Evaluating Supervised Semantic Parsing Methods on Application-Independent Data

Sebastian Beschke

Natural Language Systems Division
Department of Informatics
University of Hamburg, Germany
beschke@informatik.uni-hamburg.de

Abstract. While supervised statistical semantic parsing methods have received a good amount of attention in recent years, this research has largely been done on small and specialized data sets. This paper introduces a work-in-progress with the objective of examining the applicability of supervised statistical semantic parsing to application-independent data with linguistically motivated meaning representations. The approach discussed in this paper has three key aspects: The circumvention of data scarcity using automatic annotation, experimentation with different types of meaning representations, and the design of a suitable graded evaluation measure.

1 Introduction

We understand semantic parsing to be the task of extracting a formal meaning representation (MR) from a natural language text. Supervised statistical methods of semantic parsing are a research topic to which various approaches and formalisms have been applied over the past years. Evaluation of these methods has generally been performed on small data sets from very limited and application-specific domains. One example is Geoquery, a widely used corpus for natural language database queries on US geography [1]. A prime reason for the focus on small data sets is that the annotation of training data with full semantic MRs is laborious. These representations are even more complex than data used for many other tasks in statistical natural language processing. Therefore, fully annotated data has so far been scarce and mostly limited to application-specific data.

There is however mounting interest in application-independent semantic analysis. This task entails the creation of linguistically motivated MRs that attempt to represent certain linguistic features as completely as possible, as opposed to application-specific types of MR that only capture the amount of information that is needed for the application at hand. A prominent rule-based system performing this task is Boxer [2], while Le and Zuidema recently presented a statistical approach [3]. Both of these systems are based on Discourse Representation Theory [4].

M. Colinet et al. (Eds.): ESSLLI 2012/2013, LNCS 8607, pp. 19–25, 2014.

(a) answer(count(river(loc_2(stateid('california')))))
(b) answer(A,count(B,(river(B),loc(B,C),const(C,stateid(california))),A))
Give me the number of rivers in California.

Fig. 1. The Geoquery corpus contains two styles of meaning annotations: (a) variable-free expressions, and (b) Prolog-style expressions with variables. The meaning representations correspond directly to database queries and only contain enough information to perform the task of question answering. Linguistic details that are irrelevant to this task are not represented.

some(A,some(B,some(C,and(not(some(D,and(n12thing(D),not(r1after(A,D))))),
and(r1patient(A,B),and(r1agent(A,C),and(v1demand(A),and(n1solution(B),
and(a1global(B),and(n1problem(C),a1global(C))))))))))
After all, global problems demand global solutions.

Fig. 2. An example of the type of meaning representation created by Boxer. As Boxer's meaning representations aim to address a wider range of linguistic phenomena, they tend to be more comprehensive than typical Geoquery representations. As an example, consider the use of Neo-Davidsonian event semantics, with explicit agent and patient relations, which provides greater flexibility for semantic analysis but leads to an increase of the meaning representation size.

While the methods used for supervised semantic parsing (SSP) are in principle applicable to application-independent data, it is important to note the different characteristics of the data. While application-specific corpora such as Geoquery tend to exhibit low linguistic variability and complexity (such as consisting only of questions with short average sentence lengths), application-independent data from more open domains, such as newswire, is likely to contain longer, more varied sentences. In addition, as linguistically motivated MRs attempt to encode meaning as fully as possible, they also tend to be more complex than special-purpose MRs, which only encode information important to the application at hand. This dual increase in complexity is illustrated in Figures 1 and 2, and can also be witnessed by comparing the (application-specific) Geoquery corpus to the (application-independent) Groningen Meaning Bank [5]. It is not yet well understood how well the established SSP methods scale up to this type of data.

For this reason, we propose an experiment designed to help better understand how SSP generalizes to application-independent data. Its key aspects are the use of automatic annotation to generate open-domain test data (Section 2), experimentation on how the complexity of MRs can be adjusted to balance the expressiveness of the MR against the capabilities of the learning algorithm (Section 3), and the design of a graded measure to evaluate the performance of an SSP system (Section 4). We also present some thoughts on the possible learning framework to be used (Section 5). The paper closes with a brief discussion (Section 6).

