A Detailed Analysis of Phrase-based and Syntax-based Machine Translation: The Search for Systematic Differences

Rasoul Samad Zadeh Kaljahi[†‡] **Raphael Rubino**[†‡] **Johann Roturier**[‡] **Jennifer Foster**[†]

[†] NCLT, School of Computing
Dublin City University
Dublin 9, Ireland
firstname.lastname@dcu.ie

[‡]Symantec Research Labs
Ballycoolin Business Park
Blanchardstown, Dublin 15, Ireland
firstname_lastname@symantec.com

Abstract

This paper describes a range of automatic and manual comparisons of phrase-based and syntax-based statistical machine translation methods applied to English-German and English-French translation of user-generated content. The syntax-based methods underperform the phrase-based models and the relaxation of syntactic constraints to broaden translation rule coverage means that these models do not necessarily generate output which is more grammatical than the output produced by the phrase-based models. Although the systems generate different output and can potentially be fruitfully combined, the lack of systematic difference between these models makes the combination task more challenging.

1 Introduction

There has been a long tradition of using syntactic knowledge in statistical machine translation (Wu and Wong, 1998; Yamada and Knight, 2001). After the emergence of phrase-based statistical machine translation (Och and Ney, 2004), several attempts have been made to further augment these techniques with information about the structure of the language. Hierarchical phrase-based modelling (Chiang, 2007) emphasises the recursive structure of language without concerning itself with the linguistic details. On the other hand, syntax-based modelling uses syntactic categories in addition to recursion, in mapping from source to the target (Galley et al., 2004; Zollmann and Venugopal, 2006). Syntactic information is incorporated into the model from trees on the source side (tree-to-string), target side (string-to-tree), or both (tree-to-tree).

Utilisation of such linguistic generalisation, however, has proven to be a more complicated task than one might first imagine it to be. While relative improvements over phrase-based baselines have been reported for some language pairs, those baselines seem to remain the best option for other language pairs (DeNeefe et al., 2007; Zollmann et al., 2008).

The performance of syntax-based models is affected by errors introduced by existing imperfect syntactic parsers (Quirk and Corston-Oliver., 2006). Moreover, some non-syntactic phrases (e.g. *I'm*) identified by the phrase-based models bring useful information to the translation which are missed by syntax-based models trained on trees obtained using supervised parsing (Bod, 2007). Phrasal coherence between the two languages (Fox, 2002) is another factor affecting the performance of syntax-based models. Nevertheless, these models should in theory be better able to capture long-distance reordering — a problem for phrase-based models.

A combined framework of such varying techniques can exploit the advantages of all of them while compensating for the weaknesses of each individual method. To accomplish this goal, a more detailed insight into the characteristics of each method may be useful. Towards this objective, we look for possible systematic differences between variants of phrase-based and syntax-based systems via various analysis approaches.

In the rest of this paper, after introducing some related work, we will describe our data and baseline systems. These baseline systems will then be compared along several dimensions. Finally we will discuss our observations and conclude.

2 Related Work

DeNeefe et al. (2007) compared a string-to-tree model with a phrase-based model. While the syntax-based model performed better than the phrase-based model on Chinese-to-English translation, it was shown to be worse on Arabic-to-English translation. They found that non-lexical rules form only a small fraction of the translation rule table in syntax-based modelling. The string-to-tree modelling in this work is based on their approach.

Zollmann et al. (2008) observed that the gain achieved by hierarchical and syntax-based models could be largely compensated for by increasing the reordering limit in the phrase-based model. They also found that, for language pairs involving substantial reordering like Chinese-English, tree-based models performed better than phrase-based. However, for relatively monotonic pairs like Arabic-English, all models produced similar results.

Experimenting with French-English, German-English and English-German, Auli et al. (2009) compared a phrase-based model to a hierarchical phrase-based model by exploring as much of the search space of both types of models as was computationally feasible. Given that the search spaces were very similar, they concluded that the differences between the two types of models can be explained by the way they score hypotheses rather than by the hypotheses they produce.

