Automatic Generation of Large Knowledge Bases using Deep Semantic and Linguistically Founded Methods

Sven Hartrumpf[1], Hermann Helbig[2] and Ingo Phoenix[1]
[1]*SEMPRIA GmbH, 40237 Düsseldorf, Germany*
[2]*University at Hagen, Intelligent Information and Communication Systems Group, 58084 Hagen, Germany*

Keywords: Semantic Analysis, Knowledge Bases, Text Understanding, Natural Language Processing, Reference Resolution.

Abstract: Large-scale knowledge acquisition from texts is one of the challenges of the information society that can only be mastered by technical means. While the syntactic analysis of isolated sentences is relatively well understood, the problem of automatically parsing on all linguistic levels, starting from the morphological level through to the semantic level, i.e. real understanding of texts, is far from being solved. This paper explains the approach taken in this direction by the MultiNet technology in bridging the gap between the syntactic-semantic analysis of single sentences and the creation of knowledge bases representing the content of whole texts. In particular, it is shown how linguistic text phenomena like inclusion or bridging references can be dealt with by logical means using the axiomatic apparatus of the MultiNet formalism. The NLP techniques described are practically applied in transforming large textual corpora like Wikipedia into a knowledge base and using the latter in meaning-oriented search engines.

1 INTRODUCTION

Automatic knowledge acquisition is one of the most disturbing bottlenecks of Artificial Intelligence or, to be more specific, of Computational Linguistics. In spite of the rapid progress in the field of natural language processing (NLP), only few research teams are able to automatically build large knowledge bases from texts based on a deep semantic analysis of natural language (NL) information, and to include logical methods into the process of text understanding.

On the one hand, one meets the statistical or pattern-based approaches (Klavans and Resnik, 1996; Ravichandran and Hovy, 2002) or vector space models (Socher et al., 2012) for extracting semantic information (e.g. specific semantic relations like conceptual subordination, part-whole relations, etc.) from texts. However, they neither cover the whole spectrum of semantic relationships nor do they have a clear logic and semantic representation of the information derived from the texts. On the other hand, there are linguistically motivated approaches with a strong syntactic-semantic analysis, but very limited semantic depth (so-called *shallow* approaches, e.g. Robust Minimal Recursion Semantics (Copestake et al., 2005)).

To build a knowledge base (KB) from texts, one needs an automatic interpreter that translates NL sentences into formal meaning structures. Such an interpreter is provided by the WOCADI parser (Hartrumpf, 2003), using the MultiNet formalism for semantic representation. Since the complex knowledge representation paradigm MultiNet cannot be described on a few pages, only a short overview of the representational means of MultiNet relevant to the understanding to the paper is given in Sect. 2. The construction of a KB from the meaning structures of isolated sentences is based on an automated process, called *assimilation*, which treats all text-constituting effects (including the disambiguation of words, syntactic relations and textual references) and connects the semantic structures of single sentences of a text to a coherent KB. In this process, semantically equivalent elements of partial structures have to be identified, references must be resolved, and bridges between seemingly isolated meaning structures have to be established by means of background knowledge. This is the topic of this paper.

The problem of coreference resolution (as one of the most prominent text-constituting effects) has received plenty of scientific attention (Kamp and Reyle, 1993; Hobbs et al., 1993; Ge et al., 1998). One of

Hartrumpf S., Helbig H. and Phoenix I..
Automatic Generation of Large Knowledge Bases using Deep Semantic and Linguistically Founded Methods.
DOI: 10.5220/0004756202970304
In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence* (ICAART-2014), pages 297-304
ISBN: 978-989-758-015-4

the first approaches using background knowledge for coreference resolution was that of Hobbs et al. (Hobbs et al., 1993). Their weighted abduction scheme selects a single best interpretation, which may turn out false at a later point. This problem is avoided by model-building approaches which keep track of all alternatives simultaneously (Baumgartner and Kühn, 2000). In contrast, the system used in our approach demonstrates a rule-based method (supported by corpus-based back-off statistics) for coreference resolution of pronominal and nominal anaphors.

