LFG-DOT: a Probabilistic, Constraint-Based Model for Machine Translation

Andy Way

School of Computer Applications,
Dublin City University,
Dnblin 9, Ireland.

Abstract

We develop novel models for Machine Translation (MT) based on Data-Oriented Parsing (DOP: Bod, 1995; 1998) allied to the syntactic representations of Lexical Functional Grammar (LFG: Kaplan & Bresnan, 1982).

Introduction

It is accepted that the main paradigmatic approaches to MT—transfer, interlingua, and statistical—do not at present produce the quality of translation reqnired. There have, however, been a number of attempts at combining elements of these different approaches in an attempt to increase overall translation performance (cf. Carbonnel *et al.*, 1992; Grishman & Kosaka, 1992). Onr efforts to bring abont a better solution to the problems of MT can be viewed in this new hybrid spirit.

DOP has produced interesting results for a range of NLP problems. DOP language models consider past experiences of language to be significant in both perception and prodnction. DOP prefers performance models over competence grammars: models based on large collections of previously occnrring fragments of language are preferred to abstract grammar rules. New language fragments are handled with respect to existing fragments from the corpus, which are combined using statistical techniques to determine the most probable analysis for the new fragment.

DOP Translation Models

DOP has been used already as a basis for MT–Data-Oriented Translation (DOT: Poutsma, 1998). DOP models typically use surface PS-trees as the chosen representation for strings. The DOT translation model relates tree-fragments between two (or more) languages with an accompanying probability, linking source-target translations at all possible nodes in accordance with the principle of Compositionality of Meaning. Once the most likely parse of the source language sentence has been produced, the tree structure of the target is assembled, from which the string is (trivially) derived. Nevertheless, there are usually many different derivations for the source sentence, so many different translations may be available. As is the case when DOP is used monolingually, Poutsma shows that the most probable translation can be computed using Monte-Carlo disambiguation.

DOT is an interesting model, but it is not guaranteed to produce the correct translation when this is non-compositional and considerably less probable than the default, compositional alternative. An example is *commit suicide* ⟷ *se suicider*, where *John commits suicide* is wrongly translated by DOT as **John commet le suicide*. DOT's adherence to left-most substitution in the target given *a priori* left-most substitution in the source is too strictly linked to the linear order of words. As soon as this deviates to any significant degree between languages, DOT has a significant bias in favour of the incorrect translation (assuming the corpus to be representative). Another example is the *like* ⟷ *plaire* case, where the arguments need to be 'switched' between English and French. Even if the correct, non-compositional translation is achievable, DOT derives other wrong alternatives with higher probabilities. In such cases, the correct translation will be dismissed, unless all possible translations are inspected manually.

This is not at all surprising: being based on STSG, DOT is necessarily limited to those contextual dependencies actually occurring in the corpus, a reflection of surface phenomena only. It is well known that models based solely on CFGs are insufficiently powerful to deal with all natural language problems. In this regard, DOP models have been augmented (van den Berg *et al.*, 1994; Tugwell 1995) to deal with richer representations, but such models have remained context-free.

LFG, however, is known to be beyond context-free. It can capture and provide representations of linguistic phenomena other than those occurring at surface structure. Given this, the functional structures of LFG have been harnessed to the techniques of DOP to create a new model, LFG-DOP (Bod & Kaplan, 1998). LFG-DOP permits (via the *Discard* operator) the relaxation of certain constraints on LFG representations, thereby creating generalised fragments against which new input can be compared, and the best analysis constructed.

LFG-DOP Translation Models

We propose that LFG-DOP has the potential to be used as the basis for an innovative MT model, LFG-DOT. We have designed two LFG-DOT models:

1. a simple, linear model which builds a target f-structure from a source c-structure and f-structure, the mapping between them ϕ, and the τ-equations. This model leaves the task of generating the target string from the target f-structure to the standard LFG generation algorithms (e.g. Wedekind, 1988);

2. a more complex model, containing explicit links between both surface constituents and f-structure units in both languages, unlike the previous model which relates the languages just at the level of f-structure (via τ).

Probability models have been constructed for both translation models, and small experiments have been performed for particular cases of 'hard' translation problems. Being able to link exactly those source-target elements which are translations of each other using LFG's τ-equations, LFG-DOT overcomes some of the problems specific to the DOT system. For example, the LFG-MT solution to the *like* ⟷ *plaire* case is (1):

(1) *like*:
$(\tau\uparrow \text{ PRED FN}) = \text{plaire}$
$\tau(\uparrow \text{ SUBJ}) = (\tau\uparrow \text{ OBL})$
$\tau(\uparrow \text{ OBJ}) = (\tau\uparrow \text{ SUBJ})$

That is, the subject of *like* is translated as the oblique argument of *plaire*, while the object of *like* is translated as the subject of *plaire*. The solution to the *commit suicide* ⟷ *se suicider* problem is (2):

(2) *commit*:
$(\tau\uparrow \text{ PRED FN}) = \text{se suicider}$
$\tau(\uparrow \text{ SUBJ}) = (\tau\uparrow \text{ SUBJ})$
$(\uparrow \text{ OBJ PRED}) =_c \text{suicide}$

Where the PRED value of the OBJ of *commit* is constrained ($=_c$) to *suicide*, then the collocational units '*commit + suicide*' are translated as a whole to *se suicider*. DOP's statistical model gives a 'level of correctness' figure to alternative translations. This is useful in cases like these where the default translation in LFG-MT (and in many other systems) cannot be suppressed when the specific translation is required. We have conducted small experiments which show that for a treebank constructed from 10 sentences, despite 7 instances of *commit* ⟷ *commettre* compared to just one *commits suicide* ⟷ *se suicide* example, the correct translation *Marie commits suicide* ⟷ *Marie se suicide* is preferred by both LFG-DOT models over the wrong, compositional alternative by a factor of between 3 and 6 times, depending on which LFG-DOP definition of competition set is selected.

