Automatic Knowledge Acquisition by Semantic Analysis and Assimilation of Textual Information

Ingo Glöckner, Sven Hartrumpf, and Hermann Helbig
Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
http://pi7.fernuni-hagen.de/

Abstract

Automatic knowledge acquisition is one of the bottlenecks in artificial intelligence and large-scale applications of natural language processing (NLP). There are many efforts to create large knowledge bases (KBs) or to automatically derive knowledge from large text corpora. On the one hand, we meet KBs like CYC, where a tremendous amount of work has been invested by knowledge enterers who have manually formalized large stocks of knowledge. The other extreme are projects using flat (mostly statistically based) methods for extracting knowledge from texts. These techniques seldom produce results with a clear semantic interpretation and sufficient quality for NLP applications, however. MACQUIK is a project to automatically acquire knowledge from natural language sources (like text corpora or lexicons) by means of a deep syntactico-semantic analysis and subsequent assimilation of the generated representations into a coherent KB. The paper emphasizes the role of a homogeneous formalism for interfacing between NLP and inferential question answering, and it demonstrates its use for a deductive treatment of coreference resolution.

1 Introduction

To build real-life question answering (QA) systems or text understanding systems based on natural language processing (NLP) and logical reasoning one needs large stocks of background knowledge, be it lexical knowledge or world knowledge, encoded as a knowledge base (KB) suitable for reasoning. There are many KBs concentrating on ontological knowledge only (e.g. UNL-KB), using a restricted repertory of relations to structure their conceptual world; some of these KBs are used in an attempt to realize the so-called Semantic Web (Davies, 2003; Fensel et al., 2003). These ontological KBs are often based on some form of Description Logics (Baader et al., 2003).

Currently, there are few knowledge representation (KR) systems to offer the descriptive means needed for a comprehensive and cognitively oriented semantic description. A good formalism should be suitable both for natural language (NL) analysis and logical inference in order to avoid lossy transformations. Consider CYC, for example – one of the largest attempts to build a comprehensive KB of common sense knowledge with millions of elementary facts (Lenat and Guha, 1990). Its main problem has been its proliferation of artificial role names, which hinders the development of a computational lexicon for a coupling of CYC with NL, and apart from the ParGram project (Crouch and King, 2005), no use of CYC in large-scale NLP systems is known to us. Because of the technical difficulties associated with deep semantic text analysis and the acquisition of formalized background knowledge, many NLP systems dispense with linguistically and logically founded methods and use statistical or pattern-based methods instead (Brill et al., 2001; Lita and Carbonell, 2004). While such systems show an astonishing robustness and good results in certain applications, they will not lead to a real text understanding in the long run.

Another line of work tries a middle course in using *shallow* methods which balance the quality and effort of linguistic analysis. Such application-oriented methods are used in older text understanding systems like TACITUS (Hobbs et al., 1993) or in more contemporary QA systems, like QUETAL (Frank et al., 2005), FALCON (Harabagiu et al., 2000), and COGEX (Moldovan et al., 2003). Bobrow et al. (2005) present a basic logic for textual inference which incorporates conceptual structure (concepts and roles), contextual structure (situational embedding and propositional attitudes), and facticity. However, it is not yet clear whether or not the *contexted* description logic underlying this proposal is strong enough to capture the expected inferences.

To sum up, there are still no large-scale KBs automatically generated by NLP methods. We attribute this to the lack of formalisms which are (a) useful for the specification of lexical knowledge, for constructing the semantic representation during parsing, and for inferences in QA (homogeneity requirement), and (b) able to assign meaning representations to unconstrained text (universality requirement). In this paper, we start from the MultiNet paradigm of semantic networks (Helbig, 2006) which was specifically designed to meet these requirements in order to be suitable for the semantic description of open text. The MACQUIK (**M**ultiNet **Acqui**res **K**nowledge) approach presented here is essentially a two-step method working on existing texts.[1] Step I consists in the syntactico-semantic analysis of texts and the construction of a semantic network for each individual sentence. Step II involves the combination of the isolated semantic networks generated in Step I into a coherent large KB (so-called 'assimilation'). This step focuses on the task of knowledge acquisition and integration.

We discern the *intratextual assimilation*, which is mainly concerned with the resolution of coreferences and the reconstruction of contextual relations, and *intertextual assimilation*, which is concerned with the identification of individuals across texts and with event tracking (Ahn et al., 2006). In this paper, we only consider the intratextual case.

