Seeing the Wood for the Trees: Data-Oriented Translation

Mary Hearne & Andy Way

School of Computing,
Dublin City University,
Dublin 9, Ireland.

Email: {mhearne,away}@computing.dcu.ie

Abstract

Data-Oriented Translation (DOT), which is based on Data-Oriented Parsing (DOP), comprises an experience-based approach to translation, where new translations are derived with reference to grammatical analyses of previous translations. Previous DOT experiments [Poutsma, 1998, Poutsma, 2000a, Poutsma, 2000b] were small in scale because important advances in DOP technology were not incorporated into the translation model. Despite this, related work [Way, 1999, Way, 2003a, Way, 2003b] reports that DOT models are viable in that solutions to 'hard' translation cases are readily available. However, it has not been shown to date that DOT models scale to larger datasets. In this work, we describe a novel DOT system, inspired by recent advances in DOP parsing technology. We test our system on larger, more complex corpora than have been used heretofore, and present both automatic and human evaluations which show that high quality translations can be achieved at reasonable speeds.

1 Introduction

[Poutsma, 1998, Poutsma, 2000a, Poutsma, 2000b] presents a statistical approach to machine translation (MT) based on Data-Oriented Parsing (DOP: [Bod, 1998, Bod, Scha and Sima'an, 2003]). DOT models constitute an experience-based approach to translation, in that translations of previously unseen input are derived with reference to a set of linked ⟨source,target⟩ tree fragments in the system's database. In section 2, we present some characteristics of the DOP approach to parsing, and show how the DOP fragmentation operations of *root* and *frontier* are ported to DOT. We also show how the DOP composition operation is adapted to data-oriented models of translation, and give equations which demonstrate our probability model for DOT. We also describe the innovative notion of 'link depth' which we consider to be a more motivated method (compared to the more arbitrary notion of fragment depth) of pruning the example base from which samples are taken (the 'competition set') in order to try to derive new translations.

We describe Poutsma's DOT models of translation in section 3. A DOT system consists of a DOP parser which has been extended to handle pairs of fragments rather than single fragments. Therefore, parsing technology forms the backbone of any DOT system and all of the challenges of developing a DOP parser must also be met when implementing DOT. Advances in high-performance parsing technology are essential to any DOT system if large-scale translation experiments on complex linguistic data are to be carried out. We describe our implementation of a DOT system incorporating optimisations—inspired by those developed for DOP—in section 4. Our translation model facilitates increased ef£ciency in terms of fragment extraction, the building of a compact representation of the translation space and the selection of the most probable translation.

In section 5, we describe a set of experiments carried out on a large subset of the HomeCentre corpus to test our DOT system. The HomeCentre corpus contains 810 aligned English-French sentence pairs from Xerox documentation parsed into LFG

c(onstituent)- and f(unctional)-structure representations. This bilingual treebank provides us with a linguistically complex fragment base on which to perform experiments on a larger scale than those carried out to date. [Frank, 1999] observes that the corpus contains many 'hard' translation examples, including cases of nominalisations, relation-changing, passivisation, headswitching, complex coordination, and combinations thereof. Accordingly, the corpus would appear to present a challenge to any MT system, but given that these cases are widespread in real data, any MT system will ultimately be judged—on the level of quality, at least—on how it copes with such phenomena. We present the results of these experiments in terms of an automatic and a human evaluation of the output translations. Notably, our system achieves high quality translations in reasonable time. We contrast our results with previous data-oriented models of translation, and comment on some of the common errors that we suggest could easily be £xed by a more linguistically sophisticated system such as LFG-DOT [Way, 1999, Way, 2003a, Way, 2003b]. Finally, we conclude and provide some avenues for further research.

2 Theoretical Background

2.1 Data-Oriented Parsing

Data-oriented models of language, e.g. [Bod, 1998, Bod, Scha and Sima'an, 2003], are based on the assumption that humans perceive and produce language by availing of previous language experiences rather than abstract grammar rules. Tree-DOP models exploit treebanks comprising phrase-structure trees representing previously occurring utterances. Analyses of previously unseen input sentences are produced by combining these fragments and the most probable analysis is determined via their relative frequencies. LFG-DOP models [Bod and Kaplan, 1998] extend DOP by incorporating the representations of Lexical Functional Grammar (LFG) which can capture and represent linguistic phenomena other than those occurring at surface level.

