On the Way towards Standardized Semantic Corpora for Development of Semantic Analysis Systems

Ivan Habernal and Miloslav Konopík
Department of Computer Science and Engineering
University of West Bohemia in Pilsen, Univerzitní 8, 306 14 Pilsen, Czech Republic
E-mail: habernal@kiv.zcu.cz, konopik@kiv.zcu.cz

***Abstract*—One of the main means to achieve progress in science is cooperation. It is advantageous if the cooperation is carried among teams at different institutions. In semantics, the basic necessity for cooperation is a standardized annotated corpus. Such a corpus allows to share individual findings by the whole research community because then different systems can be tested under the same conditions. Unfortunately there is no standardized semantic corpus for the Czech language and many other languages suffer the same. Moreover the ATIS corpus set is more than ten years old and it does not meet today's trends in semantic annotation. In this article we summarize the problems of the ATIS corpora set as well as the problems encountered during our research. As a result, we provide a methodology to avoid such problems. For practical deployment of the methodology we offer a set of annotation tools. The purpose of this article is to discuss the problematic of semantic annotation and to gather other teams to create standardized shared semantic corpora.**

***Keywords*—semantic analysis; semantic corpus; ATIS.**

I. Introduction

The goal of a Spoken Language Understanding system (SLU) is to extract a meaning from natural speech. The SLU covers many subfields such as utterance classification, speech summarization, natural language understanding (NLU) and information extraction. In human-computer dialogue systems, the task of the SLU system is to process the input acoustic utterance and transform it into a semantic representation. However, this task can be split into two parts: *automatic speech recognition* (ASR) and *semantic analysis*. The purpose of a semantic analysis system is to obtain a context-independent (it depends neither on history nor context) semantic representation from a given input sentence.

There are two basic types of semantic representation: *logical structures* (e.g., First-order predicate calculus, Transparent Intentional Logic, SIL, etc.) [1], [2] and *"data" structures* (e.g., trees, frames, flat concepts, etc.) [3]. The logical structures are more suitable for complex representation of semantics while the "data" structures are better suited for automatic learning systems. The reason is that statistical learning algorithms are not capable of handling the complexity of logical structures. Our experiences with semantic analysis systems based upon logical structures [2] shown that practical deployment of such systems is complicated due to the need of creating rules manually. Therefore, we focus on automatic learning systems, that seem to be more convenient for practical applications. Hence, we have chosen the tree based semantic representation described, i.e., in [3] that was designed mainly for practical use.

During the development and testing of the system described in [4], we have used our own Czech semantic corpus [5]. However, the results are not comparable with other semantic analysis systems since most of them (e.g., [6], [7]) performed their tests on different corpora. The availability of commonly used semantic corpora is quite good for English – for example the ATIS corpus [8], which is a mixed corpus for both speech recognition and semantic analysis. The tests on this corpus were performed by many semantic analysis systems. However, there is a lack of a standard semantic corpus for the Czech language, which differs from English in many aspects (morphologically rich, free word-order, etc.).

This paper presents our proposal to start the process of creation of such a corpus. It takes into account all practical issues that a developer of a semantic analysis system must deal with. It also describes the set of tools and proposes formats of the data. In this article we focus on the Czech language but most of the principles are valid for other languages too.

II. Related work

A. ATIS corpus

One of the commonly used corpora for testing of semantic analysis systems in English is the ATIS corpus. It was used for evaluation in, e.g., [6], [9], [10] and [11]. The original ATIS corpus is divided into several parts, e.g., ATIS2 train, ATIS3 train, two test sets, etc. [8]. Unfortunately, the corpus is not directly suitable for semantic analysis system development or testing.

The two testing sets, `ATIS3 test dec94` (445 sentences) and `ATIS3 test nov93` (448 sentences), contain the annotation in the semantic frame format. Each sentence is labelled with a goal name and slot names with an associated content. The training sets `ATIS2 train` and `ATIS3 train` contain only SQL queries that carry the semantic information.