2 MEANING REPRESENTATION WITH MultiNet

One of the prominent knowledge representation paradigms used as meaning representation in NLP are *semantic networks*, which represent concepts as nodes of a graph and relations between concepts as arcs between these nodes. Multilayered Extended Semantic Networks (abbreviated MultiNet, see (Helbig, 2006)) belongs to this basic paradigm. Here are some of its key features:

1. Every node is classified according to a predefined ontology of 45 basic sorts.
2. Each node has a well-defined inner structure specified by an attribute-value structure. The attributes relevant in the context of this paper are:
   • GENER: The *degree of generality* marks a concept as generic (value: *ge*) or specific (value: *sp*). Examples: "*(A car)* [GENER *ge*] *is a useful means of transport.*" vs. "*(This car)* [GENER *sp*] *is a useful means of transport.*"
   • REFER: This attribute specifies the *determination of reference*, i.e. whether there is a determined object of reference (value: *det*) or not (value: *indet*). This information is important for the resolution of references.
   Example: "*(The man)* [REFER *det*] *observed (an accident)* [REFER *indet*]."
   • ETYPE: This is the *extensionality type* of an entity: *nil* – no extension, 0 – individual that is not a set (e.g. ⟨Elizabeth I⟩), 1 – entity with a set of [ETYPE 0] elements as extension (e.g. ⟨many houses⟩, ⟨the family⟩), 2 – entity with a set of [ETYPE 1] elements as extension (⟨many families⟩), etc.
3. The arcs may only be labeled by members of a fixed set of relations and functions. Typical relations are described in Table 1 (see (Helbig, 2006) for the complete specification). The signatures of relations and functions are defined in terms of the sorts mentioned in point 1.
4. Apart from the sorts, MultiNet provides a predefined set of semantic features (see Table 2) to check selectional restrictions during syntactic-semantic analysis.

The assimilation process as described in the paper is supported by the technological environment developed for MultiNet (comprising, among other things, a workbench for the knowledge engineer) and by the semantically based computational lexicon HaGenLex (Hartrumpf et al., 2003). The screenshots of the semantic networks in this paper are all produced by the MWR knowledge engineering workbench (Gnörlich, 2002), which can also access the parser. The development of a large semantically based computational lexicon is facilitated by LIA+, a workbench for the computer lexicographer (Hartrumpf et al., 2003).

3 TREATING TEXT-CONSTITUTING PHENOMENA BY ASSIMILATION

In this section, we discuss the most important phenomena that must be treated during the assimilation of a text from the representation of its sentences.

3.1 Grammatical and Semantical References

3.1.1 Coreference

The most important types of reference are induced by proforms, i.e. by pronouns and proadverbs. An example is given by sentence (2) below, where the phrase *ihre Mitglieder/its members* containing the possessive pronoun *ihre/its*, refers to the apposition ⟨Familie Beier⟩/⟨Beier family⟩ introduced in (1).

(1) *Familie Beier wohnt in Hoffenheim.*
The Beier family lives in Hoffenheim.

(2) *Ihre Mitglieder (R_1) sind Fans des örtlichen Fußballvereins.*
Its members (R_1) are fans of the local soccer club.

The correct resolution of reference R_1 depends on the background knowledge that Familie/family represents a collection of entities (expressed in MultiNet by [ETYPE 1]), discerning it from concepts like house with [ETYPE 0]. This information is even more important in English since on grammatical grounds

Table 1: Semantic relations of MultiNet mentioned in the text.