Using the same framework as in this work, Hoang et al. (2009) compared phrase-based, hierarchical phrase-based and string-to-tree models. While the phrase-based and hierarchical phrase-based models achieved similar results, they both performed slightly better than the syntax-based model. They argue that, in order to improve syntax-based modelling, word alignment should be amended.

There have been several efforts to exploit the difference between such models in MT system combination or multi-engine machine translation (MEMT) (Huang and Papineni, 2007). The task, however, has been shown to be difficult. Zwarts and Dras (2008) tried to identify what type of sentence could be better translated by a syntax-based model compared to a phrase-based model. Using a classification approach, they separately tested three sets of features. Sentence length and system-internal features including decoder output score did not lead to an accurate classifier. They then hypothesised that noisy parse trees may impede the performance of the syntax-based system and built another classifier based on source sentence length, parser confidence score, and linked fragment count. They found, however, no correlation. Based on their observation that most of the problems in the output were related to reordering, they assumed that the syntactic quality of the output could be discriminative in system selection. They ported the parse-quality features used on the source side to the target side, but again found no improvement.

In what follows, we build upon previous work by analysing a more comprehensive set of SMT methods and comparing them on a diverse set of evaluation data in various automatic and manual ways.

3 Data

This work takes place in the context of a wider project the ultimate aim of which is to improve the quality of English-German and English-French machine translation on content taken from Symantec user forums. This is needed since the customer service model is moving away from the traditional one in which companies provide help via technical documentation, phone lines and email to one in which customers help each other via forums. The English forum contains far more content than the French and German forums and much of this content, if translated adequately, will be useful to Symantec's French and German customers.

The translation model training data for our machine translation systems consist of English-German and English-French Symantec translation memory. These translation memories contain a mixture of Symantec content from product manuals, software strings, marketing materials, knowledge bases and websites. The English-German parallel data contains 1,029,741 sentence pairs and the English-French 975,102 pairs with no exact duplicates. Both the French and German language models are trained on a combination of the target side of their respective translation model training data and the limited amount of user forum text that is available for each language (42K sentences for French and 67K sentences for German). We have two evaluation sets for each language pair:

1. **French translation memory**: 5,000 held-out sentences from the Symantec English-French translation memory, split into development (2000) and test (3000).
2. **German translation memory**: 5,000 held-out sentences from the Symantec English-German translation memory, split into development (2000) and test (3000)
3. **French forum data**: 1,500 sentences taken from the Symantec English online forums, split into development (600) and test (900). These were automatically translated into French using an online translation tool and then post-edited by human translators.
4. **German forum data**: 1,500 sentences taken from the Symantec English online forums, split into development (600) and test (900). These were automatically translated into German using an online translation tool and then post-edited by human translators.

The translation memory data can be considered a superset of forum data in terms of subject matter. However, in terms of style, the forum data is more informal and, given that it is user-generated content, we assume that it exhibits a higher level of ungrammaticality. Because of this difference, we call the evaluation sets taken from translation memory *in-domain* and those from forum text *out-of-domain*. While the English sides of the in-domain sets are different for each of the language pairs, those of the out-of-domain sets are the same for both pairs.

4 Baseline Systems

We train the following five statistical machine translation systems:

1. **PB**: a standard phrase-based system (Och and Ney, 2004)
2. **HP**: a hierarchical phrase-based system (Chiang, 2007)
3. **TS**: a tree-to-string syntax-based system (Huang et al., 2006).
4. **ST**: a string-to-tree syntax-based system (DeNeefe et al., 2007).
5. **TT**: a tree-to-tree syntax-based system

We chose these five systems because they can be easily built using the open source Moses toolkit (Hoang et al., 2009). The PB system was trained using the `grow-diag-final-and` alignment heuristic and used the `msd-bidirectional-fe` reordering model. All other parameters were default including a maximum phrase length of 7 and a decoder distortion limit of 6 when applied. The HP system was trained using the default settings including a maximum chart span of 20. The same chart span was used for the TS,ST and TT systems. To relax the strict constraint on rule extraction in the TS, ST and TT systems, any pairs of adjacent nodes in the parse tree are combined together to form new nodes (Zollmann et al., 2008)[1]. This significantly increases the number of extracted rules and consequently the translation accuracy. All five systems are tuned using minimum error rate training (MERT) (Och, 2003) on the respective developments sets.

