The ACCEPT Portal: An Online Framework for the Pre-editing and Post-editing of User-Generated Content

Violeta Seretan
FTI/TIM
University of Geneva
Switzerland
Violeta.Seretan@unige.ch

Johann Roturier
Symantec Ltd.
Dublin, Ireland
johann_roturier@symantec.com

David Silva
Symantec Ltd.
Dublin, Ireland
David_Silva@symantec.com

Pierrette Bouillon
FTI/TIM
University of Geneva
Switzerland
Pierrette.Bouillon@unige.ch

Abstract

With the development of Web 2.0, a lot of content is nowadays generated online by users. Due to its characteristics (e.g., use of jargon and abbreviations, typos, grammatical and style errors), the user-generated content poses specific challenges to machine translation. This paper presents an online platform devoted to the pre-editing of user-generated content and its post-editing, two main types of human assistance strategies which are combined with domain adaptation and other techniques in order to improve the translation of this type of content. The platform has recently been released publicly and is being tested by two main types of user communities, namely, technical forum users and volunteer translators.

1 Introduction

User-generated content – i.e., information posted by Internet users in social communication channels like blogs, forum posts, social networks – is one of the main sources of information available today. Huge volumes of such content are created each day, reach a very broad audience instantly.[1]

The democratisation of content creation due to the emergence of the Web 2.0 paradigm also means a diversification of the languages used on the Internet.[2] Despite its availability, the new content is only accessible to the speakers of the language in which it was created. The automatic translation of user-generated content is therefore one of the key issues to be addressed in the field of human language technologies. However, as stated by Jiang et al. (2012), despite the obvious benefits, there are relatively little attempts at translating user-generated content.

The reason may lie in the fact that user-generated content is very challenging for machine translation. As shown, among others, by Nagarajan and Gamon (2011), there are several characteristics of this content that pose new processing challenges with respect to traditional content: informal style, slang, abbreviations, specific terminology, irregular grammar and spelling. Indeed, Internet users are rarely professional writers.[3] They often write in a language which is not their own, and sacrifice quality for speed, not paying attention to spelling, punctuation, or grammar rules.

The ACCEPT project[4] addresses these challenges by developing a technology integrating modules for automatic and manual content pre-editing, statistical machine translation, as well as output evaluation and post-editing. Thus, the project aims to improve the translation of user-generated content by proposing a full workflow, in which the participation of humans is essential.

The application scenario considered in the project are user communities sharing specific information on a given topic. The project focuses, more specifically, on the following use cases:

1. the commercial use case, in which the target community is the user community built around a software company in order for members to help each other with issues related to products;
2. the NGO use case, in which non-governmental organisations such as Doctors Without Borders produce health-care content for distributions in areas of need.

[1] For instance, 58 million tweets are sent on average per day (http://www.statisticbrain.com/twitter-statistics/).

[2] See http://en.wikipedia.org/wiki/Languages_used_on_the_Internet for statistics.

[3] Even when they are, as in the case of government agencies, the type of content produced (e.g., tweets) still poses "multiple challenges" to translation (Gotti et al., 2013).

[4] http://www.accept-project.eu/

Workshop on Humans and Computer-assisted Translation, pages 66–71,
Gothenburg, Sweden, 26 April 2014. ©2014 Association for Computational Linguistics

The language pairs considered in the project are English to French, German and Japanese, as well as French into English for the first use case (involving technical forum information), and French to and from English for the second use case (involving healthcare information).

Past halfway into its research program, the project has accomplished significant progress in the main areas mentioned above (pre-editing, statistical machine translation, post-editing, and evaluation). The ACCEPT technology has recently been released to the broad public as an online framework, which demonstrates the different modules of the workflow and provides access to associated software components (plug-ins, APIs), as well as to documentation. The pre-editing technology has been deployed on the targeted user forum[5], allowing users to check their messages before posting them. The post-editing technology is being used by a community of translators, which provide pro-bono translation services to the NGOs considered in our second use case.

In this paper, we describe the framework by presenting its architecture and main modules (Section 2). We discuss related work in Section 3 and conclude in Section 4.

