty utterances in unannotated online dialogue, extending a bootstrapping method previously applied to the classification of monologic subjective sentences in Riloff and Weibe 2003. We have adapted the method to fit the sarcastic and nasty dialogic domain. Our method is as follows: 1) Explore methods for identifying sarcastic and nasty cue words and phrases in dialogues; 2) Use the learned cues to train a sarcastic (nasty) Cue-Based Classifier; 3) Learn general syntactic extraction patterns from the sarcastic (nasty) utterances and define fine-tuned sarcastic patterns to create a Pattern-Based Classifier; 4) Combine both Cue-Based and fine-tuned Pattern-Based Classifiers to maximize precision at the expense of recall and test on unannotated utterances.*

2.3.3 Related method (1)

“Relevant” is meant to encompass a wide range in being similar and can be hard to list at length. However, common cases include:

1. Cases where the details of the two methods are similar but there is only high level mechanistic similarity. (Example 9b)
2. Methods are similar mechanistically at a very high level without too much similarity of details. In doing this, consider both cases where the query and candidate are working on similar problems or have similar goals (Example 8b), and the case where they work on different problems. The important thing here is to get a shared sense (based on examples and annotator discussions) of how general we can be about the notion of “very high level”.
3. If small or not-so-important mechanistic parts of the methods in two papers are similar. Note that a mere mention of a method will not merit any similarity, for example, in the case where a candidate paper analyzes a the model described in the query or vise versa.
4. Cases where query and candidate abstracts may vary in the level of granularity in which they describe a method and a high level similarity is the only one you can establish by reading the abstract. (Example 8c)

(8) a. **Query**: *“And That’s A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue”, We investigate the characteristics of factual and emotional argumentation styles observed in online debates.* ***Using an annotated set of “factual” and “feeling” debate forum posts, we extract patterns that are highly correlated with factual and emotional arguments, and then apply a bootstrapping methodology to find new patterns in a larger pool of unannotated forum posts. This process automatically produces a large set of patterns representing linguistic expressions that are highly correlated with factual and emotional language. Finally, we analyze the most discriminating patterns to better understand the defining characteristics of factual and emotional arguments.***

b. **Candidate**: *“Visual Analysis of Conflicting Opinions”, Understanding the nature and dynamics of conflicting opinions is a profound and challenging issue. In this paper we address several aspects of the issue through a study of more than 3,000 Amazon customer reviews of the controversial bestseller The Da Vinci Code, including 1,738 positive and 918 negative reviews. The study is motivated by critical questions such as: what are the differences between positive and negative reviews? ...* ***We first analyze terminology variations in these reviews in terms of syntactic, semantic, and statistic associations identified by TermWatch and use term variation patterns to depict underlying topics. We then select the most predictive terms based on log likelihood tests and demonstrate that this small set of terms classifies over 70% of the conflicting reviews correctly. This feature selection process reduces the dimensionality of the feature space from more than 20,000 dimensions to a couple of hundreds. We utilize automatically generated decision trees to facilitate the understanding of conflicting opinions in terms of these highly predictive terms. This study also uses a number of visualization and modeling tools to identify not only what positive and negative reviews have in common, but also they differ and evolve over time.***

c. **Candidate**: *“Identifying High-Level Organizational Elements in Argumentative Discourse”, Argumentative discourse contains not only language expressing claims and evidence, but also language used to organize these claims and pieces of evidence. Differentiating between the two may be useful for many applications, such as those that focus on the content (e.g., relation extraction) of arguments and those that* *focus on the structure of arguments (e.g., automated essay scoring).* ***We propose an automated approach to detecting high-level organizational elements in argumentative discourse that combines a rule-based system and a probabilistic sequence model in a principled manner.*** *We present quantitative results on a dataset of human-annotated persuasive essays, and qualitative analyses of performance on essays and on political debates.*