This brings the first practical issue: To obtain the training data, the queries must be converted back to a semantic representation (a semantic frame or an equivalent semantic description). The authors of [6] transformed the data semi-automatically into a format suitable for the HVS model. Their training data use a bracketing notation to express the concept hierarchy. However, a deep exploration of this data shows

 ISBN: 978-1-61208-104-5

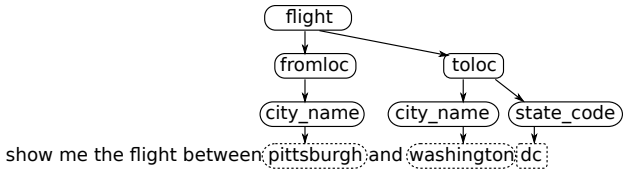

Fig. 1. An example of a semantic parse tree for a sentence from the ATIS corpus.

that a significant number of annotation break the conventions of the bracketing semantic annotation. The terminal semantic concepts (denoted as *lexical classes*) must be leaves of a semantic tree and are not allowed to contain any sub-tree (as shown in Fig. 1). However, in many cases the lexical classes act as superior concepts to other semantic concepts in the data. This inconsistency makes the data hard to use in other systems. This issue is, however, not caused by the ATIS corpus itself but re-creating the training data set from SQL queries probably always brings some sort of inconsistency.

Another issue is caused by inconsistencies in the testing data set. The following example shows a typical semantic frame for a flight query (this structure is equal to the semantic parse tree from Fig. 1).

```
show me the flight between pittsburgh and
washington dc
GOAL: FLIGHT
FROMLOC.CITY_NAME = pittsburgh
TOLOC.CITY_NAME = washington
TOLOC.STATE_CODE = dc
```

It has a very clear concept hierarchy. However, in the same testing set there also appears the following annotation:

```
what are the flight between dca and milwaukee
GOAL: FLIGHT
AIRPORT_CODE = dca
CITY_NAME = milwaukee
```

The semantic content of this sentence is rather similar to the previous one but the semantic frame is significantly different: In the second example, the concepts `AIRPORT_CODE` and `CITY_NAME` are directly inferior to the main concept (goal) without distinguishing which one is `FROMLOC` and `TOLOC`. Thus, the proper semantic frame should contain `FROMLOC.AIRPORT_CODE = dca` and `TOLOC.CITY_NAME = milwaukee`.

Another problem is how to deal with the annotation which has multiple goals. In the testing set there are about 20 sentences with two goals. The semantic interpretation part of a system (which is, in ATIS, a SQL query producer) should probably restrict the output of semantic analysis so that only one goal is allowed. Among others, there is also one typo in the testing data.

This brings two important questions: Were the testing sentences annotated according to any scheme? And how strictly was the testing set checked in a sense of inter-annotation agreement and correct semantic description?

B. Czech Semantic Corpora

Since the Czech language is morphologically rich and has a relatively free word order, it is not correct to directly adapt a semantic analysis system which is developed using an English corpus, and, obviously, a Czech semantic corpus is required. When searching for an existing suitable Czech corpus for the semantic analysis task, two significant projects must be mentioned.

The Prague Dependency Treebank (PDT 2.0) is a large corpus with morphological, syntactic and semantic (tectogrammatical) annotation. The methodology of adding the semantic layer to the PDT is described in [12]. The semantic representation formalism is based upon semantic networks and the tectogrammatical layer partially depends on syntax [13]. The tectogrammatical annotation provides a deep-syntactic (syntactical-semantic) analysis of the text. The formalism abstracts away from word order, function words (syn-semantic words), and morphological variation [14].

The DESAM corpus introduced in [15] was annotated with lemmas and gramatical categories. Subsequently, it was enriched with the semantic annotation [16]. The grammatical tagging was taken as a base and some tags were relabeled as semantic and pragmatic. The article [17] presents an attempt to combine Transparent Intensional Logic framework (which is used for capturing the semantics) with lexical units. Later, the semantic network (Czech WordNet) was enriched using morphological derivations [18].

However, the above mentioned corpora and related projects attempt to cover the semantics in a complex manner and are designed to act as a general description of semantics, in opposite to a task-oriented corpus such as ATIS.

