Perspectives of Turning Prague Dependency Treebank into a Knowledge Base

Václav Novák, Jan Hajič

Institute of Formal and Applied Linguistics
Charles University
Prague, Czech Republic
{novak,hajic}@ufal.mff.cuni.cz

Abstract

Recently, the Prague Dependency Treebank 2.0 (PDT 2.0) has emerged as the largest text corpora annotated on the level of tectogrammatical representation ("linguistic meaning") described in Sgall et al. (2004) and containing about 0.8 milion words (see Hajič (2004)). We hope that this level of annotation is so close to the meaning of the utterances contained in the corpora that it should enable us to automatically transform texts contained in the corpora to the form of knowledge base, usable for information extraction, question answering, summarization, etc. We can use Multilayered Extended Semantic Networks (MultiNet) described in Helbig (2006) as the target formalism. In this paper we discuss the suitability of such approach and some of the main issues that will arise in the process. In section 1. we introduce formalisms underlying PDT 2.0 and MultiNet, in section 2. we describe the role MultiNet can play in the system of Functional Generative Description (FGD), section 3. discusses issues of automatic conversion to MultiNet and section 4. gives some conclusions.

1. Introduction

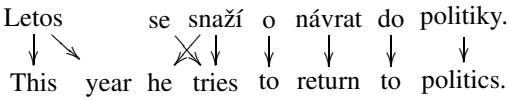
1.1. Prague Dependency Treebank

The Prague Dependency Treebank 2.0 (PDT 2.0) described in Sgall et al. (2004) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation (0.8 MW); in addition, certain properties of sentence information structure and coreference relations are annotated at the semantic level.

The theoretical basis of the treebank lies in the Functional Generative Description (FGD) of language system by Sgall et al. (1986).

PDT 2.0 is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs. The corpus itself uses the latest annotation technology. Software tools for corpus search, annotation and language analysis are included. Extensive documentation (in English) is provided as well.

An example of a tectogrammatical tree from PDT 2.0 is given in figure 1. Function words are removed, their function preserved in node attributes (*grammatemes*), information structure is annotated in terms of topic-focus articulation, and every node receives detailed semantic label corresponding to its function in the utterance (e.g., *addressee*, *from_where*, *how_often*, ...). The tree represents the following sentence:



(1)

1.2. MultiNet

The representational means of Multilayered Extended Semantic Networks (MultiNet), which are described in Helbig (2006), provide a universally applicable formalism for the treatment of semantic phenomena of natural language. To this end, they offer distinct advantages over the use of the classical predicate calculus and its derivatives. The knowledge representation paradigm and semantic formalism MultiNet is used as a common backbone for all aspects of natural language processing (be they theoretical or practical ones). It is continually used for the development of intelligent information and communication systems and for natural language interfaces to the Internet. Within this framework it is subject to permanent practical evaluation and further development.

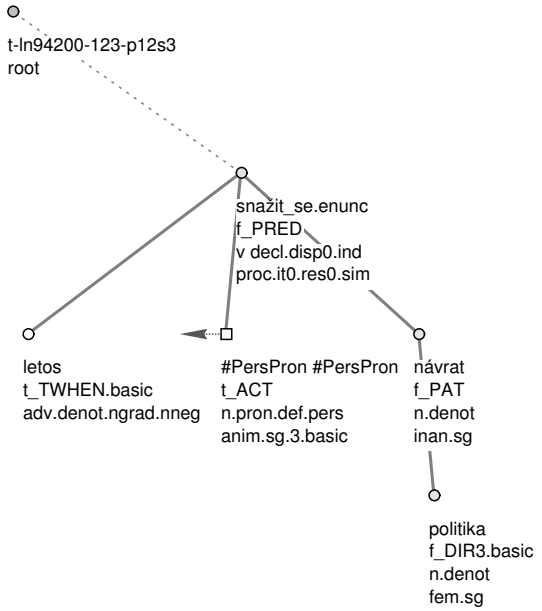


Figure 1: Tectogrammatical tree of sentence (1)

The semantic representation of natural language expressions by means of MultiNet is mainly independent of the considered language. In contrast, the syntactic constructs used in different languages to describe the same content are obviously not identical. To bridge this gap between different languages we can employ the deep syntactico-semantic representation available in the framework of FGD.