Relation	Signature	Short characteristics	Sorts used
AFF	$[dy \cup ad] \times [o \cup si]$	Affected object	dy: event; ad: abstract event
AGT	$[si \cup abs] \times o$	Agent	o: object, si: situation
ANTO	$sort \times sort$	Antonymy relation	sort: no restriction on sorts
ATTCH	$[o \setminus at] \times [o \setminus at]$	Attaching objects to objects	at: attribute
ELMT	$pe^{(n)} \times pe^{(n+1)}$ with $n \geq 0$	Element relation	$pe^{(k)}$: extensional object of type k
ORNT	$[si \cup abs] \times o$	Orientation toward something	abs: abstract situation
PARS	$[co \times co] \cup [l \times l]$	Part-whole relationship	co: concrete object, l: location
POSS	$[co \cup io] \times [co \cup io]$	Ownership relation	io: ideal object
SUB	$[o \setminus abs] \times [\overline{o} \setminus \overline{abs}]$	Subordination of objects	$\overline{o}$: generic object
SUBS	$[si \cup abs] \times [\overline{si} \cup \overline{abs}]$	Subordination of situations	$\overline{si}$: generic situation
SYNO	$sort \times sort$	Synonymy relation	$\overline{abs}$: generic abstract situation

Table 2: Typical features for the semantic fine-characterization of objects.

Semantic features		Example values	
Name	Meaning	+	−
ANIMATE	living being	tree	stone
ARTIF	artifact	house	tree
GEOGR	geographical object	the Alps	table
HUMAN	human being	student	ape
INSTIT	institution	UNO	apple
MOVABLE	object being movable	car	forest
SPATIAL	object having spatial extension	table	idea

alone (without the semantic level) the pronoun *its* could also refer to Hoffenheim.

3.1.2 Proforms (P)

The linguistic elements initiating a reference in a text are characterized in MultiNet using, among other things, the attribute [REFER *det*]. Sorts and features also play a special role in resolving references.

Pronouns (P1). Reference resolution involves a disambiguation problem, i.e. there are typically several antecedent candidates, one of which has to be chosen as the correct one. In NLP, the search problem for the antecedent fitting best the restrictions defined by the proform is mastered relatively well for pronouns compared to proadverbs. Since reference resolution is systematically treated in other publications, only the basic mechanisms shall be treated here.

Figure 1 shows the representation of two sentences after syntactic-semantic analysis and before assimilation; note that the numerical part of reading identifiers (concept IDs) like manometer.1.1 is dropped in the following if irrelevant.

(3) *Die Firma* (A_1) *hat eine neue Turbine* (A_2) *geliefert.*
The company (A_1) *delivered a new turbine* (A_2).

(4) *Sie* (P_1) *musste deren* (P_2) *Manometer auswechseln.*
It (P_1) *had to replace its* (P_2) *manometer.*

At the beginning, there are two possible antecedents for the pronoun (P_1) (word: *Sie/It*, node c13275 in Figure 1, right side): (A_1) = node c13245 and (A_2) = node c13252 in Figure 1, left side. Both are candidates for the resolution of the reference triggered by *Sie/It* because of the agreement in gender (German: feminine), number (singular), and person (3rd). Since only a company and no turbine can replace something (selectional restrictions of the verb), only node c13245 can play the semantic role of the agent (AGT) marked in event c13276, representing the meaning of the second sentence. This means, the nodes c13245 and c13275 have to be merged into one node during the assimilation of the two partial networks of Figure 1, see the result in Figure 2.

Turning to the demonstrative pronoun *deren/its* (P_2), node c13273 in Figure 1, another effect must be observed. The pronoun *deren/its* (genitive case) has a possessive meaning, whose exact interpretation requires background knowledge. This possessive aspect is expressed in Figure 2 by (c13252 ATTCH c13274), specifying a general attachment. Here, one

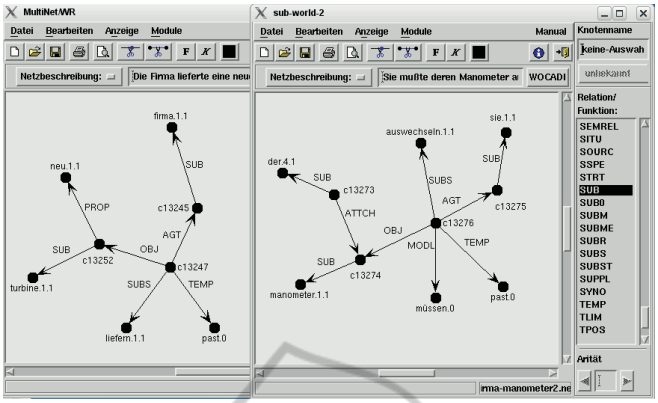

Figure 1: Ambiguities with the resolution of pronoun references.

