Towards Text-Based Machine Translation

Jörg Schütz and Bärbel Ripplinger
IAI
Martin-Luther-Straße 14
D-66111 Saarbrücken
[joerg, babs]@iai.uni-sb.de

Abstract

The increasing use of world wide networking facilities for international information exchange and service provision demands for specific applications of machine translation (MT) tools which are beyond their traditional tasks. The new areas of application of MT will have to be centered around various human-to-human communication media taking into account the different cultures inherent in natural language. Thus, MT has to aim for cross-cultural communication when being used as an information and service provider between different natural languages on the information super-highways. Most of the information electronically exchanged or accessed deals with a specialized vocabulary (terminology).

In this paper we will argue for terminology-based MT and discuss the various knowledge sources, which will also ensure an appropriate basis for text-based MT. We have restricted our work to the domain of telecommunications, in particular that of satellite communication.

1 Introduction and Motivation

Over the last decade (1984-1994) machine translation systems, commercial systems as well as research prototypes, have steadily improved regarding their linguistic capabilities. However, they still lack the command of language a human translator possesses, in particular with regard to the interpretation of the textual units to be translated in their contextual, situational and cultural background environments. This includes, for example, the correct identification of ambiguous subject/object positions, as in *'Fernübertragungsausrüstungen umfassen auch Modulationsgeräte.'* (Telecommunication equipment also comprises modulating equipment.), of metaphorical meanings, like *'Der Prozeß wurde abgeschossen.'* (The process was killed), or of pro-forms within and across sentence boundaries.

Our aim is to provide for comprehensive knowledge resources which are organized so that sublanguage information and extra-linguistic information about a specific domain is accessible in a concise, efficiently machine-tractable form, and which are formalized so as to ensure consistency across organizations of related grammatical and lexical strata. However, the depth of the natural language (NL) analysis is restricted to the needs of the translation task, i.e. the access to the different information sources is defined by the application.

To achieve this, different knowledge sorts are used to build a *sublanguage information repository* which can be applied within a constraint-based natural language processing (NLP) framework. The approach has been implemented and tested in the Advanced Language Engineering Platform (ALEP) environment ([ALEP, 1993]), a general purpose NLP development platform, based on an object-centered architecture and a typed feature logic based linguistic formalism, promoted by the European Commission (EC) for the Linguistic Research and Engineering (LRE) action line and the forthcoming Fourth Framework Programme.

The primary driving force of our MT approach is the conceptual organization of the domain of telecommunications. Its purpose is to provide domain-specific constraints that ensure the control of the analysis, translation and generation process of sublanguage expressions. On the one hand, this is done by providing a model of the domain - the ontology - which represents the concepts of the domain and their generic and partitive relationships. On the other hand, knowledge about the terminology of the domain in terms of conceptual roles and conceptual modifiers defines the multidimensional relationships of the concepts of the domain. When linked together these knowledge sorts correspond to the *intensional meaning* of a sublanguage expression (proposition).

The development of the ontology has proceeded from two global research strands: the acquisition, organization and representation of knowledge in lexical and terminological resources, and their conceptual modelling. Since the overall purpose of the conceptual structure is to be maximally supportive for the computational processing of sublanguage expressions in an NLP environment, there is a third direction from which research on this topic has proceeded: the investigation of the linguistic realization of terminological expressions in their sentential and textual context of a corpus dealing with the domain and the investigation of the question of what the sentential and textual context may contribute to the (human and computational) interpretation of terms. Here, the main focus is on an extended conceptual and linguistic analysis of the corpora, which in particular takes into account the role of domain dependent and general language verbs within the specific subject field, and on how this analysis may support the entire conceptual analysis of the domain. This has included an investigation of different cultural systems, for example, a description of the TV culture of the US cannot be directly mapped to the German TV culture (although, today, they are much closer than they were a few years ago). An example of extreme cultural divergence is the field of virtual reality (VR) where most terms have no German equivalent, e.g. eyephone, tactor, virtuphone, speaktacles, simbim, automagically, telepresence, teledildonics, goggle-roving, etc.

The overall leading idea for the integration of the conceptual (terminological) knowledge into the linguistic processes is to control a *competence grammar* for general language by means of conceptual (terminological) constraints; this we call *performance control.* The actual engineering accomplishment is carried out entirely on a lexical basis by so-called *terminological anchors,* which provide the links from the different conceptual dimensions to the general semantic relations of the competence grammar. What is new in our approach is that no specialized interface between the different sorts of knowledge has to be designed because they are modelled using the same formal device, the ALEP formalism. The advantage we gain from this approach is that we have the *full grammar* as a *fall-buck* in cases were the conceptual knowledge cannot contribute to the disambiguation process, due to ambiguities inherited from the domain itself.

