Syntax and Semantics in Quality Estimation of Machine Translation

Rasoul Kaljahi[†‡], **Jennifer Foster**[†], **Johann Roturier**[‡]
[†]NCLT, School of Computing, Dublin City University, Ireland
{rkaljahi, jfoster}@computing.dcu.ie
[‡]Symantec Research Labs, Dublin, Ireland
{johann_roturier}@symantec.com

Abstract

We employ syntactic and semantic information in estimating the quality of machine translation from a new data set which contains source text from English customer support forums and target text consisting of its machine translation into French. These translations have been both post-edited and evaluated by professional translators. We find that quality estimation using syntactic and semantic information on this data set can hardly improve over a baseline which uses only surface features. However, the performance can be improved when they are combined with such surface features. We also introduce a novel metric to measure translation adequacy based on predicate-argument structure match using word alignments. While word alignments can be reliably used, the two main factors affecting the performance of all semantic-based methods seems to be the low quality of semantic role labelling (especially on ill-formed text) and the lack of nominal predicate annotation.

1 Introduction

The problem of evaluating machine translation output without reference translations is called quality estimation (QE) and has recently been the centre of attention (Bojar et al., 2014) following the seminal work of Blatz et al. (2003). Most QE studies have focused on surface and language-model-based features of the source and target. The quality of translation is however closely related to the syntax and semantics of the languages, the former concerning fluency and the latter adequacy.

While there have been some attempts to utilize syntax in this task, semantics has been paid less attention. In this work, we aim to exploit both syntax and semantics in QE, with a particular focus on the latter. We use shallow semantic analysis obtained via semantic role labelling (SRL) and employ this information in QE in various ways including statistical learning using both tree kernels and hand-crafted features. We also design a QE metric which is based on the Predicate-Argument structure Match (*PAM*) between the source and its translation. The semantic-based system is then combined with the syntax-based system to evaluate the full power of structural linguistic information. We also combine this system with a baseline system consisting of effective surface features.

A second contribution of the paper is the release of a new data set for QE.[1] This data set comprises a set of 4.5K sentences chosen from customer support forum text. The machine translation of the sentences are not only evaluated in terms of adequacy and fluency, but also manually post-edited allowing various metrics of interest to be applied to measure different aspects of quality. All experiments are carried out on this data set.

The rest of the paper is organized as follows: after reviewing the related work, the data is described and the semantic role labelling approach is explained. The baseline is then introduced, followed by the experiments with tree kernels, hand-crafted features, the PAM metric and finally the combination of all methods. The paper ends with a summary and suggestions for future work.

2 Related Work

Syntax has been exploited in QE in various ways including tree kernels (Hardmeier et al., 2012; Kaljahi et al., 2013; Kaljahi et al., 2014b), parse probabilities and syntactic label frequency (Avramidis, 2012), parseability (Quirk, 2004) and POS n-gram scores (Specia and Giménez, 2010).

[1]The data will be made publicly available - see http://www.computing.dcu.ie/mt/confidentmt.html

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 67–77,
October 25, 2014, Doha, Qatar.

Turning to the role of semantic knowledge in QE and MT evaluation in general, Pighin and Màrquez (2011) propose a method for ranking two translation hypotheses that exploits the projection of SRL from a sentence to its translation using word alignments. They first project the SRL of a source corpus to its parallel corpus and then build two translation models: 1) translations of proposition labelling sequences in the source to its projection in the target and 2) translations of argument role fillers in the source to their counterparts in the target. The source SRL is then projected to its machine translation and the above models are forced to translate source proposition labelling sequences to the projected ones. Finally the confidence scores of these translations and their reachability are used to train a classifier which selects the better of the two translation hypotheses with an accuracy of 64%. Factors hindering their classifier are word alignment limitations and low SRL recall due to the lack of a verb or the loss of a predicate during translation.

In MT evaluation, where reference translations are available, Giménez and Màrquez (2007) use semantic roles in building several MT evaluation metrics which measure the full or partial lexical match between the fillers of same semantic roles in the hypothesis and translation, or simply the role label matches between them. They conclude that these features can only be useful in combination with other features and metrics reflecting different aspects of the quality.

Lo and Wu (2011) introduce HMEANT, a manual MT evaluation metric based on predicate-argument structure matching which involves two steps of human engagement: 1) semantic role annotation of the reference and machine translation, 2) evaluating the translation of predicates and arguments. The metric calculates the F_1 score of the semantic frame match between the reference and machine translation based on this evaluation. To keep the costs reasonable, the first step is carried out by amateur annotators who were minimally trained with a simplified list of 10 thematic roles. On a set of 40 examples, the metric is meta-evaluated in terms of correlation with human judgements of translation adequacy ranking, and a correlation as high as that of HTER is reported.