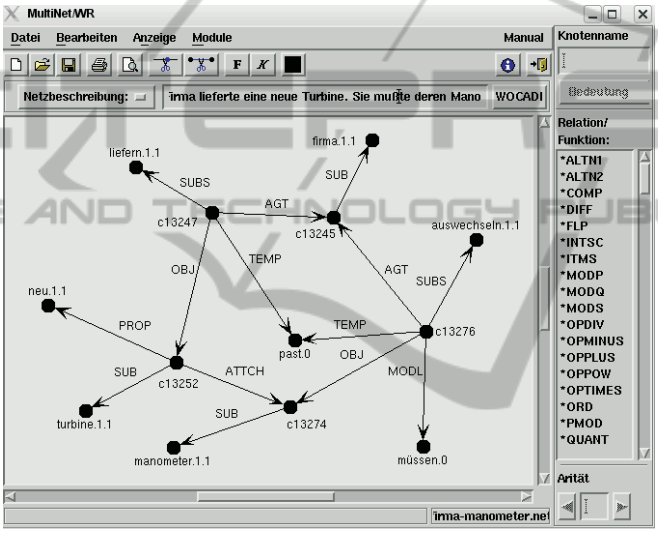

Figure 2: Result of the assimilation of the two networks from Figure 1.

has to know which of the possible antecedents possesses a manometer, either the company (c13245 POSS c13274) or the turbine (c13274 PARS c13252). Thus assimilation needs not only find the proper referential assignment, but also the correct interpretation of the underspecified relation ATTCH. The use of background knowledge necessary for assimilation will be explained in connection with inclusion and logical recurrence.

A specialty of the German word *deren* initiating a reference consists in the fact that, after having decided on the antecedent of *Sie*, the proform *deren* cannot refer to the subject of sentence (3), i.e. to A_1, for syntactical reasons. Since a reference of node c13273 to node c13245 must not be expressed by *deren*, but by *ihren*, there is no ambiguous reference in this case. Consequently, a part-whole relationship (c13274 PARS c13252) has to be established between the manometer and the turbine. Figure 2 shows the assimilation result for sentences (3) and (4).

Proadverbs (P2). It should be mentioned in advance that, even by the current state of the art, the resolution of this type of coreferences is not yet fully mastered. Nonetheless, as a special support, one has the congruency between prepositions (i.e., the congruency between proadverb and the prepositional phrase) and the agreement of sorts in general. In the following examples, it is either the MultiNet sort *si* (first example) or a local specification (MultiNet sort *l* in the second example) determining the congruency relation.

(5) *Der Kunde vertraute* **auf**
die Zusage des Händlers$_i$ [SORT *si*]
→ *Er vertraute dar***auf**$_i$ [SORT *si*].
The customer trusted **in**
the commitment of the dealer$_i$ [SORT *si*]
→ *He trusted* **in** *that*$_i$ [SORT *si*].

(6) *Der Student*$_j$ *ging in das Haus*$_i$ [SORT *l*].
→ *Dort*$_i$ [SORT *l*] *traf er seinen*$_j$ *Freund.*

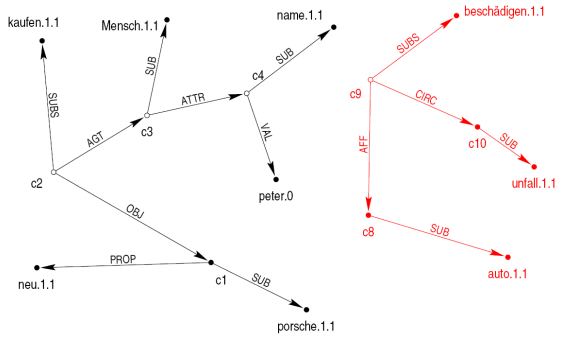

Figure 3: Inclusional reference with use of a superordinated concept.