2 The Framework

The ACCEPT technology has been made accessible to a broad audience in the form of an online framework, i.e., an integrated environment where registered users can perform pre-editing, post-editing and evaluation work. The framework – henceforth, the ACCEPT Portal – is hosted on a cloud computing infrastructure and is available at `www.accept-portal.eu`.

2.1 Architecture of the Framework

As explained in Section 1, the ACCEPT technology consists of the following main modules:

1. Pre-editing module;
2. Machine translation module,
3. Post-editing module,
4. Evaluation module.

The typical workflow is incremental, but the modules are independent. They can be used both within and outside the portal, as they are built on a REST API facilitating integration.

In the remaining of this section, we introduce each of the framework modules.[6]

2.2 Pre-editing Module

The pre-editing module leverages existing lingware which provides authoring support rules aimed at language professionals, by relying on shallow language processing (Bredenkamp et al., 2000). The existing English checker and the linguistic resources on which it relies have been extended and adapted to suit the type of data generated by community users. In particular, the software extension consisted of designing a number of pre-editing rules aimed at source normalisation, for the purpose of making the input text easier to handle by the SMT systems. In the case of French, the pre-editing rules have been designed from scratch. The pre-editing rules pertain to the levels of spelling, grammar, style and terminology. They are defined using the original lingware's rule formalism and are incorporated into a server dedicated to the project.

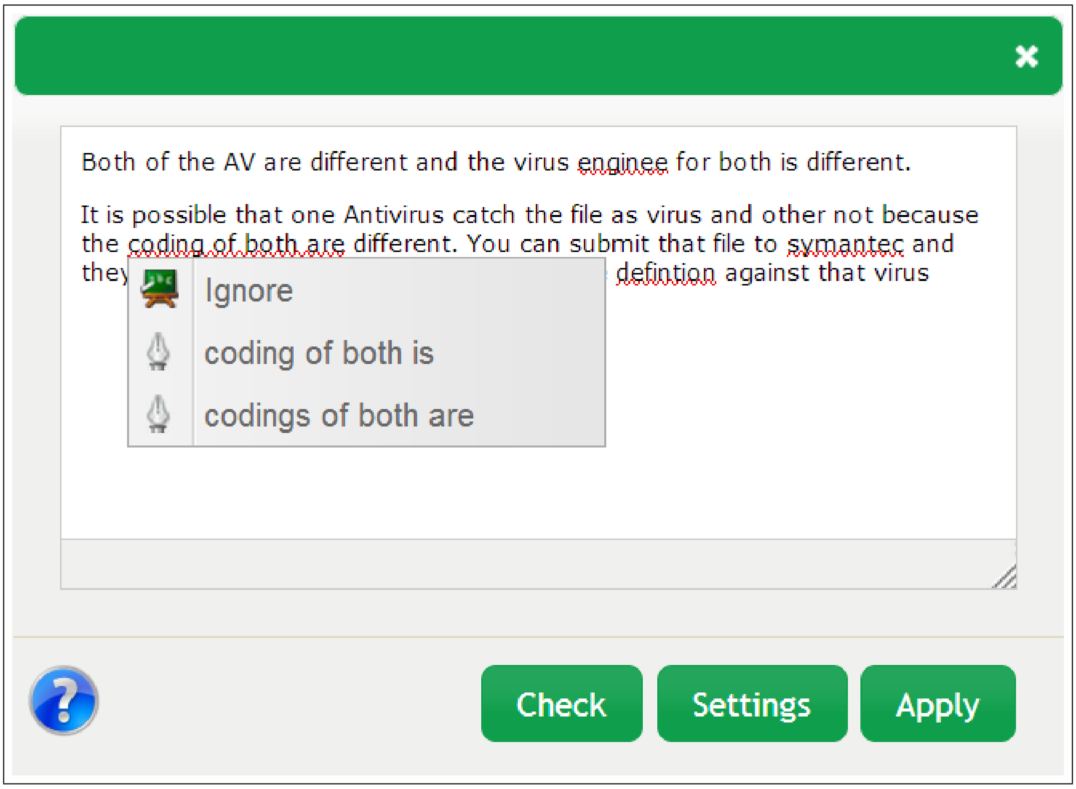
The rule development was corpus-driven and was performed on data collected for this purpose. A stable set of pre-edition rules is available in the portal for each of the domains and source languages considered (i.e., technical forum and heathcare data in English and French). The rules are described in detail in the project deliverable D 2.2 (2013).