We use our in-house C++ implementation of a PCFG-LA parser (Attia et al., 2010) to provide the parse trees for the English and French sides of the translation training data and the English sides of the evaluation data for source-syntax systems (TS and TT). The German side of the translation training data was parsed by the Berkeley parser (Petrov et al., 2006). Both parsers use the max-rule parsing algorithm (Petrov and Klein, 2007). We use the Tiger treebank (Brants et al., 2002) for training the German parsing model, the French Treebank (Abeillé et al., 2003) for training the French model and the Wall Street Journal section of the Penn Treebank (Marcus et al., 1994) for training the English model.[2]

5 System Comparison

5.1 Multiple Metrics

In order to carry out a reliable comparison, we evaluate the baseline systems at the document level using

[1] We used the SAMT-2 parse relaxation method.

[2] The two parsers achieve Parseval labelled f-scores in the 89-90 range on Section 23 of the Wall Street Journal section of the Penn Treebank. Due to some character encoding issue, our own parser could not be used to parse the German data and this is why the Berkeley parser is used instead.

	En-Fr					*En-De*				
	BLEU	NIST	TER	GTM	METEOR	BLEU	NIST	TER	GTM	METEOR
PB	0.6140	10.73	0.3584	0.6357	0.7436	0.5099	9.48	0.4911	0.5441	0.6264
HP	**0.6188**	**10.78**	**0.3535**	**0.6400**	**0.7457**	**0.5289**	**9.68**	**0.4676**	**0.5592**	**0.6408**
TS	0.5919	10.39	0.3719	0.6194	0.7284	0.4939	9.19	0.4923	0.5349	0.6146
ST	0.6013	10.53	0.3631	0.6258	0.7334	0.5086	9.43	0.4753	0.5479	0.6265
TT	0.5783	10.25	0.3842	0.6096	0.7168	0.4784	9.03	0.5059	0.5219	0.6051
Oracle 1-best	0.6658	11.38	0.2917	0.6840	0.7818	0.5739	10.32	0.3858	0.6111	0.6775
Oracle 500-best	0.7770	12.77	0.1779	0.7852	0.8616	0.6870	11.80	0.2584	0.7145	0.7712

Table 1: Baseline and oracle system combination scores on in-domain development set (translation memory)

	En-Fr					*En-De*				
	BLEU	NIST	TER	GTM	METEOR	BLEU	NIST	TER	GTM	METEOR
PB	**0.3044**	6.90	0.6024	0.3972	0.5335	**0.1681**	5.500	0.7428	0.3062	**0.4057**
HP	0.3032	**7.04**	**0.5904**	**0.4008**	**0.5341**	0.1662	5.502	0.7384	0.3082	0.4028
TS	0.2907	6.71	0.6118	0.3924	0.5202	0.1643	**5.503**	**0.7197**	**0.3128**	0.3977
ST	0.2982	6.75	0.6057	0.3952	0.5248	0.1654	5.471	0.7286	0.3117	0.3976
TT	0.2900	6.68	0.6121	0.3910	0.5166	0.1633	5.371	0.7358	0.3090	0.3966
Oracle 1-best	0.3343	7.40	0.5408	0.4265	0.5585	0.1935	5.921	0.6700	0.3376	0.4248
Oracle 500-best	0.3921	8.25	0.4717	0.4687	0.6117	0.2457	6.730	0.6049	0.3791	0.4750

Table 2: Baseline and oracle system combination scores on out-of-domain development set (forum text)

five popular metrics: BLEU (Papineni et al., 2002), NIST, TER (Snover et al., 2006), GTM (Turian et al., 2003) and METEOR (Denkowski and Lavie, 2011)[3]. The results with these metrics for in-domain and out-of-domain development sets are presented in Table 1 and Table 2 respectively. The last two rows of each table are described in section 5.5. We report scores on the development sets as our analysis has been performed on these.