Our approach is close to what is termed *'knowledge-based MT'* (KBMT). In this field, the most prominent work is that carried out at the Center for Machine Translation at Carnegie Mellon University (cf. e.g. [Nirenburg et al., 1992]), which has also influenced the DARPA project Pangloss undertaken in collaboration with the Information Sciences Institute at the University of Southern California and the Computing, Research Laboratory of New Mexico State University (cf. [Hovy and Knight, 1993]). In most KBMT projects, the focus is on *common sense world knowledge,* but not on the specific terminology of a domain as in our approach, which, however, can be seen as a specialization of world knowledge. In this sense, we are scaling up a knowledge base (KB) towards its depth, which is in contrast to the breadth scaling up of a KB in the Pangloss project ([Knight, 1994]).

2 Sublanguage Information and NLP

2.1 NL analysis controlled by conceptual information

The major concepts of the domain, i.e. those concepts which are realized by nouns, nominalized verbs and verbs (in terms of processes), are represented in the domain's ontology; they are characterized by *descriptors* which list the properties of the real world thing the concept denotes. For example, the concept TELECOMMUNICATION_EQUIPMENT can be defined as being a subconcept of EQUIPMENT (the generic relation among concepts of the ontology) with multidimensional relations, such as *input, output, location, channel-capacity, frequency-range, linearity* and *digital_rate* which must have values of type SIGNAL, WAVE, EARTH_STATION, CAPACITY_VAL, FREQUENCY_VAL, LINEARITY_VAL and DIGITAL_RATE_VAL respectively.

These relations are specified in so-called *term definition forms* according to ISO and DIN specifications, as well as existing de facto standards in the subject field, developed within the EC-sponsored ET-10 project on *'Terminology and Extra-linguistic Knowledge'([Ripplinger* et al., 1994]). Parts of the information, i.e. a general classification schema, was derived from an existing multilingual termbank of the domain of telecommunications (EIRETERM) which was developed in the context of the MT project EUROTRA.

Since this information is not sufficient for the envisaged NLP task, we have further enhanced these descriptions by a thorough textual and conceptual analysis of a domain corpus, in particular by analyzing the verbs of the domain, which enabled us to define so-called *conceptual templates* ([Schütz, 1994]). Such a template consists of a number of properties that characterize a general concept as either a type, i.e. a thing that can have instances, or a class which governs types that specialize the class. The common classes are ENTITIES, SITUATIONS and PROPERTIES. ENTITIES are those types that can have real world instances and which are realized linguistically as nouns and nominalized verbs, i.e. a subset of the elements of the ontology. SITUATIONS are facilitated by types that express *time* and *place relations,* and that identify *participants, agents* and *result roles.* PROPERTIES are types that denote verbs, i.e. the PROCESS subset of the ontology, involving a thing as subject (ACTIONS and STATES), modifiers (adjectives) that describe details of a thing (e.g. MEASURE_VAL), relationships that identify relational properties to other things, types that denote attributes that (partially) describe a thing, and types that denote constraints which are logical assertions that impose some restrictions on one or more properties of a thing. This classification schema is derived from known classifications in *compositional semantics* (cf. e.g. [Jackendoff, 1990] and [Pustejovsky, 1991]) and knowledge-based NLP (e.g. the Text And Meaning Representation Language - TAMERLAN - of [Nirenburg et al., 1992]).

The term definition forms and the conceptual templates can be automatically transformed into the typed feature structure representation of the ALEP formalism ([Ripplinger et al., 1994]). This TERM structure contains general terminological information, i.e. the classification schema as provided by the EIRETERM termbank and the concept definition, the concept feature which identifies the CONCEPT and thus provides the link to the ontology of the domain, the CONCEPT_ROLES structure which specifies the role slots of the concept, derived from the SITUATION class and parts of the PROPERTY class, and the conceptual modifiers which are listed in the CONCEPT_MODIFY structure, also derived from the PROPERTY class.

