Using a Bi-Lingual Dictionary in Lexical Transfer

— An Experience Report and Some Preliminary Findings —

Lars Nygaard[♣], Jan Tore Lønning[♣], Torbjørn Nordgård[♢], Stephan Oepen[♣♠]

[♣]Universitetet i Oslo, Boks 1102 Blindern; 0317 Oslo (Norway)

[♢]Norwegian University of Science and Technology, Dragvoll, 7491 Trondheim (Norway)

[♠]Center for the Study of Language and Information, Stanford, CA 94305 (USA)

{larsnyg | jtl | torbjorn | oe}@emmtee.net

Abstract

This paper reports on experiences with fully automatic lexical transfer rule acquisition in the domain of rule-based machine translation with deep syntactic and semantic processing. We demonstrate that if comprehensive grammars with broad lexical coverage exist for the source and target languages, then open POS class lexical transfer rules can be derived automatically if access to some bilingual dictionary is provided. We show the results in terms of extended coverage and BLEU scores.

1 Project Background

LOGON is a machine translation project (Oepen et al., 2004) which makes use of Lexical-Functional Grammar in parsing (on the XLE platform) and Head-Driven Phrase Structure Grammar in generation (via the LKB platform) to translate texts about hiking in the backcountry from Norwegian to English. A training data set containing approximately 5000 sentences has been collected and translated by professional translators (three translations for each sentence). The project has access to a comprehensive machine-readable bilingual dictionary (Kunnskapsforlaget, 2001) with more than 200.000 translation pairs (approximately 80.000 from Norwegian to English).

The Norwegian analysis component is based on an existing LFG resource grammar, NorGram, under development on the Xerox Linguistic Environment (XLE), see Dyvik (1999). The parser has a comprehensive lexicon, which consists of 80,000 lexical items derived from a modified version of the Norwegian NorKompLeks lexicon (Nordgård, 2000). For use in LOGON, the grammar has been modified and extended, and it has been augmented with a module of Minimal Recursion Semantics representations (MRSs) which are computed from the f-structures by co-description (Oepen et al., 2004).

Transfer is facilitated via Minimal Recursion Semantics representations (MRSs) (see below and Copestake, Flickinger, Sag, & Pollard, 2003). Such source language MRSs are input to a transfer module which creates target language MRSs. An example of a Norwegian MRS prior to transfer is given in Figure 1.

The generation component produces sentences in the target language from these latter MRS structures by using the English Resource Grammar (ERG) (Flickinger, 2000). ERG is in a sense a "filter" which guarantees that the target language strings are well-formed.

The project aims at producing high-quality translations in the sense that use of the ERG guarantees grammatical output and the highly structured semantic transfer makes sure that the structural source language meaning is preserved. If an error occurs in the system pipeline, for instance that an unknown word is encountered, no output is produced. Hence we do not expect that the system will be able to produce translations for every input sequence. The system architecture is visualized in Figure 2.

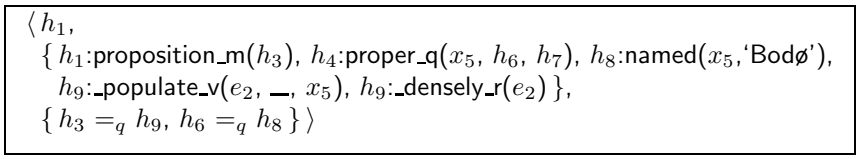

Figure 1: Simplified MRS representation for the utterance 'Bodø is densely populated.' The core of the structure is a bag of *elementary predications* (EPs), using distinguished handles ('h_i' variables) to express scopal relations, where handle identity denotes scopal conjunction and an additional set of '$=_q$' (equal modulo quantifier insertion) handle constraints enables scope underspecification. Event- and instance-type variables ('e_j' and 'x_k', respectively) capture semantic linking among EPs, where MRSs tend to use a small inventory of thematically bleached role labels (ARG_0 ... ARG_n), abbreviated through order-coding in the example above.

2 MRS-Based Transfer

Unlike in parsing and generation frameworks, there is less established common wisdom in terms of (semantic) transfer formalisms and algorithms. LOGON follows many of the main Verb*mobil* ideas—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 form of MRS transfer rules (MTRs) is as a quadruple

[CONTEXT :] INPUT [! FILTER] → OUTPUT

where each of the four components is a partial MRS. Left-hand side components 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* consume elements of M. Although our current focus is on translation into English, MTRs in principle state translational correspondence relations and, modulo context conditioning, can be reversed.