An example of a MultiNet structure is given in figure 2. The figure represents the following discourse:

(2)

Max gave his brother several apples.
This was a generous gift.
Four of them were rotten.

This is an example of interplay between the intensional and the preextensional (see Helbig (2006), p. 25) level of MultiNet.

MultiNet is not explicitely model-theoretical and the extensional level is created only in those situations where the natural language expressions require it. It can be seen that the overall structure of the representation is not a tree any more. The layer information is hidden except for the most important QUANT and CARD values. These attributes convey information that is important with respect to the meaning of the sentence. Tectogrammatical representation lacks attributes distinguishing intensional and extensional information, and there are no relations like SUBM denoting the relation between a set and its subset.

Note that the MultiNet representation crosses the sentence boundaries. First, the structure representing a sentence is created and then this structure is assimilated into existing representation.

In contrast to CLASSIC (Brachman et al., 1991) and other KL-ONE networks, MultiNet contains a predefined final set of types of relations, encapsulation of concepts and attribute layers concerning cardinality of objects mentioned in the discourse.

2. Integration of MultiNet to the FGD

PDT 2.0 contains three layers of information about the text (as described in Hajič (1998)):

Morphosyntactic Tagging. This layer will represent the text in the original linear word order with a tag assigned unambiguously to each word form occurence, much like the Brown corpus does.

Syntactic Dependency Annotation. It contains the (unambiguous) dependency representation of every sentence, with features describing the morphosyntactic properties, the syntactic function, and the lexical unit itself. All words from the sentence appear in its representation.

Tectogrammatical Representation. At this level of description we will annotate every (autosemantic nonauxiliary) lexical unit with its tectogrammatical function, position in the scale of the communicative dynamism and its grammatemes (similar to the morphosyntactic tag, but only for categories which cannot be derived from the word's function, like number for nouns, but not its case).

In the process of deeper understanding, it seems obvious that we need another layer in order to include the whole discourse into a structure allowing inferences provided by comprehensible axioms that may be both hand-made and automatically acquired. It is shown in Lin and Pantel (2001) that already the layer of syntactic dependency annotation can be successful in allowing to statistically derive useful inferences by induction. It is our hope that the structure provided by the MultiNet formalism can rival the simple syntactic structure as the input for the inference inductor.

With a sufficient initial size of the knowledge base and the dictionary we can than expand the knowledge automatically by means of bootstrapping (Eisner and Karakos, 2005) or other automatic knowledge mining methods like GUHA (Hájek and Havránek, 1978). These and other methods can operate independently and concurrently on the knowledge base.

The MultiNet format will be much more efficient and suitable for above mentioned algorithms than surface syntax or even plain text. Even if we used it directly on the level of tectogrammatical representation, we couldn't insert the results of the modules back to the representation and we wouldn't be able to operate by one module on data already enriched by another one. The resulting knowledge bases will be an invaluable resource in its own right, creating opportunities for further research.

Also the axiomatic system connected with MultiNet can not be applied to other layers of annotation, because they lack the necessary regularity with respect to inference rules (e.g., the syntactic roles can't give the kind of information that the AFF relation in MultiNet gives us – the affected object is changed by the event).

3. Conversion of TR to MultiNet

3.1. Elements of TR

The process of initial transformation from tectogrammatical representation to MultiNet is straightforward with respect to the structure of the network, but there arise many problems in transforming TR roles and attributes to MultiNet relations and layer information. In order to correctly transform the tree we sometimes need both the grammatemes and the background world knowledge.

However, by carefully studiyng the existing representation, we discovered there are still obstacles which complicate the transformation. Here is the summary and discussion of some possible solutions.

The main issues in the transformation from the TR point of view consist of:

Named entities recognition and their coreferencing in the text. This includes also numerical expressions. However, there is an ongoing project to include the named entity annotation directly into the treebank.

Topic-Focus Articulation (TFA) is annotated in PDT 2.0, but in MultiNet we need to formalize its contribution to the overall meaning. We must keep track of what is typical or characteristic for what and transform TFA into these attributes. It seems that the concept of encapsulation in MultiNet can handle much of the topic-focus distinction. Also the scope of quantificators, which is strongly influenced by TFA, can be represented formally by by interplay of layer attributes and the encapsulation.

