Subjective Attributes in Conversational Recommendation Systems: Challenges and Opportunities

Filip Radlinski,[1] **Craig Boutilier,**[1] **Deepak Ramachandran**[1] and **Ivan Vendrov**[2*]
[1] Google Research
[2] Omni Labs
{filiprad,cboutilier,ramachandrand}@google.com, ivan@omnilabs.ai

Abstract

The ubiquity of recommender systems has increased the need for higher-bandwidth, natural and efficient communication with users. This need is increasingly filled by recommenders that support natural language interaction, often conversationally. Given the inherent semantic subjectivity present in natural language, we argue that modeling *subjective attributes* in recommenders is a critical, yet understudied, avenue of AI research. We propose a novel framework for understanding different forms of subjectivity, examine various recommender tasks that will benefit from a systematic treatment of subjective attributes, and outline a number of research challenges.

Introduction

The use of descriptive item attributes or tags has been recognized for decades as useful for multiple recommendation tasks (Basu, Hirsh, and Cohen 1998; Zhang, Zhou, and Zhang 2011), allowing users to more easily express preferences, search criteria, and critiques (Zheng, Zhang, and Feng 2013; Pazzani and Billsus 2007). For instance, a user might critique a recommended camera by asking for one that is cheaper or has a more powerful zoom. However, the potential for ambiguity and disagreement across users in their usage of attributes (Lee and Yong 2007) requires methods to handle such disagreements. Most existing work in this area has been somewhat *ad hoc*; in particular, one important source of disagreement—the potential *subjectivity* of the semantics of an attribute—has yet to receive a systematic treatment. While language is a key field of AI research and has seen rapid development recently, the role of subjectivity in language has yet to receive commensurate attention.

By *subjectivity* we mean that different users may have different meanings in mind when they use an attribute or tag to describe an item. For instance, two users might interpret the term 'violent' differently as applied to movies: they may have different tolerances for the degree of violence, or be more sensitive to different forms (realistic vs. cartoonish, or physical vs. emotional vs. psychological). Critically, we distinguish subjectivity from noise, context-dependence, or other sources of disagreement. Roughly, an attribute is subjective if a significant number of users (or groups) *reliably* (w.r.t. noise) and *robustly* (accounting for context and other exogenous factors) disagree to which items the attribute applies.

We pose **the systematic, principled treatment of subjective attributes** as an important challenge for research in recommender systems, and AI more broadly. Indeed, the traditional use of attributes for search, critiquing and preference elicitation has generally avoided handling subjectivity, instead limiting interfaces to a prespecified attribute vocabulary with a fixed *extensional semantics*. More recent use of *conversational* technologies in recommenders allows users to communicate their intent or preferences more naturally and directly, yet attempts at natural interactions have been found to often fail (Grudin and Jacques 2019). Conversational recommenders call for a comprehensive treatment of attribute subjectivity—this is critical if one wants to allow people to describe items and preferences *in their own terms*, rather than shoehorning them into communicating using predefined vocabularies with rigid, hard-to-interpret semantics. This, in turn, will unlock natural, effective communication between users and recommenders.

We sketch a preliminary framework for studying subjective attributes, identifying various types of subjectivity, as well as factors that may influence subjective semantics. This sketch is not intended to be definitive, but rather to invite a structured treatment of subjectivity in future research. We also outline some concrete challenges facing RSs relating to subjectivity.

The Importance of Subjectivity

The use of *attributes* to allow users to navigate item space or express preferences is common in RSs, e.g., in *faceted search* (Zheng, Zhang, and Feng 2013) or *example critiquing* (Chen and Pu 2012). Traditionally, most systems limit users to a prespecified vocabulary of *hard* (or *catalog*) attributes (e.g., 'price', 'size', 'location'), where a definitive source of objective ground truth (the *catalog*) states which items possess which attribute values. When engaging with conversational systems using natural language, users often want to (and do) use broader terminology than supported by a rigid, incomplete catalog vocabulary. Thus we should expect more open-ended descriptions of preferences or requested items in rich language that may be ambiguous (Radlinski et al. 2019), e.g., a 'cheap' camera, a 'vibrant-colored' shirt, or a 'quiet' restaurant. The semantics of such *soft* attributes (Balog, Radlinski, and Karatzoglou 2021)—i.e., delineating

*Work done while at Google Research.

the items that exhibit that attribute—is challenging, generally requiring analysis of usage in tagging data (Gantner et al. 2010), open-ended reviews (McAuley and Leskovec 2013), or text/dialogue corpora (Radlinski et al. 2019). However, such approaches generally avoid the issue of *subjectivity*, namely, the fact that different users may have a different intent when using a given term to describe an item. For example, two users could differ on what it means for a camera to be 'cheap' if they have different budgets, or for a restaurant to qualify as 'quiet.' Understanding and accurately modeling a user's *subjective intent* when expressing such soft attributes is critical to making high-quality recommendations, requiring *personalized semantics*.