The rules proposed have been evaluated individually and in combination (Roturier et al., 2012; Gerlach et al., 2013; Seretan et al., 2014). As a general observation, it is important to notice that, for SMT, the improvement of the input text does not go hand in hand with the improvement of translation. For example, in French the rule for correcting verbal forms to the subjunctive tense had a negative impact since the subjunctive is not frequent in the training data. Conversely, it was possible to define lexical reformulations which degraded the quality of the input text, but had a positive impact on translation quality.

The combined impact of the rule application was measured in a variety of settings in a large-scale evaluation campaign involving translation students (Seretan et al., 2014). As the rules are divided into two major groups, those automatically applicable and those requiring human inter-

[5] `https://community.norton.com/`

[6] The MT module will be omitted, as it is not part of the portal. The interested reader is referred to D 4.2 (2013).



Both of the AV are different and the virus enginee for both is different.

It is possible that one Antivirus catch the file as virus and other not because the coding of both are different. You can submit that file to symantec and they defintion against that virus

Figure 1: The ACCEPT Pre-edit plug-in in action (screen capture)

vention, the evaluation was carried out for the full set of rules, as well as for the automatic rules only. In addition, the evaluation was performed in both a monolingual and a bilingual setting, i.e., with the evaluators having or not access to the source text, and it involved evaluation scales of different granularities. The evaluation results showed a systematic statistically significant improvement over the baseline when pre-editing is performed on the source content. More details about the evaluation methodology and results can be found in the project deliverable D 9.2.2 (2013).

A data excerpt illustrating the impact of pre-editing on translation quality is presented in Example 1 below. The simple correction of an accented letter, *du* → *dû*, leads to the change of several target words, and to a much better translation of the input sentence.

1. a) Source (original):
   J'ai *du* m'absenter hier après midi.

   b) Source (pre-edited):
   J'ai *dû* m'absenter hier après midi.

   c) Target (original):
   I have the leave me yesterday afternoon.

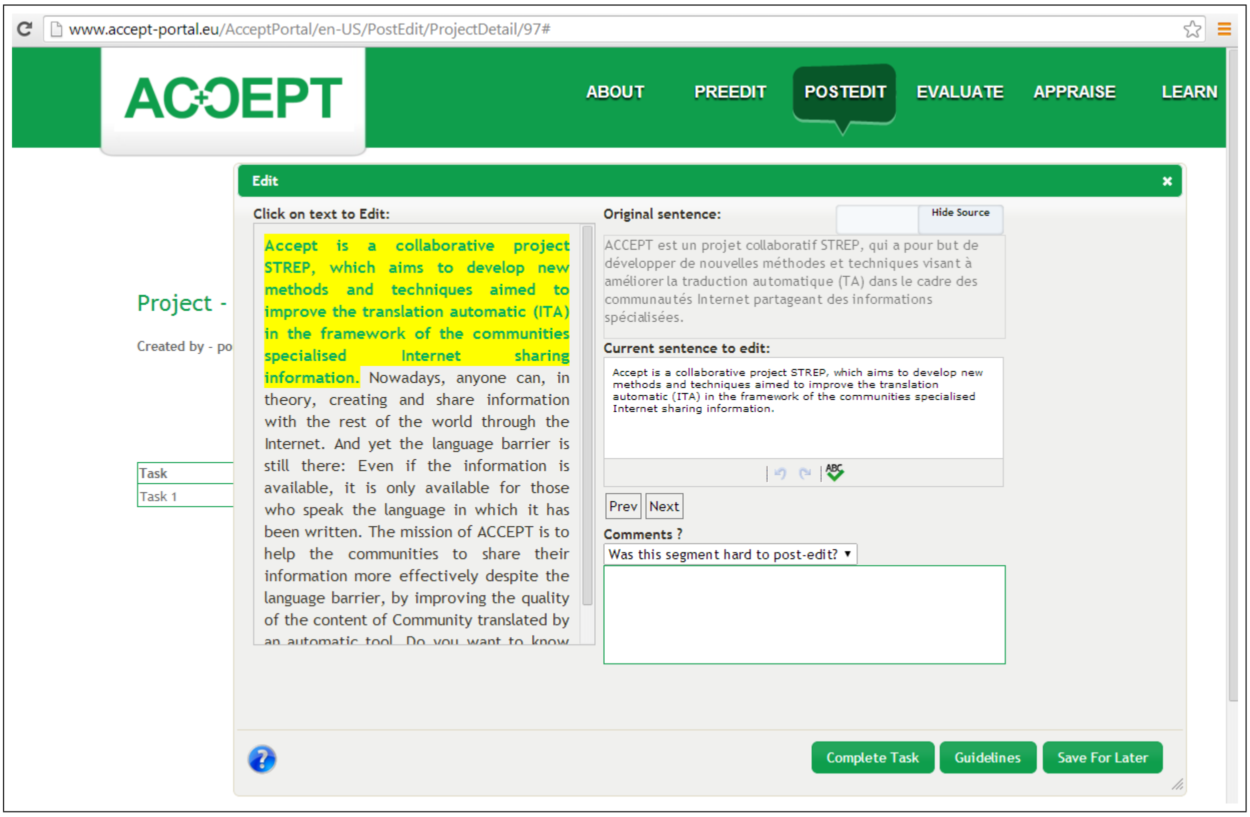
   d) Target (pre-edited):
   I had to leave yesterday afternoon.