The problem of coreference resolution has received plenty of scientific attention (Kamp and Reyle, 1993; Hobbs et al., 1993; Ge et al., 1998; Cardie and Wagstaff, 1999; Harabagiu and Maiorano, 2000). To achieve our long-term goal of generating large KBs, we need a method which scales up well and also accounts for the interactions between reasoning (knowledge) and reference resolution. One of the first approaches which integrates these aspects was presented by Hobbs et al. (1993), who propose the use of abduction for interpreting pronouns and nominal anaphora. The weighted abduction scheme selects a single best interpretation, which may turn out false at a later point in the discourse. This problem is avoided by model-building approaches which keep track of all alternatives simultaneously (Baumgartner and Kühn, 2000; Gardent and Konrad, 2000). The model construction technique also gives a natural account of bridging references (Cimiano, 2006). However, the computational effort of keeping track of *all* interpretations prevents the use of these methods for larger texts. In other words, scalable approaches must commit to a single *best* interpretation (like (Hobbs et al., 1993)), but they should improve upon ad hoc weighting methods. In any case, the selection of the *best* interpretation must be based on extra-logical criteria and it makes sense to combine deductive techniques and other symbolic approaches with numerical quality metrics (e.g. Ng and Mooney's coherence measure (Ng and Mooney, 1990)) and with statistical coreference information (Ge et al., 1998; Cardie and Wagstaff, 1999). The CORUDIS system (Hartrumpf, 2003; Hartrumpf, 2001) used by MACQUIK combines rule-based and statistical methods for coreference resolution of pronouns and nominal anaphora.

Statistical methods or pattern-directed methods of knowledge acquisition mainly work on word level (Geleijnse and Korst, 2006; Pennacchiotti and Pantel, 2006; Romano et al., 2006), and even conceptual networks like ConceptNet (Liu and Singh, 2004) do not properly deal with the disambiguation of word meanings. In contrast, MACQUIK builds on clearly distinguished word senses maintained in the computational lexicon HaGenLex (Hartrumpf et al., 2003). MACQUIK has already proved its value in the automatic creation of KBs with millions of facts, where the degree of connectedness of a KB depends only on the provision of sufficient background knowledge (the latter is especially important for the resolution of *bridging references*, see Sect. 2.3).

2 Automatic Knowledge Acquisition by Assimilation

The search for explicitly or implicitly introduced identical concepts used in different parts of a text and their fusion into one semantic representative during the successive transformation of this text into one integrated KB is the main task of the assimilation process. To discuss the subtasks to be solved and the difficulties connected with assimilation, we start with some sample sentences. The semantic representation of (S1) is assumed to be the basic information, represented already in the KB, while (S2a) and (S2b) are assumed as possible text continuations following (S1).

(S1) "*Familie Beier hat im vergangenen Jahr ein Haus gebaut.*" ("*Last year, the Beier family built a house.*")

(S2a) "*Bald danach waren sie über die Qualität des*

[1] In contrast, collaborative projects like Learner (Chklovski, 2003) aim at constructing a large KB of commonsense knowledge from volunteer contributions.

Gebäudes zerstritten." ("*Soon afterwards, they had a quarrel about the quality of the building.*")
(S2b) "*Der Keller wurde beim diesjährigen Hochwasser vollständig überflutet.*" ("*The basement was completely overflowed by this year's flood.*")
The meanings of sentences (S1) and (S2b) are represented as semantic networks in Figure 1 in the left and right window, respectively. The windows display the results that the syntactico-semantic analysis (the WOCADI parser, (Hartrumpf, 2003)) delivers for the two isolated sentences (Step I of MACQUIK). Finally, Figure 2 represents the outcome of the assimilation process (Step II of MACQUIK) after joining the semantic networks of sentences (S1) and (S2b) into one KB.