It is important to distinguish subjectivity from uncertainty, imprecision or ambiguity. The latter concepts imply that one has insufficient information to fully ground the terms being used. With subjective attributes, even with complete information, disagreement is to be expected. For example, semantic ambiguity due to imprecise terminology does not adequately explain why two people disagree on whether a given book plot is 'predictable.' Rather a model must account for why such terms are interpreted differently by different people. Morever, approaches based on, say, fuzzy sets (Cock, Bodenhofer, and Kerre 2000) are not rich enough to capture subjectivity. Some aspects of subjectivity are closely related to *polysemy* in linguistics, which is the capacity for a word or phrase to have multiple (sometimes contiguous) meanings. As linguists have recognized, a simplified model of semantics like line or ball semantics (discussed below) cannot capture the full complexity of linguistic phenomena where meaning can be transformed through independent processes such as metonymy or metaphor (Lakoff 1999). An example of this in recommender systems might be a situation where a user refers to a song as 'spicy' to communicate an emotional response through analogy rather than an ontological category.

As RSs become increasingly conversational, modeling subjectivity will take on added importance. At the same time, making recommendations is often a *small-data problem*: we usually have only a handful of potentially biased data points per user; user interests change continuously (Bernardi et al. 2015); and even the best Collaborative Filtering (CF) techniques often fail to reflect the true diversity and complexity of user preferences. Better understanding of user needs and preferences *as they are expressed* should greatly enhance recommendation quality and diversity. Current approaches for elicitation tend to focus on easily interpretable preferences; e.g., asking users about: the relevance or rating of specific items (Boutilier, Zemel, and Marlin 2003; Harper and Konstan 2015; Taijala, Willemsen, and Konstan 2018); categories of interest (Chang, Harper, and Terveen 2015); explicit pairwise comparisons (Christakopoulou, Radlinski, and Hofmann 2016); a choice from a list of items (Graus and Willemsen 2015); or yes/no category refinement questions (Zou, Chen, and Kanoulas 2020). Even approaches that nominally solicit "free-form" tags often bias users towards past labels (Harper and Konstan 2015; Vig et al. 2010). By contrast, conversational RSs aim to better understand users' needs and preferences using natural language, increasing communication bandwidth with users while reducing cognitive load. That said, nearly all conversational RSs assume that words/phrases mean the same thing to *all users* in *all contexts* at *all times*. This is often treated as a *grounding problem*, i.e., finding the *unique* mapping from strings to their meaning in the recommendation domain. This oversimplified view of language can increase the cognitive burden on users and reduces its utility. For instance, if a user states "I'd like to watch *something funny*," naïve systems may translate this to movies in the *comedy* genre (Habib, Zhang, and Balog 2020). Neural models may encode this as a single vector (Luo et al. 2020). They cannot, however, capture the fact that not all users consider the same movies funny, nor that an individual may use the term differently in different contexts (e.g., with friends vs. with children).

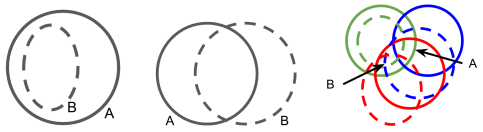
Conversational systems should be able to explicitly reason about subjectivity and exploit it to make more helpful recommendations by constructing higher fidelity user models that more accurately incorporate user feedback and preferences. By way of comparison, traditional web search has been found to fail for users who lack knowledge of the correct search terms (Aula, Khan, and Guan 2010); similarly, we argue that better handling of subjectivity will reduce recommendation failures due to users not knowing how to effectively express their preferences using a rigid vocabulary.