Authors of [7] developed an extended HVS semantic parser (based on [6]) using a Human-Human Train Timetable Dialogue Corpus [19]. The corpus is annotated at multiple levels (dimensions) where the semantic dimension uses the same abstract annotation methodology as used in [6]. The corpus contains 1109 semantically annotated dialogues.

III. Standard Czech Semantic Corpus Requirements

A. The Task Definition

One of the main purposes of this paper is to inform and get the NLP and semantic analysis community involved into our task. It can be stated as: Creating a Czech semantic corpus, which will be publicly available, with clear and sufficiently universal semantic annotation structure, which is not limited to any domain. The corpus is not intended to describe the semantics as complex as presented in Section II-A but it should be strictly task-oriented, facing the practical issues that can arise during semantic analysis system development. Moreover, it will improve cooperation among the working groups focused on semantic analysis and will allow an objective comparison of the results.

 ISBN: 978-1-61208-104-5

B. Proposed Process Description

The proposed process workflow will consist of the following steps: First, a suitable text dialog corpus must be obtained. This can be based upn a part of the corpus presented in [5]. Second, an eligible semantic representation should be chosen. We discuss it in Section III-D. Third, the data will be annotated using semi-supervised learning and supporting tools presented in Section IV. Finally, to avoid the shortcomings that are for instance pointed out for the ATIS corpus, the annotated data will be manually validated.

C. Previous Work

Our attempt to create a semantically annotated corpus is presented in [5]. The semantic representation used in this corpus is based upon abstract semantic annotation from [6]. The corpus contains written user queries in natural language entered into an intelligent web search engine. A selected part of this data can be used as a basic set for the standard Czech semantic corpus.

D. Semantic Representation

To describe the semantics of an utterance, many task-oriented semantic analysis systems (e.g., [3], [7], [23], etc.) use some formats of the frame-based structure, as shown in the ATIS example. This simple formalism offers a very clear hierarchy of semantic concepts (a semantic tree), including the lexical realizations of the lexical classes. The name *lexical class* comes originally from [6], it can be also denoted as *named entity*, etc. It is a leaf of a semantic tree and covers one or more words with a specific meaning, such as names, dates, numbers, etc.

After considering the possible issues described in II-A, our previous corpus annotation effort and semantic analysis system development was supported by using an *annotation scheme*. The annotation scheme is a hierarchical structure (a tree) that defines a dominance relationship among concepts, theme (also called *goal* in ATIS or *topic*); this is the root semantic concept of the sentence. and lexical classes. It says which concepts can be associated with which super-concepts, which lexical classes belong to which concepts, and so on.

The annotation scheme should cover the entire domain we want to annotate. Subsequently, each sentence is annotated according to the scheme. The existence of such a scheme assures that two sentences with similar semantic content (meaning) will have the same semantic representation (see II-A). Apparently, this feature is crucial for further semantic interpretation.

However, the beforementioned annotation consistency using an annotation scheme is always limited to the covered domain. Althought this is not an issue for developers of a particular semantic analysis system, it does not allow to easily extend and evolve the scheme in the future together with maintaining the semantics of the annotation. Thus, it can be also considered to use more general formalism for describing semantics, i.e., RDF/OWL. Using this formalism, the corpus can be more easily aligned to other ontologies and then used in other semantic analysis systems with arbitrary semantic annotations. Furthermore, RDF/OWL has the same ability to prevent the annotators from creating malformed annotation as the annotation scheme which has proven to be essential for semantic corpus development [4].

IV. Supporting Tools

To improve the efficiency of the annotation [5] and to facilitate the corpus processing and sharing, supporting tools are required. We have developed a complete set of software covering the data acquisition, dialog act annotation and segmentation, semantic annotation and annotation management.

The first step of the data processing is conversion of a plain text into a format suitable for further annotation. This includes the text tokenization and morphological analysis (obtaining the morphological tags and the most probable lemma) using PDT 2.0. For this task, a web service has been developed and deployed.

The dialogue act segmentation is processed by the *dialogue act editor*. The output of dialogue act segmentation is then imported into the *abstract annotation editor*. The editor supports an advanced annotation methodology based upon automatic lexical class identification and bootstrapping. Both programs are GUI applications written in Java. The usability and efficiency of the tools has been presented in [20].