The pre-editing component of the ACCEPT technology is available as a JQuery plug-in, which can be downloaded and installed by Web application owners, so that it can be used with text areas and other text-bearing elements. APIs and accompanying documentation have also been made available, so that the pre-editing rules can be leveraged in automatic steps, without the plug-in, across devices and platforms. A demo site illustrating the use of the plug-in in a TinyMCE environment is available on the portal (see Figure 1).

The latest developments of the pre-editing module include the possibility for users to customise the application of rule sets, in particular, to ignore specific rules and to manage their own dictionary, in order to prevent the activation of checking flags.

2.3 Post-editing Module

The post-editing module of the framework (see also Roturier et al., (2013)) is designed to fulfil the project's objective of collecting post-editing data in order to learn correction rules and, through feedback loops, to integrate them into the SMT engines (with the goal of automating corrections whenever possible). The project relies on the participation of volunteer community members, who are subject matter experts, native speakers of the



Click on text to Edit:

Accept is a collaborative project STREP, which aims to develop new methods and techniques aimed to improve the translation automatic (ITA) in the framework of the communities specialised Internet sharing information. Nowadays, anyone can, in theory, creating and share information with the rest of the world through the Internet. And yet the language barrier is still there: Even if the information is available, it is only available for those who speak the language in which it has been written. The mission of ACCEPT is to help the communities to share their information more effectively despite the language barrier, by improving the quality of the content of Community translated by an automatic tool. Do you want to know

Original sentence:

ACCEPT est un projet collaboratif STREP, qui a pour but de développer de nouvelles méthodes et techniques visant à améliorer la traduction automatique (TA) dans le cadre des communautés Internet partageant des informations spécialisées.

Current sentence to edit:

Accept is a collaborative project STREP, which aims to develop new methods and techniques aimed to improve the translation automatic (ITA) in the framework of the communities specialised Internet sharing information.

Comments ?

Was this segment hard to post-edit? ▾

Figure 2: The ACCEPT Portal showing the post-editing demo (screen capture)

target language and, possibly, of the source language. Accordingly, the post-editing environment (see Figure 2) provides functionalities for both monolingual and bilingual post-editing.

The post-editing text is organised in tasks belonging to post-editing projects. The latter are created and managed by project administrators, by defining the project settings (e.g., source and target languages, monolingual or bilingual mode, collaborative or non-collaborative type[7]), uploading the text for each task[8], inviting participants by e-mail, and monitoring revision progress.