Furthermore, LFG-DOT promises to improve upon the correspondence-based LFG-MT model (Kaplan *et al.*, 1989), particularly where robustness is concerned, as LFG-DOP's *Discard* function enables both nnseen and ill-formed input to he dealt with. For example, Bod & Kaplan (1998) show that given a treebank for the sentences *People walked* and *John fell*, probability models can be constructed where for the 'unseen' sentences *John walked* and *People fell*, the nnmarked interpretation is less likely that the two specific interpretations, and of these the intnitively correct ones are selected for each corresponding verb.

Problems and Future Work

The major problem with any models based on LFG-DOP is the explosion of fragments caused by *Discard*. Allowing *Discard* to operate in the unconstrained manner of Bod & Kaplan's (1998) model results in an exponential number of fragments in which the non-*Discard* fragments are overwhelmed, resulting in the probabilities of derivations via *Root* and *Frontier* being vastly outnumbered by the 'ungrammatical' alternatives. While there is a large increase in the number of fragments produced via *Discard* in LFG-DOT models, compared to the monolingnal LFG-DOP corpora from which they are derived, the explosion of fragments is nowhere near as severe. Notwithstanding this, we propose to restrict the scope of the *Discard* operator by creating two different bags of fragments: the well-formed ones (derived via *Root* and *Frontier*) and the *Discard* ones. Using Good-Tnring (cf. Bod, 2000), we can allocate a fixed, *small* probability mass to the fragments generated by *Discard* to ensure that the derivations using the 'good' non-*Discard* fragments will still be favoured.

Using different LFG-DOP probability models (in terms of which LFG grammaticality checks are enforced, and at which points in the translation process) results in different probabilities with respect to the corpus, bnt does not result in different rankings of alternative candidate translations. A potential problem, however, is that LFG-DOT models, like DOT models, show a tendency to exclude many potentially useful fragments owing to the strictness of Poutsma's (1998) definition of linked fragments. This may resnlt in translations which are theoretically describable not being achievable in practice. Only experimentation on a mnch wider scale will confirm this.

Given the small corpora from which onr findings were derived, any resnlts must be treated with some equivocatiou. Given the (relative) scarcity of some of the linguistic examples cited previously, and the subject of the tests thereon, we regret that it is nigh on impossible to derive 'representative' corpora for the examples in hand. The absence of large-scale LFG-DOP corpora cnrrently prohibits these models from being tested more widely. Nevertheless, recent work on automatic construction of the LFG-DOP corpora (Van Genabith *et al.*. 1999: Sadler *et al.*. 2000) needed for further experimentation using these techniqnes seems promising in this regard.

References

VAN DEN BERG M., BOD R. & SCHA R. (1994). A Corpus-Based Approach to Semantic Interpretation. In *9th Amsterdam Colloquium.*

BOD R. (1995). *Enriching Linguistics with Statistics: Performance Models of Natural Language.* PhD thesis, University of Amsterdam.

BOD R. (1998). *Beyond Grammar: An Experience-Based Theory of Language.* Stanford, California: CSLI Publications.

BOD, R. (2000). An Empirical Evaluation of LFG-DOP. In *Proceedings of the 19th International Conference on Computational Linguistics* (to appear).

BOD R. & KAPLAN R. (1998). A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis. In *Proceedings of the 17th International Conference on Computational Linguistics & 36th Conference of the Association for Computational Linguistics*, p. 145–151.

CARBONELL J., MITAMURA T. & NYBERG 3RD E. (1992). The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics,...). In *4th International Conference on Theoretical and Methodological Issues in Machine Translation*, p. 225–235.

GRISHMAN R. & KOSAKA M. (1992). Combining Rationalist and Empiricist Approaches to MT. In *4th International Conference on Theoretical and Methodological Issues in Machine Translation*, p. 263–274.

KAPLAN R. & BRESNAN J. (1982). *Lexical Functional Grammar: A Formal System for Grammatical for Grammatical Representation*, In *The Mental Representation of Grammatical Relations*, chapter 4. MIT Press.

POUTSMA A. (1998). Data-Oriented Translation. In *Ninth Conference of Computational Linguistics In the Netherlands.*

R.KAPLAN, NETTER K., WEDEKIND J. & ZAENEN A. (1989). Translation by Structural Correspondences. In *Fourth Conference of the European Chapter of the Association for Computational Linguistics*, p. 272–281.

SADLER L., VAN GENABITH J. & WAY A. (2000). Automatic f-structure annotation of CFGs extracted from treebank resources. In *Proceedings of LFG-2000* (to appear).

TUGWELL D. (1995). A State-Transition grammar for Data-Oriented Parsing. In *Seventh European Conference on Computational Linguistics*, p. 272–277.

VAN GENABITH J., WAY A. & SADLER L. (1999). Semi-Automatic Generation of F-Structures from Treebanks. In *Proceedings of LFG-99.*

WEDEKIND J. (1988). Generation as structure driven derivation. In *12th International Conference on Computational Linguistics*, p. 732–737.