Evaluation of MT Systems to Translate User Generated Content

Johann Roturier
Symantec Limited
Ballycoolin Business Park
Blanchardstown, Dublin 15, Ireland
johann_roturier@symantec.com

Anthony Bensadoun
Supélec
Paris
France
anthony.bensadoun@gmail.com

Abstract

This paper presents the evaluation results of a study conducted to determine the ability of various Machine Translation systems in translating User-Generated Content, particularly online forum content. Four systems are compared in this paper, focusing on the English>German and English>French language pairs, including a system called VICTOR, which is based on the Moses and IRSTLM toolkits. After describing some of the characteristics of these systems, the methodological framework used during a medium scale evaluation campaign is described. A careful analysis of both human and automated scores show that one system is overall significantly better than the other three systems for the English>German language pair, but that very little difference exists for specific post types (such as questions and solutions). The results are also much more balanced for the English>French language pair, suggesting that all systems could be useful in a multi-system deployment scenario. Our results also show that human scores and automated scores do not consistently correlate, penalizing certain systems more than others. Finally, we also show that the quality and coverage of the source posts impacts systems and language pairs differently.

1 Introduction

Software publishers used to rely on documentation sets (manuals) and online support (knowledge base) articles to assist their customers or users with the installation, maintenance or troubleshooting of products. With the advent of Web 2.0 communication channels (such as community forums or social media), users have become more active in the generation of documentation pertaining to software products. Conversations are now taking place on online forums, questions are being asked on micro-blogging platforms, and links (URLs) to blog posts containing solutions are being exchanged among savvy users. While non-technical users probably rely on alternative methods to find solutions to their problems (for example, by asking friends or family directly), savvy users can now complement their information search by taking part in online conversations.

These conversations take place in a number of environments, some of which are moderated and facilitated by software publishers. For example, Symantec started a forum[1] in 2008 to encourage Norton users to share their opinions about the products they own (including questions, answers, concerns and ideas). The initial forum was launched specifically for English-speaking users, but specific German, French, Japanese and Simplified Chinese were subsequently added to give an opportunity to global users to exchange in the same manner. One of the challenges with this type of approach, however, is that the fora are currently siloed (which means that French users cannot access the information posted by German users unless 1) they understand German or 2) rely on cross-lingual search and/or machine-translation techniques to make sense of this content). The present study tries to address the latter problem, by asking whether an MT solution could be of some use in a user-generated content scenario. This scenario poses very specific challenges due to the very nature of the content that should be translated:

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

- This content may be authored by non-professionals or people whose first language is not the language used in the forum (so its linguistic and technical accuracy may not be optimal).
- Although written, this content is similar to oral content, through "orthographic innovations that approximate characteristics of orality […]: commas appear where a pause or breath would occur in speech, and informal syntax and creative lexicon invoke spoken language and orthographic conventions." (Leblanc, 2005)
- Some of the content is authored by power users (or "techies") who "exhibit communicative techniques and practices that are guided by attitudes of technological elitism (Ibid)." These can include alternative spellings, acronyms, font change, color change, techie terms, emoticons, or representation of non-lexical speech sounds.
- This content is highly perishable (new comments are being added on a regular basis, so information may stop being relevant in a matter of minutes) and authored by a plethora of users (which increases the lexical and stylistic diversity of this content).

Bearing all of these challenges in mind, the present study attempts to answer the following questions: Do specific MT systems perform better than others when translating user-generated content? Are differences visible at the thread (conversation) level? Does the quality of the source have any impact on the translation results? And finally, are certain post types (such as questions and solutions) handled better by certain systems?

The structure of the paper is as follows. In Section 2 we review some of the work performed in the area of user-generated content translation and processing, as well as in the area of human evaluation for MT system comparison. In Section 3 the various systems used in this evaluation study are described, focusing on one of the custom systems, VICTOR. Section 4 presents the methodological framework used to conduct the evaluation while Section 5 presents the results of the evaluation. Finally conclusions are made in Section 6, which also indicates future research work.

2 Related Work

While the machine-translation of user-generated forum content has been identified as being potentially useful to allow communication between various user groups that do not share a common language (Flournoy and Rueppel, 2010), little research work has been performed in this area. Related work has been conducted in the area of chat translation (Flournoy and Callison-Burch, 2000), but the focus was on improving quality by leveraging Translation Memory technology and user feedback. Recent work has also been performed in the area of SMS normalization (Yvon, 2009) and chat normalization (Henríquez and Hernández, 2009) but results have not been directly applied to Machine Translation Evaluation.