The post-editors edit the target text in a sentence-by-sentence fashion. They have access to the task guidelines and to help documentation. The interface of the post-editing window displays the whole text, through which they can navigate with next-previous buttons or by clicking on a specific sentence. Users can check the text they are editing by accessing, with a button, the content checking technology described in Section 2.2. Their actions – in terms of keystrokes and usage of translation options – and time spent editing are recorded in the portal.[9] When they are done editing, they can click on a button marking the completion of the task. At any time, they can interrupt their work and save their results for later.

Users can enter a comment on the post-editing task they have performed. The feedback elicited from users include the difficulty of the task and their sentiment (*Was it easy to post-edit? Did you enjoy the post-editing task?*). For systematically collecting user feedback, the project administrators can specify on the project configuration page a link to a post-task survey, which will be sent to users after completing their tasks.

The post-editing module includes a JQuery plug-in for deployment in any Web-based environment; a dedicated section of the portal; APIs enabling the use of the post-editing functionality outside the portal; and sample evaluation projects for several language pairs.

The post-editing technology has been extensively used in specific post-editing campaigns involving translator volunteers and Amazon Mechanical Turk[10] workers. The campaigns, includ-

[7] In a collaborative editing scenario, users may see edits from other users and do not have to repeat them when working on the same project task. Conflicts are avoided by preventing concurrent access.

[8] Currently, the JSON format is used for the input data.

[9] The post-editing data is exported in XLIFF format.

[10] The integration was done via the ACCEPT API.

ing reports on post-task surveys, are documented *inter alia* in deliverable D 8.1.2 (2013). A notable finding was that professional translators, who were reticent towards MT before the task, had a more positive sentiment after post-editing and their motivation to post-edit in the future increased.

2.4 Evaluation Module

The role of the evaluation module is to support the collection of user ratings for assessing the quality of source, machine-translated and post-edited content, and, ultimately, to support the development of the technology created in the project.

This module groups several software components: an evaluation environment available as a section of the portal; APIs enabling the collection of user evaluations in-context; and a third component which is a customisation of the Appraise toolkit for the collaborative collection of human judgements (Federmann, 2012).

As in the case of post-editing module, this module provides functionality for creating and managing projects. Using the evaluation environment/APIs, project creators can define question categories, add questions and possible answers, and upload evaluation data (in JSON format). For traditional evaluation projects, the Appraise system is used instead.

3 Related Work

Transforming the source text in order to better fit the needs of machine translation is a well-investigated area of research. Strategies like source control, source re-ordering, or source simplification at the lexical or structural level have been largely explored; for reviews, see, for instance, Huhn (2013), Kazemi (2013), and Feng (2008), respectively.

User-generated content has been investigated in the context of machine translation in recent work dealing specifically with spelling correction (Bertoldi et al., 2010; Formiga and Fonollosa, 2012); lexical normalisation by substituting ill-formed words with their correct counterpart, e.g., *makn* → *making* (Han and Baldwin, 2011); missing word – e.g., zero-pronoun – recovery and punctuation correction (Wang and Ng, 2013).

Rather than focusing on specific phenomena or Web genres (i.e., tweets), we adopt a more general approach in which we address the problem of source normalisation at multiple levels – punctuation, spelling, grammar, and style – for any type of linguistically imperfect text.

Another peculiarity of our approach is that it is rule-based and does not require parallel data for learning corrections. In exchange, a limitation of our pre-editing approach is that it is language-dependent, as the underlying technology is based on shallow analysis and is therefore time-expensive to extend to a new language.

The post-editing technology differs from existing (standalone or Web-based) dedicated tools – e.g., iOmegaT[11] or MateCat[12] – in that it is tailored to community users, and, consequently, it is lighter, it generates more concise reports, and a simpler interface replaces the grid-like format for presenting data. Another specificity is that it is sufficiently flexible to be used in other environments (e.g., Amazon Mechanical Turk, cf. §2.3).