2 Automatic Annotation

The scarcity of corpora annotated with deep semantic representations has been a significant limit for SSP research. The widely used Geoquery corpus [1] with its 880 sentences is both small in size and narrow in scope. The same applies to most other data sources used in SSP research so far.

An important recent development in this area is presented by the Groningen Meaning Bank (GMB) [5]. Its current 2.1.0 release consists of 8,000 texts with over 1 million tokens, which are annotated in Discourse Representation Theory [4]. The annotations are first created automatically by a tool pipeline and then refined by human annotators, including both experts and non-experts, wherein gamification is employed to allow the latter to contribute their linguistic knowledge [5]. The GMB is not limited to a specific domain, containing Voice of America newswire texts, country descriptions from the CIA Factbook, texts from the Open ANC [6], and Aesop's fables. As such, it is likely to become an important data source for future SSP efforts that take an open domain approach. In fact, one such effort has already been presented [3].

While the GMB thus seems to be a very suitable data source for experiments in SSP, it also has a few drawbacks. Importantly, the linguistic complexity and average sentence length of the texts is quite high, especially when compared with special-purpose corpora such as Geoquery. This might pose problems when working with algorithms whose computational performance is not yet up to par. In addition to the GMB, we therefore plan to use data annotated using the semantic parsing tool Boxer, which is also being used in the preparation of the GMB [5]. Manual inspection suggests that the MRs generated by Boxer are of sufficient quality to serve as training material for SSP systems. This allows any corpus to be used as training data, given that it can be automatically annotated. In this way, we are able to vary the training data's complexity as seems appropriate.

Automatically generated annotations are likely to be flawed. We do not suggest that training SSP models using automatic annotation will yield systems of the highest quality. Automatic annotation should rather be seen as a crutch in developing SSP methods, which will hopefully become unnecessary as more varied training data become available.

3 Experimentation on Meaning Representations

An important open question in SSP is which type of MR is most beneficial to the task. As an example, the Geoquery corpus is annotated using two distinct types of MR: variable-free functional expressions, and Prolog-style expressions using variables (see Fig. 1). While there is of course an interaction between the type of MR and the learning algorithm used in a specific system configuration, most SSP systems are designed to be somewhat independent of the MR formalism. This allows us to study this interaction experimentally.

Some of the current SSP systems can process only variable-free forms (such as [7]), while others can process both types of MR (such as [8]). As most semantic formalisms, including Discourse Representation Theory, rely crucially on

variables (or, put differently, graphical structures such as those used in [3]), our preference should be on the latter type of learning framework. However, there is also recent work on the design of variable-free MRs with the same expressivity as lambda-calculus forms [9,10]. There are also underspecified semantic formalisms such as Lexical Resource Semantics [11]. Converting meaning representations into alternative formalisms would allow comparing these formalisms from the point of view of SSP performance.

Besides conversion to other formalisms, another likely way to improve SSP performance is the simplification of MRs. By this, we mean modifications that do not necessarily preserve the full content of an MR, but in some way make it easier to process. For instance, the use of nested logical connectives and quantifiers imposes a structure on MRs with which learning algorithms might struggle, so removing some or all of these phenomena may yield representations that are easier to learn (this can also be thought of as a kind of underspecification). The idea is that even if we remove some information from the MR, there may still be enough information left to fulfill some useful purpose. Therefore, we plan to also examine the effect of this progressive degradation.

4 Evaluation of SSP Performance

So far, the performance of SSP systems has generally been measured in terms of "complete matches", i.e. either the complete construction of the correct MR by the SSP system, or the construction of an MR that yields the same result when executed [1]. However, with meaning representations that are longer and more complex, complete and exact reconstruction of MRs becomes increasingly unlikely. It is therefore desirable to assign partial credit even to imperfect MRs.