Performance is considerably higher for the in-domain evaluation sets compared to the out-of-domain ones and for the En-Fr compared to En-De. Neither of these results are surprising since it is well known that out-of-domain translation is challenging and that English-German translation is more difficult than English-French translation. It is worth noting that the gap between En-Fr and En-De scores on out-of-domain data is bigger than on in-domain data, showing that out-of-domain En-De is a more difficult translation setting compared to the others.

The hierarchical phrase-based system (HP) performs better than the others on the in-domain data according to all metrics. This is statistically significant in the case of BLEU scores with p-value < 0.01[4]. The gap is more pronounced on the En-De pair, which is an intuitively appealing result because the hierarchical phrase-based model is in theory better able to model the systematic word order differences between English and German than the phrase-based model. The phrase-based system (PB) is the second best performing system on in-domain data.

The string-to-tree system (ST) is the best of the syntax-based systems on in-domain data according to all metrics. The tree-to-tree model (TT), on the other hand, is the worst-performing of these systems, despite its relatively larger translation rule table size. In the case of BLEU scores, these differences are also statistically significant. This shows that less useful rules are extracted by this model compared to the other two models.

On out-of-domain data however, the behaviour of systems is not consistent, with different metrics favouring different systems for different language pairs. On En-Fr, HP is still the best overall, and

[3] We used all default parameters for evaluation tools. For GTM, we used 1.2 as the exponent.

[4] We used paired bootstrap resampling with 1000 iterations and p-value = 0.01 for all BLEU significance tests. The tool used is available at http://www.ark.cs.cmu.edu/MT/

ST is the best performing syntax-based system (all statistically significant in the case of BLEU). On the other hand, more inconsistent behaviour is observed on En-De: TS scores the best of all according to most of the metrics (though marginally), and HP is no longer the best. However, the BLEU differences are not statistically significant.

5.2 One-to-one Comparison

Given the same training material, we are interested in the extent to which the methodological differences between these systems lead to different outputs. Since the output of all systems are far from a perfect translation, the more similar the outputs, the less effective the complex methods (tree-based methods here) compared to the phrase-based method which is usually a baseline in machine translation research. In addition, if systems tend to generate highly similar outputs, their combination cannot yield a noticeably better result.

To inspect this phenomenon for systems built here, we score each system against all others using the BLEU metric. In other words, each system output plays the role of reference translation for the other four systems. According to the results presented in Table 3 and Table 4:

1. HP and PB are consistently the most similar to each other (highest BLEU), whereas TT and PB are the most different (lowest BLEU).
2. The closest syntax-based system to PB and HP is TS.
3. It cannot be said which of the two syntax-based systems are the most distant ones from each other in general. It differs according to the data sets, but TT is usually one side of the pair.
4. Systems produce more divergent output on out-of-domain data and on the En-De pair than on in-domain data and on the En-Fr pair.

5.3 Sentence-level Comparison

To gain further insight into the difference between systems, we compared their output sentence-by-sentence using the TER evaluation metric (Snover et al., 2006). Table 5 shows the results of this comparison. The first row displays the number of sentences on which all systems scored the same. The second row contains the number of sentences for which all systems generated exactly the same output sentence. The following five rows, one for each system, present the number of sentence translations on which that system scores the highest, possibly along with the other systems, and the number of sentence translations on which that system scores the highest alone. We call the former *any-wins* and the latter *solo-wins*.

As presented in the table, the any-win ranking is not consistent with the solo-win ranking, especially on the in-domain sets. For example, on in-domain En-De, while HP ranks the highest in terms of any-wins (612 sentences), ST is the one with the most solo-wins (174 sentences). This may suggest that HP is mostly the best on the sentences on which the other systems perform similarly, whereas ST is capable of better translating sentences with which others have more trouble.