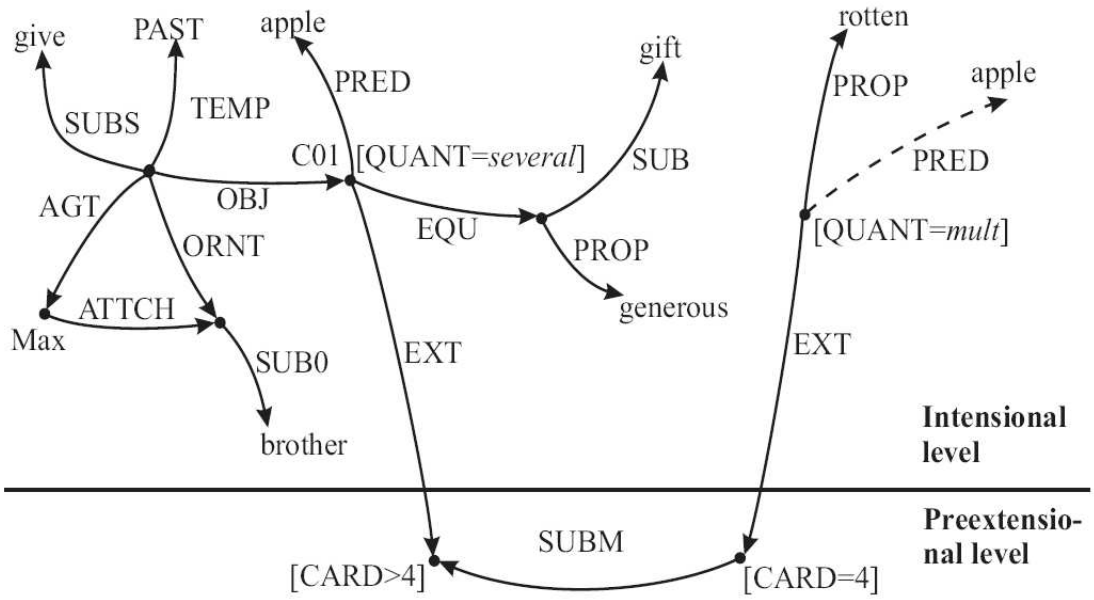


Figure 2: MultiNet representation of example discourse (2)

Attitude elimination. Especially in newspapers, the meaning is very often expressed indirectly ("X said Y", "A confirmed B", "S denied T"). In these cases we need to estimate the likelihood of Y, B and T from our knowledge of the corresponding verbs and maybe also from our knowledge about X, A and S. MultiNet itself doesn't contain a mechanism to represent this knowledge and needs to be enriched in this respect. Since it is a semantic network, there is no principal obstacle preventing us to assign weights to elements of the network.

Metadata, especially the author, the bibliographical reference and the date are often necessary for reasoning, but this must be taken into account in the process of corpora creation, so that it can be subsequently used in further processing. Multinet structures don't allow explicit metadata inclusion into the network, but it is rather a technical issue to be solved.

Grammatemes (e.g., tense, gender, iterativeness, verbal modality) must be transformed into formal modifications of the meaning and eliminated. In this process the likelihood may be affected as well as representation of temporal relations in the knowledge base. In addition, new nodes of the network will be introduced. Grammatemes can be partially transformed also to the layer information of respective nodes.

3.2. Elements of MultiNet

Some of the MultiNet elements are not present in TR. Here is a first attempt to list the main disproportions:

- Sorts of concepts (the upper conceptual ontology) are not present in TR. In order to represent the sentence we need to know the sort of every concept.
- Some of the cognitive roles in MultiNet correspond to different roles functors in TR. This relation is typically many to one:
  1. Actor in TR will usually become connected to the situation by AFF, AGT, BENF, CSTR, EXP, MEXP and SCAR relations.
  2. Patient in TR will usually become connected to the situation by AFF, ATTR, BENF, ELMT, GOAL, OBJ, PARS, PROP, SSPE and VAL relations.
  3. Functors of location will sometimes become DIRCL, ELMT, LEXT, LOC, ORIGL, ORNT, PARS, CTXT and SITU. Sometimes the right relation can be found from the grammatemes but sometimes the background knowledge is needed.

  Moreover new nodes will have to be included into the network in order to express the meaning (e.g., the preextensional nodes at figure 2).
- Attribute-Value characterization is not formalized in TR. In order to draw the inferences we need to establish the equivalence of expressions like:
  1. The color of x is y.
  2. x has y color.
  3. x is y.
  4. y is the color of x.

