NL Domain Explanations in Knowledge Based MAT

Galia Angelova, Kalina Bontcheva[1]
Bulgarian Academy of Sciences, Linguistic Modelling Laboratory
Acad. G. Bonchev Str. 25A, 1113 Sofia, Bulgaria, {galja,kalina}@bgcict.acad.bg

Abstract

This paper discusses an innovative approach to knowledge based Machine Aided Translation (MAT) where the translator is supported by an user-friendly environment providing linguistic and domain knowledge explanations. Our project aims at integration of a Knowledge Base (KB) in a MAT system and studies the integration principles as well as the internal interface between language and knowledge. The paper presents some related work, reports the solutions applied in our project and tries to generalize our evaluation of the selected MAT approach.

1. Introduction

The notion of MAT comprises approaches where - in contrast to MT - the human user keeps the initiative in translation. MAT ranges between intelligent text editors and workbenches aiming at user modelling and partial MT. A principal problem, however, is the support of domain knowledge since it affects the quality of the translated text. Moreover, the time spent for domain familiarization is estimated as 30-40% of the total translation time (KiWi90).

This paper discusses an innovative approach to knowledge based MAT: a KB is systematically integrated in a framework providing linguistic as well as domain knowledge support. Linguistic support is assured by relevant resources: grammatical and lexical data. Domain explanations are generated from a KB of Conceptual Graphs (CGs) (Sow84). The system interface offers a standard dialog: while translating, the user highlights words/texts, chooses queries from menus and receives NL answers from where new requests can be started. The results reported here were achieved in the joint German-Bulgarian project DB-MAT[2] (vHa91, vHAn94).

Depending on the viewpoint, DB-MAT can be compared to various approaches and/or systems: *(i) encoding of term's meaning:* - to lexicons and termbanks (see 2.1) and knowledge based termbanks (see 2.2); *(ii) generation of explanations* allowing follow-up questions and clarifications - e.g. IDAS (see 2.3); *(iii) NL generation from CGs* (Bon96, AnBo96).

Below we present related approaches and the DB-MAT paradigm. Our opinion about the costs and benefits of the knowledge based MAT is clearly stated.

2. Related Work

Some approaches are now discussed with comments on the rationale of a knowledge based MAT design.

2.1. Lexicons and Termbanks

In (machine-readable) terminological lexicons domain knowledge is contained in text definitions. A conceptual hierarchy is sometimes sketched by pointers like *"see.."* in the definition: the super- and sister-concepts are related to the lexical entry, i.e. to the denoted concept. An intuitive unification of the lexical units and their implicit knowledge items is assumed.

Concept-oriented termbanks support a hierarchical skeleton of underlying concepts and thus the knowledge can be treated formally during its construction, use and updating. Text definitions, however, are written manually; the progress is that the lexical entries and conceptual elements are encoded independently. In monolingual termbanks, the term is a concept label as well as a lexicon item (e.g. CoAm93). In cases of bilingualism (e.g. Fis93) every NL has its own conceptual structure and translation equivalents are mappings of one conceptual structure onto another.

An evaluation for "domain knowledge in translation" reveals knowledge *content, organization* and *usage* : *(1)* conceptual knowledge is encoded in text definitions, written manually in multilingual environment. Domain facts are not included in any definition although they are very important for the understanding of technical texts; *(2)* knowledge is artificially segmented into text fragments, organized alphabetically around lexical entries. The time-consuming search for semantically related terms is to be performed by the reader; *(3)* the user gets the whole bulk of information without any opportunity to filter for relevant aspects. The "inheritance of features" along the hierarchy is to be made by the reader.

2.2. Terminological Knowledge Bases

The term meanings are encoded in formal languages instead of text definitions. Briefly we mention: *(1)* COGNITERM (MSBE92): the term meaning is represented in a frame-like structure which is accessible by names of concepts or their characteristics. For a new NL, another KB is built up "based on the translation equivalents provided for concepts in the source language KB" (SkMe90); *(2)* Translators' Work-Bench (TWB): the meanings of each term are described by CGs. A concept is related to several terms via synonyms and foreign NL equivalents (HoAh92).

The available examples present disconnected, though formal definitions of meanings. However, there is *(1)*

[1] Current: University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello Str., Sheffield S14DP, e-mail K.Bontcheva@dcs.shef.ac.uk

[2] Funded by Volkswagen Foundation (Germany) in 1993-1995. See also www.informatik.uni-hamburg.de/Arbeitsbereiche/NATS/projects/db-mat.html

no coherent, homogenous knowledge source for a systematic conceptual evaluation; *(2)* no access by a context sensitive user interface; *(3)* no theoretically sound solution for multilingual systems.

