Automatic Detection of Machine Translated Text and Translation Quality Estimation

Roee Aharoni
Dept. of Computer Science
Bar Ilan University
Ramat-Gan, Israel 52900
roee.aharoni@gmail.com

Moshe Koppel
Dept. of Computer Science
Bar Ilan University
Ramat-Gan, Israel 52900
moishk@gmail.com

Yoav Goldberg
Dept. of Computer Science
Bar Ilan University
Ramat-Gan, Israel 52900
yoav.goldberg@gmail.com

Abstract

We show that it is possible to automatically detect machine translated text at sentence level from monolingual corpora, using text classification methods. We show further that the accuracy with which a learned classifier can detect text as machine translated is strongly correlated with the translation quality of the machine translation system that generated it. Finally, we offer a generic machine translation quality estimation technique based on this approach, which does not require reference sentences.

1 Introduction

The recent success and proliferation of statistical machine translation (MT) systems raise a number of important questions. Prominent among these are how to evaluate the quality of such a system efficiently and how to detect the output of such systems (for example, to avoid using it circularly as input for refining MT systems).

In this paper, we will answer both these questions. First, we will show that using style-related linguistic features, such as frequencies of parts-of-speech n-grams and function words, it is possible to learn classifiers that distinguish machine-translated text from human-translated or native English text. While this is a straightforward and not entirely novel result, our main contribution is to relativize the result. We will see that the success of such classifiers are strongly correlated with the quality of the underlying machine translation system. Specifically, given a corpus consisting of both machine-translated English text (English being the target language) and native English text (not necessarily the reference translation of the machine-translated text), we measure the accuracy of the system in classifying the sentences in the corpus as machine-translated or not. This accuracy will be shown to decrease as the quality of the underlying MT system increases. In fact, the correlation is strong enough that we propose that this accuracy measure itself can be used as a measure of MT system quality, obviating the need for a reference corpus, as for example is necessary for BLEU (Papineni et al., 2001).

The paper is structured as follows: In the next section, we review previous related work. In the third section, we describe experiments regarding the detection of machine translation and in the fourth section we discuss the use of detection techniques as a machine translation quality estimation method. In the final section we offer conclusions and suggestions for future work.

2 Previous Work

2.1 Translationese

The special features of translated texts have been studied widely for many years. Attempts to define their characteristics, often called "Translation Universals", include (Toury, 1980; Blum-Kulka and Levenston, 1983; Baker, 1993; Gellerstam, 1986). The differences between native and translated texts found there go well beyond systematic translation errors and point to a distinct "Translationese" dialect.

Using automatic text classification methods in the field of translation studies had many use cases in recent years, mainly as an empirical method of measuring, proving or contradicting translation universals. Several works (Baroni and Bernardini, 2006; Kurokawa et al., 2009; Ilisei et al., 2010) used text classification techniques in order to distinguish human translated text from native language text at document or paragraph level, using features like word and POS n-grams, proportion of grammatical words in the text, nouns, finite verbs, auxiliary verbs, adjectives, adverbs, nu-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 289–295,
Baltimore, Maryland, USA, June 23-25 2014.

merals, pronouns, prepositions, determiners, conjunctions etc. Koppel and Ordan (2011) classified texts to original or translated, using a list of 300 function words taken from LIWC (Pennebaker et al., 2001) as features. Volanski et al. (2013) also tested various hypotheses regarding "Translationese", using 32 different linguistically-informed features, to assess the degree to which different sets of features can distinguish between translated and original texts.

2.2 Machine Translation Detection

Regarding the detection of machine translated text, Carter and Inkpen (2012) translated the Hansards of the 36th Parliament of Canada using the Microsoft Bing MT web service, and conducted three detection experiments at document level, using unigrams, average token length, and type-token ratio as features. Arase and Zhou (2013) trained a sentence-level classifier to distinguish machine translated text from human generated text on English and Japanese web-page corpora, translated by Google Translate, Bing and an in-house SMT system. They achieved very high detection accuracy using application-specific feature sets for this purpose, including indicators of the "Phrase Salad" (Lopez, 2008) phenomenon or "Gappy-Phrases" (Bansal et al., 2011).