(9) a. **Query**: *"The BURCHAK corpus: a Challenge Data Set for Interactive Learning of Visually Grounded Word Meanings", We motivate and describe a new freely available human-human dialogue dataset for interactive learning of visually grounded word meanings through ostensive definition by a tutor to a learner.* ***The data has been collected using a novel, character-by-character variant of the DiET chat tool (Healey et al., 2003; Mills and Healey, submitted) with a novel task, where a Learner needs to learn invented visual attribute words (such as"burchak"for square) from a tutor.*** *As such, the text-based interactions closely resemble face-to-face conversation and thus contain many of the linguistic phenomena encountered in natural, spontaneous dialogue. These include self-and other-correction, mid-sentence continuations, interruptions, overlaps, fillers, and hedges. We also present a generic n-gram framework for building user (i.e. tutor) simulations from this type of incremental data, which is freely available to researchers. We show that the simulations produce outputs that are similar to the original data (e.g. 78% turn match similarity). Finally, we train and evaluate a Reinforcement Learning dialogue control agent for learning visually grounded word meanings, trained from the BURCHAK corpus.*

b. **Candidate**: *"SWITCHBOARD: telephone speech corpus for research and development", SWITCHBOARD is a large multispeaker corpus of conversational speech and text which should be of interest to researchers in speaker authentication and large vocabulary speech recognition.* ***About 2500 conversations by 500 speakers from around the US were collected automatically over T1 lines at Texas Instruments.*** *Designed for training and testing of a variety of speech processing algorithms, especially in speaker verification, it has over an 1 h of speech from each of 50 speakers, and several minutes each from hundreds of others. A time-aligned word for word transcription accompanies each recording.*

2.4 Result Similarity

In the case of result similarities it will often be important to consider the remainder of the abstract to determine similarity.

2.4.1 Near-Identical Results (3)

"Near-Identical" implies the same finding or conclusion, this can take one of two forms depending on the nature of the paper –

1. Papers demonstrate a specific observation, or prove or disprove a hypothesis (X made system more efficient, annotation study found Y, system/model component X behaves like Y, X causes Y).
2. Evaluate on a similar task (e.g. for ML methodological papers).

Mark as "Near-Identical" even if findings are alike but opposite to one another (Q: X helps predict Y, C: X does not help predict Y or vise versa).

Example 10b demonstrates the ability of multiple embeddings to be helpful to find similar words in a sentential context which is very similar to a lexical substitution task. On the other hand Example 10c presents state of the art on the lexical substitution task.

(10) a. **Query**: *"Learning Topic-Sensitive Word Representations", Distributed word representations are widely used for modeling words in NLP tasks. Most of the existing models generate one representation per word and do not consider different meanings of a word. We present two approaches to learn multiple topic-sensitive representations per word by using Hierarchical Dirichlet Process. We observe that by modeling topics and integrating topic distributions for each document we obtain representations that are able to distinguish between different meanings of a given word.* ***Our models yield statistically significant improvements for the lexical substitution task indicating that commonly used single word representations, even when combined with contextual information, are insufficient for this task.***

b. **Candidate**: *"Multi-Prototype Vector-Space Models of Word Meaning", Current vector-space models of lexical semantics create a single "prototype" vector to represent the meaning of a word. However, due to lexical ambiguity, encoding word meaning with a single vector is problematic. This paper presents a method that uses clustering to produce multiple "sense-specific" vectors for each word. This approach provides a context-dependent vector representation of word meaning that naturally accommodates homonymy and polysemy.* ***Experimental comparisons to human judgements of semantic similarity for both isolated words as well as words in sentential contexts demonstrate the superiority of this approach over both prototype and exemplar based vector-space models.***

c. **Candidate**: *"Learning to Rank Lexical Substitutions", The problem to replace a word with a synonym that fits well in its sentential context is known as the lexical substitution task. In this paper, we tackle this task as a supervised ranking problem. Given a dataset of target words, their sentential contexts and the potential substitutions for the target words, the goal is to train a model that accurately ranks the candidate substitutions based on their contextual fitness. As a key contribution, we customize and evaluate several learning-to-rank models to the lexical substitution task, including classification-based and regression-based approaches.* ***On two datasets widely used for lexical substitution, our best models significantly advance the state-of-the-art.***

2.4.2 Similar Results (2)

"Similar" implies the results are generally similar but lack similarity in the specifics or show partial overlaps of a larger salient finding.

