A New Corpus for Context-Dependent Semantic Parsing

Andreas Vlachos
Computer Laboratory
University of Cambridge
av308@cam.ac.uk

Stephen Clark
Computer Laboratory
University of Cambridge
sc609@cam.ac.uk

Abstract

Semantic parsing is the task of translating natural language (NL) utterances into a machine-interpretable meaning representation (MR). Most approaches to this task have been developed and evaluated on a small number of existing corpora. While these corpora have made progress in semantic parsing possible, most of them cover rather narrow domains and context is rarely considered. In this paper we present a new set of guidelines for context-dependent semantic parsing and describe the annotation of a semantic parsing corpus. This new corpus covers a wider domain, namely tourism-related activities in a city, and consists of 17 dialogs containing 2374 user utterances.

1 Introduction

Semantic parsing is the task of translating natural language (NL) utterances into a machine-interpretable meaning representation (MR). Progress in semantic parsing has been greatly facilitated by the existence of corpora containing NL utterances annotated with MRs, the most commonly used corpora being ATIS (Dahl et al., 1994) and GeoQuery (Zelle, 1995). However, these existing corpora have some important limitations. In most cases each utterance is interpreted in isolation, i.e. independently of its context. Thus, utterances that use coreference or whose semantics are context-dependent are typically ignored. Another limitation of existing corpora is that they cover narrow domains such as U.S. geography questions or flight reservations, and hence the types of entities encountered and the interactions among them are relatively limited. Furthermore, the MRs are typically restricted to some form of database query; thus they cannot represent meanings which are not related to the database under consideration.

In this paper we present a new MR language (MRL) and a new corpus annotated with it. The proposed MRL covers the domain of tourism-related activities in a city, such as navigation and information requests. It can handle dialog context such as coreference and can accommodate utterances that are not interpretable according to a database. The utterances were collected in experiments with human subjects, and thus contain phenomena such as ellipsis and disfluency. We developed guidelines and annotated 17 dialogs containing 2374 utterances, and conducted what we believe is the first inter-annotator agreement study on semantic parsing annotation, reporting 0.829 exact match agreement between two annotators. We argue that this new, publicly available corpus[1] is likely to be more challenging than existing corpora as it covers a wider domain and the MRL uses a larger controlled vocabulary.

2 Commonly Used Corpora

GeoQuery (Zelle, 1995) is the most commonly used semantic parsing corpus. It consists of questions about U.S. geography annotated with expressions in a Prolog-style MRL. The questions are independent; hence the MRL does not need to capture context-dependent semantics. Furthermore, the database supporting the dataset contains only about 800 facts. The relative simplicity of GeoQuery, combined with improvements in semantic parsing methods, has resulted in accuracies for supervised (Kwiatkowski et

[1] The corpus can be downloaded from https://sites.google.com/site/andreasvlachos/resources.

al., 2011) and response-based (Liang et al., 2011) methods exceeding 90% on the English language version of the dataset. Andreas et al. (2013) showed that simplifying the MRL to a sequence of predicates and treating GeoQuery as a machine translation task results in near state-of-the-art performance.

Another commonly used corpus is the airline travel information system (ATIS) corpus (Dahl et al., 1994). It consists of dialogs between a user and a flight booking system collected in Wizard-of-Oz experiments (Kelley, 1983) in which users seeking information on flights interacted with a human performing the role of an automated booking system (the "wizard"). Each utterance is annotated with the SQL statement that would return the requested piece of information from the flights database. The utterance interpretation is context-dependent. However, the original SQL annotation is rarely considered, and the only two studies which have considered the context-dependent parts are Miller et al. (1996) and Zettlemoyer and Collins (2009). Instead, most studies focus on the context-independent utterances, on which exact match accuracies have exceeded 0.8 (Kwiatkowski et al., 2011).

Other commonly used corpora include TownInfo (Mairesse et al., 2009), Jobs640 (Tang and Mooney, 2001) and RoboCup (Kuhlmann et al., 2004). However, none of them requires context-dependent utterance interpretation.

3 Meaning Representation Language

Our proposed MRL was designed in the context of a portable, interactive navigation and exploration system The users of this system — typically tourists visiting a city — through which users can obtain information about places and objects of interest, such as monuments and restaurants, as well as directions (see sample dialog in Fig. 1). The system is aware of the position of the user (through the use of GPS technology) and is designed to be interactive; hence it can initiate the dialog by offering information on nearby points of interest and correcting the route taken by the user if needed. A first version of the system was described and evaluated by Janarthanam et al. (2013).