4 Conclusion

The technology outlined in this paper demonstrates a specific case of human-computer interaction, in which, for the first time, several modules are integrated in a full process in which human pre-editors, post-editors and evaluators play a key role for improving the translation of community content. The technology is freely accessible in the online portal, has been deployed on a major user forum, and can be downloaded for integration in other Web-based environments. Since it is built on top of a REST API, it is portable across devices and platforms. The technology would be useful to anyone who needs information instantly and reliably translated, despite linguistic imperfections.

One of the main future developments concerns the further improvement of SMT, by exploring, in particular, the use of text analytics and sentiment detection. In addition, by incorporating post-editing rules and developing techniques to change the phrase table and system parameters dynamically, it will be possible to reduce the amount of error corrections that human post-editors have to perform repeatedly. Another major development (joint work with the CASMACAT European project) will focus on novel types of assistance for translators, aimed specifically at helping translators by identifying problematic parts of the machine translation output and signalling the paraphrases that are more likely to be useful.

[11] http://try-and-see-mt.org/

[12] http://www.matecat.com/

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n^o 288769.

References

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2010. Statistical machine translation of texts with misspelled words. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 412–419, Los Angeles, California.

Andrew Bredenkamp, Berthold Crysmann, and Mirela Petrea. 2000. Looking for errors: A declarative formalism for resource-adaptive language checking. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece.

2013. ACCEPT deliverable D 2.2 Definition of pre-editing rules for English and French (final version). `http://www.accept.unige.ch/Products/D2_2_Definition_of_Pre-Editing_Rules_for_English_and_French_with_appendixes.pdf`.

2013. ACCEPT deliverable D 9.2.2: Survey of evaluation results. `http://www.accept.unige.ch/Products/D_9_2_Survey_of_evaluation_results.pdf`.

2013. ACCEPT deliverable D 4.2 Report on robust machine translation: domain adaptation and linguistic back-off. `http://www.accept.unige.ch/Products/D_4_2_Report_on_robust_machine_translation_domain_adaptation_and_linguistic_back-off.pdf`.

2013. ACCEPT deliverable D 8.1.2 Data and report from user studies - Year 2. `http://www.accept.unige.ch/Products/D_8_1_2_Data_and_report_from_user_studies_-_Year_2.pdf`.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 98:25–35.

Lijun Feng. 2008. Text simplification: A survey. Technical report, CUNY.

Lluís Formiga and José A. R. Fonollosa. 2012. Dealing with input noise in statistical machine translation. In *Proceedings of COLING 2012: Posters*, pages 319–328, Mumbai, India.

Johanna Gerlach, Victoria Porro, Pierrette Bouillon, and Sabine Lehmann. 2013. La préédition avec des règles peu coûteuses, utile pour la TA statistique des forums ? In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 539–546, Les Sables d'Olonne, France.

Fabrizio Gotti, Philippe Langlais, and Atefeh Farzindar. 2013. Translating government agencies' tweet feeds: Specificities, problems and (a few) solutions. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 80–89, Atlanta, Georgia.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon.

Jie Jiang, Andy Way, and Rejwanul Haque. 2012. Translating user-generated content in the social networking space. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2012)*, San Diego, California.

Arefeh Kazemi, Amirhassan Monadjemi, and Mohammadali Nematbakhsh. 2013. A quick review on reordering approaches in statistical machine translation systems. *IJCER*, 2(4).

Tobias Kuhn. 2013. A survey and classification of controlled natural languages. *Computational Linguistics*.

Meenakshi Nagarajan and Michael Gamon, editors. 2011. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Portland, Oregon.

Johann Roturier, Linda Mitchell, Robert Grabowski, and Melanie Siegel. 2012. Using automatic machine translation metrics to analyze the impact of source reformulations. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, California.

Johann Roturier, Linda Mitchell, and David Silva. 2013. The ACCEPT post-editing environment: a flexible and customisable online tool to perform and analyse machine translation post-editing. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice, France.

Violeta Seretan, Pierrette Bouillon, and Johanna Gerlach. 2014. A large-scale evaluation of pre-editing strategies for improving user-generated content translation. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

Pidong Wang and Hwee Tou Ng. 2013. A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 471–481, Atlanta, Georgia.