2.3. Generation of Explanations

There are similarities between DB-MAT and other NL generation systems, e.g. IDAS which produces technical documentation from a domain KB and linguistic and contextual models (RMLe95). In a sense IDAS builds an on-line user interface to KBs and provides system answers by NL generation. The system generates hypertext nodes (both text and links) with relevant follow-up questions. The following particularities however display the differences between the systems: *(1)* IDAS is a full-scale application, its KL-ONE like KB contains more domain information and the proper system evaluation can be performed; *(2)* the hypertext links act as visual hints for the available relevant information, while the user should "guess" that in DB-MAT; *(3)* the DB-MAT KB pretends to be arbitrary, i.e. we investigated the integration of arbitrary domain KB into applications in the humanities; as a contrast the IDAS KB contains fixed number of (task-adequate) "conceptual relations" and supports fixed query types; *(4)* the systems certainly have different interface design oriented towards different goals and user types.

3. Benefits of the Knowledge Based MAT

The optimal separation of domain knowledge as an independent source facilitates its structuring and processing and makes its theoretical foundation sound. KBs seem difficult to acquire (compared to informal textual lexicons), but this is not true with respect to formal and heterogeneous lexicon structures. Moreover, formal descriptions are built up increasingly both for research and industrial applications, e.g. formal specifications are developed by various acquisition tools. Thus DB-MAT aims at the reuse of KBs in a MAT system.

Keeping knowledge in a separate structure enables its processing with formal operations. Especially formal consistency and semantic coherency can be best achieved in a well-defined representation language. In DB-MAT the NL explanation semantics is kept as CGs as long as possible: we tailor the explanation to the users' expectations by the formal operations *projection* and *join*. By *inheritance* the adequate degree of detailness in the generated answer is provided (AnBo96).

In multilingual MAT, the CG type hierarchy proved to be particularly useful in case of e.g. terminological gaps. For missing translation equivalents, the type hierarchy provides NL explanations about the "relative position" of the highlighted term. The attributes of the node are verbalized in the source language to facilitate paraphrases in the target one (WiAn93).

4. DB-MAT

DB-MAT is a knowledge-based TWB providing linguistic as well as domain knowledge support. The system has a user-friendly interface, with a main window separated into two scrollable regions for the source and target texts. The translator selects the explanation language (*German, Bulgarian*) and the detailness of the generated explanations (*Less, More*) with radio buttons. DB-MAT provides figures as well, to facilitate the user's understanding of the domain. Currently all figures are associated to lexicon entries.

4.1. Main menu

Except **File** and **Edit** with their standard functionality, the main menu contains three task-specific items:
- Under **Note** the user selects flags (*<Check later>*, *<Gap>*, etc.) to be inserted in the text as reminders;
- **Information** provides monolingual support and access to available figures. Grammatical data from the lexicon is shown to the user. Under the submenu **Explanations**, a NL explanation is generated for terms while for non-terms a textual definition is given instead (the user should always get something without bothering from where the answer comes);
- **Multilingual** offers bilingual data.Under **Translations** the lexicon correspondences are presented. The other subitems are **Idioms** and **Examples**.

