LOGON - a Norwegian MT effort

Jan Tore Lønning[♣], Stephan Oepen[♣♠], Dorothee Beermann[♡],
Lars Hellan[♡], John Carroll[△], Helge Dyvik[◇],
Dan Flickinger[♠], Janne Bondi Johannessen[♣], Paul Meurer[◇],
Torbjørn Nordgård[♡], Victoria Rosén[◇] Erik Velldal[♣]

[♣]Universitetet i Oslo, Boks 1102 Blindern; 0317 Oslo (Norway)
[◇]Universitetet i Bergen, Sydnesplassen 7, 5007 Bergen (Norway)
[♡]Norges Teknisk-Naturvitenskapelige Universitet, 7491 Trondheim (Norway)
[♠]Center for the Study of Language and Information, Stanford, CA 94305 (USA)
[△]University of Sussex, Falmer, Brighton BN1 9QH (UK)

all@emmtee.net

1 Introduction

The LOGON project, focusing on machine translation from Norwegian to English, involves three universities (Oslo, Bergen and NTNU Trondheim) and has a public funding from the Norwegian Research Council's program for language technology (KUNSTI, 2002–2006) on the order of 20 millions NOK. KUNSTI emerged as a reaction to the global growth in language technology at the turn of the millennium with the following two concerns. Firstly, language technology applications should not only be available for English and other large languages but also for Norwegian. Secondly, Norway as a country should not fall behind in this new and growing industry. As a response to the first concern, KUNSTI asked for projects with a focus on Norwegian language, over time giving rise to reusable language technology resources. As a reaction to the second concern, it was important to raise the national competence, in particular educate more PhD's within the field. It was also a goal to initiate large projects resulting in working demonstrators that involve several sub-areas of the field and a diversification of methods. Machine translation turned out to be a suitable task.

2 The approach

We have chosen a traditional semantic transfer-based approach as our starting point. In spite of strong winds in the direction of statistical methods during the last decennium, in language technology in general, and machine translation in particular, we are still firm believers in symbolic and 'deep'

linguistic methods. Although statistical approaches can deliver good initial results to MT, they seem to sooner or later suffer from 'ceiling' effects in performance and ask to be augmented with more linguistic structure.

Our approach is to start with a firm and theoretically sound symbolic backbone, while augmenting this with probabilistic methods to direct the choices where the symbolic methods fan out. The transfer-based approach falls into three steps. (i) An in-depth grammatical and semantic analysis of Norwegian resulting in language-specific logical semantic representations. (ii) A transfer of these representations into language-specific English representations. (iii) And, finally, generation from the semantic representations to English sentences.

A central locus in the project is the format of the semantic representations, where our starting point is Minimal Recursion Semantics (MRS, Copestake et. al. 2003). MRS provides logical representations which both have model interpretations and allow underspecification of scope. The formalism is specially developed with translation and transfer in mind. Moreover, there is an available public domain implementation of HPSG with MRSes in LKB which allow parsing and generation from MRSes. In particular, we are using the LiNGO English Resource Grammar (ERG) as the target language grammar.

When it comes to the Norwegian analysis, we have not used the HPSG formalism, however, but an LFG grammar for Norwegian, NorGram, which has already been under development in Bergen for several years (Dyvik 1999). This has activated new research topics, viz. on the integration of the MRS framework, developed within HPSG originally, with LFG; and more generally, on the relationship between grammatical formalisms and semantic representations.

The project is scheduled for the period 2003–2006. Among our central goals is a functional demonstrator, and we decided to have a first end-to-end baby demonstrator by March this year to test the viability of the approach and to direct the course for the rest of the project period. We decided to use tourist texts as our domain and started with 100 sentences extracted from Norwegian tourism brochures. As it took some time to get the project started, we have worked effectively for 9 months on the baby demonstrator.

3 Representations: MRS

MRS is not a theory of semantics but rather a system of semantic representations that features a flat semantics, expressed as a bag of elementary predications (EPs) which compositionally express the meaning of the structure to be described. Three of the basic components of an MRS structure should be mentioned:

(L)TOP the handle of the highest-scoping EP in the scope partial order.

RELS(list) a bag of elementary predications each with a handle (a special type of a variable used, for example, in the expression of scopal dependencies) and an array of semantic argument slots.

HCONS(list) a set of constraints on possible scopes between the EPs.

EPs in the MRS universe are typed, featuring a small set of underspecified semantic arguments (ARG1, ARG2, ARG3, ...) with variables as their values. Also variables are typed, allowing a distinction between handles on the one side and referential and event variables on the other. MRS structures obey a set of wellformedness constraints regulating the binding of arguments (handle connectivity, binding of variables).

A simplified MRS for the Norwegian sentence *Hunden bjeffer* (The dog barks) can be expressed as:

$\langle\, h_1,$
$\{\, h_1{:}\mathsf{prpstn\_m}(h_7),\ h_8{:}\mathsf{bjeffe\_v}(e_9,\ x_5),\ h_3{:}\mathsf{def}(x_5{,}h_2,\ h_4),\ h_6{:}\mathsf{\_hund\_n}(x_5)\,\},$
$\{\, h_2 =_q h_6,\ h_7 =_q h_8\,\}\,\rangle$

In this example 'bjeffe_v(e_9, x_5)' is one of the EPs, while h_1 is the TOP.

4 Analysis

One challenge has been the augmentation of the LFG resource grammar NorGram with MRS-representations. NorGram assigns the usual LFG representations c-structure(PS tree) and f-structure (attribute-value matrix) to sentences. The f-structure is derived by co-description: partial descriptions of f-structures are associated with c-structure rules and lexical entries. The LFG architecture allows the projection of new representations by similar co-description, and the MRS-structure is projected off the f-structure in this way.