Lo et al. (2012) propose MEANT, a variant of HMEANT, which automatizes its manual steps using 1) automatic SRL systems for (only) verb predicates, 2) automatic alignment of predicates and their arguments in the reference and machine translation based on their lexical similarity. Once the predicates and arguments are aligned, their similarities are measured using a variety of methods such as cosine distance and even Meteor and BLEU. In computation of the final score, the similarity scores replace the counts of correct and partial translations used in HMEANT. This metric outperforms several automatic metrics including BLEU, Meteor and TER, but it significantly under-performs HMEANT and HTER. Further analysis shows that automatizing the second step does not affect the performance of MEANT. Therefore, it seems to be the lower accuracy of the semantic role labelling that is responsible.

Bojar and Wu (2012) identify a set of flaws with HMEANT and propose solutions for them. The most important problems stem from the superficial SRL annotation guidelines. These problems are exacerbated in MEANT due to the automatic nature of the two steps. More recently, Lo et al. (2014) extend MEANT to ranking translations without a reference by using phrase translation probabilities for aligning semantic role fillers of the source and its translation.

3 Data

We randomly select 4500 segments from a large collection of Symantec English Norton forum text.[2] In order to be independent of any one MT system, we translate these segments into French with the following three systems and randomly choose 1500 distinct segments from each.

- ACCEPT[3]: a phrase-based Moses system trained on training sets of WMT12 releases of Europarl and News Commentary plus Symantec translation memories
- SYSTRAN: a proprietary rule-based system augmented with domain-specific dictionaries
- Bing[4]: an online translation system

These translations are evaluated in two ways. The first method involves light post-editing by a professional human translator who is a native

[2] http://community.norton.com

[3] http://www.accept.unige.ch/Products/D_4_1_Baseline_MT_systems.pdf

[4] http://www.bing.com/translator(on24-Feb-2014)

	Adequacy	Fluency
5	All meaning	Flawless Language
4	Most of meaning	Good Language
3	Much of meaning	Non-native Language
2	Little meaning	Disfluent Language
1	None of meaning	Incomprehensible

Table 2: Adequacy/fluency score interpretation

	1-HTER	HBLEU	HMeteor	Adq	Flu
1-HTER	-	-	-	-	-
HBLEU	0.9111	-	-	-	-
HMeteor	0.9207	0.9314	-	-	-
Adq	0.6632	0.7049	0.6843	-	-
Flu	0.6447	0.7213	0.6652	0.8824	-

Table 3: Pearson r between pairs of metrics on the entire 4.5K data set

French speaker.[5] Each sentence translation is then scored against its post-edit using BLEU[6](Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Denkowski and Lavie, 2011), which are the most widely used MT evaluation metrics. Following Snover et al. (2006), we consider this way of scoring MT output to be a variation of *human-targeted* scoring, where no reference translation is provided to the post-editor, so we call them HBLEU, HTER and HMETEOR. The average scores for the entire data set together with their standard deviations are presented in Table 1.[7]

In the second method, we asked three professional translators, who are again native French speakers, to assess the quality of MT output in terms of adequacy and fluency in a 5-grade scale (LDC, 2002). The interpretation of the scores is given in Table 2. Each evaluator was given the entire data set for evaluation. We therefore collected three sets of scores and averaged them to obtain the final scores. The averages of these scores for the entire data set together with their standard deviations are presented in Table 1. To be easily comparable to human-targeted scores, we scale these scores to the [0,1] range, i.e. adequacy/fluency scores of 1 and 5 are mapped to 0 and 1 respectively and all the scores in between are accordingly scaled.

The average Kappa inter-annotator agreement for adequacy scores is 0.25 and for fluency scores 0.19. However, this measurement does not differentiate between small and large differences in agreement. In other words, the difference between scores of 5 and 4 is the same as the difference between 5 and 2. To account for this, we use weighted Kappa instead. Specifically, we consider two scores of difference 1 to represent 75% agreement instead of 100%. All the other differences are considered to be a disagreement. The average weighted Kappa computed in this way is 0.65 for adequacy and 0.63 for fluency. Though the weighting used is quite strict, the weighted Kappa values are in the substantial agreement range.

Once we have both human-targeted and manual evaluation scores together, it is interesting to know how they are correlated. We calculate the Pearson correlation coefficient r between each pair of the five scores and present them in Table 3. HBLEU has the highest correlation with both adequacy and fluency scores among the human-targeted metrics. HTER on the other hand has the lowest correlation. Moreover, HBLEU is more correlated with fluency than with adequacy which is the opposite to HMeteor. This is expected according to the definition of BLEU and Meteor. There is also a high correlation between adequacy and fluency scores. Although this could be related to the fact that both scores are from the same evaluators, it indicates that if either the fluency and adequacy of the MT output is low or high, the other tends to be the same.

