Parser Accuracy in Quality Estimation of Machine Translation: A Tree Kernel Approach

Rasoul Samad Zadeh Kaljahi[†‡], **Jennifer Foster**[†], **Raphael Rubino**[†‡], **Johann Roturier**[‡] and **Fred Hollowood**[‡]
[†]NCLT, School of Computing, Dublin City University, Ireland
{rkaljahi, jfoster, rrubino}@computing.dcu.ie
[‡]Symantec Research Labs, Dublin, Ireland
{johann_roturier, fhollowood}@symantec.com

Abstract

We report on experiments designed to investigate the role of syntactic features in the task of quality estimation for machine translation, focusing on the effect of parser accuracy. Tree kernels are used to predict the segment-level BLEU score of English-French translations. In order to examine the effect of the accuracy of the parse tree on the accuracy of the quality estimation system, we experiment with various parsing systems which differ substantially with respect to their Parseval f-scores. We find that it makes very little difference which system we choose to use in the quality estimation task – this effect is particularly apparent for source-side English parse trees.

1 Introduction

Much research has been carried out on quality estimation (QE) for machine translation (MT) (Blatz et al., 2003; Ueffing et al., 2003; Specia et al., 2009; Callison-Burch et al., 2012), with the aim of solving the problem of how to accurately assess the quality of a translation without access to a reference translation. Approaches differ with respect to the nature of the quality scores being estimated (binary, 5-point or real-valued scales; human evaluations versus automatic metrics), the learning algorithms used or the feature set chosen to represent the translation pairs. The aspect of the task that we focus on is the feature set, and, in particular, the role of syntactic features. We ask the following: *To what extent is QE for MT influenced by the quality of the syntactic information provided to it? Does the accuracy of the parsing model used to provide the syntactic features influence the accuracy of the QE system?* We compare two pairs of parsing systems which differ with respect to their Parseval f-scores by around 17 absolute points in a QE system for English-French MT and find that it makes little difference which system we use.

2 Related Work

Features extracted from parser output have been used before in QE for MT. Quirk (2004) uses a feature which indicates whether a full parse for a sentence can be found. Gamon et al. (2005) use part-of-speech (POS) tag trigrams, CFG production rules and features derived from a dependency analysis of the MT output. Specia and Giménez (2010) use POS tag language model probabilities of the MT output 3-grams. Hardmeier et al. (2012) combine syntactic tree kernels with surface features to produce a system which was ranked second in the WMT 2012 shared task on QE for MT (Callison-Burch et al., 2012). Rubino et al. (2012) explore source syntactic features extracted from the output of a hand-crafted broad-coverage grammar/parser and a statistical constituency parser. Avramidis (2012) builds models for estimating post-editing effort using syntactic features such as parse probabilities and label frequency. Like Hardmeier et al. (2012), we use tree kernels to represent the output of a parser, but unlike all the previous works, we explicitly examine the role of parser accuracy.

There have been some attempts to investigate the role of parser accuracy in downstream applications. Johannson and Nugues (2007) introduce an English constituency-to-dependency converter and find that syntactic dependency trees produced using this converter help semantic role labelling more than dependency trees produced using an older converter despite the fact that trees produced using the older converter have higher attachment scores than trees produced using the new converter. Mollá and Hutchinson (2003) find significant differences between two dependency parsers in a task-based evaluation involving an answer extraction system but bigger differences be-

International Joint Conference on Natural Language Processing, pages 1092–1096,
Nagoya, Japan, 14-18 October 2013.

tween the two parsers when evaluated intrinsically. Quirk and Corston-Oliver (2006) demonstrate that a syntax-enhanced MT system is sensitive to a decrease in parser accuracy obtained by training the parser on smaller training sets. Zhang et al. (2010) experiment with a different syntax-enhanced MT system and do not observe the same behaviour. Both Miyao et al. (2008) and Goto et al. (2011) evaluate a suite of state-of-the-art English statistical parsers on the tasks of protein-pair interaction identification and patent translation respectively, and find only small (albeit sometimes statistically significant) differences between the parsing systems. Our study is closest to that of Quirk and Corston-Oliver (2006) since we are taking one parser and using it to train various models with different training set sizes.

3 Parsing

For parsing we use the LORG parser (Attia et al., 2010)[1] which learns a latent-variable probabilistic context-free grammar (PCFG-LA) from a treebank in an iterative process of splitting the treebank non-terminals, estimating probabilities for the new rules using Expectation Maximization and merging the less useful splits (Petrov et al., 2006), and which parses using the max-rule parsing algorithm (Petrov and Klein, 2007).