The student$_j$ went into the house$_i$ [SORT *l*].
→ *There$_i$* [SORT *l*], *he met his$_j$ friend.*

3.1.3 Inclusion (I)

Characteristic of inclusional references is the use of subordination relations between concepts to establish a coherent text, as shown here:

(7) *Peter (A_1) kaufte einen neuen Porsche$_i$ (A_2).*
Peter (A_1) bought a new Porsche$_i$ (A_2).

(8) *Der Wagen$_i$ (R) wurde bei einem Unfall beschädigt.*
The car$_i$ (R) has been damaged in an accident.

The semantic structures of both sentences are shown in Figure 3. Please, take into consideration that node auto.1.1 in this figure arises from a normalization process transforming the concept wagen.1.1 (one meaning of the German word *Wagen*) into the synonymous concept auto.1.1. It is then up to the assimilation process to determine which of the theoretically possible antecedents (A_1) and (A_2), represented by the object nodes c1 and c3, are coreferential with the semantic representative c8 of the noun phrase (R) initiating the reference.[1] The general approach for solving inclusional references is the following:

1. Let C_R denote the semantic representative of the phrase R initializing the reference and S_R the superordinated concept used to describe C_R. S(R) be the sentence containing R with semantic description $C_S(R)$. The node C_R bears the attribute [REFER *det*]. At the beginning of assimilation, a logical query form (? SUB S_R) is automatically generated, where the question mark ? stands for the semantic representative of the antecedent searched for and also for C_R, since both refer to the same object.

2. The query (? SUB S_R) has to be answered by logical means over the given knowledge base containing the meaning of the foregoing sentences and all background knowledge. The question mark is interpreted as a variable to be substituted during the inference process.

3. At the end of the inferential question answering, if successful, the substitute found for the variable ?, i.e. a node from the knowledge base, has to be merged with C_R. Thus, the new piece of knowledge $C_S(R)$ containing the node C_R is integrated (assimilated) into the existing KB.

In sentences (7) and (8), the query mentioned has to be derived from the semantic description of node c8 of the partial network N_I at the right side of Figure 3 since this node represents the entity with layer attribute [REFER *det*] initializing the reference. Thus we get as query form (? SUB wagen.1.1) or English: (? SUB car.1.1). As already emphasized, the answer again can generally only be found by means of background knowledge. In this case, the computer has to know that a Porsche is a car and that the subordination of concepts, i.e. the relation SUB, is transitive. The answer, in this case can be derived by means of the knowledge represented by the partial network N_{II} on the left side of Figure 3.

The knowledge needed for treating sentences (7) and (8) comprises:
(1) (? SUB wagen.1.1) :: query generated from network N_I
(2) (c1 SUB porsche.1.1) :: from network N_{II}
(3) (porsche.1.1 SUB wagen.1.1) :: background knowledge
(4) (x SUB y) ∧ (y SUB z) → (x SUB z) :: axiom for the relation SUB
Additionally, the following constraints have to be observed:
(C1) [GENER(?) *sp*]
(C2) [GENER(porsche.1.1) *ge*]
(C3) [GENER(wagen.1.1) *ge*]
From this, the following conclusion can be drawn:
(5) (c1 SUB wagen.1.1) :: from (2), (3) and, (4)

By unifying (1) and (5), substituting c1 for ?, the answer and the solution of the assimilation problem can be found: c1 has to be merged with c8.[2] This solution is also intuitively understandable since node c1 represents the only car in the knowledge base that c8 could refer to.

Inclusions of situations (events) also play a role in reference resolution:

[1] One should not be misled by implicit usage of full human knowledge. Without knowledge what is a car, for a machine, Peter could have been referred to by (R).

