

Dialogue Analysis with Graph Databases: Characterising Domain Items Usage for Movie Recommendations

Martina Di Bratto, Maria Di Maro, Antonio Origlia, Francesco Cutugno

URBAN/ECO Research Center - University of Naples "Federico II"

`martina.dibratto@unina.it, maria.dimaro2@unina.it,`

`antonio.origlia@unina.it, cutugno@unina.it`

Abstract

Nowadays, the use of graph databases combined with textual corpora analysis seems to play a pivotal role in supporting dialogue systems design and implementation. However, dialogues are rarely put in an explicit relationship with the graph structures representing the knowledge domain. In this work, we show how native graph databases provide a framework for a deeper understanding about the use of domain items during dialogue. We describe a multiple-source data collection procedure and we describe how linguistic concepts related to common ground can be found in graph structures. We also describe different patterns that can be detected in the obtained graph structures and discuss their implications in the design of dialogue systems for the movie recommendation task.

1 Introduction

Graph-based data have become popular as support tools for a number of Natural Language Processing tasks, from Word Sense Disambiguation using node embeddings (Yao et al., 2017), to Knowledge Base Collection (Yu et al., 2020) and Fraud Detection (Srivastava and Singh, 2018; Stray, 2019). Most recent approaches involving the use of graph databases converge towards numerical representations of the included items for use by machine learning algorithms (Chanpuriya et al., 2020; Yang et al., 2020b). Knowledge structures using Labelled Property Graphs, however, are typically designed to be more interpretable by human researchers, who can also setup informative queries to extract latent knowledge from a number

of cross-referenced resources. This suggests that the same resources supporting dialogue systems can be queried by human experts to extract deeper understanding about the use of domain items, supporting explainability. This implies that it is necessary to explicitly cross-reference dialogue corpora with domain knowledge, while the two are typically considered separately.

In this paper, we describe a multiple-source data collection procedure to cross-reference an annotated English dialogue corpus covering the movie recommendation task, the Internet Movie Database and Wikidata. We present the relationship *common ground* concepts (see Section 3) have with graph structures and we present a set of examples concerning dialogue analysis performed in the graph. The main research questions are:

- Q1: How to represent dialogues and domain knowledge in a single graph structure?
- Q2: Is it possible to formalise common ground concepts guiding dialogue analysis in the resulting graph?
- Q3: Is it possible to use the resulting graph to extract interpretable patterns to guide dialogue systems design?

The paper is organised as follows: in Section 2 we describe the considered resources and how they were cross-referenced in a graph database. In Section 3, we provide a deeper discussion about Common Ground representations and their relationship with graph databases. Section 4 shows how dialogue history, connected with domain knowledge, can be queried to extract knowledge characterising the use of domain items in the considered (sub-)dialogues. This is of interest for the open challenge concerning the study of Argumentation-Based Dialogue which, as opposed to Argumentation-Based Inference, is an

area that is still lacking a theoretical framework of reference (Prakken, 2018).

2 Materials and Methods

To demonstrate how a single, graph-based, resource can be built to support corpus-based linguistic analyses, we present, first of all, the assembling procedure of a graph database in the movie recommendation domain. Specifically, we present how freely available information about the movies domain were imported and integrated with data coming from a corpus of chat-based interactions between human subjects.

2.1 The Inspired Corpus

For the purpose of this study, the Inspired corpus is considered (Hayati et al., 2020). It is a recommendation dialogue dataset of two-paired crowd-workers who chat in a natural setting in English. In each conversation, different roles have been assigned to each participant: one acts as the recommender, while the other acts as the movie seeker. The aim of the recommenders is to recommend a movie to the seekers following their preferences, thus achieving the conversational goal successfully. Sociable recommendation strategies are annotated by two experts with a linguistic background, based on past social science studies. The annotation schema is composed by a set of persuasive strategies that has been divided in two categories: preference elicitation strategies and sociable strategies. The whole dataset consists of 1,001 dialogues, with a total of 35,811 utterances, where each recommender’s utterance is manually annotated with the corresponding sociable strategies. Domain items like movies, people, genres and plots are tagged in the transcriptions. The dataset presents a average of turns per dialogue equal to 10.73 since recommenders are asked to continue the chat for a minimum of 10 turns.