MultiNet is a semantic network formalism whose nodes describe conceptual entities and whose arcs correspond to relations between these entities. (For a detailed description of MultiNet, see (Helbig, 2006).) Every arc is labeled by a member of a fixed set of relations and functions. Every node is classified according to a predefined conceptual ontology forming a hierarchy of sorts. The nodes also have rich descriptions in terms of predefined *layer attributes* which determine the kind of reference (REFER), the extensionality type (ETYPE) and the generality type (GENER), for example. In addition to the sorts, MultiNet allows for characterization of nodes by semantic features like [ANIMATE $+/-$], [GEOGRAPHIC $+/-$], or [MOVABLE $+/-$], which are also used in the computational lexicon for describing selectional restrictions (valencies). The knowledge about a given concept represented by a node c is enclosed in a conceptual capsule which is divided into three parts: the categorical knowledge holding unrestrictedly, the prototypical knowledge interpreted as default knowledge, and the situational knowledge. The first two parts together constitute the definitional knowledge $D(c)$ which is important for the resolution of references (see Sect. 2.3).

2.1 The Resolution of References Induced by Proforms

The most important types of reference are anaphoric (backward pointing), cataphoric (forward pointing), and deictic (pointing to the situational context). All of these types are often expressed by proforms (pronouns and proadverbs). An example of an anaphoric reference is "*sie*"/"*they*" in (S2a), which refers to the antecedent *Familie Beier*/*Beier family* in (S1). To resolve this reference, one needs the background knowledge that *family* represents a collection (expressed in MultiNet by the layer attribute constraint $[\text{ETYPE} = 1]$, see right part of Figure 2), as opposed to ordinary entities (e.g. *basement*) with $[\text{ETYPE} = 0]$). This information is provided by HaGenLex.

Proadverbs, like "*here*" (local deixis), and semantically related expressions, like "*last year*" (temporal deixis) as in (S1), often refer to elements not explicitly introduced by the foregoing text. By decrementing the current year Y, which the system fetches from its dialog model or meta-knowledge about the publication time of the text, "*last year*" is correctly interpreted as the year $Y-1$ (see node $c1512$ in Figure 2). More details about the resolution of temporal deictics within the MACQUIK setting can be found in (Hartrumpf and Leveling, 2006). If the situational context itself is described by a larger MultiNet network, the agreement of sorts (sort [t] – *temporal entity* – for temporal deixis, sort [l] – *local entity* – for local deixis) or even semantic features, like [GEOGRAPHIC $+$] or [HUMAN $+$], attached to concepts in the KB can be of great help to disambiguate between multiple reference candidates. The latter feature would help to find referents for pronouns like "*he*" or "*she*" because the selectional restrictions imposed on the pronoun and its antecedent must be compatible.

2.2 Ontologically Based References

References in a text are often characterized by the use of hypernyms or synonyms to remention entities already introduced in the discourse. Since lexical relations like conceptual subordination (relation SUB) and synonymy (SYNO) are characteristic of ontologies, references based on them will be called 'ontological references'. The relation SUB is essential for handling expressions of the form ⟨definite article/demonstrative determiner⟩ ⟨noun denoting a superordinated concept⟩. This reference type is met in (S2a): the phrase "*des Gebäudes*"/"*the building*" points to the house introduced in (S1). In many cases, such a reference (also called *inclusion*) is mediated over several steps in the subordination hierarchy and the transitivity of SUB must be taken into account:
(A1) $\text{SUB}(x,y) \wedge \text{SUB}(y,z) \rightarrow \text{SUB}(x,z)$
If one substitutes the word "*Gebäudes*"/"*building*" in (S2a) by "*Heims*"/"*home*" one must utilize synonymy to find the antecedent. These references can be handled by the deductive method presented in Sect. 2.3.