Transfer rules deploy a multiple-inheritance hierarchy with strong typing and appropriate feature constraints (the LKB formalism; Copestake, 2002) both for elements of MRSs and MTRs themselves. In close analogy to constraint-based grammar, typing facilitates generalizations over transfer regularities—hierarchies of predicates or common MTR configurations, for example—and aids development and debugging. The following shows the type for simple noun transfer:

```
noun_mtr := monotonic_mtr &
[ INPUT.RELS <
  [ LBL #h1,
    ARG0 #x1 ] >,
  OUTPUT.RELS <
  [ LBL #h1,
    ARG0 #x1 ] > ].
```

First this shows that the type inherits from the more general type `monotonic_mtr`. Expressions like `#h1` and `#x1` are variables and show that these are shared between the representation before and after and application of a rule based on this type. A simple instance of this rule is illustrated by

```
ras_avalanche_n := noun_omtr &
[ INPUT.RELS <
  [ PRED _ras_n_rel ] >,
  OUTPUT.RELS <
  [ PRED _avalanche_n_1_rel ] > ].

ras_landslide_n := noun_mtr &
[ INPUT.RELS <
  [ PRED _ras_n_rel ] >,
  OUTPUT.RELS <
  [ PRED _landslide_n_1_rel ] > ].
```

There are two alternative translations which is catered for by making the first rule optional, as is indicated by the `_omtr` suffix in the type name.

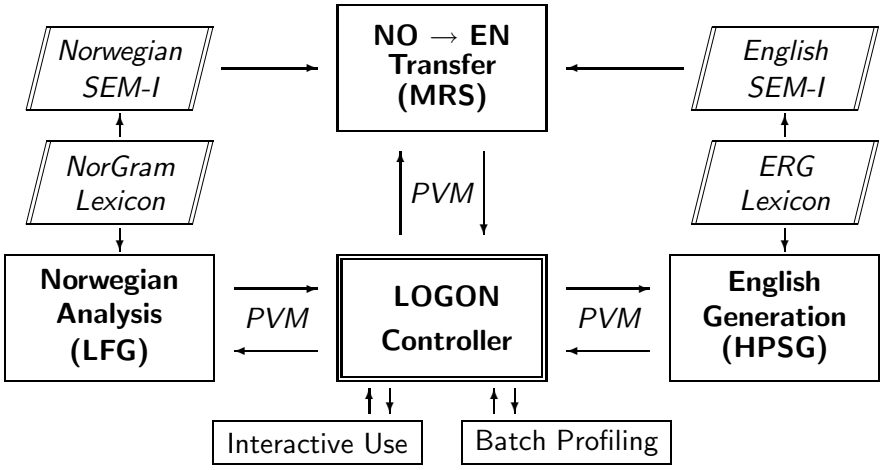

Figure 2: Schematic LOGON system architecture: the three core processing components are managed by a central controller that passes intermediate results (MRSs) through the translation pipeline. Both the analysis and generation grammars 'publish' their interface to transfer—i.e. the inventory and synopsis of semantic predicates—in the form of a Semantic Interface specification ('SEM-I'), such that transfer can operate without knowledge about grammar internals.

```
n+n_nominalization_mtr := monotonic_mtr &
[ INPUT [ RELS < [ LBL #h1, ARG0 #x2 ],
                [ LBL #h3, ARG0 #x4 ],
                [ PRED unspec_rel, LBL #h3, ARG1 #x4, ARG2 #x2 ],
                [ PRED udef_q_rel, ARG0 #x2, RSTR #h5 ]>,
          HCONS < qeq & [ HARG #h5, LARG #h1 ] > ],
  OUTPUT [ RELS < [ LBL #h6, ARG0 #e6 & e_no_tense & [ PROG + ] ],
                  [ PRED nominalization_rel, LBL #h3, ARG0 #x4, ARG1 #h4  ],
                  [ PRED prpstn_m_rel, LBL #h4, ARG0 #e6, MARG #h5 ] >,
           HCONS < qeq & [ HARG #h5, LARG #h6 ] > ] ].

aksjon+radius_n_rel_2 := n+n_nominalization+n_mtr &
[ INPUT.RELS < [ PRED _aksjon_n_rel ],
               [ PRED _radius_n_rel ], ... >,
  OUTPUT.RELS < [ PRED _cruise_v_1_rel ],
                [ PRED _range_n_of_rel ], ... >,
```

Figure 3: Correspondence type and sample translation rule for the compound *aksjonsradius*.

For a slightly more interesting example, consider the following correspondence type and sample instance in Figure 3.

This example demonstrates how two input EPs, viz. the decomposed semantics corresponding to the compound *aksjonsradius* is translated as a compound whose left member is actually a de-verbal nominalizaton, e.g. *crusing range*.