While Arase and Zhou (2013) considered MT detection at sentence level, as we do in this paper, they did not study the correlation between the translation quality of the machine translated text and the ability to detect it. We show below that such detection is possible with very high accuracy only on low-quality translations. We examine this detection accuracy vs. quality correlation, with various MT systems, such as rule-based and statistical MT, both commercial and in-house, using various feature sets.

3 Detection Experiments

3.1 Features

We wish to distinguish machine translated English sentences from either human-translated sentences or native English sentences. Due to the sparseness of the data at the sentence level, we use common content-independent linguistic features for the classification task. Our features are binary, denoting the presence or absence of each of a set of part-of-speech n-grams acquired using the Stanford POS tagger (Toutanova et al., 2003), as well as the presence or absence of each of 467 function words taken from LIWC (Pennebaker et al., 2001). We consider only those entries that appear at least ten times in the entire corpus, in order to reduce sparsity in the data. As our learning algorithm we use SVM with sequential minimal optimization (SMO), taken from the WEKA machine learning toolkit (Hall et al., 2009).

3.2 Detecting Different MT Systems

In the first experiment set, we explore the ability to detect outputs of machine translated text from different MT systems, in an environment containing both human generated and machine translated text. For this task, we use a portion of the Canadian Hansard corpus (Germann, 2001), containing 48,914 parallel sentences from French to English. We translate the French portion of the corpus using several MT systems, respectively: Google Translate, Systran, and five other commercial MT systems available at the http://itranslate4.eu website, which enables to query example MT systems built by several european MT companies. After translating the sentences, we take 20,000 sentences from each engine output and conduct the detection experiment by labeling those sentences as MT sentences, and another 20,000 sentences, which are the human reference translations, labeled as reference sentences. We conduct a 10-fold cross-validation experiment on the entire 40,000 sentence corpus. We also conduct the same experiment using 20,000 random, non-reference sentences from the same corpus, instead of the reference sentences. Using simple linear regression, we also obtain an R^2 value (coefficient of determination) over the measurements of detection accuracy and BLEU score, for each of three feature set combinations (function words, POS tags and mixed) and the two data combinations (MT vs. reference and MT vs. non reference sentences). The detection and R^2 results are shown in Table 1.

As can be seen, best detection results are obtained using the full combined feature set. It can also be seen that, as might be expected, it is easier to distinguish machine-translated sentences from a non-reference set than from the reference set. In Figure 1, we show the relationship of the observed detection accuracy for each system with the BLEU score of that system. As is evident, regardless of the feature set or non-MT sentences used, the correlation between detection accuracy and BLEU

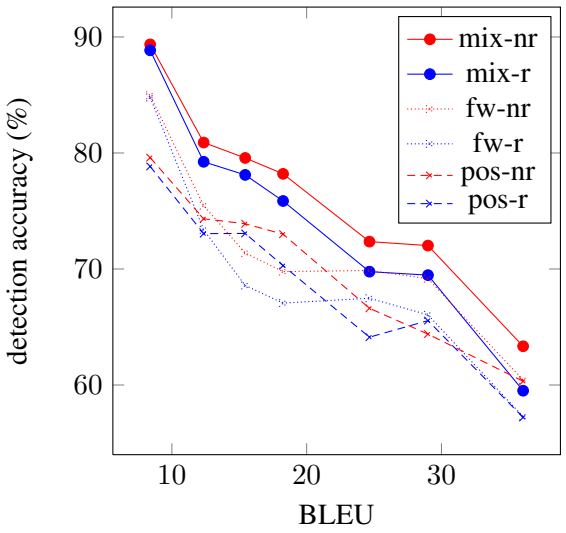

Figure 1: Correlation between detection accuracy and BLEU score on commercial MT systems, using POS, function words and mixed features against reference and non-reference sentences.

score is very high, as we can also see from the R^2 values in Table 1.