On other hand, Machine Translation evaluation is an active field of research, with large scale evaluation campaigns being conducted on a regular basis. These campaigns tend to compare multiple MT systems using a ranking approach. For example, Callison-Burch et al. (2010) used Amazon Mechanical Turk to collect non-expert ratings, leveraging the same interface they used with traditional evaluators (2007). The present work differs from these previous studies because of the type of content being evaluated (user-generated forum content instead of news content). While news content has traditionally been used in the shared tasks of MT workshops on Statistical Machine Translation, this year's workshop features a task on SMS translation[2]. This confirms the trend outlined earlier in this section.

3 Systems Description

Four systems were used in this evaluation, focusing on the English>French (en-fr) and English>German (en-de) language pairs.

3.1 Commercial Systems

The first system is a freely available generic MT system: Microsoft Translator (which was accessed using the second version of their API[3]).

The second system is a customized version of SYSTRAN Enterprise Server 6[4] (the customization

[2] http://www.statmt.org/wmt11/
[3] http://msdn.microsoft.com/en-us/library/ff512419.aspx
[4] http://www.systransoft.com/translation-support/enterprise-server-6

being achieved with the use of 10K+ dictionary entries for the security and availability domains). As described in Roturier (2009), this system has been mostly used to translate structured content.

The third system is a third-party commercial SMT system that was customized using Symantec translation memories (up to ~2 million translation units) as well as monolingual forum data (up to ~40K German and ~20K French sentences). In the remainder of this paper, the following acronyms are used to refer to these systems: CSMT (Custom SMT system), MS2 (Microsoft Translator V2) and CSYS6 (Custom SYSTRAN Enterprise Server 6).

3.2 The VICTOR System

The fourth system is a standard phrase-based SMT system trained using the Moses (Koehn et al, 2007) and IRSTLM (Federico et al., 2008) toolkits and tuned using MERT on a development set (optimized in terms of BLEU scores). Trigram language models were created using Witten-Bell smoothing. This system was trained using the data described in Section 3.1 and supplemented by one million translation units from the Computer Software domain for the English>French language pair. These additional translation units were obtained from the TAUS Data Association[5]. Since this system is supplemented with two additional components, it is given its own name: VICTOR. The first custom component of this system is an enhanced tokenizer to make sure that certain entities do not get split (and broken) during the decoding phase (such as URLS, file paths and registry keys). The second component is a source text pre-processor that performs simple string replacements using regular expressions and a context-free dictionary lookup.

The lookup dictionary used by the pre-processor was created in the following manner. An open-source spell-checker (Mudge, 2009) was run on all tokens from legacy English forum data. Tokens that were flagged as mispelt words were then imported into a database and made accessible for human validation through a Web interface. A human reviewer was then asked to determine whether the spell-checker's most frequent flags (frequency of 4 or more in a corpus of 20-million words) were legitimate or false positives, with the possibility to skip ambiguous ones. To do, they were provided with sample sentences showing the context in which words occurred. When flags were legitimate, a preferred form had to be provided by the reviewer. In total 2931 items were identified with preferred forms, with some examples shown in Table 1:

Mispelt Word	Preferred Form
abou	about
aare	Are
Thxs	Thanks
Taks	Tasks
THAY	that
Symantecx	Symantec's
Suze	Suse
Suspecious	suspicious
Suscription	subscription

Table 1: Mispelt words and corresponding preferred forms

This section presented the four systems used in a user-generated content MT evaluation study, whose experimental design is discussed in Section 4.

4 Experimental Design

4.1 Evaluation Portal

To collect human ratings, we developed an online tool (called the Evaluation Portal). This portal has the following characteristics: It is Web-based, so designated users can access it remotely once they have registered (and specified which languages they are capable of evaluating). It supports the upload of XLIFF[6] files containing evaluations sets. Once the XLIFF files are uploaded into a project, a random evaluation task is generated every time an evaluator logs in.