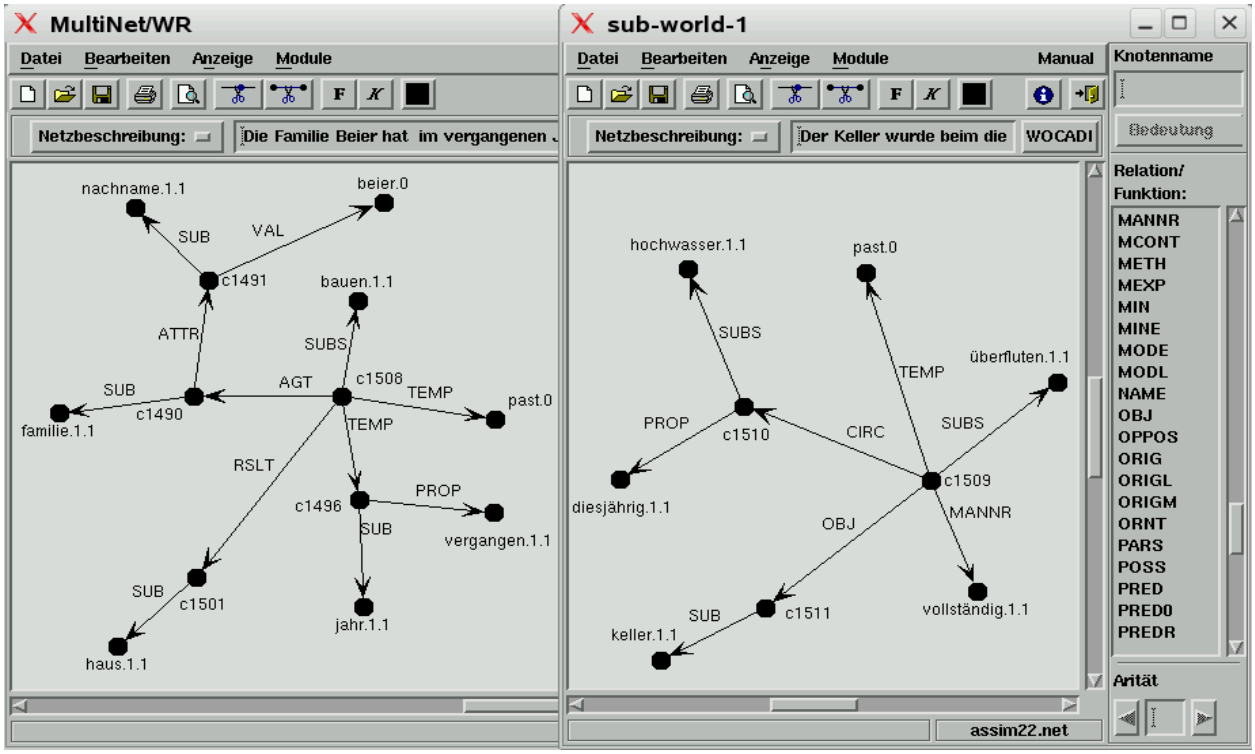

Figure 1: The semantic representation of sentences (S1) and (S2b) before assimilation