3.3 In-House SMT Systems

	Parallel	Monolingual	BLEU
SMT-1	2000k	2000k	28.54
SMT-2	1000k	1000k	27.76
SMT-3	500k	500k	29.18
SMT-4	100k	100k	23.83
SMT-5	50k	50k	24.34
SMT-6	25k	25k	22.46
SMT-7	10k	10k	20.72

Table 3: Details for Moses based SMT systems

In the second experiment set, we test our detection method on SMT systems we created, in which we have control over the training data and the expected overall relative translation quality. In order to do so, we use the Moses statistical machine translation toolkit (Koehn et al., 2007). To train the systems, we take a portion of the Europarl corpus (Koehn, 2005), creating 7 different SMT systems, each using a different amount of training data, for both the translation model and language model. We do this in order to create different quality translation systems, details of which are described in Table 3. For purposes of classification, we use the same content independent features as in the previous experiment, based on function words and POS tags, again with SMO-based SVM as the classifier. For data, we use 20,000 random, non reference sentences from the Hansard corpus, against 20,000 sentences from one MT system per experiment, again resulting in 40,000 sentence instances per experiment. The relationship between the detection results for each MT system and the BLEU score for that system, resulting in $R^2 = 0.774$, is shown in Figure 2.

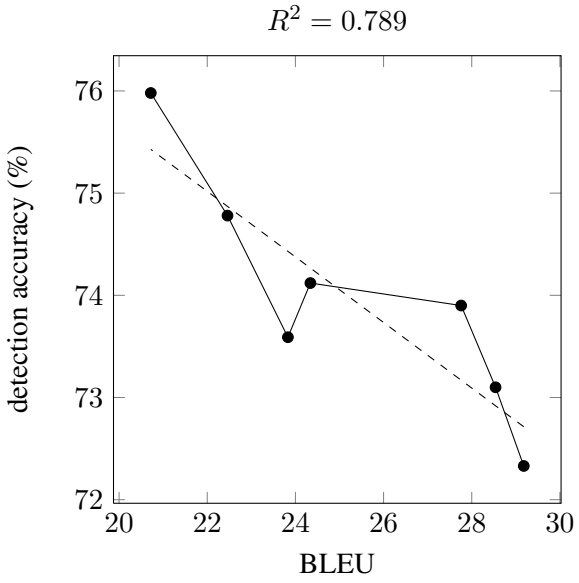

Figure 2: Correlation between detection accuracy and BLEU score on in-house Moses-based SMT systems against non-reference sentences using content independent features.

4 Machine Translation Evaluation

4.1 Human Evaluation Experiments

As can be seen in the above experiments, there is a strong correlation between the BLEU score and the MT detection accuracy of our method. In fact, results are linearly and negatively correlated with BLEU, as can be seen both on commercial systems and our in-house SMT systems. We also wish to consider the relationship between detection accuracy and a human quality estimation score. To do this, we use the French-English data from the 8th Workshop on Statistical Machine Translation - WMT13' (Bojar et al., 2013), containing outputs from 13 different MT systems and their human evaluations. We conduct the same classification experiment as above, with features based on function words and POS tags, and SMO-based SVM as the classifier. We first use 3000 refer-

Features	Data	Google	Moses	Systran	ProMT	Linguatec	Skycode	Trident	R^2
mixed	MT/non-ref	**63.34**	**72.02**	**72.36**	**78.2**	**79.57**	**80.9**	**89.36**	0.946
mixed	MT/ref	59.51	69.47	69.77	75.86	78.11	79.24	88.85	0.944
func. w.	MT/non-ref	60.43	69.17	69.87	69.78	71.38	75.46	84.97	0.798
func. w.	MT/ref	57.27	66.05	67.48	67.06	68.58	73.37	84.79	0.779
POS	MT/non-ref	60.32	64.39	66.61	73	73.9	74.33	79.6	**0.978**
POS	MT/ref	57.21	65.55	64.12	70.29	73.06	73.04	78.84	0.948

Table 1: Classifier performance, including the R^2 coefficient describing the correlation with BLEU.

MT Engine	Example
Google Translate	"**These** days, **all but one were** subject **to a** vote, **and all had a** direct link **to the** post September 11th."
Moses	"**these** days , **except one were the** subject **of a** vote , **and all had a** direct link **with the after** 11 September ."
Systran	"**From these** days, **all except one were the** object **of a** vote, **and all were** connected **a** direct link **with after** September 11th."
Linguatec	"**Of these** days, **all except one were** making **the** object **of a** vote **and all had a** straightforward tie **with after** September 11."
ProMT	"**These** days, **very** safe **one all** made object **a** vote, **and had a** direct link **with after** September 11th."
Trident	"**From these all** days, **except one** operated object voting, **and all had a** direct rope **with after** 11 septembre."
Skycode	"**In these** days, **all** safe **one** made **the** object **in a** vote **and all had a** direct connection **with him after** 11 **of** September."