The data is split into train, development and test sets of 3000, 500 and 1000 sentences respectively.

4 Semantic Role Labelling

The type of semantic information we use in this work is the predicate-argument structure or semantic role labelling of the sentence. This information needs to be extracted from both sides of the translation, i.e. English and French. Though the SRL of English has been well-studied (Màrquez et al., 2008) thanks to the existence of two major hand-crafted resources, namely FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005), French is one of the under-studied languages in

[5]The post-editing guidelines are based on the TAUS/CNGL guidelines for achieving "good enough" quality downloaded from https://evaluation.taus.net/images/stories/guidelines/taus-cngl-machine-translation-postediting-guidelines.pdf.

[6]Version 13a of MTEval script was used at the segment level which performs smoothing.

[7]Note that HTER scores have no upper limit and can be higher than 1 when the number of errors is higher than the segment length. In addition, the higher HTER indicates lower translation quality. To be comparable to the other scores, we cut-off them at 1 and convert to 1-HTER.

	1-HTER	HBLEU	HMeteor	Adequacy	Fluency
Average	0.6976	0.5517	0.7221	0.6230	0.4096
Standard Deviation	0.2446	0.2927	0.2129	0.2488	0.2780

Table 1: Average and standard deviation of the evaluation scores for the entire data set

this respect mainly due to a lack of such resources.

The only available gold standard resource is a small set of 1000 sentences taken from Europarl (Koehn, 2005) and manually annotated with Propbank verb predicates (van der Plas et al., 2010). van der Plas et al. (2011) attempt to tackle this scarcity by automatically projecting SRL from the English side of a large parallel corpus to its French side. Our preliminary experiments (Kaljahi et al., 2014a), however, show that SRL models trained on the small manually annotated corpus have a higher quality than ones trained on the much larger projected corpus. We therefore use the 1K gold standard set to train a French SRL model. For English, we use all the data provided in the CoNLL 2009 shared task (Hajič et al., 2009).

We use LTH (Björkelund et al., 2009), a dependency-based SRL system, for both the English and French data. This system was among the best performing systems in the CoNLL 2009 shared task and is straightforward to use. It comes with a set of features tuned for each shared task language (English, German, Japanese, Spanish, Catalan, Czech, Chinese). We compared the performance of the English and Spanish feature sets on French and chose the former due to its higher performance (by 1 F_1 point).

It should be noted that the English SRL data come with gold standard syntactic annotation. On the other hand, for our QE data set, such annotation is not available. Our preliminary experiments show that, since the SRL system heavily relies on syntactic features, the performance considerably drops when the syntactic annotation of the test data is obtained using a different parser than that of the training data. We therefore replace the parses of the training data with those obtained automatically by first parsing the data using the Lorg PCFG-LA parser[8] (Attia et al., 2010) and then converting them to dependencies using `Stanford` converter (de Marneffe and Manning, 2008). The POS tags are also replaced with those output by the parser. For the same reason, we replace the original POS tagging of the French 1K data with those obtained by the MElt tagger (Denis and Sagot, 2012).

The English SRL achieves 77.77 and 67.02 labelled F_1 points when trained only on the training section of PropBank and tested on the WSJ and Brown test sets respectively.[9] The French SRL is evaluated using 5-fold cross-validation on the 1K data set and obtains an F_1 average of 67.66. When applied to the QE data set, these models identify 9133, 8875 and 8795 propositions on its source side, post-edits and MT output respectively.

5 Baseline

We compare the results of our experiments to a baseline built using the 17 baseline features of the WMT QE shared task (Bojar et al., 2014). These features provide a strong baseline and have been used in all three years of the shared task. We use support vector regression implemented in the `SVMLight` toolkit[10] with Radial Basis Function (RBF) kernel to build this baseline. To extract these features, a parallel English-French corpus is required to build a lexical translation table using GIZA++ (Och and Ney, 2003). We use the Europarl English-French parallel corpus (Koehn, 2005) plus around 1M segments of Symantec translation memory.

Table 4 shows the performance of this system (`WMT17`) on the test set measured by Root Mean Square Error (RMSE) and Pearson correlation coefficient (`r`). We only report the results on predicting four of the metrics introduced above, omitting `HMeteor` due to space constraints. C and γ parameters are tuned on the development set with respect to `r`. The results show a significant difference between manual and human-targeted metric prediction. The higher `r` for the former suggests that the patterns of these scores are easier to learn. The RMSE seems to follow the standard deviation

[8] `https://github.com/CNGLdlab/LORG-Release`.

[9] Although the English SRL data are annotated for noun predicates as well as verb predicates, since the French data has only verb predicate annotations, we only consider verb predicates for English.