2.2 Cross-Domain Graph Design

In this work, we adopt Neo4J (Webber, 2012): an open source graph database manager that has been developed over the last 16 years and applied to a high number of tasks related to data representation (Dietze et al., 2016), exploration (Drakopoulos et al., 2015) and visualisation (Jiménez et al., 2016). Neo4j is characterised by high scalability and ease of use. It is a native graph database using data structures that, differently from other

graph based approaches, like the ones based on RDF, are designed for performance speed and optimised for graph traversal operations. In general terms, approaches based on RDF are designed for compatibility and general purpose knowledge representation while graph databases are more application oriented. Neo4j has been used in (Sanonetti et al., 2019) as part of a social recommender system based on friends networks extracted from Facebook and on cultural heritage data coming from DBPedia and Europeana. Automatically collecting and organising large amounts of data is relatively easy, nowadays, given the availability of Linked Open Data (LOD). Graph databases, moreover, provide the necessary flexibility to cross-reference different resources, enabling dialogue analysis in relationship with domain knowledge representation. The knowledge base for the movies domain is built by collecting data from different sources and organising it so that cross-referencing is possible. The first step consists in importing the data provided by the Internet Movie Database (IMDB), which is the most structured data source available for the specific domain. The main node labels imported after this step are MOVIE and PERSON. Moreover, for the MOVIE nodes, genres are also represented as labels (e.g. WESTERN, THRILLER, etc...). MOVIE and PERSON nodes are linked by WORKED_IN, WROTE and DIRECTED relationships, mainly. IMDB does not report information about awards won by people and movies. This however, is an important information to consider when ranking potential recommendations. Such information is found in Wikidata, which also contains the IMDB ids for both movies and persons (P345). Connecting Wikidata to IMDB is straightforward, given the already existing alignment. The *award_received* (P166) relationship in Wikidata connects movies and awards while also providing optional qualifiers to further detail the relationship. The qualifier we consider for *award_received* is *winner*, connecting the relationship to one or multiple people. Awards are represented by AWARD nodes and each specific award is represented by a node labelled as AWARDINSTANCE. This way, each AWARDINSTANCE can be connected to multiple winners, to the movie it was awarded for and to the AWARD node it represents. The second part of the procedure covers the import of dialogue data from the Inspired corpus.

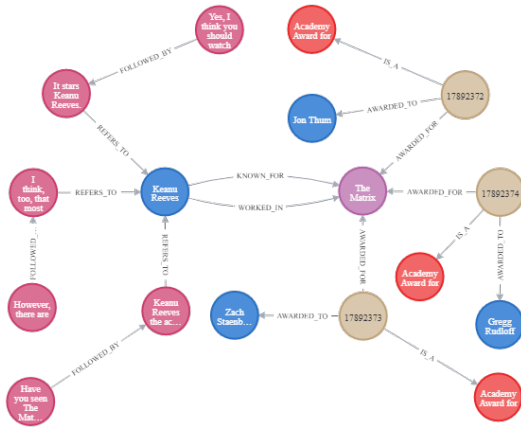


Figure 1: An extract of the final graph structure, including MOVIES (purple), PERSONS (blue), AWARDINSTANCES (brown) of AWARDS (red).

Since Inspired is a dialogue corpus constituted by a sequence of turns between two subjects acting either as an information seeker or as a domain expert recommender, the most natural way to represent dialogue history in a graph form is by a nodes chain. Therefore, each turn is represented by an UTTERANCE node also having a RECOMMENDER/SEEKER label representing who stated the sentence. Since there are no returning users in Inspired, there is no need to model them as separate nodes. UTTERANCE nodes expressing specific intents, were marked with labels corresponding to the Inspired sociable strategies. UTTERANCE nodes were linked using FOLLOWED_BY relationships to keep track of the turns sequences. Connecting Inspired to the data collected from the web is performed using text matches between substrings marked as named entities in the original dataset (e.g. movies, people, etc...) and the name fields collected from IMDB and Wikidata. At this time, potential ambiguities are introduced in the graph due to homonyms: these are later solved during graph analysis, as discussed in Section 4. The final structure of the graph is shown in Figure 1. We are not able to share the database, as we cannot redistribute IMDB data. However, the code to generate it, once the data is obtained from the official source, is freely available¹. The final graph contains approximately 17 million nodes and 75 million relationships.