Table 2: Outputs from several MT systems for the same source sentence (function words marked in bold)

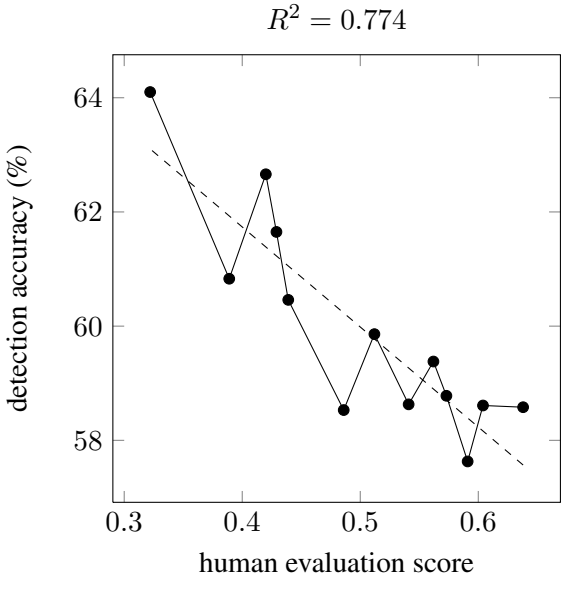

Figure 3: Correlation between detection accuracy and human evaluation scores on systems from WMT13' against reference sentences.

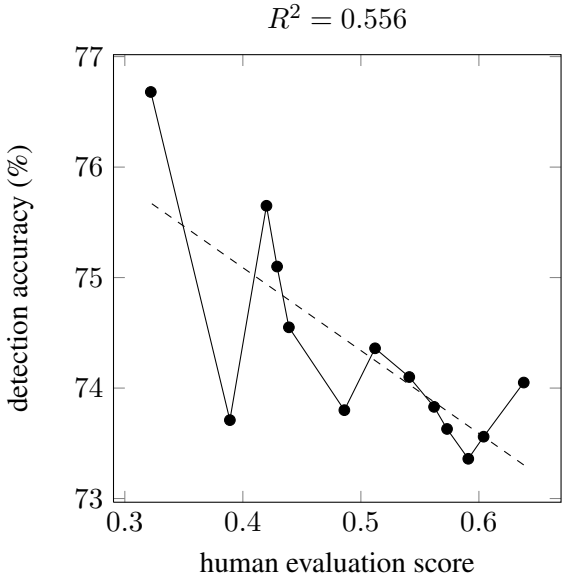

Figure 4: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences.

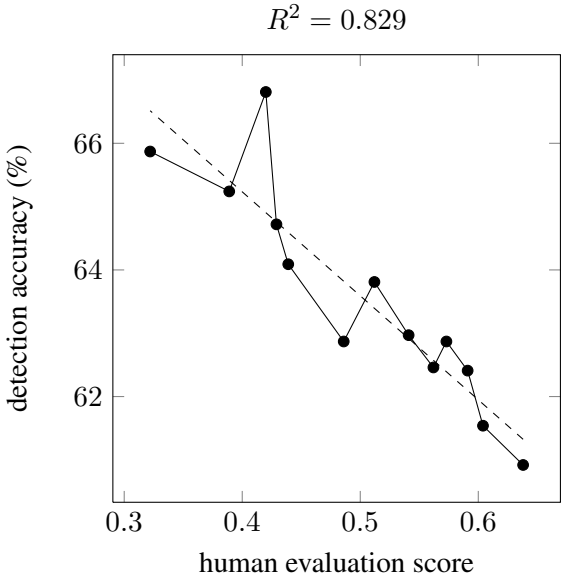

Figure 5: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using the syntactic CFG features described in section 4.2

ence sentences from the WMT13' English reference translations, against the matching 3000 output sentences from one MT system at a time, resulting in 6000 sentence instances per experiment. As can be seen in Figure 3, the detection accuracy is strongly correlated with the evaluations scores, yielding $R^2 = 0.774$. To provide another measure of correlation, we compared every pair of data points in the experiment to get the proportion of pairs ordered identically by the human evaluators and our method, with a result of 0.846 (66 of 78). In the second experiment, we use 3000 random, non reference sentences from the newstest 2011-2012 corpora published in WMT12' (Callison-Burch et al., 2012) against 3000 output sentences from one MT system at a time, again resulting in 6000 sentence instances per experiment. While applying the same classification method as with the reference sentences, the detection accuracy rises, while the correlation with the translation quality yields $R^2 = 0.556$, as can be seen in Figure 4. Here, the proportion of identically ordered pairs is 0.782 (61 of 78).