A proper understanding of subjectivity will play a role in many different *types* of recommenders. In faceted search, even "fixed-vocabulary" systems will be made more powerful by adopting user-specific interpretations of qualitative attributes like 'cheaper,' 'more colorful' or 'quieter.' Likewise, critiquing interfaces to traditional recommenders will benefit, when allowing users access to a broader set of attributes, from the ability to handle open-ended critiques more precisely for individual users. Tagging-based recommenders can personalize the retrieval and recommendation of items to a user's precise intent. With a proper, personalized grounding of subjective attributes, the attribute vocabulary used by interactive preference elicitation techniques will not only expand, but provide a more engaging experience by asking users questions in terms of the soft attributes over which they most naturally conceive of their preferences. Finally, treating subjectivity adequately is critical for conversational recommenders if they are to speak to users in their own terms.

Though we primarily focus on challenges and solutions for the recommendations field, our discussion of subjectivity is broadly applicable to many other fields of AI. Indeed, a broad definition of RSs includes many search tasks, with many direct analogs between search and RS operations. For instance, critiquing in recommendation tasks is strongly related to query refinement in search.

A Framework for Analyzing Subjectivity

We outline different forms of subjectivity in recommender settings, and briefly sketch some model formulations. We emphasize that this short treatment is merely suggestive, intended to spur future research, as opposed to limiting attention to any particular approach. We first identify three distinct forms of subjectivity, those pertaining to *degree*, *semantics* and *composition* (see Figure 1). While some attributes may



(a) Degree (b) Semantic (c) Compositional

Figure 1: Different types of subjectivity: users A and B disagree to which items a soft attribute applies. Ovals represent the set of items to which the user applies the attribute.

exhibit elements of all three, these forms represent independent properties that likely require different modeling and learning methods. We then discuss how *contextual* factors may influence the assessment of subjectivity.

Degree Subjectivity. The simplest form of subjectivity is that of *degree*. It arises when users translate a scalar or ordinal attribute A_s (with an induced ordering $<_{A_s}$ over items) into a boolean A_b. The boolean attribute exhibits *degree subjectivity* when most users agree on the meaning of A_s, but apply different *thresholds* when translating it to A_b. For instance, two users u_1, u_2 may each want an 'inexpensive' camera, treating this as a boolean attribute A_b translated from the scalar 'price' attribute A_s. However, if they have different needs or budgets, they may disagree on the price threshold that qualifies as inexpensive. Degree subjectivity bears some connection to *fuzzy sets* (Zadeh 1975), which have been applied to linguistic representations (Cock, Bodenhofer, and Kerre 2000).

We note that attributes like price are not only *objective*, but have a ground-truth source (e.g., the item catalog). We refer to such as *hard attributes*, in contrast to *soft attributes* whose application to items must be learned/inferred from noisy and/or incomplete data (e.g., tagging data or review text). Soft scalar attributes include 'spiciness' (recipes/menu items), 'degree of violence' (movies), 'up-tempo' (music), etc. Nevertheless, concepts derived from both hard attributes (such as 'inexpensive') and soft attributes (such as 'spicy') may exhibit degree subjectivity.

Semantic Subjectivity. An especially challenging form of subjectivity is *semantic subjectivity*, where the same attribute or tag is imbued with different meaning by different users. For example, two users may use the term 'funny' (or the ordinal 'funnier') to refer to different movies in a way that cannot be explained by degree subjectivity. User u_1 may use 'funny' to refer to films with clever dialogue, dark humor or political satire, while u_2 may use the same term for more physical or slapstick comedic stylings. In contrast to degree subjectivity, semantic subjectivity implies that two users may disagree on the ordering $<_{A_s}$ induced by the subjective ordinal attribute—in the case of 'funny,' there will exist pairs of movies (i_1, i_2) for which u_1 would assess i_1 to be funnier than i_2, while u_2 believes the opposite. We refer to these different meanings as *senses* of the term. Other examples of such subjectivity include 'thought-provoking' films, 'healthy' restaurants, and 'interesting' museums.

We believe that such senses will exhibit some structure—the different senses of funny (or violent, etc.) are likely to have much more pairwise agreement (e.g., exhibit higher *rank correlation*) than those of two random attributes. Moreover, we expect that for some attributes/terms, large subgroups of users will share (roughly) the same semantics. Both properties (partial alignment of senses and common sense adoption among subgroups) should prove useful by allowing generalization across users when learning the meaning of a soft attribute and its senses. At the same time, the former may make it more challenging to disentangle different senses in a personalized way.