An evaluation task corresponds to a set of documents that must be rated. In the present study, an evaluation task is based on a forum thread (which contains multiple posts that have been translated with one of the systems presented in Section 3). Depending on the languages specified during registration, users are presented with one of the following three tasks:

1. A monolingual evaluation task to rate the comprehensibility of the source posts.

[5] http://www.tausdata.org/

[6] http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html

2. A monolingual evaluation task to rate the comprehensibility of the translated posts.
3. A bilingual evaluation task to rate both the comprehensibility and fidelity of the translated posts.

Once a task is completed, the user is presented with another task, which will be different in terms of type (monolingual source, monolingual target or bilingual) and system (CSMT, MS2, CSYS6 or VICTOR).

4.2 Evaluation Criteria

Comprehensibility refers "the extent to which the text as a whole is easy to understand (…) and the extent to which valid inferences can be drawn by combining information from different parts of the document" (Hovy et al., 2002). It is evaluated in this study using a 1-5 point scale. Evaluators are instructed to pick the rating that corresponds best to the document (post) they have just read:

1. Hopelessly incomprehensible: It appears that no amount of study and reflection would reveal the thought of the document.

2. Generally incomprehensible: The document tends to read like nonsense, but with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the document.

3. Almost immediately comprehensible: The comprehension of the document is distinctly interfered with by poor style, poor word choice, alternative expressions, and incorrect grammatical arrangements.

4. Generally clear and comprehensible: The document is very clear but contains minor grammatical problems, and/or unusual word usage and/or wrong word order.

5. Perfectly clear and comprehensible: The document reads like ordinary text.

To evaluate the fidelity (or semantic accuracy) of the translation (Ibid), the following criteria are used:

False: The target document does not convey the meaning of the original document.

True: The target document conveys the meaning of the original document.

It was decided to use this true/false rating instead of a traditional 1-5 scale. This was done to avoid having to decide whether missing information (even if minimal) would impact monolingual users (which can be difficult to determine when the domain knowledge of these target users is unknown).

4.3 Evaluation Data

The evaluation data was harvested from the English Norton forum[7], by selecting threads from multiple boards. These threads were different from the legacy data described in Section 3. To make this sample as representative as possible, both solved and unsolved threads were selected. Solved threads contain a post that has been marked as a solution by one of the forum moderators. Table 2 shows a number of characteristics of the evaluation data (with HTML markup removed):

Number of threads	14
Number of posts	111
Average thread length (sentences/words)	36/697
Type/Token ratio	18.35% (1791/9761)

Table 2: Evaluation Set Characteristics

The evaluation set was translated using the four systems presented in Section 3, wrapped into an XLIFF format, and uploaded to the Evaluation Portal.

It was also decided to generate reference translations for the evaluation set. These were obtained by machine-translating the evaluation posts using an online system (different from the four systems used in this study) and asking translators to post-edit them using the TAUS/CNGL "Good enough" quality guidelines[8]. These reference translations were also uploaded to the Evaluation Portal in order to collect additional ratings.

These reference translations were used to score each evaluated system using the following automated metrics: BLEU (Papineni et al., 2002), GTM (Turian et al., 2003)[9], Meteor (Banerjee & Lavie, 2005) and TER (Snover et al., 2006).

In total, a maximum of 154 evaluation tasks had to be completed per user: 14 monolingual source tasks, 70 (14*5) monolingual target tasks and 70 (14*5) bilingual tasks. To perform the evaluation, the following evaluator profiles were considered:

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

[8] http://www.translationautomation.com/machine-translation-post-editing-guidelines.html

[9] An exponent of 1.2 is used.

1. Paid linguistic reviewers and translators (bilingual speakers)
2. Technical Support engineers (bilingual speakers)
3. Community members (monolingual speakers)

Specific persons were then contacted and asked to sign up on the Evaluation Portal. Once they were logged in, they gained access to online instructions detailing what was expected of them. These instructions (as well as the whole Web interface) were made available in their language of choice (English, French or German). Evaluators could complete several tasks in a row or log off whenever they wanted (by saving their progress). They were not instructed to complete all tasks in a specific timeframe, but were encouraged to complete as many tasks as possible. While most evaluators did not have a specific number of tasks to complete, we made sure that each thread was evaluated at least four times for each system and task configuration (monolingual source, monolingual target, and bilingual). In total, 13,953 post ratings were collected (90% of these ratings originated from linguists and 10% from technical support engineers). Figure 1 shows the Web interface that was used.

