Combining Search with Structured Data to Create a More Engaging User Experience in Open Domain Dialogue

Kevin K. Bowden, Shereen Oraby, Jiaqi Wu, Amita Misra, and Marilyn Walker
Natural Language and Dialogue Systems Lab
University of California, Santa Cruz
{kbowden,soraby,jwu64,amisra2,mawalker}@ucsc.edu

ABSTRACT
The greatest challenges in building sophisticated open-domain conversational agents arise directly from the potential for ongoing mixed-initiative multi-turn dialogues, which do not follow a particular plan or pursue a particular fixed information need. In order to make coherently contributing to the conversation in real-time. This paper describes how we leveraged search and structured data from different sources to help SlugBot produce dialogic turns and carry on conversations whose length over the semi-finals user evaluation period averaged 8:17 minutes.

CCS CONCEPTS
• Human-centered computing → Natural language interfaces; Ubiquitous and mobile computing systems and tools; • Computing methodologies → Natural language processing; Natural language generation; • Information systems → Web searching and information discovery; Ontologies;

KEYWORDS
Dialogue; Personal Assistants; Conversational Agents; Chatbots; Search-Oriented Dialogue; Ontologies

1 INTRODUCTION
The Alexa Prize funded 12 international teams to compete to create a conversational agent that can discuss any topic for at least 20 minutes. UCSC’s Slugbot was one of these funded teams. The greatest challenges with the competition arise directly from the potential for ongoing mixed-initiative multi-turn dialogues, which do not follow a particular plan or pursue a particular fixed information need. This paper describes some of the lessons we learned building Slugbot for the 2017 Alexa Prize, particularly focusing on the challenges of integrating content found via search with content from structured data in order to carry on an ongoing, coherent, open-domain, mixed-initiative conversation. Slugbot’s conversations over the semi-finals user evaluation averaged 8:17 minutes.

Unlike much previous work on conversational AI, Slugbot could not and did not assume that the user had an “information need” [8, 23, 33]. Rather, the design of the Alexa Prize was aimed at open conversations that could engage the user, through any type of dialogue or chitchat, discussing films and books, gossiping about celebrities, playing verbal games, telling stories or sharing experiences, or any other of many different types of activities that conversation is often used for.

This open design foregrounds many longstanding challenges that have not been solved even for task-oriented dialogue systems. These include:
• Modeling discourse coherence;
• Supporting mixed-initiative dialogue;
• Generating contextualized and stylistically appropriate natural language responses.

This paper is structured around the “lessons learned” with respect to these challenges from our experience building Slugbot. To be clear, we are not offering a solution to these problems: instead our intention is simply to highlight the difficulties with developing adequate computational models of these phenomena that particularly arise in the context of open-domain conversations, where users cannot be assumed to be pursuing a particular task or information need. We will attempt to motivate our hypothesis that a comprehensive solution to these challenges for open-domain dialogue requires a much deeper understanding and utilization of the semantic relations that underly dialogue coherence.

For example, consider dialogue focused on content related to the movie domain. This should be one of the easiest domains because it is well-structured, and there are existing systems handling conversations where there is a specified user information need or task, such as finding films with particular properties, finding out what is

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICTIR ’17 Workshop on Search-Oriented Conversational AI (SCAI’17), October 1, 2017, Amsterdam, Netherlands
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-5386-1730-6/17/10
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The competition is still ongoing and its conditions prohibit us from reporting detailed information about the evaluation results or Slugbot’s system architecture.
An example of a recently developed MovieBot that tries to support free-ranging conversations can be found at [https://www.amazon.com/dp/B01MRKGF5W](https://www.amazon.com/dp/B01MRKGF5W).

Playing and where, or booking a movie ticket [6, 14, 22]. Moreover, the Internet Movie Database (IMDB) [20] provides information on plot, rating, and actors that can be leveraged to support conversations. IMDB also makes use of the Schema.org [39] structure to connect common entities to their related attribute types (such as Actor → Person → birthDate), allowing the system to retrieve a large set of possible next topics and related facts and entities.

However, remember that SlugBot is based on the assumption that the user might simply enjoy talking about films and related entities and therefore may freely move the conversational focus among different movie entities, along with the vast array of semantically-associated movie attributes: movies have actors, genres, plots, and awards; actors have names, affiliations, other movies they were in, awards, etc. Actors are people, who have spouses, families and friends, and engage in other life activities besides acting, such as political advocacy.