Compositional Subjectivity. A third form is *compositional* subjectivity: a user may use an attribute as shorthand for some combination of more fundamental attributes; e.g., a user may use 'safe' to express vehicle preference, or 'family-friendly' for restaurant recommendations. Each user may have a specific, but different, concept in mind when using that term. For instance, a 'safe' vehicle for a young parent may mean adequate child restraints, traction control and a good crash-test rating. For a sports-car enthusiast it may mean high-performance brakes, racing seats and a heads-up display. A key distinction between compositional and semantic subjectivity is that the former can often be grounded by referring to more primitive attributes, without referring to specific items, while the latter cannot. For example, a conversational recommender may ask the user to "define" what they mean by a 'safe' vehicle in terms of known vehicle attributes (e.g., "Does crash-test rating matter to you?" or "Do you require ABS?") (Boutilier, Regan, and Viappiani 2009): this can be accomplished, in principle, without referring to any particular vehicle. By contrast, extracting personal meaning for semantically subjective attributes will generally require some assessment of specific items by the user.

Contextual Factors. It is well-understood that user preferences are conditional on the context in which a recommendation is made; for instance, a user may have very different restaurant preferences for a family meal vs. a business dinner. But it is less well-appreciated that such contextual factors may also influence a user's *subjective use of certain terms*. For instance, location (a 'posh' restaurant in a user's home town may not be considered 'posh' while abroad), time of day (different music may be considered 'chill' at 6PM and at 2AM), current activity ('upbeat' music may differ when exercising vs. driving children to school), and many other types of context may well influence a user's intended meaning of a specific term. This can complicate the naïve application of all forms of subjectivity discussed above. For example, a given user's threshold (degree subjectivity) of whether a dish is spicy or a movie is violent may change if the recommendation is for a family night with small children.

Another type of context emerges when we consider a user's *interactions* with a recommender. In faceted search or example critiquing, a user may choose the attribute to adjust (and its direction) in a way that depends on the current slate of items being presented or recommended. This context may also influence a user's subjective assessment of an attribute. For instance, a user's assessment of whether a restaurant is 'trendy' or a camera is 'expensive' may be influenced

by other items examined at the same time (e.g., those in a slate of recommendations), items recommended earlier, or by phrasing or positioning of the item. Such cognitive biases and heuristics and their influence on user choice and decision making—such as anchoring, framing, hyperbolic discounting, and endowment effects—have been studied extensively in behavioral psychology, decision theory and economics (Tversky and Kahneman 1974; Camerer, Loewenstein, and Rabin 2003) and are likely to influence user's subjective attribute assessments.

Finally, a user's subjective semantics is likely to be *dynamic*—beyond changes in context—with both subjective assessments and language usage evolving over time as experience and knowledge changes. For instance, increasing familiarity with a product domain might influence the grounding of both degree and semantic subjectivity. Consider that, for example, after significant exposure to or training in a specific musical genre, a user's meaning for terms like an 'advanced' instructional video or 'sophisticated' piece of music may change.

Research Challenges

Addressing attribute subjectivity presents a number of interesting research challenges for a number of AI subdisciplines. We briefly outline some of these in this section.

Representing Subjectivity. The most fundamental question is that of *representing subjective attributes*. Two broad classes of representations are possible (i) treating subjective attributes as distinct concepts that can be related to items; or (ii) treating them as inherent properties of items. Since many recommenders learn item and user *embeddings*, methods that relate attributes to embedding representations are promising (Rendle and Schmidt-Thieme 2010; McAuley and Leskovec 2013; Wu et al. 2019; Nema, Karatzoglou, and Radlinski 2021). One natural model class uses *ball semantics*. Here, attributes are embedded as points in the item embedding space and a distance function $d(i, a)$ between the embedding i of an item and a of an attribute measures the degree to which the item satisfies the (scalar) attribute A_s. An alternate *line semantics* treats a soft attribute as a direction $\vec{a}$ in embedding space. A key advantage of line semantics is that composition of soft attributes can be interpreted as intersections of hyperplanes (guaranteed to have support if the number of attributes composed is less than the dimensionality of the embedding space). In ball semantics, composition requires intersections, which may result in empty support regions. Furthermore, line semantics allows a symmetric treatment of boolean attributes (e.g., 'spicy' vs. 'bland' as opposite directions in embedding space), whereas in ball semantics, if 'spicy' is a ball, its complement 'bland' is not.

An adequate representation of degree subjectivity should: (i) accurately associate soft attributes, and their relative degree, with items; (ii) adequately distinguish subjective degree implicit in the usage of different users; (iii) support elicitation or active learning of user-specific, *personal* semantics; and (iv) support effective update of user representations. In ball semantics, u's personal boolean-attribute semantics is captured by a distance threshold (or hyper-circle centered on the attribute embedding point a). In a line semantics, u's semantics can be viewed as a threshold on the projection of an item's representation onto the attribute direction.