In example 11 one states that style is a stronger indicator of community identity the other makes a statement about long term members developing forum specific jargon – similar findings if one applied an inference that long term members feel a stronger sense of community identity.

(11) a. **Query**: *"Characterizing the Language of Online Communities and its Relation to Community Reception", ... Experiments with several Reddit forums show that style is a better indicator of community identity than topic, even for communities organized around specific topics. Further, there is a positive correlation between the community reception to a contribution and the style similarity to that community, but not so for topic similarity.*

b. **Candidate**: *"Language use as a reflection of socialization in online communities", In this paper we investigate the connection between language and community membership of long time community participants through computational modeling techniques. We report on findings from an analysis of language usage within a popular online discussion forum with participation of thousands of users spanning multiple years.* ***We find community norms of long time participants that***

are characterized by forum specific jargon and a style that is highly informal and shows familiarity with specific other participants and high emotional involvement in the discussion. We also find quantitative evidence of persistent shifts in language usage towards these norms across users over the course of the first year of community participation. Our observed patterns suggests language stabilization after 8 or 9 months of participation.

2.4.3 Related Results (1)

"Related" results cases will also involve several kinds of similarity. This includes the set of cases where results are similar in very coarse or under-specified ways, the broader contexts of the results indicated in the backgrounds and methods are only coarsely related, and the results are similar only in non-salient ways.

In example 12, the findings are similar only in a very vague sense given the high level similarity of the word similarity and lexical substitution tasks, further the methods lack any specific sense of similarity in the background and method facets which might have helped draw more similar conclusions.

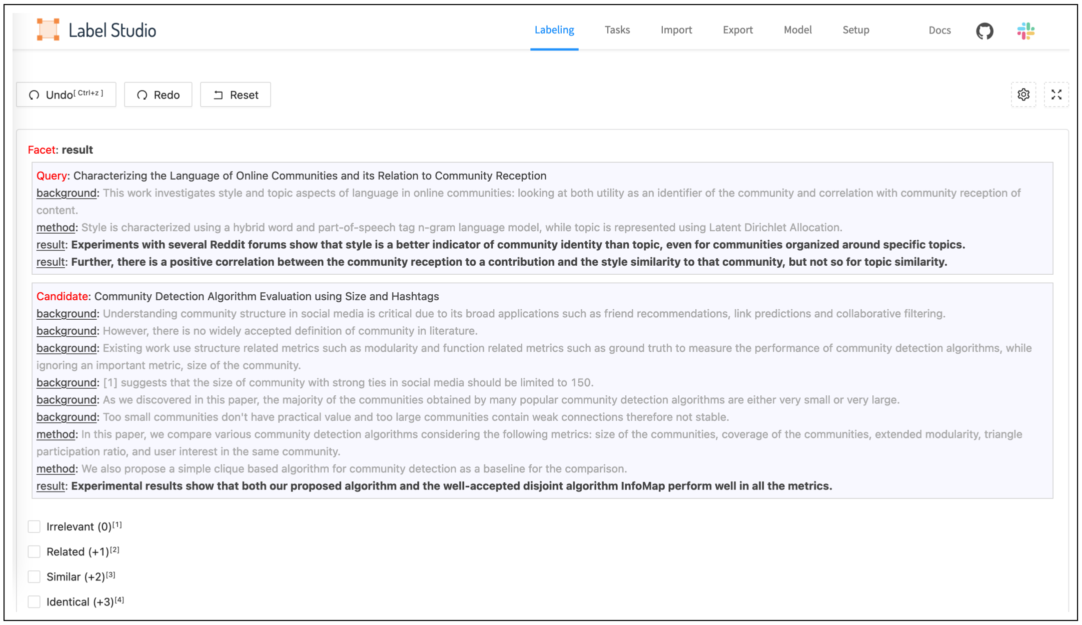
(12) a. **Query**: *"Learning Topic-Sensitive Word Representations", ... Our models yield statistically significant improvements for the lexical substitution task indicating that commonly used single word representations, even when combined with contextual information, are insufficient for this task.*