In order to investigate the effect of parsing accuracy, we train two parsing models – one "higher-accuracy" model and one "lower-accuracy" model – for each language. We use training set size to control the accuracy. For English, the higher-accuracy model is trained on Sections 2-21 of the Wall Street Journal (WSJ) section of the Penn Treebank (PTB) (Marcus et al., 1994) (approx 40k sentences). For French, the higher-accuracy model is trained on the training section of the French Treebank (FTB) (Abeillé et al., 2003) (approx 10k sentences). For the lower-accuracy models, we first select four random subsets of varying sizes from the larger training sets for each language[2] and measure the performance of the resulting models on the standard parsing test sets[3] using Parseval F_1 – see Table 1. All parsing models are trained with 5 split/merge cycles.

The worst-performing models for each language are those trained on 100 training sentences. However, these models fail to parse about 10 and 2 percent of our English and French data respectively. Since the failed sentences are not necessarily parallel in the source and translation sides, this could affect the downstream QE performance. Therefore, we opt to employ as our "lower-accuracy" models the second smallest training set sizes, which are 1K sentences for English and 500 for French. For both languages, the difference in F_1 between the lower-accuracy and higher-accuracy models is about 17 points. In order to measure how different the parses produced by these models are on our QE data, we compute their F_1 relative to each other. The F_1 for the English model pair is 71.50 and for French 63.19.

4 Quality Estimation

To minimise the effect of domain variation, we use a QE dataset for the domain on which our parsers have been trained (newswire). Since there are very few human QE evaluations available for English-French in this domain, we instead attempt to predict automatic metric scores. We experiment with BLEU, METEOR and TER, but due to space restrictions and the similar behaviour observed, we report only BLEU score predictions. We randomly select 4500 parallel segments from the News development data sets released for the WMT13 translation task.[4] To remain independent of any one MT system, we translate the dataset with the following three systems, randomly choosing 1500 distinct segments from each:

- ACCEPT[5]: a phrase-based Moses system trained on training sets of WMT12 releases of Europarl and News Commentary plus data from Translators Without Borders (TWB)
- SYSTRAN: a proprietary rule-based system
- Bing[6]: an online translation system

The translations are scored at the segment level using segment-level BLEU. The data set is randomly split into 3000 training, 500 development, and 1000 test segments. Model parameters are tuned using the development set.

We encode syntactic information using tree kernels (Collins and Duffy, 2002; Moschitti, 2006) because they allow us to use all subtrees of the

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

[2] Each smaller subset is contained in all the larger subsets.

[3] WSJ Section 23 and the FTB test set.

[4] http://www.statmt.org/wmt13

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

[6] http://www.bing.com/translator

	English					French				
Training size	100	**1K**	10K	20K	**40K**	100	**500**	2.5K	5K	**10K**
F_1	51.06	72.53	87.69	88.47	89.55	52.85	66.51	78.55	81.85	83.40

Table 1: Parser F_1s for various training set sizes: the sizes in bold are selected for the experiments.

parsed sentences as features in an efficient way, thus obviating the need for manual feature engineering. We use SVMLight-TK[7] (Moschitti, 2006), a support vector machine (SVM) implementation of tree kernels. The trees we use are constituency trees obtained by the parsing models described in Section 3, and their conversion to dependency trees using the Stanford converter for English (de Marneffe and Manning, 2008) and `Const2Dep` (Candito et al., 2010) for French. The labels must be removed from the arcs in the dependency trees before they can be used in SVMLight-TK – the nodes in the resulting tree representation are word forms and dependency relations, omitting part-of-speech tags.[8] Based on preliminary experiments on our development set, we use subset tree kernels.

We build a baseline system with features provided for the WMT 2012 QE shared task (Callison-Burch et al., 2012): we use Europarl v7 and News Commentary v8 (Koehn, 2005) to extract n-gram frequency, language model and word alignment features. This is considered a strong baseline as the system that used just these features was ranked higher than many of the other systems.