Drawbacks of the DOP approach centre around issues of ef£ciency. Recent advances in parsing have sought—with some success—to address these issues. As the set of fragments extracted from a treebank of reasonable size is generally both extremely large and extremely redundant, pruning strategies have been developed in an attempt to constrain the number of fragments without reducing parse accuracy [Bod, 2001]. This work has led to the formation of the DOP hypothesis, which states that parse accuracy increases with increasing fragment size. Optimised algorithms to compute the parse space of an input sentence over large fragment bases have also been developed [Goodman, 1996, Sima'an, 1999]. Extraction of the most probable parse constitutes an NP-complete problem [Sima'an, 1999] as many different derivations can result in the same parse and, therefore, the most probable derivation (MPD) and the most probable parse (MPP) are not necessarily the same. Monte-Carlo sampling involves searching over a reduced random sample of the search space. It has been proposed as a method for calculating the MPP in DOP [Bod, 1998] and the approach has been further re£ned by [Chappelier and Rajman, 2003].

2.2 Data-Oriented Translation

Data-Oriented Translation exploits bilingual treebanks comprising linguistic representations of previously seen translation pairs, as well as explicit links which map the translational equivalences present within these pairs at sub-sentential level. Analyses and translations of the input are produced simultaneously by combining source and target language fragment pairs from the treebank. That is, there is no distinction between the separate phases of analysis, transfer and generation as in transfer-based MT, for instance. In this sense, a DOT system can be viewed as a DOP parser which has been adapted to process fragments which consist of pairs of subtrees rather than single subtrees.

The tree fragment pairs used in Tree-DOT are called *subtree pairs*. The two decomposition operators, which are similar to those used in Tree-DOP but are re£ned to take the translational links into account, are as follows:

- the *root operator* which takes any pair of *linked* nodes in a tree pair to be the roots of a subtree pair and deletes all nodes except these new roots and all nodes dominated by them;

- the *frontier operator* which selects a (possibly empty) set of *linked* node pairs in the newly created subtree pairs, excluding the roots, and deletes all subtree pairs dominated by these nodes.

The DOT composition operator is de£ned as follows. The composition of tree pairs $<s_1,t_1>$ and $<s_2,t_2>$ ($<s_1,t_1> \circ <s_2,t_2>$) is only possible if

- the leftmost non-terminal frontier node of s_1 is of the same syntactic category (e.g. S, NP, VP) as the root node of s_2, and
- the leftmost non-terminal frontier node of s_1's *linked counterpart* in t_1 is of the same syntactic category as the root node of t_2.

The resulting tree pair consists of a copy of s_1 where s_2 has been inserted at the leftmost frontier node and a copy of t_1 where t_2 has been inserted at the node linked to s_1's leftmost frontier node.

As in DOP, the DOT probability of a translation derivation is the joint probability of choosing each of the subtree pairs involved in that derivation. The probability of selecting a subtree pair is its number of occurrences in the corpus divided by the number of pairs in the corpus with the same root nodes as it:

$$P(< e_s, e_t >) = \frac{|<e_s,e_t>|}{\sum_{<u_s,u_t>:r(<u_s,u_t>)=r(<e_s,e_t>)} |<u_s,u_t>|}$$

The probability of a derivation in DOT is the product of the probabilities of the subtree pairs involved in building that derivation. Thus, the probability of derivation $<s_1,t_1> \circ ... \circ <s_n,t_n>$ is given by

$$P(< s_1, t_1 > \circ ... \circ < s_n, t_n >) = \prod_i P(< s_i, t_i >)$$

Again, a translation can be generated by many different derivations, so the probability of a translation $w_s \longleftrightarrow w_t$ is the sum of the probabilities of its derivations:

$$P(< w_s, w_t >) = \sum_{<t_{s_i},t_{t_i}> yields <w_s,w_t>} P(< t_{s_i}, t_{t_i} >)$$

While the translation process under DOT clearly mirrors the DOP parsing process, DOT fragments suffer from limited compositionality where DOP does not [Way, 2003a, Way, 2003b]. In DOP, a fragment with root category NP can be freely composed with any fragment whose leftmost substitution site is also of category NP. Under DOT, however, a source fragment with root category NP can only be composed with a fragment whose leftmost substitution site is of category NP if their target categories also correspond—for example, a pair with roots <NP,PP> cannot be composed with a pair whose leftmost substitution categories are <NP,NP>. Thus, the number of potential compositions is reduced. Our DOT model is no different from those of Poutsma in this respect.