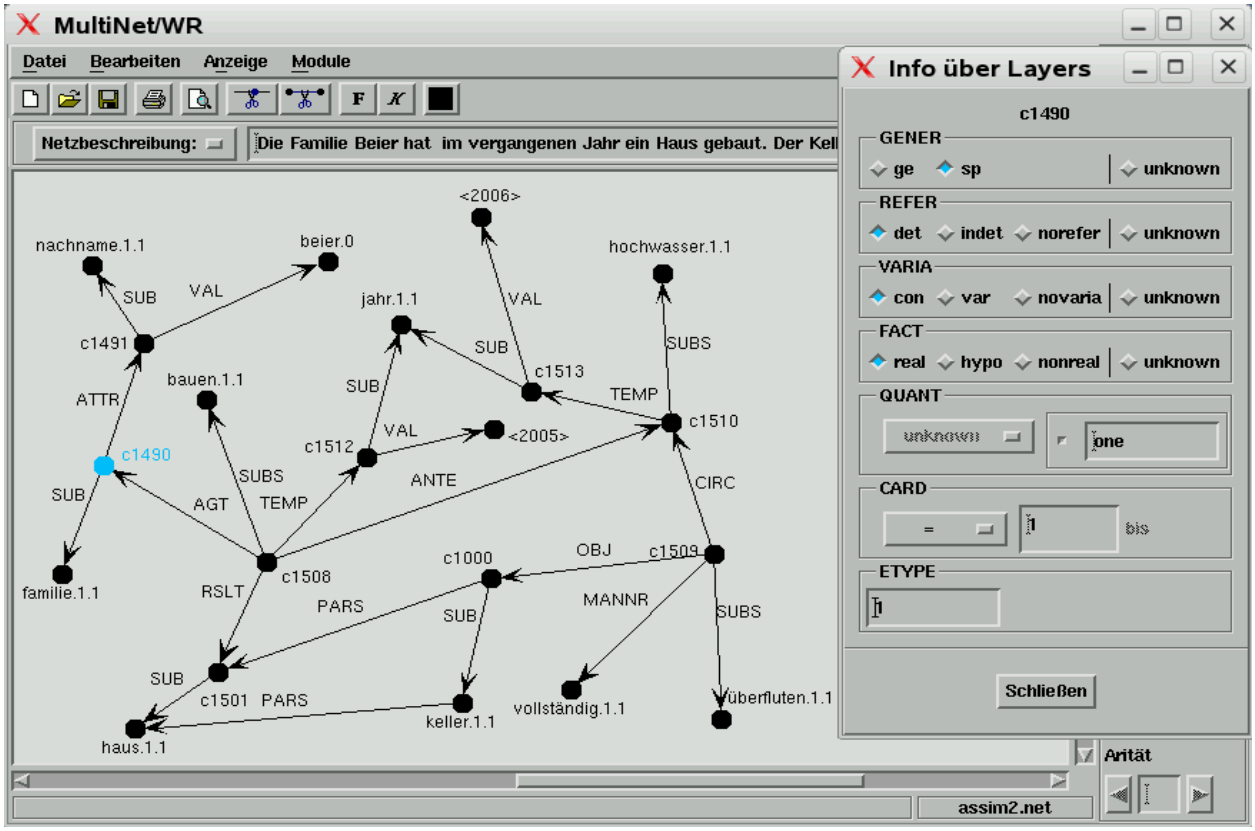

Figure 2: The semantic representation after assimilation

The coreference phenomena for NPs described in Sect. 2.1 and 2.2 are tackled in MACQUIK by the CORUDIS system (see Sect. 1). CORUDIS determines possible coreference alternatives which can serve as a starting point for the deductive techniques described below. If background knowledge is missing, computation is too slow, or other problems occur, the results of CORUDIS can be seen as a fallback strategy.

2.3 Logical Recurrence and Bridging References

The basic technique used to resolve non-pronoun references can be paraphrased as 'reference resolution by deduction'. The idea underlying this process and also the assimilation process as a whole is the following.[2] Consider a sentence S, whose semantic interpretation $\mathrm{sem}(S)$ has to be assimilated into an already existing KB K. MultiNet employs the layer attribute constraint $[\mathrm{REFER} = \mathit{det}]$ for a semantic network node $c_r \in \mathrm{sem}(S)$ which signals that c_r stems from a definite description in S and must hence be resolved from K (or from the dialog model). The node c_r is characterized by its definitional knowledge $D(c_r)$ which comprises all edges needed to express what the node stands for.[3] The logical expression describing $D(c_r)$ is interpreted as a question to be answered or, in technical terms, a theorem to be proved over K. In this inference process, c_r is a variable to be substituted by a known node c_a (the antecedent) contained in K. The central step in assimilating $\mathrm{sem}(S)$ into K is the identification of the nodes c_r and c_a and merging them into one node c of the extended KB arising from K and $\mathrm{sem}(S)$. In this sense the assimilation is a function $A : \mathscr{K} \times \mathscr{T} \to \mathscr{K}$ mapping an old KB $K \in \mathscr{K}$ and a meaning representation $\mathrm{sem}(S) \in \mathscr{T}$ to another (extended) KB $K' \in \mathscr{K}$.

For *bridging* references (Asher and Lascarides, 1999; Clark, 1977), the antecedent node c_a is not explicitly contained in K but must be inferred from the implicit knowledge provided by axioms. In order to cope with these references, coreference resolution must therefore incorporate background knowledge and logical inference. A typical example is given by sentence (S2b), where meronymic knowledge is needed to find the antecedent c_a for $c_r = c1511$ described by "*der Keller/the basement*". The semantic description $D(c_r)$ of this phrase involving the variable c_r is given by $\mathrm{SUB}(c_r, \mathit{basement})$. This is also the theorem to be proved from the semantic network $\mathrm{sem}(S1)$ shown in Figure 1 (left side). The inheritance of the part-whole relation PARS within a hierarchy of conceptual subordinations is defined by the axiom:

(A2) $\mathrm{SUB}(d_1, d_2) \wedge \mathrm{PARS}(d_3, d_2) \rightarrow$
$\exists d_4[\mathrm{SUB}(d_4, d_3) \wedge \mathrm{PARS}(d_4, d_1)]$[4]

The following inference steps lead to the antecedent in $\mathrm{sem}(S1)$:

(1) $\mathrm{SUB}(c_r, \mathit{basement})$ – Start with the question.