[10] http://svmlight.joachims.org/

of the scores as the same ranking is seen in both.

6 Tree Kernels

Tree kernels (Moschitti, 2006) have been successfully used in QE by Hardmeier et al. (2012) and in our previous work (Kaljahi et al., 2013; Kaljahi et al., 2014b), where syntactic trees are employed. Tree kernels eliminate the burden of manual feature engineering by efficiently utilizing all subtrees of a tree. We employ both syntactic and semantic information in learning quality scores, using the SVMLight-TK[11], a support vector machine (SVM) implementation of tree kernels.

We implement a syntactic tree kernel QE system with constituency and dependency trees of the source and target side, following our previous work (Kaljahi et al., 2013; Kaljahi et al., 2014b). The performance of this system (TKSyQE) is shown in Table 4. Unlike our previous results, where the syntax-based system significantly outperformed the WMT17 baseline, TKSyQE can only beat the baseline in HTER and fluency prediction, with neither difference being statistically significant and it is below the baseline for HBLEU and adequacy prediction.[12] It should be noted that in our previous work, a WMT News data set was used as the QE data set which, unlike our new data set, is well-formed and in the same domain as the parsers' training data. The discrepancy between our new and old results suggests that the performance is strongly dependent on the data set.

Unlike syntactic parsing, semantic role labelling does not produce a tree to be directly used in the tree kernel framework. There can be various ways to accomplish this goal. We first try a method inspired by the PAS format introduced by Moschitti et al. (2006). In this format, a fixed number of nodes are gathered under a dummy root node as slots of one predicate and 6 arguments of a proposition (one tree per predicate). Each node dominates an argument label or a dummy label for the predicate, which in turn dominates the POS tag of the argument or the predicate lemma. If a proposition has more than 6 arguments they are ignored, if it has fewer than 6 arguments, the extra slots are attached to a dummy null label. Note that these trees are derived from the dependency-based SRL of both the source and target side (Figure 1(a)). The results are shown in Table 4 (D-PAS). The performance is statistically significantly lower than the baseline.[13]

	1-HTER	HBLEU	Adq	Flu
	RMSE			
WMT17	0.2310	**0.2696**	**0.2219**	0.2469
TKSyQE	**0.2267**	0.2721	0.2258	**0.2431**
D-PAS	0.2489	0.2856	0.2423	0.2652
D-PST	0.2409	0.2815	0.2383	0.2606
C-PST	0.2400	0.2809	0.2410	0.2615
CD-PST	0.2394	0.2795	0.2373	0.2578
TKSSQE	0.2269	0.2722	0.2253	0.2425
	Pearson r			
WMT17	0.3661	**0.3806**	**0.4710**	0.4769
TKSyQE	**0.3693**	0.3559	0.4306	0.5013
D-PAS	0.1774	0.1843	0.2770	0.3252
D-PST	0.2136	0.2450	0.3169	0.3670
C-PST	0.2319	0.2541	0.2966	0.3616
CD-PST	0.2311	0.2714	0.3303	0.3923
TKSSQE	0.3682	0.3537	0.4351	**0.5046**

Table 4: RMSE and Pearson r of the 17 baseline features (WMT17) and tree kernel systems; TKSyQE: syntax-based tree kernels, D-PAS: dependency-based PAS tree kernels of Moschitti et al. (2006), D-PST, C-PST and CD-PST: dependency-based, constituency-based *proposition subtree* kernels and their combination, TKSSQE: syntactic-semantic tree kernels

In order to encode more information in the trees, we propose another format in which *proposition subtrees* (PST) of the sentence are gathered under a dummy root node. A dependency PST (Figure 1(b)) is formed by the predicate label under the root dominating its lemma and all its arguments roles. Each of these nodes in turn dominates three nodes: the argument word form (the predicate word form for the case of a predicate lemma), its syntactic dependency relation to its head and its POS tag. We preserve the order of arguments and predicate in the sentence.[14] This system is named D-PST in Table 4. Tree kernels in this format significantly outperform D-PAS. However, the performance is still far lower than the baseline.

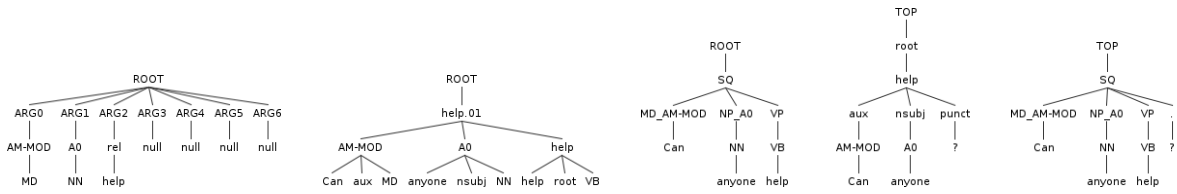
The above formats are based on dependency trees. We try another PST format derived from constituency trees. These PSTs (Figure 1(c)) are the lowest common subtrees spanning the predicate node and its argument nodes and are gathered under a dummy root node. The argument role

[11] http://disi.unitn.it/moschitti/Tree-Kernel.htm

[12] We use paired bootstrap resampling Koehn (2004) for statistical significance testing.