3 Automatic Transfer Rule Acquisition

While we believe that hand-crafted lexical transfer rules are a necessary component in precision-oriented MT, writing such rules is of course time-consuming. In the LOGON set-up, we experienced transfer rule creation as a bottleneck to system building, as we had started out with two grammars that already had substantial lexica. But it is not necessary to hand-write all the transfer rules. Using our external bi-lingual dictionary and building on hand-built transfer rules as a seed set of templates, large numbers of lexical transfer rules can be constructed by analogy. For elementary, one-to-one mappings between a source language and target language word, the rule acquisition task is largely mechanic—but still presenting interesting design choices in relating semantic predicates (as they comprise input and output MRSs to transfer) to citation forms (as they are found in the dictionary). A large number of dictionary translations, however, are complete syntactic units, and here the automatic creation of semantic transfer rules becomes a more interesting task.

In a nutshell, the transfer-level mapping that needs to be established is three-fold: (a) relating a source language MRS predicate to surface forms and thus dictionary entries; (b) looking up candidate translations from the dictionary and analyzing their internal structure (if any); and (c) finding target language predicates that correspond to each dictionary translation(s) and its components. Besides the dictionary (accessed as a relational database), the source and target language Semantic Interfaces (called SEM-Is; see Figure 2 and Flickinger, Lønning, Dyvik, Oepen, & Bond, 2005 for background) are central knowledge sources in this process. For both the analysis and generation grammars, their SEM-Is spell out the inventory of valid semantic predicates, jointly with information about the 'terms of use' for each predicate, e.g. its range of semantic roles, value constraints, indication of optionality, et al. Based on certain naming conventions in semantic predicates and the typing of MRS variables, we can further read out a limited number of syntactic properties for the class of words introducing each predicate from its SEM-I entry.

To faciliate the mapping from semantic predicates to dictionary entries, each SEM-I is augmented with links to candidate surface forms and an indication of the underlying morphological process, e.g.

```
_cruise_v_1_rel :
  base 'cruise',
  gerund: 'cruising',
  passive: 'cruised', ...
```

Given all these pieces, assume the bilingual dictionary prosed the following lexical mappings for the Norwegian *avl*

```
avl :
 breeding | animal husbandry |
 agricultural production |
 crop | ...
```

Auto-generating transfer rules on this basis demonstrates both the basic mechanics as well as the use of more involved transfer correspondence types. Looking up the base form *avl* in the Norwegian SEM-I yields the semantic predicate _avl_n_rel; likewise, English *crop* maps to _crop_n_1_rel. The full SEM-I entries for these predicates are fairly simple:

```
_avm_n_rel : ARG0 x.

_crop_n_1_rel : ARG0 x.
```

Interpreting the variable type of the inherent argument (where x denotes referential indices in the MRS universe) and the _n_ sense tag in the predicate name, we can clearly treat this mapping in analogy to the elementary *rasfare – avalanche* transfer rule

example from Section 2 above. Thus, the appropriate correspondence type should be `noun_mtr`, and to create the full rule we only need to supply the pair of source and target language predicates.

Moving on to other known lexical mappings for *avl*, lookup of English *breeding* yields no predicates for which it is the base form, but *breeding* maps to _breed_v_1_rel in its gerund form. Pairing a source language (non-relational) noun with a de-verbal nominalization in the target language requires the use of a different transfer correspondence type, viz. `n_nominalization_mtr`; our semantic transfer approach makes it possible to, again, simply 'plug' the corresponding pair of semantic predicates into an existing type definition. The correspondence type, in turn, arranges for the nominalization to take the correct semantic form and identity, in terms of its logical variables.

Other candidate translations of *avl*, viz. *animal husbandry* and *agricultural production* show multi-word lexical outputs, where in the first case the dictionary translation is an N−N compound (a syntactic process in English), and in the second it is an $\overline{\text{N}}$ with an attributive adjectival modifier. Our transfer rule acquisition machinery uses PoS lookup and simple pattern-based analysis to determine the specific syntactic structure and then, again, can choose appropriate correspondence types (e.g. `n_n+n_mtr` and `n_adj+n_mtr`, respectively). Obviously, when outputting a compound semantics, either of the elements could itself be a nominalization (thus giving rise to different correspondence types, once again).