In addition, it can be observed that, on about one third of the in-domain sets, systems achieve the same scores (score ties), most of them being exactly the same translations (real ties). The ratio is, however, far less for out-of-domain data sets: only about %4. Given the performance gap between these two domains (Table 1 and Table 2), this discrepancy is expected to some degree: the closer the outputs to the reference, the less divergent they can be. However, this large ratio disparity does not seem to be only justified by this fact, suggesting that the real difference between systems is revealed on more difficult tasks. Consequently, more gain is expected from combined systems on out-of-domain data than on in-domain data.

5.4 N-best Comparison

So far our analysis has been carried out on the highest ranked translation returned by each system. We now compare the 500-best (distinct) output of systems. An interesting observation is that the size of the n-best lists is the largest for the least constrained system in terms of rule extraction (PB) and smallest for the most restricted one (TT). For each evaluation set, Table 6 shows some statistics on the overlap between the n-best outputs of the five systems. The figures show that there is larger overlap between the n-bests of the in-domain data than the out-of-domain data and the En-Fr pair than the En-De one. This is

	En-Fr				*En-De*			
	PB	HP	TS	ST	PB	HP	TS	ST
HP	**0.8535**	-	-	-	**0.7576**	-	-	-
TS	0.7799	0.7958	-	-	0.6817	0.7071	-	-
ST	0.7769	0.7917	0.7980	-	0.6624	0.6940	0.6778	-
TT	**0.7339**	0.7430	0.8065	0.7893	**0.6405**	0.6484	0.7113	0.7068

Table 3: One-to-one BLEU Scores on in-domain development set (translation memory data)

	En-Fr				*En-De*			
	PB	HP	TS	ST	PB	HP	TS	ST
HP	**0.7501**	-	-	-	**0.6207**	-	-	-
TS	0.6640	0.7028	-	-	0.6023	0.6162	-	-
ST	0.6618	0.6959	0.6731	-	0.5700	0.5802	0.6191	-
TT	**0.6165**	0.6344	0.7122	0.6764	**0.5211**	0.5365	0.6027	0.6014

Table 4: One-to-one BLEU Scores on out-of-domain development set (forum text)

consistent with our other observations and appears to suggest that the more difficult the sentences are to translate, the more differently the systems perform on them.

5.5 Oracle Combination

Using the sentence-level TER scores for each data set, we select the best translation for each sentence and form the oracle combined output of all systems. In case of score ties, we choose the output of systems in this order: PB, HP, ST, TS, and TT. The list is sorted by the computational cost of training and translating with each system. We also build an oracle by merging and reranking n-bests of all systems using TER scores.

The oracle combination outputs are evaluated using all the metrics. The scores are presented in the last two rows of Table 1 and Table 2. As expected, there are large gaps between the best performing systems on each data set and the oracle combinations, especially those of 500-best lists. In the case of BLEU and for the 1-best combination, the gaps are %7 and %9 on in-domain En-Fr and En-De and %10 and %15 on out-of-domain En-Fr and En-De respectively. For 500-best combination, these figures are %16, %19, %17, and %27. Apparently, the benefit from combination increases as the level of translation difficulty increases. This is further confirmation that the different systems built here behave more differently on more difficult data.

5.6 Sentence-level Manual Evaluation

Previous sections compared systems based on scores generated using automatic metrics. It is interesting and useful to know how these different systems handle various linguistic phenomena in translation. For example, one common argument in comparing syntax-based and phrase-based systems is that the former generates better word order in the output. In order to investigate such behaviours, we select 100 sentences from each development set, and compare the outputs of two of the systems, namely HP and ST, for each of these sentences. 50 of the selected sentences are the solo-win cases of HP and the other 50 are those of ST (see section 5.3). The reason why these two systems are selected is that HP is the overall best performing system, and ST is the best performing syntax-based systems according to the various comparisons so far.