The purpose of the proposed MRL and corpus is to facilitate the development and evaluation of a semantic parser for this application. The MRs returned by the semantic parser must represent the user utterances adequately so that the system can generate the appropriate response. Thus the MRL must be able to abstract over multiple ways of expressing meanings that are considered the same w.r.t. the application.

The MRL uses a flat syntax composed of elementary predications, based loosely on minimal recursion semantics (Copestake et al., 2005), but without an explicit treatment of scope. Each MR consists of a dialog act representing the overall function of the utterance, followed for some dialog acts by an unordered set of predicates. All predicates are implicitly conjoined and the names of their arguments specified to improve readability and to allow for some of the arguments to be optional. The argument values can be either constants from the controlled vocabulary, verbatim string extracts from the utterance (enclosed in quotes) or variables (`Xno`). Negation is denoted by a tilde (`~`) in front of predicates. The variables are used to bind together the arguments of different predicates within an utterance, as well as to denote coreference across utterances.

Dialog acts The dialog acts are utterance-level labels which capture the overall function of the utterance in the dialog, for example whether an utterance is a statement of information, an acknowledgement, or a repetition request (`inform`, `acknowledge` and `repeat` in Figure 1). The dialog acts in the MRL are divided into two categories. The first category contains those that are accompanied by a set of predicates to represent the semantics of the sentence, such as `set_question` and `inform`. For these acts we denote their focal points — for example the piece of information requested in a `set_question` — with an asterisk (`*`) in front of the relevant predicate. The focal point together with the act provide similar information to the intent annotation in ATIS (Tur et al., 2010). The second category contains dialog acts that are not accompanied by predicates, such as `acknowledge` and `repeat`. These are used to annotate utterances whose function in the dialog is clear and simple, even if their actual semantics might be rather complex and possibly beyond what the controlled vocabulary of the MRL.

USER what's the nearest italian, em, for a meal?

```
dialogAct(set_question)
*isA(id:X1, type:restaurant)
def(id:X1)
hasProperty(id:X1, property:cuisine,
            value:"italian")
distance(location:@USER,
         location:X1, value:X2)
argmin(argument:X1, value:X2)
```

WIZARD vapiano's.

```
dialogAct(inform)
isA(id:X4, type:restaurant)
*isNamed(id:X4, name:"vapiano's")
equivalent(id:X1, id:X4)
```

USER take me to vapiano!

```
dialogAct(set_question)
*route(from_location:@USER,
       to_location:X4)
isA(id:X4, type:restaurant)
isNamed(id:X4, name:"vapiano")
```

WIZARD certainly.

```
dialogAct(acknowledge)
```

WIZARD keep walking straight down clerk street.

```
dialogAct(instruct)
*walk(agent:@USER, along_location:X1,
      direction:forward)
isA(id:X1, type:street)
isNamed(id:X1, name:"clerk street")
```

USER yes.

```
dialogAct(acknowledge)
```

USER what is this church?

```
dialogAct(set_question)
*isA(id:X2, type:church)
index(id:X2)
```

WIZARD sorry, can you say this again?

```
dialogAct(repeat)
```

USER i said what is this church on my left!

```
dialogAct(set_question)
*isA(id:X2, type:church)
index(id:X2)
position(id:X2, ref:@USER,
         location:left)
```

WIZARD it is saint john's.

```
dialogAct(inform)
isA(id:X3, type:church)
*isNamed(id:X3, name:"saint john's")
equivalent(id:X2, id:X3)
```

USER A sign here says it is saint mark's.

```
dialogAct(inform)
isA(id:X4, type:church)
*isNamed(id:X4, name:"saint mark's")
equivalent(id:X2, id:X4)
```

Figure 1: Sample dialog annotated with MRs

Predicates The MRL contains predicates to denote entities, properties and their relations:

- Predicates introducing entities and their properties: `isA`, `isNamed` and `hasProperty`.
- Predicates describing user actions, such as `walk` and `turn`, with arguments such as `direction` to express different modes of action.
- Predicates describing geographic relations, such as `distance`, `route` and `position`. The latter uses the argument `ref` in order to denote relative positioning.
- Predicates denoting whether an entity is introduced using a definite article (`def`), an indefinite (`indef`) or an indexical (`index`), which are useful in determining which real-world entity is being referred to.
- Predicates expressing numerical relations such as `argmin` and `argmax`, which are used to denote superlatives.