The *annotation manager* software helps to deal with an extensive semantic data. Some selected features are: A distribution of the sentences for the annotation among the annotators; annotation merging including conflict checking; various statistics (corpus statistics, annotation statistics, inter-annotation agreement, and annotator statistics). Again, this is a GUI Java application (see Figure 2).

All presented software tools are licenced under GPL licence and are publicly available from `http://liks.fav.zcu.cz`. At the same web page you can find information about the current state of the corpus, join the e-mail conference and get involved into the process of creating the standard Czech semantic corpus.

V. Conclusions

In this article, we have proposed an activity to create a standardized semantic corpus. We discussed issues that are connected with the annotation process of such a corpus. The basic parameters of the corpus to be created were described together with the process of how to create it. The expected impact of this article is to open a discussion of measuring the performance of systems for semantic analysis so that the results have an informative value.

Many recent articles about semantic analysis (e.g., [21], [22], [23], including ours) were published with the results measured on a private corpus. Our effort is to change this state by introducing a standardized semantic corpus. In order to be successful we, however, need a broad agreement on the details of the corpus to be created. That prevents us from creating such a corpus by ourselves and forces us to publish a work in progress.

 ISBN: 978-1-61208-104-5

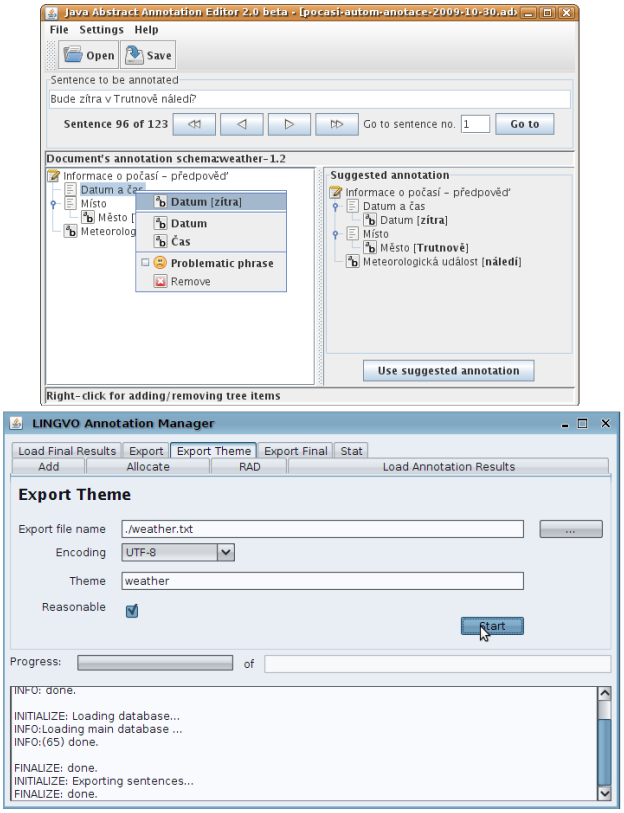

Fig. 2. A screenshots of the annotation editor and the annotation manager.

ACKNOWLEDGEMENTS

This work was supported by grant no. 2C06009 Cot-Sewing.

REFERENCES

[1] Allen J.: Natural Language Understanding, Benjamin/Cummings, Redwood City, CA, 1995.

[2] Ocelíková J., Matoušek V., and Krutišová J.: Design and implementation of a dialog manager. In: Text, Speech, Dialogue: Proceedings of the First Workshop on Text, Speech, Dialogue - TSD'98. Brno, Masaryk university. ISBN 80-210-1899-2, p. 257-262, 1998

[3] Young S.: The Statistical Approach to the Design of Spoken Dialogue Systems. Tech Report CUED/F-INFENG/TR.433, Cambridge University Engineering Department, 2002.

[4] Konopík M. and Habernal I.: Hybrid Semantic Analysis. In Proceedings of the 12th international Conference on Text, Speech and Dialogue. Lecture Notes In Artificial Intelligence, vol. 5729. Springer-Verlag, 2009, Berlin, Heidelberg, 307-314. ISBN 978-3-642-04207-2.