5 Experiments and Results

We build a QE system using constituency and dependency parse tree kernels of the source and translation sides, exploring first the higher-accuracy parse trees. Table 2 shows the performance of this system (*CD-ST$_H$*) compared to the system trained on the baseline features (*B-WMT*). We also compare to another baseline (*B-Mean*) which always predicts the mean of the segment-level BLEU scores of the training instances. We evaluate performance using Root Mean Square Error (RMSE) and Pearson correlation coefficient (r). To test the statistical significance of the performance differences (at $p < 0.05$), we use paired bootstrap resampling (Koehn, 2004).

CD-ST$_H$ achieves statistically significantly better RMSE and Pearson r than both baselines, which shows the usefulness of tree kernels in QE. We combine *CD-ST$_H$* and *B-WMT*[9] – this system *B+CD-ST$_H$* performs statistically significantly better than both systems individually, suggesting that tree kernels can also be useful in synergy with non-syntactic features.

	RMSE	**Pearson r**
B-Mean	0.1626	0
B-WMT	0.1601	0.1766
CD-ST$_H$	0.1581	0.2437
B+CD-ST$_H$	0.1570	0.2696

Table 2: Baselines, higher-accuracy parse tree kernels and combinations

We now investigate the impact of the intrinsic quality of the parse trees on the QE system. We build a similar model to *CD-ST$_H$* but with the lower-accuracy model described in Section 3. This system is named *CD-ST$_L$* in Table 3. *CD-ST$_H$* is also presented in this table for ease of comparison. Surprisingly, *CD-ST$_L$* performs only slightly lower than *CD-ST$_H$* and the difference is not statistically significant.

To better understand the behaviour of these systems, we break them down into their components: source constituency trees, target constituency trees, source dependency trees and target dependency trees. We first split based on the parse type and then based on the translation side.

C-ST$_H$ and *C-ST$_L$* in Table 3 are the systems with the constituency trees of both source and translation sides with higher- and lower-accuracy parsing models respectively. Although the difference is not statistically significant, the system with lower-accuracy parse trees achieves better scores than the system with higher-accuracy trees. *D-ST$_H$* and *D-ST$_L$* are built with the dependency trees of the higher- and lower-accuracy parsing models respectively. Unlike the constituency systems, the system with higher-accuracy parses performs better. However, the difference is not statistically significant. These results suggest that the intrinsic accuracy of neither the constituency

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

[8] A word is a child of its dependency relation to its head and this dependency relation is the child of the head word.

[9] The combination is carried out using vector summation.

nor the dependency parses is crucial to the performance of the QE systems. We now further split these systems based on the translation sides.

C-S_H and *C-S_L* use the higher- and lower-accuracy constituency trees of only the source side. Similar to when constituency trees of both sides were used (*C-ST_H* and *C-ST_L*), the system built on the lower-accuracy parses performs better although the difference is not statistically significant. The system using higher-accuracy constituency trees of the translation side (*C-T_H*) achieves better scores than the one using the lower-accuracy ones (C-T_L), but, again, this difference is not statistically significant.

D-S_H and *D-S_L* are the systems using the dependency trees of only the source side. Again, there is a small, statistically insignificant gap between the scores of these systems. On the other hand, there is a bigger performance gap between the systems built on the higher- and lower-accuracy dependency trees of the translation side: *D-T_H* and *D-T_L*. Although this is the only large difference observed among all settings, it is surprisingly not statistically significant.[10]

	RMSE	Pearson r
CD-ST_H	0.1581	0.2437
CD-ST_L	0.1583	0.2350
C-ST_H	0.1584	0.2307
C-ST_L	0.1582	0.2348
D-ST_H	0.1591	0.2103
D-ST_L	0.1597	0.1902
C-S_H	0.1583	0.2312
C-S_L	0.1582	0.2335
C-T_H	0.1608	0.1479
C-T_L	0.1616	0.1204
D-S_H	0.1598	0.1869
D-S_L	0.1601	0.1780
D-T_H	0.1598	0.2102
D-T_L	0.1604	0.1679

Table 3: QE systems with higher- and lower-accuracy trees (C: constituency, D: dependency, ST: Source and Translation, H: Higher-accuracy parsing model, L: Lower-accuracy parsing model)

One may argue that the way the parser accuracy is varied here could impact the results – a parser with similar F_1 but different output may lead to a different conclusion. It is possible to test this by using the parsing model from a lower split/merge (SM) cycle. For example, the models from the first SM cycle with a 10K training set size for English and a 2.5K training set size for French score 73.04 and 70.22 F_1 points on their respective test sets. While these scores are close to those of the lower-accuracy models used above, their outputs are different: the parses with the two lower-accuracy English models achieve only 66.46 F_1 against each other and with the two French ones 66.51 F_1. We use the parse trees of these alternative lower-accuracy parsing models to build a new QE system. The RMSE is 0.1585 and Pearson r is 0.2316. These scores are not statistically significantly different compared to *CD-ST_H*, strengthening our conclusion that intrinsic parse accuracy is not crucial for QE.