A potential dialogue is shown in Table 1. The interaction might appear to be simple enough: the user chooses to discuss movies, and selects Jason Bourne as the specific movie she is interested in, the system finds the movie in IMDB, and then provides information on its rating, lead actor, and plot. The user then changes the topic to other movies with the same actor, and the conversation continues.

Even with the availability of IMDB, however, the interaction is not totally straightforward. The RHS of Table 1 describes some of the required competencies and decisions Slugbot must make. First, Slugbot must be able to perform coreference resolution and recognize that the movie and it in turns U6 and U8 are coreferential. We estimate the accuracy of noun-phrase coreference resolution to only be about 70% for off-the-shelf tools applied to dialogue, since most of them are targeted to text [5, 12, 27, 30, 32, 35, 36, 49].

More challenging is that at each system turn, there are a large number of conversational moves that are possible. Making good decisions about what to say next requires balancing a dialogue policy as to what dialogue acts might be good in this context, with real-time information as to what types of content might be possible to use in this context. Slugbot could offer an opinion as in turn S5, ask a follow-up question as in S3, ask the user to provide information, as in U4, or decide, e.g. in the case of the user’s request for plot information, to use search to retrieve some relevant content. Search cannot be used effectively here without constructing an appropriate query, or knowing in advance what plot information might be available. In a real-time system, live search may not be able to achieve the required speed and efficiency, so preprocessing or caching of relevant information may be necessary. Finally, most other domains do not have such high quality structured data available, leaving us to develop or try to rely on more general models of discourse coherence.

### 2 Modeling Discourse Coherence

In open-domain conversation, dialogue coherence between related turns must be maintained. What underlies dialogue coherence goes beyond simple word overlap or similarity, and its clear that neural models of open-domain conversational dialogue do not yet capture it. Theories of discourse posit that there are a small number of semantic relations that can hold between adjacent turns: at the most general level these are CONTINGENCY, COMPARISON, EXPANSION, and TEMPORAL ORDER [18, 19, 31]. We posit that one way to allow Slugbot to take the initiative and produce a turn that maintains discourse coherence is to find content to use in Slugbot’s next turn that instantiates a valid semantic relation between the current

| Table 1: Sample Dialogue about Movies. System content indicated as coming from search. |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| S1 | System | What do you want to talk about? | Opening prompt, expects to map an answer to a known topic. |
| U2 | User | Let’s talk about movies. | Response to open prompt, topic “movies”. |
| S3 | System | I love movies! Which movies have you seen recently? | Take initiative to provide an opinion, then ask followup question. |
| S4 | System | Oh yes, according to ratings, Jason Bourne is a pretty good movie. | SlugBot could have suggested talking about a particular movie, or asked if the user had scene a particular movie. |
| U5 | User | I watched Jason Bourne recently. | Names a particular film, named entity recognition must map “Jason Bourne” to a movie entity. |
| S6 | System | It stars Matt Damon. | SlugBot takes initiative, provide ratings information, not necessarily unique or what is asked for, conversational marker “Oh yes”. |
| U7 | User | Have you heard much about it in terms of the plot? | Need to know that the first sentence is usually good by itself. |
| S8 | System | The CIA’s most dangerous former operative is drawn out of hiding to uncover more explosive truths about his past. | System selects first sentence of plot synopsis that is returned by search. |
| U9 | User | Sounds about right. What other movies has Matt Damon been in? | User requests topic chaining by actor to other movie entities. System must know that actor can be a key to find other movies that star that actor. |

Some theories rely on the plan-based coherence of discourse [13, 26], but we believe that this can only be easily applied to task-oriented dialogue.
user turn and SlugBot’s next turn. One of the strongest bases for such semantic relations are the relations captured by ontologies or frames, which give us related entities, e.g. movies have actors and directors [14, 16]. These types of relations can be used to instantiate the expansion relation, which basically captures moving to strongly related subtopics, often by chaining off a particular discourse entity. To find content to instantiate the expansion relation to use in Slugbot’s next turn (taking the initiative), we carry out the following pipeline:

1. Perform coreference resolution on the user’s turn
2. Identify relevant entities in the user query, i.e. a movie name or sports team, by parsing and checking entities using Google Knowledge Graph.
3. Look up information on the entity to assign relevant attributes, using tools like Schema.org, YAGO, or DBPedia (for example, a Sports Team has players).
4. Retrieve relevant information about the entity. In the easiest scenario, we might have a structured source of information, like IMDB. In other cases, we need to use search. This often then requires further processing, such as parsing unstructured information on Wikipedia or in search results and attempting to extract relevant content.
5. Select content and produce an utterance to give back to the user. This may be accomplished through natural language generation, string selection, or sentence compression [9, 24, 28, 29, 54]. We discuss NLG in more detail in Section 4.