Each data set was evaluated by a linguist using eight error categories. These categories are adapted from those used by Dugast et al. (2007) to evaluate post-editing changes. The evaluators were asked to count the number of errors in each output sentence under each category. While they were given the reference translation, they were not constrained to it and were allowed to compare against the closest correct translation to the output itself. We believe that this can better reflect the real performance of the systems, as it is not limited to a single reference, though we might lose some correlation with automatic met-

	In-domain				*Out-of-domain*			
	En-Fr		*En-De*		*En-Fr*		*En-De*	
Score ties	740		627		32		35	
Real ties	738		578		26		16	
PB Any/Solo wins	582	**130**	513	123	190	71	163	51
HP Any/Solo wins	**586**	95	**612**	125	**244**	**88**	172	43
TS Any/Solo wins	489	103	514	116	173	56	**208**	73
ST Any/Solo wins	517	125	572	**174**	177	60	205	**78**
TT Any/Solo wins	394	94	447	100	160	62	196	**78**

Table 5: Sentence-level TER-based System Comparison

	In-domain		*Out-of-domain*	
	En-Fr	*En-De*	*En-Fr*	*En-De*
# of sentences with a common n-best translation per sentence	1579	1367	202	169
% of sentences with a common n-best translation per sentence	%78	%68	%33	%28
Average number of common n-best translations per sentence	17	17	4	5

Table 6: N-best overlaps

rics. The following are the categories used in the evaluation, the first half of which can be considered to be grammar-related and the second half lexical.

1. *Verb tense:* number of wrong verb tense translations
2. *Gender/number agreement:* number of wrong gender and number agreements (e.g. adjective/noun gender in German)
3. *Local word order:* number of wrong local word orders
4. *Long-distance word order:* number of wrong long-distance word orders
5. *Mis-translated:* number of wrong word or phrase translations including wrong sense and unusual usage
6. *Untranslated:* number of words or phrases transferred to the output without translation
7. *Spurious translation:* number of words or phrases added to the output without any reference in source
8. *Missing translation:* number of words or phrases in source ignored by the system

The results of the manual evaluation are shown in Table 7. We observe the following:

- French word order (both local and long distance) is better handled by ST and German word order by HP.

- Since verb tense and gender/number agreement are handled in a methodologically similar way, the two categories can be collapsed for the purposes of comparison. From this point of view, HP generates better output. This gap is more pronounced on in-domain data.

- Though no generalizable pattern is seen for mis-translation, it can roughly be said that ST is less erronous than HP on this category.

- ST outputs overall fewer untranslated words. However, the gap is marginal. It, on the other hand, tends to generate more spurious translations. The only exemption is on in-domain En-De. On the other hand, HP misses more words and phrases.

It appears that no confident conclusion can be made based on the above observations. However, contrary to what one might expect, the syntax-based model is not necessarily better than the hierarchical model in treating syntactic phenomena in translation. The next section provides a closer scrutiny of the internal behaviour of the systems.

	In-domain				*out-of-domain*			
	En-Fr		*En-De*		*En-Fr*		*En-De*	
	HP	ST	HP	ST	HP	ST	HP	ST
Verb tense	13	**11**	**1**	3	25	25	21	**20**
Gender/number agreement	**27**	34	**25**	31	**64**	66	**60**	63
Local word order	29	**25**	**24**	31	53	**47**	**93**	99
Long-distance word order	7	**6**	**3**	4	8	**5**	**70**	82
Mis-translated	85	**84**	61	**55**	185	**177**	207	207
Untranslated	10	**9**	15	**13**	**92**	95	81	**72**
Spurious translation	**21**	26	29	**22**	**16**	21	**25**	35
Missing translation	**35**	41	42	**31**	18	**13**	140	**122**
Sum	**227**	236	201	**194**	461	**449**	**697**	700

Table 7: Manual evaluation results: number of translation errors by each system on each data set in each of 8 error categories

5.7 Error Analysis

Since all systems are built upon the same word alignment, under the same framework, and make use of the same training data for translation and language models, it is not surprising that their outputs are largely similar to each other.

Looking at the translation rule tables of each system and following their decoding process, it can be observed that relaxation blurs the boundaries between the phrase-based and syntax-based models. As in the phrase-based models, the rules in the syntax-based models are based on ad-hoc phrases and the only difference is in the set of nonterminals.