4.2. The Lexicon

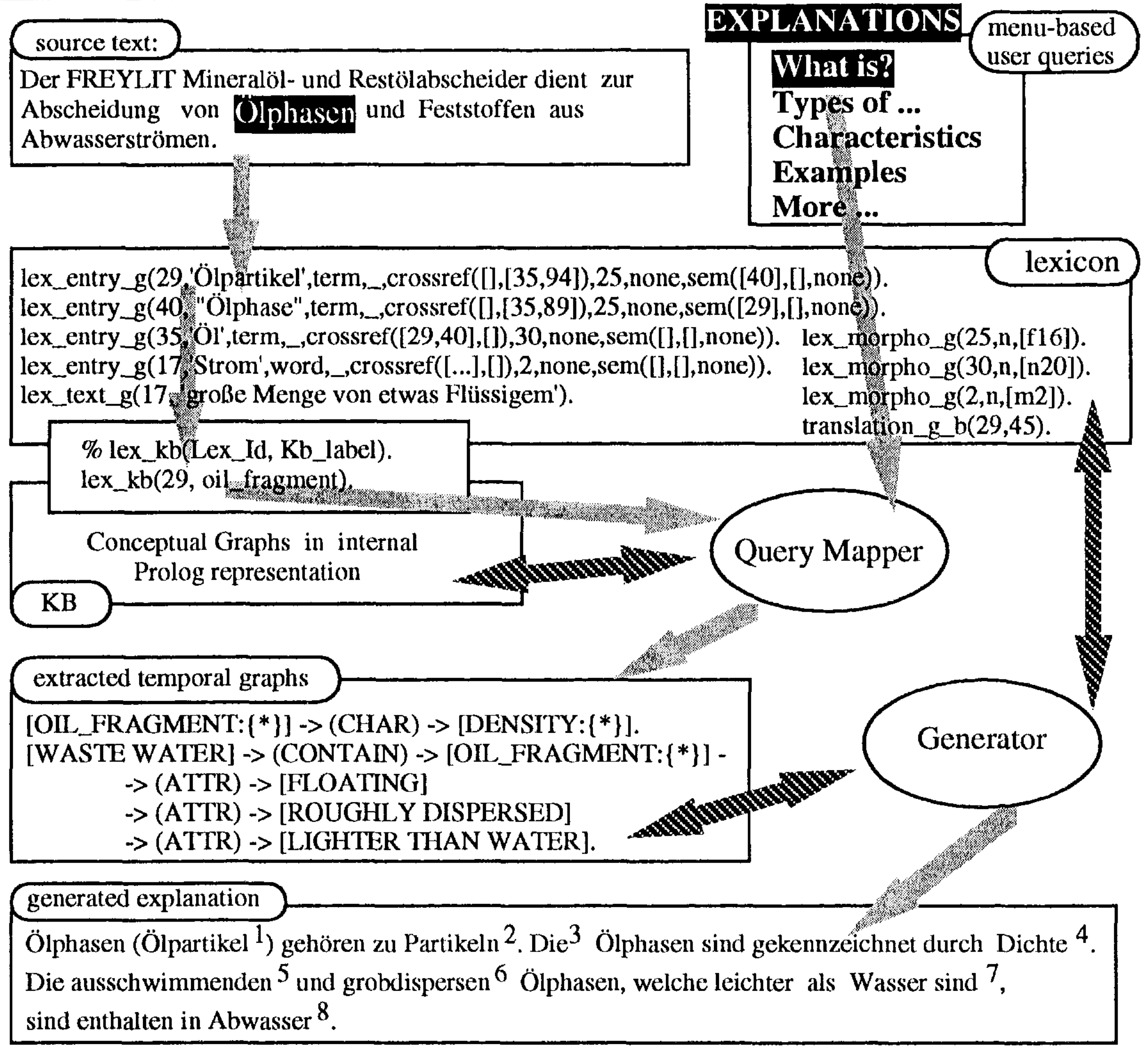
DB-MAT uses one lexicon, i.e. general vocabulary and terms are distinguished by checking whether text definitions or KB-pointers are available. There is one entry per meaning. Special links *contained in* and *consists of* (Fig.1, "crossref" of Ids #35, 29, 40), acquired semi-automatically, provide precise mappings of the chosen text segments onto lexicon items.

The lexicon contains (BoEu95): *(1)* morphological data organized in morpho-groups (part of speech, inflection class, verb types, etc); *(2)* syntactic information - syntax groups used by the NL generator only and some text strings (e.g. list of collocations); *(3)* synonyms (Ids #29, 40), antonyms, abbreviations; *(4)* text definitions for general vocabulary (e.g. Id #17); *(5)* references for bilingual correspondences.

4.3. The KB and the Query Mapper (QM)

The KB consists of concepts, a type hierarchy and conceptual graphs. Each graph is either a semantic definition of a term or contains certain factual knowledge. The QM, our "what to say" component, extracts as temporary graphs (by CG *projection*) knowledge fragments to be verbalized. There is no fixed predefined schema mapping a user request to some knowledge fragments. Given a highlighted term (i.e. its KB concept), and the user request for domain knowledge, the QM searches the KB on the fly and extracts all relevant facts according to the conceptual relations. Depending on the detailness level, all attributes and characteristics are inherited from a more generic node.