2.3 Pruning: link depth

The re£nement of the fragmentation process to account for translational links may (and often does) result in a smaller number of DOT fragment per tree pair than would be the case with DOP. However, pruning methods to constrain the size of the fragment base are still necessary. Several pruning criteria have been proposed [Bod, 2001], one of which involves restricting the fragment base with respect to depth: fragments above a certain depth are excluded from the fragment base. Since, for fragments consisting of a single tree, any node can be designated a substitution site, such fragments can be pruned at any node. However, the de£nition of fragment depth becomes less obvious when the fragments in question consist of pairs of linked subtrees. For linked subtree pairs, only linked nodes can be designated substitution sites and, therefore, such fragments can only be pruned at linked nodes—to do otherwise would result in source substitution sites with no linked counterpart in the associated target trees. Furthermore, as linked source and target trees frequently differ with respect to depth, an arbitrary decision would have to be taken as to whether depth is calculated over the source or target trees. Consequently, we replace the notion of fragment depth—the greatest number of steps taken to get from the root node to any frontier node—with the notion of *link depth* for fragments comprising linked subtree pairs. The *link depth* of a fragment is the greatest number of steps taken *which depart from a linked node* to get from the root node to any frontier nodes. This yields the same result whether calculated over

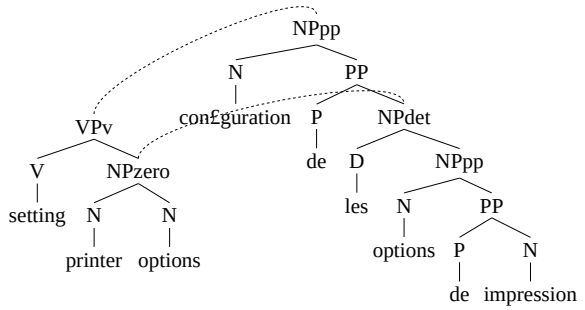

Figure 1: source depth = 3, target depth = 6, *link depth* = 2

the source or target fragment, as shown in Figure 1.

3 Poutsma's original DOT systems

In developing his DOT2 system,[1] Poutsma does not take advantage of the optimisations which have been developed for DOP. Rather, fragments are created explicitly and converted to rewrite rules of the form

$$<root_s, root_t> \longrightarrow$$
$$<(frontier_{s_1} ... frontier_{s_n}>),(<frontier_{t_1} ... frontier_{t_n})>$$

with links indicating translational correspondences between frontiers. He then relies on standard parsing algorithms to generate the chart representing the parse space. He uses Monte-Carlo sampling during disambiguation in order to determine the most probable translation, limiting the sample size to 1500 derivations.

Poutsma's model [Poutsma, 2000b] is tested via experiments carried out on a subset of the Verbmobil corpus, which contains transcribed spoken dialogues in the domain of appointment scheduling in German, English and Japanese. Poutsma uses just the English and German strings in his experiments—the English Verbmobil strings are annotated with Penn-II Treebank annotations, and the German strings with the Tübingen scheme. He manually analysed each tree pair and inserted translational links where necessary. In total, his dataset comprises 266 tree pairs yielding a maximum of 33,479 fragments.

[1]For the purposes of this discussion, we ignore here the original DOT1 model [Poutsma, 1998] which was shown to derive wrong translations where the ⟨source,target⟩ word order differed to any degree [Way, 1999]. This was due to the fact that the recombination operator in DOT was de£ned on source trees only rather than ⟨source,target⟩ pairs of trees. Poutsma named his subsequent model of translation DOT2, which overcomes this particular problem of DOT1.