[13] Note that the only lexical information in this format is the predicate lemma. We tried replacing the POS tags with argument word forms, which led to a slight degradation.

[14] This format is chosen among several other variations due to its higher performance.



(a) D-PAS (b) D-PST (c) C-PST (d) D-TKSSQE (e) C-TKSSQE

Figure 1: Semantic tree kernel formats for the sentence: *Can anyone help?*

labels are concatenated with the syntactic non-terminal category of the argument node. Predicates are not marked. However, our dependency-based SRL is required to be converted into a constituency-based format. While constituency-to-dependency conversion is straightforward using head-finding rules (Surdeanu et al., 2008), the other way around is not. We therefore approximate the conversion using a heuristic we call (D2C).[15] As shown in Table 4, the system built using these PSTs C-PST improves over D-PST for human-targeted metric prediction, but not manual metric prediction. However, when they are combined in CD-PST, we can see improvement over the highest scores of both systems, except for HTER prediction for Pearson r. The fluency prediction improvement is statistically significant. The other changes are not statistically significant.

An alternative approach to formulating semantic tree kernels is to augment syntactic trees with semantic information. We augment the trees in TKSyQE with semantic role labels. We attach semantic roles to dependency labels of the argument nodes in the dependency trees as in Figure 1(d). For constituency trees, we use the D2C heuristic to elevate roles up the terminal nodes and attach the labels to the syntactic non-terminal category of the node as in Figure 1(e). The performance of the resulting system, TKSSQE, is shown in Table 4. It substantially outperforms its counterpart, CD-PST, all differences being statistically significant. However, compared to the plain syntactic tree kernels (TKSyQE), the changes are slight and inconsistent, rendering the augmentation not useful. We consider this system to be our syntactic-semantic tree kernel system.

[15]This heuristic (D2C) recursively elevates the argument role already assigned to a terminal node (based on the dependency-based argument position) to the parent node as long as 1) the argument node is not a root node or is not tagged as a POS (possessive), 2) the role is not an AM-NEG, AM-MOD or AM-DIS adjunct, and 3) the argument does not dominate its predicate's node or another argument node of the same proposition.

	1-HTER	HBLEU	Adq	Flu
	RMSE			
WMT17	**0.2310**	**0.2696**	**0.2219**	**0.2469**
HCSyQE	0.2435	0.2797	0.2334	0.2479
HCSeQE	0.2482	0.2868	0.2416	0.2612
	Pearson r			
WMT17	**0.3661**	**0.3806**	**0.4710**	**0.4769**
HCSyQE	0.2572	0.3080	0.3961	0.4696
HCSeQE	0.1794	0.1636	0.2972	0.3577

Table 5: RMSE and Pearson r of the 17 baseline features (WMT17) and hand-crafted features

7 Hand-crafted Features

In our previous work (Kaljahi et al., 2014b), we experiment with a set of hand-crafted syntactic features extracted from both constituency and dependency trees on a different data set. We apply the same feature set on the new data set here. The results are reported in Table 5. The performance of this system (HCSyQE) is significantly lower than the baseline. This is opposite to what we observe with the same feature set on a different data set, again showing that the role of data is fundamental in understanding system performance. The main difference between these two data sets is that the former is extracted from a well-formed text in the news domain, the same domain on which our parsers and SRL system have been trained, while the new data set does not necessarily contain well-formed text nor is it from the same domain.

We design another set of *feature types* aiming at capturing the semantics of the source and translation via predicate-argument structure. The feature types are listed in Table 6. Feature types 1 to 8 each contain two features, one extracted from the source and the other from the translation. To compute argument span sizes (feature types 4 and 5), we use the constituency conversion of SRL obtained using the D2C heuristic introduced in Section 6. The proposition label se-

1	Number of propositions
2	Number of arguments
3	Average number of arguments per proposition
4	Sum of span sizes of arguments
5	Ratio of sum of span sizes of arguments to sentence length
6	Proposition label sequences
7	Constituency label sequences of proposition elements
8	Dependency label sequences of proposition elements
9	Percentage of predicate/argument word alignment mapping types