Table 8 illustrates an example in which neither of the rules used by systems to translate *you will need* is built upon a syntactic phrase. Nevertheless, unlike ST, HP translates it correctly. It is worth noting that there were eight similar rules in the rule table of ST (including the one used in the example) covering the span *, you will need X*, half of which could translate it correctly. However, due to a higher score, this rule was selected.

Another example concerning output word order, which is a major motivation behind incorporating syntax in machine translation, is presented in Table 9. Although the spans on which the ST rules have been applied are syntactic in this case, the first two rules have been wrongly chosen resulting in an invalid output word order. On the other hand, HP has correctly parsed the input and applied appropriate rules, leading to a correct output word order.

Despite the pitfalls of relaxation, without it, the syntax-based models suffer from limited translation rule coverage and produces significantly lower results. This confirms the need for a syntactic structure specialized for SMT.

6 Conclusions

We compared commonly used phrase-based and syntax-based models in SMT research in the context of a study on translating technical forum data from English into German and French. The results of various automatic evaluations showed that hierarchical phrase-based models are overall slightly better than others. One-to-one and sentence-by-sentence comparison and oracle combination of the output of all models showed that the more difficult the translation problem, the more different their output and the greater the gain to be achieved by combining outputs.

Manual analysis of the outputs and translation process showed that there was no obvious systematic difference between syntax-based and non-syntax-based modelling, mostly due to the relaxation of syntactic constraints on translation rule extraction. This makes it difficult to find features to be exploited in combining these models, despite the potential gain which was observed in their oracle combination. In the future, we hope to perform successful system combination by exploring the space of features used in our recent work on quality estimation (Rubino et al., 2012).

Another avenue for future work is to focus on improving parser accuracy on our datasets by leverag-

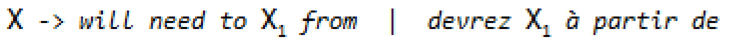
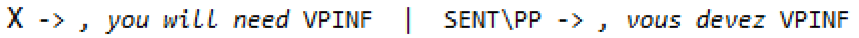
<table>
<tr><td>Source</td><td>If you choose to continue, **you will need to** set the options manually from the Altiris eXpress Deployment Server Configuration control panel applet.</td></tr>
<tr><td>Reference</td><td>Si vous décidez de continuer, **vous devrez** configurer les options manuellement à partir de l'applet du panneau de configuration Altiris eXpress Deployment Server.</td></tr>
<tr><td>HP output</td><td>Si vous décidez de continuer, **vous devrez** configurer les options manuellement à partir de l'applet Altiris eXpress Deployment Server Configuration Control Panel.</td></tr>
<tr><td>HP rule application</td><td>X -> will need to X_1 from | devrez X_1 à partir de</td></tr>
<tr><td>ST output</td><td>Si vous décidez de continuer, **vous devez** configurer les options manuellement dans l'applet de panneau de configuration de Altiris eXpress Deployment Server.</td></tr>
<tr><td>ST rule application</td><td>X -> , you will need VPINF | SENT\PP -> , vous devez VPINF</td></tr>
</table>

Table 8: Example of verb tense translation by two systems.

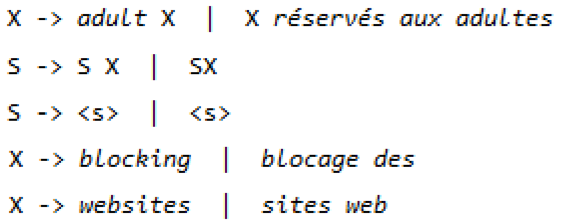
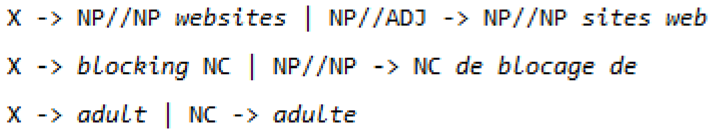
<table>
<tr><td>Source</td><td>blocking adult websites</td></tr>
<tr><td>Reference</td><td>blocage des sites web pour adultes</td></tr>
<tr><td>HP output</td><td>Blocage des sites Web réservés aux adultes</td></tr>
<tr><td>HP rule application</td><td>X -> adult X | X réservés aux adultes
S -> S X | SX
S -> &lt;s&gt; | &lt;s&gt;
X -> blocking | blocage des
X -> websites | sites web</td></tr>
<tr><td>ST output</td><td>Adulte de blocage de sites Web</td></tr>
<tr><td>ST rule application</td><td>X -> NP//NP websites | NP//ADJ -> NP//NP sites web
X -> blocking NC | NP//NP -> NC de blocage de
X -> adult | NC -> adulte</td></tr>
</table>