To date we have applied the technique sketched here to all nominal and adjectival elements from the (limited) LOGON vocabulary. For a total of some 2500 input words, we arrived at a little over 6000 transfer rules. Close to fiften per cent of the auto-generated rules are non-trivial in nature, i.e. output more than a single target language predicate. While we expect to provide more detailed and up-to-date evaluation results in a full version of the paper, we were able to confirm the effectiveness of our method in an early experiment. When we first started work on our 5000-sentence development corpus, initial end-to-end coverage of the LOGON demonstrator was very low: only eight per cent of the inputs could be translated by the system (recall that the system will only produce output when parsing, transfer and, generation succeed). Counting sentences with zero outputs as a zero contribution, this initial system configuration achieved a BLEU score of 0.04. Adding auto-generated transfer rules into the LOGON pipeline, end-to-end coverage increased to 22.4 per cent, and the BLEU average went up to 0.13. In a sense these figures indicate the influence of basic lexical coverage in transfer (and in subsequent work, primarily harmonizing meaning representations across grammars and adding syntactic and structural transfer coverage, we are aiming to approximately double end-to-end coverage). When averaged only over sentences that were actually translated, our BLEU score was 0.56 for both versions, i.e. there was no measurable loss of 'per-sentence' output quality brought about by the inclusion of auto-generated lexical transfer rules.

4 Compound Translation

Compounding is a highly productive morphological process in Norwegian, and a (surprisingly) large number of compound words are listed in both our bi-lingual dictionary and the mono-lingual lexicon of the analysis grammar; however, the two resources differ substantially in the choice of compounds that are considered lexicalized, thus suggesting a certain degree of arbitrariness in the vocabulary choice of existing lexica. In related work, we have identified candidate sets of compounds that require idiosyncratic translation; compounding often constrains the range of available word senses for compound elements, and what is a compound in one language may be lexicalized or a different syntactic structure in another language.

Section 2 already gave an example of compound translation based on the bilingual dictionary, viz. Norwegian *aksjonsradius* ('action' plus 'radius') as English *cruising range*. Obviously, our technique of combining hand-written correspondence types with pairings of source and target lan-

guage semantic predicates extends straightforwardly to such examples: here, the correspondence holds between an (ordered) set of source predicates and a corresponding set of target predicates. Each individual transfer type determines the specific relationship across the two sets and also among elements within each.

5 Outlook and Summary

We have so far applied the method to nouns and ajectives. As common nouns is by far the largest word class, this is already useful. The method can be applied to other word classes, as well. Hence, we would be able to produce transfer rules also for verbs which have argument structures of the same types in the two languages.

Some care should be taken, though. Even though the input predicate and the output predicate may be of the same semantic type, there might be some sort of argument switching involved as in the pair *smake* and *enjoy* which switch subject and object. Our current procedure will not detect this since it is not easily read off the dictionary. This illustrates that even when we use automatic rule acquisition, the results should be checked manually to give optional results.

In conclusion, we have demonstrated that a "traditional" rule-based MT system can overcome at least parts of the lexical acquisition bottleneck by utilizing wide coverage grammars and lexicons for the source and target languages and a comprehensive bilingual dictionary.

References

Copestake, A. (2002). *Implementing typed feature structure grammars.* Stanford, CA: CSLI Publications.

Copestake, A., Flickinger, D., Sag, I. A., & Pollard, C. (2003). *Minimal Recursion Semantics. An introduction.* In preparation, CSLI Stanford, Stanford, CA.

Dyvik, H. (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.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG)*, 15 – 28.

Flickinger, D., Lønning, J. T., Dyvik, H., Oepen, S., & Bond, F. (2005). SEM-I rational MT. Enriching deep grammars with a semantic interface for scalable machine translation. In *Proceedings of the 10th Machine Translation Summit* (pp. 165 – 172). Phuket, Thailand.

Kunnskapsforlaget (Ed.). (2001). *Engelsk stor ordbok.* Oslo: Kunnskapsforlaget.

Nordgård, T. (2000). A Norwegian Computational Lexicon. In *Proceedings of COMLEX 2000.* Patras, Greece.

Nygaaard, L., & Oepen, S. (2006). Identifying transparent compounds in a large computational lexicon. In *Proceedings of the 5th International Conference on Language Resources and Evaluation.* Genoa, Italy.

Oepen, S., Dyvik, H., Lønning, J. T., Velldal, E., Beermann, D., Carroll, J., Flickinger, D., Hellan, L., Johannessen, J. B., Meurer, P., Nordgård, T., & Rosén, V. (2004). Som å kapp-ete med trollet? Towards MRS-based Norwegian – English Machine Translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation* (pp. 11 – 20). Baltimore, MD.

Wahlster, W. (Ed.). (2000). *Verbmobil. Foundations of speech-to-speech translation.* Berlin, Germany: Springer.