  All these examples are cases of simple attribute-value assignments that should result in a single MultiNet representation although their TRs differ considerably.
- Some nodes from TR will have no counterpart in the network. They will be represented by an edge like ANLG, CORR, or CTXT.
- Unlike TR, in MultiNet one object can be connected more than once to a situation (typically, AGT will be also a CSTR).

- The temporal relationship between events expressed by TEMP, ANTE, DUR, STRT and FIN can not be trivially transferred from grammatemes of TR denoting verbal tenses.

On the other hand, there are some TR functors corresponding quite straightforwardly to MultiNet relations. The most typical examples are listed in table 1.

TR functor	MultiNet Relation
ADDR	ORNT
DIR1	ORIGL
DIR2	VIA
DIR3	DIRCL
MANN	MANNR
MEANS	INSTR
SUBS	SUBST
TPAR	DUR
TSIN	STRT
TTILL	FIN
TWHEN	TEMP

Table 1: TR functors corresponding to a single MultiNet relation

3.3. Additional requirements

In addition, these are some of the components necessary in the process of including new knowledge into the knowledge base:

Spatio-Temporal representation is necessary for relating places and times mentioned in the text.

Calendar is important for mapping the language expression into the temporal representation.

Ontology will be useful in further processing of the knowledge although it is not necessary in the transformation itself. It will however create the network that will connect the isolated pieces of information into one network. The advantage of using MultiNet lies in the fact that the ontological hierarchy is an inherent part of the semantic network. There is a single SUB relation connecting both instances of concepts with the concept and concepts in different levels of the ontology.

We have already certain suggestions how to tackle these issues; their solution enables to automatically turn text annotated on the tectogrammatical level of PDT to a practically usable knowledge base, which can be further processed and enriched with inferences. Such a base can be searched for answers and can serve as a precise source for information extraction.

4. Conclusions

We discussed usage of Multilayered Extended Semantic Networks (MultiNet) as an extension of the existing framework of Functional Generative Description (FGD). We discussed the suitability of such approach and some of the main issues that will arise in the process. We described the role of MultiNet in the system of Functional Generative Description and discussed issues of automatic conversion to MultiNet.

Let us also remark that from the historical perspective, the MultiNet formalism is a natural continuation of the effort to analyze natural language. Its philosophical foundation (meaning as the use of language in (Wittgenstein, 1953)) has much in common with the notion of function–form assymetry in FGD dating back to Karcevskij (1929).

5. Acknowledgements

This work was supported by Czech Academy of Science grant 1ET201120505 and by Czech Ministry of Education, Youth and Sports project LC536. The views expressed are not necessarily endorsed by the sponsors.

6. References

Ronald J. Brachman, Deborah L. McGuinness, Peter F. Patel-Schneider, Lori Alperin Resnick, and Alex Borgida. 1991. Living with classic: When and how to use a kl-one-like language. In John Sowa, editor, *Principles of Semantic Networks: Explorations in the representation of knowledge*, pages 401–456. Morgan-Kaufmann, San Mateo, California.

Jason Eisner and Damianos Karakos. 2005. Bootstraping without the boot. In *Proceedings of HLT/EMNLP*, pages 395–402, Vancouver, Canada, October.

Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Karolinum, Charles University Press, Prague, Czech Republic.

Jan Hajič. 2004. *Complex Corpus Annotation: The Prague Dependency Treebank*. Jazykovedný ústav Ľ. Štúra, SAV, Bratislava, Slovakia.

Hermann Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer-Verlag, Berlin Heidelberg.

Petr Hájek and Tomáš Havránek. 1978. *Mechanizing Hypothesis Formation; Mathematical Foundations for a General Theory*. Springer-Verlag, Berlin, Heidelberg, New York.

S. Karcevskij. 1929. Du dualisme asymétrique du signe linguistique. *Travaux du Cercle linguistique de Prague*, 1:88–93.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing company, Dodrecht, Boston, London.

Petr Sgall, Jarmila Panevová, and Eva Hajičová. 2004. Deep syntactic annotation: Tectogrammatical representation and beyond. In A. Meyers, editor, *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ludwig Wittgenstein. 1953. *Philosophische Untersuchungen*. Suhrkamp Verlag, Frankfurt am Main.