Ideally, we would like to compare two meaning representations in terms of the similarity of their meaning. Since such a notion is inaccessible even from a theoretical point of view, we are left with the choice of a suitable proxy [12]. Logical equivalence is an option, but still undecidable. For lack of alternatives, we therefore decide to state a similarity measure for a pair of meaning representations in purely syntactic terms.

It seems natural to use a measure that exploits the graphical nature of MRs by searching for a node-to-node assignment between gold-standard annotation and SSP output. In fact, [13] presents such a measure, where an assignment's score is determined by matching node labels as well as the number of matching edges on nodes that are assigned to each other. The score is then defined to be the highest score achieved by any assignment. In [3], a similar measure is introduced based on a maximum common subgraph alignment.

Instead of maximum common subgraph alignment, we have opted to adopt a measure based on solving an assignment, or bipartite matching, problem. As the underlying graphical structure, we use a syntax tree of the MR. The final score is made up of two components: a node score and a variable score. Both of them are determined by the weight of an optimal assignment of certain components of the MR under evaluation to their counterparts in the gold standard MR.

In the calculation of the node score, the inner nodes – i.e., predicate names, quantifiers, and logical connectives – are assigned to each other. A weight is calculated for each pair of a single node in the test MR and a node in the gold-standard MR, based on the following factors: whether the node types match (i.e. they represent the same predicate, quantifier, or connective), whether the parents' node types match, and whether their depth in the MR syntax tree is similar.

The variable score is derived from the best assignment between the variables in the two MRs, based on the following factors: whether the variables are bound by the same type of quantifier, whether the quantifier appears in the same polarity, and how many of their occurrences match regarding name of the predicate governing the occurrence, the argument place that is filled by the occurrence, and the polarity of the occurrence.

A combined score is then derived through the multiplication of the node and variable scores. It is 1 if the MR under test equals the gold-standard, and strictly less than 1 otherwise. From manual inspection we gather that the measure seems to reflect human judgement quite well, assigning high scores to MRs that contain large sub-structures of the gold-standard.

5 The Learning Framework

Initial experimentation with the two state-of-the-art SSP systems WASP [1] and UBL [8] has revealed, not surprisingly, that the application of SSP to larger and more complex data sets requires addressing computational issues first. It will therefore be necessary to produce an implementation of an SSP system that is capable of dealing with sufficient amounts of more complex data. While this problem has prompted Le and Zuidema to invent a completely new learning framework and underlying formalism [3], we instead plan to follow the line of work represented by Kwiatkowski et al. [8]. In addition to achieving state-of-the-art performance on the Geoquery data set, it is based on combinatory categorial grammar (CCG)[14], which has a solid foundation in linguistic theory. Additionally, the existence of the rule-based Boxer system, which is also based on CCG, suggests the suitability of CCG-based models for the task.

As it is common in CCG, meaning representations are constructed using lambda-calculus. This means that any MR formalism can be used as long as it supports this construction method. Of course, this is not to say that there were no interaction between the semantic parsing model, the mode of construction, and the MR formalism used. However, as we consider CCG-based models a promising approach to SSP, we think it makes sense to evaluate the various types of MR with regards to this type of model.

The main computational problem lies in searching the space of possible splits of meaning representations over CCG items. Kwiatkowski et al. address this by limiting the size of the portion of the meaning representation that is split off. However, this strategy proves too restrictive for the large meaning representations that are generated by Boxer. We suggest that heuristics may instead be used to define the space of splits that is searched. E. g., one plausible heuristic would place split points at the boundaries of constituents generated by an

external syntactic parser. This could be supplemented by a heuristic based on word-to-predicate alignment, similar to [3].

6 Discussion and Outlook

We have introduced a research project towards the extension of SSP methods to application-independent data. An important motivation is that we believe that the consideration of more complex data in SSP is crucial for its evolution to become a more general problem-solving tool. Being able to work with application-independent data means that costly annotation efforts do not need to be repeated for every potential application of semantic parsing. This will reduce the cost of exploring further potential applications.