¹https://github.com/antori82/MS_MovieGraph

3 Common Ground and Dialogues

Conversing is a joint activity, for which goals and roles of the interlocutors must be identified in order to reach the conversational targets (Macagno and Bigi, 2017). For this purpose, mutual understanding is so fundamental that messages need to be encoded upon a *common ground*, defined as the “[...] presumed background information shared by participants in a conversation” (Stalnaker, 2002, p. 1). Common ground, as (Clark, 2015) acknowledged, can be of four main types: personal, local, communal and specialised. In this work, we focus on personal and communal common ground. *Personal Common Ground* (PCG) is established collecting information over time through communicative exchanges with an interlocutor and it can be considered as a record of shared experiences with that person. This specific set of information can also be considered, as in this work, as part of what builds the *Personal Experience* (PE) of an interlocutor, useful in the future steps of the interaction or for future interactions. *Communal Common Ground* (CCG) refers to the amount of information shared with people belonging to the same community, such as general knowledge, knowledge about social background, education, religion, nationality, and language(s).

In this work, we present how different sets of shared knowledge can be represented in the form of graphs to be analysed in depth. While the proposal draws significant inspiration from older systems based on inference engines, the aim is to re-interpret the approach using modern technologies that naturally blend with machine learning. Graph-based representations have gained popularity, in recent years, due to their flexibility and expressiveness. In (Yang et al., 2020a), knowledge graphs were used for intent identification in multi-turn dialogues. (Lei et al., 2020) proposes a graph based framework to detect user preferences in multi-turn dialogues. This approach covers Knowledge Graph navigation but it does not represent dialogues in the graph. Differently, (Wang et al., 2021) proposes the use of graph structures both to represent interactions and learn intents automatically. User-item interactions were part of the graph but they were treated as isolated items. We include sequence dynamics as the recorded order from Inspired is explicitly represented.

Graph structures obviously open new perspectives in automatic dialogue management and it is,

therefore, important to investigate how linguistic concepts can be used to describe with deeper details the recurring structures forming between dialogues (PE) and domain knowledge (CCG). While the CCG contains entities and relationships between entities that belong to the common knowledge of a particular domain (i.e., in the movie domain: ‘directors direct movies’, ‘George Lucas directed Star Wars’, etc.), the PCG is built considering the CCG to select specific entities during the interaction as being related to one another according to particular relationships. The PCG records the interaction flow with a specific user to also extract possible personalised strategies for future interactions with the same user (i.e., ‘The user X is looking for Sci-fi movies; the user X decides to watch The Matrix’). This sub-graph has a temporary status as, at the end of the interaction, it is incorporated in the PE, which contains information about entities and their relationships extracted from various interactions (i.e., ‘Users who usually look for Sci-fi movies also choose to watch The Matrix’). The PE sub-graph, therefore, is complementary to the information contained in the CCG (i.e., the fact that The Matrix is a Sci-fi movie) for other interactions or different interaction phases, such as the *persuasion dialogue* phase, used to “resolve a conflict of opinion between real agents, who can ask for and provide substantive reasons for their claims” (Prakken, 2018). In the graph structure, we now identify the different types of CG as sub-graph structures, supporting a linguistically oriented view of the obtained resource based on mathematical notations. First of all, we identify the CCG in the set of data that are obtained by importing commonly available resources so that

$$CCG = \langle V_{lod}, E_{lod} \rangle \quad (1)$$

where V_{lod} is the set of nodes imported from LOD and E_{lod} are the reported relationships among these nodes. PE refers to recorded interactions between RECOMMENDERS and SEEKERS, so

$$PE = \langle V_d, E_d \rangle \quad (2)$$

where V_d is the set of nodes representing utterances from the Inspired corpus and E_d is the set of chain relationships recording the dialogue evolution. PE contains information that is not accessible by other interlocutors, as it concerns the owning speaker, who can disclose parts of it and provide unverifiable information that other involved actors

should trust. The PCG is a subset of the PE involving the dialogue chain representing the ongoing interaction so that $PCG \subseteq PE$. Named entities in the PE graph are linked to items in the CCG by a ϕ function representing the result of the *grounding* process. The domain of the ϕ function D_ϕ consists of utterances containing named entities, while its range R_ϕ consists of CCG items. Links defined by ϕ generate a *reference graph* R defined as

$$R = \langle V_\phi, E_\phi \rangle \quad (3)$$

where $V_\phi = \{D_\phi \cup R_\phi\}$ and $E_\phi = \{(v_d, v_{lod}) | \phi(v_d) = v_{lod}\}$ where $v_d \in V_d$ and $v_{lod} \in V_{lod}$. The reference graph is represented in Neo4j using REFERS_TO relationships between UTTERANCE nodes and nodes representing named items, like MOVIEs and PERSONs.