Poutsma uses 226 of the 266 Verbmobil tree pairs as a training set, holding out 40 tree pairs to test the system. He translates both from English to German and from German to English, and each time uses a different training–test split to provide more representative results. In a manual evaluation, he provides £gures for exact match, alternate (a different, though reasonable translation from the one given), wrong (invalid translations), and partial exact/alternate. Comparative results are provided using *Babel£sh* as a baseline. Poutsma also provides separate results depending on what depth of fragment is included in the system database.

For both language pairs, DOT2 generated more exact match translations than *Babel£sh*—about 2.5% more for German–English and 8-10% more for English–German. Where *Babel£sh* outperforms DOT is in the 'alternate' translation category—15-23% more for German–English and 32-35% more for English–German. DOT also produces many more ungrammatical translations for English–German (14% more), but far fewer for German–English (8% fewer). As for 'wrong' translations, *Babel£sh* generates far more of these—for English–German, 17-30% more, and for German–English, 28-34% more. It is impossible to provide comparative results for 'partial' translations as *Babel£sh* does not produce these.

It is reasonably easy to provide an explanation as to why DOT outperforms *Babel£sh* in certain categories: given that DOT was trained on Verbmobil data whereas *Babel£sh* is a general purpose MT system, one would expect DOT to do well when confronted with similar test data. Nevertheless, it seems from Poutsma's £gures that for German–English, DOT is about as likely to produce an exact (13-15%) translation as an ungrammatical (13%) one, while 'wrong' translations also appear maximally 13% of the time. For English–German, DOT is more likely to generate an ungrammatical translation (32%) than an exact one (19%), with 'wrong' translations appearing around 5% of the time. Disappointingly, and contrary to the DOP hypothesis, the performance of the DOT model does not improve when fragments of greater depth are included in the system database. Poutsma explains this by the fact that "the trees in our corpus contained a lot of lexical content ... at very small tree depths".

Poutsma's system would appear to be feasible solely because his experiments were carried out on a small scale. However, having implemented a DOP parser using similar rewrite rules and conventional parsing techniques, it is clear to us that such systems cannot work with larger fragment bases. Therefore, we feel that it would not be possible to carry out further experiments with a larger fragment base using his approach. His system is simplistic, and yet despite this, the results in [Poutsma, 2000b] are not overly encouraging.

4 Optimised Data-Oriented Translation

As the driving technology behind any DOT system is a DOP parser, DOT can only be implemented ef£ciently and robustly if advances in DOP technology are incorporated.

4.1 Creation of the fragment base

In order to test our MT system, we employed a dataset comprising the £rst 605 aligned English-French sentence pairs from the HomeCentre corpus which we manually annotated with translational links. In total, our dataset yields in excess of 343 million fragments. Although fewer fragments are extracted per tree pair than for DOP, the number of bilingual fragment pairs extracted is still signi£cant and, as we are dealing with pairs of trees, the number of actual fragments created is double this £gure. Clearly, generating, storing and searching this number of fragments, as well as gathering frequencies of occurrence for each subtree pair, is a non-trivial task.

We have developed a dynamic method to generate a compact representation of all fragments which can be derived from a particular tree pair [Hearne and Way, 2003]. Fragments generated by the *root* operation are extracted as usual. These fragments are then decorated with a set of unique identi£ers referring to each fragment which can be extracted via the *frontier* operation. Frequency information is calculated by recursively comparing all decorated trees and identifying duplicates. This method allows us to store and access only the original treebank trees, thus alleviating the need to explicitly create the fragment base—a task which, given a corpus of reasonable size and complexity, quickly becomes unfeasible. Instead, we can ef£ciently retrieve only those fragments directly useful in translating the given input string.

4.2 Construction of the translation space

A chart built during the analysis phase is a compact representation of all possible derivations leading to valid parsed translations of the input string, which can be constructed either bottom-up or top-down. In order to build a translation chart using conventional chart-parsing techniques, each fragment pair must be expressed as a rewrite rule where links between frontiers are preserved and a direct reference to the original fragment structure is be retained—this is the approach taken by Poutsma. However, these approaches are not equipped to handle the sheer numbers of fragments involved in large-scale translation within the data-oriented framework.