[2] Note that the question mark ? also represented node c8 from network N_I.

(9) *Peter schnitzte (A) eine Figur aus Eichenholz.*
Peter carved (A) a figure from oak wood.

(10) *Während er arbeitete (R), hörte er ein neues Radio-Hörspiel.*
While working (R) he listened to a new radio play.

Here the inclusion is mediated by the relation SUBS instead of SUB, to be more specific, by the relationship (schnitzen/carve SUBS arbeiten/work). The inclusion of situations can be treated analogously to the inclusion of conceptual objects.

3.1.4 Semantic Recurrence (S)

The inner coherence of many texts or partial texts can only be established by including the semantic level and using logical inferences. Semantic gaps seemingly encountered during this process can often be closed only by background knowledge. Analogous mental activities occur also with human beings. But these activities mostly remain unconscious. Thus, they are difficult to model, which is aggravated by two circumstances: a great amount of common sense knowledge is needed; and the automatic inference processes involved are not yet sufficiently mastered. Nevertheless, the basic mechanisms are already well understood and can be properly formalized. Here, we use the representational means of MultiNet to show the working of these mechanisms. Typical relations which often play an essential part in this context are the following:

- The synonymy relation (MultiNet relation: SYNO). Example:
  (11) *The writer (A) brought a new book on the market. Immediately afterwards a new biography of the author (R) was published.*
  Background knowledge:(writer SYNO author);
- The antonymy relation (ANTO). Example:
  (12) *During the day (A) he carried out regular activities. During the night (R) he stole cars.*
  Background knowledge needed to recognize the contrast: (day ANTO night);
- The part-whole relationship (PARS). Example:
  (13) *The department bought a new computer (A). The monitor (R) had to be reclaimed.*
  Backgr. knowledge: (monitor PARS computer);
- The relationship of set membership (ELMT). Example:
  (14) *The department (A) bought an expensive computer. The coworkers (R) were provided with an Internet access (by that).*
  Background knowledge needed: (coworker$_{EXT}$ ELMT department$_{EXT}$)[3]

Ontologically based References. Some references are based on ontological knowledge. Such an ontology is given, for instance, by the sort hierarchy of MultiNet. Since, besides of the sort symbols used in the signatures, there are also NL terms labeling the ontological classes (e.g. [SORT *dy*] for *Ereignis/event*, [SORT *l*] for *Ort/location*, or [SORT *p*] for *Eigenschaft/property*), these sorts are anchored in NL. Therefore, in some cases, they can be seen as mediators of references.

(15) *On March 11, 1997 the best students of the annual contest had been found out.* ([SORT *dy*] for the whole event)

(16) *Many parents were present during this event.*

Since the term *event* bears the sort label *dy*, the phrase *this event* refers to the whole situation described by the first sentence. There is a close connection to reference phenomena dealt with under the headline *inclusion* since hierarchies of concepts of this kind can be represented by the relations SUB or SUBS.

As already mentioned, references in a text are often characterized by the use of superordinated concepts, synonyms, or antonyms. Since relations like subordination (SUB/SUBS), synonymy (SYNO), and antonymy (ANTO) are characteristic for ontologies, references built on them are called *ontological references*. With references induced by constructs like
⟨definite article⟩ ⟨noun denoting a superordinated concept⟩
⟨demonstrative determiner⟩ ⟨noun denoting a superordinated concept⟩
the hierarchy of concept subordination carried by the relation SUB comes into play (see axioms (A1) and (A2) below):

(17) *Familie Beier hat im vergangenen Jahr ein neues Haus (A) gebaut.*
Last year, the Beier family built a new house (A).

(18) *Das Gebäude (R_1) wurde von allen bewundert.*
The building (R_1) was admired by everyone.

(19) *Leider wurde der Keller (R_2) durch das Hochwasser überflutet.*
Alas, the basement(R_2) has been overflowed by flood.

Depending on the continuation (18) or (19) of sentence (17), one meets different types of references (R_i) to (A) and needs different inferences and pieces of background knowledge to resolve the references.