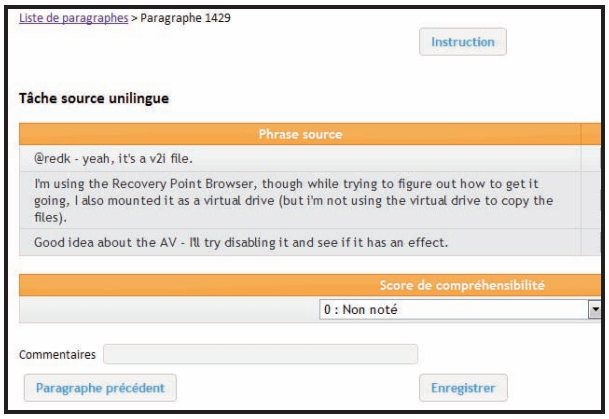

Figure 1: Evaluation Interface for the Monolingual Source Task Type

5 Results

5.1 Human Scores

We report the average post scores per evaluation category, language pair and system. In Table 3, and the remainder of this paper, Monolingual is shortened to Mono., Bilingual is shortened to Bil., Comprehensibility is shortened to Comp., and Fidelity is shortened to Fid.). Also, we introduce a combined score (Bil. CompFid.), which is the sum of the bilingual comprehensibility and fidelity scores. Apart from the scores given to the human translations, the highest scores are in bold.

en-de	Mono. Comp.	Bil. Comp.	Bil. Fid.	Bil. CompFid.
Human	4.63	4.68	0.98	5.67
CSMT	2.37	2.32	0.39	2.71
MS2	**2.60**	**2.63**	**0.55**	**3.18**
CSYS6	2.45	2.50	0.47	2.97
VICTOR	2.21	2.21	0.31	2.52
en-fr				
Human	4.43	4.37	0.97	5.34
CSMT	2.35	2.44	0.43	2.87
MS2	**2.45**	**2.48**	**0.45**	**2.93**
CSYS6	2.36	2.40	0.44	2.85
VICTOR	2.37	2.38	0.38	2.75

Table 3: Average Human Scores per Post and Evaluation Category

The first observation is that most systems fail to reach an average quality which makes posts "immediately comprehensible" (preserving the meaning of the original post in less than 50% of cases). The second observation is that the MS2 system has the highest average scores in all evaluation categories. To determine whether score differences between systems are statistically significant we use a t-Test (two-sample assuming unequal variances). For the en-de language pair, results are clear-cut in all categories. The best system is MS2, which is significantly better than CSYS6, which is in turn significantly better than CSMT, which is in turn significantly better than VICTOR. For the en-fr language pair, however, differences exist at the evaluation category level. While for the Mono. Comp. category, MS2 is significantly better than the other three systems, it is not significantly better than CSMT for the Bil. Comp. category. However, CSMT is not significantly better than CSYS6 and VICTOR. For the Bil. Fid. category, MS2, CSMT and CSYS6 are all significantly better than VICTOR. Finally, for the Bil. CompFid. category, only MS2 is significantly better than VICTOR.

While Table 3 reported overall average scores (regardless of the evaluator profile - linguist or technical support engineer), Table 4 shows averages scores per evaluator type.

en-de	Bil. Comp.	Fid.
Linguist	2.40	0.41
T.S. Engineer	2.55	0.63
en-fr	Bil. Comp.	Fid.
Linguist	2.36	0.39
T.S Engineer	2.72	0.72

Table 4: Average Scores per Evaluator Type

The differences in scoring are statistically significant, technical support engineers scoring generally higher than linguists. One of the main reasons for these differences lies in the way documents with low comprehensibility scores are perceived. For the en-fr language pair, when a linguist scores between 1 and 3 in comprehensibility, they are 67% likely to score the fidelity as 0, whereas a T.S engineer is only 23% likely to score the fidelity as 0. This suggests that T.S engineers see more value than linguists in documents whose linguistic form presents comprehensibility challenges.