NorGram has been developed under the Xerox Linguistic Environment (XLE) and we use this software package for the analysis. We have managed to integrate this in the overall pipeline with some post-processing of the output before it is fed into the transfer module.

5 Semantic transfer

The transfer component, as defined within the LOGON project, is a resource sensitive rewriting process over MRS structures. It follows many of the main ideas from VerbMobil—transfer as a resource-sensitive rewrite process, where rules replace MRS fragments (SL to TL) in a step-wise manner (Wahlster 2000)—but adds two innovative elements to the transfer component, viz (i) the use of typing for hierarchical organization of transfer rules and (ii) a chart-like treatment of transfer-level ambiguity.

The general layout of an MRS transfer rule (MTR) in LOGON is illustrated by the most general MTR type definition

```
mrs_tranfer_rule := top &
[ FILTER mrs,
  CONTEXT mrs,
  INPUT mrs,
  OUTPUT mrs ]
```

The INPUT feature and the optional features CONTEXT and FILTER are unified against an input MRS M and, when successful, trigger the rule application; elements of M matched by INPUT are replaced with the OUTPUT component, respecting all variable bindings established during unification. The optional CONTEXT and FILTER components serve to conditionalize rule application (on the presence or absence of specific aspects of M), establish bindings for OUTPUT processing, but do *not* (contrary to INPUT) consume elements of M.

Transfer rules deploy a multiple-inheritance hierarchy with strong typing and appropriate feature constraints both for elements of MRSs and MTRs themselves. In close analogy to constraint-based grammar, typing facilitates generalizations over transfer regularities. A transfer rule for intransitive verbs, arg1_v_mtr, inherits from the most general rule above, while the specific instance for the pair 'bjeffe' → 'bark', in turn, inherits from arg1_v_mtr.

```
arg1_v_mtr := mrs_transfer_rule &
[ INPUT.RELS < [ LBL #h, ARG0 #e, ARG1 #x ] >,
  OUTPUT.RELS < [ LBL #h, ARG0 #e, ARG1 #x ] > ].

bjeffe := arg1_v_mtr &
[ INPUT.RELS < [ PRED "_bjeffe_v_rel" ] >,
  OUTPUT.RELS < [ PRED "_bark_v_rel" ] > ].
```

On top of this, rules are marked as either optional or obligatory (the default). This makes the rule ordering critical for the outcome of the transfer step, but is required to avoid a combinatorial explosion of spurious transfer ambiguity.

6 Implementation and software engineering

Even though the translation machine at the current stage translates only a limited number of sentences, we have implemented a solid software architecture which we believe will stand up to extensions. The three core components (analysis, transfer, generation) are implemented as separate processes managed by a central controller which passes intermediate results through

the translation pipeline. Use of the Parallel Virtual Machine (PVM) protocol facilitates flexibility and robustness.

A central tool in the project so far, where we have developed two grammars and transfer rules at the same time, has been the [incr tsdb()] profiling package (Oepen & Carroll, 2002). The profiling methodology and tool make it possible to assess progress and keep track of a multitude of central system measures like coverage, ambiguity and speed over successive system revisions. In particular with a project involving several sites and researchers it is difficult to make progress evaluation precise, and impressionistic methods run short. It is our general experience so far that a project this size must take software engineering aspects seriously.

7 Further perspectives

All three stages of the basic translation system fan out. Ambiguous sentences have more than one analysis; transfer produces several English MRSs for each Norwegian MRS, and each English MRS has several realizations, in average in fact 30. We will use probabilistic methods to rank output at all stages. Preliminary experiments with ranking the output from the generator against the BNC using an n-gram method show promising initial results.

So far, we have not taken the problem of lexical selection seriously and it is obvious that as the coverage grows so will the transfer ambiguities and the number of bad translations. We will face this problem in the next phase of the project and work on it from two sides. We will on one hand try and fine-tune the transfer rules, and on the other hand expand the work on ranking the outputs.

8 Outlook

In addition to the work on the core MT prototype, LOGON pursues more basic, PhD-level research (on disambiguation techniques, soft constraints, WSD, and the syntax – semantics interface) as well as resource creation (adaptation of a large computational lexicon and associated tools and the production of a parallel domain corpus) and evaluation activities.

We expect to broaden the scope of our prototype continually, specifically in terms of transfer coverage, and in parallel plan to pursue a few in-depth feasibility studies, for example on the use of more 'geometric' semantic accounts of temporal relations or modal operators.

For component-level evaluation we have revised the competence and performance profiling methodology [incr tsdb()], but also foresee a round of end-to-end, black-box evaluation to assess the utility of currently fashionable, n-gram based similarity metrics.

References

Ann Copestake, Daniel Flickinger, Ivan A. Sag and Carl Pollard: 2003, 'Minimal Recursion Semantics. An Introduction', CSLI, Stanford, CA, submitted for publication.

Helge Dyvik; 1999, 'The Universality of F-Structure. Discovery or Stipulation? The Case of Modals', in *Proceedings of the 4th International Lexical Functional Grammar conference*, Manchester, UK.

Stephan Oepen and John Carroll: 2002, 'Efficient Parsing for Unification-based Grammars', in Stephan Oepen, Daniel Flickinger, J. Tsujii and Hans Uszkoreit (eds.), *Collaborative Language Engineering. A Case Study in Efficient Grammar-based Processing*, CSLI Publications, Stanford, CA, 195–225.

Wolfgang Wahlster (ed.): 2000, *VerbMobil. Foundations of speech-to-speech translation*, Berlin, Germany, Springer.