Coreference In order to model coreference we adopt the notion of discourse referents (DRs) and discourse entities (DEs) from Discourse Representation Theory (DRT) (Webber, 1978; Kamp and Reyle, 1993). DRs are referential expressions appearing in utterances which denote DEs. DEs are mental entities in the speaker's model of discourse (which do not necessarily correspond to real-world entities). Note also that multiple DEs can refer to the same real-world entity; for example, in Figure 1 "vapiano's" refers to a different DE from the restaurant in the previous sentence ("the nearest italian"), even though they are likely to be the same real-world entity. We considered DEs instead of actual entities in the MRL because they allow us to capture the semantics of interactions such as the last exchange between the wizard and user, in which the disagreement would not have been possible to represent without using DEs. The MRL represents multiple DEs referring to the same real-world entity through the predicate `equivalent`(see e.g. the first wizard utterance in Figure 1).

Coreference is indicated by using identical variables across predicate arguments within an utterance or across utterances. The main principle in determining whether DRs corefer is that it must be possible to infer this from the dialog context alone, with-

out using world knowledge. For example, in Figure 1 "vapiano's" and "vapiano" are assumed to refer to the same DE even though they are different names, because it is clear from the dialog that the user is referring to the DE mentioned by the wizard. Compared to the coreference annotation in ATIS in SQL, our approach avoids repeating the MR of previous utterances, thus resulting in shorter forms that are likely to align better with NL utterances.

4 Data Collection

The NL utterances were collected using Wizard-of-Oz experiments with pairs of human subjects. In each experiment, one human pretended to be a tourist visiting Edinburgh (by physically walking around the city), while the other performed the role of the system using a suitable text-to-speech interface. Each user-wizard pair was given one of two scenarios involving requests for directions to different points of interest and information about them, as well as the system offering information considered of interest. Each experiment formed one dialog which was manually transcribed from recorded audio files. 17 dialogs were collected in total, seven from the first scenario and 10 from the second.

Each scenario was designed to last approximately one hour, but the actual execution time (and number of utterances collected) varied in each experiment depending on the amount of interaction between the user and the wizard. The users were encouraged to ask for information according to their interests, which resulted in a wide range of tourism-related discussions, such as foreign currency exchange rates and architecture. Furthermore, allowing the wizards to answer in natural language instead of restricting them to responding via database queries as in ATIS led to more varied dialogs. However, this also resulted in some of the user requests not being within the scope of the system. Furthermore, the proposed MRL has its own limitations; e.g. it does not have predicates to express temporal relationships. Therefore, we filtered the utterances collected [2] and decided not to annotate those falling into the following categories:

[2] A similar filtering process was used for GeoQuery (Section 7.5.1 in Zelle (1995)) and ATIS (principles of interpretation document (`/atis3/doc/pofi.doc`) in the NIST CDs).

	new corpus	GeoQuery	ATIS
user utterances	2374	880	5871
utterances/dialog	139.7	1	8.8
unique NL words	896	280	611
MRL vocabulary	115	35	85

Table 1: Corpus comparison.

vocabulary type	number of terms
dialog acts	15
predicates	19
arguments	41
constants	9
entity types	26
properties	4

Table 2: MRL vocabulary used in the annotation

- Utterances that are not human-interpretable, e.g. utterances that were interrupted.
- Utterances that are human-interpretable but outside the scope of the system, e.g. exchange rates.
- Utterances that are within the scope of the system but too complex be represented by the MRL, e.g. utterances expressing temporal relations.

Note that when the core of an utterance can be captured adequately by the MRL, we opt to annotate it with an appropriate MR even if some of the entities mentioned are outside the scope of the system, or if some of the language used is too complex. For example, the final utterance in Figure 1 mentions a sign which is outside the scope of the system, but we still annotated the utterance since its interpretation with respect to the application is not affected. We argue that determining which utterances should be translated into MRs is an important subtask for real-world applications of semantic parsing and hence decided to keep these utterances in the corpus.

5 Annotation

The annotation was performed by one of the authors and a freelance linguist with no experience in semantic parsing. As well as annotating the user utterances, we also annotated the wizard utterances with dialog acts and the entities mentioned, as well as their names and `def`, `indef` and `index` predicates as they provide the necessary context to perform context-dependent interpretation. In practice, though, we expect this information to be used by a

natural language generation system to produce the system's response and thus be available to the semantic parser.

The total number of user utterances annotated was 2374, out of which 1906 were annotated with MRs, the remaining not translated due to the reasons discussed in Sec. 4. Tbl. 1 has more statistics, including a comparison with GeoQuery and ATIS. The number and types of the MRL vocabulary terms used appear in Tbl. 2. The new corpus has a larger controlled vocabulary, which is indicative of its wider domain and, although the new corpus has fewer utterances than ATIS, it has a larger NL vocabulary and the utterances themselves are more varied since they do not consist of database queries exclusively.