5.2 Overall Automated Scores

Some of the results described in Section 5.1 are supported by the scores provided by automated metrics, as shown in Table 5, with system MS2 coming on top in both language pairs for most metrics.

en-de	BLEU	GTM	Meteor	TER
CSMT	0.32	0.39	0.28	0.60
MS2	**0.34**	**0.44**	**0.32**	**0.57**
CSYS6	0.31	0.41	0.27	0.63
VICTOR	0.30	0.39	0.24	0.65
en-fr	BLEU	GTM	Meteor	TER
CSMT	**0.40**	0.45	0.21	0.56
MS2	0.39	**0.46**	**0.21**	**0.54**
CSYS6	0.35	0.42	0.17	0.61
VICTOR	0.37	0.43	0.19	0.58

Table 5: Automated Scores for Whole Evaluation Set (en-fr)

However, the difference between the CSYS6 and CSMT systems observed in Section 5.1 for the en-de language pair is only mirrored by one metric, GTM (0.41 > 0.39). This is also the case for the en-fr language pair, where differences in human scores showed that CSYS6 was significantly better than VICTOR. However, all automatic metrics suggest the opposite in Table 5. These results confirm one of the conclusions from Callison-Burch et al. (2007), that the choice of automatic metrics can have a significant impact on comparing systems.

5.3 Thread-level Scores

In the remainder of this paper, we focus on the bilingual combined scores introduced in Section 5.1. The inconsistencies between the English>German and English>French results are confirmed when examining the correlation[10] between human and automated scores, using the average scores obtained for each of the 14 threads (instead of the average post scores used in Section 5.1). Table 6 shows that some systems (especially the rule-based CSYS6 system for the en-fr language pair) appear to be heavily penalized by automated scores (which confirms the findings from Callison-Burch et al., 2006).

en-de	BLEU	GTM	Meteor	TER
CSMT	0.64	0.35	0.60	-0.38
MS2	0.36	0.49	0.47	-0.46
CSYS6	0.43	0.35	0.22	-0.28
VICTOR	0.67	0.53	0.41	-0.47
en-fr	BLEU	GTM	Meteor	TER
CSMT	0.60	0.76	0.64	-0.69
MS2	0.17	0.51	0.35	-0.36
CSYS6	-0.11	-0.16	-0.14	0.26
VICTOR	0.39	0.31	0.46	-0.24

Table 6: Correlation between Combined Scores and Automated Scores

In order to find out whether certain threads are translated more effectively by certain systems, the approach used by Callison-Burch et al. (2010) is used. A win in a thread is obtained "when no other system is statistically significantly better at p-level≤0.1 in pairwise comparison". Pairwise com-

[10] Pearson correlation coefficients are used throughout this analysis.

parisons between systems are made using all of the combined scores collected for a given thread. The results are presented in Table 7:

Thread Wins	en-de	en-fr
CSMT	4	10
MS2	13	11
CSYS6	8	9
VICTOR	4	9

Table 7: Distribution of Thread Wins

These results confirm the picture that has emerged in the previous sub-sections: for the English>German language pair, MS2 is significantly better than the other three systems (apart from one thread where CSYS6 is actually the better system). For the English>French language pair, however, the results are mixed, each system managing to provide relatively good results for most of the threads. This finding also suggests that two ore more systems can be complementary: for example, the three threads where MS2 does not record a win (with average combined scores of 2.34, 2.625 and 2.5 respectively), higher quality is produced by CSMT (2.92), VICTOR (3.47) and CSYS6 (3.02) respectively.

5.4 Impact of Source on Scores

Even though the VICTOR system was equipped with the ability to perform string replacements to handle certain spelling mistakes, no replacement was actually made by the pre-processor when translating the evaluation set. This suggests that the quality of the evaluation content was quite high (from an orthographic perspective). This is confirmed by the relatively high comprehensibility scores attributed to the source posts (with an average score of 4.18). However, the source comprehensibility scores did not correlate consistently with the systems' combined scores. Interestingly, only the rules-based system (CSYS6) showed moderate (0.58) to strong (0.74) correlation for the en-fr and en-de language pairs respectively.

Finally, perplexity scores were also calculated for each thread after building language models using the source segments from the translation memories for each language pair. The perplexity scores obtained showed a negligible correlation with the Combined scores (0.10) for the English>French language pair but a moderate correlation for the English>German language pair (0.48) for the VICTOR system. The difference between the two language pairs may be due to the fact that different training resources were used (additional third-party translation units were used for the en-fr language pair, possibly adding translation alternatives which affected the consistency of the translation output). Further analysis at the segment level will be required to support this explanation. Moderate correlations can be also observed between the perplexity scores and the BLEU scores (0.48 and 0.67 for en-de and en-fr respectively).