To evaluate the applicability of a state-of-the-art semantic parsing algorithm to application-independent data, we performed a preliminary test using UBL and automatically annotated data. While annotated newswire texts proved computationally infeasible, we were able to run a test using the Geoquery dataset. The Geoquery sentences were annotated using Boxer, yielding MRs formulated in first-order logic that were considerably longer and more complex than the original Geoquery annotations. These annotations were recovered by UBL with F1-scores between 30% and 50%. Compared to the F1-score of 89% reported on the original annotations, these figures appear very low. However, we still consider this result encouraging considering that the amount of training data was very small, and that the re-annotation of the corpus increased the variance of the annotated MRs. The Geoquery corpus contains many sentences where different natural language formulations are used for expressing the same semantic content, which will however be assigned different MRs by Boxer. In addition, inspection of the parser output suggested that in some cases where MRs could not be exactly recovered, important MR components were nonetheless present.

As has already been detailed, computational issues need to be addressed when dealing with input data of higher complexity. Our current main concern is therefore the design of suitable algorithms, notably for the induction of CCGs for semantic parsing.

The results of this work will be beneficial to various endeavors related to SSP, such as improving existing SSP systems, developing new SSP methods, and applying SSP to other tasks in natural language processing. An example for such a task is the development of hybrid syntax/semantics-based machine translation systems.

References

1. Wong, Y.W., Mooney, R.: Learning synchronous grammars for semantic parsing with lambda calculus. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 960–967. Association for Computational Linguistics (June 2007)

2. Bos, J.: Wide-coverage semantic analysis with boxer. In: Semantics in Text Processing, STEP 2008 Conference Proceedings. Research in Computational Semantics, vol. 1, pp. 277–286. College Publications (2008)
3. Le, P., Zuidema, W.: Learning compositional semantics for open domain semantic parsing. In: Proceedings of COLING 2012, Mumbai, India, pp. 1535–1552. The COLING 2012 Organizing Committee (December 2012)
4. Kamp, H., Reyle, U.: From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer, Dordrecht (December 1993)
5. Basile, V., Bos, J., Evang, K., Venhuizen, N.: Developing a large semantically annotated corpus. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 3196–3200. European Language Resources Association (ELRA) (2012); ACL Anthology Identifier: L12-1299
6. Ide, N., Baker, C., Fellbaum, C., Passonneau, R.: The manually annotated subcorpus: A community resource for and by the people. In: Proceedings of the ACL 2010 Conference Short Papers, Uppsala, Sweden, pp. 68–73. Association for Computational Linguistics (July 2010)
7. Lu, W., Ng, H.T., Lee, W.S., Zettlemoyer, L.S.: A generative model for parsing natural language to meaning representations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 783–792 (2008)
8. Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., Steedman, M.: Inducing probabilistic CCG grammars from logical form with higher-order unification. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, pp. 1223–1233. Association for Computational Linguistics (October 2010)
9. Liang, P., Jordan, M., Klein, D.: Learning dependency-based compositional semantics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 590–599. Association for Computational Linguistics (June 2011)
10. Alshawi, H., Chang, P.C., Ringgaard, M.: Deterministic statistical mapping of sentences to underspecified semantics. In: Bos, J., Pulman, S. (eds.) Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), Oxford, UK, pp. 15–24 (2011)
11. Richter, F., Sailer, M.: Basic concepts of lexical resource semantics. In: Collegium Logicum. ESSLLI 2003 - Course Material I. Collegium Logicum, vol. 5, pp. 87–143. Kurt Gödel Society, Wien (2004)
12. Shieber, S.M.: The problem of logical-form equivalence. Computational Linguistics 19(1), 179–190 (1993)
13. Allen, J.F., Swift, M., de Beaumont, W.: Deep semantic analysis of text. In: Bos, J., Delmonte, R. (eds.) Semantics in Text Processing, STEP 2008 Conference Proceedings. Research in Computational Semantics, vol. 1, pp. 343–354. College Publications (2008)
14. Steedman, M.: The Syntactic Process. MIT Press, Cambridge (2000)