Table 6: Semantic feature types

quence (feature type 6) is the concatenation of argument roles and predicate labels of the proposition with their preserved order (e.g. A0-go.01-A4). Similarly, constituency and dependency label sequences (feature types 4 and 5) are extracted by replacing argument and predicate labels with their constituency and dependency labels respectively. Feature type 9 consists of three features based on word alignment of source and target sentences: number of non-aligned, one-to-many-aligned and many-to-one-aligned predicates and arguments. The word alignments are obtained using the `grow-diag-final-and` heuristic as they performed slightly better than other types.[16]

As in the baseline system, we use SVMs to build the QE systems using these hand-crafted features. The nominal features are binarized to be usable by SVM. However, the set of possible feature values can be large, leading to a large number of binary features. For example, there are more than 5000 unique proposition label sequences in our data. Not only does this high dimensionality reduce the efficiency of the system, it can also affect its performance as these features are sparse. To tackle this issue, we impose a frequency cutoff on these features: we keep only frequent features using a threshold set empirically on the development set.

Table 5 shows the performance of the system (`HCSeQE`) built with these features. The semantic features perform substantially lower than the syntactic features and thus the baseline, especially in predicting human-targeted scores. Since these features are chosen from a comprehensive set of semantic features, and as they should ideally capture adequacy better than general features, a probable reason for their low performance is the quality of the underlying syntactic and semantic analysis.

8 Predicate-Argument Match (PAM)

Translation adequacy measures how much of the source meaning is preserved in the translated text. Predicate-argument structure or semantic role labelling expresses a substantial part of the meaning. Therefore, the matching between the predicate-argument structure of the source and its translation could be an important clue to the translation adequacy, independent of the language pair used. We attempt to exploit predicate-argument match (PAM) to create a metric that measures the translation adequacy.

The algorithm to compute PAM score starts by aligning the predicates and arguments of the source side to its target side using word alignments.[17] It then treats the problem as one of SRL scoring, similar to the scoring scheme used in the CoNLL 2009 shared task (Hajič et al., 2009). Assuming the source side SRL as a reference, it computes unlabelled precision and recall of the target side SRL with respect to it:

$$UPrec = \frac{\#\ aligned\ preds\ and\ their\ args}{\#\ target\ side\ preds\ and\ args}$$

$$URec = \frac{\#\ aligned\ preds\ and\ their\ args}{\#\ source\ side\ preds\ and\ args}$$

Labelled precision and recall are calculated in the same way except that they also require argument label agreement. $\texttt{UF}_1$ and $\texttt{LF}_1$ are the harmonic means of unlabelled and labelled scores respectively. Inspired by the observation that most source sentences with no identified proposition are short and can be assumed to be easier to translate, and based on experiments on the dev set, we assign a score of 1 to such sentences. When no proposition is identified in the target side while there is a proposition in the source, we assign a score of 0.5.

We obtain word alignments using the Moses toolkit (Hoang et al., 2009), which can generate alignments in both directions and combine them using a number of heuristics. We try intersection, union, source-to-target only, as well as the `grow-diag-final-and` heuristic, but only the source-to-target results are reported here as they slightly outperform the others.

Table 7 shows the RMSE and Pearson `r` for each of the unlabelled and labelled F_1 against ade-

[16] It should be noted that a number of features in addition to those presented here have been tried, e.g. the ratio and difference of the source and target values of numerical features. However, through manual feature selection, we have removed features which do not appear to contribute much.

[17] We also tried lexical and phrase translation tables for this purpose in addition to word alignments but they do not outperform word alignments.

	1-HTER	HBLEU	Adq	Flu
	RMSE			
1 UF_1	**0.3175**	**0.3607**	**0.3108**	0.4033
LF_1	0.4247	0.3903	0.3839	**0.3586**
	Pearson r			
UF_1	**0.2328**	**0.2179**	**0.2698**	**0.2865**
LF_1	0.1784	0.1835	0.2225	0.2688

Table 7: RMSE and Pearson r of PAM unlabelled and labelled F_1 scores as estimation of the MT evaluation metrics

	1-HTER	HBLEU	Adq	Flu
	RMSE			
PAM	**0.2414**	0.2833	0.2414	0.2661
HCSeQE	0.2482	0.2868	0.2416	0.2612
$HCSeQE_{pam}$	0.2445	**0.2822**	**0.2370**	**0.2575**
	Pearson r			
PAM	0.2292	0.2195	0.2787	0.3210
HCSeQE	0.1794	0.1636	0.2972	0.3577
$HCSeQE_{pam}$	**0.2387**	**0.2368**	**0.3571**	**0.3908**

Table 8: RMSE and Pearson r of PAM scores as features, alone and combined (PAM)

quacy and also fluency scores on the test data set.[18] According to the results, the unlabelled F_1 (UF_1) is a closer estimation than the labelled one. Its Pearson correlation scores are overall competitive to the hand-crafted semantic features (HCSeQE in Table 5): they are better for the automatic metric cases but lower for manual ones. However, the RMSE scores are considerably larger. Overall, the performance is not comparable to the baseline and other well performing systems. We investigate the reasons behind this result in the next section.