4 Patterns in Common Ground

Using graphs to analyse PE and CCG together allows to extract interpretable knowledge to obtain deeper understanding of dialogue phenomena and support an better informed design of dialogue systems with respect to solutions relying on machine learning only. The examples we provide concern items characterisation in relationship with dialogue strategies. Since the presented graph structure captures dialogue dynamics in the form of FOLLOWED_BY relationship chains, we use it to analyse dialogue dynamics and to characterise items selected for the employed strategies w.r.t. previous exchanges. This can put in relationship the PE and the CCG explicitly in the form of ring-like patterns crossing both the PE and the CCG. Formally, let’s consider a path $p = [u_n, c_1, \dots, c_w, u_{n-k}, \dots, u_n]$ where $u_i \in V_d$, $c_j \in V_{lod}$, $(u_n, c_1) \in E_\phi$, $(u_{n-k}, c_w) \in E_\phi$ and $\forall 0 \leq m < k, (u_{n-m-1}, u_{n-m}) \in E_d$. The pattern can be used both for disambiguation purposes and to extract the context that led interlocutors to discuss a certain item. In the case shown in Figure 2, first the MOVIE *Knives Out* was named and, after a few turns, a person named *Chris Evans*. There are 103 people, in the database, named *Chris Evans* but only one of them can be directly related to the previous subject, *Knives Out*, removing the ambiguity. The pattern involving the naming of a MOVIE followed by a PERSON can be further studied considering the annotated sociable strategies for the RECOMMENDER.

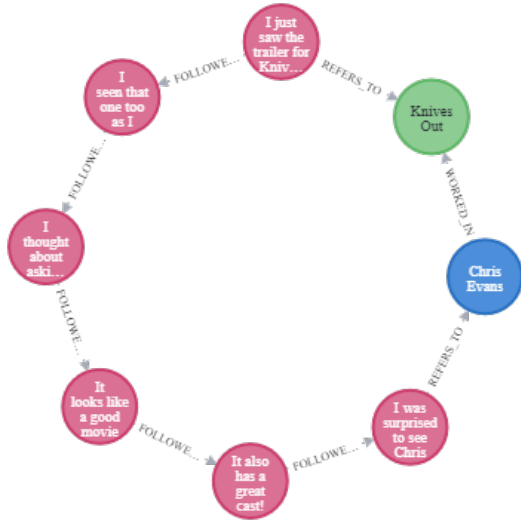


Figure 2: A ring-like pattern found in one of the Inspired dialogues.

Let’s consider the case of ring-like patterns involving a PERSON who can be related to a previously named MOVIE by a WORKED_IN relationship. In the database, there are 122 patterns matching these constraints. In this case, MOVIES involved in this pattern are used to support mainly sociable strategies such as *experience inquiry* but also *encouragement* to watch the previously proposed movie. PERSON items, instead, seem to support mostly *credibility*, used by the RECOMMENDER to show expertise and trustworthiness in providing information. The opposite pattern, where PERSONs are named before the MOVIE they WORKED_IN, is less frequent, with 66 occurrences, but presents more variability in the involved strategies. PERSONs support, in this case, *credibility*, *experience inquiry* and *personal opinion* strategies mostly, while MOVIEs support *personal opinion* and *encouragement* strategies. Figure 3 shows a summary of the observed distributions over the considered patterns. We interpreted these data as an indication of the transition between two different recommendation dialogue phases. In particular, PERSON items are mainly used, by RECOMMENDERS, to persuade the SEEKER by showing competence about MOVIEs mentioned during the Exploration phase. Here, strategies like *experience inquiry* are more frequent given that information about SEEKERs’ tastes and experiences need to be collected (Gao et al., 2021). An interesting observation arises from the analysis of the inverse pattern as it

seems to present a continuity with the previous one. Also here PERSON items are mainly used to support *credibility*, introducing the Exploitation phase, whose aim is to take advantage of the collected information to recommend a potentially valuable item. Indeed, the following MOVIE items involved in the pattern seem to occur in strategies typical of this phase such as *personal opinion* or *encouragement*, mainly used by the RECOMMENDERS to persuade the SEEKER to watch a proposed MOVIE. The pattern that seems to emerge suggests that, during the Exploration Phase, RECOMMENDERS build a model of the SEEKER by collecting previously watched MOVIEs and talking about potentially interesting ones. They show competence about MOVIEs by referencing PERSONs involved in them. The dialogue then blends into the Exploitation phase, where persuasion strategies are adopted.