Another question is to what extent we require a linguistically realistic syntactic structure which retains some form of regularity no matter how accurate. To answer this question, we build random tree structures for source and translation segments. The random tree for a segment is generated by recursively splitting the sentence into random phrases and randomly assigning them a syntactic label.[11] We parse the source and translation segments using this method and build a QE system with the output trees. The RMSE and Pearson r are 0.1631 and -0.0588 respectively. This shows that tree kernels still require the regularity encoded in the lower- and higher-accuracy trees.

6 Conclusion

We explored the impact of parse quality in predicting automatic MT evaluation scores, comparing the use of constituency and dependency tree kernels built from the output of parsing systems with a large accuracy gap when measured using Parseval F_1. This large difference in F_1 did not have a knock-on effect on the QE task. Our next step is to carry out the experiments in the opposite direction (French-English) so that we better understand why the translation side trees were not as useful as the source side trees. Using other intrinsic parser evaluation metrics might also prove useful.

Acknowledgements

This research has been supported by the Irish Research Council Enterprise Partnership Scheme (EPSPG/2011/102 and EPSPD/2011/135) and the computing infrastructure of the Centre for Next Generation Localisation at Dublin City University.

[10] The high scores of *D-T_H* seem to be happening by chance, because on the development set, on which the parameters are tuned, the scores are much lower.

[11] The English random model achieves an F_1 of around 0.5 and the French model an F_1 of 0.2.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In Anne Abeille, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Springer.

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*.

Eleftherios Avramidis. 2012. Quality Estimation for Machine Translation Output Using Linguistic Analysis and Decoding Features. In *Proceedings of WMT*.

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*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of WMT*.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French Dependency Parsing: Treebank Conversion and First Results. In *Proceedings of LREC'2010*.

Michael Collins and Nigel Duffy. 2002. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In *Proceedings of ACL*.

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*.

Mollá Diego and Ben Hutchinson. 2003. Intrinsic versus Extrinsic Evaluation of Parsing Systems. In *Proceedings of EACL*.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-Level MT Evaluation without Reference Translations: Beyond Language Modeling. In *Proceedings of EAMT*.

Isao Goto, Masao Utiyama, Takashi Onishi, and Eiichiro Sumita. 2011. A Comparison Study of Parsers for Patent Machine Translation. In *Proceedings of MT Summit*.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Tree Kernels for Machine Translation Quality Estimation. In *Proceedings of WMT*.

Richard Johansson and Pierre Nugues. 2007. Extended Constituent-to-dependency Conversion for English. In *Proceedings of NODALIDA*.

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 *Proceedings of MT Summit*.

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 ARPA Speech and Natural Language Workshop*.

Yusuke Miyao, Rune Saetre, Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii. 2008. Task-oriented Evaluation of Syntactic Parsers and their Representations. In *Proceedings of ACL*.

Alessandro Moschitti. 2006. Making Tree Kernels Practical for Natural Language Learning. In *Proceedings of EACL)*.

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 COLING-ACL*.

Chris Quirk and Simon Corston-Oliver. 2006. The Impact of Parse Quality on Syntactically-informed Statistical Machine Translation. In *Proceedings of EMNLP*.

Chris Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of LREC*.

Raphael Rubino, Jennifer Foster, Joachim Wagner, Johann Roturier, Rasoul Kaljahi, and Fred. Hollowood. 2012. DCU-Symantec Submission for the WMT 2012 Quality Estimation Task. In *Proceedings of WMT*.

Lucia Specia and Jesús Giménez. 2010. Combining Confidence Estimation and Reference-based Metrics for Segment Level MT Evaluation. In *Proceedings of AMTA*.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the Sentence-level Quality of Machine Translation Systems. In *Proceedings of EAMT*.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence Measures for Statistical Machine Translation. In *Proceedings of MT Summit*.

Hao Zhang, Huizhen Wang, Tong Xiao, and Jingbo Zhu. 2010. The Impact of Parsing Accuracy on Syntax-based SMT. In *Proceedings of the International Conference on NLP-KE*.