Another way to employ the PAM scores in QE is to use them in a statistical framework. We build a SVM model using all 6 PAM scores The performance of this system (PAM) on the test set is shown in Table 8. The performance is considerably higher than when the PAM scores are used directly as estimations. Interestingly, compared to the 47 semantic hand-crafted features (HCSeQE), this small feature set performs better in predicting human-targeted metrics.

We add these features to our set of hand-crafted features in Section 7 to yield a new system ($HCSeQE_{pam}$ in Table 8). All scores improve compared to the stronger of the two components. However, only the manual metric prediction improvements are statistically significant. The performance is still not close to the baseline.

[18]Precision and recall scores were also tried. Precision proved to be the weakest estimator, whereas recall scores were highest for some settings.

8.1 Analyzing PAM

Ideally, PAM scores should capture the adequacy of translation with a high accuracy. The results are however far from ideal. There are two factors involved in the PAM scoring procedure, the quality of which can affect its performance: 1) predicate-argument structure of the source and target side of the translation, 2) alignment of predicate-argument structures of source and target.

The SRL systems for both English and French are trained on edited newswire. On the other hand, our data is neither from the same domain nor edited. The problem is exacerbated on the translation target side, where our French SRL system is trained on only a small data set and applied to machine translation output. To discover the contribution of each of these factors in the accuracy of PAM, we carry out a manual analysis. We randomly select 10% of the development set (50 sentences) and count the number of problems of each of these two categories.

We find only 8 cases in which a wrong word alignment misleads PAM scoring. On the other hand, there are 219 cases of SRL problems, including predicate and argument identification and labelling: 82 cases (37%) in the source and 138 cases (63%) in the target.

We additionally look for the cases where a translation divergence causes predicate-argument mismatch in the source and translation. For example, *without sacrificing* is translated into *sans impact sur* (*without impact on*), a case of *transposition*, where the source side verb predicate is left unaligned thus affecting the PAM score. We find only 9 such cases in the sample, which is similar to the proportion of word alignment problems.

As mentioned in the previous section, PAM scoring has to assign default values for cases in which there is no predicate in the source or target. This can be another source of estimation error. In order to verify its effect, we find such cases in the development set and manually categorize them based on the reason causing the sentence to be left without predicates. There are 79 (16%) source and 96 (19%) target sentences for which the SRL systems do not identify any predicate, out of which 64 cases have both sides without any predicate. Among such source sentences, 20 (25%) have no predicate due to a predicate identification error of the SRL system, 57 (72%) because of the sentence structure (e.g. copula verbs which are not labelled

	1-HTER	HBLEU	Adq	Flu
	RMSE			
WMT17	0.2310	0.2696	0.2219	0.2469
SyQE	0.2255	0.2711	0.2248	0.2419
SeQE	0.2249	0.2710	0.2242	0.2404
SSQE	0.2246	0.2696	0.2230	0.2402
SSQE+WMT17	**0.2225**	**0.2673**	**0.2202**	**0.2379**
	Pearson r			
WMT17	0.3661	0.3806	0.4710	0.4769
SyQE	0.3824	0.3650	0.4393	0.5087
SeQE	0.3884	0.3648	0.4447	0.5182
SSQE	0.3920	0.3768	0.4538	0.5196
SSQE+WMT17	**0.4144**	**0.3953**	**0.4771**	**0.5331**

Table 9: RMSE and Pearson r of the 17 baseline features (WMT17) and system combinations

as predicates in the SRL training data, titles, etc.), and the remaining 2 due to spelling errors misleading the SRL system. Among the target side sentences, most of the cases are due to the sentence structure (65 or 68%) and only 14 (15%) cases are caused by an SRL error. In 13 cases, no verb predicate in the source is translated correctly. Among the remaining cases, two are due to untranslated spelling errors in the source and the other two due to tokenization errors misleading the SRL system.

These numbers show that the main reason leading to the sentences without verbal predicates is the sentence structure. This problem can be alleviated by employing nominal predicates in both sides. While this is possible for the English side, there is currently no French resource where nominal predicates have been annotated.

9 Combining Systems

We now combine the systems we have built so far (Table 9). We first combine syntax-based and semantic-based systems individually. SyQE is the combination of the syntactic tree kernel system (TKSyQE) and the hand-crafted features (HCSyQE). Likewise, SeQE is the combination of the semantic tree kernel system (TKSSQE) and the semantic hand-crafted features including PAM features (HCSeQE_{pam}). These two systems are combined in SSQE but without syntactic tree kernels (TKSyQE) to avoid redundancy with TKSSQE as these are the augmented syntactic tree kernels. We finally combine SSQE with the baseline.