In the case of movies, the structure of IMDB, as discussed above, allows us to link between related entities and attributes using various DB keys. However other conversational domains do not have freely available richly structured information such as this. It is rare for a single resource to aggregate all the information that might be useful, so SlugBot must be able to leverage information and integrate information from multiple sources. But state-of-the-art knowledge bases and ontologies are still limited. Table 2 lists some of the resources that we have found to be most useful for search and structured information.

| # | Tool | Description |
|---|------|-------------|
| 1 | Wikipedia [56] | Multi-lingual, web-based, free-content encyclopedia. |
| 2 | YAGO [34] | Semantic knowledge base from Wikipedia [56], WordNet [44], and GeoNames [38], with over 10 million entities and over 120 million facts about them. |
| 3 | DBPedia [2] | Crowd-sourced semantic knowledge graph using Wikipedia data (around 4.22 million entities in the ontology). |
| 4 | Google Knowledge Graph [11] | API for finding entity information include type and details, with some relevance score. |

Table 2: Search and Structured Information Resources

Like movies, sports is another domain that has rich structure, and in which there is broad user interest. Search results for a query about “Madison Bumgarner” are in Figure 1, showcasing a sample of the different information retrievable from each source (Step 2 of the pipeline).

From the Google Knowledge Graph (Figure 1a result we are able to ascertain the entity type, a brief description, and a relevant Wikipedia page (Figure 1b) which we can use to find accurate structured information. We may further augment our knowledge by using the information returned by the Google Knowledge Graph as parameters to our YAGO or DBPedia query which can more easily extract specific relationships between an entity-attribute. For example, the results returned by YAGO for the “Madison Bumgarner” query contains a connection to the headline *Struggling Madison Bumgarner might not garner next start*, which is contextually relevant data not encapsulated anywhere in the previously examined results.

There, however, exists a disconnect between the resources, i.e. some entities are available in one resource and not another, or there may be inconsistent information across resources. While it would be nice not to have to anticipate the types of integration that are needed, our take-away from this, is that at present, it appears we have to accomplish the steps in our pipeline by integrating knowledge from different resources in advance, even though projects such as YAGO have already been working on such integration for at least ten years.

Other discourse coherence relations besides expansion are also viable candidates for selecting content for next turns, but finding content that instantiates these relations can be a challenging problem in itself. For example, in casual conversation, it is common to provide opinions and then perhaps further take the initiative and justify them. The justification of an opinion is a type of contingency relation: we describe how we curate content to provide justifications in Section 3.

We have also been able to use the temporal relation in a limited way by drawing on narratively structured sources, such as personal stories in blogs. Since these stories are told in temporal order, we can repurpose the content of these blogs to tell stories, maintaining pre-existing narrative coherence when the system produces a sequence of turns [4]. However, we posit that there is much more that could be done to make better use of deep semantic discourse relations for recognizing discourse relations and generating coherent conversational turns.

3 MIXED INITIATIVE DIALOGUE

Mixed Initiative dialogue is key to a natural conversational interaction [1, 3, 7, 15, 33, 55], and this is even more important for open domain dialogue than it is for task-oriented or information seeking dialogue. One of our primary hypotheses, as described above, is that good models of discourse coherence will help SlugBot identify content that can be used to take the initiative. However, models of discourse coherence have been rarely applied to conversation [37, 40, 43] and thus there is considerable work to be done simply in understanding how these relations can be instantiated in dialogue.

In addition, a further challenge arises from the fact that both system and user options for dialogue acts are extremely varied at each turn, e.g. user intents can be to provide opinions, give or solicit information, contrast two possibilities, request the system to perform an action, and more. One reasonable taxonomy for the types of dialogue acts that might be available to SlugBot could be based for example on the dialogue act annotations in the Switchboard corpus [21].
Table 3: Example opinion data used in Table 4 and 5

| Entity | Bin | Sentiment | Justifications |
|--------|-----|-----------|----------------|
| Magneto | comics | 5 | he can control metal |
| Aliens | movie | 4 | well cast, action packed |
| Dracula | monster | 5 | scary |

Table 4: Sample Dialogue about Comic Books. System content based on either search† or structured data‡.

Here, we consider a simple case combining discourse relations and dialogue acts that we have implemented in Slugbot in order to take the initiative in a way that we hoped the user would find interesting. Our aim was to utilize the CONTINGENCY discourse relation to connect a statement of opinion and its justification. We designed a template containing both arguments of the CONTINGENCY relation, namely I think {entity} is {sentiment} because {justification}. We construct a table of argument pairs that can instantiate this relation, as shown in Table 3. This table can be populated by crowd-sourcing or by using search as a pre-processing step.