For each query type, the QM maintains a list of relevant conceptual relations. So far, the QM has a fixed scope of extraction: for most of the conceptual relati-



Der FREYLIT Mineralöl- und Restölabscheider dient zur Abscheidung von Ölphasen und Feststoffen aus Abwasserströmen.

```
lex_entry_g(29,'Ölpartikel',term,_,crossref([],[35,94]),25,none,sem([40],[],none)).
lex_entry_g(40, "Ölphase",term,_,crossref([],[35,89]),25,none,sem([29],[],none)).
lex_entry_g(35,'Öl',term,_,crossref([29,40],[]),30,none,sem([],[],none)).   lex_morpho_g(25,n,[f16]).
lex_entry_g(17,'Strom',word,_,crossref([...],[]),2,none,sem([],[],none)).   lex_morpho_g(30,n,[n20]).
lex_text_g(17,'große Menge von etwas Flüssigem').                            lex_morpho_g(2,n,[m2]).
                                                                             translation_g_b(29,45).
```

```
% lex_kb(Lex_Id, Kb_label).
lex_kb(29, oil_fragment).
```

```
[OIL_FRAGMENT:{*}] -> (CHAR) -> [DENSITY:{*}].
[WASTE WATER] -> (CONTAIN) -> [OIL_FRAGMENT:{*}] -
        -> (ATTR) -> [FLOATING]
        -> (ATTR) -> [ROUGHLY DISPERSED]
        -> (ATTR) -> [LIGHTER THAN WATER].
```

Ölphasen (Ölpartikel [1]) gehören zu Partikeln [2]. Die[3] Ölphasen sind gekennzeichnet durch Dichte [4]. Die ausschwimmenden [5] und grobdispersen [6] Ölphasen, welche leichter als Wasser sind [7], sind enthalten in Abwasser [8].

Figure 1. Generation of explanations: 1 - a synonym from the lexicon; 2 - the supertype from the KB type hierarchy, 3 - definite article because [OIL FRAGMENT] is already present in the context; 4 - the characteristic [DENSITY]; 5,6 - the attributes [FLOATING] and [ROUGHLY DISPERSED]; 7 - a relative clause for [LIGHTER THAN WATER] since it cannot be verbalised as an adjective; 8 - the second graph with focused [OIL FRAGMENT], in passive voice because the graph is traversed against the arcs.

ons it is "one step around" the selected concept. Nested graphs (e.g. situations) are extracted as unbreakable knowledge fragments due to their specific meanings. The explanation semantics is under certain control: the QM does not allow trivial answers like "Oil separator is a non-animated physical object" etc. Detailed discussion is given in (AnBo96).

4.4. The Generator (EGEN)

The generation algorithms are strongly influenced by some features of the CGs and their well-defined operations. An important asset of the CGs proved to be their non-hierarchical structure, allowing for the generation to start from any KB node without any graph transformations. Thus EGEN may select the subject and the main predicate of each sentence from a linguistic perspective rather than being influenced by the structuring of the underlying semantics (as with the frequently used tree-like notations).

EGENs input is: *(1)* the relevant knowledge pool; *(2)* the explanation language; *(3)* the highlighted concept(s) (the corresponding term(s) will become the global focus of the generated explanation); *(4)* the query type (necessary for the selection of an appropriate text-organisation schema); *(5)* an iterative call flag indicating a request for further clarification.

In order to produce a coherent explanation, EGEN orders the CGs by applying a suitable text organisation schema (AnBo96) - definition, similarity or difference (similar to those in McKe85). Afterwards the generator breaks some CGs into smaller graphs or joins similar ones into a single graph to ensure that each CG is expressible in a single sentence. Finally EGEN verbalizes the CGs by applying the utterance path approach - the algorithm searches for a cyclic path which visits each node and relation at least once. If a node is visited several times then grammar rules determine when and how it is verbalized. As proposed in (Sow84), concepts are mapped into nouns, verbs, adjectives and adverbs, while conceptual relations are mapped into "functional words" or syntactic elements. The Sowa's algorithm is extended (AnBo96) to: *(1)* process extended referents (e.g. measures, conjunctive and disjunctive sets, etc.); *(2)* group relevant features together (e.g. first utter all "dimension" attributes, then all "weight" attributes, etc. instead of mixing them up); *(3)* introduce relative clauses and conjunctions; *(4)* generate a sentence tree allowing for future transformations. The APSG grammar used by EGEN is implemented in Prolog.

Additionally, EGEN keeps all uttered concepts in a stack and later refers to them using a definite article or a pronoun. This stack is cleared in the end of the explanation, unless there is an iterative request.

A request for definition ("What is?") of "Ölphasen" is given on Fig. 1. Some relevant lexicon entries are presented. The QM extracts the supertype and the conceptual relations ATTR, CHAR and RESULT (AnBo96). The extracted temporal graphs are shown in linear notation. They contain all occurrences of the "highlighted" concept and the necessary conceptual relations. The QM has applied the *type contraction* operation in order to "simplify" the graphs. Thus there are complex concepts like [LIGHTER THAN WATER] which have corresponding type definitions.