We have developed a two-phase analysis component based on the DOP optimisation proposed by Sima'an [Sima'an, 1999]. However, we have optimised for top-down computation of the most probable translation rather than bottom-up computation of the most probable derivation. The £rst phase of analysis involves using the context-free grammar underlying the source side of the corpus to compute an approximation of the parse space for the input using the CKY algorithm. Given that the grammar underlying the English section of the HomeCentre corpus comprises just 2606 rules, this clearly constitutes a dramatic reduction of the initial search space. During the second phase, the set of bilingual fragments is applied to this reduced parse space to generate the exact DOT translation space for the given input. In order to do so, a correspondence is drawn between the context-free grammar rules used during the £rst phase and the tree fragments we wish to insert into the chart during the second phase. The fragmentation process described in the previous section provides these correspondences because they allow the extraction of unique identi£ers for all fragments associated with each context-free grammar rule. The appropriate fragments are rebuilt using these identi£ers, thus allowing for a highly optimised second analysis phase. Further details can be found in [Hearne and Way, 2003].

4.3 Computation of the output translation

Disambiguation, the £nal stage in the translation process, involves selecting the most probable translation or derivation from the translation chart. Monte-Carlo sampling has been proposed as a method for maximisation of the MPP in the DOP framework [Bod, 1998] and we have applied this technique to selection of the MPT for DOT. A fragment is chosen at random from the top of the chart. Fragments chosen at random from appropriate chart positions and which have appropriate root categories are then successively composed with this fragment until there are no open substitution sites left, at which point the derivation is complete. When suf£cient samples have been seen, the translation which occurs most frequently in the sample corresponds to the MPT.

5 Experiments and Results

Having manually aligned the ⟨source,target⟩ tree fragments from the £rst 605 aligned English-French sentence pairs from the HomeCentre corpus, we divide our dataset into 8 different training/test set splits, where each training set contains 545 parsed sentence pairs and each test set 60 sentence pairs. One restriction was placed on the training/test splits, namely that all words occurring in the source side of the test set had to also occur in the source side of the training set, but not necessarily with the same lexical category. All translations carried out were from English into French. Finally, we limited the number of samples taken during the disambiguation process to 5000.

Link Depth	1	2	3
No. fragments	4,506	23,478	104,400
Secs/sentence	17.80	16.27	15.33
Coverage (%)	66.47	67.92	67.92
Type 1 fail (%)	11.46	11.46	11.46
Type 2 fail (%)	1.04	1.04	1.04
Type 3 fail (%)	21.04	19.58	19.58

Table 1: Quantitative evaluation of DOT on the HomeCentre Corpus

5.1 Coverage

As the size of the fragment base increases, the number of sentences for which translations can be produced remains relatively steady. As can be seen in Table 1, there is a slight increase in coverage from 66.47% at link depth 1 to 67.92% at link depth 2 and no increase at link depth 3. There are 3 possible reasons why a particular sentence cannot be translated, which we have classi£ed as types 1, 2 and 3.

- A type 1 failure occurs where a complete parse space cannot be constructed for the source sentence using the CFG underlying the source side of the training set. As all words in the test set also occur in the training set, this generally indicates a word of unknown category—this is also a major problem for DOP [Bod, 1998].

- A type 2 failure occurs where a complete parse space can be constructed for the source sentence using the CFG but not using the fragments extracted from the source side of the training set. This situation does not arise using a monolingual fragment base as the minimal set of depth 1 fragments corresponds exactly to the set of underlying CFG rules. This is not the case for DOT, however, as the minimal set of fragments is of *link depth* 1 rather than depth 1 (cf. Figure 1).

- A type 3 failure occurs where a complete parse space can be constructed for the source sentence using both the CFG and the fragment base extracted from the source side of the training set but a complete translation space cannot be constructed using the bilingual fragment base as DOT fragments suffer from reduced compositionality.