Table 4 illustrates how this is used in our conversations about comics. At Line 6, when the user asks Who is your favorite character?, it is most appropriate to provide an opinion. It is difficult to imagine retrieving search-based data which contains a contextually relevant opinion, but it is even more difficult to imagine that if search had returned such an opinion, that search could be used a second time in order to retrieve a justification for the provided opinion and answer the user’s follow-up question in Line 8, Okay why?: The source text for the search would have to be annotated for the type of content that could be used to provide justifications, and search would have to support these types of semantic relations.

4 NATURAL LANGUAGE GENERATION

The current challenges for natural language generation, in our view, arise from the need to combine information from structured and unstructured sources when producing conversational utterances. SlugBot currently uses a combination of pre-written templates, sentence selection, and techniques for telling stories that are based on converting monologic stories to dialogic sequences [4].

Structured data, when available, can do more than structure a search result: it can also be easier to use within a conversation because it provides the necessary structure needed for high precision natural language generation [41, 54]. More precisely, a small set of generic templates with various slots can be filled with information from structured data sources to insure high quality, accurate responses. These generic templates can be hand crafted, or prepared in advance by learning natural language generation templates automatically from appropriate conversational domain sources such as different types of user-generated content [17, 29], as illustrated in our justification initiatives above in Section 3.

For general fact-based questions, on the other hand, search content can be used directly. For example, at line 14 in Table 5 when the user asks What was the first movie to feature a vampire?, search provides us with a good response. This introduces however the
challenge of updating the discourse context with the right representation of the two movies under discussion, so that they can then be available for follow-on coreference. This is an open problem.

It is clear that in order to use a semi-structured approach, we need to determine when to utilize each source. Structured data can be easier to formulate into system responses and can often more easily handle on-topic follow-up questions, but is more limited in scope. An obvious approach, also used in the Watson Jeopardy system [10], is to pool responses from both sources and rank them. We have not, to date, collected enough data to build a ranker.

Our plan is to apply a combination of reinforcement learning and learning of ranking functions for utterance variants in a particular context to SlugBot conversations as we move forward with our own data collection, outside of the Alexa Prize competition [25, 48, 50, 51, 57]. The first step however is to use the Alexa Prize competition data to learn a Paradise-Open-Domain evaluation function, with additional metrics relevant to open-domain dialogue, e.g. independent variable metrics that predict overall dialogue quality such as response delay, vocabulary diversity, dialogue act sequence n-grams [45], conversational depth, number of reponses [47], and other measures that can be automatically logged. Many of the required measures have been used over the last 20 years in Paradise to evaluate task-oriented dialogue systems and they remain highly relevant to overall dialogue quality in open-domain dialogue systems [46, 52, 53]. We predict this can potentially improve the overall performance of the system as demonstrated in Table 6. Here, the structured data is sparse, resulting in an uninteresting response, while search returns a very robust answer. Our Paradise-Open-Domain evaluation function would need to learn to place priority on the result returned by search, through ranking, despite having structured data.

For open domain NLG, we have also conducted experiments with neural sequence to sequence approaches using open domain corpora such as film dialogue, Big Bang theory scripts, and open subtitles. These approaches to date do not produce interesting utterances that maintain discourse coherence. It is possible that further curation and semantic annotation of these resources, e.g. by labelling semantic roles and identifying dialogue acts and discourse relations might be helpful, but this could also introduce data sparsity. For example in Switchboard the dialogue act distribution is highly skewed. Integrating information across multiple sources could also be further explored [4]. Recent work on hybrid neural generation approaches that use knowledge of sentence and discourse planning structures also seem promising [28, 42, 50].

5 CONCLUSIONS

In this paper, we describe some of the challenges we encountered building SlugBot, an open domain conversational agent funded by the Amazon Alexa Prize. We have introduced more problems than we have solved, and we have attempted to support our hypothesis that we need richer models of discourse coherence and discourse semantics to allow a conversational agent to take the initiative in open domain conversations. We illustrated how search and structured information can be combined in order for SlugBot to find content to use to take the initiative and respond to the user’s utterances. We propose a hybrid approach for language generation that which combines templates to generate responses with sentence selection from search, through ranking, despite having structured data.