5. Costs of the knowledge based MAT

It is difficult to acquire the interrelated lexicon/KB although once the KB is acquired, the reuse effect will decrease the costs of adding a new NL to the system. In DB-MAT we used special lexicon acquisition tools and we plan to develop tools with restricted NL interface for future KB acquisition. Our estimation is that DB-MAT resources are not more complicated than the lexicons in sophisticated MT systems, e.g. the KBMT lexicon and ontology (GoNi91). However, the proper use of AI-methods requires additional study, design efforts and evaluation experiments oriented towards knowledge based NLP.

6. Implementation and Conclusion

The DB-MAT demo is implemented in LPA MAC Prolog32. Special lexicon acquisition tools were developed. The German lexicon contains about 900 entries. The KB (about 300 concept nodes and 30 conceptual relations) was manually acquired from a textbook and encyclopedias in admixture separation. The lexicon covers a demo text but any MacProlog readable file demonstrates the DB-MAT features if it contains the basic terminology (enabling requests for domain explanations).

DB-MAT studies one of the possible applications of KB-methods to computational terminology and translation aid tools. Further research is aimed at: *(1)* building a larger KB; *(2)* development of a general methodology relating the terminology to the corresponding conceptual knowledge; *(3)* experiments with the role of negation; *(4)* improvement of the generation to ensure more elaborate and coherent output combining textual and graphical information.

Acknowledgments: We are grateful to the DB-MAT project leader Prof. Dr. Walther von Hahn (Hamburg University) for his support during all project stages. DB-MAT would not have been possible without his efforts concerning the general design and the German language data. We also thank all those people who contributed to the project and/or to the quality of our papers and presentations.

References:

[AnBo96] Angelova, G. and K. Bontcheva. *DB-MAT: Knowledge Acquisition, Processing and NL Generation using CGs.* To appear Proc. of ICCS-96, Sydney, Australia, August 1996 (Lecture Notes AI).

[BoEu95] Boynov, N. and L. Euler. *The Structure of the Lexicon and its Support in DB-MAT.* Report 1/95, Project DB-MAT, Univ. Hamburg, 5/1995.

[Bon96] K. Bontcheva. *Generation of Multilingual Explanations from CGs.* To appear in Mitkov, Nikolov (eds.), *'Recent Advances in NLP'*, Current Issues in Linguistic Theory 136, Benjamins Press.

[CoAm93] Condamines, A. and P. Amsili. *Terminology Between Language and Knowledge: an example of terminological knowledge base.* In [Schm93].

[GoNi91] Goodman, K. and S. Nirenburg (eds). *A Case study to KBMT.* Morgan Kaufmann Pub. 1991.

[Fis93] D. Fischer. *Consistency Rules and Triggers for Multilingual Terminology.* In [Schm93].

[HoAh92] Hook, S. and K. Ahmad. *Conceptual Graphs and Term Elaboration: Explicating (Terminological) Knowledge.* Tech.report, ESPRIT II No. 2315: TWB Project, University of Surrey, 1992.

[KiWi90] Kieselbach,C., H. Winshiers. *Studie zur Anforderungsspezifikation einer computergestuetzten Uebersetzerumgebung.* Studienarb.,Univ. Hamburg.

[McKe85] K. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate NL Text.* Cambridge Univ. Press, 1985.

[MSBE92] Meyer, I., D.Skuce, L.Bowker, K.Eck. *Towards a new generation of term. resources: an experiment in building a TKB.* COLING-92, 956-960.

[RMLe95] E. Reiter, C. Mellish and J. Levine. *Automatic Generation of Technical Documentation.* Applied AI, Vol. 9, No. 3, 1995, pp. 259-287.

[Schm93] K. Schmitz (Ed.), *Terminology and Knowledge Engineering*, Proc. 3rd Int. Congress, Cologne, Germany, August 1993.

[SkMe90] Skuce, D. and I. Meyer. *Concept Analysis and Terminology: A Knowledge-Based Approach to Documentation.* COLING-90, 56-58.

[Sow84] J. Sowa. *Concept. Structures: Information Processing in Mind and Machine.* Add.Wesley, 1984.

[vHa91] W. von Hahn. *Innovative Concepts for MAT.* Proceedings VAKKI, Vaasa 1992, pp. 13-25.

[vHAn94] v. Hahn, Walther and G. Angelova. *Providing Factual Information in MAT.* In Proc. Int. Conf. *MT: Ten Years On*, Cranfield, UK, Nov.1994.

[WiAn93] Winschiers, H. and G. Angelova. *Solving Translation Problems of Terms and Collocations Using a Knowledge Base.* Techn. report 3/93, Project DB-MAT, University of Hamburg, December 1993.