5.2 Automatic evaluation of quality

Table 2 shows IBM Bleu scores using the NIST MT Evaluation Toolkit[2] for DOT at each link depth. The Bleu scores – calculated over translations actually produced – range from 0.7018 when only fragments of link depth 1 are considered, to 0.7838 when all fragments up to link depth 3 are included in the competition set. The absolute bleu scores range from 0.2911 to 0.3472. Such scores are possible given the linguistic sophistication of the treebank—the availability of good contextual information ensures that only suitable fragments are considered

[2] http://www.nist.gov/speech/tests/mt/mt2001/index.htm

where translations are derived by recombining different subtree pairs. Of course, this is only achievable given the effort taken to manually construct the set of ⟨source,target⟩ tree fragments in the system's database. However, we are con£dent that better Bleu scores are achievable when we augment our translation models with the syntactic information contained in the LFG f-structures in the Homecentre corpus.

Link Depth	1	2	3
Score 4 (%)	60.12	74.13	75.52
Score 3 (%)	27.32	14.18	13.22
Score 2 (%)	8.40	7.38	5.95
Score 1 (%)	4.15	4.31	4.31
BLEU score	0.7018	0.7456	0.7838

Table 2: Qualitative evaluation calculated over translations produced

5.3 Manual evaluation of quality

In order to manually evaluate the quality of our MT system, we assigned each translation produced to one of the following categories:

- Category 4: perfect translation (exact/alternative);
- Category 3: good quality translation with minor syntactic or translation errors;
- Category 2: partially intelligible translation with major syntactic or translation errors;
- Category 1: unintelligible.

Two native speakers of French with ¤uent English carried out this task. As shown in Table 2, translation quality improved consistently as the size and complexity of the fragment base increased. Perfect translations ranged from 60.12% to 75.52% as link depth increased. Note that the Bleu scores in Table 2 are quite similar to these Category 4 manual evaluations, which bears out the claim that Bleu scores are intended to correlate highly with those of human evaluators. Furthermore, minor and major grammatical and translation errors decreased, ranging from 27.32% to 13.22% and from 8.40% to 5.95% respectively, as more fragments were included. A good example is *the page is printed.* ↔ *le page est imprimé.* Here we see two agreement errors: between the determiner and noun, and between the subject NP and the ending on the past participle. Both errors would be easy to £x in LFG-DOT given the availability of syntactic information in the f-structures. The number of translations so poor as to be unintelligible remained relatively stable, ranging from 4.15% to 4.31%. These results appear to con£rm that the DOP hypothesis also holds for DOT as we have observed that translation accuracy also increases as larger subtree pairs are included in the fragment base.

5.4 Time

From Table 1 we observe that, contrary to intuition, the average time taken to translate each sentence decreases as more fragments are included in the fragment base. During the disambiguation process, fragments are sampled from the chart and substituted into the current derivation until no open substitution sites remain in that derivation. Where large fragments are selected, fewer fragments are subsequently sampled in completing the derivation, thus resulting in reduced disambiguation time. It is unclear—and, indeed, unlikely—that this trend will continue as link depth is increased; further experiments at greater link depths will be required to verify this.

Given that most criticisms of DOP-based approaches centre on problems of ef£ciency, we consider the translation times of between 15–18 seconds per sentence to be quite reasonable, particularly when the translation quality is taken into account. These were achieved on a Pentium 4 with 1.7GHz CPU and 750Mb RAM.

5.5 Contrasting Results

In terms of quality, we achieve perfect exact or alternative translations in 60.12%–75.52% of cases, whereas Poutsma reports results of 18.92%–24.33% for the same category. Our results, which also show increased quality as fragment depth increases, provide initial con£rmation that the DOP hypothesis also holds for DOT, contrary to Poutsma's £ndings. He suggests that this is due to the fact that his dataset contained much lexical context at small tree depths, and also that his dataset was small and of poor quality [Poutsma, 2000a]. Our £ndings would appear to con£rm this conclusion as our dataset is of high

quality and contains a greater degree of linguistic complexity.

Our innovation of *link depth* may also be important in con£rming the DOP hypothesis for DOT as Poutsma does not describe how he calculates the depth of a linked subtree pair. While these issues go some way towards explaining why our results have improved on those of Poutsma, it is also the case that our experiments have been performed on a different language pair. Therefore, we intend to extend our experiments both by translating from French to English and by working with the English-German section of the HomeCentre corpus.