Table 9: Example of output word order of two systems.

ing our recent work in parsing user-generated content (Foster et al., 2011; Le Roux et al., 2012) in the hope that better phrase-structure trees will lead to better syntax-augmented machine translation.

Finally, the work that we have presented here has used syntax-augmented systems that can be conveniently built using the Moses toolkit. From the point of view of how best to integrate syntactic knowledge into the machine translation process, we are interested in investigating methods which can relax the syntactic constraint on rule extraction while retaining the constituency structure. This may involve using syntactic knowledge as a soft rather than a hard constraint (Marton and Resnik, 2008) or investigating translation models based on tree sequences (Zhang et al., 2008).

Acknowledgements

The research reported in this paper has been supported by the IRCSET Enterprise Partnership Scheme (EPSPG/2011/102 and EPSPD/2011/135). We are grateful to the reviewers for their insightful comments and to Linda Mitchell and Susanne Beekes for evaluating machine translation output.

References

A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a treebank for French. In *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

Mohammed Attia, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi, and Josef van Genabith. 2010. Handling unknown words in statistical latent-variable parsing models for Arabic, English and French. In *Proceedings of SPMRL*.

Michael Auli, Adam Lopez, Hieu Hoang, and Philipp Koehn. 2009. A systematic analysis of translation model search spaces. In *Proceedings of WMT*.

Rens Bod. 2007. Is the end of supervised parsing in sight? In *Proceedings of ACL*.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of TLT*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *Proceedings of EMNLP-CoNLL*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of WMT*.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of WMT*.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From news to comment: Benchmarks and resources for parsing the language of Web 2.0. In *Proceedings of IJCNLP*.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of HLT-NAACL*.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of IWSLT*.

Fei Huang and Kishore Papineni. 2007. Hierarchical system combination for machine translation. In *Proceedings of EMNLP-CoNLL*.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of AMTA*, July.

Joseph Le Roux, Jennifer Foster, Joachim Wagner, Rasul Samed Zadeh Kaljahi, and Anton Bryl. 2012. DCU-Paris-13 systems for the SANCL shared task. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the 1994 ARPA Speech and Natural Language Workshop*, pages 114–119.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrase-based translation. In *Proceedings of ACL-08:HLT*.

Franz Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

Franz Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact and interpretable tree annotation. In *Proceedings of the 21st COLING and the 44th ACL*.

Chris Quirk and Simon Corston-Oliver. 2006. The impact of parse quality on syntactically-informed statistical machine translation. In *Proceedings of EMNLP*.

Raphael Rubino, Jennifer Foster, Joachim Wagner, Johann Roturier, Rasul Samed Zadeh Kaljahi, and Fred Hollowood. 2012. DCU-Symantec submission for the WMT 2012 quality estimation task. In *Proceedings of WMT*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Joseph Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *In Proceedings of MT Summit IX*.

Dekai Wu and Hongsing Wong. 1998. Machine translation with a stochastic grammatical channel. In *Proceedings of ACL*.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL*.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree-sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-08:HLT*.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of WMT*.

Andreas Zollmann, Ashish Venugopal, Franz Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of COLING*.

Simon Zwarts and Mark Dras. 2008. The impact of parse quality on syntactically-informed statistical machine translation. In *Proceedings of COLING*.