REFERENCES

[1] Mixed-initiative interaction. IEEE Intelligent Systems, 14(5):14–23, September 1999.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference, ISWC’07/ASWC’07, pages 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] Dan Bohus and Alexander J. Rudnicky. Ravenclaw: dialog management using hierarchical task decomposition and an expectation agenda. In INTERSPEECH. ISCA, 2003.
Jerry R. Hobbs. Topic drift. In Bruce Dorval, editor, J. Li and D. Jurafsky. Neural Net Models for Open-Domain Discourse Coherence.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: Modeling unrestricted coreference on onomotones. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–27. Association for Computational Linguistics, 2011.

R. Prasad, N. Dinesh, A. Lee, E. Mitsakakis, L. Rohalbo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. In Proc. of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pages 2961–2968, 2008.

Ettel M. Prince and Marilyn Walker. A statistical approach to givenness: a hearer-status algorithm and a centering algorithm. In Proc. of the 4th International Pragmatics Conference. Benjamin, 1993.

Crapnell Radlinski and Nick Crapnell. A theoretical framework for conversational search. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR ’17, pages 117–126, New York, NY, USA, 2017. ACM.

Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Edrad Kuzey, and Gerhard Weikum. YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames, pages 177–185. Springer International Publishing, Cham, 2016.

Marta Recasens. Coreference: Theory, Annotation, Resolution and Evaluation. PhD thesis, Universitat de Barcelona, 2010.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. The life and death of discourse entities: Identifying singleton mentions. In HLT-NAACL, pages 627–633, 2013.

Giuseppe Riccardi, Evgeny A Stepannot, and Shammar Abas Chowdhury. Discourse connective detection in spoken conversations. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pages 6095–6099. IEEE, 2016.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Spering. Twitternet: News in tweets. In Proceedings of the 17th ACM SIGKDD International Conference on Advances in Geographic Information Systems, GIS ’09, pages 49–58, New York, NY, USA, 2009. ACM.

Schema.org. Schema.org, http://www.schema.org/.

Amanda Stent. Rhetorical structure in dialog. In Proceedings of the first international conference on Natural language generation—Volume 14, pages 247–252. Association for Computational Linguistics, 2000.

Amanda Stent and Martín Molina. Evaluating automatic extraction of rules for sentence plan construction. In Proc. of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 299–297, 2009.

Amanda Stent, Marilyn Walker, Steve Whittaker, and Preetam Maloor. User-tailored generation for spoken dialogue: An experiment. In ICSLP, 2002.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Arvind K. Joshi. Annotation of discourse relations for conversational spoken dialog. In LREC, 2010.

Princeton University. About Wordnet. http://wordnet.princeton.edu, 2010.

M. Walker and R. Passonneau. DATE: A dialogue act tagging scheme for evaluation. In Human Language Technology Conference, 2001.

M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potaminas, R. Passonneau, S. Roukos, G. Sanders, S. Sennoff, and D. Stallard. DARPA communicator: Cross-system results for the 2001 evaluation. In ICSLP 2002, 2002.

M. A. Walker, I. Langelie–Geary, H. Wright Hastie, J. Wright, and A. Gorin. Automatically Training a Problematic Dialogue Predictor for a Spoken Dialogue System. Journal of Artificial Intelligence Research, 16:293–319, 2002.

Marilyn Walker, Julie Boland, and Candace Kann. The utility of elapsed time as a usability metric for spoken dialogue systems. In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU’39/99, 1999.

Marilyn A. Walker. Limited attention and discourse structure. Computational Linguistics, 22:2:255–264, 1996.

Marilyn A. Walker. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. Journal of Artificial Intelligence Research, 12:387–416, 2000.

Marilyn A. Walker, Candace A. Kann, and Diane J. Litman. Towards developing general models of usability with PARADISE. Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems, 2000.

Marilyn A. Walker, Diane Litman, Candace Kann, and Alicia Abella. Evaluating interactive dialogue systems: Extending component evaluation to integrated system evaluation. In Proc. of the ACL-EACL Workshop on Interactive Spoken Dialogue Systems, 1997.

Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In Proc. of the Meeting of the Association for Computational Linguistics, ACL 2001, 2001.

Marilyn A. Walker, Bernard Graziano, François Mairese, and Rashmi Prasad. Individual and domain adaptation in sentence planning for dialogue. Journal of Artificial Intelligence Research (JAIR), 30:413–456, 2007.

Marilyn A. Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In Proc. 28th Annual Meeting of the ACL, pages 70–79, 1990.

Wikipedia. Wikipedia. https://www.wikipedia.org/.

Zhou Yu, Ziyu Xu, Alan W. Black, and Alexander J. Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In 5th Annual Meeting of SIGDIAL, pages 404–412. The Association for Computer Linguistics, 2016.