A model for semantic subjectivity should allow the representation of different senses for a given attribute and, again, support effective learning of senses from data. Possible approaches include (i) using a small, discrete set of s senses, with each user adopting one such sense; (ii) using a continuum of senses defined directly on the embedding space; or (iii) using a mixture of discrete *proto-senses*, where each user is some distribution over these proto-senses.

Models for compositional subjectivity may benefit from techniques from concept learning (Angluin 1988) to uncover user-specific definitions of soft attributes (e.g., 'safe' car) as logical combinations of hard attributes, where precise definitions vary across users (Boutilier, Regan, and Viappiani 2009).

We note also that there may be other types of subjectivity than those we consider here, or an alternative ontology of types which invites completely different treatments.

Uncovering Subjectivity in Data. Key questions regarding the practical import of subjectivity in RSs must ultimately be answered with data. These include the degree to which subjectivity arises in attribute usage in recommenders (Balog, Radlinski, and Karatzoglou 2021), the structure of subjectivity, its overall impact on recommendation quality and user experience, and validity of our proposed taxonomy. We describe some promising data sources and directions.

Social tagging datasets (e.g., MovieLens Tags (Harper and Konstan 2015), Bibsonomy (Jäschke et al. 2009), Goodreads shelves (Wan and McAuley 2018)) provide **(user, item, tag)** triplets, and can be used to study subjectivity in the relationship between items and tags. Most work has modelled subjectivity as a simple scalar degree of agreement across users (Vig, Sen, and Riedl 2012; Kobren et al. 2019), without learning patterns in how users interpret attributes. Richer models have been proposed for *personalized tag recommendation*, e.g., recommending tags for a **(user, item)** pair (Jäschke et al. 2009; Rendle and Schmidt-Thieme 2010).

Natural language corpora can support learning attribute subjectivity using language models trained on those corpora (Devlin et al. 2019; Raffel et al. 2020). These may encode, say, that 'funny' is more subjective than 'violent' if organic text shows more disagreement in the usage of the former. Welch et al. (2020) show that *personalized word embeddings* can be built using text from different users. Subjectivity could be extended to richer phrases with more expressive power (e.g., 'reminds me of my childhood'). Other natural language sources include user-contributed reviews (McAuley and Leskovec 2013; Ni, Li, and McAuley 2019) or conversations/dialogs (Byrne et al. 2019; Radlinski et al. 2019).

Learning Personalized Semantics. Understanding what a *particular* user means when she uses a subjective attribute is a "small data" problem. CF (Su and Khoshgoftaar 2009) and few-shot learning (Wang et al. 2020) are two effective approaches for these types of problems in other fields, but to our knowledge neither has been applied to this task. In cases where data is not available, *active learning* or *concept elicita-*

tion might efficiently query a user for their intended meaning (Boutilier, Regan, and Viappiani 2009, 2010). The design of user interfaces for eliciting the meaning of subjective concepts, and their evaluation in user studies, is an important understudied problem. Standard approaches soliciting positive and negative labels to learn the meaning of an attribute (as for RSs (Boutilier, Regan, and Viappiani 2010) or image classifiers (Kim et al. 2018)) could be extended to also capture subjectivity of terminology across a user population.

Finally, techniques used to model disagreements in crowdsourcing tasks (e.g., identifying subgroups of raters with common interpretations (Kairam and Heer 2016)) might be adapted to our setting to provide initial steps towards learning the different forms of subjectivity and their application to specific domain attributes.

Metrics and Methodology. Open benchmarks (datasets combined with a suite of standard metrics) have historically been a key driver of research progress in many fields of AI. It is important to design benchmarks that measure the extent to which recommenders properly manage subjectivity. Jäschke et al. (2009) present one such benchmark for the related problem of personalized tag recommendation. Balog, Radlinski, and Karatzoglou (2021) propose novel measures of tag subjectivity. In crowd-sourcing, which has been widely used in many AI domains to collect labeled datasets (Sheng and Zhang 2019), quality control usually involves measuring inter-annotator disagreement (Cohen 1960; Welty, Paritosh, and Aroyo 2019), or label noise (Passonneau and Carpenter 2014). New methods may be required for validating crowdsourced subjectivity data, since *disagreements* carry meaningful information in our setting. Specifically, if the goal is to measure *disagreements* then *agreement*-based metrics may not be useful for data validation.