For the semantic descriptions of general language we have used the semantic *relations* (SRs) approach developed in the MT project EUROTRA for German. The SRs define the SEM feature structure of an HPSG inspired competence grammar for German, which specifies a *functor-argument-modifier* structure. The domain-specific conceptual information is associated with these relations: the concept type (CONCEPT) is associated with the semantic functor, the conceptual frame elements (CONCEPT_ROLES) with the semantic arguments and the conceptual modifiers (CONCEPT_MODIFY) with the semantic modifiers. The TERM structure is embedded in the SEM structure to permit the testing and evaluation of different sublanguage templates in an appropriate way (modularization).

The semantic and conceptual information structure (SEM with TERM) is embedded together with the syntactic (SYN) and phonological information (PHON) in the overall SIGN feature structure. On the one hand, this organization establishes the global structure of a KB entity, and, on the other hand, the complete lexical information for the implementation. With this organization of information the NL analysis is controlled either by unification or by inferences on the sentence level (for which the competence grammar is designed) in order to check, for example, *selectional restrictions* and *subcategorization frames* based on general semantic and domain-specific information, *type coercion* for the identification of metaphorical senses, or *conceptual classification* information (generic and partitive relations).

The application of these information structures to the analysis process results in a language-independent representation of the intension expressed in a sentence by means of a conceptual organization. We call this the *micro-structure* of the sentence; adopting the term from evaluation strategies of human translations, where similar representations are used (cf. [Gerzymisch-Arbogast, 1994]). This micro-structure can then be used as input for multilingual language processing such as translation (cf. below).

In accordance with the work of [Gerzymisch-Arbogast, 1994] it is possible to extend our approach to the text level by introducing further relations which address the conceptual macro-structures of a given text by so-called *discourse grammar rules.* A *discourse grammar* is simply the sentence grammar augmented by a set of discourse rules. At present, we have applied this only to the resolution of anaphora across sentence boundaries as, for example, in: *The following section describes modulating equipment. It superimposes the audio-frequency signals on the IF-carrier. This equipment extracts them from the IF-carrier.*

Research in this direction addresses in particular the application of modified unification processes, such as higher-order unification (e.g. [Dalrymple et al., 1991]), which, however, is beyond the scope of this paper.

2.2 Conceptual information and translation

In general terminography, such as the EIRETERM database, the focus is on concepts and their linguistic form expressed in terms which are extracted from texts (term identification). In translation the focus is on *production,* i.e. a dynamic process, concerned with the movement from the textual substance in one language to the textual substance in another language. Inside this process there is a procedure in which *units of meaning* of one culture are matched with those of another before finding their textually and situationally appropriate linguistic realization. In view of terminology these units are not of interest because they are temporary and casual collocations of concepts brought into a particular relationship by an author. Translation has to work with concepts and terms in context, whereas terminology isolates terms from their context (decontextualization) and then associates them with concepts, i.e. matching between term and concept vs. matching between textual units through concepts.

Concept correspondence is discovered when comparing the terminologies of different languages, subject fields and cultural systems. Based on this assumption there are thus four possibilities for the process of translation based on the *intension* of a conceptual representation. By intension we mean the set of characteristics, i.e. the formal representation of the properties of an object serving to form and delimit its concept, which constitutes the concept. We distinguish:

1. *Complete co-incidence* of intensions, i.e. the conceptual meaning can be expressed in the languages under consideration in terms of a linguistically realized proposition.

2. *Inclusion* of one intension in the other, i.e. there are conceptual meanings of a concept which do not exist in another language, for example, the concept PALACE has one specific meaning which is only valid in a monarchy. Another example of this kind is the process DIE in its metaphorical meaning in telecommunications and computer science: in English we may have the realization with an active verb *'The signal died.'* but in German this has to be realized by the ergative verb 'abbrechen' (break down), i.e. *'Das Signal brach ab.'.* This in contrast to *'Hans brach das Signal ab.' (*'Hans broke down the signal.')* vs. *'Hans killed the signal.'*

3. *Overlapping* of intensions, i.e. there are in either language conceptual meanings of a concept which do not have a corresponding value. For example, the concept PICTURE, which in English is a superconcept of PAINTING, DRAWING and PHOTOGRAPH, has no direct correspondence in Japanese. Only the subconcepts have such a correspondence.

4. *No co-incidence* of intensions, i.e. either the concept does not exist in another language, or the conceptual meanings are different. For example, the term *zapping* with its meaning of the frequent switching between TV channels didn't exist in German a few years ago. In the field of VR, we can find many of these examples.

The cases 2, 3 and 4 above are called *conceptual* or *intensional mismatches.* Mismatches are mostly caused on social, political and cultural backgrounds, although the conceptual structures are not bound to particular languages.