We assessed the quality of the annotation through an inter-annotator agreement study in which the two annotators annotated one dialog consisting of 510 utterances. Exact match agreement at the utterance level, which requires that the MRs by the annotators agree on dialog act, predicates and within-utterance variable assignment, was 0.829, which is a strong result given the complexity of the annotation task, and which suggests that the proposed guidelines can be applied consistently. We also assessed the agreement on predicates using F-score, which was 0.914.

Variable assignment was more challenging to assess (beyond exact match) since measuring the agreement between two annotations relies on measuring the identity between the variables assigned to different arguments rather than the variable names themselves. Since variable names cannot be taken into account, we decided to treat each variable as a cluster of argument slots and evaluate variable assignment as a clustering task. We chose to use information-theoretic clustering evaluation measures which avoid the problem of cluster mapping; in particular we used the adjusted mutual information (AMI) measure (Vinh et al., 2010) as implemented by Pedregosa et al. (2011). AMI scores range from 0 to 1 and, unlike the more commonly used V-measure (Rosenberg and Hirschberg, 2007), are adjusted for chance, assigning scores close to 0 for random clusterings. The AMI was found to be 0.974 at the utterance level, while for the whole dialog it was 0.948. Note that the commonly used Kappa statistic (Carletta, 1996) could not have been used for any of the evaluations given above, since it can only be applied to classification tasks.

6 Conclusions

In this paper we presented a new MRL which covers the domain of tourism-related activities and a new corpus annotated with it. The proposed MRL can handle dialog context such as coreference and can accommodate utterances that are not interpretable according to a database. The annotated corpus consists of 17 dialogs containing 2374 user utterances which were collected using Wizard-of-Oz experiments with human subjects. We conducted an inter-annotator agreement study and found 0.829 exact match agreement between two annotators. As recent approaches to semantic parsing have achieved rather high accuracies on existing corpora, we believe the new corpus will be a useful resource in making further progress thanks to its wider domain, greater variety of utterances and longer average dialog length.

Acknowledgements

Andreas Vlachos is funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270019 (SPACEBOOK project `www.spacebook-project.eu`). The authors would like to thank Diane Nicholls for her efforts and feedback in corpus annotation and Robin Hill for the data collection, as well as all SpaceBook project participants for their feedback.

References

Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (short papers)*.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Ann Copestake, Dan Flickinger, Ivan Sag, and Carl Pollard. 2005. Minimal recursion semantics: An introduction. *Research in Language and Computation*, 3(2–3):281–332.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In *Proceedings of the Workshop on Human Language Technology*, pages 43–48, Plainsboro, New Jersey.

Srinivasan Janarthanam, Oliver Lemon, Phil Bartie, Tiphaine Dalmas, Anna Dickinson, Xingkun Liu, William Mackaness, and Bonnie Webber. 2013. Evaluating a city exploration dialogue system with integrated question-answering and pedestrian navigation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1660–1668, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

John F. Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 193–196.

Gregory Kuhlmann, Peter Stone, Raymond J. Mooney, and Jude W. Shavlik. 2004. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *The AAAI-2004 Workshop on Supervisory Control of Learning and Adaptive Systems*, San Jose, California.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical generalization in CCG grammar induction for semantic parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1512–1523, Edinburgh, UK.

Percy Liang, Michael Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 590–599, Portland, Oregon.

François Mairesse, Milica Gasic, Filip Jurcícek, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2009. Spoken language understanding from unaligned data using discriminative classification models. In *Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4749–4752, Taipei, Taiwan.

Scott Miller, David Stallard, Robert Bobrow, and Richard Schwartz. 1996. A fully statistical approach to natural language interfaces. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 55–61, Santa Cruz, California.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420, Prague, Czech Republic.

Lappoon R. Tang and Raymond J. Mooney. 2001. Using multiple clause constructors in inductive logic programming for semantic parsing. In *Proceedings of the 12th European Conference on Machine Learning*, pages 466–477, Freiburg, Germany.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What's left to be understood in ATIS? In *IEEE Workshop on Spoken Language Technologies*.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854.

Bonnie Lynn Webber. 1978. *A Formal Approach to Discourse Anaphora*. Ph.D. thesis, Harvard University.

John M. Zelle. 1995. *Using Inductive Logic Programming to Automate the Construction of Natural Language Parsers*. Ph.D. thesis, Department of Computer Sciences, The University of Texas at Austin.

Luke S. Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 976–984, Singapore.