b. **Candidate**: *"Learning Better Embeddings for Rare Words Using Distributional Representations", There are two main types of word representations: low-dimensional embeddings and high-dimensional distributional vectors, in which each dimension corresponds to a context word. In this paper, we initialize an embedding-learning model with distributional vectors. Evaluation on word similarity shows that this initialization significantly increases the quality of embeddings for rare words.*

In example 13, the results are related only so far as both of them speak about the relation of the topic and community engagement, potentially also presenting slightly contradictory results.

(13) a. **Query**: *"Characterizing the Language of Online Communities and its Relation to Community Reception", ... Experiments with several Reddit forums show that style is a better indicator of community identity than topic, even for communities organized around specific topics. Further, there is a positive correlation between the community reception to a contribution and the style similarity to that community, but not so for topic similarity.*

b. **Candidate**: *"Extracting Topics with Focused Communities for Social Content Recommendation", A thorough understanding of social media discussions and the demographics of the users involved in these discussions has become critical for many applications like business or political analysis. Such an understanding and its ramifications on the real world can be enabled through the automatic summarization of Social Media. Trending topics are offered as a high level content recommendation system where users are suggested to view related content if they deem the displayed topics interesting. However, identifying the characteristics of the users focused on each topic can boost the importance even for topics that might not be popular or bursty. We define a way to characterize groups of users that are focused in such topics and propose an efficient and accurate algorithm to extract such communities.* ***Through qualitative and quantitative experimentation we observe that topics with a strong community focus are interesting and more likely to catch the attention of users.***



Facet: **result**

Query: Characterizing the Language of Online Communities and its Relation to Community Reception
background: This work investigates style and topic aspects of language in online communities: looking at both utility as an identifier of the community and correlation with community reception of content.
method: Style is characterized using a hybrid word and part-of-speech tag n-gram language model, while topic is represented using Latent Dirichlet Allocation.
result: **Experiments with several Reddit forums show that style is a better indicator of community identity than topic, even for communities organized around specific topics.**
result: **Further, there is a positive correlation between the community reception to a contribution and the style similarity to that community, but not so for topic similarity.**

Candidate: Community Detection Algorithm Evaluation using Size and Hashtags
background: Understanding community structure in social media is critical due to its broad applications such as friend recommendations, link predictions and collaborative filtering.
background: However, there is no widely accepted definition of community in literature.
background: Existing work use structure related metrics such as modularity and function related metrics such as ground truth to measure the performance of community detection algorithms, while ignoring an important metric, size of the community.
background: [1] suggests that the size of community with strong ties in social media should be limited to 150.
background: As we discovered in this paper, the majority of the communities obtained by many popular community detection algorithms are either very small or very large.
background: Too small communities don't have practical value and too large communities contain weak connections therefore not stable.
method: In this paper, we compare various community detection algorithms considering the following metrics: size of the communities, coverage of the communities, extended modularity, triangle participation ratio, and user interest in the same community.
method: We also propose a simple clique based algorithm for community detection as a baseline for the comparison.
result: **Experimental results show that both our proposed algorithm and the well-accepted disjoint algorithm InfoMap perform well in all the metrics.**

☐ Irrelevant (0)[1]

☐ Related (+1)[2]

☐ Similar (+2)[3]

☐ Identical (+3)[4]

Figure 1: Example annotation task in the Labelstudio annotation interface. Tasks indicate, a target facet for annotation (here, result), a query abstract, a candidate abstract, and a set of sentences indicative of the target facet in the query and candidate. Relevance judgements must be made while considering the context provided in the remainder of the abstracts and the target facet sentences in accordance with the guidelines outlined in this document.

References

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *ACL*.

Ian Soboroff. 2017. Building test collections: An interactive guide for students and others without their own evaluation conference series. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1407–1410.