To further detail the choice of actors involved in dialogues, we consider the relationship actors have with movie genres. Dialogue analysis alone cannot go beyond frequency observations for specific tokens: the cross-referenced database allows to relate named items to the categories items of interest fall into. Specifically, for each PERSON who WORKED_IN a MOVIE, we consider the skewness of the corresponding GENRE distribution. Highly skewed distributions are obtained for PERSONs specialised in a specific GENRE while low skewed distributions are obtained for more *versatile* PERSONs. Interestingly, people with mid-skewed distributions are used more often both w.r.t. specialised PERSONs and w.r.t versatile ones (Figure 4). This suggests that people who work in hybrid genres (e.g. DRAMA/COMEDY) are more *informative* items to discuss. One possible explanation would be that these items provide indications about genre *nuances* or sub-genres.

5 Discussion and Conclusions

Graph databases have a great potential as linguistic tools for dialogue analysis, modelling interaction dynamics and domain knowledge in the same resource. While this representation provides a powerful support to dialogue systems through machine learning, it can be integrated by more informed design based on a qualitative analysis of the same graph structures. We presented a data collection procedure for a large graph database representing dialogue dynamics explicitly linked

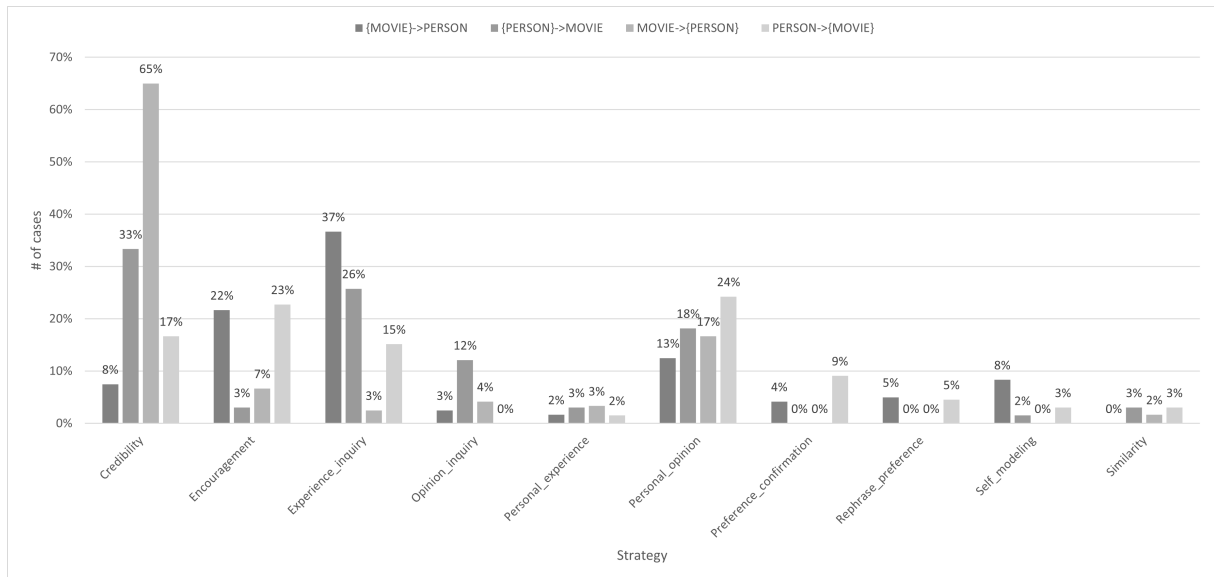


Figure 3: Strategies distribution in ring-like patterns for UTTERANCES using MOVIE and PERSON items. For each pattern, the item and its position in the considered pattern are found between brackets.

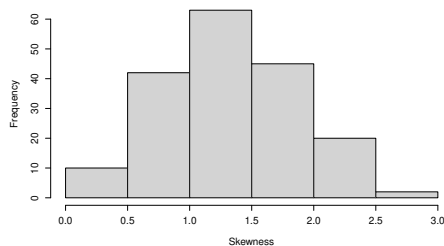


Figure 4: Skewness values distribution.