Preference Elicitation. *Preference elicitation* is often used to improve understanding of a user's preferences (Pu and Chen 2008; Viappiani and Boutilier 2010; Bonilla, Guo, and Sanner 2010). Proper handling of subjectivity should make such techniques more natural and effective. Since a user's response to a question involving a subjective attribute may have multiple interpretations, RSs may be served well by maintaining some measure of uncertainty over the *user's semantics*, not just her preferences, and the ability to query the user for semantic information ("what do you mean by funny?" or "do you consider movie m_1 to be funnier than m_2?"). The *joint distribution* of semantics and preferences is also of interest: can knowledge of the user's preferences tell us about their semantics and vice versa?

Another challenge is developing *user choice models* (Luce 1959) for preference queries that involve subjective attributes. No standard model exists to interpret a user's intent when she calls a movie 'violent': is it more violent than the average movie? a recently watched movie? or popular or protypical movie? There is also evidence that the details of the elicitation protocol affect the degree to which users employ subjective attributes (Radlinski et al. 2019).

Explanations. Explanations have a large impact on how users receive recommendations. Methods for generating explanations have been designed for various objectives, such as decision-making efficiency or establishment of trust (Herlocker, Konstan, and Riedl 2000; Sinha and Swearingen 2002; Nunes and Jannach 2017; Balog and Radlinski 2020). However, to the best of our knowledge, current methods for generating explanations in RSs do not incorporate the user's subjective interpretation of terminology. Rather, natural language explanations derived from reviews typically rely on the *reviewer's* interpretation (e.g., (Musto et al. 2019; Donkers, Kleemann, and Ziegler 2020)). As such, current explanation techniques are insensitive to attribute subjectivity. This leaves an important opportunity to tailor explanations to a specific user's interpretation of attributes, i.e., that mirror the subjective language the user employs to describe their intents/preferences ("you may find this movie thought-provoking"). Such explanations may help build trust and invite users to express their preferences in similarly sophisticated and personalized language.

Difficulties are likely to arise if subjective attributes adopted in explanations are presented to the user as if they are objective. Careful phrasing strategies may be appropriate to reinforce the subjective nature of particular explanations, especially when a subjective attribute is used in a way that may conflict with a user's own interpretation (e.g., "*Many reviewers found* this movie to be thought-provking."). This requires the ability to identify which statements refer to subjective concepts and which do not.

Conclusion

The emergence of conversational recommenders offers tremendous opportunities to increase engagement with users by allowing more direct and natural interactions, but requires managing the inherent ambiguity in natural language. We have argued that effective modeling of *subjectivity* in a user's terminology is a critical, yet understudied, part of this program, and that a more rigorous and intentional treatment of subjectivity is a key technological advance needed to support conversational recommenders. Beyond recommendation, most AI domains and sub-disciplines that either consume or produce natural language should benefit from a more principled treatment of subjectivity.

We categorized three forms of subjectivity and outlined a rough formalization intended to motivate new research on this topic. We expect future work will develop a more nuanced understanding of these notions and identify new types of subjectivity. Finally, we outlined a set of broad research challenges including: (i) the development of suitable representations and formal semantics for subjective attributes in RSs; (ii) techniques for identifying attribute subjectivity from existing datasets or the generation of new datasets for this purpose, and assessing the prevalence of subjectivity in current and future modes of interaction with RSs; (iii) the development of methodology and metrics for assessing the capabilities of any proposed treatment of subjectivity; and (iv) new models and algorithms that account for and exploit subjectivity in important recommender sub-tasks such as preference elicitation and explanation. Advances in these directions will greatly enhance our understanding and adoption of subjectivity attributes in RSs.

References

Angluin, D. 1988. Queries and Concept Learning. *Machine Learning*, 2(4): 319–342.

Aula, A.; Khan, R. M.; and Guan, Z. 2010. How does search behavior change as search becomes more difficult? In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 35–44.

Balog, K.; and Radlinski, F. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 329–338.

Balog, K.; Radlinski, F.; and Karatzoglou, A. 2021. On Interpretation and Measurement of Soft Attributes for Recommendation. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*.

Basu, C.; Hirsh, H.; and Cohen, W. 1998. Recommendation as classification: Using social and content-based information in recommendation. In *Proc. AAAI/IAAI*, 714–720.