[3]The index EXT refers to the fact that, strictly speaking, the element relation ELMT holds between the extensions of the concepts.

In (18), *das Gebäude/the building* (R_1) points to the house (A) introduced in (17). Such a reference often spans several steps in the subordination hierarchy and the transitivity of SUB must be considered:
(A1) (x SUB y) $\wedge$ (y SUB z) $\rightarrow$ (x SUB z)
For the referent *der Keller/the basement* (R_2) in sentence (19), there is no immediate antecedent in sentence (17). Here we need an axiom governing the inheritance of part-whole relationships:
(A2) (d1 SUB d2) $\wedge$ (d3 PARS d2) $\longrightarrow \exists$d4 [(d4 SUB d3) $\wedge$ (d4 PARS d1)]
and the common sense knowledge (Keller PARS Haus) or (basement PARS house), i.e. a typical house has a basement.[4]

Logical Recurrence and Bridging References. Bridging references are a type of reference where the antecedent is not directly mentioned in the foregoing text, i.e. an antecedent implicitly introduced has to be made explicit by logical inferences and background knowledge. A typical example is given by sentences (17) and (19), where meronymic knowledge (general properties of the part-whole relation PARS, and a part-whole relationship of two generic concepts) is needed to find the antecedent c_a for the concept c_r = c1511 described by *der Keller/the basement*. The semantic description D(c_r) of this phrase with the variable c_r is represented by (c_r SUB Keller/basement); this is also the question to be answered over the semantic network shown in Figure 4, where the meaning of sentence (17), in the following shortly denoted by sem(17), is represented on the left side by node c1508. The meaning of sentence (19) is represented on the right side by node c1509 (before the assimilation, the partial networks represented by nodes c1509 and c1508 are separated, and especially (c1511 PARS c1501) is missing).

The background knowledge of the previous paragraph and (A2) lead to the antecedent in sem(17) by means of the following backwards deduction:
(1) (c_r SUB Keller/basement) (Start with question)
(2) Unification of (1) with the right side of (A2), substituting basement for d_3 and a fresh constant c1000 for c_r, yields the new goal (d_1SUB d_2) $\wedge$ (basement PARS d_2).
(3) The first literal can be proved from the network sem(17) by the arc (c1501 SUB house) of sem(17), substituting c1501 for d_1 and house for d_2.
(4) The second literal can be derived from the meronymic background knowledge that (basement PARS house).

[4]Axiom (A2) means: If a concept d_2 superordinated to a concept d_1 is known to have a part d_3, then there must exist a more specific part d_4 of d_1 subordinated to d_3.

Applying the proposed assimilation mechanism to the inclusion reference for *das Gebäude/the building* in sentence (18), D(c_r) = (c_r SUB building), and using as a KB sem(17), axiom (A1), and the relationship (house SUB building), one obtains node c1501 of representation (17) (left side in Figure 4) as the antecedent c_a to be identified with c_r.

From the above, it can easily be seen that assimilation itself heavily depends on the availability of background knowledge, especially common sense knowledge. Thus, in building a large KB, one has to use a kind of bootstrapping process. Starting with some kernel of knowledge which is manually prepared using the workbench of the knowledge engineer, NLP techniques based on MultiNet technology can be used to automatically enlarge the background KB (vor der Brück and Helbig, 2010; vor der Brück, 2010). And this knowledge again can be used in the assimilation process to build even larger KBs.

4 CONCLUSIONS

The assimilation of knowledge derived from pieces of textual information into existing KBs plays a crucial role in AI. In this task, the knowledge representation formalism MultiNet and its software tools can be used as the central technological means. To the best of our knowledge, there is no other approach integrating so seamlessly and consistently all linguistic and logical processes as well as the computational lexicon and the background knowledge into one complex system for automatically building large KBs from textual archives. The power of this approach is witnessed by several real-life NLP applications developed in this framework, like question answering systems (Hartrumpf, 2005) based on corpora with millions of sentences, and NL interfaces to data bases (Leveling, 2006).