(2) Unification of (1) with the right-hand side of axiom (A2) based on the substitution $\sigma_0 = \{c_r/c1000, d_4/c1000, d_3/\mathit{basement}\}$ for an arbitrary fresh constant $c1000$, yields $\mathrm{SUB}(d_1, d_2) \wedge \mathrm{PARS}(\mathit{basement}, d_2)$ as the new subgoal to be derived.

(3) The first literal can be proved from the arc $\mathrm{SUB}(c1501, \mathit{house})$ of the network $\mathrm{sem}(S1)$, using the substitution $\sigma_2 = \{d_1/c1501, d_2/\mathit{house}\}$.

(4) Due to the substitution σ_2, the second literal of the subgoal now becomes $\mathrm{PARS}(\mathit{basement}, \mathit{house})$. It can be derived from the meronymic background knowledge.

Since the left-hand side of (A2) has been proved by steps (3) and (4), the right-hand side of (A2) must also hold because of modus ponens. According to the substitutions σ_0 through σ_2, we obtain the derived literals $\mathrm{SUB}(c1000, \mathit{basement}) \wedge \mathrm{PARS}(c1000, c1501)$. These literals which describe $c1000$ will be added to K, and the referring node $c1511$ will be merged with $c1000$ obtained by resolving the bridging reference.

Applying the assimilation mechanism described above to the inclusion reference induced by the phrase "*this building*" (see Sect. 2.2), with $D(c_r) = \{\mathrm{SUB}(c_r, \mathit{building})\}$, and using as a KB $\mathrm{sem}(S1)$, axiom (A1), and the background knowledge $\mathrm{SUB}(\mathit{house}, \mathit{building})$, one obtains node $c1501$ of representation $S1$ (Figure 1) as antecedent node c_a to be identified with c_r.

[2] In the following, the assimilation process is denoted by A; $\mathscr{K}$ is the set of KBs successively generated by A; $\mathscr{T}$ is the set of all meaning representations of isolated sentences to be assimilated into that $K \in \mathscr{K}$ which has been generated last by A; $K \in \mathscr{K}$ consists of the knowledge assimilated so far, the general background knowledge, and the axioms.

[3] For quantifying nodes, the defining edges correspond to the restriction of the quantifier, while the non-defining or *assertional* edges form its nuclear scope.

[4] This axiom means: If a superconcept d_2 of concept d_1 is characterized by having a part d_3, then there must exist a more specific part d_4 of d_1 subordinated to d_3.

2.4 Thematic Roles and Textual Coherence

It is typical of languages like German and English that the names of concepts superordinated to the entities filling a certain participant role of an activity can be systematically derived by special axiom schemata of morpho-lexical character.

(A3) $\mathrm{SUBS}(v, \langle \textit{activity} \rangle) \wedge \mathrm{AGT}(v, p)$
$\rightarrow \mathrm{SUB}(p, \langle \textit{activity} \rangle \mathrm{er})$
with $\langle \textit{activity} \rangle \in \{\textit{build}, \textit{teach}, \textit{work}, \ldots\}$

Schema (A3) says that the agent of a building activity is a builder, the agent of a singing activity is a singer, etc. Thus, if we know that some x teaches, then we can refer to x by the phrase "*the teacher*". Using axiom schema (A3) one can employ the same inference procedure as described in Sect. 2.3. The only preparation step is to generate a corresponding axiom from (A3) by substituting *teach* for ⟨activity⟩. Which activities are ruled by (A3) is anchored in HaGenLex.

2.5 Spatio-temporal Structure Inherent in Textual Information

There are also language phenomena not connected with reference mechanisms which are still important for textual coherence. The prime example are temporal or spatial relationships between events, described in the text by adverbial constructs or (in the case of implicit temporal relationships) by the deliberate use of different tenses of the verb (see, for instance, (Reichenbach, 1947)). Even the simple succession of sentences in a text (narrative order) often establishes a temporal relationship between the events described by them.[5] One example is given by sentences (S1) and (S2b). After resolving the deictic references induced by the phrases "*im vergangenen Jahr*"/"*last year*" of (S1) and "*diesjährig*"/"*this year's*" of (S2) one gets the corresponding representations of nodes $c1512$ and $c1513$ in Figure 2. The values of the attribute *Jahr*/*year* attached to these nodes give rise to a temporal relation between the nodes $c1508$ and $c1510$, which is deduced by MACQUIK and represented by the ANTE arc in Figure 2. This ANTE relation could be further transferred to the nodes $c1508$ and $c1509$ by means of an axiom relating the time of an event and its circumstance:

(A4) $\mathrm{CIRC}(v, w) \wedge \mathrm{TEMP}(w, t) \rightarrow \mathrm{TEMP}(v, t)$

This axiom allows the resolution of temporal expressions like "*two years after the overflowing*".