Bernardi, L.; Kamps, J.; Kiseleva, J.; and Müller, M. J. I. 2015. The Continuous Cold-start Problem in e-Commerce Recommender Systems. In Bogers, T.; and Koolen, M., eds., *Proc. Workshop on New Trends on Content-Based Recommender Systems at RecSys*, volume 1448, 30–33.

Bonilla, E. V.; Guo, S.; and Sanner, S. 2010. Gaussian Process Preference Elicitation. In *Advances in Neural Information Processing Systems 23 (NIPS-10)*, 262–270.

Boutilier, C.; Regan, K.; and Viappiani, P. 2009. Online feature elicitation in interactive optimization. In *Proc. Annual International Conference on Machine Learning*, 73–80.

Boutilier, C.; Regan, K.; and Viappiani, P. 2010. Simultaneous elicitation of preference features and utility. In *24th AAAI Conference on Artificial Intelligence*, AAAI '10.

Boutilier, C.; Zemel, R. S.; and Marlin, B. 2003. Active Collaborative Filtering. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI-03)*, 98–106.

Byrne, B.; Krishnamoorthi, K.; Sankar, C.; Neelakantan, A.; Duckworth, D.; Yavuz, S.; Goodrich, B.; Dubey, A.; Kim, K.-Y.; and Cedilnik, A. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Camerer, C. F.; Loewenstein, G.; and Rabin, M., eds. 2003. *Advances in Behavioral Economics*. Princeton, New Jersey: Princeton University Press.

Chang, S.; Harper, F. M.; and Terveen, L. 2015. Using groups of items for preference elicitation in recommender systems. In *Proc. ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1258–1269.

Chen, L.; and Pu, P. 2012. Critiquing-based Recommenders: Survey and Emerging Trends. *User Modeling and User-Adapted Interaction*, 22(1): 125–150.

Christakopoulou, K.; Radlinski, F.; and Hofmann, K. 2016. Towards Conversational Recommender Systems. In *Proc. SIGKDD International Conference on Knowledge Discovery and Data Mining*, 815–824.

Cock, M. D.; Bodenhofer, U.; and Kerre, E. E. 2000. Modelling Linguistic Expressions Using Fuzzy Relations. In *Proc. International Conference on Soft Computing*, 353–360.

Cohen, J. A. 1960. Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 37–46.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Donkers, T.; Kleemann, T.; and Ziegler, J. 2020. Explaining Recommendations by Means of Aspect-Based Transparent Memories. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, 166–176.

Gantner, Z.; Drumond, L.; Freudenthaler, C.; Rendle, S.; and Schmidt-Thieme, L. 2010. Learning Attribute-to-Feature Mappings for Cold-Start Recommendations. In *IEEE International Conference on Data Mining (ICMD-10)*, 176–185.

Graus, M. P.; and Willemsen, M. C. 2015. Improving the user experience during cold start through choice-based preference elicitation. In *Proc. ACM Conference on Recommender Systems*, 273–276.

Grudin, J.; and Jacques, R. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proc. CHI Conference on Human Factors in Computing Systems*, 1–11.

Habib, J.; Zhang, S.; and Balog, K. 2020. IAI MovieBot: A Conversational Movie Recommender System. In *Proc. ACM International Conference on Information & Knowledge Management*, 3405–3408.

Harper, F. M.; and Konstan, J. A. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Herlocker, J. L.; Konstan, J. A.; and Riedl, J. 2000. Explaining Collaborative Filtering Recommendations. In *Proc. ACM Conference on Computer Supported Cooperative Work*, 241–250.

Jäschke, R.; Eisterlehner, F.; Hotho, A.; and Stumme, G. 2009. Testing and Evaluating Tag Recommenders in a Live System. In Benz, D.; and Janssen, F., eds., *Workshop on Knowledge Discovery, Data Mining, and Machine Learning*, 44–51.

Kairam, S.; and Heer, J. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proc. ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1637–1648.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; and sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proc. 35th International Conference on Machine Learning*, 2668–2677.

Kobren, A.; Barrio, P.; Yakhnenko, O.; Hibschman, J.; and Langmore, I. 2019. Constructing High Precision Knowledge Bases with Subjective and Factual Attributes. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2050–2058.

Lakoff, G. 1999. Cognitive models and prototype theory. *Concepts: Core Readings*, 391–421.

Lee, S.; and Yong, H. 2007. Component Based Approach to Handle Synonym and Polysemy in Folksonomy. In *7th IEEE International Conference on Computer and Information Technology (CIT 2007)*, 200–205.