Semantic representations by means of the MultiNet formalism are applicable across different languages, which is investigated in a machine translation project (German – Chinese) and in a prototype of a semantically based search engine working on English documents. The MultiNet paradigm was also used for building large semantically based computational lexica (Hartrumpf et al., 2003). The techniques described in the paper were utilized for automatically translating the German Wikipedia with its 60 million sentences into a coherent MultiNet KB.

The tremendous amount of information contained in such KBs is also the reason why it is practically impossible to use traditional measures from information retrieval (like precision and recall) to directly evalu-

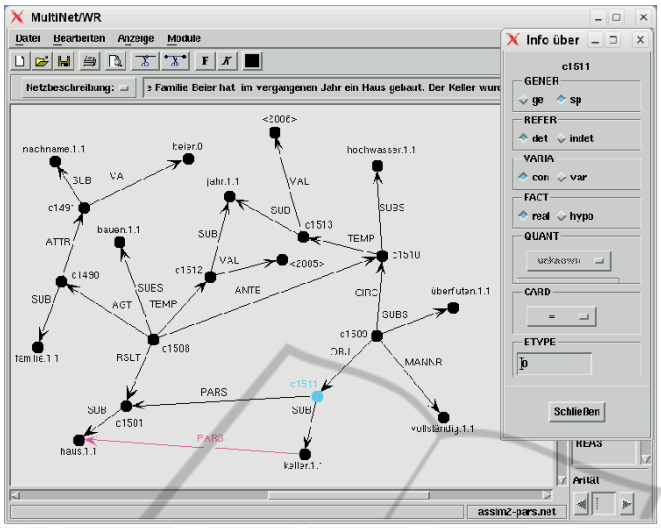

Figure 4: The semantic representation after the assimilation.

ate the performance of the analysis or the quality of the resulting KB, since nobody has a correct annotation of really large KBs with semantic networks as their meaning representation for comparison. So, the best way in the future seems to be to indirectly measure the quality of the processes described and the quality of the resulting KBs by judging the improvement of the application systems based on them. For example, precision and recall of a meaning-oriented search engine increases by about 10% (depending on the test set) when using a KB derived from the German Wikipedia. Similar improvements are observed in a deep, MultiNet-based question answering system.

REFERENCES

Baumgartner, P. and Kühn, M. (2000). Abducing coreference by model construction. *Journal of Language and Computation*, 1:193–209.

Copestake, A., Flickinger, D., Sag, I., and Pollard, C. (2005). Minimal recursion semantics. *Journal of Research on Language and Computation*, 3:281–332.

Ge, N., Hale, J., and Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proc. 6th Workshop on Very Large Corpora*.

Gnörlich, C. (2002). *Technologische Grundlagen der Wissensverwaltung für die automatische Sprachverarbeitung*. PhD thesis, FernUniversität Hagen.

Hartrumpf, S. (2003). *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany.

Hartrumpf, S. (2005). Question answering using sentence parsing and semantic network matching. In Peters et al., C., editor, *5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, pages 512–521. Springer.

Hartrumpf, S., Helbig, H., and Osswald, R. (2003). The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105.

Helbig, H. (2006). *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin.

Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Klavans, J. L. and Resnik, P., editors (1996). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Language, Speech, and Communication. MIT Press, Cambridge, Massachusetts.

Leveling, J. (2006). *Formale Interpretation von Nutzeranfragen für natürlichsprachliche Interfaces zu Informationsangeboten im Internet*. Der andere Verlag, Tönning, Germany.

Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–47, Philadelphia, Pennsylvania.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic Compositionality Through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1201–1211.

vor der Brück, T. (2010). Hypernymy extraction using a semantic network representation. *International Journal of Computational Linguistics and Applications*, pages 243–250.

vor der Brück, T. and Helbig, H. (2010). Retrieving meronyms from texts using an automated theorem prover. *Journal of Language Technology and Computational Linguistics*, 25(1):57–81.