Case 1 needs no specific translation rule. Cases 2 and 3 need inferencing capabilities over the concept system for the identification of *common* superconcepts, which, however, will cause a degradation of the granularity of the concept's intensional description. In order to keep the granularity of the source and target language as close as possible, as well as to save costly inferences during generation, it might be worth considering the application of explicit translation rules, as this is done for case 4. In the actual implementation (cf. below), we have applied the latter approach, due to the missing inference capabilities in the current ALEP system.

3 Demonstrator Implementation

In the previous sections we have briefly outlined the theoretical framework for the integration of different sorts of knowledge into the analysis and translation process of an NLP system; in this section we describe the actual implementation in the ALEP framework.

3.1 Implementation Overview

The general architecture of our analysis module is based on *staged processing,* which was selected for reasons of efficiency (runtime behaviour). In our approach analysis is therefore composed of two steps: 1. *shallow syntactic analysis* for efficient parsing with a competence grammar for German, and 2. *conceptual refinement* of the parsing result as performance control.

With the second step we achieve a sublanguage-specific filtering of the parsing results. For parsing we have used the grammar and the parts of the lexical entries which specify the syntactic and phonological information, including the terms of the domain, but without any particular domain information. For the refinement process (filtering) we have used those parts of the lexical entries which specify the general semantics and the domain-specific information. In this step the grammar rules function as the navigator through the parsing structures; the actual filtering process is done by unification (cf. below).

For the translation module which has been designed for mapping German analysis output (so-called *linguistic structures)* to English synthesis input, we have adopted an approach which calls translation on a specific type contained in the top-most feature structure of the input linguistic structure, i.e. the conceptual (sub-) feature structure. At the moment, compared to the German analysis module, the transfer module as well as the English synthesis modules have a limited coverage. This is mainly due to the fact that the focus of our work was on the conceptual organization of the domain and the performance control of the analysis process through conceptual knowledge.

3.2 Knowledge Sorts

The formal specifications for the conceptual and sortal (semantics) organization can be directly expressed in terms of the *type system* facility of the ALEP formalism (cf. above).

In the parsing grammar we have specified the information distribution of the semantic feature structure (SEM) which includes as a substructure the conceptual knowledge organization (TERM) about the domain. During parsing these information slots are opened and during refinement they are filled in by the appropriate information by unification. Unification failure then triggers the disambiguation process in the refinement phase and thus the performance control in analysis.

In the refinement part of the lexicon we have stated the selectional restrictions for different semantic and conceptual reading distinctions, as well as the appropriate subcategorization frames and type coercion information. This information is used during the refinement process to identify valid parsing results by unification. The result of the refinement process is a fully specified intensional representation according to the selected semantic and conceptual information.

Consider, for instance, the lexicon entry for *adaptieren (adapt);* in the entry the semantic subject *agent* is linked to the conceptual role *agent* which is of type EQUIPMENT, which is a type of the domain's ontology, and the semantic object *affected* is linked to the conceptual role *result,* which is of type SIGNAL.

Selectional restrictions based on specific domain information for nouns are linked to the noun's sub-categorisation frame and which can be associated with the appropriate prepositions, such as *von (of)* and *mit (with)* which have a specific interpretation in the domain, e.g. '... *die Abstimmung von Hochfrequenzträgern mit Niedrigfrequenzsignalen . . . '* (...the modulation of very high-frequency carriers with low-frequency signals ...).

Similar to these selectional restrictions, domain dependent restrictions, for example, for the subject/object identification, can be formulated, e.g. *'Fernübertragungsausrüstungen umfassen* *auch Modulationsgeräte.'* (Telecommunication equipment also comprises modulating equipment.). In this example, the concept associated with the object must be more specific than the concept assigned to the subject (generic relationship).

According to the domain-specific information, the sentence *'Diese Geräte überlagern die Audiofrequenzsignale auf der IF-Trägerwelle.'* (This equipment superimposes the audio-frequency signals on the IF-carrier.) is well-formed, as opposed to the sentence *'Diese Geräte überlagern die Audiofrequenzsignale auf der Erde.'* (This equipment superimposes the audio-frequency signals on the earth.) which is not well-formed, although grammatically correct.

3.3 Translation Relations

Within the translation module there is one rule for initializing the translation process. Once translation is called on the conceptual (sub-) feature structure specified as the value of the linguistic structure's top-most SEM structure, translation is called recursively on type SEM and all subordinate types respectively.