SyQE significantly improves over its tree kernel and hand-crafted components. It also outperforms the baseline in HTER and fluency prediction, but is beaten by it in HBLEU and adequacy prediction. None of these differences are statistically significant however. SeQE also performs better than the stronger of its components. Except for adequacy prediction, the other improvements are statistically significant. This system performs slightly better than SyQE. Its comparison to the baseline is the same as that of SyQE, except that its superiority to the baseline in fluency prediction is statistically significant.

The full syntactic-semantic system (SSQE) also improves over its syntactic and semantic components. However, the improvements are not statistically significant. Compared to the baseline, HTER and fluency prediction perform better, the latter being statistically significant. HBLEU prediction is around the same as the baseline, but adequacy prediction performance is lower, though not statistically significantly.

Finally, when we combine the syntactic-semantic system with the baseline system, the combination continues to improve further. Compared to the stronger component however, only the HTER and fluency prediction improvements are statistically significant.

10 Conclusion

We introduced a new QE data set drawn from customer support forum text, machine translated and both post-edited and manually evaluated for adequacy and fluency. We used syntactic and semantic QE systems via both tree kernels and hand-crafted features. We found it hard to improve over a baseline, albeit strong, using such information which is extracted by applying parsers and semantic role labellers on out-of-domain and unedited text. We also defined a metric for estimating the translation adequacy based on predicate-argument structure match between source and target. This metric relies on automatic word alignments and semantic role labelling. We find that word alignment and translation divergence only have minor effects on the performance of this metric, whereas the quality of semantic role labelling is the main hindering factor. Another major issue affecting the performance of PAM is the unavailability of nominal predicate annotation.

Our PAM scoring method is based on only word matches as there are no constituent SRL resources available for French – perhaps constituent-based arguments can make a more accurate comparison between the source and target predicate-argument structure possible.

Acknowledgments

This research has been supported by the Irish Research Council Enterprise Partnership Scheme (EPSPG/2011/102) and the computing infrastructure of the CNGL at DCU. We thank the reviewers for their helpful comments.

References

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 the 1st Workshop on Statistical Parsing of Morphologically Rich Languages*.

Eleftherios Avramidis. 2012. Quality estimation for Machine Translation output using linguistic analysis and decoding features. In *Proceedings of the 7th WMT*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet project. In *Proceedings of the 36th ACL*.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for Machine Translation. In *JHU/CLSP Summer Workshop Final Report*.

Ondřej Bojar and Dekai Wu. 2012. Towards a predicate-argument evaluation for MT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on Statistical Machine Translation. In *Proceedings of the 9th WMT*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of the COLING Workshop on Cross-Framework and Cross-Domain Parser Evaluation*.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.*, 46(4):721–736.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th WMT*.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Tree kernels for machine translation quality estimation. In *Proceedings of the Seventh WMT*.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Rasoul Kaljahi, Jennifer Foster, Raphael Rubino, Johann Roturier, and Fred Hollowood. 2013. Parser accuracy in quality estimation of machine translation: a tree kernel approach. In *International Joint Conference on Natural Language Processing (IJCNLP)*.

Rasoul Kaljahi, Jennifer Foster, and Johann Roturier. 2014a. Semantic role labelling with minimal resources: Experiments with french. In *Third Joint Conference on Lexical and Computational Semantics (*SEM)*.

Rasoul Kaljahi, Jennifer Foster, Raphael Rubino, and Johann Roturier. 2014b. Quality estimation of english-french machine translation: A detailed study of the role of syntax. In *International Conference on Computational Linguistics (COLING)*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*.

LDC. 2002. Linguistic data annotation specification: Assessment of fluency and adequacy in chinese-english translations. Technical report.

Chi-kiu Lo and Dekai Wu. 2011. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of the Seventh WMT*.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. Xmeant: Better semantic mt evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, June.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: An introduction to the special issue. *Comput. Linguist.*, 34(2):145–159, June.

Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2006. Tree kernel engineering for proposition re-ranking. In *Proceedings of Mining and Learning with Graphs (MLG)*.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *Proceedings of EACL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311–318.

Daniele Pighin and Lluís Màrquez. 2011. Automatic projection of semantic structures: An application to pairwise translation ranking. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–9.

Chris Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of LREC*.

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*.

Lucia Specia and Jesús Giménez. 2010. Combining confidence estimation and reference-based metrics for segment level MT evaluation. In *Proceedings of AMTA*.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*.

Lonneke van der Plas, Tanja Samardžić, and Paola Merlo. 2010. Cross-lingual validity of propbank in the manual annotation of french. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.