6 Conclusions and future work

We have developed a high-performance data-oriented MT system which incorporates and adapts optimisations originally developed for DOP. We have tested this system on the complex and challenging HomeCentre corpus and have achieved promising results, both in terms of results and ef£ciency. We intend to perform further experiments—using alternative translation directions, language pairs and pruning parameters—in order to test our system comprehensively and, consequently, establish the data-oriented translation models as viable approaches to MT.

As the corpus is aligned at sentence level, subsentential translational equivalences must be inserted manually—to date we have completed 75% of the alignment process. Despite the reduced number of fragments produced for DOT, pruning of the search space is still essential. This involves rede£ning pruning parameters used for DOP—such as max. depth, max. no. of lexical entries, max. no. of substitution sites etc.—to render them functional with DOT fragments. We intend to complete this alignment process and test our system on the whole of the Homecentre corpus to see whether our good, interim results can be maintained.

We provided instances of translation errors which would be corrected in an LFG-DOT system. Errors of determiner-noun and subject-verb agreement, for example, would not be made if the syntactic information available in the LFG f-structures were available in the translation model. In addition, [Way, 2003a, Way, 2003b] has shown (albeit on small datasets) that the DOT problem of limited compositionality, whereby fragments cannot be adequately generalised and are therefore only reusable in very restricted circumstances with very small probabilities, can be avoided in LFG-DOT. We intend in further work to create a parser and translation system based on LFG-DOP [Bod and Kaplan, 1998], where the full LFG representations are allied with the techniques of DOP. We hope that the extension of our prototype DOT system to LFG-DOT will imporove upon the encouraging results achieved here when experiments are carried out using the f-structure annotations provided in the HomeCentre corpus.

7 References

Rens Bod. 1998. *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Publications, Stanford, CA.

Rens Bod. 2001. What is the Minimal Set of Fragments that Achieves Maximal Parse Accuracy? In *39th ACL/10th EACL*, Toulouse, France, pp.66–73.

Rens Bod and Ron Kaplan. 1998. A Probabilistic Corpus-Driven Model for Lexical Functional Analysis. In *COLING-ACL '98*, Montreal, Canada, pp.145–151.

Rens Bod, Remko Scha and Khalil Sima'an. eds. 2003. *Data-Oriented Parsing*. CSLI, Stanford CA.

Jean-Cédric Chappelier and Martin Rajman. 2003. Parsing DOP with Monte-Carlo Techniques. In Bod *et al*., eds. (2003).

Anette Frank. 1999. LFG-based syntactic transfer from English to French with the Xerox Translation Environment. *ESSLLI'99 Summer School*, Utrecht, The Netherlands.

Joshua Goodman. 1996. Ef£cient Algorithms for Parsing the DOP Model. *Proceedings Empirical Methods in Natural Language Processing*, Philadelphia, PA, pp.143–152.

Mary Hearne and Andy Way. 2003. A Robust, Wide-Coverage Data-Oriented Parser for Tree-Based Translation. Working paper, School of Computing, Dublin City University, Ireland.

Arjen Poutsma. 1998. Data-Oriented Translation. In *Proc. 9th CLIN*, Leuven, Belgium.

Arjen Poutsma. 2000a. *Data-Oriented Translation: Using the Data-Oriented Parsing framework for Machine Translation*. MSc thesis, University of Amsterdam, The Netherlands.

Arjen Poutsma. 2000b. Data-Oriented Translation. In *18th COLING*, Saarbrücken, Germany, pp.635–641.

Khalil Sima'an. 1999. *Learning Ef£cient Disambiguation*. PhD Thesis, University of Utrecht, The Netherlands.

Andy Way. 1999. A Hybrid Architecture for Robust MT using LFG-DOP. *Journal of Experimental and Theoretical Arti£cial Intelligence* **11**:441–471, Taylor and Francis, London.

Andy Way. 2003a. Machine Translation using LFG-DOP. In Bod *et al*., eds. (2003).

Andy Way. 2003b. Translating with Examples: The LFG-DOT Models of Translation. In *Recent Advances in Example-Based Machine Translation*, M. Carl and A. Way, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.443–472.