[5] See (Cimiano, 2006) for a method which reconstructs rhetorical relations including explanation and narration.

3 Conclusion

MultiNet has already proved its value in several real-life NLP applications, like QA systems (Hartrumpf, 2006) or NL interfaces to data bases and to the internet (Leveling and Helbig, 2002). It has also been used for building large semantically based computational lexica (Hartrumpf et al., 2003) and for the semantic annotation of different text corpora. Such quantitatively demanding tasks can not be solved without technological support.

The assimilation process as described in the paper is supported by the technological environment developed for MultiNet and by the computational lexicon HaGenLex. The semantic networks shown in Figure 1 and 2 have been generated by the MWR tool, a workbench for the knowledge engineer (Gnörlich, 2002), which has access to the WOCADI parser and therefore also to CORUDIS. Lexicon development is facilitated by LIA, a workbench for the computer lexicographer (Osswald, 2002).

The software tools support the main steps in our MACQUIK approach: (I) Translating single sentences of a large text into their meaning representations, expressed in the MultiNet formalism. This step is carried out by the WOCADI parser. (II) Integration of semantic networks representing the meaning of single sentences into a larger KB representing the meaning of whole texts. The software tool supporting this assimilation step is the knowledge engineering workbench MWR.

On the basis of these tools one can build KBs[6] and implement intelligent QA by combining methods of NLP with logical inference. This also allows a more organic inclusion of background knowledge which is not so easy for flat methods or even impossible for statistically based techniques. The InSicht QA system demonstrates the utility of anaphora resolution for improving QA results when the answer depends on more than one sentence in the text. The MultiNet KB used by InSicht was generated from the 5 million sentences in the QA@CLEF corpus and further elaborated by assimilation (Hartrumpf, 2006). InSicht also serves as a testbed for temporal deixis resolution within the MACQUIK setting. The benefits of deixis resolution for QA, which involves determining publication dates of texts from metadata or the texts themselves, were evaluated in (Hartrumpf and Leveling, 2006). Finally MAVE

[6] For example, we are building a large coherent KB from the German Wikipedia. Several smaller MultiNet KBs were constructed from biographical, juridical, and medical texts.

(Glöckner, 2006), a system for logical answer validation, uses assimilation (based on the CORUDIS data) and logical inference on the resulting MultiNet KB for enhancing the results of QA systems by a subsequent plausibility filter. Based on these techniques, MAVE scored best in the CLEF 2006 answer validation exercise for German.

The layered structure of a MultiNet KB, created by MACQUIK, which clearly discerns the generic and episodic information contained in a text, makes it possible to extract the generic knowledge for use in NLP applications, keeping it as a general knowledge background. It should be noted that, while using MultiNet as a KR language, the phenomena discussed and the methods to deal with them are of general importance and independent of the KR formalism.

References

D. Ahn, S. Schockaert, M. De Cock, and E. Kerre. 2006. Supporting temporal question answering: Strategies for offline data collection. In *Proc. of ICoS-5*, Buxton, UK.

N. Asher and A. Lascarides. 1999. Bridging. *Journal of Semantics*, 15:83–113.

F. Baader, D. Calvanese, D. McGuiness, D. Nardi, and P. Patel-Schneider, editors. 2003. *The Description Logic Handbook*. Cambridge University Press, Cambridge, UK.

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

D. Bobrow, C. Condoravdi, R. Crouch, R. Kaplan, L. Karttunen, T.H. King, V. de Paiva, and A. Zaenen. 2005. A basic logic for textual inference. In *Proc. AAAI Workshop on Inference for Textual Question Answering*, Pittsburgh, PA.

E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. 2001. Data-intensive question-answering. In *Proc. of TREC-10*.

C. Cardie and K. Wagstaff. 1999. Noun phrase coreference as clustering. In *Proc. of EMNLP/VLC-99*, pages 82–89, College Park, MD.