In cases of complete co-incidence of source and target structures, no specific translation rule is applied; only in cases of mismatches are explicit translation rules applied. In this case, when translation is called on type SEM, the predicate string specified by the *pred*-attribute of the functor feature structure is translated from one language into the other. For the translation of the appropriate conceptual information rules for the different conceptually dependent arities are then used. This approach also allows for a straightforward account of instances of complex transfer where changes have to be performed according to the argument structure of the predicate that has to be translated.

A domain-specific role structure of a concept, identified by the TERM attribute *concept_roles,* is translated by a rule dedicated to the relevant subtypes of type CONCEPT_ROLES. For instance, the role structure assigned to the predicate is translated by a rule operating on the conceptual role subtypes and calling recursively for translation on type CONCEPTUAL_FS which is the type assigned to the roles of a concept. Type CONCEPTUAL_FS will, then, be translated by a rule which, in turn, calls for translation on type TERM again.

The translation of the modifier-list of a concept in TERM, finally, is performed by distinct rules with each of them accounting for a specific number of elements specified in the modifier list (including the empty modifier list).

In each case, the result of the translation is a fully specified conceptual representation of the intension of the analyzed sentence. In cases of mismatches, the representation is augmented by an appropriate semantic description for ease of generation.

3.4 Synthesis

Ideally, the basic SIGN feature structure and, more specifically, the conceptual feature structures should be the same for all languages. With this assumption, it should only be the syntactic feature structure which has to be revised in designing the type and feature specification for an English synthesis grammar.

Since no refinement can be applied in synthesis (in the current ALEP release), the English synthesis grammar operates in one step. Here, the conceptual descriptions (in some cases augmented by general semantics) trigger the access to the (generation) lexicon.

4 Conclusions

In this paper we have briefly described a new approach which certainly deserves the attention of the machine translation community and further exploration in additional subject domains.

One important advantage of the suggested approach is its modularity. The basic linguistic knowledge is represented in a competence grammar. The terminological knowledge for the subject domain is encoded in a hierarchy of typed feature terms. This specialized knowledge for a sublanguage constrains grammatical analysis and transfer. The depth of semantic/conceptual analysis is restricted to the needs of the translation task. The separation of the knowledge sources facilitates extensibility and portability to other sublanguages as well as to the macro-structural handling of texts.

Our approach is innovative because we do not need a specially designed interface between the different knowledge sorts, because each sort is realized by means of the ALEP formalism.

Acknowledgements

Thanks are due to our anonymous reviewers. Their comments were valuable for improving our paper according to the intention and the audience of the AMTA conference.

References

[ALEP, 1993] ALEP Documentation Package, Vol. I and II. CEC and PE International, Luxemburg.

[Dalrymple et al., 1991] M. Dalrymple, S. M. Shieber and F. C. N. Pereira, 1991. Ellipsis and higher-order unification. Linguistics and Philosophy 14(4).

[Gerzymisch-Arbogast, 1994] H. Gerzymisch-Arbogast, 1994. **Übersetzungswissenschaftliches Propädeutikum**. UTB 1782, Francke Verlag Tübingen, Germany.

[Hovy and Knight, 1993] E. Hovy and K. Night, 1993. Motivating Shared Knowledge Resources: An Example from the Pangloss Collaboration. In: Proceedings of the IJCAI workshop on Shared Knowledge, Chambery, France.

[Knight, 1994] K. Knight, 1994. Building a Large-Scale Knowledge Base for Machine Translation. In: Proceedings of AAAI-94.

[Jackendoff, 1990] R. Jackendoff, 1990. **Semantic Structures**. MIT Press, Cambridge, Massachusetts.

[Nirenburg et al., 1992] S. Nirenburg, J. Carbonell, M. Tomita and K. Goodman, 1992. **Machine Translation: A knowledge-based Approach**. Morgan Kaufmann Publishers, San Mateo, California.

[Pustejovsky, 1991] J. Pustejovsky, 1991. The Generative Lexicon. In: Computational Linguistics, 17(4).

[Ripplinger et al., 1994] B. Ripplinger, J. Schütz, G. Talbot, 1994. Terminology and Extra-Linguistic Knowledge, ET-10-66 Final Report. European Commission, Luxembourg.

[Schütz, 1994] J. Schütz, 1994. **Terminological Knowledge in Multilingual Language Processing**. Studies in Machine Translation and Natural Language Processing, Volume 5, Office for Official Publications of the European Communities, Luxembourg.