Luce, R. D. 1959. *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Luo, K.; Yang, H.; Wu, G.; and Sanner, S. 2020. Deep Critiquing for VAE-Based Recommender Systems. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 1269–1278.

McAuley, J.; and Leskovec, J. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proc. ACM Conference on Recommender Systems*, RecSys '13, 165–172.

Musto, C.; Lops, P.; de Gemmis, M.; and Semeraro, G. 2019. Justifying Recommendations Through Aspect-based Sentiment Analysis of Users Reviews. In *Proc. 27th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '19, 4–12.

Nema, P.; Karatzoglou, A.; and Radlinski, F. 2021. Disentangling Preference Representations for Recommendation Critiquing with β-VAE. In *Proc. ACM International Conference on Information and Knowledge Management*.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 188–197.

Nunes, I.; and Jannach, D. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Model. User-Adap. Inter.*, 27(3): 393–444.

Passonneau, R. J.; and Carpenter, B. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2: 311–326.

Pazzani, M. J.; and Billsus, D. 2007. Content-based Recommendation Systems. In Brusilovski, P.; Kobsa, A.; and Nejdl, W., eds., *The Adaptive Web*, volume 4321, 325–341. Springer.

Pu, P.; and Chen, L. 2008. User-involved Preference Elicitation for Product Search and Recommender Systems. *AI Magazine*, 29(4): 93–103.

Radlinski, F.; Balog, K.; Byrne, B.; and Krishnamoorthi, K. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proc. SIGdial Meeting on Discourse and Dialogue*, 353–360.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rendle, S.; and Schmidt-Thieme, L. 2010. Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation. In *Proc. ACM International Conference on Web Search and Data Mining*, 81–90.

Sheng, V. S.; and Zhang, J. 2019. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proc. AAAI Conference on Artificial Intelligence*, volume 33, 9837–9843.

Sinha, R.; and Swearingen, K. 2002. The Role of Transparency in Recommender Systems. In *CHI Extended Abstracts on Human Factors in Computing Systems*, 830–831.

Su, X.; and Khoshgoftaar, T. M. 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009.

Taijala, T. T.; Willemsen, M. C.; and Konstan, J. A. 2018. MovieExplorer: building an interactive exploration tool from ratings and latent taste spaces. In *Proc. ACM Symposium on Applied Computing*, 1383–1392.

Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157): 1124–1131.

Viappiani, P.; and Boutilier, C. 2010. Optimal Bayesian Recommendation Sets and Myopically Optimal Choice Query Sets. In *Advances in Neural Information Processing Systems 23*, 2352–2360.

Vig, J.; Sen, S.; and Riedl, J. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Trans. Interact. Intell. Syst.*, 2(3).

Vig, J.; Soukup, M.; Sen, S.; and Riedl, J. 2010. Tag expression: Tagging with feeling. In *Proc. ACM Symposium on User Interface Software and Technology*, UIST '10, 323–332.

Wan, M.; and McAuley, J. J. 2018. Item recommendation on monotonic behavior chains. In *Proc. ACM Conference on Recommender Systems*, 86–94.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3): 1–34.

Welch, C.; Kummerfeld, J. K.; Pérez-Rosas, V.; and Mihalcea, R. 2020. Exploring the Value of Personalized Word Embeddings. In *Proc. International Conference on Computational Linguistics*, 6856–6862.

Welty, C.; Paritosh, P.; and Aroyo, L. 2019. Metrology for AI: From Benchmarks to Instruments. arXiv:1911.01875.

Wu, G.; Luo, K.; Sanner, S.; and Soh, H. 2019. Deep language-based critiquing for recommender systems. In *Proc. 13th ACM Conference on Recommender Systems*, 137–145.

Zadeh, L. A. 1975. The concept of a linguistic variable and its application to approximate reasoning - I. *Inf. Sci.*, 8: 199–249.

Zhang, Z.-K.; Zhou, T.; and Zhang, Y.-C. 2011. Tag-aware recommender systems: a state-of-the-art survey. *Journal of computer science and technology*, 26(5): 767.

Zheng, B.; Zhang, W.; and Feng, X. F. B. 2013. A Survey of Faceted Search. *Journal of Web Engineering*, 12(1&2): 041–064.

Zou, J.; Chen, Y.; and Kanoulas, E. 2020. Towards question-based recommender systems. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 881–890.