T. Chklovski. 2003. LEARNER: A system for acquiring commonsense knowledge by analogy. In *Proc. of 2nd Int. Conf. on Knowledge Capture (K-CAP 2003)*.

P. Cimiano. 2006. Ingredients of a first-order account of bridging. In *Proc. of ICoS-5*, Buxton, UK.

H.H. Clark. 1977. Bridging. In P.N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press.

R. S. Crouch and T. H. King. 2005. Unifying lexical resources. In *Proc. Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 32–37.

J. Davies. 2003. *Towards the Semantic Web*. John Wiley & Sons.

D. Fensel, W. Wahlster, H. Lieberman, and J. Hendler. 2003. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, Cambridge, MA.

A. Frank, H.-U. Krieger, F. Xu, H. Uszkoreit, B. Crysmann, B. Jörg, and U. Schäfer. 2005. Querying structured knowledge sources. In *Proc. of AAAI-05, Workshop on Question Answering in Restricted Domains*, pages 10–19, Pittsburgh, PA.

C. Gardent and K. Konrad. 2000. Interpreting definites using model generation. *Journal of Language and Computation*, 1:193–209.

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

G. Geleijnse and J. Korst. 2006. Learning effective surface text patterns for information extraction. In *Proc. of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, pages 1–8, Trento, Italy.

I. Glöckner. 2006. University of Hagen at QA@CLEF 2006: Answer validation exercise. In *Working Notes for the CLEF 2006 Workshop*.

C. Gnörlich. 2002. *Technologische Grundlagen der Wissensverwaltung für die automatische Sprachverarbeitung*. Ph.D. thesis, FernUniversität Hagen, Hagen, Germany.

S. Harabagiu and S. Maiorano. 2000. Multilingual coreference resolution. In *Proc. of (ANLP-NAACL'2000)*, Seattle, WA.

S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu. 2000. FALCON: Boosting knowledge for answer engines. In *Proceedings of Text REtrieval Conference (TREC-9)*.

S. Hartrumpf and J. Leveling. 2006. University of Hagen at QA@CLEF 2006: Interpretation and normalization of temporal expressions. In *Working Notes for the CLEF 2006 Workshop*.

S. Hartrumpf, H. Helbig, and R. Osswald. 2003. The semantically based computer lexicon Ha-

GenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105.

S. Hartrumpf. 2001. Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proc. of CoNLL-2001*, pages 137–144, Toulouse, France.

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

S. Hartrumpf. 2006. Extending knowledge and deepening linguistic processing for the question answering system InSicht. In *Accessing Multilingual Information Repositories*, volume 4022 of *LNCS*, pages 361–369. Springer, Berlin.

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

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

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

D.B. Lenat and R.V. Guha. 1990. *Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, Reading, MA.

J. Leveling and H. Helbig. 2002. A robust natural language interface for access to bibliographic databases. In *Proc. of SCI 2002*, volume XI, pages 133–138, Orlando, FL.

L.V. Lita and J. Carbonell. 2004. Instance-based question answering: A data-driven approach. In *Proc. of EMNLP 2004*, pages 396–403.

H. Liu and P. Singh. 2004. Commonsense reasoning in and over natural language. In *Proc. of 8th Int. Conf. on Knowledge-Based Intelligent Information & Engineering Systems (KES 2004)*.

D. Moldovan, C. Clark, S. Harabagiu, and S. Maiorano. 2003. COGEX: A logic prover for question answering. In *Proc. of the North American Chapter of the ACL on Human Language Technology*, volume 1, pages 87–93, Morristown, NJ.

H.T. Ng and R.J. Mooney. 1990. On the role of coherence in abductive explanation. In *Proc. 8th National Conference on AI*, pages 337–342, Boston, MA.

R. Osswald. 2002. *A Logic of Classification – with Applications to Linguistic Theory*. Ph.D. thesis, FernUniversität Hagen, Hagen, Germany.

M. Pennacchiotti and P. Pantel. 2006. A bootstrapping algorithm for automatically harvesting semantic relations. In *Proc. of ICoS-5*, Buxton, UK.

H. Reichenbach. 1947. *Elements of Symbolic Logic*. Free Press, New York.

L. Romano, M. Kouylekov, I. Szpektor, I. Dagan, and A. Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proc. of EACL-2006*, pages 409–416.