to domain knowledge. In the representation of such data, different kinds of Common Ground were identified as sub-graphs. We then, used it to investigate the choice of domain items supporting sociable conversational strategies. Concerning the research questions we were interested in investigating, the data collection procedure presented in Section 2.2 presents how a graph database can organise and connect, at the same time, both dialogue history and domain knowledge, providing an example answering Q1 for the specific case of the movies recommendation dialogues. This is, then, generalised in Section 3 by identifying linguistic concepts related to common ground that are described in terms of sub-graphs, independently of the specific domain. Recurring patterns involving both dialogue and common ground structures can, then, be identified using graph traversal queries. Specifically, we have charac-

terised the different use of MOVIE and PERSON nodes depending on the order of appearance in a pattern where the referring nodes are linked by the relationship WORKED_IN, in the CCG. By analysing the sociable strategies involved in the patterns in which the specific domain items may occur, we described how the different use of the same node categories in opposite patterns (i.e., MOVIE followed by PERSON; PERSON followed by MOVIE) can be typical of a specific recommendation dialogue phase. The obtained data describe a way of managing Exploration and Exploitation phases informing the design of conversational recommender systems. We have also shown that, in selecting which PERSON items to use when adopting sociable strategies, RECOMMENDERS seem to prefer PERSONS with mid-skewed distributions over movie genres, suggesting that opinions about these PERSONS may give insight about sub-genres. Future work will consist of empowering the analyses by relating more complex numerical item representations, like node embeddings, to dialogue dynamics.

Acknowledgments

The authors would like to thank Sabrina Mennella for her support in correcting and analysing the Inspired dataset. Maria Di Maro's work is supported by the Italian PON I&C 2014-2020 within the BRILLO project, no. F/190066/01-02/X44.

References

- Sudhanshu Chanpuriya, Cameron Musco, Konstantinos Sotiropoulos, and Charalampos Tsourakakis. 2020. Node embeddings and exact low-rank representations of complex networks. *Advances in Neural Information Processing Systems*, 33.
- Eve V. Clark. 2015. Common ground. In *The Handbook of Language Emergence*, page 328–353. Wiley, Chichester, UK.
- Felix Dietze, Johannes Karoff, André Calero Valdez, Martina Zieffle, Christoph Greven, and Ulrik Schroeder. 2016. An open-source object-graph-mapping framework for neo4j and scala: Renesca. In *International Conference on Availability, Reliability, and Security*, pages 204–218. Springer.
- Georgios Drakopoulos, Andreas Kanavos, Christos Makris, and Vasileios Megalooikonomou. 2015. On converting community detection algorithms for fuzzy graphs in neo4j. In *Proceedings of the 5th International Workshop on Combinations of Intelligent Methods and Applications, CIMA*.
- Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *arXiv preprint arXiv:2101.09459*.
- Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. *arXiv preprint arXiv:2009.14306*.
- Pablo Jiménez, Javier Villalba Diez, and Joaquin Ordieres-Mere. 2016. Hoshin kanri visualization with neo4j. empowering leaders to operationalize lean structural networks. *Procedia CIRP*, 55:284–289.
- Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2073–2083.
- Fabrizio Macagno and Sarah Bigi. 2017. Analyzing the pragmatic structure of dialogues. *Discourse Studies*, 19(2):148–168.
- Henry Prakken. 2018. *Historical overview of formal argumentation*, volume 1. College Publications.
- Giuseppe Sansonetti, Fabio Gasparetti, Alessandro Micarelli, Federica Cena, and Cristina Gena. 2019. Enhancing cultural recommendations through social and linked open data. *User Modeling and User-Adapted Interaction*, 29(1):121–159.
- Sakshi Srivastava and Anil Kumar Singh. 2018. Graph based analysis of panama papers. In *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 822–827. IEEE.
- Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.
- Jonathan Stray. 2019. Making artificial intelligence work for investigative journalism. *Digital Journalism*, 7(8):1076–1097.
- Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat-Seng Chua. 2021. Learning intents behind interactions with knowledge graph for recommendation. In *Proceedings of the Web Conference 2021*, pages 878–887.
- Jim Webber. 2012. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pages 217–218. ACM.
- Kai Yang, Xinyu Kong, Yafang Wang, Jie Zhang, and Gerard De Melo. 2020a. Reinforcement learning over knowledge graphs for explainable dialogue intent mining. *IEEE Access*, 8:85348–85358.
- Luwei Yang, Zhibo Xiao, Wen Jiang, Yi Wei, Yi Hu, and Hao Wang. 2020b. Dynamic heterogeneous graph embedding using hierarchical attentions. *Advances in Information Retrieval*, 12036:425.
- Liang Yao, Yin Zhang, Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang, and Qinfei Chen. 2017. Incorporating knowledge graph embeddings into topic modeling. In *Thirty-first AAAI conference on artificial intelligence*.
- Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. 2020. A domain knowledge graph construction method based on wikipedia. *Journal of Information Science*, page 0165551520932510.