[5] Habernal I. and Konopík M.: Semantic Annotation for the LingvoSemantics Project. In Proceedings of the 12th international Conference on Text, Speech and Dialogue. Lecture Notes In Artificial Intelligence, vol. 5729. Springer-Verlag, 2009, Berlin, Heidelberg, 299-306. ISBN 978-3-642-04207-2.

[6] He Y. and Young S.: Semantic processing using the Hidden Vector State model. Computer Speech and Language, Volume 19, Issue 1, 2005, 85–106.

[7] Jurčíček F.: Statistical approach to the semantic analysis of spoken dialogues . Ph.D. thesis, p. 137, University of West Bohemia, Faculty of Applied Sciences, Pilsen, Czech Republic, Plzeň, 2007

[8] Deborah A. Dahl, et al., ATIS3 Test Data, Linguistic Data Consortium, Philadelphia, 1995

[9] Iosif E. and Potamianos A.: A soft-clustering algorithm for automatic induction of semantic classes. In *Interspeech-07*, pages 1609–1612, Antwerp, Belgium, August 2007.

[10] Jeong M. and Lee G.: Practical use of non-local features for statistical spoken language understanding. Computer Speech and Language, 22(2):148–170, April 2008.

[11] Raymond Ch. and Riccardi G.: Generative and discriminative algorithms for spoken language understanding. In *Interspeech-07*, pages 1605–1608, Antwerp, Belgium, August 2007.

[12] Novák V.: Semantic Network Manual Annotation and its Evaluation, Ph.D. thesis, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 129 pp., Sep 2008

[13] Cinková S.: Semantic Representation of Non-Sentential Utterances in Dialog, in Proceedings of SRSL 2009, the 2nd Workshop on Semantic Representation of Spoken Language, Athina, Greece, pp. 26-33, 2009

[14] Novák V., Hartrumpf S., and Hall K.: Large-scale Semantic Networks: Annotation and Evaluation, in Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, Boulder, CO, USA, ISBN 978-1-932432-31-2, pp. 37-45, 2009

[15] Pala K., Rychlý P., and Smrž P.: DESAM - Annotated Corpus for Czech. In Proceedings of SOFSEM 97. Heidelberg : Springer Verlag, 1997. pp. 523-530. ISBN 3-540-63774-5.

[16] Pala K.: Semantic Annotation of (Czech) Corpus Texts. In Proceedings of the Second Workshop on Text, Speech and Dialogue. Berlin : Springer Verlag, 1999. pp. 56-61. Lecture Notes in Artificial Intelligence 1692. ISBN 3-540-66494-7.

[17] Pala K.: Word Senses and Semantic Representations. In Proceedings of the Third international Workshop on Text, Speech and Dialogue. Lecture Notes In Computer Science, vol. 1902. Springer-Verlag 2000, London, pp. 109-114. ISBN 3-540-41042-2.

[18] Pala K. and Sedláček R.: Enriching WordNet with Derivational Subnets. In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing CICLING 2005. Springer Verlag, 2005. pp. 305-311, ISBN 3-540-24523-5.

[19] Jurčíček F., Zahradil J., and Jelínek L.: A Human-Human Train Timetable Dialogue Corpus. In Proceedings of EUROSPEECH, Lisboa, Portugal, 2005.

[20] Habernal I. and Konopík M.: JAAE: the Java Abstract Annotation Editor, In *INTERSPEECH-2007*, 1298-1301, 2007.

[21] Jurcicek F., Gasic M., Keizer S., Mairesse F., Thomson B. and Young S. Transformation-based learning for semantic parsing In: Proceedings of Interspeech 2009, 10th Annual Conference of the International Speech Communication Association, 6-10 Sept 2009, Brighton, UK.

[22] Wu W., Lu R., Duan J., Liu H., Gao F., and Chen Y. 2010. Spoken language understanding using weakly supervised learning. Comput. Speech Lang. 24, 2 (Apr. 2010), 358-382

[23] Zhou D. and He Y. 2009. Discriminative Training of the Hidden Vector State Model for Semantic Parsing. IEEE Trans. on Knowl. and Data Eng. 21, 1 (Jan. 2009), 66-77

 ISBN: 978-1-61208-104-5