5.5 Results per Post Type

As described in Section 3, some posts in a thread are more important than others because they contain the solution to a user's question or problem. Translating these posts accurately would therefore be an advantage. We focused on 14 initial posts (one in each thread) and 10 solution posts (not all threads have solutions). However, no statistical difference was found in any pairwise comparison using the combined scores at the thread-level for the en-fr language pair. For the en-de language pair, the difference between the averaged combined scores for MS2 (3.16) and VICTOR (2.56) was statistically significant for the solution posts. For the remainder of the posts (which would be of lesser value), similar pairwise comparisons were made, but none showed statistically significant difference for the English>French language pair. For the English>German language pair, however, the following comparisons showed wins for both MS2 (average 3.25) and CSYS6 (average 3.01) over both CSMT (average 2.66) and VICTOR (average 2.53).

6 Conclusion and Future Work

This paper presented the results of a study focusing on the evaluation of MT systems in a user-generated forum context. The lack of impact of VICTOR's pre-processor confirms the lexical diversity problem caused by user-generated content, as spelling mistakes seen in training data do not necessarily appear in evaluation. A more robust approach would be required to deal with these issues more effectively.

Our results also showed that overall, one of the systems (MS2) outperformed all systems for the English>German language pair, but that the differ-

ences with the CSMT and CSYS6 systems were not significant for high-value posts (questions and solutions). For the English>French language pair, however, results were more balanced, each system producing significantly better results for some of the threads.

These findings suggest that further work is required in source analysis to determine whether certain posts present linguistic characteristics that help MT systems perform better. The moderate correlation between the thread perplexity scores and the VICTOR combined scores for the en-de language pair also suggests that further confidence estimation work should be envisaged with a view to distribute the translation process across various MT systems.

Finally, while this study has provided some insight into the potential suitability of various MT systems for the translation of user-generated content, it should be supported by a pilot project with a view to collect actual user ratings. Such a project may help determine whether a specific usability threshold should be met before publishing machine-translated content.

Acknowledgements

The authors would like to thank Dr. Fred Hollowood, Robert Leyden and Amanda Henry for their help in making this evaluation study possible.

References

Banerjee, S., & Lavie, A. (2005). METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* (pp. 65-72).

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. *Proceedings of EACL* (pp. 249-256).

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) evaluation of machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07* (pp. 136-158). Morristown, NJ, USA.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., & Zaidan, O. F. (2010). Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. *Proceedings of Workshop on Statistical Machine Translation (WMT10).*

Federico, M., Bertoldi, N. and Cettolo, M. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. *Proceedings of Interspeech-2008* (pp.1618–1621). Brisbane, Australia.

Flournoy, R. S., & Callison-Burch, C. (2000). Reconciling User Expectations and Translation Technology to Create a Useful Real-World Application. *Proceedings of the Twenty-second International Conference on Translating and the Computer*.

Flournoy, R., & Rueppel, J. (2010). One Technology : Many Solutions. *Proceedings of AMTA 2010: the Ninth Conference of the Association for Machine Translation in the Americas*. Denver, Colorado.

Henríquez Q., Carlos A. Hernández, A. (2009). A Ngram-based Statistical Machine Translation Approach for Text Normalization on Chat-speak Style Communications. *Proceedings of the CAW2 Workshop*.

Hovy, E., M. King, and A. Popescu-Belis. (2002). Principles of Context-Based Machine Translation Evaluation. Machine Translation 17 (1), 43-75.

Koehn, P., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., et al. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. *ACL-2007: Proceedings of demo and poster sessions*. Prague, Czech Republic.

Leblanc TR. (2005). Is There A Translator In Teh House?": Cultural And Discourse Analysis Of A Virtual Speech Community On An Internet Message Board.

Mudge, Raphael S. (2009). After the Deadline – Language Checking Technology. Automattic. http://open.afterthedeadline.com.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* (pp. 311-318). Philadelphia, PA.

Roturier, J. (2009). Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. MT Summit XII: Proceedings of the twelfth Machine Translation Summit. Ottawa, Canada.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*. Cambridge, Massachusetts.

Turian, J., Shen, L., & Melamed, I. (2003). Evaluation of machine translation and its evaluation. Proceedings of MT Summit IX. New Orleans, LA.

Yvon, F. (2010). Rewriting the orthography of SMS messages. Natural Language Engineering, 16(02), 133-159.