4.2 Syntactic Features

We note that the second leftmost point in Figures 3, 4 is an outlier: that is, our method has a hard time detecting sentences produced by this system although it is not highly rated by human evaluators. This point represents the Joshua (Post et al., 2013) SMT system. This system is syntax-based, which apparently confound our POS and FW-based classifier, despite it's low human evaluation score. We hypothesize that the use of syntax-based features might improve results. To verify this intuition, we create parse trees using the Berkeley parser (Petrov and Klein, 2007) and extract the one-level CFG rules as features. Again, we represent each sentence as a boolean vector, in which each entry represents the presence or absence of the CFG rule in the parse-tree of the sentence. Using these features alone, without the FW and POS tag based features presented above, we obtain an $R^2 = 0.829$ with a proportion of identically ordered pairs at 0.923 (72 of 78), as shown in Figure 5.

5 Discussion and Future Work

We have shown that it is possible to detect machine translation from monolingual corpora containing both machine translated text and human generated text, at sentence level. There is a strong correlation between the detection accuracy that can be obtained and the BLEU score or the human evaluation score of the machine translation itself. This correlation holds whether or not a reference set is used. This suggests that our method might be used as an unsupervised quality estimation method when no reference sentences are available, such as for resource-poor source languages. Further work might include applying our methods to other language pairs and domains, acquiring word-level quality estimation or integrating our method in a machine translation system. Furthermore, additional features and feature selection techniques can be applied, both for improving detection accuracy and for strengthening the correlation with human quality estimation.

Acknowledgments

We would like to thank Noam Ordan and Shuly Wintner for their help and feedback on the early stages of this work. This research was funded in part by the Intel Collaborative Research Institute for Computational Intelligence.

References

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1597–1607, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: in honour of John Sinclair*, 233:250.

Mohit Bansal, Chris Quirk, and Robert C. Moore. 2011. Gappy phrasal alignment by agreement. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 1308–1317. The Association for Computer Linguistics.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *LLC*, 21(3):259–274.

Shoshana Blum-Kulka and Eddie A. Levenston. 1983. Universals of lexical simplification. *Strategies in Interlanguage Communication*, pages 119–139.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

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 the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Dave Carter and Diana Inkpen. 2012. Searching for poor quality machine translated text: Learning the difference between human writing and machine translations. In Leila Kosseim and Diana Inkpen, editors, *Canadian Conference on AI*, volume 7310 of *Lecture Notes in Computer Science*, pages 49–60. Springer.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95.

Ulrich Germann. 2001. Aligned hansards of the 36th parliament of canada release 2001-1a.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *CICLing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 1318–1326. The Association for Computer Linguistics.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic Detection of Translated Text and its Impact on Machine Translation. In *Conference Proceedings: the twelvth Machine Translation Summit*.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical report, IBM Research Report.

J.W. Pennebaker, M.E. Francis, and R.J. Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, pages 404–411.

Marius Popescu. 2011. Studying translationese at the character level. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *RANLP*, pages 634–639. RANLP 2011 Organising Committee.

Matt Post, Juri Ganitkevitch, Luke Orland, Jonathan Weese, Yuan Cao, and Chris Callison-Burch. 2013. Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation, August 8-9, 2013.*, pages 206–212. Association for Computational Linguistics.

Gideon Toury. 1980. *In Search of a Theory of Translation*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *IN PROCEEDINGS OF HLT-NAACL*, pages 252–259.

Hans van Halteren. 2008. Source language markers in europarl translations. In Donia Scott and Hans Uszkoreit, editors, *COLING*, pages 937–944.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Literary and Linguistic Computing*.